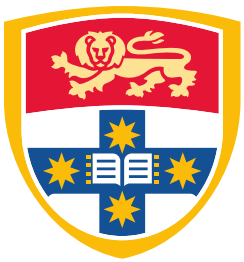
THE UNIVERSITY OF
SYDNEY

GROUNDING EVENT REFERENCES IN NEWS

Joel Nothman



THE UNIVERSITY OF
SYDNEY

A thesis submitted
in fulfilment of the requirements
for the degree of Doctor of Philosophy

School of Information Technologies
Faculty of Engineering & IT
The University of Sydney
2014

Abstract

Events in the world around us are frequently discussed in natural language, and their accurate identification is central to tasks from intelligence analytics to question answering. They are nonetheless very diverse, both in ontology and in how they are referred to, and have complex but ill-defined structures. For these reasons, computationally identifying and characterising events by how they are referred to proves very challenging. We establish this argument through a broad survey of computational tasks that identify and characterise references to events.

News and social media play an important role in informing the public of events as they occur, thus providing a shared foundation for communication. This work presents several studies into interpreting references to the breadth of events covered in news, before proposing and evaluating a new model for computationally grounding newsworthy event references.

We perform two annotation studies over broad-coverage news text, the first applying a coarse-grained event typology to mark event-referring sentences. Its results concur with our analysis of inter-annotator agreement and type distribution in the ACE05 corpus (Walker et al., 2006): both highlight the brittleness of such type schemas which leave a long, heavy tail of news events unmarked. The second annotation introduces a new approach to event typology, which employs a hierarchy of types that is extended over the course of annotation. This yields more complete coverage of newsworthy events, but suggests that there is no definitive means of structuring the space of events.

The second annotation is also novel in characterising each news report in terms of its *update event* and *topic event*, focusing only on the notable event content of news. We find that topic is too ambiguously characterised as a single event, suggesting that we should instead consider the explicit references to related events that a journalist provides as background.

Despite Wikipedia's utility for processing entity references, an analysis of its event articles shows that they are unsuitable for informing news event detection and disambiguation: they are strongly biased towards those that are enumerable (e.g. sports series), or that are collections of newsworthy sub-events (e.g. Vietnam War); their distribution, granularity and referential forms are mismatched to news media's event coverage.

In this context, we propose the *event linking* task. By analogy with named entity linking or disambiguation (Bunescu and Paşca, 2006), event linking models the grounding of references to notable events. It treats a news archive as a proxy for the set of events it reports, and defines the disambiguation of a newsworthy event reference as a link to the article that first reports it.

When two references are linked to the same article, they need not be references to the same event. We argue that precise event coreference is often too strict: it does not account for the intricate structure of events, nor the flexibility of referential language. By considering events at the granularity in which they are reported, we hope to provide a more intuitive

approximation to coreference, erring on the side of over-generation in contrast with the literature. When compared with other work in cross-document event coreference (Allan, 2002; Lee et al., 2012), event linking is also distinguished in considering a breadth of event references from multiple perspectives spanning a long period of time.

We perform a diagnostic evaluation of the task by first annotating a corpus of event links: references to past, newsworthy events are identified in a sample of news and opinion pieces and linked to an archive of the Sydney Morning Herald spanning 24 years. To perform this annotation, we employ non-expert annotators hired through an online freelancer marketplace, and present what is to our knowledge the first discussion of this mode of outsourced linguistic annotation in comparison to traditional expert annotation and popular crowdsourcing methods.

The intensive nature of our annotation task results in only a small corpus of 150 documents with 229 distinct links. However, we observe that a number of hyperlinks targeting online news correspond to event links. We thus acquire two large corpora of hyperlinks at very low cost – one of links internal to Fairfax Digital news sites; the other, Wikipedia's citations of news – and apply minimal, heuristic filtering. From these we learn weights for temporal and term overlap features in a link candidate generation system. These noisy sources of event linking knowledge lead to significant performance gains over a bag-of-words baseline. While our initial system can accurately predict many manual event links, a larger portion of references will require deep linguistic processing for their disambiguation.

Acknowledgements

In the years leading up to this point at which I find my *i*'s sufficiently dotted and my *t*'s sufficiently crossed, many people have helped make that a possibility. I hereby praise them.

The work was financially supported by a University of Sydney Vice Chancellor's Scholarship, and by a Capital Markets CRC PhD scholarship.[1] The latter had a far-reaching impact on the work, in that it drove our partnership with Fairfax Media that in turn framed directions for research, while creating a lot of additional applied work. Although impeding at times, this collaboration was a valuable experience. Much of my gratitude goes to the Computable News team who went the distance with me. I save especial praise for Will Radford, whose generosity with his time and sense of adventure really buoyed the project (and the lab), while keeping it real. I also highlight James, Will Cannings and Candice Loxley, who regularly dealt with the messy stuff in that collaboration, clearing the way for research. The project also depended on George Wright, who I thank for championing his Fizzing Panda within Fairfax, and introducing us to a variety of language-related real-world problems.

There were a lot of dead ends before the event linking idea got rolling. I particularly want to thank Matt Honnibal for managing much of the corpus annotation and for claiming (perhaps prematurely) that working on this new task was *even enjoyable*, particularly when compared to preceding event annotation slog. Thanks also to Ben Hachey who was co-supervising my work at the time, and providing an alternative perspective. My work rests on that of my annotators: within ə-lab, that includes particular contributions from Matt, Ben and David Vadas; otherwise, Eleanor Robertson, Jonathan Thambyrajah, and especially Kate Cousino who was a diamond in the Freelancer rough. Cheers to Richard Billingsley for ensuring my pronumerals add up.

Overall, ə-lab has been a warm, supportive and engaging research environment and community, and I thank all its members (and some adjuncts in SIT 4E) for numerous conversations on and off topic, donations of chocolate, etc. I note Tim Dawborn and Will Radford's unofficial roles in keeping the lab's machinery well-oiled, and various meeting *maîtres d'* for ensuring the same of our minds. I am grateful to many lab members for reviewing chapters of this work.

I attribute a lot of my interest in language to Mum, and to the influence of my high-school maths teacher, the late James Taylor; Dad I thank for his encouragement in computing, from my first loaning a Make-Your-Own-Computer-Games book from Waverley Library as a small child. But James Curran has spent the last decade nurturing my love of research and of education, my diverse interest in computational linguistics, and my passion for programming, particularly with Python which he forced upon me at our first encounter in 2004. Thank you for establishing this vibrant research environment, using your keen eye for talent; for finding

[1] Yet I acknowledge that without the priveleges afforded me by my parents and ancestors, and indeed the caretakers of this land, I would have unlikely got this far.

the right words to prod or lift me; for always keeping the end in sight.

A few seasoned academics have advised or influenced me at key points along the way – among them are Steven Bird, Alan Fekete and Jon Patrick. Their wisdom is keenly appreciated and fondly recalled, even though I have not always heeded it.

Hearing *I'm busy; just wait a moment* must get very tiring when that moment is an ever-lengthening doctorate. Many thanks to my friends and family who have patiently supported and endured me over these five years. Jenny, who I certainly don't praise enough, bore the brunt more than any other, and provided me with companionship, love, food, counselling, home management, and many other good things along the way. Plus, she has often brought me rejuvinating fresh air from outside the Ivory Tower, in the form of music, animals, dinner guests, a garden, getaways, and our most enthralling daughter, Kinneret.

— Joel Nothman, 7 May, 2014

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

(*Joel Nothman*)

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

> It is not incumbent upon you
> to complete the work,
> but neither are you free
> to ignore it.
>
> Attributed to Tarfon (Avot 2:21)

Natural language processing (NLP) models aspects of linguistic communication so that computational systems may understand texts as a competent human communicator does. This involves decoding not only the syntax and semantics of a text, but also its *reference*: what entities – and what types of entities – are being discussed? What does the text predicate or assume about the interaction and attributes of these discourse entities? Events play a focal role in discourse both by (a) predicating interactions among other entities; and by (b) being entities of interest in their own right. Thus:

(1) Rudd returned to the top job in June after challenging Gillard in a caucus ballot.

entails facts about an entity referred to as Rudd and another referred to as Gillard, while detailing the temporal (and perhaps causal) interaction between three events – a challenge, a ballot and a return – and stating that the latter took place in June.

Discourse entities need not correspond to entities in the real world. The reference in Example 1 could be understood (at least partially) in an entirely hypothetical world, or were the name Rudd replaced with any other. However, effective communication often relies on the interlocutors' shared familiarity with some common set of entities (e.g. Microsoft and September 11), entity types (organisation or software company; event or terrorist attack) and stereotypes (organisations tend to have leaders; terrorist attacks have perpetrators, victims and reprisals). This shared knowledge, and the ability to refer to its elements succinctly, compresses and thus enables the communication of complex ideas.

News and social media play a foundational role in public discourse: they establish shared entities and points of reference for communication among their readers and contributors. It

is therefore essential for a language processing system to correctly interpret (a) the media's introduction of entities and knowledge about them to popular discourse; and (b) reference to entities that the media has shared with the public. This motivates *entity detection and characterisation*,[1] a broad family of tasks that attempt to infer structured knowledge about entities from free text, and to recall entities previously referenced or externally known.

While a vast literature addresses with some success the detection and characterisation of selected types of discourse entities – notably *named entities* such as people, organisations and locations – several features make the interpretation of *event reference* especially challenging. Some are due to the complexity of event ontology:

- Distinguishing between events and non-events is not trivial or definitively agreed upon in the way that distinguishing, say, a location from a non-location generally is.
- Events are not readily partitioned into discrete sub-types. As discussed in later chapters, there seems to be no fundamental feature of events by which they are naturally grouped, in contrast with how we might group organisations by their industry or function.
- As an abstract entity, the bounds of an event are often fuzzily defined, untrue of people.[2] Thus the leadership spill of Example 1 may be understood to conclude with the returning of ballot results, or with the resignations and appointments that ensued.
- In this manner, events also have internal and temporal/causal structure that describe a complex interrelation among events.

There is corresponding complexity in the language of event reference:

- Few events mentioned in text have a canonical designation as proper names often are for people and organisations. While proper name forms are an easy way to introduce a known entity into a discourse, event references often need deeper interpretation and disambiguation.
- To achieve succinct communication, precise disambiguating information may be elided, such that Kevin Rudd's leadership challenge might be assumed to refer to an event of June 2013, though it ambiguously refers to an event of March 2013. This also illustrates that the relationship between reference and referent is highly influenced by context and perspective: between March and June 2013, the referent is unambiguous.
- Events can be referred to through diverse syntactic constructs, including verb phrases (e.g. Rudd returned), noun phrases (Rudd's return) and adjectival phrases (the recently returned prime minister). The textual extent of a reference may also be ambiguous.

[1] This includes but is not limited to the Automatic Content Extraction evaluation task of the same name (in title-case). Here we consider a broader space of entities, and a more general goal in their characterisation.

[2] Some other named entities may be fuzzily bounded. For example, one might not be able to precisely define the extent of Mount Everest.

- References vary in their salience. For instance, a human annotator can easily fail to identify events that are mentioned deep within a syntax tree rather than forming a main clause; some events may themselves be insignificant to a reader.

- The close tying of reference to predicates allows reference to be easily confused with the predicate's denotation. Thus assassinate, murder, kill and poison may refer to more-or-less the same event, but the distinct semantics of each predicate may complicate determining their coreference.

A number of these features make it especially difficult to identify that two event references are *coreferent*, by which we mean they indicate the same event: references may employ vastly different language; they may refer to closely related events or aspects of an event, and identity may therefore be difficult to decide; and although the entities participating in an event often help to identify it, Hasler and Orăsan (2009) show that this is not a reliable indicator of event coreference.

This thesis presents a data-driven view of news events. Our approach contrasts with seminal prior work that has been largely motivated by particular applications (e.g. Grishman and Sundheim, 1996; Allan, 2002) or theories of event understanding (e.g. Baker et al., 1998). We study event reference by analysing annotated corpora, human disagreement within that annotation, and the ability of computational systems to replicate human annotation. We ultimately investigate the impact of events' structure and interrelation on their presentation in news and reference in later discourse, with a focus on determining event identity.

The centrepiece of this thesis is a new task, *event linking*, which reformulates event reference understanding as the selection of a canonical identifier from among a fixed set of candidates. In contrast with approaching coreference as partitioning the set of references according to their referents, this resembles the communicative act of grounding a reference in shared knowledge. The task is analogous to named entity disambiguation or linking (NEL; Bunescu and Paşca, 2006; Cucerzan, 2007; Ji et al., 2011; Hachey et al., 2013) which involves grounding references to a Wikipedia-derived knowledge base of famous entities. By pre-specifying the space of candidate referents, and including structured knowledge to assist in the disambiguation, NEL has revolutionised cross-document named entity coreference resolution.

Our task similarly selects a set of notable, or *newsworthy*, referents as disambiguation candidates: all events reported in a selected news archive.[3] Each news report thus becomes a proxy for the events it is first to report within the archive – as if an idealised news reader associated each fact with where they learnt it – so that event references within other texts can be identified canonically.

[3]In this work we focus on the granularity of event that is reported in a single article, like an athletic record being broken, but unlike an Olympic Games.

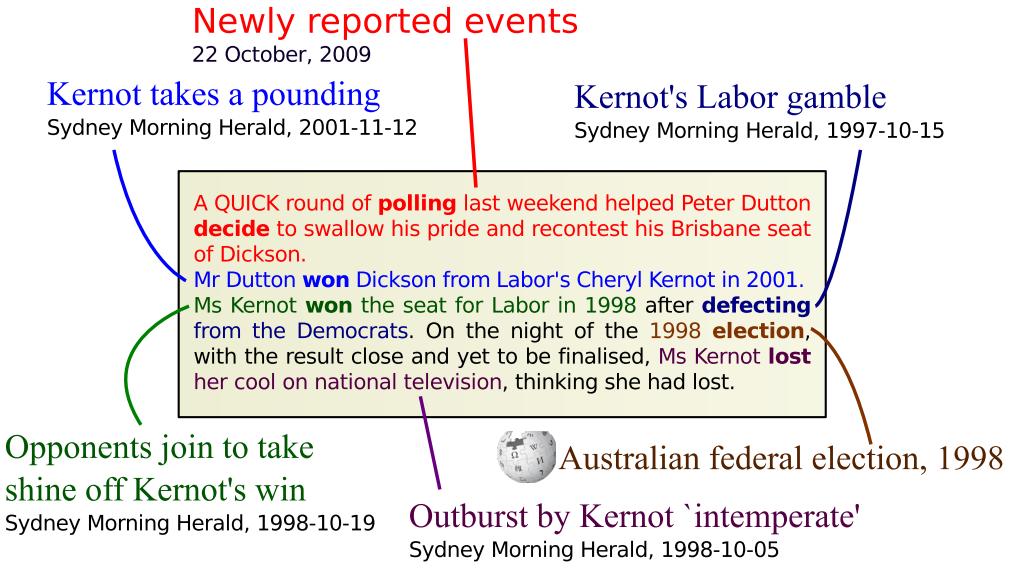


Figure 1.1: Linking background event references to a news archive: an excerpt from *Lib polling helps sway Dutton's return to old seat*, Sydney Morning Herald, 22 October, 2009.

Figure 1.1 illustrates an example of event linking annotation. It takes the Sydney Morning Herald as its source of event knowledge, and can therefore identify archival stories to represent the referent of Ms Kernot won the seat for Labor in 1998 and . . . defecting from the democrats, for example. The concept of newsworthiness is relative to a subjective frame; we adopt the frame of the selected news source, and had we selected another source where these events were not reported, their references would be labelled $\varnothing$ ("nil"). Not all events are reported in a single article as these are; a complex event such as the 1998 election might be better linked to the article titled *Australian federal election, 1998* in English Wikipedia. When linking news text, we further make a distinction between references to background events that a reader may know of and others that the report is introducing to public discourse. Therefore the events of our example's first sentence cannot be linked to archival stories; as the article is added to the news archive, these references determine its scope as an event link target.

Most existing computational approaches to event reference – together with NEL – seek a form of coreference that may be impractically strict and not reflective of language use and perception; other work such as Bejan and Harabagiu (2008) attempts to identify fine-grained relations between events (such as *part of*, *reason for*), but we argue that these labels are often undecidable. Some recent work has emphasised accounting for partial or near identity of referents (Recasens et al., 2011, 2012; Hovy et al., 2013a). Under our model, the relationship obtained by linking two references to the same report A need not be precise coreference: the texts may refer to different events or different parts of the same event reported in A, a phenomenon we denote *co-reporting*. While this admits the conflation of tangentially-related events, it is also able to provide a sense of near-identity that approximates human perception and discourse surrounding events.

This work describes the event linking task, supported by the annotation of an evaluation corpus, a system to benchmark its feasibility and a diagnostic evaluation.

A number of smaller experiments into event characterisation influence our design of event linking. We report on annotation experiments that explore existing and novel approaches to event typology. They also consider the contrast between marking a specific anchor for an event reference, and considering the event reference of a news story as a whole. In the latter scenario, we assess the idea that each news story pertains to a broad *topic event* and a specific *update event*. We find identifying a single implied topic event particularly problematic, from which derives the notion of finding and linking explicit background event references in the event linking task. Similarly, updates often consist of multiple related, co-reported events and their sub-events.

Cutting across this thesis is a theme of minimising dependence on costly expert linguistic annotation. This follows on from our work in exploiting Wikipedia as a large quantity of high-quality training data for named entity recognition (Nothman, 2008; Nothman et al., 2008, 2009, 2013). Thus in Section 3.1 we consider Wikipedia's applicability to similar processing of events, but find its coverage is skewed away from the general application we seek. In Section 5.2 we describe a manual corpus annotation for event linking, and relate our experience of working with non-expert annotators individually hired through an online freelancing marketplace. While crowd-sourced linguistic annotation and system evaluation with tools like Amazon Mechanical Turk has become popular in recent years, we know of no prior work discussing our approach to outsourcing for linguistic annotation. We suggest that the applicability of this outsourcing model is task-dependent, and may not be especially suited to a newly developed and intricate task such as event linking. Our manually annotated corpus remains small due to its intensive annotation procedure, so for our final event linking evaluation, we employ two corpora of hyperlinks to online news to train our system (see Section 7.1). We believe that hyperlinks to online news present an under-exploited indicator of cross-document event (co)reference, and implicitly exploit this fact in using them to learn an event linking model. Despite the quantity of noise in these data sources, they yield a significant improvement in event linking performance over a textual similarity baseline. Nonetheless, a large quantity of newsworthy event references remain difficult to link and will require deeper linguistic disambiguation techniques.

1.1 Contributions and outline

Our work begins by surveying a vast array of NLP tasks addressing event language. Chapter 2's review is novel in its attempt to capture the common threads of event characterisation present in the information retrieval, information extraction and computational lexical semantics literature. It focuses in particular on three challenges in processing references to news

events that motivate event linking and our explorations in Chapter 3. One (Section 2.4) relates to the fact that event references may be prominent or peripheral, and that peripheral events may be difficult for readers to identify. This suggests that tasks and systems should account for varying salience, and leads to our focus on newsworthy events in event linking. Another considers the difficulty of grouping events into types, despite typed events being the focus of information extraction work (Section 2.5). Following the annotation exercises of the following chapter, for event linking we opt to categorise events only by the manner in which the news media reports them, rather than by their ontology. The third discussion (Section 2.6) delves into the problem of determining whether two references indicate an identical event, which motivates our relaxed notion of co-reporting in event linking.

In Chapter 3 we detail three explorations of event referential language oriented towards the event content of broad-coverage news. Drawing on our other work in exploiting links between Wikipedia articles as indicators of named entity reference (Nothman et al., 2013; Hachey et al., 2013), we appraise its comparable application to event detection. Although we later take an alternative approach to harnessing event references from Wikipedia in Section 7.1, this exploration finds Wikipedia's topical coverage of events to be highly skewed, making it unsuitable for general application to their detection. We therefore proceed to consider the manual annotation of events. The experiment in Section 3.2 adapts an existing event identification approach to annotate a news corpus with broader domain coverage, and provides a qualitative analysis of its feasibility, including new empirical insights into annotator disagreement in an existing corpus. This analysis leads us to a second experiment (Section 3.3) involving a novel approach to event salience in news, in which a story is characterised in terms of its update and topic events. In this context we also propose and qualitatively assess a new method for annotating event types, employing a dynamic hierarchy to delay specification of typological granularity. The conclusions of these experiments underlie the event linking model of reference introduced in the following chapter.

The event linking task first described in Nothman et al. (2012) is given a fuller consideration in Chapter 4. We motivate the new task in terms of prior literature in event detection and characterisation, and as an event-oriented analogue of named entity linking. We list possible applications of event links under the assumption that they may be determined at scale. The remainder of the thesis works towards evaluating this assumption.

Chapter 5 describes the manual annotation of a corpus of event links. It introduces some of the considerations when developing an annotated corpus for this and related tasks. One factor is the selection of event references to disambiguate, discussed in Section 5.1, while another is the hiring of annotators. The chapter thus contributes to a growing literature on outsourced, non-expert linguistic annotation. In Section 5.2 it informally compares the high redundancy, low interaction and retention model employed by Amazon Mechanical Turk with our less reported approach of hiring annotators through an online freelancer marketplace. We

also introduce an extensible, web-based tool that we have collaboratively developed to enable this approach to annotation, which has been applied to a number of information extraction-related tasks. Finally, Section 5.4 analyses annotator agreement and the resulting corpus, providing an initial appraisal of event linking's feasibility and challenges.

The following two chapters further this diagnosis by detailing an initial system to perform event linking (Chapter 6) and its evaluation (Chapter 7). While we anticipate that precise event linking will rely on deep linguistic inference and disambiguation, this entails having a shortlist of candidates to select among. This encourages us to initially consider a retrieval approach to linking. Unlike ad-hoc information retrieval, a query in event linking is an event reference, not a selection of keywords. We therefore describe a means of deriving a term-based query from an event reference. The system further incorporates features designed to retrieve candidates matching the *who*, *what*, *when* and *where* of an event reference. Since event linking targets the story that first reports an event, we introduce term weighting schemes and textual zoning intended to emphasise this characteristic of a candidate. Lastly we introduce a component accounting for textually-derived and prior expectations of when an event link target is likely published, and describe a method for estimating model parameters from annotated examples.

Since our manually annotated examples are few, we consider in Section 7.1 the idea that a portion of existing hyperlinks to online news archives will constitute valid event links. We analyse a sample of in-text hyperlinks from an online news source and find that this is roughly true of half the sample, suggesting this is an under-exploited source of event coreference knowledge. Performing some filtering to remove the most egregious violations of this assumption, we derive two noisy, "silver standard" training corpora: one of hyperlinks between the articles of an online news source, and the other consisting of citations of news in Wikipedia. These corpora result in two parametrisations of the system described in Chapter 6, which we evaluate on gold-standard event links in Section 7.3. We proceed to identify and analyse event links that are easily predicted, and others that will require a more nuanced or intelligent approach.

Before concluding this work in Chapter 8, we discuss the limitations and further possibilities uncovered through it. As the next chapter makes clear, we contribute to an endeavour of understanding event language that is well established, while diverse in its goals and methods. While it does not complete the broader undertaking, this work takes a new direction in computational analysis of event reference.

Chapter 2

Computationally characterising event reference

Reference to events pervades linguistic communication. It allows us to describe the state of the world, how it came to be, and what it might become. As a chronicle of notable events, news media play an important role: they share knowledge of events that becomes the foundation of public discourse. Identifying and characterising references to events within and with respect to news is hence essential to many NLP applications.

Unlike other entities referred to in language, events are intangible and often intricate, and references to them may be structurally complex. To illustrate this and to provide a basis for comparing the accounts of disparate approaches in the literature, we follow Ashish et al. (2006) in adopting an example from a Voice of America news report:[1]

(2) Somali Gunmen Release Ship Carrying Tsunami Aid

The United Nations says Somali gunmen who hijacked a U.N.-chartered vessel carrying food aid for tsunami victims have released the ship after holding it for more than two months. The World Food Program hired the Kenyan vessel to carry 850 metric tons of rice donated by Japan and Germany. The ship and its 10-person crew was hijacked by pirates as it sailed from Kenya to Somalia in June.

From the opening sentence, a competent English reader is able to infer, given sufficient background knowledge, the following sequence of events (as analysed by Fillmore et al. (2006)): a tsunami, the chartering of a ship, the transporting of food aid, the hijacking, the illegal retention of the ship, the ship's release, and the announcement about the release. Ideally, a computational event characterisation system should be able to identify these events from text, the participants of each, the temporal (and other) relations between events, and whether the text asserts that each event has occurred, will occur, may occur, etc. To capture this and

[1] This is the opening sentence of an article published on 15 September, 2005, available in full at `http://www.voanews.com/content/a-13-2005-09-15-voa20-67540612/285900.html`. Ashish et al. (2006) only proposes the first sentence, which is insufficient for adequately comparing approaches presented here.

other information across multiple references, it is important to identify that two expressions are referring to the same event, or more generally to identify any inferable relationship between multiple referents: that one causes another, or is part of another, for example. For a particular application, or to assist in other characterisation, it may also be necessary to determine whether a referent belongs to an interesting class of events: identifying a *transportation* event might allow us to seek arguments for participant roles such as sender, recipient, vehicle, cargo, origin, destination and route. The many linguistic forms that event references can take, coupled with the diversity of information relevant to particular event types, leads to very long-tailed distributions that make automatic event characterisation a sophisticated and challenging task.

This survey seeks to unite disparate approaches to event language, including work in information extraction, retrieval and lexical semantics. In Section 2.2 we use the above example to delve into applications of event characterisation that motivate developing models and systems, and then consider specific representations of event reference through their analyses of that example in Section 2.3. We ultimately select three challenges that are treated in recent work and which guide our own. These we approach with in-depth analysis, studies of annotated corpora, and reference to engineering solutions, as appropriate: the variable salience and notability of events (Section 2.4); event diversity and taxonomy (Section 2.5); and determining identity of referent (Section 2.6). Since the space of literature tackling this and related problems is vast, we begin by describing some limits on the scope of our review.

2.1 Scope

Event is difficult to define, although events are the subject of many genres of discourse. The term is thus used to refer to:

a shooting — with a perpetrator, a victim, a timespan and a place, or
    a death — with a patient, and an instant in time;
an anticipated election — perhaps at one time, or including its lead-up and counting disputes, occurring across a whole country, or
    an unexpected tsunami — appearing in many places and times over a few days;
a change in a stock-market index — aggregated from individually-insignificant trading events, which in turn consist of
        bidding,
        matching, and
        settlement;
a friend's birthday party
    or a Royal Birthday Event

a (state of) recession — which may exist unidentified until economic statistics are
released, and
the vague notion of "downturn" that signalled the recession's arrival;
a series of sports matches, or
a single match, or perhaps even
a single winning goal;
talking to a friend,
negotiating with an enemy,
debating with an opponent,
announcing to a press club, or
swearing at your boss;
the construction of a building — from start to end, or,
its discontinuation,
resumption, and
completion (though that may never happen);
the Wicked Witch of the West melting, albeit not in our world;
(in more technical domains:)
a generalised interaction of chemicals, or
transcription of genes,
the click of a mouse,
or the arrival a package on a network interface.

Although impressionistic, this listing conveys the diversity of the space of referents that may constitute events: some are closer to the prototypical event, having a specific place and time and perhaps a name or near-canonical description, while others may occur across an ill-defined span of time, or in many discrete times or places, or at no specific time or place; events may be changes of state (e.g. a death) or states (e.g. a war, arguably); may be realised or hypothetical; may cause or be caused by other events, or have complex internal structure. *Event* tends to be defined through notions of "situation", "happening", "change in state", and typical temporal-spatial properties (e.g. Pustejovsky et al., 2003b; LDC, 2005). Although a prototypical event may be indisputably identified, the bounds of *event* are ill-defined, incorporating a long tail of discourse entities that are referred to and manipulated in the manner of more typical events.

This work focuses on the kinds of event that underlie the reporting of news for popular consumption. Specifically, we exclude the notion of *event* as applied in the extensive literature on processing the language of technical domains such as biological text (e.g. Kim et al., 2008, 2009). Although some technology may be transferable between the two applications, news events present particular challenges in comparison to these restricted domains, including: their diversity and structure (temporal or otherwise), the reliance on world knowledge

and common sense for their interpretation, and the many perspectives from which they are perceived and reported. Henceforth, our use of *event* will assume this domain selection.

We concentrate on news events, but do not confine our attention to events *as mentioned in news*. News events are referenced in many genres, including traditional opinion writing, diplomatic and military messages, blogs and microblogs, histories, encyclopædias and popular literature. While some methods of processing event knowledge rely on genre-specific structure, we should be aware that similar references must be understood from other genres.

This work also presents a bias towards resources in the English language, although we assume that events as discourse entities are common and present similar challenges across many languages.

We do not directly address the linguistic and cognitive literature on events, as we consider the issues they raise through computational models. The interested reader is referred to Mani et al. (2005) which anthologises seminal work in the formal analysis of event reference and temporal reasoning, along with related computational literature. That collection largely omits background to some focal aspects of this work, including event identity as discussed within philosophy (e.g. Davidson, 1969; Kim, 1973; Peterson, 1997; Bennett, 1998), and referential language and grounding (e.g. Donnellan, 1966; Webber, 1987; Kronfeld, 1990; Clark and Brennan, 1991; Clark and Bangerter, 2004).

2.2 Motivating event detection and characterisation

We approach event reference largely from the perspective of information extraction (IE), which seeks to obtain structured knowledge from free text; Yangarber and Grishman (1997) describe its goal as the "selective extraction of meaning". A focus of IE is to enable quantitative analysis and aggregation of data that is difficult to count without linguistic interpretation and data normalisation. Event details need to be represented in a common form to be aggregated or retrieved from a database.

For example, an event IE system might be used to chart trends in piracy as reported by news, with the ability to segment the data by region and time, as well as assailant or damages. Other applications include real-time analysis of disease outbreaks (Grishman et al., 2002a; Patwardhan and Riloff, 2007), violent activity and natural disasters (Piskorski et al., 2008), or notable corporate activities for business intelligence analysis (Aggour et al., 2006). In order to gather such information from our example, a system first must identify:

what That hijacked is a reference to an instance of piracy. This is a matter of lexical semantics, although hijacked – like many event verbs – may also be used metaphorically to denote other forms of occupation or seizure; and it is by no means the only way of referring to piracy.

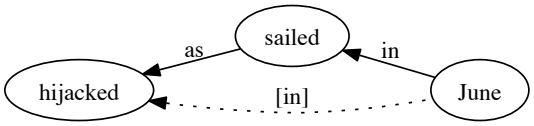


Figure 2.1: A fragment of the dependency graph of was hijacked by pirates as it sailed from Kenya to Somalia in June with an inferred arc between hijacked and in June: event attributes may not be syntactically local.

when In this story, the hijacking assailants and target are syntactically local to the main event verb making them relatively easy to extract (if not interpret); but identifying the event's time of occurrence for indexing is not as straightforward. Given only the first sentence, a system could deduce that the hijacking happened a little over two months before the present news report. While the last sentence specifies in June, this preposition phrase attaches to hijacked only via a semantic coordination facilitated through *as* it sailed (see Figure 2.1).

where From the fact that it sailed from Kenya to Somalia, a system might infer that the event took place near the East African coast; this information, however, is not local to the initial event reference.

which In order to incorporate the information from a later sentence into an earlier sentence, information extraction must deduce that the two uses of hijacked refer to the same event; while a *one referent per lexeme per discourse* assumption[2] might apply in the present example, this is often not the case since news reports frequently reuse the focal event's vocabulary to describe related events, such as other recent hijackings. Conversely the article could equally have used other parts of speech (e.g. the hijack, hijacker, hijacked vessel, it) or vocabulary (e.g. attacked, seized, took control) with an identical or near-identical referent event. Further, when aggregating details of distinct events across multiple texts, references to the same event must be identified and combined into a single record. This is more difficult than treating references within a single document because the texts are composed in different contexts; the manner of referring to an event may change over time, such that in our example tsunami is presumed unambiguous in its temporal context without further specification.

Collating structured records of events therefore requires tackling the complexities of event-referential language, even when focusing on a single, specific event type.

Distilling event structure from news reports may also enable their textual and visual summarisation. Rather than extracting events of a given *type*, this family of applications

[2]By this we mean that repetitions of a referring word such as hijacked are assumed to be coreferent. A similar heuristic is introduced in the context of word sense disambiguation – rather than reference – by Gale et al. (1992).

seeks references to events related to a given *topic*, such as constructing a timeline or biography of a particular entity. For instance, a system may include our example news story when summarising the effects of the December 2004 Indian Ocean Earthquake. In order to do so, it may identify basic attributes of events and event coreference as above. In addition, to produce a coherent timeline (Swan and Allan, 2000; Chieu and Lee, 2004; Yan et al., 2011) or summary (Filatova and Hatzivassiloglou, 2004; Li et al., 2006; Liu et al., 2007) it needs to discriminate salient events (such as the hijacking and release) from others (perhaps the hiring of the vessel), and may identify relationships between events, including:

temporal relations where referent events are in sequence or overlapping. For example, the release ended the holding of the vessel over two months after the hijacking.

structural relations where one referent is part of or has a non-empty intersection with another. For example, the referent of hijacked might be construed as a super-event of holding.

logical relations where one event results from or is enabled by another. For example, the chartering of an aid ship was in response to the damage caused by the tsunami.

Recognising event interrelations allows a summary to avoid redundancy, but also to focus on diverse, relevant sub-events for a broad event being summarised (Daniel et al., 2003); by timestamping (Filatova and Hovy, 2001) or obtaining an ordering of events (Bethard et al., 2012), a chronological summary is feasible. Here we have considered the specific relationships between events within our story, but summarisation may also exploit general knowledge of the relationships commonly found between types of events, such as natural disasters causing death and damage, which in turn leads to international aid (and perhaps also increased local crime and anarchy). Such typical sequences of events are known as *scripts* (Schank and Abelson, 1977).[3]

Similar extracted knowledge is required for specialised event and temporal reasoning modules in question answering (Bruce, 1972; Saurí et al., 2005; Schockaert et al., 2006; Harabagiu and Bejan, 2006; Saquete et al., 2009), temporally-aware information retrieval (Alonso et al., 2007, 2011) and more general textual inference (Wang and Zhang, 2008; Im and Pustejovsky, 2010). These applications treat events as facts, and therefore are also concerned with evidentiality and modality in event reference: did or will an event happen certainly, probably or possibly? did it certainly *not* happen? according to whom? As Hovy et al. (2013a) note, recognising such attribution may be utilised to identify coreference for the same event reported from multiple perspectives (with varying death toll counts, etc.).

[3]Ours is a relatively uncommon example of an event script. More commonly, they are described in terms of criminal activity, investigation, arrest, trial, conviction, etc., in which agents and themes of each event can be traced through the script.

Incident	Date: *?–15 Sep 2005*; Location: *Kenya–Somalia*;
	Type: *Hijacking*; Stage: *Accomplished*;
	Instrument id:— Instrument type: *Gun.*
Perpetrator	Category: Terrorist Act; ID: gunmen;
	Organization:—; Org. confidence: —.
Physical target	ID: UN-chartered vessel; Type: *Transport vehicle*;
	Number: 1; Foreign nation: —.
Human target	Name: —; Description: crew; Type: *Civilian*;
	Number: 10; Foreign nation: —; Effect of incident: *No death or injury.*

Figure 2.2: A MUC-3terrorist incident template filled from the hijacking example. Most slots contain normalised value (e.g. Incident Date) or an element of a predefined set (e.g. Incident Type), while a few accept raw textual fragments (e.g. Perpetrator ID).

News writing focuses on events; thus journalistic convention introduces a news *story* with its *what*, *where*, *when* and *who*. Since news is a major textual repository of common knowledge, knowledge-based systems depend on the ability to extract event knowledge from news sources. Despite this clear demand for event information, research progress has been held up by their linguistic and ontological complexity. As such, over two decades since the Message Understanding Conference (MUC) sought a shared evaluation of the technology, it is an area still very much under exploration.[4]

2.3 A brief overview of event characterisation

In order to understand event reference, much of the literature is focused on schematising aspects of event knowledge for annotation and machine replication. The centrality of events to many discourses has led to their treatment under many rubrics, making various schemas difficult to compare. To simplify the literature on event characterisation, we group it into three categories. The first treats an event as the instantiation of a template or predicate; the second treats event instances or mentions as arguments of relational predicates;[5] and a final group is concerned with the typical characteristics of event types.

Events as template instantiations Information extraction traditionally treats an event as a predicate with a number of argument entities or values. Evaluations in the Message Understanding Conferences (MUC; Grishman and Sundheim, 1996) define templates for event types or scenarios of interest – such as *terrorist incident*, as exemplified in Figure 2.2, with

[4]The conference's original name, MUCK, foreshadowed the duration and labour of this quest.

[5]This is not to say that the two are incompatible, and schemas generally include aspects of both; in some schemas, event templates admit other events as arguments, predicating their interrelation.

Event reference	Arguments
transfer-ownership$_{\text{v1}}$([who] hijacked)	buyer: *person*$_{\text{e1}}$(gunmen),
	artifact: *vehicle*$_{\text{e2}}$(vessel)
transfer-ownership$_{\text{v2}}$(chartered)	buyer: *person*$_{\text{e3}}$(UN),
	artifact: *vehicle*$_{\text{e2}}$(vessel)
transfer-ownership$_{\text{v3}}$([have] released)	seller: *person*$_{\text{e1}}$(gunmen),
	artifact: *vehicle*$_{\text{e2}}$(the ship)
transfer-ownership$_{\text{v2}}$(hired)	buyer: *organization*$_{\text{e4}}$(World Food Program),
	artifact: *vehicle*$_{\text{e2}}$([Kenyan] vessel)
transfer-ownership$_{\text{v1}}$([was] hijacked)	buyer: *person*$_{\text{e1}}$(pirates),
	artifact: *vehicle*$_{\text{e2}}$(The ship),
	time-within: *timex*(June)
transportation$_{\text{v4}}$(sailed)	artifact: *vehicle*$_{\text{e2}}$(The ship),
	origin: *location*$_{\text{e5}}$(Kenya),
	destination: *location*$_{\text{e6}}$(Somalia),
	time-within: *timex*$_{\text{-}}$(June)

Figure 2.3: Event references and arguments extracted from the hijacking example according to ACE05 (NIST, 2005; LDC, 2005). References are shown as *type*$_{\text{referent index}}$(textual anchor). At least as notable are the events the schema does not annotate, including those predicated by tsunami, says, carrying, holding, donated. It also does not mark the subordination of to carry by hired, nor the nature of the hijacking as violent and its effect on 10 crew.[7]

slots for the perpetrator's name, weapon, targeted facility, time and place among others – to be filled from each document.[6] Many domain-specific event extractors follow this approach by defining a small set of templates (e.g. Grishman et al., 2002b; Kim et al., 2008; Reuters OpenCalais, 2009); Aone and Ramos-Santacruz (2000) hand-code 61 events with verb-centered templates, while the Automatic Content Extraction evaluation (ACE05; NIST, 2005) broadens the syntactically-unconstrained task to 33 fine-grained types. Despite their breadth, these types capture few, select aspects of the event content of a text, as exemplified in Figure 2.3. Like the original MUC task, these templates consist of normalised values and are to be merged from multiple references within a text. The type-specific templating

[6]The original task involved categorical, numerical and textual slots, sometimes multivalued, but more recent literature has focused on string-based template extraction as an encapsulated task (e.g. Patwardhan and Riloff, 2007). Sundheim (1995) notes that categorical slots were difficult to fill: "Two of the slots, VACANCY_REASON and ON_THE_JOB, had to be filled on the basis of inference from subtle linguistic cues in many cases. An entire appendix to the scenario definition is devoted to heuristics for filling the ON_THE_JOB slot. These two slots caused problems for the annotators as well as for the systems."

[7]Some of these predicates are not markable because ACE05 does not include a relevant type, such as *natural-disaster*; others are due to rules in type definitions requiring them to have an artifact argument that is a markable ACE05 entity, excluding food aid. hijacked might be marked as an *attack* event, marking the ten crew as a target, if not for the schema requiring each predicate to anchor at most one event. We assume

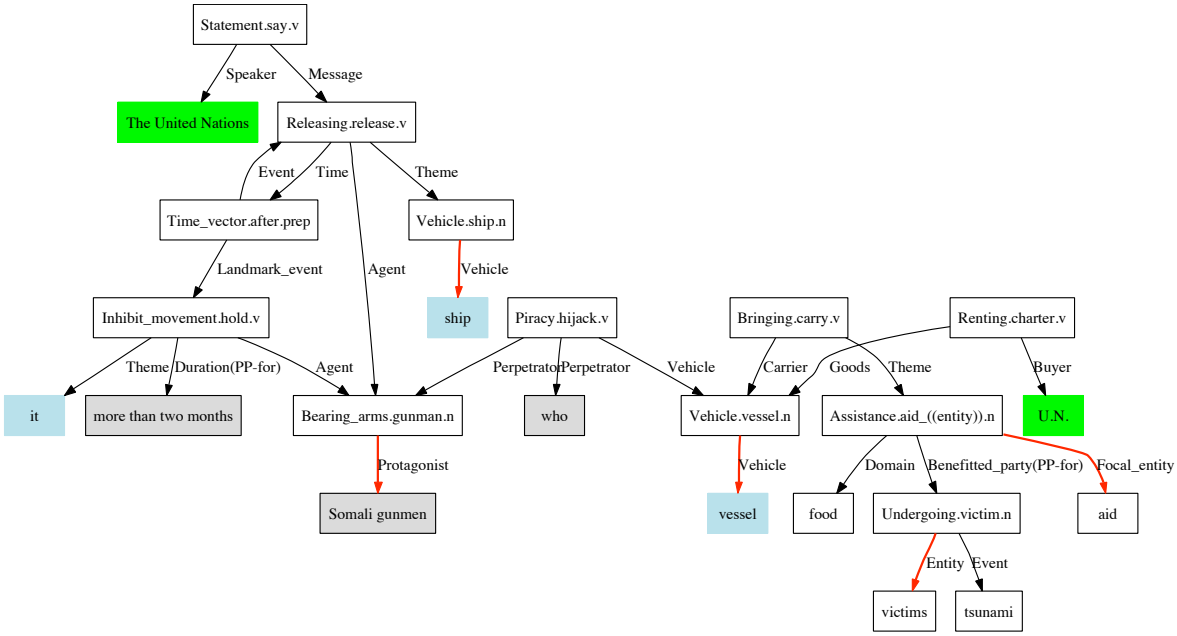


Figure 2.4: The frame dependency graph of the first sentence of our hijacking example (from Fillmore et al., 2006). The FrameNet (Fillmore et al., 2003) schema relates events to arguments – including other events – where there is a syntactic dependency between them.

approach is further generalised in semantic role labelling (SRL; Gildea and Jurafsky, 2000; Palmer et al., 2005; Fillmore et al., 2003), which focuses on the syntactically-local content around each predicate rather than event references in particular. A manual application of FrameNet (Baker et al., 1998; Fillmore et al., 2003) to the first sentence of our example is shown in Figure 2.4. Their schema assigns generalised frames to prototypical event expressions (hijacked instantiates the *Piracy* template) and less event-like predicates (aid instantiates the *Assistance* template), acknowledging their similar argument structures; unlike ACE05, it also allows events to be arguments of events, modelling relationships present in syntactic subordination such as released . . . after holding. SRL ultimately focuses on predicate semantics, and so is not concerned with identifying entity or event coreference.

Open Information Extraction (Etzioni et al., 2008) similarly captures events among other predications, but does not vary the slots available to fill for each extracted relation. This approach does not (initially) attempt to resolve, classify or normalise the event or its participants, representing them through raw strings. Its strategy is to focus on simple, precise extraction at large scale, using information redundancy as a measure of importance; it is therefore not designed to apply to isolated documents as we do for our example in Figure 2.5). Responding to an annotation experiment, Filatova and Hatzivassiloglou (2003) similarly describe a redundancy-driven shallow event extraction model, relying on the presence of at least two named entity, location or time references in context of a hyponym of *event* or *activity*.

hijacked constitutes *transfer-ownership* since it is similar to the example seized the building (LDC, 2005).

subject	predicate	object	confidence
its 10-person crew	was hijacked by	pirates	0.86
The World Food Program	hired	the Kenyan vessel	0.62
The United Nations	says	Somali gunmen	0.43
tsunami victims	have released	the ship	0.30
it	sailed from	Kenya	0.10
Somali gunmen	hijacked	a U.N.-chartered vessel	0.02

Figure 2.5: Automatic Open IE extractions from ReVerb (Fader et al., 2011) in the unintended setting of extracting from a single document. The extractions emphasise the linguistically shallow features used in ReVerb; the confidence scores identify a bias against non-proper nouns and against lone verb relations, often being unreliable shallow extractions.

Like Open IE, the 5W task (Parton et al., 2009) considers a one-template-fits-all approach to extract the key elements describing the main event mentioned in a sentence, motivated by the need to capture the same information in multiple languages.

In all these approaches, the participants in an event and the roles they fill are central to the extraction. They differ in the extent of typological specification and selection, the use of normalisation and coreference as opposed to surface forms, and constraints on the textual locality of participating references.

Inter-reference relations Another family of approaches focusses on the relationships among discourse entities represented by event references. The referent identity relationship (i.e. coreference as between chartered and hired in our example) is the focus of much work, including coreference across documents between MUC (Bagga and Baldwin, 1999) or ACE05 templates (Ji et al., 2009), while ACE05 includes within-document coreference annotation.[8] These stand out as determining the relationship between structured and selectively-typed representations of event. Thus OntoNotes comprehensively annotates coreference between noun phrases within a document, as well as any verb phrases coreferent with noun phrases (Pradhan et al., 2007b), thus including event references. Extending this to cross-document identity, Lee et al. (2012) similarly mark up clusters of same-topic news stories that had been more sparsely annotated by Bejan and Harabagiu (2010). At a coarser referential granularity, the main task of the Topic Detection and Tracking evaluations (TDT; Allan, 2002) – grouping news stories by topic – may also be considered a form of event coreference detection.

Yet event dependencies other than identity may also be of interest. This includes logical relationships – e.g. a ship's sailing enables its hijacking – and temporal relations. Nallapati et al. (2004) granulate the topic detection task by threading clustered stories into a hierarchy

[8] MUC's template extraction relies on identifying coreference, but does not annotate it explicitly.

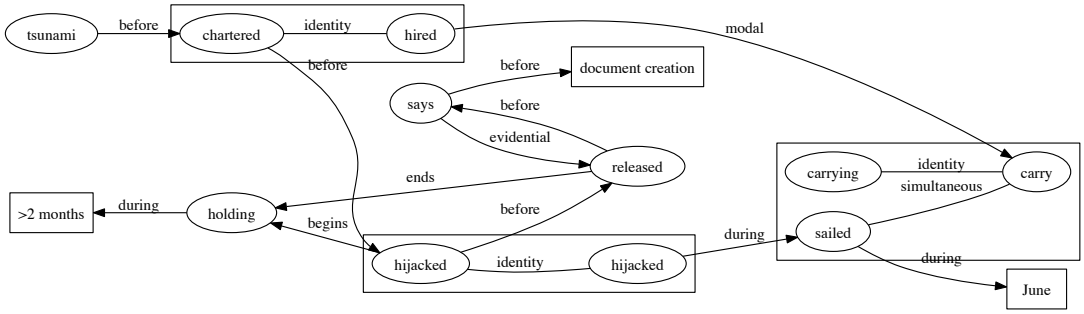


Figure 2.6: Part of a TimeML (Pustejovsky et al., 2003b) temporal and subordination graph annotated over the hijacking example.

corresponding broadly to event dependencies, with Feng and Allan (2007, 2009) labelling relationships between news passages from a set of referential, logical and rhetorical relationships, akin to the annotations of cross-document structure theory (Radev, 2000). In a similar vein, but working with selected event predicates rather than all passages, Bejan and Harabagiu (2008) annotate structural (identity, sub-event), logical[9] (purpose, enablement, related) and temporal (precedence) relations.

TimeML (Pustejovsky et al., 2003b) standardises earlier work (e.g. Setzer and Gaizauskas, 2000; Mani and Wilson, 2000; Harper et al., 2001) by considering events – broadly defined to include stative predicates – as temporal entities. As shown in Figure 2.6, it relates their times of occurrence to mentioned timestamps and to each other. Recent work has suggested that a dependency tree might be a more appropriate temporal structure for annotation and inference (Bethard et al., 2012; Kolomiyets et al., 2012). Others focus on schematising spatial rather than temporal relations between events (Roberts et al., 2012).

These schemas and tasks share a focus on representing interrelation, rather than characterising the semantics of individual event references.

Meta-event templates Further work explores the typical attributes of and relationships between types rather than instances of event, effectively a second-order abstraction of other event characterisations. Many schema and ontology construction efforts can thus be construed in this category. Thus the script knowledge that FrameNet includes in its ontology – that *arrest* precedes *arraignment* precedes *trial* and so on – is the target of learning in Chambers and Jurafsky (2008, 2009) and Bejan (2008). As suggested by the example output in Fig-

[9]Logical relations between events and the discourse relations in their description can easily be confused. As such, there is a large literature on detection of explicit and implicit rhetorical indication of causation (early work includes Grishman and Kslezyk, 1990; Amsili and Rossari, 1998; Khoo et al., 1998), with some particular attention given to event causation as it blurs into other logical relations (e.g. Do et al., 2011). Schematic analysis of causation between events is more applicable to domains such as environmental chemistry (Ji et al., 2010) where one or more chemical events may indeed be a necessary and sufficient direct cause of another.

police	arrest	suspect	company	produce	film	company	issue	report
police	search	suspect	company	direct	film	report	found	company
police	detain	suspect	company	develop	film	company	release	report
police	charge	suspect	company	sell	film	report	criticize	company
police	found	suspect	company	plan	film	report	cite	company
police	identify	suspect	company	write	film	report	recommend	company
police	raid	suspect	company	release	film	report	note	company
police	release	suspect	company	base	film	company	receive	report
police	seize	suspect	company	adapt	film	report	praise	company
police	question	suspect	company	star	film	report	pan	company

Figure 2.7: Sample scripts including the verb release induced from unlabelled text by Chambers and Jurafsky (2009). The highest scoring subject and object label is given for each verb. Verbs are ordered by their centrality to the script; while the approach also attempts to induce temporal ordering (Chambers et al., 2007), that data is pairwise between predicates and very sparse.

ure 2.7, these approaches probabilistically model the occurrence of multiple event predicates with the same arguments. Such data is then able to predict reported events from partial event sequences (Chambers and Jurafsky, 2008, 2009; Radinsky et al., 2012). Similar knowledge may also be exploited by lower orders of event extraction, with global extraction models benefiting from cooccurrence knowledge, such as the expectation that an ACE05 *life:die* reference is likely to appear in a sentence containing a *conflict:attack* reference (Liao and Grishman, 2010). Learning typical event sequences is also similar to learning new event templates, including slots for arguments, from on-topic text (Yangarber et al., 2000; Filatova et al., 2006; Sekine, 2006; Chambers and Jurafsky, 2011), and to learning which predicates tend to occur in a causal relationship (Tanaka et al., 2012). Other work learns typical attributes of event types, such as aspectual properties (Siegel and McKeown, 2000) and duration (Gusev et al., 2011). These tasks appreciate that much of event understanding derives from our expectations – knowledge and common sense – of events, and consider methods of modelling these expectations from textual corpora.

Each family of model captures a part of event understanding, while no approach alone captures the full breadth of event semantics, reference and discourse. There is therefore space to explore systems that integrate a number of these representations to solve an applied task. However, these models also overlap to some extent in struggling with fundamental characteristics of event reference. So having summarised a range of proposals for event representation, we now delve into some universal challenges presented by event reference.

Computational models follow human annotators in struggling to model events and their reference discretely. As suggested above, some referents seem more prototypical than others on a spectrum of event-likeness; the same applies to referential language, such that some

references are more easily noticed than others. Event models must therefore contend with event salience – let alone pertinence to a particular task – varying widely. A second problem faced with events is the difficulty of grouping them into types, as is traditionally considered in information extraction. Events are naturally intangible and diverse; their ontology has a level of complexity not present for other entity types such as people, organisations and locations. In a similar vein, whether two referents are the same event can be difficult to determine, complicated by near-identity relationships between events, such as containment and causation, the interference of semantics when using an event reference from a particular perspective, and ambiguity or vagueness in reference. The following sections therefore focus on these problems – salience, diversity and identity – with reference to annotated corpora, schemas and systems, providing a point of departure for our own exploration and proposal.

2.4 Balancing salience and recall

Within any type of discourse, some referenced events tend to be more focal than others. While one might distinguish the notion of *important events* in a particular discourse context from *prominent event references*, there tends to be a lot of overlap. Thus unimportant events may be indicated through linguistic features such as topicalisation on the one hand and relativisation on the other; adjectival forms may generally be less salient as references than tensed verbs. Some classes of event and reference are also more pertinent to particular applications of language understanding than others. Nonetheless, event schemas tend not to directly address salience, though some incorporate it implicitly.

News stories in particular include peripheral detail surrounding key events. In a seminal study of news discourse structure, Van Dijk (1988) notes that as well as relating events and facts, news structure and detail serve to make a story "noticed, understood, represented, memorized, and finally believed and integrated" by emphasising factuality, providing attitudinal or emotional dimensions and building a strong relational structure for facts (Van Dijk, 1988). He further cites studies suggesting that readers of news rely on an internal schema for expected content structure, and shows that recall is higher for focal events than their reported causes or consequences. Hence some detail may be included for largely rhetorical purposes. Information retrieval and extraction applications would often benefit from distinguishing event references that are central to an article.

Systems may choose to consider only events that are pertinent to some purpose and expressed saliently, or all events referenced. This trades off coverage for substance. Where systems rely on human annotation, they may already be subject to the limited human recall of events as described by Van Dijk (1988). His psychological observations are reflected in linguistic annotation according to Filatova and Hatzivassiloglou (2003) and our own corpus analysis. Filatova and Hatzivassiloglou (2003) perform a study in which students are asked to

Mention class	#	verb	noun	adj.	adv.	pron.	pos.	ass'td	spec.	singleton	no args
Ann. 1 only	341	42.5	51.9	5.0	0.3	0.3	95.3	78.6	79.5	68.3	12.0
Ann. 2 only	495	40.8	50.1	6.3	0.6	0.6	95.8	72.3	83.2	70.7	14.5
Agreed	1331	48.8	49.1	1.7	0.3	0.0	97.1	84.1	89.6	58.1	9.2
Adjudicated	2642	45.7	50.1	3.3	0.3	0.2	96.4	80.5	85.4	63.9	10.7

Table 2.1: Comparison of agreed and disagreed single-token event references in news portions of the ACE05 corpus: the total number of marked references; proportion of references by part-of-speech; the proportion that are annotated as positive in polarity, asserted modality and specific (as opposed to generic); references with no coreference in the document, and those with no arguments in the same sentence. Statistics over the adjudicated corpus are also shown for reference.

mark news story passages that describe events, without further defining the task. The authors notice substantial disagreement, such as in marking statements that describe a continuation of state (while a change of state is often considered a necessary component of an event), and in marking certain types of events such as reporting, also reporting frequent inconsistency of judgement within an annotator's work.

To better comprehend the problem of event reference recall, we perform an analysis of inter-annotator agreement in the ACE05 training corpus (Walker et al., 2006; LDC, 2005; NIST, 2005). In this data, annotators mark usually a token[10] per sentence as anchoring reference to an event of targeted type, aiming for complete coverage. They link this anchor to the event's arguments within the sentence, and annotate it with attributes: tense, polarity, modality and genericity. The data is released with the annotations of two first-pass annotators and the adjudicated corpus. To analyse problems in recall, we identify two classes of event references in the data, discarding other types of disagreement:

Agreed references where there is a unanimous decision on the single anchor token, event type and subtype.

Disagreed references where one annotator did not mark an event[11] and the adjudicator agreed with the other annotator, who marked a single token.

Considering only the news text portions of the corpus, we calculate the proportion of references in each class with particular properties[12] as shown in Table 2.1. While annotator two

[10]Over the entire corpus, 95.6% of references have a single token anchor; another 3.9% are two tokens. Multi-token anchors are most often verb-particle constructions (e.g. thrown out), idiom (bitten the dust) or proper names (Operation Iraqi Freedom). See LDC (2005).

[11]To be sure, we discard any annotated references in the same sentence of the same type, and on the same anchor word of a different type. This still admits both false positives and false negatives for lack of recall, but this should not substantially affect our understanding of the distributions.

[12]We adopt the attribute values and arguments of the final adjudicated annotation.

seems to have substantially higher recall of adjudicator-accepted references, results for the two annotators are consistent in the types of reference that are easy to miss, including:

- those not part of coreference chains (disagreeing is 11% more frequent than agreeing);
- generic references such as have carried out *bombings* in the past (9.1%);
- non-asserted modalities including hypothetical events (7.2%);
- references not expressed through verbs or nouns (4.4%), as determined automatically using the C&C tools (Curran et al., 2007) POS tagger trained on the Wall Street Journal portion of the Penn Treebank;
- references with no arguments in the sentence (3.5%); and
- negated events (1.7%).

Logically, these are atypical of salient event references in news, and are therefore not picked up by annotators, despite their seeking such references more assiduously than the common reader. Unlike marking named entities, there are no simple orthographic cues (in English, at least) to ensure full coverage of event reference.

Generic references are a particularly tricky case. These have referents that cannot be understood as "a singular occurrence at a particular place and time, or a finite set of such occurrences" (LDC, 2005). Among their examples are:

(3) a. Salat Hassen called on countries that give *aid*.

b. The group specialized in *transporting* illegal weapons

c. There have been concerns the *clashes* in southern Serbia could explode into *violence* similar to the 1999 conflict in Kosovo.

In the ACE05 data there is a substantial discrepancy between the two first-pass annotators, with annotator 1 more frequently marking generic events that are accepted by the adjudicator, in contrast with the usual adjudication preference towards annotator 2. The use of clashes in Example 3c gives us some idea of the difficulty of identifying generics: it can be difficult to distinguish them from reference to a finite set of specific referents without further context that explicitly states such. The examples of violence, aid and transporting seem less-certainly references to events. TimeML annotation has avoided generic event mentions – by which they mean references to events that cannot be temporally positioned with respect to other temporal references – and temporally located events with generic complements (Saurí et al., 2009), such as:

(4) He *said* students are prohibited from fighting with each other

UzZaman and Allen (2010) suggest that generic references have application in question answering, and so propose additional annotation of generic events in TimeML. This illustrates competing interests within the field, between having good coverage of references to events for a complete textual understanding, and seeking only those that are most prototypically events with singular references.

MUC-style template extraction generally avoids the issue of salience through precise specification of the sought event type, such as *Latin American terrorist incidents*, wherein each matching event is presumed important to the task. Template-based approaches may additionally utilise the extent to which a template can be filled, and its entity or quantity arguments as a means to select important extractions. In its attempt to make MUC's event templating more diverse and widely-applicable, ACE05 loses some of this property: an analyst is unlikely to be equally interested in gunfire, sexual abuse and the Holocaust, all of which are annotated *conflict:attack* events in the ACE05 corpus. In the terminology of Ji et al. (2009), ACE05 produces extractions that are *unranked*.

Noisy event detail may also be avoided by exploiting typical news discourse structure. Rather than interpreting events mentioned throughout a story, an approach may treat each story as a single unit for event reference, as in topic detection and tracking (Allan, 2002). Topic detection classifies each article as belonging to or briefly mentioning a topic revolving around a set of related events. This allows systems to be penalised differently for missing a focal or a brief mention. Other schemas extract at most one event per article, even considering only its headline or opening sentences (*lead*) as a proxy for the main content (e.g. Piskorski et al., 2008; Radinsky et al., 2012).[13] The threaded topic detection of Feng and Allan (2009) works at the paragraph level, considering all paragraphs as of equal importance, and ignoring paragraph-internal event relations. The 5W task presented by Parton et al. (2009) makes a similar simplification at a finer grain, extracting details of at most one event per sentence. While TimeML may annotate multiple event references in a sentence, the notion of each sentence's *main event* is utilised to limit the set of candidate temporal relations for evaluation (Verhagen et al., 2010).[14] Annotation at low resolution focuses on salient events but it may sacrifice the ability to pinpoint explicit event references or relations where multiple events are salient.

Other forms of text may provide different, if not structured, means of conveying notability. For example, Wikipedia's almanac-like listing of notable occurrences by date represents a small fraction of the total set of events described in the remainder of the encyclopædia. Ahn et al. (2006) harness the implied significance of such events when constructing an event-

[13] This is a feature of news text also exploited for some notions of sentence centrality in automatic summarisation (e.g. Radev et al., 2004b). It seems to be truer of syndicated news content than of the popular press, where the "human interest" factor and engaging readers may be more pertinent to sub-editors.

[14] This is necessary given that many temporal relations are not stated explicitly, and the number of possible temporal relations is quadratic in the number of events.

temporal knowledge base for question answering.

Recent work has attempted to overcome issues of salience by relying on textual redundancy: frequent referents (or their sub-events) may be presumed salient. This is a necessary assumption when seeking events in microblog streams such as Twitter, where the level of noise is high. Yet frequency of reference and re-publishing alone are not sufficient to filter out non-notable, but popular, noise; Osborne et al. (2012) show that correlation between topics viewed in Wikipedia concurrent with Twitter reports provide a better indication of newsworthy content. In general, media processing may readily harness redundancy, since the same content tends to be redacted in multiple sources. Redundant reports have been considered as a means of improving low confidence extractions: this is fundamental to Open IE (Banko et al., 2007; Etzioni et al., 2008), which primarily considers redundancy as a predictor of correctness rather than salience (Downey et al., 2005). Yangarber (2006) pose that information extraction should involve reference to multiple sources, allowing the cross-referencing of related reporting (in concord or discord) and events, while harnessing non-local content for a better understanding just as a human reader might; and Ji and Grishman (2008) apply this notion when filling ACE05 event templates. More explicitly seeking to rank events by salience, (Ji et al., 2009) instead utilise the frequency of named entities in a document collection to determine the weight of the events they participate in. Rather than redundancy in concurrent publications, Gaugaz et al. (2012) argue that the longevity of an event's coverage in news as an indicator of its importance. Since this measure is only determinable in hindsight, they attempt to predict it from the initial news of the event using a regression model.

We have considered the balance between characterising focal event content and all event content in a text. It is clear from textual annotation and psychological studies that readers do not perceive or recall references that are peripheral – by way of atypicality, syntactic form and discourse structure – as easily as direct, topical reference. While the treatment of events in schemes such as TimeML and OntoNotes largely ignores salience, previous work has used various means of targeting only focal references, reliant on: detailed specification of a targeted event type; properties of discourse structure and the structured content of Wikipedia; and measures related to textual redundancy. Our work is guided by this understanding to consider event topicality in terms of news reporting structure.

2.5 Capturing events in their diversity

Different types of events have different properties and participant roles, and may assume distinctive linguistic forms when referenced. Yet the vast and ill-defined semantic space covered by notable events is difficult to capture or partition into a finite taxonomy. In

one morning's *Sydney Morning Herald*,[15] news deemed worthy of publication includes: the lodging of a complaint, a plane crash, releasing study results, awarding a prize, approval of development plans, opening of a new restaurant, initiation of a lawsuit, appointment of a chief justice and the opening of a house for public view; and this before reaching the sport, business and entertainment sections with their own arrays of stereotypical and unusual events. A broad range of events are salient in news, some unforeseeable, presenting challenges of sparsity and fuzzy (and therefore noisy) semantic boundaries to extraction tasks. We critique two attempts to assign a set of broad-coverage, mutually-exclusive types to events, and then consider such typologies in the context of alternatives applied in the literature.

2.5.1 Partitioning events into types

The challenge presented by diversity is exhibited in the transition from MUC to more general type-based extraction in the ACE programme. MUC was defined by its selectiveness; it targeted a "fixed and closely circumscribed subject domain" (Yangarber and Grishman, 1997) for each evaluation (for instance, *management succession* and *aircraft accident*. Through iterative refinement of templates and detailed annotation specifications, this yielded human inter-annotator agreement from 70 to 90 percent (Will, 1993) with the best system in each evaluation performing in the 50-60% F_1 range (Chinchor, 1998b).[16] The ACE programme was the natural descendent of MUC evaluations, in terms of its tasks and participants; it further specified the extraction of entities and other values, entity coreference considered by MUC-6 (Sundheim, 1995) and entity-entity relations of MUC-7 (Chinchor, 1998a), before considering events (Doddington et al., 2004).

One outcome of MUC was the understanding that targeted applications could utilise small, portable, textually fine-grained components, to be determined and benchmarked separately. (Grishman and Sundheim, 1996). In an attempt to parallel entity and relation extraction, ACE thus targeted more general notions of event extraction than application-specific scenario templates considered in MUC. Hence the first ACE attempt at event annotation considered five very broad event types (LDC, 2003):

- destruction/damage;
- creation/improvement;
- transfer of possession or control;
- movement; and
- interaction of agents.

[15] The edition for 10-11 September 2010.

[16] Note this agreement is calculated over detailed templates rather than whether an event of the target type is present.

Event type	Event subtype
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

Table 2.2: Event types and subtypes in the ACE05 evaluation (NIST, 2005).

The annotation guide (LDC, 2003) provides *arrest* and *winning an election* as examples of *transfer of possession*; presumably *hijacking* falls into this category as well, rather than with other attack-like events in *destruction/damage*, where tsunami might also reside. So while these categories capture ontological families of event and may represent a vast proportion of newsworthy events, such broad types naturally hide distinguishing features of event semantics. This pilot annotation produced much lower inter-annotator agreement than entity or relation detection tasks (Strassel et al., 2004) and so the schema was reinvented for future evaluations.

ACE05 introduced events into the evaluation, categorising those of interest into eight more thematic types which break down further into 33 sub-types[17] listed in Table 2.2. It distinguishes, for instance, birth of a person and the creation of an organization where the 2003 schema did not, but it does not completely cover that earlier typology. For example, it cannot mark the creation of an interesting artifact.

To consider the heterogeneity of ACE05 event sybtypes we plot the frequency of each in an annotated corpus (Walker et al., 2006) against the average length of its coreference chains. As shown in Figure 2.8, frequencies of event subtypes vary from two *justice:pardon* to 1119 *conflict:attack* events. The distribution of the *conflict* type is also clearly imbalanced between its two constituent subtypes, with *attack* over ten times more frequent than *demonstrate*. The infrequent types are too too scarce for supervising a learnt extractor, while the most frequent types are impractically broad for application, with annotated *movement:transport* instances include withdrawal of troops, climbing Mount Everest, a Mars Rover voyage, swimming and weapons smuggling. Even so, numerous interesting events are missed by the schema, from natural disasters to construction to legislation and other publication. The frequency variation is notably present in a corpus that was not sampled randomly from its sources, but selected to ensure sufficient instances of targeted types within a corpus of predetermined size.[18] Variation

[17] All systems known to the author focus on the sub-types, ignoring the broader groupings.

[18] Targeted types include entities such as *vehicle* and *weapon*, relations such as *investor-shareholder*, values

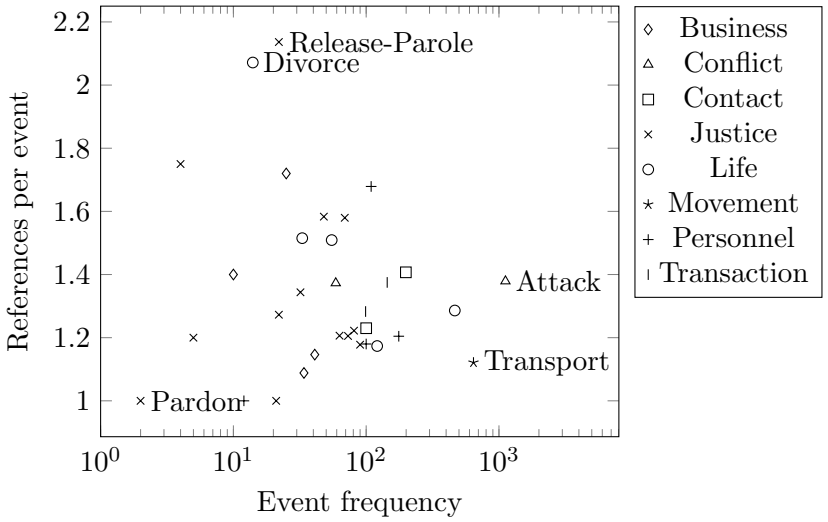


Figure 2.8: Frequencies of event subtypes in all 600 ACE05 training documents.

Evaluation	*attack*	*transport*	*die*	*meet*	*injure*	*charge-indict*
# gold references	984	472	392	160	87	85
Annotator 1	84	78	89	80	89	89
Annotator 2	88	85	92	79	88	87
Inter-annotator	73	61	82	64	76	76
Naughton et al. Trigger-based	25	20	80	65	65	80
Naughton et al. SVM	70	40	75	70	60	80

Table 2.3: Human and system (Naughton et al., 2010) performance (F_1) on a sentence-level event type identification task, over six frequent event types in the newswire portion of the ACE05 corpus (Walker et al., 2006).

on the other axis, the number of references per distinct event, indicates a few categories of event subtypes: *divorce* and *release-parole* are both subjects of documents, with a number of references to the same generic concept of such events, rather than specific referents; in contrast, *pardon*'s annotations tend to be single references in passing; while *attack* exhibits a mix of single focal events and cases where a number of distinct events of that type are mentioned in an article.

By reducing ACE05 event detection to a sentence-level classification task Naughton et al. (2010) illustrate the difficulty of identifying such broad event types. Despite its relative infrequency, a homogeneous type like *charge-indict* is reliably recognised by the human annotators

such as *phone number* and events as in Table 2.2; some of these may substantially bias the corpus domain. Type-targeted sampling was first adopted for the 2005 evaluation in place of random sampling (Walker et al., 2006), and follows from MUC evaluations where corpora were selected to match the target event domain: the MUC-3corpus includes only documents matching topical keywords (Chinchor et al., 1993); MUC-6collates an equal proportion of on-topic and off-topic documents (Sundheim, 1995).

(see Table 2.3), while broader types of event such as *meet* and *transport* are recognised with reasonably high precision, resulting in high annotator F_1 with respect to the final corpus, but lower recall, such that the annotators fail to mark the same sentences, presumably due to sub-salient references. Using support vector machines (SVM), Naughton et al. (2010) are able to approach inter-annotator performance well for most types, but perform half as well for *transport* as for *charge-indict*; for the latter type, using a small list of trigger words is equally effective, while for the former trigger terms perform only half as well as SVM.[19] The *attack* type is also notable for being identifiable with a machine-learning model, but not with a word list, suggesting that unlike four of the six types that Naughton et al. (2010) consider, this type is lexically diverse.[20]

Having reviewed two (correlated) attempts to schematise broad-coverage event types, the extreme variability within and across types suggests that this approach does not readily generalise to the breadth of events. Although we again consider such a typology in Section 3.1, the data presented here suggest that this approach is flawed: while considering a few prescribed event types may be suited to specific applications, alternatives must be considered for more general event processing.

2.5.2 Working with event diversity

To obtain broad coverage of the events that underlie news, a number of computational approaches to event typology have been utilised, which we group into:

- predicates as events;
- adaptation to new event types;
- unsupervised type acquisition; and
- no event types.

We outline the schemas and technologies in each of these areas, reflecting on their relevance to the present work.

[19] We note as a caveat to these conclusions that Naughton et al. (2010) only train (in the case of SVM) and evaluate their classifiers on sentences in documents containing positive instances in the adjudicated corpus. They report that average document containing *charge-indict* consists of 14.8 sentences, and 32.8 for *transport*, while other types have around 30 sentences per document, suggesting results may not be comparable. However, their systems also perform well on *die* which is more comparable to *transport* in terms of document length and class balance (8% positive for both).

[20] Although we can confirm that the *attack* (and *transport*) is relatively lexically and ontologically diverse – covering strangulation, jihad and the Civil War – the poor performance of the lexicon-based system may be in some part due to nuances of the ACE05 schema (LDC, 2005). In particular, annotations are anchored in one or more words, but each word may anchor at most one entity, value or event, such that predicates like murder are annotated as a *crime* value (not an event) or a *die* event; only where separate words in one sentence may indicate the associated death and attack events are they both annotated. (However there appear to be annotated instances of kill in the adjudicated corpus that do not adhere to this rule.) Thus while models accounting for ACE05 type co-occurrence may improve performance (Liao and Grishman, 2010; Li et al., 2011, 2013), this is at least in part influenced by the models learning the specific annotation interdependencies prescribed in this schema.

Predicates as events Applying an inclusive definition of event together with no need to directly address salience, any lexico-syntactic predicate may be considered a potential event reference.

Focusing on the referents' temporal aspects, TimeML (Pustejovsky et al., 2003b) consider every verbal predicate and, selectively, other parts of speech as event references, providing linguistic criteria for their inclusion. As noted above, generic references that cannot be temporally related are not presently marked.[21] As such, bracketed terms in the following are all marked as events: (Saurí et al., 2009):

(5) a. The US economic and political [state embargo] has [state kept] Cuba [state in] a box.

b. All non-essential personnel should [aspectual begin] [occurrence evacuating] the sprawling base.

c. Israel will [i_action ask] the United States to [i_action delay] a military [occurrence strike] against Iraq.

d. "They don't [i_state want] to [occurrence play] with us," one U.S. crew chief [reporting said].

e. Witnesses[22] [reporting tell] Birmingham police they [perceptual saw] a man [occurrence running].

Their seven classes of event distinguish the temporal and evidential effect of predicates, with all but *state* and *occurrence* taking another event as their complement.[23] Chance-corrected inter-annotator agreement over these class labels is $\kappa = 0.67$ (Pustejovsky et al., 2006), which is high but represents a task that is far from trivial for humans.[24] A 2013 evaluation of TimeML technology saw seven sites compete in event detection, yielding up to 81% F_1 for recognition and 89% classification accuracy (UzZaman et al., 2013).[25] The top system (Jung and Stent, 2013), from AT&T Labs, learns a sequence tagger over tokens with morphosyntactic and semantic role features from gold annotations and a large corpus automatically annotated by an ensemble of earlier state-of-the-art systems (Llorens et al., 2012). This high performance suggests their broad-coverage event definition is reasonably robust.

[21] There are of course remaining ambiguities, and we are unsure of whether aid in a U.N.-chartered vessel carrying food aid for tsunami victims: aid, like embargo, may broadly apply as an event predicate; its sense in this context seems not to fit their criteria, though Saurí et al. (2009) do not directly discuss (subtle) polysemy in noun-anchored events. This case might also be considered a generic reference.

[22] According to Saurí et al. (2009), agentive nominals – among other sortal states – are marked only when acting as predicative complements, e.g. They were witnesses to a theft.

[23] Thus saw in Example 5e defines both an evidential and temporal scope for its complement running event.

[24] For comparison, the same annotation effort identified event part-of-speech tags with $\kappa = 0.96$ and tense with $\kappa = 0.93$, determining the normalised value of temporal expressions with $\kappa = 0.89$ (Pustejovsky et al., 2006).

[25] The University of Alicante's top-performing system (Llorens et al., 2013) in the TempEval 2010 evaluation (Verhagen et al., 2010) was used to pre-annotate test data for 2013, and so may not be compared in that evaluation on different data. In the previous evaluation it yielded 83% F_1 for recognition and 79% classification accuracy.

Semantic role labelling (Gildea and Jurafsky, 2000; Baker et al., 1998, SRL) similarly analyses a broad range of predicates, driven by lexical and compositional semantics rather than temporal reference. It uses lexicalised semantic frames to extract syntactically local predicate-argument structures. The extractions are directed at denotational semantics rather than reference; as with the shallower approach of Open IE (Banko et al., 2007), further processing is needed to group extractions and interpret their reference. This caveat and syntactic constraints notwithstanding, the extractions are very similar to MUC and ACE in identifying parse constituents corresponding to a predicate's arguments (its agent, theme, instrument, time, place, etc.), abstracting over semantically-equivalent syntactic forms.[26] Two main families of resources drive SRL work. PropBank (Kingsbury and Palmer, 2003; Palmer et al., 2005) and NomBank (Meyers et al., 2004) are lexicons of verbs and their nominalisations, respectively, with mappings of semantic roles to syntactic realisations.[27] These resources are built to exhaustively cover the Wall Street Journal of the Penn TreeBank for statistical learning, FrameNet (Baker et al., 1998) more closely resembles IE in constructing selected templated *frames*[28] with reference to corpus evidence.[29] Although a lexicon attaches realisations to the frame ontology, it is ambivalent to part of speech, and the role labels assigned are specific to each denotation, rather than each lexeme. Thus the agent of buy and the beneficiary of sell are both assigned the *buyer* role, and their *commerce_buy* and *commerce_sell* frames are both labelled as *perspectives on* the *Commerce_goods-transfer* frame. This frame inherits from a more general *transfer* frame, mapping *buyer* to *recipient*, and so on. Thus FrameNet grounds an intricate type system in surface linguistic structures; while this results in very sparse data for portable tagging, one may coarsen the ontology on an application-specific basis (Ruppenhofer et al., 2010). PropBank-style role labelling was evaluated in the CoNLL 2004-5 and 2008-9 shared tasks (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009), with further related evaluations in SemEval (Pradhan et al., 2007a; Màrquez et al., 2007). Shared evaluations of FrameNet SRL were conducted by Litkowski (2004) and Baker et al. (2007). For both these tasks, the coverage of lemmas is limited by existing annotation, and some work extends existing frames to unannotated predicates through mapping lexical resources (Shi and Mihalcea, 2005; Giuglea and Moschitti, 2006) or statistical language models (Honnibal and Hawker, 2005; Das and Smith, 2011); Palmer and Sporleder (2010) discuss FrameNet evaluation issues related to limited lexical-conceptual coverage. In terms of capturing event knowledge, this may limit the application of such frameworks in comparison to TimeML's less-constrained annotation.[30] The FrameNet

[26] Thus the window is the theme of John broke the window and the window broke despite being respectively the object and subject of the same verb.

[27] Each mapping is known as a *roleset*.

[28] Initially, the following semantic domains were considered: health care, chance, perception, communication, transaction, time, space, body, motion, life stages, social context, emotion, cognition. (Baker et al., 1998)

[29] See Palmer et al. (2005) for further discussion of differences between the two approaches.

[30] However, their purpose differs in that TimeML is not concerned directly with the arguments of each detected predicate, except where those arguments are events.

ontology and mapping of PropBank to WordNet also provide hierarchical semantics of covered predicates, which may serve to inform structured event types.

SRL systems have been applied widely; Surdeanu et al. (2003) in particular suggest SRL's application to IE and Christensen et al. (2011) evaluates SRL in an Open IE context; McCracken et al. (2006) consider SRL as a domain-independent event extraction solution which may then be adapted to a domain-specific application. In this vein, Grishman et al.'s (2005) ACE05 event extraction system is founded on a lexical semantics module incorporating SRL. A hybrid approach is also possible, in that predefined domain-specific templates can be filled, with generic predicate-argument structures employed as an extraction solution for unknown types (Srihari et al., 2003; Reuters OpenCalais, 2009).

By specialising in the interface between syntax or surface representation and semantics, these approaches are able to independently underlie referential models. While we do not make use of them in the present work, their incorporation is likely to benefit most event characterisation systems.

Learning new types Information extraction templates may be distinguished from predicate-argument and semantic role structures by: (a) grouping related predicate referents into types of interest; (b) labelling arguments with meaningful roles; (c) not being bound to local syntax and gathering relevant content from a larger scope; (d) being referential rather than semantic, and thus directed at identity of arguments, rather than their linguistic expression.[31] These differences underlie techniques for generating IE templates for new event types. Where extraction systems may be customised (Yangarber and Grishman, 1997; Chiticariu et al., 2010) or may learn (Riloff, 1993; Glickman and Jones, 1999) a new task from gold-standard templates, the manual analysis and annotation of large quantitative training data remains expensive, and later work attempts to minimise supervision. Riloff (1996) and Yangarber et al. (2000) initiated a thread of research that learns a template for a single event type given a corpus of text describing instances of that type,[32] often with seed lexicons or patterns to bootstrap learning of relevant extraction patterns. As with predicate-argument extractions, these initial approaches require human evaluation and induction of roles from patterns, which further work attempts to automate (Riloff and Schmelzenbach, 1998; Filatova et al., 2006; Sekine, 2006; Chambers and Jurafsky, 2011). Recent work reduces the set of seed patterns to nouns for each role (Huang and Riloff, 2012), while Chambers and Jurafsky (2011) begin with only an in-domain corpus and acquires further instances from unclassified corpora. Huang and Riloff (2013) suggest that a target event type could be characterised by its *who*, *what* and

[31] The distinction is somewhat blurred: FrameNet's ontology groups related predicates and provides meaningful role labels; following Patwardhan and Riloff (2007), many MUC-style information extractors focus on filling fields with surface strings rather than abstracted referents.

[32] The literature talks of an "in-domain corpus" without describing how focused the texts need to be on those events. Presumably the notion is tied to MUC corpus selection, where documents tend to be focused descriptions of the target event.

why,[33] but only use it to construct a lexicon for future template construction. Although the family of minimally-supervised bootstrapping approaches can be brittle to their input and may quickly diverge from it (McIntosh and Curran, 2009), this approach provides solutions for acquiring templates where there is a particular event type in mind.

The main constraint of these approaches is the need to have a particular application in mind. Our work intends to address event reference more generally.

Unsupervised type acquisition A related group of work seeks to group predicates in domain-independent text into likely event types. This approach basically relies on the assumption that multiple references to an event will include the same argument entities, and that similar sets of predicates (with different arguments) express instances of the same event type (Filatova and Hatzivassiloglou, 2003; Shinyama and Sekine, 2003, 2006; Li et al., 2010).[34] Such a system could therefore learn that x won against y is referentially equivalent to y lost to x. There are obvious cases where these assumptions do not apply, as some entity pairs or groups interact in a variety of events; Filatova and Hatzivassiloglou (2003) suggest a reduced weight for frequent entities in this approach. Another concern is the polysemy of event predicates and the same entities coocurring in sequences of related events. Thus Bejan (2008) and Chambers and Jurafsky (2008, 2009) learn scripts of event predicates that describe sequences of interactions among a set of entities; importantly, the same predicate is probabilistically a member of multiple scripts. By framing types around the arguments, these approaches are not necessarily constrained to particular surface forms of predicates, but may prefer the precision of local syntax when learning from large corpora. As with the popular topic modelling technique (Blei, 2012) from which these derive, the output of such processes applies to various tasks; thus Roberts and Harabagiu (2011) employs the Bejan (2008) scripts for TDT's first story detection task.

While these approaches aim to learn type knowledge that may be inspected and reused, a related technique in applied event characterisation represents event types as latent variables. For example, Alfonseca et al. (2013) associate observed surface patterns connecting sets of entities with latent event variables in a complete bipartite Bayesian network, in which inference allows the generation of a succinct headline from a cluster of texts.

The latent nature of event typology in these approaches makes them particularly suitable to tasks where no type schema is formally required.

Type-free event characterisation The set of tasks that consider event characterisation without typology naturally focus on reference more than semantics, and are designed for

[33] The authors use *event phrases*, *agent terms* and *purpose phrases* which, although more precise, do not give the impression of *type* abstracting away the specification of *when*, *where*, *whom* and *how* of the 5 W's

[34] This notion is comparable to the Distributional Hypothesis, that words occurring in the same context tend to have similar meanings (Harris, 1954).

broad application. As such, identifying that a collection of documents (Allan, 2002) or phrases (Pradhan et al., 2007b) refer to the same event need not depend on type. Systems in this area instead rely on clustering or matching surface forms as well as participating entities and the event's location and time, which we detail below with respect to coreference identification. While the TimeML event classes described above are useful for distinguishing the relationship between some events, particularly those that have events as complements, for the most part temporal (Pustejovsky et al., 2003b) and spatial (Roberts et al., 2013) event interrelation operate without types. Type-free approaches can also make use of lexical-semantic resources and latent types when modelling more general interrelation.

After initially exploring flat and structured typologies for event characterisation, we too take the approach of discarding type to focus on reference.

2.6 Elusive event identity

Identifying that a group of references indicate the same referent is central to information extraction. Inferring that events are identical may require inferring that their participants, time and place are identical (or at least compatible), a challenge compounded by variations in reference, internal event structure and changes in perspective. We go into these issues in detail through examples before discussing existing systems and schemas.

Same meaning ⇔ same referent? The difficulty of identifying event coreference is exemplified in the following extracts from Example 2, where a subscripted letter indexes references by their referent:

(6) a. ... Somali gunmen who [hijacked]$_a$ a U.N.-[chartered]$_b$ vessel ...

b. The ship and its 10-person crew was [hijacked]$_a$ by pirates ...

c. The ship and its 10-person crew was [seized]$_a$ by pirates ...

d. The World Food Program [hired]$_b$ the Kenyan vessel ...

It is clear from the three references to event a that coreference may involve paraphrase, being language with the same meaning. Yet, with these fragments disconnected from a coherent discourse, their paraphrase is an insufficient indicator of their coreference: they could be references to two distinct hijackings. Where available in the text, identity of paraphrastic references' arguments (including time and place) is a good indicator of identical events. Event b shows that this is not necessary, either: The World Food Program and U.N. are distinct but related entities, but are agents of the synonymous and coreferent predicates chartered and hired. From an annotation study of within-document event coreference (Hasler et al., 2006), Hasler and Orăsan (2009) find that only 21% of annotated coreference chains include only

coreferential arguments, while 22% have exclusively non-coreferent arguments. This suggests that coreference detection requires identifying arguments that may be substitutions for one another, often an entity that represents or is part of another entity.

(7) a. The United Nations [says]$_c$ ...

b. A United Nations spokesperson [announced]$_c$ ...

Similarly, our text uses the formulation in Example 7a, but 7b is an equivalent reference. However the UN and its spokesperson are not identical entities. One analysis of the semantics involved is that says selects an animate agent, coercing United Nations into a metonymic interpretation; Pustejovsky et al. (2010) establish the detection of these type coercions as a tagging task, in which a single participating system (Roberts and Harabagiu, 2010) performed apparently well, detecting the source and target types of coercions with 96% precision and recall. Since our understanding of both expressions is that the UN is the source of the statement, reference must transfer in the other direction such that in Example 7b the announcement's source is identified as the UN, not merely its spokesperson.

Nor is paraphrase necessary for coreference. We may consider other ways of referring to event a, given appropriate context:

(8) a. [It]$_a$ was altogether unanticipated.

b. ... [another hijacking]$_a$ off the East African coast ...

c. ... authorities must put an end to [this piracy]$_a$ ...

A pronominal reference like 8a need not describe the event, while examples 8b and 8c refer specifically to event a among a family of similar events.[35] Thus while paraphrase is commonly associated with coreference, its identification and the identification of coreferent arguments is neither sufficient nor necessary to determine coreference.[36]

Event structure and coreference These last two examples also highlight the linguistic entanglement of event identity and other structural relationships such as membership and containment. Although very similar to 8c example 9 does not explicitly refer to a alone, but employs a generic reference to a family of events in which a is a member.

(9) ? ... authorities must put an end to [Somali piracy]$_a$ ...

This formulation seems pragmatically equivalent to 8c, suggesting that it is necessary to identify a relationship between the reference in 9 and a.[37] Similarly, a complex or script-like

[35] This dual reference is also associated with constructions such as the first X since Y, as well as the morpheme re- in vocabulary such as reelected.

[36] Recasens and Vila (2010) further discuss the relationship between paraphrase and coreference and its pertinence to NLP.

[37] A further complication is found in ... [other hijackings]$_?$ off the East African coast which includes a reference to that family of events, but explicitly excludes a specific event such as a. To what extent is it coreferent with references to hijackings off the East African coast?

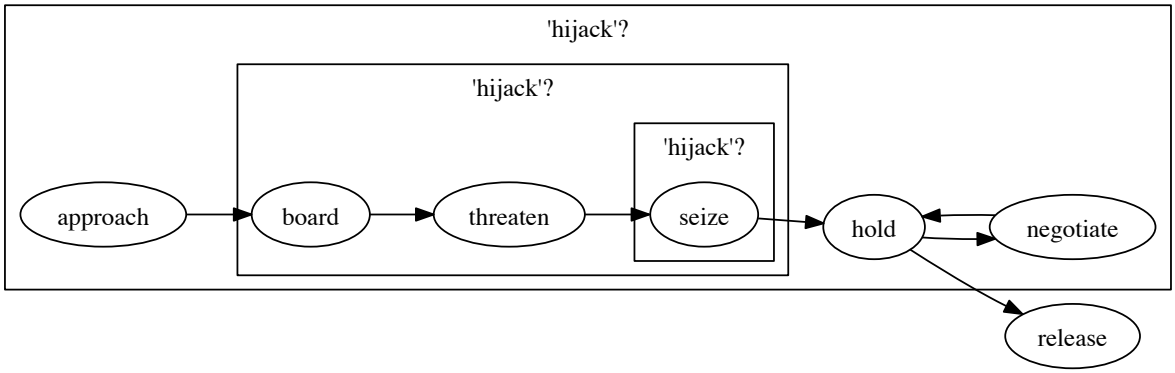


Figure 2.9: Wide and narrow readings of hijacking: a script for hijacking, with three nested units that each might be referred to by The ship was hijacked by pirates. The narrowest reading is required in the interpretation of Example 10d, while either of the two proposed wider readings satisfy 10a-10c.

event referent may be referred to by its constituent events, which may be referred to as a unit (10a) or individually (10b, 10c):

(10) a. The ship was [boarded by pirates and held for two months]$_a$ ...

b. ? The ship was [boarded]$_a$ by pirates ...

c. ? The ship was [held]$_a$ by pirates for two months ...

d. After [approaching]$_?$, the pirates [hijacked]$_a$ the ...

By considering *approaching* as a pre-condition, rather than a sub-event of *hijacking*, Example 10d demonstrates a phenomenon associated with many predicates that denote scripts: they may adopt wider or narrower readings, which we illustrate in Figure 2.9.[38] Hovy et al. (2013a) suggest that within a discourse, a reference will generally adopt a single semantic scope – not wide and narrow readings simultaneously – for the purpose of coreference; without additional references, the intended scope may remain ambiguous. This presents a complication for schemas like Bejan and Harabagiu's (2008) that distinguish sequential-logical (*board* then *hijack*) and structural (*board* is part of *hijack*) relations between events without determining the scope of a reference. Event modality adds another dimension of complexity:

(11) Summers [died]$_d$ from a gunshot to the head. Police suspect [suicide]$_{?d}$ but are investigating other possibilities.

[38]For a more familiar example that makes the need for an event and its sub-event to be recognised as (nearly) identical, consider the reference September 11 and its referent's most prominent sub-event, the twin towers attack. The latter might be understood as identical to a narrow reading of September 11; for many purposes the two may be used interchangeably.

Summers' death is asserted, but the suicide event may not have happened according to the author. In practice, relationships other than precise identity often hold between references that are to be understood as coreferent. Thus Recasens et al. (2011) present a cognitive model in which the identity of discourse referents should be interpreted on a continuum rather than as binary judgements. They describe a typology of near-identity relations – for instance, the same event at a different time – but annotators under this typology still agree only moderately on precise identity (Recasens et al., 2010); a further annotation experiment suggests that explicitly marking *near-identity* may be difficult, but that it can be approximated by agreement levels between redundant annotations (Recasens et al., 2012). While that work applies to general coreference, Hovy et al. (2013a) adopt a categorical approach to events in particular, defining strict identity and two categories of *quasi-identity* – *membership* and *sub-event* – which they claim incorporate the majority of partial identity cases and which, by being distinguished from full identity, assist in the accurate annotation of coreference. Although only treating certain identity and certain non-identity may simplify the task computationally – partitioning references into neat equivalence classes – due to the complexity of event ontology, it is an insufficient account of the phenomenon and may lead to inter-annotator discrepancy.

Changing perspectives When events are referred to within a single discourse by a single author, we may assume coreference is intentional. That is, the discourse introduces events and may refer to them repeatedly with minimal description and expect the reader/perceiver to understand the references. When this constraint is removed, as it is for cross-document coreference, we encounter further complexity: references must be recontextualised according to assumed shared knowledge; authors have different perspectives on what the event is; and language about or knowledge of an event may change over time.

When a discourse first introduces an entity that the reader is familiar with, it may use a canonical form, such as a proper name with minimal disambiguating detail. Most events are not afforded names; they must be described in as much or as little detail as is necessary to specify the referent given knowledge shared between the author and reader. Cross-document named entity coreference (Bagga and Baldwin, 1998; Gooi and Allan, 2004; Haghighi and Klein, 2007) is much simplified by considering only the entities contained in a knowledge base derived from Wikipedia, a task known as named entity linking or disambiguation (Bunescu and Paşca, 2006; Cucerzan, 2007; Hachey et al., 2013). With a much vaster and fuzzily-defined space of notable events, the same quality of knowledge base is not easily constructed; in the next chapter we discuss the poor coverage of events as standalone Wikipedia articles. Alternative to this use of Wikipedia, identifying cross-document event coreference is more feasible for a limited scope of salient events, constrained by type (Bagga and Baldwin, 1999), topical-temporal document clusters (Feng and Allan, 2007, 2009; Bejan and Harabagiu, 2010;

Lee et al., 2012), or salience within a news source (Allan, 2002; Nothman et al., 2012, and this work).

Since canonical forms are rarely available for event reference, the form of reference may differ according to author perspective:

(12) a. Two have died after [an explosion at the Boston Marathon]$_e$.

b. Three have died after [a terror attack at the Boston Marathon]$_e$.

These may be coreferent, despite different estimates of the number of casualties – or that number changing over time – and assessments (known or assumed) of the agency and intent of the explosion. This same event eventually becomes known as [the Boston Marathon bombings]$_e$ with little ambiguity, just as tsunami is used in our hijacking example without explication. But unambiguity may not be relied on over time, as is the case with the attack on the NYC World Trade Center which was unambiguous in news from March 1993 to August 2001; this also exemplifies the need for world knowledge in a successful coreference resolution approach. Finally, modality in event reference presents a further problem for coreference resolution over time. In a single closed discourse, once a hypothetical (or future, or possibly-occurred) event has been introduced, it can be repeatedly referenced. Across multiple perspectives, the factuality of events may change:

(13) a. A Somali pirate group has claimed they [hijacked]$_a$ a U.N. ship.

b. The ship and its 10-person crew was [hijacked]$_a$ by pirates.

c. When Yeltsin [visits]$_f$ in July, ...

d. Despite Yeltsin not being able to [attend]$_f$ due to ill health, ...

An event that is only rumored to occur in Example 13a is presumed by 13b; similarly, a certainly anticipated event in 13c had equally-certainly not come to pass according to 13d. Cross-document event coreference may therefore need to account for the frame of reference of the author, the expectation of changed detail as a story unfolds, and the formation and deprecation of near-canonical designations for notable events. Similarly, tasks evaluating event reference identification should consider a breadth of perspectives, and not merely contemporary news.

2.6.1 Schemas and systems

To estimate how much this complexity affects event coreference in practice, we review shared corpora and give an impression of state-of-the-art performance at replicating human annotations.[39]

[39] Along with the much larger literature on entity coreference, a review of the many metrics applied to coreference resolution is out of scope of this work, limiting our interpretation of quantitative results. Unless otherwise specified, we report coreference performance over gold event extractions.

Although its literature tends to be treated separately, topic detection and tracking (Allan, 2002) – with seven shared evaluations from 1998 to 2004 – should be construed as a form of event coreference detection. The task groups news stories from multiple concurrent sources into partially-overlapping topical clusters.[40] Ignoring its treatment of story segmentation within broadcast news, it labels each document-topic pair with one of *on topic*, *off topic* and, in some evaluations, *brief mention of topic*.[41] Its notion of *topic* effectively encompasses a loose sense of event identity, incorporating a seminal event, all its necessary preconditions and unavoidable consequences, and all directly related events and activities (LDC, 2004).[42] To better address event structure and blurred topical boundaries, the final evaluation considers a hierarchical topic detection task rather than evaluating over binary output (Feng and Allan, 2005).[43] From the earliest evaluations, most approaches to the TDT tasks have revolved around clustering with measures of textual similarity – adopting and adapting standard techniques from information retrieval – and temporal proximity, which is more distinctive to events and the news genre (Allan et al., 1998). Later work distinguishes types of event and incorporates named entities (Yang et al., 2002; Kumaran and Allan, 2004); Makkonen and Ahonen-Myka (2003) extend this with spatio-temporal information, features that are more particular to event content than ad-hoc retrieval. First Story Detection was early found to be most challenging of the TDT tasks (Allan et al., 2000). It has recently seen renewed interest in the context of identifying breaking stories in microblogs where scalable similarity techniques are required (Petrović et al., 2010), and in order to incorporate advances in the processing of event and spatio-temporal reference (Roberts and Harabagiu, 2011).

For MUC-style scenario template extraction, coreference within documents is implicit: systems are expected to extract knowledge and group these partial structured representations from multiple references into a single event template. One approach is described by Humphreys et al. (1997) that involves considering each reference in the order of appearance and checking for attribute and type consistency with prior extractions. A more recent approach models identification, extraction and merging of event references jointly (Reichart and Barzilay, 2012). Bagga and Baldwin (1999) consider the further task of matching coreferent templates across documents, considering those describing resignations, elections and espi-

[40]Regarding topical overlap, in the TDT3 corpus with 180 topics, 34% (excluding brief mentions: 30%) of stories are assigned to exactly one topic, 4% (2%) are assigned to two, with 4% (0.3%) assigned to more than two. The most topically-prolific story, the New York Times' *Clinton urges unified attack on global economic crisis* (5 October 1998) briefly references six topics: November APEC Summit Meeting; Anwar Ibrahim Case; Brazilian Presidential Elections; Russian Financial Crisis; IMF Bailout of Brazil; and G-7 World Finance Meeting.

[41]The evaluations reformulate this task as: Topic Tracking, which finds other on topic stories given a temporally-prefixing sample; Topic Detection, which clusters stories by topic; First Story Detection, which identifies stories that are off topic for all topics previously identified; and Link Detection, document-document pairwise classification of co-topicality. All have parallels in the linguistic coreference literature.

[42]The annotation guidelines also prescribe *rules of interpretation* that explicate the bounds of this definition according to the type of seminal event (LDC, 2004).

[43]Note this change does not affect the annotated data, but evaluation is dependent on the relative positions of stories within the hierarchical output.

onage. They account for vastly lower performance on the elections data as complexity due to sub-events (a general election may incorporate congressional and presidential elections), as well as superficial similarities across instances of that event type, given that elections may have common participants and issues.

ACE05 marks references at a finer textual and typological granularity and coindexes those that should be merged into a single template. It explicitly does not mark structural, temporal or logical relations between mentioned events (Walker et al., 2005), and takes a conservative approach to identity, stating "When in doubt, do not mark any coreference." (LDC, 2005) However, it does coindex mentions where participants are different entities, as with U.N. [chartered]$_b$ and The World Food Program [hired]$_b$, and where they are "modally questionable": Maddux was [killed]$_g$ in Philadelphia. ... Einhorn is accused of [killing]$_g$ Maddux (LDC, 2005). It only allows for coreference between events of identical type, so Maddux was [attacked]$_{*g}$ in Philadelphia may not join that coreference chain. Ahn (2006) applies Florian et al.'s (2004) method for entity coreference[44] to ACE05 events, reporting 66% F_1[45] for a pairwise classification model, against a 29% baseline where all events of the same type are considered coreferent.[46] Tailoring a solution to the task, although still applying a pairwise resolution model and evaluating with different metrics and a non-trivial baseline, Chen et al. (2009) report 72% MUC F_1,[47] and perform slightly worse on a different metric when using spectral graph clustering (Chen and Ji, 2009). Ji et al. (2009) and Li et al. (2011) propose the cross-document aggregation of ACE05-style extractions and explore similarity and diversity among events for user-facing summaries, while avoiding the more explicit and fraught task of coreference identification.

The OntoNotes corpus annotates coreference between all noun phrases within a document, including those referring to events.[48] Where a noun phrase is coreferent with a verbal predicate, the latter is also indexed as part of the same coreference chain (Pradhan et al., 2007b): the annotation guidelines (BBN Technologies, 2011) include an example connecting grew, rose and the strong growth, but do not treat the subject in depth.[49] As the focus of the 2011 and 2012 CoNLL shared tasks (Pradhan et al., 2011, 2012), the task of predicting this

[44]Florian et al. (2004) do not separately report performance for entity coreference, disallowing task comparison under the same algorithm.

[45]The evaluation metric is unspecified: it may be pairwise binary classification or MUC F_1.

[46]He also reports results on the official ACE metric, while substituting gold and learnt pipeline components. Compared to a baseline where no coreference is predicted, error on this metric reduces 40% where gold events are used (absolute 7.5% increase in ACE value), but only 1% with detected events (absolute 0.6% increase).

[47]Similar to Ahn (2006), this represents a 55% error reduction on their baseline that is reduced to a 2% error reduction when using automatically detected events.

[48]Data is annotated in English, Chinese and Arabic.

[49]Note also that by tying reference to syntactic units, the dual reference in lexemes like another is problematic. The OntoNotes corpus treats phrases specified by another as a reference to that specific "other" entity, rather than the generic family of entities indicated by the remainder of the phrase. The guidelines give particular attention to the *generic reference* problem by considering all generic nominal references not coreferent: "Generic nominal mentions are linked to referring pronouns and other definite mentions, but not to generic nominal mentions." (BBN Technologies, 2011)

data has recently received much attention. Pradhan et al. (2011) report 80.9% MUC F_1 agreement (with annotators obtaining 85% and 88% on the adjudicated corpus) on the English newswire portion of the corpus, with about 3% of disagreements resulting from one annotator marking a verb where the other did not. The best system in the 2012 evaluation (Fernandes et al., 2012) achieves 62.7% MUC F_1 on the English newswire portion by inducing complex features over latent "coreference trees", yet, like most participating systems, Fernandes et al. (2012) do not incorporate verbs into this model. Chen et al. (2011) specifically evaluate performance on the subset of coreference chains that include a verbal reference, which constitute 15% of all coreference chains for their evaluation; while the full task includes non-events, this excludes events that are only mentioned nominally.

As noted above, cross-document coreference is difficult without some constraint on the event scope, making the OntoNotes document selection unsuitable for this purpose. To evaluate detection of cross-document event identity and other structural and logical relations, Bejan and Harabagiu (2008) collect and annotate 482 documents from Google News clustered into 43 topics. They implement a variant of the hierarchical Dirichlet process with rich features[50] to learn cross-document coreference without supervision, achieving 90% B^3 F_1 on the annotated corpus which marks selected references to the events considered most important (Bejan, personal communication). Lee et al. (2012) re-annotate this corpus exhaustively with event and entity coreference, distinguishing these two classes of chains for error analysis. They extend the OntoNotes guidelines more particularly to their task, including chains containing only verbs, and report a chance-corrected agreement between four annotators of $\alpha = 0.55$ (Krippendorff, 1980), highlighting the artificially high performance on the original evaluation corpus. Further, they distinguish event references from others in their annotation, allowing them to be evaluated separately. They apply a variant of their top-scoring CoNLL 2011 system (Lee et al., 2011): rather than treating each reference and its candidate antecedents individually, it makes multiple passes over the data, applying a series of rules (or *sieves*) with decreasing precision to merge clusters of references, propagating extracted knowledge about the referent throughout the cluster. To incorporate events and cross-document resolution, they: (a) predict only within document clusters[51] in order to apply single-discourse assumptions relied on for in-document coreference; and (b) following the intuition that event coreference is informed by the involved entities and vice-versa (e.g. Shinyama and Sekine, 2003), they greedily cluster sieve-predicted coreference chains and verbs using a learnt similarity metric that incorporates SRL features. Their approach is thus able to harness the information available in some descriptions of an event and not others, similar to Ji and Grishman's (2008) inference from unlabelled documents to bolster ACE05 event extraction

[50]These include surface lexical forms, part-of-speech, TimeML event class, lexical semantics and SRL-based features.

[51]It is unclear whether the clusters are smaller than those selected by Google News that are the basis of their corpus.

from similar texts. This model improves upon a Lee et al. (2011) baseline with verb lemma matching in all metrics, whether considering events only or all references, achieving a MUC F_1 of 67.8% (62.7% for events only) and a CoNLL F_1 of 55.9% (54.8%). Although appropriate for some applications, the use of Google News clusters creates an unrealistic cross-document coreference task: the documents are already linguistically and temporally clustered, presumably employing TDT techniques; they are more likely to refer to the same events in similar ways than in a more general or longitudinal approach to cross-document reference.[52]

We also note that TimeML (Pustejovsky et al., 2003b) annotates an *identity* relation between events, among others that may be a result of structural and logical relations (*simultaneous*, *during*, *overlaps*, *before*, etc.), but it is not the particular focus of any work in our knowledge.

While great progress is being made on computational identification of coreference, theoretical issues in the determination of abstract referents' identity from reference continue to make it a challenging task, even for linguistically competent humans. In particular, as suggested by Recasens et al. (2011), conservative notions of identity may be over-specified relative to how reference is utilised in communication. Cross document coreference presents adds problems related to changing perspective onto the in-document task, and for broad application requires a model that is not biased to contemporary points of view. The event linking task introduced in Chapter 4 thus poses an approximation to coreference within a model that considers references many years after the referent event occurs.

2.7 Conclusion

The challenges presented here merely skim the surface of research into event characterisation. We might otherwise detail work on identifying event arguments that are not syntactically local, or that are implicit; or the great mass of work in identifying temporal and logical relations between events; or models of the factuality of an event reference.Understanding events often involves common sense and world knowledge, and we have only briefly touched on work to acquire this information. And while we focus on reference to mostly past or anticipated events, another branch of literature analyses past events in order to predict potential events in the future. We have not considered the particular problems of event reference in languages other than English or in harnessing cross-lingual reference.

We no longer need mention a final challenge in computational approaches to event reference: the vast arena of related work. This chapter has attempted to identify structure and common ground among disparate approaches to event detection and characterisation, while providing a foundation and context for the remainder of this thesis. It first considered some

[52] In a similar critique, in performing a related paraphrase detection task, predicate-argument alignment (Roth and Frank, 2012), Wolfe et al. (2013) evaluate on the Lee et al. (2012) corpus assuming coreference implies paraphrase, and suggest the dataset is easy (a trivial same-lemma baseline obtains 63% on this task; their system achieves 74%) in comparison to a translation-based corpus (42% baseline; 59% system).

applications that motivate event characterisation from an engineering perspective. Then it reviewed the major attempts to schematise aspects of event language, dividing them into those that focus on events as abstracted predicate-argument structures, others trying to understand the relationships between event references, and still others using text to model types rather than instances of event. We have selected three features that make event reference challenging to process, underpinning on our observations in coming chapters.

We intend to depart from traditional IE to model a broad range of news event references – rather than application-dependent types – and yet are faced with the question of whether events should still be grouped into types, and how best to do so. This is informed by our study of the approaches to type used in MUC and ACE evaluations, and an evaluation of event processing literature that ranges from being entirely conditioned on lexical choice to being completely independent of lexicon and ontology. In the following chapter we explore this question, first questioning whether event type information can be obtained from Wikipedia, then considering an ACE05-like typology, and a novel approach to a structured typology, before ultimately doing away with traditional conceptions of event type to focus on reference.

When a group of annotators set out label event references in text, it is immediately apparent that annotators often recall different references, and have admit different references that are distant from the prototypical event (cf. Filatova and Hatzivassiloglou, 2003). This contrasts with marking all peoples' names, for instance; properties of event ontology and reference make some references more readily identified and agreed upon than others. Hence we analyse inter-annotator disagreements in the ACE05 corpus that point to some features that make an event reference less noticeable than others. We also describe means by which existing approaches to event reference select the most notable instances, relying on pre-specified type filters, discourse structure and redundancy of reference. Similar concerns cause us to ask: when communicating in public, what events are we expected to be familiar with? This directs our concern towards *newsworthy* events, how that notion is defined by particular sources of news, and how news text serves to convey event knowledge.

The third challenge we discuss, event identity and coreference, is the focus of our proposed event linking task. The lack of rigid designators for most events sets them apart from many of the other entities we refer to in discourse, such that event language is almost always entangled in the semantics of the reference. Together with the intricate structure of events and the imprecision that may be used in reference, this suggests that annotating only precise identity is too strict, while similar problems are inherent in typologies of other event-event relationships. We have therefore reviewed some of limitations of the literature in this area, and find in particular that broad-coverage cross-document coreference remains to be evaluated in a manner that may connect references from vastly different times and perspectives. With insights gleaned from the upcoming chapter's experiments into event characterisation, this work establishes the event linking task in which such evaluation is possible.

Chapter 3

Experiments in event representation

Events of popular interest are presented to the public in the form of *news text*: each document with its headline and body text, possibly a timestamp and reporter location, allocated to a broad editorial section, and given prominence in a publication according to editorial valuation. These are features of traditional print text that have been transferred online: with some exceptions – notably finer topical groupings or mentioned entities in publications like the online New York Times; and social media distribution and feedback – these same features define a reader's news navigation and article selection. In particular, readers cannot trivially elect to view or follow articles reporting an event with particular characteristics, such as all references to future acquisitions of exploration companies. The previous chapter discussed event detection and extraction tasks that could underlie such a navigation system; a chief difficulty remains in the grouping of salient events into useful types. Where much previous work in event extraction targets specific domains of event and text, we make fewer domain assumptions:

- we target a news archive with broad topical coverage;
- we should be able to identify any type of event that is frequently reported; and
- we will not rely on the sought events being redundantly reported in multiple news sources.

This chapter presents two experiments into labelling news events with their type (such as *acquisition* or *natural disaster*), and a brief analysis of Wikipedia's suitability as a repository of knowledge for event characterisation. Appreciating that event ontology is intricately linked to world knowledge, we consider Wikipedia – a free, growing and semi-structured repository of world knowledge – could be exploited to inform event characterisation, in a similar vein to our work with named entity recognition (Nothman et al., 2008, 2009, 2013). Yet where

famous people and organisations are well-covered by Wikipedia's articles, the types of events that attract similar Wikipedia coverage are skewed, for instance, towards extensive historical events such as military operations or elections.

Resorting to manual annotation of news text, we take the ACE05 (NIST, 2005) approach to event reference as our point of departure due to its broad-coverage typology of interesting events, although we apply it to broader-domain news text and attempt to address some of its flaws in terms of type sparsity and the interdependence of its annotation layers (see Section 2.5.1). Modifying its set of coarse types, we investigate whether human annotators are able to readily agree on the identification and classification of event references. Our results suggest the need for structured types, and a better understanding of the centrality of an event to a news story. This guides the second experiment towards a dynamic hierarchy of types, which annotators apply to the main event triggering a news story, and a background event closer to the notion of topic applied in TDT (Allan, 2002). Hierarchy provides better coverage and greater purity within type groupings, yet we find the placement of types within the hierarchy is not definitive, although one might be designed for a particular application. This second annotation also shows that topics are often not construed as single events. Nonetheless, a journalist may provide a sense of topic by referring to several background events; it is such references that become the focus of the event linking task of the next chapter, which further builds upon the centrality of an event to a news story. As such, we do not pursue these approaches to event detection further, but they give us a better understanding of the challenges in characterising the events – in contrast with other entities – that underlie broad-domain news text, and provide a basis for the design of the event linking task (next chapter).

The contributions of this chapter are thus established through three distinct studies: The first is an analysis of Wikipedia's applicability to general news event detection (Section 3.1). The second (Section 3.2) adapts an existing event identification approach to annotate a news corpus with broader topical coverage; its analysis leads to new empirical insights into annotator disagreement in the ACE05 corpus. Our final study (Section 3.3) proposes and qualitatively assesses a novel method for labelling event types, employing a dynamic hierarchy to delay specification of typological granularity. It also takes a new approach to event salience in news, characterising a news report as a reference to *update* and *topic* events.

3.1 Events in Wikipedia

An event is implicitly important – for a particular audience – if it is reported in popular news. Similarly, we may consider the set of events pertinent to an encyclopædia. Among encyclopædias, Wikipedia is a useful source of world and linguistic knowledge: it is large but restricted in genre; it grows and is updated; it combines structured and free-text content; and it is freely accessible (Hovy et al., 2013b; Medelyan et al., 2009; Nothman, 2008). Where

Subtype	Popular sample					Random sample				
	Class	Instance	Subtopic	Total	Recall	Class	Inst.	Sub.	Total	Recall
Sports	6	17	7	30	30	11	41	9	61	55
Conflict	0	8	0	8	6	0	7	0	7	3
Disaster	0	3	0	3	2	0	2	0	2	1
Performance	0	0	0	0	0	2	3	0	5	4
Other	4	4	0	8	2	4	10	0	14	4
Total	10	32	7	49	40	17	63	9	89	67

Table 3.1: Sub-type frequencies for English Wikipedia articles manually classified as events by Nothman et al. (2013), with recall of the *Event* type using their best classifier.

annotating news with event characteristics may be costly and difficult, using existing latent annotations in Wikipedia – by way of its semi-structured content – could yield models of event characterisation at a much smaller cost. Having successfully applied this approach to learn the extraction of named entities (Nothman et al., 2013), we consider whether it may also be applicable to inform event-related processing tasks.

Each article in Wikipedia focuses on a single topic; some topics correspond to events. The structured content of these articles and the disambiguated links to them could be exploited to make generalisations about event reference.[1] Yet this is greatly limited by event articles' infrequency and typological skew.

3.1.1 Distribution of event articles

We may estimate the type distribution and frequency of event-related articles by considering a manually-labelled sample. In Nothman et al. (2013) we describe sampling two sets of articles from English Wikipedia and manually classifying them into fine-grained entity-oriented types. Of the 2322 "popular"[2] articles labelled, 49 (2%) are labelled *Event* or a subtype. Of a uniform random sample of 2531 articles from a English Wikipedia snapshot,[3] 89 (3%) are labelled as events.[4] On this basis, we may estimate that 140,000 Wikipedia articles (of four million) are about events. This population may suffice to learn a model of event reference, if its distribution can approximate our target domain, a broad coverage of newsworthy events; hence we drill down into its content.

[1]This goal is in contrast to other work in extracting temporally grounded event knowledge (Ahn et al., 2006; Kuzey and Weikum, 2012) or temporally bounded relations (Wang et al., 2010) and learning their generalised identification (Garrido et al., 2012; Surdeanu et al., 2011) from Wikipedia.

[2]This sample was made up of March 2009 English Wikipedia articles that had inter-language equivalents in all ten of the largest Wikipedias (German, French, Italian, Dutch, Spanish, Japanese, Polish, Portuguese and Russian) and were also among the most frequently-accessed pages of August 2008 or those with the highest number of incoming Wikipedia-internal links. See Tardif et al. (2009).

[3]Dated 30 July 2010.

[4]Events represent a similar proportion of articles in other languages examined in Nothman et al. (2013). The manually-labelled datasets are available from `http://resources.schwa.org`.

The classification detailed in Nothman et al. (2013) distinguishes between some event types such as sports events, conflicts, etc. It does not make a distinction between articles pertaining to a class of repeated event (e.g. Summer Olympic Games or Halloween) and those pertaining to an instance (e.g. 2000 Sumer Olympics) or even those discussing a non-event subtopic of an instance (e.g. Athletics at the 2000 Summer Olympics). We augment the data with this classification and show the subtype distribution in Table 3.1. There we see that only a third of articles labelled *Event* actually consider a single event, and more than half of those – almost two-thirds in the random sample – are sports events, be they series or contests. This leads us to believe that non-sports event instance articles account for approximately 0.9% of English Wikipedia, or ≈35,000 articles. Among those labelled in our sample are a census, a lunar eclipse, a military operation, a riot, a plane crash and the Great Plague of Vienna. Not all events are in recent history; some, such as the mythical Norse apocalypse, Ragnark, are yet to happen. Wikipedia policy tends to disprefer very small articles on related topics, so many event instances are themselves collections of sub-events, such as 2009 Formula Le Mans Cup season, Taste of Chaos Tour 2008, Mid-December 2007 North American Winter storms, Vietnam War and Atomic bombings of Hiroshima and Nagasaki. Although these are spoken of as single events, often with designating names, from the perspective of news reporting most are considered in terms of their sub-events.

Predicting the *Event* label[5] on the random sample achieves a recall and precision of 80% under ten-fold cross-validation of a state-of-the-art 18-class classifier[6] trained on a combined (popular and random) dataset (Nothman et al., 2013). However, Table 3.1 indicates that only the *Sports* and *Performance* (mostly concert tours) subtypes exhibit high recall: recall for non-*Sports* subtypes is 40%. Although our training instances are few, this may indicate that event articles are not homogeneous enough – in terms of the feature families found effective for Wikipedia article classification – to be easily distinguished from other article types.

We see that Wikipedia's event articles are diverse topically, yet skewed: in quantity, Wikipedia tends towards event types that are notable but enumerable, such as sport contests; in granularity, towards events that are named collections of newsworthy sub-events; in popularity, towards lengthy, complex events such as wars; and in time, towards recent events.[7] Some of these articles include structured information such as timespan and primary participants that could be harnessed to interpret references to such events, yet such an approach is encumbered by a mismatch between the more atomic events described in news, and the broad topical coverage provided by Wikipedia.

[5] We do not attempt to predict its subtypes as they have too few training instances.

[6] We use a ℓ_2-regularised logistic regression classifier with one-versus-all multi-class operation. It learns from features corresponding to the categories assigned to the article in Wikipedia, and bags of words from: the article's title, first sentence, first paragraph, and included template names, keys and values.

[7] The focus on recent events is mostly a function of circumstance rather than editorial policy, although the need to contextualise historical events for a modern audience may prefer their placement within a broader topical or chronological coverage.

3.1.2 Structured content and event articles

Given a set of Wikipedia articles describing events, linguistic knowledge can be inferred by utilising a combination of the free text and structured content in Wikipedia. The exploitable structure takes two main forms: structure within the articles, and links to the articles.

Most Wikipedia articles contain some structured content, such as *infoboxes* that provide MUC-like templated content, and categories that may incorporate some ontological information but are generally noisy (Medelyan and Legg, 2008; Ponzetto and Strube, 2011). As of January 2013, a generic infobox for news events is instantiated 276 times and may report details including the date and time of occurrence, participants, causes and outcomes of an event. An infobox for military conflict has over 11,000 instances, with more particular fields such as for combatants. By exploiting the redundant presence of facts in infoboxes and free text within the same article, Wu and Weld (2007) train information extractors; with the further ability to unify infobox templates by identifying their equivalent fields (Wu et al., 2008), this results in MUC-style template extraction for individual relations expressed in Wikipedia. Later work enhances these extractors by bootstrapping their training from the web (Mintz et al., 2009; Hoffmann et al., 2010), while the Slot Filling task specifically evaluates the extraction of such knowledge from textual corpora other than Wikipedia (Ji and Grishman, 2011), but only considers attributes of person and organisation entities,[8] not events. Kuzey and Weikum (2012) use similar methods to extract temporal knowledge specifically about events, while Freebase (Bollacker et al., 2008) contains such structured knowledge on 114,000 events using a mix of automatic and manual curation. While extractors of frequent facts such as time of occurrence may be learnt using these methods, we know of no work that specifically targets extraction of other fields of event article infoboxes; to a great extent these are still affected by Wikipedia's topical biases in their applicability to more general event reference.

Where links point to event articles, we hypothesise that they may exhibit various ways of referring to a single event. The anchor texts linked to a Wikipedia article attest to aliases for the corresponding entity, as applied to named entity linking (Cucerzan, 2007) and learning named entity recognition (Nothman et al., 2013), although there we find they can also reduce performance since they are noisier aliases than canonical title forms. We might similarly use them to identify event coreference or references to particular types of event, however since we are not only interested in *named* events, we specifically seek non-canonical forms of reference, generally employing common nouns or verbs. We consider the set of links targeting articles that were manually identified as event instances from a sample of Wikipedia articles.[9] We collect the set of anchor texts found within paragraph text, ignoring those consisting only of title-case words, stop-words and numbers. The two examples shown in Figure 3.1 exhibit

[8] Up to and including the 2013 evaluation.

[9] The sampling and labelling of event articles is described in more detail in the previous subsection. The present analysis is performed on a Wikipedia snapshot from 30 July 2010.

Caucasian War	**1992 Summer Olympics**
campaigns	'92 games
Caucasian resistance	1992 Barcelona games
conquer the Caucasus	1992 Olympic games
conquered by Russia	1992 games
conquest of the Caucasus	Barcelona Olympic games
expeditions	Barcelona games
local Caucasian peoples	Eight years later
post-war	first Olympic Games
Russian conquest of the Caucasus	Four years later
Russian conquest of the Northern Caucasus	games
Russian military policies in Caucasus	last Olympics
Shamil rebellion	Olympic final 1992
waging the war	reception of the Olympic torch in Barcelona
war in the Caucasus	Summer Games since 1992
	Summer Olympic games
	That summer
	that year's Summer Olympics

Figure 3.1: Non-name anchor texts for links in English Wikipedia to two articles that are manually classified as event instances.

some of the diversity in referring to each event: links to Caucasian War include predicates conquer, wage, campaign and rebel, while mentions of 1992 Summer Olympics are often implied by way of temporal references and relative expressions (e.g. [. . . she travelled to her] first Olympic Games). Both examples exhibit sub-event references, such as Shamil rebellion, expeditions in took part in a series of *expeditions* against the Dagestani tribes, and reception of the Olympic torch in Barcelona, while instances like 1992 Barcelona games reflect inconsistent use of capitalisation in Wikipedia. Further, the event reference implied by local Caucasian peoples must be understood in context: in . . . the territories acquired by Russia in a series of wars with the *Ottoman Empire*, *Persia*, and the *local Caucasian peoples*, each theme of wars is separately linked. Hence link anchors present multiple ways of referring to an event or its subevent that could be exploited by an event detection or coreference system.

However, Cucerzan (2007) suggests that link anchor texts are too noisy to rely on a single instance as support for an informative alias. Our procedure to filter out named event forms retains mostly infrequent aliases: for 63 event instance articles identified in a random sample, we find an average of 18.6 incoming links from body text, with 4 distinct (case-sensitive) anchor texts, of which 2.0 anchors (in 7.3 links) pass our filter to remove named event forms. Of these, 1.4 anchor texts occur only once; 0.4 occur twice or thrice. We can estimate

around 60,000[10] non-named anchor texts supported more than once in Wikipedia that link to event instance articles. Yet only 12 (20%) of the articles sampled have a non-named anchor supported more than once; and only 4 have two, leaving about 6300 Wikipedia articles from which we can learn a model of non-named event coreference. Combined with our knowledge that the event space covered by Wikipedia articles is very skewed, the sparseness of non-trivial link anchors limits the feasibility of exploiting links to learn event coreference or type knowledge.

3.1.3 Conclusion

We have investigated Wikipedia as a source of linguistic and world knowledge regarding events and their reference, finding it not as straightforward to use as for some entity-related learning tasks. Although there may be other avenues for exploiting Wikipedia's event data – such as collecting a knowledge base of temporally-grounded events (Ahn et al., 2006; Kuzey and Weikum, 2012), or using it to model event salience for summarisation (Biadsy et al., 2008) – the structured information available tends to focus on very broad events such as wars, on the one hand, and essential facts such as birth, death and marriage for particular entities. These classes of events do not neatly accord with news event granularity, so linguistic knowledge in references to these events are unlikely to broadly generalise to news events. This leads us to consider more costly annotation approaches to model news events.

3.2 Type-driven annotation experiment

Without requiring a definition of *event*, annotators may be able to consistently mark all events *of a particular type*; thus typologies are able to effectively constrain an ill-defined annotation space. In contrast to information extraction tasks that follow MUC in targeting a few specified types of events, we are driven by the content of a news corpus, aiming to group the events it describes into types. Where TimeML and SRL provide event characterisations suitable to broad domain coverage, they focus on low-level distinctions and ignore the relative importance of referenced events. In particular, we are interested in whether annotators are able to agree on what is a reference to a news event of a particular type, and how well a small but broad typology is able to cover the diverse range of events constituting the news. The task is analogous to named entity recognition (NER) with events; it also approximates a sub-task of ACE05 event detection (LDC, 2005), ignoring aspects such as participant identification. Part of the intention of a simplified, underspecified task is to allow minimally-trained annotators to produce a redundantly-annotated corpus. Our pilot annotation suggests, however, that searching a document for relevant events still takes a lot of annotator effort (in comparison

[10]This is calculated as 0.6 anchor texts with event instances constituting 63 out of 2531 articles sampled from 4M Wikipedia articles.

The United Nations says Somali gunmen who [Conflict; realised hijacked]$_a$ a U.N.-[Transaction; realised chartered]$_b$ vessel carrying food aid for [Disaster; realised tsunami]$_c$ victims have [Conflict; realised released]$_d$ the ship after [Conflict; realised holding]$_{e \subset a}$ it for more than two months. The World Food Program [Transaction; realised hired]$_b$ the Kenyan vessel to carry 850 metric tons of rice [Transaction; realised donated]$_f$ by Japan and Germany. The ship and its 10-person crew was [Conflict; realised hijacked]$_a$ by pirates as it sailed from Kenya to Somalia in June.

Figure 3.2: A possible annotation of the hijacking story of Chapter 2 under our type-driven annotation scheme.

to seeking entity references) for low agreement. We argue that this is due to fundamental problems in applying a flat, broad-coverage typology to textually fine-grained event references.

3.2.1 Task definition

We adapt and simplify the ACE05 event detection task which seeks textual references to events of broad-but-predetermined type, and extracts their arguments, tense, polarity, modality and genericity (NIST, 2005; LDC, 2005). An example annotation is shown in Figure 3.2. Our annotation differs from ACE05 in a number of ways, as prescribed in the annotation guidelines reproduced in Section A.1:

Corpus sampling ACE05 biases its corpus to targeted entity types (Walker et al., 2006); we instead sample the corpus for annotation at random from Sydney Morning Herald stories of 2009. This adopts some of the biases already present in the publication, including a local focus. Only news stories are considered: annotators are instructed to only annotate articles which approximate the inverted pyramid style common to news journalism, to the exclusion of opinion, analysis, reviews, etc.

Reference granularity Although our use case primarily targets document-level retrieval, we identify event references at finer granularity to focus the annotation and enable local detection and user-facing snippets. Like ACE05, we mark individual words or phrases as event anchors, but consider the event reference's extent to be a single sentence since precisely determining a referential segment may itself be controversial. Practically, this means that a single event mentioned multiple times needs only be tagged once per sentence and that evaluation at the sub-sentence level is not meaningful.[11]

Unlike ACE05, we allow a single word to anchor multiple events, for instance enabling a *murder* to be characterised as both an attack and a death. Another example: had we the typological granularity of ACE05, we could mark references to a player defecting to another

[11] We do not go into detail on anchor word selection, as we believe it is often difficult to select a single word referring to an event, or to describe its consistent selection to minimally-trained annotators.

team as both the end and start of their employment. In this case we consider the referent events to be distinct (and they may be coreferent with distinct chains of reference within the document), although indicated by the same text. While this has the potential to complicate evaluation,[12] we think it is truer to the relationship between lexical semantics and event reference.

Arguments We do not mark the arguments of events (their participants, time and place) on the basis that: (a) this requires a prior specification of argument roles for each type; and (b) using the binary presence of an entity along with an event type might suffice for an initial event retrieval system. Although we briefly experiment with a single, underspecified *has named entity participant* attribute per event, it is often difficult to determine the scope of *participant*. For example, an individual politician mentioned in a story with a legislation event may not be a focal participant of that event.

Typology As with popular NER schemas (Chinchor and Robinson, 1998; Tjong Kim Sang, 2002), we do not attempt to build an exhaustive, detailed taxonomy, but focus on broad categories of event. We adopt a variant of ACE05's eight *coarse* event types. Their division into 33 fine-grained subtypes results in very few instances of some types: despite the sampling biases, more than half of them have fewer than 50 instances in the 600 ACE05 training documents, with low inter-annotator agreement for some infrequent subtypes,[13] despite its detailed annotation scheme and annotator training regime. Moreover, since all events marked in the ACE05 annotation must be labelled with a subtype, the coarse types are in fact the union of their subtypes; here we instead describe the types of interest and allow miscellanies excluded by ACE05, such that:

- a corporate expansion or establishment of a subsidiary is included under the *Organisation Lifecycle* type, where ACE05's *Business:Start-Org* requires detailed distinction to only annotate the birth of entirely *new* organisations;
- bidding for a transaction, the bid's acceptance, and the transaction's settlement – which may be difficult to distinguish in textual reference – are incorporated under the *Transaction* type;
- the release of a hijacked vehicle, as with the surrender of a hostage-taker, are included within *Conflict*, though a surrender may also be annotated as a *Justice* event. In ACE05, the hijack and release of a vehicle would be marked *transaction:transfer-ownership*.

[12] Inter-annotator confusion may be harder to evaluate where units may infrequently have more than one annotation, but this is already difficult given that the anchor word to mark is underspecified.

[13] For the newswire texts, disagreement was more frequent than agreement for *Personnel:Nominate* and *Business:Merge-Org*, due in part to their infrequency.

Type	Subsumes ACE05	Example text
Conflict	*conflict*	scam, attack, threats
Correspondence	*contact*	meeting, warned
Disaster		explosion, global financial crises
Employment / Award	*personnel*	defecting, won, loss
Finance		risen, dividend
Governance		intervention, inquiries, ban
Justice	*justice*	seized, sentence, summoned
Lifecycle	*life*	born, back injury, die
New Release		published, report, announced
Organisation Lifecycle	*business*	relaunch, expansion, takeover
Real Estate / Development		rebuilt, new, duplication [of a highway]
Sports Match		match, ride, [beat France and] England
Transaction	(partially *transaction*)	pay, funding, fiscal stimulus

Table 3.2: Event types considered in our type-driven annotation. We also indicate ACE05 event types wholly subsumed by our types (cf. Table 2.2; *movement:transport* is not subsumed), and examples of event reference anchors marked in our corpus. See the annotation guidelines (Section A.1) for prescriptive type definitions.

To develop this typology, we examine a large sample of first sentences from Sydney Morning Herald articles and their most frequent verbs to identify key events, compare these to ACE05 (NIST, 2005) and OpenCalais (Reuters OpenCalais, 2009) event typologies, and revise through a pilot annotation of 18 articles. The full set of types annotated and a summary of their relation to ACE05's typology is shown in Table 3.2.

Newsworthy events in a broad-domain corpus include many not covered by ACE05's type system, such as real estate developments and product releases. The following examples illustrate some of the new annotation types:

(14) a. Jean Le Cam was running third in the [Sports match Vendee Globe] race around the word when his boat, VM Materiaux, [Disaster overturned] in high winds west of Cape Horn.

b. Voters have swept Japan's conservative government [Employment from office], the first exit [New release polls] from yesterday's elections show.

c. Nick Xenophon said principles the Federal Government [Governance adopted] last year for investments by foreign government-owned enterprises were not working.

d. Century Funds Management, a subsidiary of the listed Over Fifty Group, has given the thumbs-up to the [Real estate new] Chatswood to Epping rail link after the $2 million [Real estate upgrade] of 9 Help Street, Chatswood (pictured), located nearby.

Because we do not mark the arguments of each event, we remove and rearrange some ACE05 subtypes that are only useful with knowledge of the participant entities: we remove *Movement:Transport* which is overloaded with many meanings; and we include corporate takeovers in *Organisation Lifecycle* alongside mergers and closures, although they are classified in ACE05 as *Transaction:Transfer-Ownership*.[14]

Factuality As well as events that have happened or are happening, event reference encompasses hypothetical, anticipated, rumoured, explicitly negated and other modalities of referent. It may be useful to precisely select a category of factuality in navigating the space of events, and FactBank (Saurí and Pustejovsky, 2009) schematises the evidential aspects of English propositions on top of fine-grained TimeML tense and aspect annotations in TimeBank (Pustejovsky et al., 2003a); ACE05 provides a simplified coverage of this content, annotating polarity (positive or negative), tense (past, present, future) and modality (asserted or other) (LDC, 2005). To maximise utility while simplifying the interpretation of these semantic categories for minimally-trained annotators, we cut this back to a binary decision of whether or not the event has been *realised* according to the news story. Thus absorbs tense, aspect, modality (grammatical and attributional) and polarity in such a way that users might distinguish events that are reported as having happened or presently happening.

The following example illustrates mentions of three employment events: a realised abandonment of an electoral position, a later re-nomination to that position, and a future election win.

(15) Mr Dutton [Employment; realised walked]$_a$ away from the marginal seat to [Employment; realised seek]$_b$ [Employment; not realised preselection]$_c$ for a safe seat on the Gold Coast... Mr Dutton had some chance of [Employment; not realised holding]$_c$ on because he still had a strong personal following, despite [Employment; realised abandoning]$_a$ the seat.

It is clear from the example of preselection that event realisation is not always evident from grammar, but may require the full discourse to decide: out of context, the first sentence could imply a *realised* preselection event; this necessitates identifying coreference and, for the present example, marking figurative references.

Coreference Annotation of within-document coreference enables aggregation of sentential context across multiple mentions of the same event. Hence where multiple references in a document are to the same event, they are labelled as such; we notate this with subscripted indexation of references as between walked and abandoning in Example 15. We extend upon ACE05's strict coreference annotation by allowing annotators to mark an event (a chain of

[14]Note that ACE05 arguably makes similar exceptions: the arguable transaction involved in employment becomes a *Personnel* event; transfers of property that involve notable movement are instead transportation events (LDC, 2005).

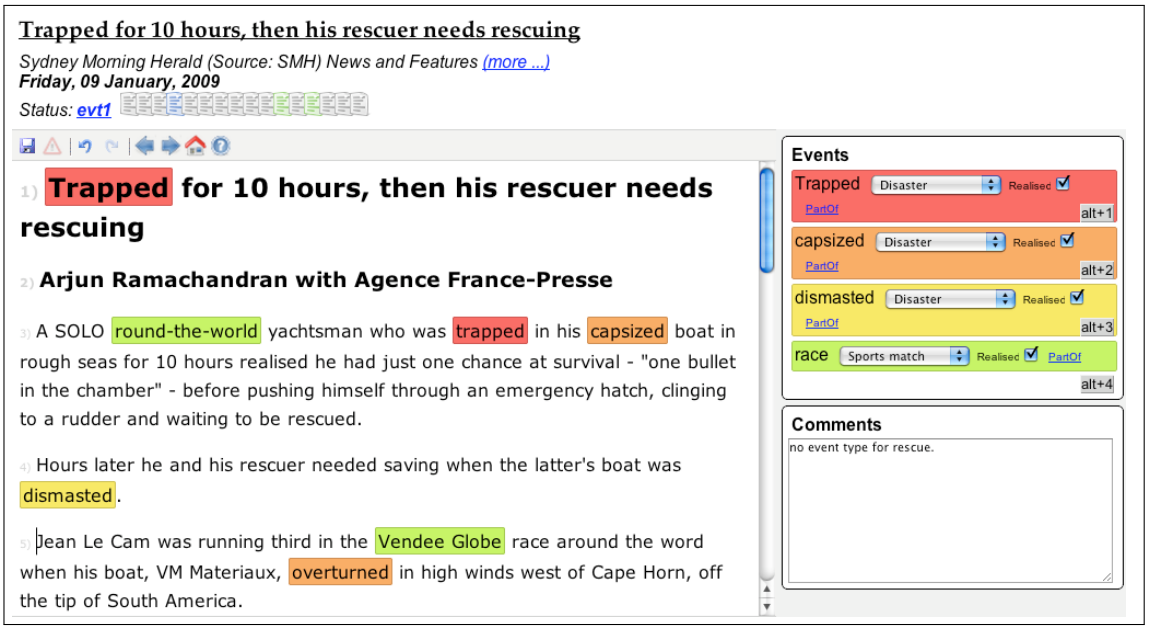


Figure 3.3: A screenshot of the annotation tool for the type-driven annotation

references) as *part of* another referent. This encompasses both sub-event and member relationships (see Hovy et al., 2013a) as in the following examples, in which *part of* is represented by $\subset$:

(16) a. . . . the [Conflict, Lifecycle; realised murdered]$_a$ Frisoli brothers . . . they had been [Conflict; realised bashed]$_{b \subset a}$ with a blunt implement and [Conflict; realised stabbed]$_{c \subset a}$.

b. . . . winning their first five successive [Sports Match; realised Tests]$_a$, including against [Sports Match; realised South Africa]$_{b \subset a}$ and [Sports Match; realised New Zealand]$_{c \subset a}$.

c. . . . two consecutive quarters of negative [Finance; realised growth]$_a$. The state registered minus 0.4 per cent [Finance; realised growth]$_{b \subset a}$ in the December quarter . . .

Explicitness We allow references to employ figurative language but, like the TempEval-2010 TimeML annotation (Saurí et al., 2009), only mark references to specific events, such that we mark become a victim but not rationalisation as a reference to a non-realised *Organisation Lifecycle* event:

(17) Can AMP survive as an independent group or will it [Organisation lifecycle; not realised become] a victim in the inevitable rationalisation that will sweep the insurance industry?

Annotation interface Annotators are provided with a web-based tool shown in Figure 3.3 which we had initially developed for named entity linking annotation – with extensibility in mind – and customise to this task. Each reference may be annotated by selecting a word and using the mouse or keyboard to add it to an existing coreference chain or to create

a new one, each chain (corresponding to a single referent event) being assigned a distinct colour. Attributes are assigned to each event in a sidebar, with a comment field for the entire document. We only annotate documents that already have manual named entity linking annotations to complement the event content.

For the type-driven task, two computational linguistics PhD candidates marked up 103 and 109 documents respectively, with an overlap of 15 documents.[15] Despite a coarser annotation granularity – in both text and typology – than ACE05, annotations were still sparse, and produced moderate inter-annotator agreements, to the extent that agreement is reliable over such a small sample.

3.2.2 Inter-annotator agreement

In our task as in ACE05 it is difficult to aggregate inter-annotator agreement directly. Token annotation is underspecified in both schemas, although ours more so, leading to sparse agreements.[16] We therefore assess inter-annotator agreement as a sentence-level (or document-level) binary decision of whether it refers to an event of a particular type, in accordance with Naughton et al. (2010). This fails to account for many aspects of the task, including:

- multiple events of the same type marked in the evaluation unit;[17]
- type confusion for an agreed event;[18]
- the relative salience of references and referents: one annotator missing an adverbial reference is considered no less an error than missing a typical verbal reference; and
- marked attributes other than type, particularly in ACE05.

Though the sample is very small, from the raw agreement counts in Table 3.3 we can see annotators often disagree on the identification or classification of events. Annotator B is alone in identifying *Correspondence*, *Disaster* and *New Release* events, such as those in the following examples:

(18) "Ninety-five million tonnes will be rolled out when the market needs it," Mr Forrest told a $[_{\text{Correspondence}}$ conference$]^{B}$ of the Securities and Derivatives Industry Association in Sydney yesterday.

[15] These counts – and all statistics presented here – exclude the five annotator training documents.

[16] Micro-averaged event type (and subtype) F_1 agreement between the first-pass annotators for the full ACE05 training corpus (Walker et al., 2006) is 79% (75% for subtypes) at the document level, 65% (64%) at the sentence level and 57% (57%) at the anchor-text (i.e. token) level. This 8% drop from sentence to token accounts for both anchor choice and disagreement in the number of events of the same type.

[17] In our annotation, the number of distinct events of the same type per document is 2.1 for the median type. Over the ACE05 training corpus (Walker et al., 2006) the equivalent statistic is 1.7 for subtypes, and 2.7 for coarse types.

[18] Our schema does not strictly allow us to determine when annotators have identified the same referent, even in token-level evaluation, since the anchor to mark is underspecified and multiple references with the same attributes may appear within a sentence. In ACE05 the specification of event participants makes this less problematic.

Type	Documents				Sentences				ΔF_1
	$\|A \cap B\|$	$\|A \backslash B\|$	$\|B \backslash A\|$	F_1	$\|A \cap B\|$	$\|A \backslash B\|$	$\|B \backslash A\|$	F_1	
Conflict	4	0	4	0.7	10	1	18	0.5	-0.2
Correspondence	0	0	5	0.0	0	0	17	0.0	0.0
Disaster	0	0	2	0.0	0	0	3	0.0	0.0
Employment / Award	2	3	1	0.5	5	5	3	0.6	0.1
Finance	2	3	0	0.6	4	12	5	0.3	-0.3
Governance	0	1	3	0.0	0	1	7	0.0	0.0
Justice	1	0	2	0.5	5	1	3	0.7	0.2
Lifecycle	2	0	0	1.0	7	1	3	0.8	-0.2
New Release	0	0	5	0.0	0	0	5	0.0	0.0
Organisation Lifecycle	2	3	0	0.6	1	12	1	0.1	-0.4
Sports Match	1	0	1	0.7	1	0	2	0.5	-0.2
Transaction	4	2	1	0.7	7	8	10	0.4	-0.3

Table 3.3: Agreement and disagreement counts for our type-driven annotation and the derived F measure (F_1) for document and sentence-level binary event type identification. Here A and B are the sets of (unit, type) pairs produced respectively by our annotators, with $F_1 = \frac{2|A \cap B|}{|A|+|B|}$. ΔF_1 is the gain in F measure when moving from document to sentence-level evaluation.

(19) The two poor divisional performances were triggered by the harsh $[_{\text{Disaster}}$ recession$]^B$ in New Zealand and by the $[_{\text{New Release}}$ relaunch$]^B$ of the Dick Smith brand to compete better with the market leader, JB Hi-Fi.

Excluding these types, our annotators produce the same quantity of distinct document-level type annotations, but with substantial disagreement: for each type there are at least as many documents annotated in disagreement as there are documents annotated in agreement. Except where document-level agreement is already poor, agreement measured with F_1[19] at the sentence level is consistently much lower than document-level agreement (see Table 3.3, column ΔF_1). Some of these errors may result from insufficient specification in the annotation schema, or real differences in interpreting the annotated texts; others result from the difficulty of identifying all relevant event references.

Various types of disagreement can be identified in the story shown in Figure 3.4, exemplifying the difficulty of such an annotation task. In sentences 1. and 4., the transfer of money into a fund is disputed as either a *Finance* or *Transaction* event, according to the respective annotations of A and B; our annotation guidelines provide no clear reference for such an example. Annotator B alone marks an investigation in sentence 1.; A alone marks the *Employment* event of hiring administrative oversight (sentence 2.). Both seem to be clearly

[19]Equivalently, Dice's coefficient over the set of (sentence id, type) pairs.

	Annotator A	Annotator B
1.	On June 30, managers acting for Trio Capital $[_{\text{Finance}}$ poured$]^A_a$ \$47 million of their investments into a fund that is now being investigated for the whereabouts of \$118 million in hedge fund investments.	On June 30, managers acting for Trio Capital $[_{\text{Transaction}}$ poured$]^B_a$ \$47 million of their investments into a fund that is now being $[_{\text{Justice}}$ investigated$]^B_b$ for the whereabouts of \$118 million in hedge fund investments.
2.	Administrators called in before Christmas to $[_{\text{Employment}}$ oversee$]^A_b$ Trio Capital say they are unable to determine what assets have been $[_{\text{Transaction}}$ bought$]^A_c$ with \$118 million invested through the Astarra Strategic Fund.	Administrators called in before Christmas to oversee Trio Capital say they are unable to determine what assets have been bought with \$118 million $[_{\text{Transaction}}$ invested$]^B_a$ through the Astarra Strategic Fund.
3.	Inquiries have focused on a company $[_{\text{Org. Lifecycle}}$ registered$]^A_c$ in the British Virgin Islands, EMA International, which has provided statements but no proof of investments in hedge funds.	Inquiries have focused on a company registered in the British Virgin Islands, EMA International, which has provided statements but no proof of investments in hedge funds.
4.	The annual report of Astarra Strategic Fund, one of 24 Trio Capital managed investment schemes now under administration, reveals that on June 30 its assets were $[_{\text{Finance}}$ topped$]^A_d$ up with a \$47 million transfer of assets.	The annual report of Astarra Strategic Fund, one of 24 Trio Capital managed investment schemes now under administration, reveals that on June 30 its assets were topped up with a \$47 million $[_{\text{Transaction}}$ transfer$]^B_a$ of assets.
5.	On December 22, Astarra Asset Management was $[_{\text{Org. Lifecycle}}$ placed into voluntary administration$]^A_e$ at the request of creditors.	On December 22, Astarra Asset Management was $[_{\text{Org. Lifecycle}}$ placed$]^B_c$ into voluntary administration at the request of creditors.

Figure 3.4: An extract from a news story with two predominantly differing annotations (Stuart Washington, *Mystery deepens over Trio's missing $118m*, Sydney Morning Herald, 2009-12-30). Annotators agree that all marked events are *realised*.

events within the typology, but each annotator failed to identify one. Other superficially similar instances – what assets have been [Transaction bought] (2.), the $118 million [Transaction invested] (2.) and a company [Org. Lifecycle registered] in ... (3.) – seem more dubious in terms of their reference to a specific event. It is unclear for instance whether registered refers to the act or the state of registration. We discuss similar examples below. There are apparent errors in the annotation of coreference, where A marks poured in 1. as not being coreferent with topped in 4.; B incorrectly coindexes invested in 2. with poured, although a partitive relation might hold. Finally, in sentences 4. and 5. we find disputes with respect to the marked span, which is not considered disagreement within our schema.

3.2.3 Recall in type-driven annotation

We focus on the frequent occurrence of recall errors – where one annotator marks an event that the other misses – and compare it to the ACE05 annotations. Although there are simple cases where an annotator's failure to mark an event is not easily explained (perhaps the investigation in sentence 1. of Figure 3.4 is one such instance), more often they seem attributable to atypical event-referential language or typologically/ontologically-borderline events.

Recall of atypical references Cases where annotators agree on an event at the document level but disagree with regard to sentences often correspond to sub-salient, if not oblique, references:

(20) Speaking at ANZ Stadium, where the Bulldogs winger will be hoping to score the seven points he needs to break the record in Saturday's $[_{\text{Sports match}}\ \text{game}]_a^{A,B}$ against Manly, Greenberg said El Masri deserved a share of the spotlight no matter what else was going on... "I'm trying not to think about it too much, but it would be nice if it all fell into place this $[_{\text{Sports match}}\ \text{week}]_a^{B}$, at our home stadium," El Masri said.

(21) When asked if he knew what had happened to Albert and Mario Frisoli, who were found $[_{\text{Lifecycle, Conflict}}\ \text{murdered}]_b^{A,B}$ in their Rozelle home last week, Mr Di Cianni's son Robert said he did not know... Yesterday police said their inquiries were focused on three or four lines of inquiry and the family appealed for help in finding the brothers' $[_{\text{Lifecycle}}\ \text{killer}]_b^{B}$.

In the former example, annotators agree on the reference to a sports match in the article. However, annotator B alone understands this week, at our home stadium as referring to the game mentioned in an earlier sentence. The typical characterisation of *event* as having a particular time and place of occurrence is applied to the extent that the *what* of the event is elided in context. The latter example similarly includes an antecedent marked by both annotators, but the use of an agentive nominalisation, killer, arguably *assumes* the death event rather than asserting it. Such indirectness may cause A to miss references.

Recall of borderline events A large portion of the disagreement results from the difficulty of determining whether a fact is to be considered a specific event or of a particular type. Annotators are particularly prone to disagree when the referent in question diverges from the prototypical "event" in not having a clear time or place of occurrence. For example, compare annotation of the following contiguous sentences:

(22) Shares in the retailer [Finance lost]^{A,B} $1.86, or 7 per cent, to close at $26.14.

(23) The company has been able to [Finance deliver]^{A} double-digit profit growth every year since 1999 and its share price is based on this.

(24) Mr Luscombe assured investors he could continue to [Finance deliver]^{A} double-digit growth over the medium term in all businesses.

The latter two sentences were not annotated by B as referring to *Finance* events, perhaps because they are not clearly specific (or not events). Similarly, while B sees sanctions as anchoring both *Conflict* and *Governance* events in the following, A probably disregards the sanctions as an event:

(25) In return for a tougher array of United Nations [Conflict, Governance sanctions]^{B} against Iran targeting the country's vast oil and gas reserves, ...

Referents such as these make event extraction difficult: there is a long and heavy tail – in comparison to named entity recognition, for instance[20] – of references and referents that are not prototypical events, or not prototypical to their type. When performing type-driven annotation, it is therefore easy for annotators to drift in leniency towards atypical events and atypical type instances, resulting in disagreement.

Comparison to ACE05 The caveat of sample size notwithstanding, we may compare these results to annotator agreement on the English newswire portions of ACE05. We have previously reviewed inter-annotator agreement in ACE05 in relation to subtype homogeneity and predictability (Section 2.5.1), and to identify features of low-salience references (Section 2.4). Here we more directly consider agreement and adjudication of coarse event type identification at the document and sentence levels. The ACE05 corpus is annotated by two independent, "first-pass" annotators, $fp1$ and $fp2$, and these are adjudicated by adj. We hence report the chance-corrected agreement (Cohen's κ) between the independent annotations, and F measure between all pairs in Table 3.4.[21]

[20] Even in that task, a long but lighter tail of named entities is accounted for in the CONLL task through a MISC label (Tjong Kim Sang, 2002). Due to a lack of syntactic and orthographical cues, a similar type for events would be difficult to scope.

[21] Chance-corrected metrics are not applicable to adjudication where the annotations are not independent.

Type	Documents				Sentences			
	κ	F_1			κ	F_1		
	$\frac{fp1}{fp2}$	$\frac{fp1}{fp2}$	$\frac{fp1}{adj}$	$\frac{fp2}{adj}$	$\frac{fp1}{fp2}$	$\frac{fp1}{fp2}$	$\frac{fp1}{adj}$	$\frac{fp2}{adj}$
Business	0.6	0.7	0.8	0.8	0.4	0.5	0.6	0.7
Conflict	0.7	0.9	1.0	1.0	0.7	0.7	0.8	0.9
Contact	0.7	0.8	0.9	0.9	0.6	0.6	0.7	0.7
Justice	0.8	0.9	0.9	1.0	0.8	0.8	0.9	0.9
Life	0.8	0.9	0.9	0.9	0.6	0.7	0.8	0.8
Movement	0.6	0.8	0.9	0.9	0.5	0.6	0.7	0.8
Personnel	0.8	0.8	1.0	0.9	0.5	0.5	0.7	0.8
Transaction	0.5	0.6	0.9	0.9	0.4	0.5	0.7	0.6

Table 3.4: Inter-annotator event type agreement in newswire portions of the ACE05 corpus. We report Cohen's κ between the two first-pass annotators ($fp1$ and $fp2$) and F measure (F_1) between all pairs including the adjudicator (adj).

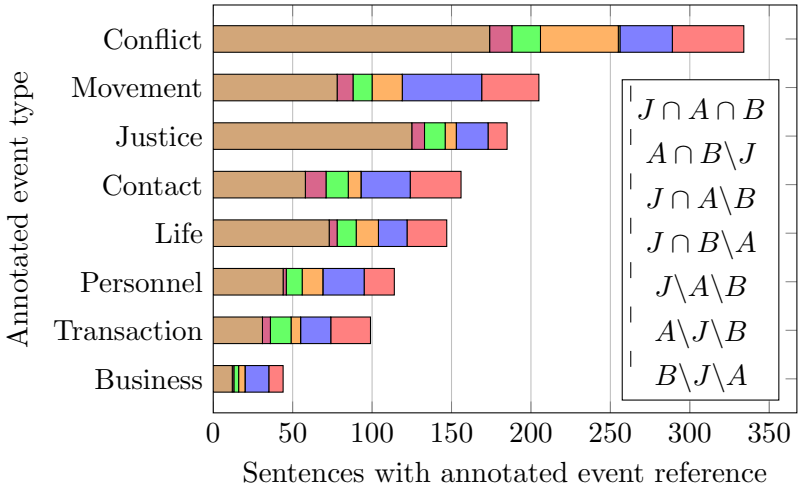


Figure 3.5: Sentence-level event-type annotation contingency for the most frequent types annotated in the newswire portion of the ACE05 evaluation corpus. A, B and J are the sets of sentences labelled by each annotator. A and B correspond to first-pass annotators $fp1$ and $fp2$ respectively. J corresponds to adjudication (adj).

Document-level agreement between first-pass annotators is quite high, the lowest-agreement being *Transaction* which is infrequent in the corpus.[22] In contrast to our annotation, the high agreement in ACE05 reflects the annotators being more thoroughly trained to a detailed schema, and only marking events that fall under sought subtypes and with marked participants. The news text being annotated is also much more topically homogeneous, so annotators are able to focus on the types of event relevant to the topic, as evidenced by the outlying high frequency of *Conflict* events.

At the document level, the adjudicator almost entirely agrees with the first-pass annotators. However, agreement drops substantially when considered at the sentence level, for independent annotation and for adjudication. This suggests the difficulty of identifying all references of an event type (or their correct classification), and indeed Figure 3.5 shows that for the *Conflict* type, the adjudicator adopts many annotations from $fp2$ not marked by $fp1$, and a few in the opposite direction. This recall problem we have identified in our own annotation was further discussed with respect to ACE05 in Section 2.4.

More surprisingly, in all types, more sentences are rejected in adjudication than are adopted from either single first-pass annotator in a case of disagreement: both first-pass annotators substantially overgenerate annotations with respect to the schema. This is suggestive of the annotation drift noted in our task.

We proposed the use of a pre-specified typology as a way to constrain the referent space to make the event detection task feasible, reliable and useful. Despite this, in our work and in ACE05, it seems that the complexity and variability among events and their references results in both undergeneration – we suggest due to atypical, low-salience references – and overgeneration – due to borderline, perhaps unsought referents – of manual annotations.

3.2.4 Annotation analysis

A summary of the full set of annotations is shown in Table 3.5. The most frequent event types occur in 32% of all documents annotated, and are notably among the least frequent types in the ACE05 training corpus (see Figure 3.5). Many of the frequent *Transaction* events seem salient within their story, while *Employment* events may appear within peripheral descriptions of salient entities and events as in the following examples:

(26) a. The drone attacks on Pakistani territory have continued since Barack Obama [Employment became] president in January.

b. That frontrunners are suited by the Caulfield conditions isn't new but it will be different on Saturday, says the studious Craig Williams, [Employment deposed] as stable jockey this week [Employment for] David Hayes and also serving a suspension.

[22] It also represents a class of events that are referred to in very diverse ways, being those that "refer to the buying, selling, loaning, borrowing, giving, or receiving of artifacts or organizations ..., [or] money when it is not in the context of purchasing something." (LDC, 2005)

Event type	Doc freq	(%)	Evts / doc	% subevent	% realised	Sents / evt
ANY	176.5	(90)	6.50	10	62	1.6
Transaction	62.5	(32)	2.46	5	49	1.4
Employment / Award	62.0	(31)	2.30	6	57	1.4
Conflict	54.0	(27)	2.36	13	73	1.7
Justice	45.0	(23)	2.38	4	75	1.5
Sports Match	43.5	(22)	4.93	22	56	1.6
Governance	39.0	(20)	1.77	1	43	2.0
Lifecycle	33.0	(17)	1.98	4	92	1.7
Finance	31.5	(16)	3.25	9	79	1.3
Organisation Lifecycle	29.5	(15)	1.63	1	41	1.8
New Release	27.5	(14)	1.04	0	68	1.3
Correspondence	25.5	(13)	1.80	11	63	1.7
Disaster	14.0	(7)	1.54	0	77	2.2
Real Estate	6.0	(3)	1.33	0	38	1.3

Table 3.5: Statistics of events annotated in our corpus by type, including: *document frequency*, the number of documents annotated with a given type; *events per doc*, the average number of distinct events per document where that type is marked, and under *not part*, the quantity that are not marked as part of another event; *% realised*, the proportion of events marked has happened or happening; and *sentences per event*, the average number of sentences with coreferent annotations. For double-annotated portions, counts are averaged.

Employment events are also prominent in sports and business news, which frequently focus on changes in personnel and personal awards. Our data also demonstrates ACE05's bias towards *Conflict* and *Justice*: while frequent in broad coverage news, there is an apparent skew towards these types in the ACE05 corpus sampling relative to a random sample of the SMH. Our least frequent event types, appearing in as few as 3 and 8 documents are *Real Estate / Development* and *Disaster*; neither is considered by ACE05.

Counting the number of distinct events per document rather than types, we find there are more *Sports Match* events across our corpus than any other type. We find the number of distinct events per document is very type dependent: if a *New Release* or *Real Estate / Development* event is mentioned in a document, it is almost always the only instance. Many (22%) of the *Sports Match* events are sub-events of other *Sports Match* events, accounting for their high number. In contrast, documents with *Finance* events such as price movements tend to mention many such events together that are not frequently annotated as sub-events of others (9%). Since the *part of* annotation only applies when the super-event is also marked, sibling events are under-reported. This suggests that, at least for some types, structure and interrelation is frequent among news events, which neither strict coreference nor perhaps a simple *part of* label will suffice to capture.

We also note the variation in the proportion of events marked as having been realised. The death and injury events that constitute much of the *Lifecycle* type are rarely anticipated, in contrast with *Organisation Lifecycle* events.

The highly type-dependent nature of event reference distribution, also identified in ACE05 (see Section 2.5.1), complicates building generic event detection models.

3.2.5 Discussion

While an attempt to codify types of event that prominently feature in news, our schema leaves some room for improvement. Our broad event types are often almost thematic, and could do with further specification; similarly, greater clarity in the definition of *specific event* targeted by the annotation might improve inter-annotator consistency. Nonetheless, our annotation experience, together with our analysis of inter-annotator agreement and type variability in our corpus and in ACE05 suggest wider concerns regarding broad-coverage, type-driven event detection tasks. We consider two primary issues: the distillation of the event space into a small, fixed typology, and a typology's application to text.

Typology construction An ideal event typology would capture all interesting event types and discriminate between those that are perceived as different, while remaining compact for application.

The gaps left by our annotation show that broad coverage is limited given the long tail of newsworthy events. Many of the 10% of news articles without annotations from our typology

are investigative analyses that do not refer to specific events, or reports of economic trends that our typology does not consider. However, we also find our typology fails to cover key news events such as a rescue mission, a reconciliation, and non-transaction business interactions that are mentioned in our corpus.

Assuming we ignore some infrequent portion of newsworthy events, the requirements of coverage and compactness ultimately result in types that are semantically specific, but which apply so broadly as to actually be perceived as different events. The pilot ACE03 event typology (LDC, 2003) exemplifies this in grouping the creation of a business, an artifact or a human life into one event type. The ACE05 typology reduces this problem, but it remains, particularly in types like *Transaction:Transfer-Ownership*: an international purchase of arms, capturing a building or vehicle, theft of a single weapon and a corporate takeover are fundamentally very similar at the level of abstraction that the event type describes, but are perceived as vastly different events. We have similar fundamental problems in our *Transaction* and *New Release* types, at a minimum.

Other types may better reflect how events are perceived, but do so by being broadly thematic groupings of events, such as our *Finance* and *Governance* types, or by being evaluative in that they depend on a particular perspective on the event's purpose or outcome, such as *Conflict* as it appears both here and in ACE05, or our *Disaster*. We would like to suggest that the human perception of newsworthy events and their categorisation is highly determined by function, although this is not necessarily a property of the event itself. Often its function or domain, rather than its type, makes an event notable: speech is uninteresting, but argument or allegation is. Yet the annotation of *Disaster* events including an organisational mishap, a financial recession and a bushfire still suggests the category requires further subdivision in terms of magnitude and domain in order to match perception.

We are thus presented with competing aspects of event ontology when designing a broad-coverage typology. In theory, this could be resolved by allowing events to fall into many categories, but this reduces typology compactness and creates an explosion in the number of annotation decisions and hence the difficulty of annotation. We allow multiple annotations per token where ACE05 does not, and find instances with multiple types that are emblematic of this problem, including [Conflict, Lifecycle assassination], [Disaster, Finance the economic downturn], [Correspondence, Conflict debate], [Conflict, Justice take action] and [Employment, Justice banned], but mark aspects of the denotational semantics due to language choice just as often as aspects of the referent; we think the 0.5% of mentions (and fewer distinct events) assigned multiple types are an under-representation of this problem, limited by the difficulty of considering all types for all events and event references. Where multiple types can be assigned, the notion of broad coverage also becomes harder to define. One solution is to construct multiple orthogonal partitions of the event space corresponding to action, domain, function, etc., such that each event acquires a type label from each partition. In our next annotation attempt, we pilot

this concept together with a dynamic approach to typology to accommodate fine-grained distinctions.

Granular annotation Identifying typed events within a document is inhibited by the lack of well-defined annotation units together with the varying salience of event references and typological distribution. For instance, if we compare this task with the identification and classification of proper noun-named entities into *Person*, *Organisation* and *Location* categories, the unit of annotation is still not as well defined as in document classification or part-of-speech tagging, but the vast majority of entities can be identified readily – with bearable overgeneration – from their capitalisation; even extending such an annotation to non-proper noun references is largely feasible with noun phrase identification tools. The small number of entity types is easy to recall and evaluate for each candidate, aided by the fact that they have similar frequency in text. In contrast, events are referred to with great linguistic variability, employing many grammatical structures, figurative language and idiom, making annotation units hard to identify; since annotators are faced with both understanding an article and identifying its event references, those that are focal to the discourse are much easier to identify than the same events mentioned peripherally; and the typology's size and skewed distribution leaves infrequent types easily forgotten during annotation.

Our annotation sheds light on problems in applying a flat, broad-coverage typology to event annotation, most notably the difficulty of fitting events to a typology, and of exhaustively identifying matching event references in plain text. To alleviate these particular problems, we propose a novel approach to news event annotation that applies a dynamic notion of typology to characterise the events that make a story newsworthy.

3.3 Story-driven annotation experiment

We present another approach to selecting news by event attributes, but avoid problems due to a prescribed typology and searching for all events mentioned in a document. We intend to focus exclusively on the most salient events in a news story, as we think this is the content that most readers take away; and we adopt a more inclusive and expressive approach to event typology that may better match the way similar events are grouped in the reader's perception. Our second event annotation approach is thus novel in two ways:

Firstly, rather than constraining our annotation to a typology and marking every reference, we exploit (and make assumptions about) news discourse structure to consider only the central event of a news story, without concern for detailed background and peripheral event references. The central event of a story may be the event that was impetus for it to be written; however, we find many stories report new information primarily because of their relation to some broader topical event. We hence describe articles as reporting one or more

Somali Gunmen Release Ship Carrying Tsunami Aid		
Attribute	Update event	Topic event
Type	Interaction:Resolution:Release	Interaction:Conflict:Physical:Capture
Domain	Crime	Crime
Tense	Past	Past
Modality	Asserted	Asserted
Polarity	Happening	Happening

Figure 3.6: A possible story-driven event annotation for the hijacking example. In this instance, the event type has been added to our otherwise deficient typology. An alternative understanding would place the tsunami as the topic event.

update events in the context of a *topic event*.

Secondly, we again attempt to apply the notion of event type to abstract over the space of newsworthy events and to group similar incidents. However, instead of pre-specifying events of interest, we presume all topic and update events are notable and must be assigned a type. We therefore allow annotators to extend a type hierarchy as necessary to accommodate previously-unseen event types, allowing us to delay discrete event type groupings until after a substantial corpus is annotated.

These changes may also enable faster annotation, since an annotator need not read an entire story to identify its key events, nor consternate over the correct event type when one can be created. Our pilot annotation investigates whether the central event content of a news story can indeed be captured in the notions of topic and update, and highlights the particular difficulty of identifying a single topic event, while a number of stories in our corpus do not have a clear update event. The annotation also explores the extent to which these events may be characterised through a dynamic type hierarchy, finding that the notion of event hierarchy may have many interpretations, so that while a hierarchy may capture event type relations, another structure may better represent event type.

3.3.1 Task definition

We adapt the previous type-driven task to annotate only the central events of each news story and remove the constraint of a fixed typology. We illustrate the task with an annotation of our running example in Figure 3.6. It differs from the type-driven annotation in the following aspects:

Reference granularity The many events that may be referred to in a news article are not of equal importance: some provide additional detail or attribution to a focal event. Often the focal event, such as the (announced) release of the aid ship in the hijacking example, is

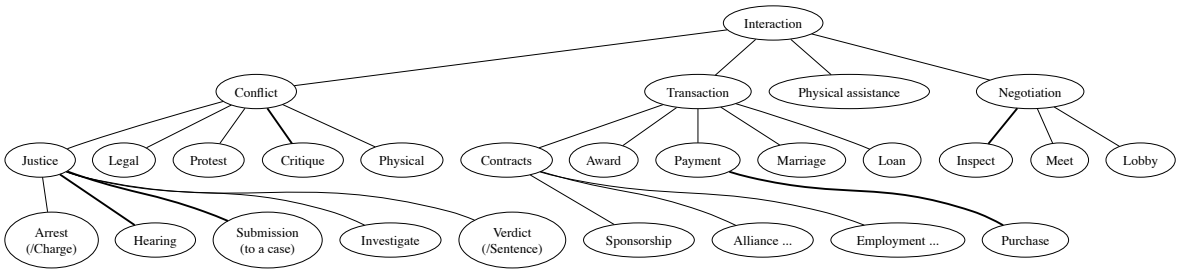


Figure 3.7: An excerpt from a dynamic type hierarchy used in the story-driven annotation. Arcs in bold indicate additions during the annotation procedure. Examples of non-*Interaction* event types are: *Death*, *Invention*, *Disaster* and *Measured change*. Children of the *Employment* and *Alliance* nodes are not shown. See the full hierarchy in Section A.2.

the impetus for the particular story being written. This annotation task therefore aims only to describe a news story in terms of its focal event. Each story should be annotated with its impetus, called an *update event*, as well as a broader *topic event*, which may correspond to a script including the update event.[23] Thus, release of a hijacked ship is the primary event of our ongoing example story, suggesting an update event corresponding to the release and a topic event corresponding to the hijacking. Alternatively, the broader event that ultimately results in the hijacking and release is a tsunami, and this might also be considered a topic event. Similarly, this model is able to describe update-topic relationships such as: a legal challenge of a proposed policy; or the opening night of a musical theatre run.

In our annotation, the events are not anchored in a fine-grained unit of text, allowing the annotators to interpret the impetus even where it is not explicitly referenced. Thus where an article describes a changed state as in Kevin Rudd is again the Prime Minister of Australia, the event bringing about that state, such as an election, may be inferred without pinpointing a reference in text.

For this pilot annotation, multiple values in each slot are permitted as the annotator deems necessary.

Typology We apply a dynamic typology: annotators extend the existing type hierarchy as they feel is appropriate. This assigns each event a fine-grained type, while accounting for the similarities between event types through hierarchy, delaying the determination of a more pragmatic, coarse set of types, or leading to a similarly adaptive automatic prediction framework. At the same time, this ensures full coverage of relevant events; unlike the type-driven annotation, the typology does not constrain the referent event. We have successfully applied a similar dynamic typology to the entity type classification of Wikipedia articles (Nothman et al., 2013), enabling the training of named entity recognisers with application-specific type

[23] We adopt this terminology by analogy with *update summarisation* (Dang and Owczarzak, 2008) which seeks to summarise the new information in one corpus in relation to a background corpus, and *topic detection* (Allan, 2002) which effectively groups stories with the same topic event.

schemas. The use of a hierarchy clearly allows for more expressive types: we are able to adopt a very broad definition of *Transaction* to incorporate many specific subtypes that correspond more closely to our perception of distinct event types. The typology constructed during our annotation (prior to any potential refinement) is found in Figure A.2; an excerpt is shown in Figure 3.7.

The most frequent modification applied to the hierarchy is subtype insertion. For example, the final typology contains no definitive representation for hijacking, the closest type being *Interaction:Conflict:Physical*. In annotating the hijacking-release story that featured in the previous chapter, we assume that it would be useful to distinguish such capture of property from other physical conflict. Hence we create a subtype *Interaction:Conflict:Physical*; similarly, for the release of a hijacked boat or another captured artifact, we create the new type *Interaction:Resolution:Release*, adding the *Resolution* type as well to potentially include *Rescue* or similar subtypes. In a similar manner, the application of a type to a story may result in an intermediate node in the hierarchy being introduced; thus we introduced *Measured change* as a parent of *Trend* and price or index movements. Other types may be moved within the hierarchy or deleted to be subsumed by another type; thus we deleted *Inauguration* which was a subtype of *Interaction:Transaction:Contracts:Employment*, and introduced a *Ceremony* node to contain its events.

With reference to our discussion on event typology definitions in the previous section, the main event type is intended to be abstracted from the thematic domain of the event, which we mark using a separate dynamic hierarchy, shown in Appendix Figure A.1. In this way a death and a business closure share the same event type (*Lifecycle:End*), but a different event domain (*Business* vs. *Human interest*); similarly, the commonality between rule enforcement on the cricket pitch and criminal justice (even with the same offender) is identified through a shared subtype of the *Justice* event type, with different domains.

Initial hierarchies were constructed from an examination of prior schemas, including annotations and comments from the previous task. These and the modifications applied to them during annotation are shown in Appendix Section A.2.

Factuality As well as its type and domain, each event is labelled with its tense (of *past* (default), *ongoing*, *future*)[24], modality (of *certain* (default), *probable*, *possible*) and polarity (of *happening* (default), *cancelling*). Where the previous task simplified this annotation to a single attribute, we expect they are more decidable and informative within the present task's focus on few, salient events.

[24]We do not use *tense* in its grammatical sense. Tense is here conflated with relative time of occurrence and aspect; a feature based on aspect – selecting from among *completed*, *ongoing* and *anticipated* – may have produced a clearer annotation.

Coreference No explicit relations between events or event references are annotated, although all references to an event within the news story may be used to determine the annotations. The topic detection and tracking task (Allan, 2002) may be construed as determining coreference between topic events of multiple stories. Further, by characterising both a topic and update event where available, partial scripts of related event types are implicitly constructed.

Annotation interface For this pilot, we use existing Google Spreadsheets as an annotation tool, providing us with the required features without substantial development effort. Within one spreadsheet, each annotator works within a single worksheet, with additional worksheets to store the current typologies and for inter-annotator review. The software handles concurrent updates to the typologies to ensure redundant nodes are not added by multiple users, while also storing the revision history of typologies and annotations.[25] Each annotator worksheet consists of rows corresponding to news stories, with columns for: update and topic event annotations; the first two sentences of each story, which are usually sufficient for annotation; and story details such as database identifier and headline which connect the spreadsheet to the full article text. Only non-default tense, modality and polarity values are annotated. Although probably inappropriate for a larger-scale annotation, this simple interface encourages quick annotator decisions.

After developing the task with a committee including the annotators, one annotator labelled 60 news stories, almost all from the Sydney Morning Herald of 2nd April, 2009, with a few from the following day's edition. Another annotator produced second annotations on a quarter of these stories, which provides insight into the task's uncertainties, but is insufficient for quantitative agreement evaluation. We analyse the annotation in terms of the two novel components of our approach, asking: can the event content of a news story be distilled into an update and a topic event? can the character of the event be succinctly summarised through event type and domain hierarchies? We stress that this is a largely unconstrained and unadjudicated pilot annotation over a very small sample of documents; its analysis may lead to only preliminary conclusions, while suggesting worthwhile directions for further investigation.

3.3.2 News stories as two events

By distilling stories into these key events, our annotation highlights patterns in the relationship between update and topic in news: the topic is most often a past event and the update a follow-up,[26] as in our hijacking example; the update may instead anticipate a follow-up

[25] By immediately sharing modifications to the typology, annotations from multiple annotators are not entirely independent; while this is a realistic setting for annotation at greater scale, it presents as a caveat for inter-annotator evaluation.

[26] About 7 in 10 articles annotated with a single update and topic have both events in the past, with most other topics ongoing and most other updates in the future, but this may also be an artifact of the ambiguities

such as legislation in response to some past event; it may be part of an ongoing topic, as with a hearing in a trial; or prepare for a future topic event, as with a qualifying game for a competition.[27] However, we identify a single focal topic and update event for about half of all stories annotated; often the selection of a single topic event is ambiguous. At the same time, the impetus for a story is not always apparent, such as in human interest angles on current affairs, and the update content need not constitute an event.

3.3.2.1 Identifying topic events

A single topic event is often difficult to identify because the story's topic consists of multiple events, or none in particular. In other cases, a story may introduce its impetus into discourse without relating it to an existing topic, or providing reference to related stories that nonetheless do not seem to constitute the topic of the story.

Unexpected events such as sudden crimes, accidents or disasters are generally reported as standalone events, with some references to or statistics of similar occurrences in the past. The following stories describe their impetus as independent of other events in the contemporary news discourse:

(27) BEACHES were pounded by five-metre waves, city streets were flooded and blackouts hit inner-city suburbs yesterday as the Bureau of Meteorology warned more wet weather was on the way.
Update event: type *Disaster:Natural*; domain *Disaster*; tense *Ongoing*.

(28) ADVERTISERS will no longer be able to use images of nature and call themselves "environmentally friendly" unless they can back up any green claims under new proposals put forward by the advertising industry. The new self-regulatory green marketing code – thought to be the first of its kind in the world – will also prevent companies from passing off a mandated environmental initiative as something it has voluntarily adopted.
Update event: type *Regulation*; domain *Business*; tense *Future*; aspect *Probable*.

Although these could be connected to very broad topics such as *weather* and *product labelling*, such topics have no seminal event. Instead, we construe the flood in (27) and the regulation in (28) as initiating new topics. Analogous to the concept of the *first story* of a topic in TDT (Allan, 2002), such cases should be specifically labelled during annotation.

In other cases, a journalist may connect some new event to a topic when the two are only tenuously related, to frame the update event's interpretation. This contextualisation is often applied to update events that report the publication of new data, which are generally

involved in annotating future update events, discussed below.

[27] Annotating the relationship between the update and topic events may also be an interesting task, but one that we have only performed ad hoc for the purpose of analysis; it depends on the integrity of update and topic event annotation.

uninteresting without some interpretation and context. Thus in the following example, the reported metric is provided in the context of economic stimulus measures, which may or may not have been a substantial contributor to the measured outcome.

(29) ANYONE wondering what will happen to retail sales when the Government stimulus handouts end has only to look at yesterday's official reading on the sector. The retail sales figures showed department sales fell a whopping 9.8 per cent, compared to the increase of 8.3 per cent in December, when the payments were made.
Topic event: type *Regulation*; domain *Gov't:Federal*; tense *Past*.
Update event: type *Measured change*; domain *Business*; tense *Past*.

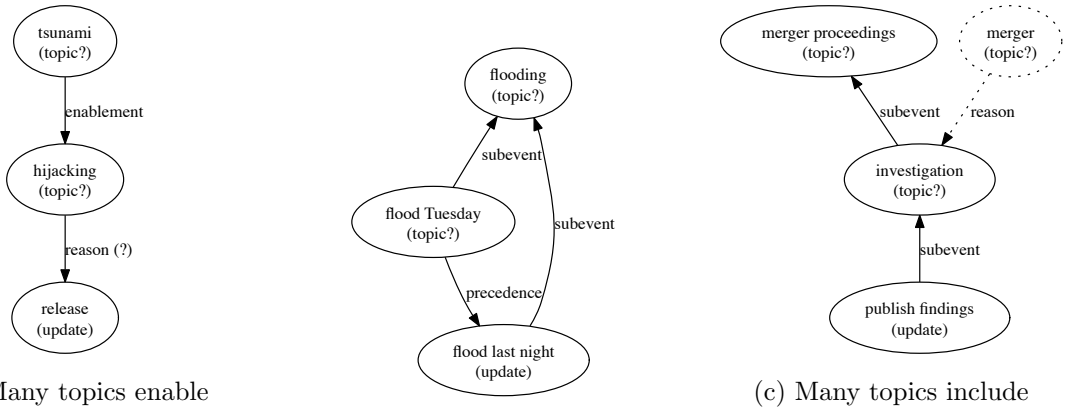
In such cases, the topic event is rhetorically rather than ontologically related to the update event.

There are frequently multiple candidates for the topic event, since topics tend to have nested structure as suggested by TDT's change from flat to hierarchical topic detection (Feng and Allan, 2005; Allan et al., 2005). Thus for the release of a hijacked aid boat, one might choose between the hijacking or the tsunami as the topic (illustrated in Figure 3.8a), and we find similar ambiguity in the following examples:

(30) LARGE areas of the Mid-North Coast – already declared a natural disaster zone after thousands of people were trapped by floods – were battening down last night for a king tide that could cause swollen rivers and creeks to devastate even more homes. The once-in-a-hundred-year storm that hit on Tuesday has already flooded 198 houses in Coffs Harbour alone.
Topic event: type *Disaster:Natural*; domain *Past*.
Alternative topic event: type *Disaster:Natural*; domain *Disaster*; tense *Ongoing*.
Update event: type *Disaster:Natural*; domain *Disaster*; tense *Future*; aspect *Possible*.

(31) THE consumer watchdog is worried the proposed merger between Australia's third and fourth largest mobile phone companies could force customers to pay higher prices. The Australian Competition and Consumer Commission has handed down preliminary findings as it ponders whether to allow the marriage between Vodafone Australia and Hutchison Telecommunications.
Topic event: type *Merger*; domain *Business*; tense *Future*; aspect *Possible*.
Alternative topic event: type *Investigation*; domain *Business*; tense *Ongoing*.
Update event: type *Publication*; domain *Business*; tense *Past*.

In (30), the storm ... on Tuesday is explicitly referred to as the background event. Yet it is unclear whether this should constitute the topic, or rather, that the upcoming flood is part of a collection of floods that together constitute a topic event: as illustrated in Figure 3.8b, is the update *part* of the topic, or is it in *sequence* with the topic? In the final example, the



(a) Many topics enable update

(b) Topic includes or precedes update

(c) Many topics include update

Figure 3.8: Ambiguous choices of topic event in relation to an update event, labelled according to the schema of Bejan and Harabagiu (2008).

update is part of an investigation, which in turn is part of proceedings related to a merger (see Figure 3.8c); in addition, the tense of the *merger* topic event differs if we read the term widely as including the proceedings, or narrowly as the conclusion of the merging process.

In other cases, the topic may not naturally be seen as a single event, but as a series of related events. This is our interpretation of Example (32), where a series of violent incidents amount to a topic of note.

(32) THE Federal Government will support states in cracking down on bikie gangs, suggesting it could make it easier for police to tap phones. The Attorney-General, Robert McClelland, said Australia was facing an "immediate crisis" following the outbreak of violence between the gangs but the Federal Government had no jurisdiction to police them.
Topic event: type *?Conflict:Physical*; domain *Crime*; tense *?Past*.
Update event: type *Regulation*; domain *Gov't:Federal*; tense *Future*; aspect *Possible*.

We introduced the notion of topic event to help distinguish two groups of event content of a news story. These annotations illustrate that topics represent sometimes-structured collections of events, such that a single seminal topic event can be difficult to select for a given story. Allan et al. (2005) reflect on unrealistic assumptions in most Topic Detection and Tracking evaluations, wherein topics were: unstructured; not overlapping; and assumed to focus on a seminal event. With the loosening of these constraints for the 2004 evaluation, the topical structure is more expressive. However, as suggested in example (29), we also see that the journalist may provide a topic to contextualise new information, even when not intrinsic background to that new information. Hence an alternative approach might not employ TDT's implicit notion of topic, instead focusing on precisely the events that a journalist refers to. Under this model, developed in the remainder of this thesis, the notion of topic may emerge from the set of *explicit* event references that are background to an update event.

3.3.2.2 Identifying update events

As with topic, the assumption that information being reported reflects an event is simplistic. Generally the update involves newly-public facts; when those facts state that a single event that recently took place, the identification of an update event is straightforward, as in the following example:

(33) TWO partners of PricewaterhouseCoopers no longer face the prospect of a judicial inquiry into their conduct as liquidators of Reynolds Wines following a NSW Court of Appeal decision in their favour.
Topic event: type *Conflict:Legal*; domain *Business*; tense *Past*.
Update event: type *Verdict*; domain *Business*; tense *Past*.

The update event is less clear when the new knowledge regards a future event, as with the advertising regulation in (28), a fact released in a report as in (29), or a long-past event as in (34).

(34) THE Chinese-born businesswoman Helen Liu paid for Joel Fitzgibbon to travel first-class to China in 1993 before they had ever met, giving him access to top Communist Party officials.
Topic event: type *Scandal*; domain *Politics:Federal*; tense *Ongoing*.
Update event: type *?Payment*; domain *?*; tense *Past*.

In (29), the fact happens to be reported as if it is an event: fell is used to describe the downward change in an economic metric. Although the event that is actually impetus for writing such articles is an announcement, a publication, a decision or private revelation, this is not always cited (and that event may not be newsworthy in itself); besides, the same can be said of most news, since a journalist rarely directly witnesses the event that she reports. Therefore, identifying the central fact being reported rather than an event might constitute a more straightforward task.

Even with this approach, in the following story it is not clear what newsworthy knowledge is introduced:

(35) JASMIN COMMOR needed little persuasion to realise she and her infant brothers were in imminent danger. Her mother, Melissa Commor, had rung from the Urunga supermarket at 5.45pm to tell her she could not get home because rising flood waters had cut streets.
Topic event: type *Disaster:Natural*; domain *Disaster*.
Update event: type *?*.

Such stories provide a personal perspective on the topic event rather than a news update; it is more likely to instill an image or emotion in the reader rather than a fact.[28] It may be

[28] In this particular case, we are somewhat compromised by only considering the opening sentences of the story, since it doesn't adhere to the inverted pyramid structure. The complete article provides a variety of updates about the after-flood response, together with anecdote.

adequate to not mark an update event in such cases, but – though we have not identified such cases in our annotation – there are likely to be other stories that blur this boundary.

In addition, some stories appear to have a number of equally focal events (or facts) that often occur simultaneously. When both an anticipated event and its declaration are reported focally, selecting between them may be difficult. This is perhaps the case in the following, in which cut and lashed out are salient updates, but so is agreed, given that the article's title is QBE bends to pressure and caps chief's pay:

(36) THE insurance company QBE has agreed to cut the retirement payout to its chief executive, Frank O'Halloran, but has lashed out at a "jaundiced" report by shareholder advisory firm RiskMetrics that was critical of its remuneration plans.
Topic event: type *Critique*; domain *Business*; tense *Past*.
Update event: type *Payment*; domain *Business*; tense *Future*; aspect *Certain*; polarity *Cancelling*.

Similarly, the news in the following example consists of a number of related events:

(37) ONE of Pakistan's most powerful politicians, Shahbaz Sharif, has been reinstated as chief minister of the influential Punjab province, easing a political controversy that has destabilised the country. There were celebrations across Punjab following the decision by Pakistan's Supreme Court to suspend a February court order that banned Shahbaz Sharif, along with his brother and former Prime Minister Nawaz Sharif, from holding elected office because of prior convictions. The suspension of the ban means Shahbaz Sharif can resume office immediately while the court review continues.
Topic event: type *Conflict:?*; domain *Politics:World*; tense *Past*.
Update 1 event: type *Employment*; domain *Politics:World*; tense *Past*.
Update 2 event: type *Verdict*; domain *Politics:World*; tense *Past*.

Underlying this annotation task is the assumption that a news story generally reports or results from a single central event that can be identified by readers. Although the new content is often clearly marked, our annotation suggests that this is frequently false as when multiple events or some fact other than an event's occurrence is focal.

3.3.3 Dynamic hierarchies for event typing

The novel aspects of our type system with respect to our previous experiment are threefold: types may be added or modified during annotation; types form a hierarchy, identifying interesting shared aspects of events; and event type is abstracted from event domain.

3.3.3.1 Working with a dynamic typology

Our annotation scheme's typological dynamism permits the assignment of every event to a type, with the risk that annotators produce types that are very fine-grained, poorly speci-

fied or redundant; problems which add work to the annotation process, and which grow in proportion to the number of concurrent annotators. Granularity may be reduced through a hierarchy of types, since infrequent branches may be pruned; and redundancy can be avoided through clear specification and regular checks for semantic integrity of the typology. Under-specification is a particular problem for this pilot annotation as our typology does not provide descriptions and examples of each type, in contrast to our previous work with a dynamic typology (Nothman et al., 2013). In contrast, a small set of static types can be documented and specified with relatively little effort. Types within a hierarchy can be easily merged, or leaf-nodes moved, but the process of modifying the type system amid annotation may also involve substantial additional work.

For an example of the complexity involved in altering the typology, consider the creation of the *Ceremony* node, during our annotation, to accommodate the following example:

(38) THE strains of a lone didgeridoo have welcomed an estimated 2500 people, including Prince Charles and Camilla, Duchess of Cornwall, to London's Westminster Abbey to remember those who died in the Black Saturday Victorian bushfires.
Topic event: type *Disaster:Natural*; domain *Disaster*; tense *Past*.
Update event: type *Ceremony*; domain *Disaster*; tense *Past*.

The new type was deemed a child of the *Structured event* type that held *Contest*, *Show* and *Election*. It is conceivable that some ceremonies and contests are also shows, and this distinction might be clear were *Show*'s bounds prescribed in more detail. The annotator may have also appreciated that *Ceremony*, *Contest* and *Show* share more than just being structures, and could perhaps have inserted a node *Event with an audience*, but in our case did not. The annotator noted, however, that *Inauguration* was incorrectly included under its thematic parent, *Employment* – descendant from *Transaction* and *Structured event* – when it is merely a functionally-specific ceremony. They could have chosen to move it to be a subtype of *Ceremony*, but without any existing *Inauguration* annotations, they instead deleted it. In this way, modifying the typology may involve reconsidering large portions of it as part of the annotation process.

Our application of the typology allows events to be assigned to internal or leaf nodes. Yet when a new type is added, as *Critique* was to *Conflict*, any existing annotations assigned to the parent must be flagged for review to move them to the created subtype where necessary. We did not perform such bookkeeping – our sample size and typology are small enough to review all annotations – but this is an important exception to the notion that inconsistencies may be resolved within the typology alone, without reference to the annotated samples. Requiring that all samples be assigned to leaf nodes minimises this problem, but requires that a node's contents be completely partitioned into its children when it acquires its first child node, with further caveats for concurrent annotation. Many additions to the typology

are nonetheless unaffected by this issue, such as adding internal nodes together with their children, or adding new subtypes to types where no events are assigned. In this way *Measured change* was inserted as a new parent to *Trend* together with its children *Gain* and *Loss*;[29] and *Submission* was added to *Justice* which only has annotations assigned via its descendants. Hence while the addition of new types can require inefficient review of previous annotations, this situation can often be avoided.

Our rapid construction of the typology prior to and during annotation results in a number of types that could be better named or positioned within the hierarchy. We nonetheless hold that as long as the types are clearly and finely specified, inconsistencies may largely be resolved by intermittently moving nodes within the typology.

3.3.3.2 Expressiveness and integrity of the type hierarchy

Hierarchical typing is necessary to avoid sparseness given the dynamic creation of fine-grained types, but it is not automatically meaningful or useful; nor is the single-parent approach we adopt fully expressive of all type relations.

Our typology shown in Figure A.2 incorporates some incorrect taxonomy. We initially founded our typology on a distinction between structured events – events thought of or often described as a sequence of sub-events – and unstructured events, but this distinction was not well adhered to. For example, *Lifecycle:Bankruptcy/Liquidation*, *Justice:Hearing*, *Conflict:Physical* and *Disaster:Natural* appear as unstructured events and in at least some senses might be construed with structure. This distinction may be inappropriate: as discussed in Section 8, many event predicates have wide (often structured) and narrow readings; a wedding is structured, but a marriage is not; merger may refer to an involved proceedings or the mere act of two businesses becoming one (see Figure 3.8c above). Rather, the events under *Structured event* – *Ceremony*, *Contest*, *Show* and *Election* – might be better described as *Calendar events*, contrasted with others in our typology. Similarly, while we included *Justice* under *Conflict* to emphasise its adversarial nature, it is hard to justify all of its subtypes, notably *Verdict*, as conflicts. While each type may seem to have a locally appropriate parent, we have not ensured that more distant ancestor types are appropriate.

Our hierarchy forms a tree: each type has 0 or 1 parent type. This is an unrealistic simplification of event semantics; types might be better grouped by a set of ontological features, and while this would make the taxonomy more explicit and expressive, it would be more complex to build and modify. For example, we suggested above that *Ceremony*, *Contest* and *Show* are all events with audiences, yet *Election* which has no audience as such also shares attributes of *Contest* such as being a competition with a winner, features not shared by *Ceremony* and *Show* in the sense of a theatre or concert performance. Although

[29] The event types *Gain* and *Loss* would not be distinguished from one another in a refined typology, but are an example of excess granularity within a hierarchy doing little harm. It is less clear whether their distinction from *Trend* is a real one, which would be clearer if their scopes were explicitly specified.

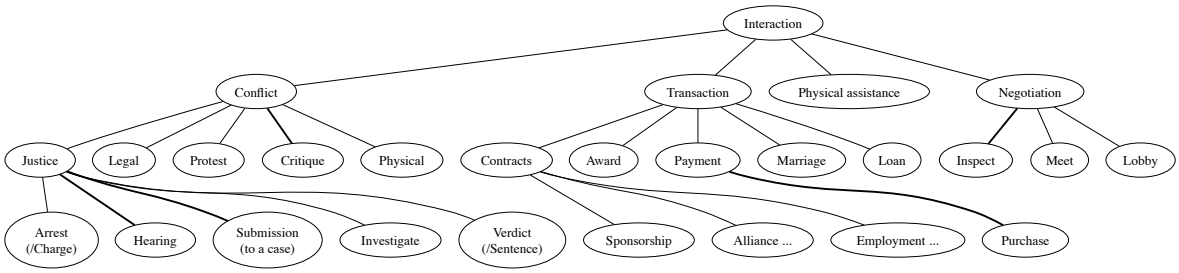


Figure 3.9: The *Interaction* branch of the event typology in story-driven annotation, a duplicate of Figure 3.7.

FrameNet (Baker et al., 1998) allows multiple inheritance between its frames (analogous to event types), this feature is infrequently used, an mostly applies for very fundamental properties of the frame, such as *Intentionally affect* inheriting from *Intentionally act* and *Transitive action*. As of July 2013, FrameNet does not represent the shared competitive nature of its *Change of leadership*, *Competition Fighting activity*: determining appropriate features by which to group event types is itself an open problem. This suggests that finding that a hierarchy – especially a tree – under-fits the space of event types, but may be a practical approximation.

Aside from these inconsistencies, we can locate parts of the typology where its hierarchy is effective in the sense that collapsing types into their parent p produces a set of events that is meaningfully cohesive with respect to events in nodes that are siblings and cousins of p. Consider examples of subtype instances where p is *Unstructured event:Interaction:Conflict* (see Figure 3.9):

(39) THE corporate watchdog is preparing to make its own submission to the Victorian Supreme Court in the legal battle between Brisbane toll-road builder BrisConnections and a rebel shareholder, Nicholas Bolton.
Topic event: type *Conflict:Legal*; domain *Business*; tense *Ongoing*.
Update event: type *Conflict:Justice:Submission*; domain *Business*; tense *Future*.

(40) POLICE have charged an alleged member of the Bandidos outlaw motorcycle gang with a drive-by shooting in Lalor Park last December, bringing to three the shootings he allegedly committed. The man, Todd Obierzynski, 21, had already been charged - with three fellow Bandidos - with firearms offences and two other drive-by shootings at Lurnea and Sadlier on December 10.
Topic event: type *Conflict:Physical*; domain *Crime*; tense *Past*.
Update event: Charge *Conflict:Justice:Arrest*; domain *Crime*; tense *Past*.

(41) Andrew Michelmore, the harried chief of mining debacle OZ Minerals, has become the butt of a series of cruel jibes in the rough and tumble world of mining. Specialist resources websites have been lampooning the OZ boss for weeks over his role in the demise of the

company, formed from a merger of the "growth-rich Oxiana and cash-rich Zinifex" less than a year ago.
Topic event: type *Lifecycle:Bankruptcy*; domain *Business*; tense *Past*.
Update event: type *Conflict:Critique*; domain *Business*; tense *Past*.

Here we have instances of a lawsuit, a drive-by shooting, an arrest, and the online smearing of a reputation. Although these may not generally be construed as the same category of event, the common adversarial nature of these events is apparent. This can be contrasted with *Conflict*'s major sibling (under the parent, *Interaction*), *Transaction* in which mutual benefit for the interacting entities is assumed,[30] as in the following *Purchase*, and the *Employment* (37) and *Merger* (31) events mentioned above.

(42) OZ MINERALS expects to remake itself as a one-mine company - and potential takeover target - on par with copper miner Equinox Minerals, after it agreed to sell the bulk of its assets to China Minmetals for $US1.2 billion ($1.75 billion). The deal would wipe out OZ's debt burden except for $US105 million of convertible notes and leave it with $600 million of cash, the Prominent Hill copper-gold mine in South Australia and a stake in uranium explorer Toro Energy.
Topic event: type *Lifecycle:Bankruptcy*; domain *Business*; tense *Past*.
Update event: type *Transaction:Payment:Purchase*; domain *Business*; tense *Past*.

These are again contrasted with the sibling *Negotiation*, which is neither adversarial nor transactional, and *Physical assistance* which may be construed as the opposite of *Conflict*. Hence we see that a tree-structured typology is able to capture similarities among events as well as fine-grained distinctions not available with fixed type systems, producing an expressive typology for broad coverage. Yet constructing a meaningful hierarchy of types still involves discretising a very complex space, and describing event ontology effectively and compactly remains a topic for future work, while we opt for approaches that leave the event space (largely) untyped.

3.3.3.3 Event type and domain

We characterise events as having both a type and a domain, in order to abstract the ontological semantics of an event from its theme. Top-level event domains are listed in Figure 3.10 and their hierarchy is shown in Figure A.1. The notion of event domain or some other feature/category orthogonal to event type requires further specification than provided in the current annotation. As a result, some event types are not completely abstracted from domain. This is most apparent from *Disaster*'s appearance in both hierarchies; furthermore,

[30] *Unemployment* is a clear exception to this, again a problem where local placement of types does not ensure distant inheritance within the taxonomy.

• Business/Finance	• Entertainment	• Politics
• Civic	• Environment	• Property
• Disaster	• Government	• Science and technology
• Economy	• Human interest	• Sport

Figure 3.10: Top-level event domains after dynamic expansion during annotation

the event type *Disaster* has children *Natural* and *Accident* which might be considered analogous to the *Environment* and *Civic* domains. Only in Example (38) do we find the *Disaster* domain applying, but not the *Disaster* type, because the event in question is a ceremony commemorating a disaster; this knowledge is already conveyed in the distinction of the update from the topic event. Nonetheless, the *Disaster* event type is differentiated across domain in our annotation, such that when combined with *Economy*, it may characterise the Global Economic Crisis; with *Business*, the collapse of the car manufacturing industry. The domains as produced in this annotation are poor at distinguishing a gas pipe explosion from flooding, which are both assigned the *Disaster* type and *Disaster* domain in the current annotation. Hence, while this distinction increases the schema's expressiveness, developing the type and domain hierarchies in parallel does not ensure their clear separation.

3.3.4 Conclusion

Characterisations of news events need to distinguish events that are newsworthy from those are merely peripheral. We have explored this notion by labelling news stories in terms of their focal *update event*, as well as the *topic event* it builds on. Despite using a dynamic hierarchy of event types, the labels do not capture much useful information, which could be richer if a large ontology – not constrained by single-parent hierarchy – were used instead, or if augmented with a set of roles to be filled from the story. Although the notion of an update event seems generally useful, some task clarification is required to distinguish between the event that caused a story to be written, which may not be explicitly mentioned, and the key fact that the story brings to public knowledge, which is not necessarily an event. Our annotation leads us to believe that topics are rarely encapsulated as single events, but that topical context is in itself a complex structure. Hence, where background events are of interest, seeking explicit references and relating them to the update event may be more manageable and meaningful.

3.4 Conclusion

We believe that a characterisation of news events must be driven by data as well as application. As such, this chapter has discussed preliminary explorations into event reference and

structure in news publications and in an online encyclopædia.

We have described two annotation tasks, one closely related to previous work in typed event reference detection, and another more novel in trying to get to the essence of each news story in identifying the event it reports. The results of our type-driven task suggest that such type schemas are brittle and unable to cover the range of news events; and while this task acknowledges that stories may report multiple related events, it fails to account for salience. The second task recognises that news stories tend to be borne of a particular *update event*, which is newsworthy and therefore salient by definition. However, our attempts to grasp a *topic event* as well suggests that topics have complex event structures that may be best understood from news stories in terms of explicit (but not typed) event references.

We find a mismatch between the types of events for which Wikipedia provides structured content and those of interest in the news that necessitates seeking other event characterisation approaches. Nonetheless, it is the notion of event reference as a hyperlink occasionally applied within Wikipedia that inspires our novel approach to characterising event reference.

Chapter 4

Grounding event references in a news archive

References to past, newsworthy events are common in public discourse, whether as background to new information, as a subject of discussion, or in substantiating an opinion. The general difficulty of determining event identity as discussed in Section 2.6 makes computationally emulating the interpretation of such references a challenge, which is exacerbated by their appearance in diverse texts with differing authorial frames of reference. We might go further to suggest that events that are familiar or salient within a public discourse may be referred to with elision of disambiguating detail: in late 2013, informed Australians will understand the reference Rudd's reaccession without explication of the protagonist's full name, when or why the event occurred or what role he attained. Thus a coreference model that relies on locally extracting a structured representation of the event (e.g. Bagga and Baldwin, 1999) might be least applicable to events that are part of popular knowledge. Further, if a terse reference can be associated with a news article in which the referent event is reported, or an encyclopædia article on the topic, a language understanding system may harness these more detailed event descriptions.

Departing from previous work in event semantics and coreference, we introduce *event linking* which focuses on grounding references in a news archive. Grounding concerns the mutual knowledge underlying communication, and the use of language to refer to a particular cognitive entity (Clark and Brennan, 1991). Upon reading a reference to a past news event, an idealised reader will recall it, having read its initial reporting. Given a news archive, *event linking* considers each story initially reporting some past events as a proxy for those events. When an event is later mentioned, an *event link* grounds the reference to its referent's proxy story.

This task is orthogonal to much of the work in the previous chapters; it does not target event detection directly, instead focusing on event identity, enhancing the representation of past events by matching references to a canonical news story. Yet it accounts for a number

of issues identified from our previous annotation experiments and the broader literature:

No typology Event linking is modeled on how reported events become common knowledge through the news publication process. Thus it avoids a brittle event typology[1] in order to capture the diverse events that constitute news.

Update events It adopts the idea from our story-driven annotation that each news story may introduce update events, yet it leaves their explicit characterisation to other extraction models. Unlike that task, it handles references to those events from non-news text.

Salience Like our story-driven annotation, it targets only notable events: importantly, the data determines prominence, rather than annotators; this is in contrast to ACE05 and our type-driven annotation, as well as other minimal-typology event characterisation such as TimeML or OntoNotes coreference. At the same time, the event coverage is atomic with respect to news updates, complementing the much coarser prominent events covered as Wikipedia articles, such as sports contests and military operations.

Liberal identity By considering each update event or events as a single referent, problems inherent in strict coreference are avoided: references that differ in their denotation but refer to the same event, such as *X* murdered *Y* and *Y* died acquire the same event link.

Explicit Unlike our story-driven annotation and Topic Detection, event linking accounts for only explicit references to topical events at a fine textual granularity.

This chapter defines and analyses the event linking task. In the following section, we describe event linking as an abstraction of cognitive grounding. Section 4.2 appraises the task in terms of its potential applications, and Section 4.3 compares event linking to related work in identifying cross-document structure. Finally, we attempt to make the task tangible through an annotated corpus (Chapter 5).

4.1 The event linking task

In event linking, a news archive becomes a discrete proxy for the set of past events that are newsworthy according to a particular news publication. For the set of events within its scope, event linking assigns each a canonical identifier corresponding to the first story where that event is reported in the archive. Thus multiple event references to the same event – and to different components of that event – are linked to the same canonical story.

We continue with the running example from Chapter 2 to illustrate aspects of event linking: salience relative to a publication, reference enhancement via linking, and a focus on

[1] By this we mean types in the sense of MUC and ACE05; we make broad categorical distinctions in terms of the way events are reported in news.

referent over semantics. Reconsider the event references in this Voice of America report from 15 September, 2005:

(43) The United Nations says Somali gunmen who hijacked a U.N.-chartered vessel carrying food aid for tsunami victims have released the ship after holding it for more than two months.

In mapping the event space onto a news archive, event linking only provides non-nil labels to events that are newsworthy in the sense that they are reported in the given archive. For instance, of the events mentioned in Example 43, only the tsunami is reported in the Sydney Morning Herald, albeit across many articles describing its development over place and time; the hijacking and release, but not the chartering, are additionally reported by Agence France Presse (AFP). Just as a reader who acquires her event knowledge from the Herald alone will not know of these events, we cannot link them when that archive constitutes our linking target. In the present definition of event linking, we would not link tsunami to the news archive, since we only consider the class of event that is reported in a single news item, rather than in terms of sub-events across many articles; yet this reference corresponds to and could be linked to a Wikipedia topic, complementing the events of the news. Otherwise, in reflecting the knowledge accumulated by an avid and exclusive reader of a particular news source, event linking is able to ground references to the hijacking and release with respect to the AFP archive.

Where little detail about the hijacking appears in the Voice of America source, a link to the AFP report augments the event reference with further knowledge. The initial report about the hijacking, published on June 30 that year,[2] opens:

(44) A UN-chartered vessel carrying aid for Somali tsunami victims has been hijacked off the coast of Somalia amid a flurry of new piracy warnings for the area, the World Food Programme (WFP) said Thursday.

This article becomes the target of an event link corresponding to the *hijacking* event if the AFP archive is used as the linking domain. It details the place and time of the hijacking, and situates it in the context of other events. Thus the event link enriches the source reference with detail; yet the knowledge of the target article is also enriched by being connected to later comments and updates about the release of the ship. Maximally, assuming appropriate event references can be identified, event linking can induce a directed graph over the stories in a news archive, with arcs corresponding to explicit event reference, as shown in Figure 4.1. For example, the following specific references in the June 30 AFP article have earlier antecedents in that archive:[3]

(45) a. Before Monday, the last reported attack took place on June 7 off Mogadishu when three

[2]Identified as `AFP_ENG_20050630.0069` in Gigaword (Parker et al., 2011).
[3]Identified as `AFP_ENG_20050608.0075` and `AFP_ENG_20050317.0335` respectively in Gigaword.

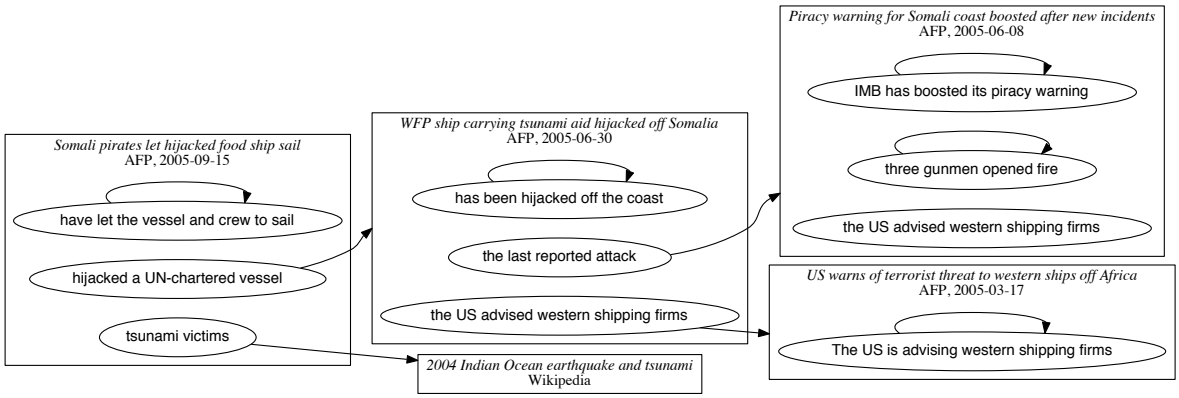


Figure 4.1: Part of an event link graph induced among AFP stories by manually identifying linkable events. Edges from a reference to itself indicate that the containing article is the correct event link target, since it newly reports that event.

gunmen in a white speedboat opened fire with automatic weapons on an unidentified bulk carrier, according to the IMB.

b. In March, the United States advised western shipping firms of possible speedboat-launched terrorist attacks on vessels in the Indian Ocean off the coast of east Africa, including Somali waters.

The event link graph groups references to an event, but also tracks topics through stories with related background. Through both local content and an induced link graph, event links are able to enrich an event reference with further knowledge.

The *release* event can similarly be linked to an article from AFP. However, the target article[4] describes the event as Somali pirates let hijacked food ship sail, rather than a release, saying that United Nations and Kenyan officials refused to confirm the vessel had been definitively released, given the unreliability of the Somali hijackers. Although the implications of release in Example 43 and let sail here differ, it is clear that they refer to the same event. Thus event linking intends to avoid some of the interference of mismatched semantics in identification of event coreference.

By following through an example, we have illustrated some properties of event linking. We now proceed to define it more formally, among a more general set of grounding tasks, and to make its scope more explicit.

4.1.1 Event linking as a grounding task

Event linking is among a family of computational grounding tasks that approximate identity within a non-enumerable referent space with a discrete, finite set of representatives. Thus where full language understanding involves interpreting references to fictitious and famous

[4]Identified as `AFP_ENG_20050915.0197` in Gigaword.

entities alike, these focus on referents in shared knowledge. This complements traditional information extraction tasks for determining reference-related semantics and attributes.

Wikipedia's representation of shared knowledge encouraged a similar approach to the computational processing of named entity references. Prior to Wikipedia's prominence, systems sought to recognise, classify and identify coreference between entity mentions (Sundheim, 1995; Bagga and Baldwin, 1998). While Wikipedia's knowledge has been harnessed for traditional named entity recognition and classification (Nothman et al., 2013), it is often used as a knowledge base (KB) when matching names to referents in the named entity disambiguation or linking task (Bunescu and Paşca, 2006; Cucerzan, 2007; Hachey et al., 2013).

Where a referent is matched, traditional extraction tasks are implicitly solved: Wikipedia's data can be used to infer fine-grained type information (Nothman et al., 2013), and multiple links to the same entity entail cross-document coreference.[5] But NEL has a major limitation: it can only match *notable* entities represented in the knowledge base.[6] Since KB entries are discrete, NEL may also be underspecified for near-matches:[7] should Elsevier Australia be linked to Elsevier, its global parent company? Links may not distinguish contextual semantics or metonymy, such as whether a particular mention of University of Sydney denotes a location or an organisation. NEL therefore provides reduced coverage but enhanced, fine-grained knowledge with respect to traditional entity language processing.

Though analogous to NEL, the event linking task differs in the types of expressions that may be linked, and the manner of determining the correct KB node to link to, if any.

In event linking, a corpus of articles published as news acts as a KB for grounding. Just as a NEL KB represents notable entities, an event linking KB represents *newsworthy events*; a particular archive implicitly incorporates its publisher's notion of newsworthiness. A caveat is that the threshold for publication is relative to other newsworthy content at the time; unlike Wikipedia, many news publications are guided by a quota as well as notability.

Event linking defines a canonical link target for each event: *the earliest story in the archive that reports the given event happened or is happening*; for brevity, we say this story *first reports* the event. The model therefore excludes predictions or descriptions of future and hypothetical events. Each archival story implicitly represents zero or more (but often uncountably many) related events, just as Wikipedia entries represent zero or one entity in NEL. The intention is that a story represents the event(s) that caused it to be written, as well as their sub-events.

[5]The entailed cross-document coreference has the caveat of only including named references; references such as the Prime Minister or my alma mater would generally not be considered linkable.

[6]Wikipedia editors are expected to apply notability guidelines (for English Wikipedia at `http://en.wikipedia.org/wiki/WP:N`) to determine whether a topic deserves its own article. Acting as a gazetteer, this tends towards completeness for geopolitical districts; in general, the systemic biases in English Wikipedia limit its usefulness as a KB to entities that are not notable within populous, educated regions of the English-speaking world.

[7]A similar issue arises in attempts to map an infinite semantic space onto a taxonomy or ontology, such as in named entity recognition.

4.1.2 Event scope limitations

Archive scope As in NEL, the KB (i.e. the news archive) used implicitly defines the scope of events – in time, type, and newsworthiness – that may be linked to it. In the present work, we only consider a news archive from a single publisher: combining multiple could increase coverage, but increases the difficulty in ensuring that each reported event corresponds to a single canonical article;[8] concatenating corresponding articles from multiple archives could produce a KB with higher coverage.

Story events scope In order that targets can be identifiably canonical, we must limit a story s to representing events which are:

reported in s news; not merely commented on, assumed knowledge, or background;

reported as fact avoiding the troublesome space of hypothetical and speculative events;[9] and

completed or happening since reports of future events are effectively hypothetical.

Reference scope When considering the breadth of expressions that refer to events, it becomes apparent that many references to past, newsworthy events cannot be linked to a KB entry under the above regime. Among these are references to:

- *multiple* or underspecified past events such as four separate punishments or his punishment continues, US drone attacks in Pakistan since 2004;
- a *compound* event reported over multiple articles in terms of its sub-events, such as 2012 election, 2009 FIFA Soccer World Cup or Watergate scandal; or
- an *aggregate* event that emerges from other events reported over time, such as the GDP grew 15%, the rise of the United States, or scored 100 goals.

The former two categories may be better understood as topics or named events whose notable instances are often represented by entries in Wikipedia. In each of these cases the referent event space may overlap with events first reported in one or more articles, but is not a subset of any. Rather, event linking specialises in the granularity of incident that directly triggers news through the occurrence of the event.

[8] We note however that even within a single news archive, there may be multiple articles concurrently reporting aspects of the same event. We leave this issue for future consideration, although we account for ambiguity within the same day's publication in our inter-annotator agreement evaluations in Section 5.4.2.

[9] Within the context of grounding events in an archive, it is reasonable to assume that events reported as fact actually occurred: the reader is expected to think so, such that it becomes mutual knowledge in later communication. However, since much news reporting includes attribution of event knowledge to external sources, we may need to assume these sources are credible, and their statements factual, unless suggested otherwise.

4.1.3 Co-reporting as approximate coreference

References in NEL that are linked to the same KB entry are coreferent. In event linking, for a reference r to be linked to some target story s, its referent need only be a subset of the event space reported in s. Hence, multiple references linked to the same target may overlap in the events they refer to, but may be entirely disjoint.

As an example, consider a KB entry s first reporting a particular attack that consisted of missiles being fired 3 minutes apart on two different locations L_1 and L_2. Now consider the following decontextualised references:

(46) a. last week's attack

b. the direct hit on a house in L_1

c. three were killed in L_2

These may be interpreted such that the referents of 46b and 46c are each subsets of 46a's referent, although 46b and 46c do not overlap. Indeed, 46b and 46c are neither co-located nor cotemporaneous, and the references would not have the same event type under the ACE05 guidelines (LDC, 2005), which would type them as *conflict:attack* and *life:death* respectively. Nonetheless, they have the same link target; we say their referents are *co-reported*.

Not all co-reported events are as closely related as those in our example. When contrasted with defining an intricate ontology of relationships between event references and detecting them, we find co-reporting a worthwhile approximation to coreference. While recent work has sought to annotate near-identity as well as identity (Recasens et al., 2012; Hovy et al., 2013a), our approach adopts a broader unit for event identity in accordance with how it is reported. Hovy et al. (2013a) find that within a document, events tend to be near-identical by way of *membership* in a series of events (or generic event reference) and one event being a *sub-event* of another. While such relationships may be inferred from closely reading a single text, similar inference of event relations across documents is likely to find more diverse relations, e.g. *sibling*, which the co-reporting approximation, albeit underspecified, may account for.

4.1.4 Summary

Event linking models a limited form of cross-document event coreference that is concerned only with disambiguating references to newsworthy events. Grounding events to canonical story identifiers ensures they can be identified across perspectives and applications. Under the event linking model, each article in a news archive is a proxy for the events it reports, providing an approximate form of event identity that we argue is reflective of flexible event-referential language, and limiting the task's scope to notable events as defined by a news source. This associates references to covered events with a canonical identifier corresponding

to a particular archival story. Hence references to events that are reported together, including an event and its sub-events, are grouped together.

4.2 Utilising event links

Having motivated event linking by comparison and contrast with other NLP tasks, and in relation to a cognitive model, we extend its motivation in terms of some potential applications.

Semantic hypertext construction Traditional news media consist of text and images, using physical placement for readers to navigate the news. Despite many traditional news sources being available online, few of them at present take advantage of hypertext to connect in-text references to past stories.[10] A straightforward application of event linking would allow online news journalists and editors to select event references to be automatically hyperlinked to past stories from the same publication; or news publishers could provide a tool to automatically link from external publications, such as blogs, to their news archive. Readers who are unaware of an event, or interested to see how it was first reported, may then follow links as references to additional information.[11] Since event linking prescribes a specific relationship between anchor text and link target, such hyperlinks describe a precise semantic relation that could also be represented as a Semantic Web RDF triple.

Applied exhaustively with manual or automatic reference selection, event links may transform a news archive into a web of textually connected documents. This web may be "surfed" by interested readers in much the same way as topics are within Wikipedia, providing a new means of accessing historical news.

Adaptive background referencing in news An analysis of event reference could enable journalists to model what events their readers are likely familiar with (for example, what has been referred to recently) and what requires extended contextualisation. Indeed, event links could drive the summarisation of past news for insertion as background reference in new content. Similarly, readers could suppress or expand descriptions of background events.

Topic tracking and threading Exhaustively linking past references within a news archive to itself produces a directed acyclic graph among news stories, under certain conditions including that a reference in some story may only be linked to a story with a strictly earlier publication timestamp. Graph-based analysis could then be applied to identify hub events or event-oriented topical threads, or to construct timelines of related events.

[10] It seems much more frequent to include hyperlinks to related stories from outside the story text.

[11] The event linking target, while canonical, may not be the ideal text for a reader to learn about an event. A more ideal target might be identifiable through topic tracking and network analysis methods, discussed below.

Corpus analysis A corpus annotated with event links could be used to analyse the discourse structure of event reference. For example, one could explore factors in the distribution of time differences between each event reference and its referent event. One might also examine how a particular event – or class of events – is referred to, and how this changes with respect to time and sociolinguistic variables.

Improved event extraction By interpreting references that are within event linking's scope, we come to a better understanding those event references that are out of scope. Shared evaluations of named entity linking have introduced a *nil clustering* task, which seeks to identify cross-document coreference for entities not within their KB. We intuit that knowledge can be transferred from in-KB references to others, enabling us to learn models of co-reference for less-notable events, and exploiting details from lint target articles to inform other event characterisation.

4.3 Related work

Although we reviewed the characterisation of event reference in Chapter 2, there is substantial literature relevant to event linking that does not focus on event characterisation. This spans from work in grounding references to the induction of graphs over topically-related news text.

Event linking is situated among a family of disambiguation tasks that identify that a fragment of unstructured text indicates one of a set of referents. Connecting medical terms with an ontology (Aronson, 2001) and the resolution of place names (Leidner, 2004) are now long-established tasks. The growth of Wikipedia, which provided alias knowledge and disambiguated references in context, resulted in its use for more general disambiguation of references to notable named entities (Bunescu and Paşca, 2006; Cucerzan, 2007). In all these cases, linking textual references to the knowledge base establishes coreference and enables inference from related structured knowledge. Performing any of these tasks involves recalling a set of candidates for a term, and disambiguating the referent by using context – Sydney should be understood differently when mentioned near New South Wales as opposed to Nova Scotia – and encoded prior expectations – e.g. prefer a capital city to a town of the same name. All of the above consider more homogeneous forms of reference, and a more structured knowledge base, than are available for event reference.

More recent work by Fokkens et al. (2013) grounds textual event references to Semantic Web descriptions within the Simple Event Model (van Hage et al., 2011). This framework, established through an initial annotation of references to earthquakes, ties in but is not limited to textual representations of events. It attempts to capture event interrelation and disputed factuality, and has the potential to provide for event-oriented inference. It is therefore much more expressive and structured than our event linking model, but it remains to be seen how

well the breadth of event reference can be encoded and processed in this framework, or how references might be automatically resolved to existing referents.

Arapakis et al. (2014) perform a task resembling event linking, albeit motivated by user interface more than linguistics. Searching within news text, they identify candidate newsworthy event references consisting of a named entity adjacent to a past-tense, transitive "action verb"[12]. They query an index of news articles from a single source with the verb and words from its subject and object noun phrases, scoring each candidate according to the length-normalised dot-product of its term frequency vector[13] with the query. A link to the top candidate is accepted if its score exceeds a fixed threshold. As a task theirs differs to ours by: only considering particular forms of event reference; and seeking an article describing the query event, not a canonical target. In this vein, their evaluation centered on the usefulness of the hyperlinks to readers: Arapakis et al. (2014) compare their output to that of independent professional editors who annotated news articles with links "that were perceived as related and newsworthy and that would provide interesting insights with respect to the main article". Applying no score threshold, only 7% of automatic links overlapped with the manual annotations, producing a similar number of links in 75 articles, but no links in the other 125 articles manually annotated. In the opinion of Amazon Mechanical Turk workers, the manual links were significantly better located, related, newsworthy and insightful, although the effect is small, accounting for less than 10% of the variance in the sample. An alternative evaluation found 30% of automatic links as good or excellent, with another 35% fair and 35% bad, according to professional editors. Overall the evaluation suggests that their approach could scalably support manual curation in creating links between articles to enhance the news reading experience.

Other work focuses on connecting new media to more traditional news sources. Guo et al. (2013) predict hyperlinks pointing from 17 days of Twitter content to news sources CNN.com and NYTimes.com over the same period, after removing tweets trivially consisting of news headlines or summaries. They propose a system that first reduces both texts to a latent dense vector representation in which they are then compared, using the textual and temporal neighbourhood of a tweet to enhance its representation. Their model improves on a term-based retrieval baseline, achieving a mean reciprocal rank of 0.490 from 0.463.[14] Note that they mention no handling of overlapping content between the two news sources; they do include non-report articles such as reviews and presumably opinion, and so do not exclusively deal with event reference. The recently-concluded Sync3 project attempts to

[12]They describe this as excluding "be, become, seem, grow, etc." but do not describe how these exclusions are determined.

[13]Terms in the candidate article's title are weighted 3x those in the body, selecting for articles focusing on the sought event.

[14]They optimise performance for the average top-k hit rate, a metric designed to elegantly handle missing data in recommender systems (Steck, 2010). Under this metric their absolute gains are greater with comparable relative gain.

connect event references within the blogosphere to clustered news abstracts, and is reported to correctly match the mentioned event in 73% of the instances evaluated (Bounegru and Karstens, 2012), although a detailed methodology is not yet published. As with other real-time processing, both these tasks may rely heavily on the assumption that social media and blogs predominantly refer to very recent news. This is an assumption we would like to avoid in event linking.

Our work begs comparison to the news event model considered by topic detection and tracking (TDT). Its goal is to group news stories mentioning the same topic, and thus relating to "a [seminal] event or activity, along with all directly related events and activities"[15] (Allan, 2002), although the precise relationship between topic and event varied across evaluations and corpus annotation efforts (Allan et al., 2005). One component of TDT– First Story Detection – involves recognising an article that introduces a new topic relative to prior news, which resembles but is not identical to event linking's concept of *first reporting*. TDT's approach to reference is coarse-grained relative to event linking let alone more precise coreference approaches, both in terms of the referring linguistic unit (document) and the referent (broadly related events are conflated). Topics from three different evaluation iterations include: *Boris Yeltsin's illness*; *Taiwanese Premier Tang Fei Resigns*; and *Murder of the Palestinian Child Mohammed El Dorra*. In this last case, reactions to the incident and retaliations are grouped together; yet this topic is still occasionally in the news a decade later, through a defamation case and a series of appeals. At what point this becomes a separate topic is unclear, with similar issues motivating hierarchical topic detection (Feng and Allan, 2005). Feng and Allan (2009) further granulate the task by considering labelled relationships (e.g. *follow-up*, *prediction*) between sub-document passages within TDT topic clusters, which compares to more linguistically oriented and specified approaches to labelling cross-document rhetorical structure (Radev et al., 2004a) or event reference (Bejan and Harabagiu, 2008). Other differences from TDT include that the three or five month periods of news covered in its evaluations do not afford vast changes of reference frame, allowing topical clustering to rely heavily on textual similarity with recency heuristics; TDT only considers news text, which projects like Sync3 mentioned above rectify to some extent; and TDT does not provide a notion of canonical identification of referent independent of the collection of news sources.

Techniques from TDT are applicable to this task; indeed, one event linking solution may involve indentifying the appropriate cluster and selecting its first story. Yet the present task requires connecting more precise and numerous references to finer archival events. Hence, rather than First Story Detection (FSD), a system might identify and index only the event content that a particular article newly introduces with respect to the archive. This relates closely to the Novelty Detection task in information retrieval, which seeks to highlight the

[15] *Rules of Interpretation* conditioned on broad event theme (*elections*, *natural disasters*, *new laws* etc.) attempt to specify the extent of relatedness.

content in a retrieved result that is relevant to the query and unknown from higher-ranked documents (Soboroff and Harman, 2005).

Much work in TDT has focused on an online or streaming setting, wherein a system may only harness a limited amount of context published after the document being clustered. A similar setting is also pertinent to event linking, again given the ability to detect linkable references. Online TDT systems tend to employ single-pass clustering, particularly for FSD, which determines if each incoming document is similar to any indexed content above a set threshold, using some task-oriented document representation. If not, it is identified as a "first story" and indexed; otherwise, the indexed representation may be augmented by the incoming content, building a clustered representation, or the incoming document may be indexed separately (Allan et al., 1998; Yang et al., 1998; Roberts and Harabagiu, 2011). As with recent work on FSD in Twitter (e.g. Petrović et al., 2010), the granularity of event links may motivate scalable techniques, which avoid this pairwise comparison with all previous input. Petrović et al. (2010) thus employ Localitly Sensitive Hashing (LSH) to estimate cosine similarity for approximate nearest neighbour retrieval (Charikar, 2002), but backoff to an exhaustive pairwise comparison with a fixed number of recent documents. They also constrain the LSH index to constant space, discarding the oldest documents upon overflow. This has the intended side-effect of preferring to match incoming documents to topics that have been mentioned recently. Similar techniques could well apply to event linking, where we see in the next chapter that event references within a news publication are most often to recently reported events. However that annotation also shows that many references are to events of the more distant past, including many that are unlikely to be often referenced in the intervening period. While there is great similarity between TDT tasks and event linking, the long tail of event references we consider here may not admit some of the engineering techniques applied successfully to that work.

4.4 Conclusion

This chapter has introduced a new task, event linking, that addresses the problem of identifying cross-document reference to the same event. Other models of cross-document event coreference have been limited to: a structured representation of a particular type of event (Bagga and Baldwin, 1999); evaluation only within clusters of articles reporting roughly the same story, albeit without semantic restriction (Bejan and Harabagiu, 2008; Lee et al., 2012); or a notion of topical relevance rather than explicit event reference (Allan, 2002). Event linking considers only the set of events reported in a given news archive, particularly the granularity of event that is reported in a single news story, without some limitations present in earlier work: it is able to consider unstructured references to an event through vast changes in perspective and time. Just as named entity linking focuses on notable entities, event linking

focuses on notable, or newsworthy, events, while considering more fine-grained events than the topical coverage of Wikipedia or Freebase.

The set and granularity of events targeted by event links depends on the selected archive. This reflects the role of news media in defining the shared event knowledge of its readership. By opting for a data-driven definition of notability, we avoid schematising this complex referent space, while reducing the potential for annotator subjectivity and recall problems.

The event linking model does not attempt to represent each reported event individually. Not doing so allows event linking to account for approximate event identity, such as references to the same event structure with different connotations or perspectives (e.g. X murdered Y vs Y died), without constructing an ontology of various often-indeterminate relationships between events and their mentions as in Bejan and Harabagiu (2008). It also allows newsworthy events to be assigned canonical identifiers – being unique IDs of archived articles – independent of processing that determines the precise set of events reported in a text, which we believe cannot be done reliably. Instead, event linking conflates events that are reported together, which may or may not constitute a single event from the perspective of a reader; this errs in the opposite direction of previous event coreference work, which tends to adopt a strict approach to event identity.

A mention of an event may be locally poor on detail; linking it and other mentions of that event to a canonical article may provide further knowledge for understanding the local reference, just as a link to Wikipedia provides further knowledge about the mentioned entity. This is clearly the case when the canonical article details the event, although further processing may be required to identify relevant attributes. Still further knowledge could be obtained from the set of documents linking to that article; similar cross-document inference has been found beneficial for typed event extraction (Ji and Grishman, 2008). As event links are only indicative of co-reporting, rather than identity, event linking introduces the new problem of interpreting the relationships – event-event relations as well as differences of perspective – among the set of references that share a link target. Thus, like named entity linking to Wikipedia or Freebase, event linking can provide enhancement in addition to coreference, but in this case obtaining structured knowledge depends on further processing.

We have defined the event linking task, its key characteristics, its relationship to existing literature, and some potential applications of its reference model. The remainder of this thesis evaluates the feasibility of event linking, while identifying the challenges it poses. In the next chapter we describe compiling a corpus of manually-annotated event links. Replicating these annotations then becomes the goal of a system described in Chapter 6, which allows us to identify event references that are easy or more difficult to disambiguate through the evaluation in Chapter 7. Searching for and disambiguating among event link candidates turns out to involve great effort for human annotators, so a system that is able to replicate manual annotations may also assist in the construction of future event link corpora.

Chapter 5

Annotating a corpus of event links

In seeking a deeper understanding of the event linking task introduced in the previous chapter, we have developed a collection of documents annotated with event links. In abstract, event linking may pertain to any reference to past news events; for annotation we must apply it to a particular situation. We therefore consider the task of linking all past newsworthy event references within a sample of a news publication to previous articles from that same source, and design an annotation task schema to do so. To facilitate the annotation work, an extensible web-based annotation tool is built for a family of reference annotation tasks, and we customise it to our schema.

Having successfully outsourced named entity linking annotation to non-expert freelance annotators, we attempt the same with this task, but find it is not as well suited. Hence our work also contributes a discussion of this approach to hiring annotators in qualitative comparison to a more orthodox *expert annotation* approach and recent literature exploiting crowd redundancy to inexpensively annotate corpora through very small subtasks.

The task turns out to be very labour-intensive. It involves identifying all linkable references, constructing and refining search queries, and filtering through a potentially large quantity of candidates in search of a single target – or else to assure oneself that the target does not exist. These together require extended critical focus; correspondingly, we only obtain a small manually annotated corpus.

Over a collection of 150 annotated news and opinion articles, we analyse disagreement among annotators, and identify interesting aspects of the event linking task through the annotations. We find that although annotators are not able to reliably identify event references that may be linked, a canonical link target is usually determinable given an agreed reference. An analysis of the identified links highlights that in many cases, mere temporal and lexical similarity between the source and target of an event link is not necessary, although this assumption underlies much previous work in tracking events over time.

This chapter first describes our approach to event link annotation as a task (Section 5.1), before detailing the practical aspects of working with non-expert annotators to produce a cor-

pus (Section 5.2), and the news archive we both mark up and link to (Section 5.3). Section 5.4 then analyses the resulting corpus and the disagreements identified between annotators, before Section 5.5 concludes with a discussion of issues relating to our annotation task's design.

5.1 Exhaustive event linking annotation within a news archive

Event linking – as an approach to coindexing reference to events – does not entail a method of sampling references to annotate; nor does it prescribe a particular news archive to link to. These parameters to the event linking task may have a number of possible settings, and in a similar manner, named entity linking and related tasks have been developed and evaluated using: pre-existing hyperlinks to Wikipedia articles (Bunescu and Paşca, 2006; Cucerzan, 2007); post-hoc analysis of system output (Cucerzan, 2007); exhaustive joint annotation of entity references (unitising) and their disambiguation (Kulkarni et al., 2009; Radford, 2014); annotation of disambiguation overlaid on all named entities marked in an existing corpus (Bentivogli et al., 2010; Hoffart et al., 2011); and annotation of selected references only in order to to focus on ambiguous references (Simpson et al., 2010), there identified by a review of automatic named entity recognition, and later by annotators manually searching for interesting references (Ellis et al., 2012). We opt to focus on the way events are referred to in a news publication – though not exclusively in the news report genre – and hence our annotation task extends upon basic event linking in the following ways:

Archive-internal linking The documents annotated with event links are sampled from the target archive B. Hence some event expressions may be first reported in the document being annotated. One advantage of this setting is that annotators become familiar with the content and style of that publication while both annotating and linking to it.

Exhaustive linking All appropriate event references are annotated in each document considered, in contrast with the pure event linking definition above which seeks a target for a given reference.

No future links References to events that happen in the future with respect to the referring document are not annotated. Although one could link future-time references once the referent event had occurred, such cases present both semantic and technical difficulties.

The annotation is performed following the schema in Appendix B.2. The schema is intended to be accessible to non-linguists who were not otherwise involved in the project.

Annotation proceeds one source document at a time, with annotators seeking to mark a single word (a *mention*) corresponding to each newsworthy, completed-or-ongoing event reference identified.[1] If the annotator knows the reference is co-reported with one already

[1] We do not attempt a precise definition of *event*, but describe *newsworthy events* as "Things that happen and directly trigger news". We ask that annotators only identify explicit references to events, providing a counter-example.

marked, they may assign it to the same chain of references and proceed to the next reference. For each event marked, the annotator decides whether the event is within the event linking reference scope (*linkable*), or otherwise selects from among the unlinkable categories of *multiple*, *compound* and *aggregate* mentioned in Section 4.1.2.[2]

For each *linkable* event identified, the annotator may identify a canonical link target, or mark the event as *not found*, *precedes archive* or *reported here* (i.e. the document being annotated is reporting the event and no earlier news story does so; an event link targeting the document being annotated). To identify a canonical link target, the annotator performs a keyword-based retrieval over the archive,[3] optionally constraining the publication date of the target to within, before or after a given year, month or day. The annotator may select from up to 100 search results, ordered by ascending or descending publication date, or by descending score[4] with respect to the query. Only candidates whose publication date equals or precedes that of the source document are available to choose from.

For each *compound* event reference, the annotator may link it to a Wikipedia article specifically about that event. We expect the replication of such links may feasibly exploit the Wikipedia features discussed in Section 3.1, but these annotations are largely intended to assist annotators in event categorisation, and are not utilised in the present work.

5.2 Annotators and adjudication

5.2.1 Non-expert annotation and outsourcing

In addition to the author and some local undergraduates, we hired annotators using Freelancer.com. Due to the cost of expert linguistic annotators, a lot of recent work has considered acquiring annotations from non-experts through crowdsourcing platforms such as Amazon Mechanical Turk (AMT; Snow et al., 2008; Callison-Burch and Dredze, 2010) or through *gamification* of annotation tasks (e.g. Chamberlain et al., 2008). In contrast with popular crowdsourcing techniques that focus on small, independent annotation tasks, our annotation process involves understanding whole documents and an intricate schema. Aware of this complexity, we sought to employ non-experts who we could retain through lengthy periods of training, annotation and task revision. Since this approach is less familiar than others in the literature, we provide a qualitative summary comparing aspects of these annotation outsourcing models in Table 5.1.

[2] In practice *aggregate* was labelled *trend/change* and also incorporated references to price movements and the like; *multiple* was labelled *many/generic* and *linkable* was labelled *basic*.

[3] To provide keyword search semantics similar to popular search engines, we use the DisMax query parser (`http://wiki.apache.org/solr/DisMaxQParserPlugin`) within Apache Solr. By default a conjunction of entered keywords or phrases is matched, with the `OR` operator available for arbitrary boolean combinations. Stop words are removed and terms are Porter2-stemmed to match against the index.

[4] Lucene's Practical Scoring Function – an efficient approximation of cosine similarity with tf.idf (Apache Software Foundation, 2012) – is applied to match the query q against a candidate c's body and title texts, producing $\sigma_{body}(q, c)$ and $\sigma_{title}(q, c)$. Then the overall score is $\sigma(q, c) = max\{\sigma_{body}(q, c), 5 \cdot \sigma_{title}(q, c)\}$.

	Expert annotators	Online freelancing	Human Intelligence Task-style
•	High **cost per annotation**	Moderate-to-low cost per annotation	Low cost per annotation
•	Paid at award **rates**	Paid according to bid per hour (or per task)	Paid fixed price per task unit
•	Low **annotation redundancy**	Moderate annotation redundancy	High annotation redundancy
•	Moderate **training cost**: less training effort (and possibly face-to-face) at a higher pay rate.	Relatively high training cost: intermittent supervision and review is required; failing candidates may require pay.	Low training cost: training materials and qualifier tests are set up in advance and require little supervision.
•	**Annotator selection and retention** is controlled by the hiring process, and may be bound by a predetermined contract. Experts may not be available for intermittent work while revising the task.	Annotator selection is flexible, but quality may be hard to predict from bid. Can contract small tasks and rehire quality annotators over a long period of time, although availability and pay rates vary.	Annotators are anonymous, but may be selected by personal attributes and certification, custom qualification tasks and outlier removal. Cannot retain quality annotators.
•	Annotators may work for long stretches on **large tasks** with sophisticated annotation guidelines.	Large tasks can be assigned, but work and attention to them may be intermittent, and annotation schemas may need simplification.	Suitable for small, straightforward tasks to be completed in seconds or minutes. Once submitted, annotators may not reconsider their work.

Table 5.1: Comparison of three annotator hiring models: traditional hiring of an expert; online, low-skill freelance marketplaces; and the Amazon Mechanical Turk model. We do not focus on hiring of local non-experts, which we also employed for this task, and which shares properties with the first two columns.

Under this model, annotators are selected individually after they nominate a rate of pay and introduce themselves. A strategy for hiring annotators given their bids is not obvious; it is hard to convey the nature of a task to candidates such that they can accurately estimate their aptitude and pay rate.

After hiring, annotators are then trained and their work reviewed with the help of instant messaging. This personal approach allows quality assurance and training investment in a way that AMT does not, yet the initial supervision also tends to be vastly more time-consuming than face-to-face or completely crowd-based approaches. This also allowed us to approach annotators who were competent at our other annotation tasks and hence already familiar with linguistic annotation and our user interface. Despite the personalised training, quality of independent annotations is not as well assured with freelance annotators as with experts; yet while redundancy underlies quality control with AMT, setup costs under the freelancing model make producing similar redundancy expensive. The remoteness of the annotators – while possibly also applicable to an expert annotator context – further impedes training. In both these approaches, best practices for annotator training, including an iterative procedure of training and testing, and conferencing between annotators, are not readily accomplished.

This outsourcing approach was successful for other tasks, such as quote attribution (O'Keefe et al., 2012) and named entity recognition and linking (Radford, 2014), but may have not been well-suited to event linking annotation. Annotator quality varied in all tasks, with some correlation to their background and hiring costs, but most annotators who attempted training for those tasks continued to annotate further documents with high inter-annotator agreement.[5] The same was not true of event linking: of nine annotators hired using Freelancer.com only three produced training annotations of sufficient quality to proceed to further documents. It also became apparent that a number of annotators – including some with the highest qualifications and work quality – were only available to work intermittently amid other professional or domestic activities; the event linking task requires prolonged concentration relative to the other tasks where decisions are more independent and a surface understanding of the document suffices. Annotators who had successfully performed the other annotation tasks therefore opted out of the more involved annotation for event linking.

Further, working with freelancers without expertise in linguistics also entails a simplified annotation schema,[6] which may require translating technical, precise concepts into compact, simpler, but more ambiguous statements. Thus although our pilot schema in Section B.1 was underspecified and yielded a lot of variation among annotators, our impression is that most found our later schema in Section B.2 burdensome in its size and detail, despite being small

[5]For a named entity linking annotation of Sydney Morning Herald 2009 documents, we calculated annotator-pairwise chance-corrected Cohen's κ agreement of 0.80 to 0.88 over token entity type annotations and 0.85 to 0.89 over token entity link annotations. For quotation attribution, O'Keefe et al. (2012) report raw agreement of 0.98 over 400 double-annotated articles.

[6]This is an assumption of ours that would be worthwhile testing in future work: what level of quality can we expect from non-expert annotators working with existing, linguistically detailed annotation schemas?

relative to linguistic annotation schemas designed for experts.

The choice of annotation outsourcing approach is therefore task-dependent: some tasks that are too involved for an AMT-style crowdsourcing setup may also be too involved for the casual freelance market. The experimental nature of this task's design and the depth of understanding required for annotation thus made event linking unsuitable for a freelancing approach.

Thus the work of Freelancer.com annotators often required substantial revision and correction. Fortunately, web-based interaction ensures this supervision is recorded. Annotators were asked to correct their work for:

- marking references to non-events, such as emissions trading scheme;
- annotating very sparsely, missing many newsworthy event references;
- labelling *multiple* references as *linkable*, even when as generic as dealings;
- linking to articles that do not report the event happening, including news reports where the event is background, and opinion articles;
- constraining the date of the target too tightly and thus missing the correct link;
- using phrasal search where terms were unlikely to appear as a phrase, and thus missing the correct link;
- searching for the expected headline of an article identified using a web search, rather than entering keywords directly into our annotation tool, often resulting in false *not found* annotations because the online headline differed from our archive.

While in general it can be difficult to understand an isolated news story, some of these annotators also lacked sufficient domain knowledge to identify, for example, indirect references to the global financial crisis of 2007-8.

Ultimately, we employed local undergraduate humanities students to double-annotate the work of Freelancer.com workers (predominantly that of a single reliable annotator). Like with the freelancers, this required direct supervision during training, easier in a face-to-face interaction, but without the costly false starts experienced in the online outsourcing approach; the local annotators produced similar quality of annotation at a higher cost per document.

5.2.2 Annotation procedure

For annotation, each annotator was guided through the worked example in Appendix Section B.3[7] and their work was reviewed on at least two training documents, before annotating other document collections more independently.

[7]The worked example was not available during pilot annotation attempts.

To ensure high recall and consistency, each document was reviewed at least twice. All documents were annotated by a single external annotator, and a portion was double-annotated (detailed below). All annotations were then adjudicated or corrected by the author.

5.2.3 Annotation tool

To facilitate whole-document annotation, we have developed an extensible, web-based annotation tool for reference extraction tasks, currently supporting annotation for named entity recognition and linking, type-driven event detection as discussed in the previous chapter, direct and indirect quotation attribution, and event linking. Common to all of these tasks are:

- annotators mark up whole documents at a time;
- markable units are non-overlapping tokens or strings of tokens within sentences;
- units cannot be completely specified in advance, although hints may be provided, such as marking all strings of capitalised words as candidates;
- units form coreference chains; and
- most of the annotation occurs with respect to a chain of units (e.g. a named entity link or attribution of a sequence of quotes), rather than annotating attributes of units.

Each task is configured by a JavaScript object describing: the interface widgets assigning attributes to a chain, which define their own validation; the classes of unit that can join a chain[8]; and a function to mark up potential candidates.[9]

Sampled documents to annotate are grouped into *tasks*, and a number of these are assigned to each annotator. The annotator may proceed through documents in order and return to selected documents, which are colour-coded according to their completion status (unannotated, valid, or invalid).

For a given document, the interface (see Figure 5.1) allows a user to mark a textual span as a unit assigned to a new or existing chain using a mouse-triggered menu or a keyboard shortcut, with each chain assigned a distinct colour.[10] Chains are annotated in a sidebar and the document is validated for each modification; the annotator may also add document-level comments. The user may also cycle through viewing all units assigned to a chain by clicking the chain's title.

[8] For event annotation, there is a single class of mention. Otherwise this might distinguish between proper or pronominal entity mentions; or whether the quotation unit selected is direct, indirect or the reference to speech introducing the quotation.

[9] Although we considered approaches to marking event candidates, we do not use this feature for the present work.

[10] Units may also be modified (expanded or contracted textual spans) or removed.

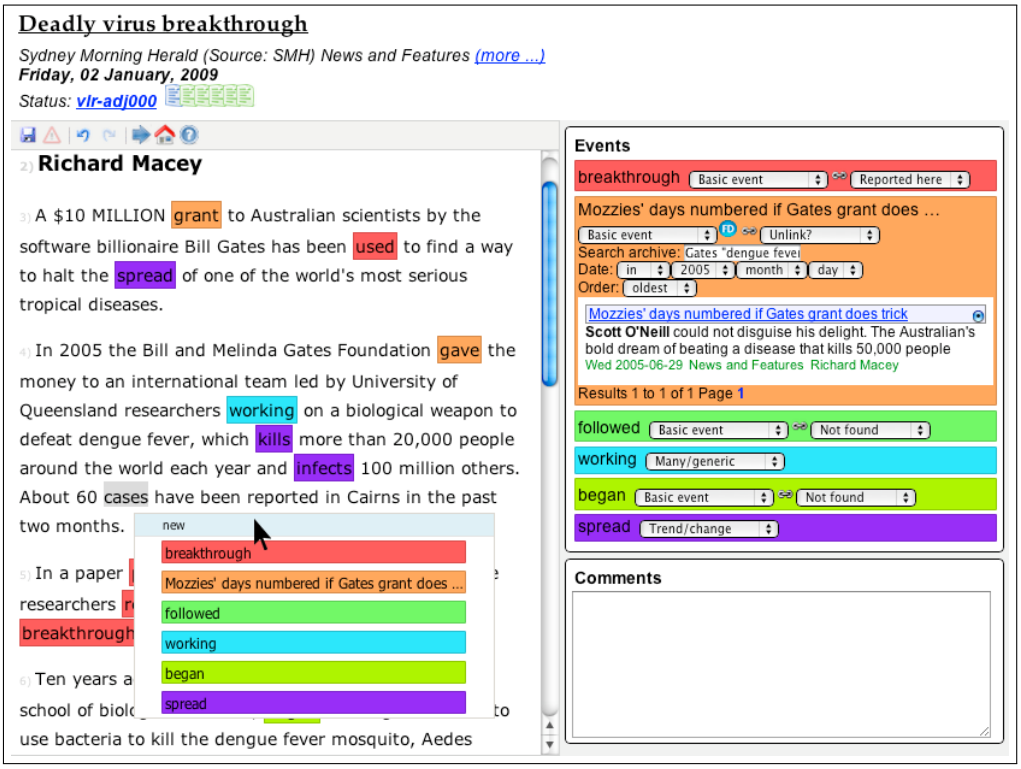


Figure 5.1: The event linking annotation interface showing the worked example in Appendix Section B.3. The window is divided into: the top pane listing document metadata and providing navigation to other documents; the left pane, where the story is read and event references may be marked; and the right pane where referent events are categorised and linked to the archive. In the bottom-right, the annotator may store arbitrary comments on the document. A context menu is shown for marking the word cases, allowing the user to join it to an existing chain of references or to create a new one.

For event linking, the tool provides a drop-down selector for the category of event (such as *linkable* or *multiple*), which conditions the availability of sub-categorisation or link searching tools. After entering search keywords with an optional date constraint, the user is shown the title, snippet with highlighted query matches, publication date and the newspaper section for each link candidate, and may click its title to view the entire story with query terms highlighted.

When saved, the input HTML document is marked up with units and their assignment to chains, while chain attributes are stored in JavaScript object notation (JSON) within the HTML. As well as the annotations themselves, our tool records the retrieval queries used for linking. Staff may view the annotations of any user through the web interface, or may export the unit data as HTML or tokenised IOB format (Tjong Kim Sang and Veenstra, 1999) alongside the JSON chain-level annotations.

The use of a web-based tool allows for distributed annotation with no installation costs, while providing familiar interface components to users without technical experience or knowledge of the underlying format.

5.3 The underlying corpus

We link to a digital archive of the Sydney Morning Herald (SMH), which consists of Australian and international news and commentary published as a daily newspaper, Monday through Saturday. The digital archive contains articles published from 1986 to present.[11] We annotate a randomly sampled corpus from its 2009 *News and Features* and *Business* sections including news reports, op-eds and letters. We exclude some regular columns that rarely refer to news events, such as humour and puzzles.

Over the fourteen years we consider, the archive includes 471k articles, as summarised in Figure 5.2. Figure 5.3 depicts the genre distribution within the 2009 *News and Features* and *Business* sections, according to a rough manual classification of the 150 articles.

Annotating documents from diverse genres introduces complexity to the task: news reports have a relatively formal structure; other genres may introduce less regular forms of event reference. Considering references outside of the news genre is a key aspect of the event linking task, which intends to apply broadly to references to past news events. Using a 24-year archive allows us to annotate references over long spans of publication time.

Using a daily publication rather than newswire as a KB may simplify the task: we need not consider minor updates to reports present in syndicated news feeds, and determining the earliest reporting of an event is only a matter of date comparison. On the other hand, more articles are published simultaneously, meaning that multiple stories discussing the same event need to be weighed as to which is most clearly reporting the event. Working with an archive

[11] The archive may be searched at `http://newsstore.smh.com.au/apps/newsSearch.ac`.

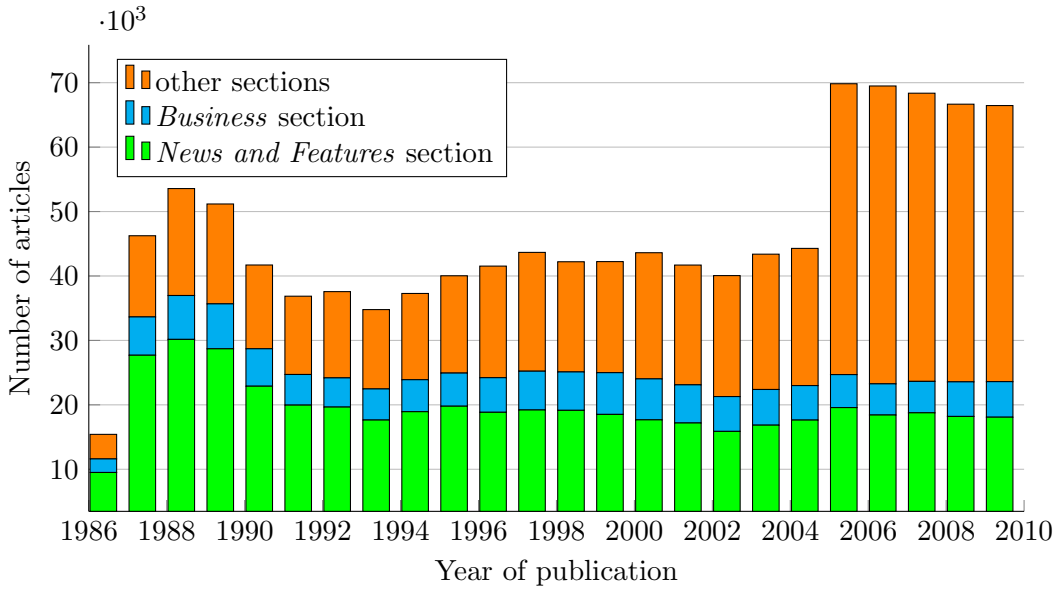


Figure 5.2: Quantity of articles (thousands) in the SMH archive, 1986–2009, highlighting the proportion in *News and Features* (green) and *Business* (blue) sections.

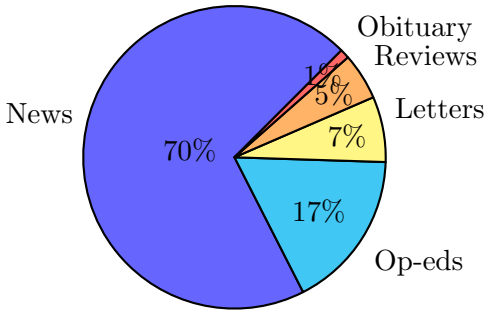


Figure 5.3: Approximate genre distribution in a sample of 150 documents from the 2009 *News and Features* and *Business* sections of the SMH archive. The Op-Eds category includes commentary and editorial columns. News incorporates some instances that have aspects of commentary or review, as well as popular interest stories and gossip columns.

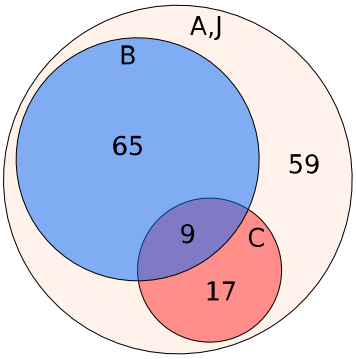


Figure 5.4: A Venn diagram illustrating multiply-annotated portions of our corpus. In total, A and J annotated/adjudicated 150 documents, B annotated 74 and C annotated 26.

Category	Mentions	Types	Docs
Any markable	2136	655	149
linkable	1399	417	144
linked	501	229	99
reported here	667	111	111
nil	231	77	77
compound	220	79	79
multiple	328	102	102
aggregate	189	57	57

Table 5.2: Annotation frequencies: number of mentions, distinct per document, and document frequency

with a continuous publication cycle will present different challenges.

5.4 Inter-annotator agreement and corpus analysis

With annotators completing up to four documents in an hour, we produce a modest corpus of 150 documents.

The corpus size is limited by the hard work inherent in the task, particularly in the process of retrieving and reviewing candidates for an event link, and refining the query until a target is found or the annotator is confident of its absence. Adjudication is also not trivial and requires that source and target articles be carefully considered, if not alternative candidates.

Labelling our primary external annotators A, B and C, and the adjudicator J, we illustrate the multiply-annotated portions of corpus in Figure 5.4. Overall annotation frequencies are shown in Table 5.2. Before reviewing quantitative measures of inter-annotator agreement in Section 5.4.2, we consider some of the challenges that result in disagreement through particular cases. We then assess the adjudicated corpus in terms of the relationship between link source and target in Section 5.4.3.

5.4.1 Disagreement case study

Since several annotators underwent training but did not mark up other articles, we use these redundant annotations over training documents to illustrate variability, with the caveat that some annotators have not sufficiently grasped the task, but that others are likely to be highly attentive to the annotation guidelines.

One training document contains the following background to news regarding Australian Federal Lower House preselection in the electorate of Dickson:

(47) Mr Dutton [won]$_a$ Dickson from Labor's Cheryl Kernot in 2001. Ms Kernot [won]$_b$ the seat for Labor in 1998 after [defecting]$_c$ from the Democrats.

Out of 14 annotations of this text,[12] 13 marked a *linkable* event for a and b,[13] while only 8 marked c, reflecting either its lesser salience due to its subordinate clausal position, or annotators deeming the event not newsworthy.

The linked article for c was identified with highest consistency, with six in agreement on the news report, one finding an analysis article from the same day of publication, and the last selecting *not found*.[14] The two selected articles both mention Kernot's defection, but the canonical reporting is highly evident, opening with similar language to the 2009 article, In a stunning political coup, the Leader of the Australian Democrats, Senator Cheryl Kernot, has defected to the Labor Party...; the analysis article presupposes that fact: Ms Cheryl Kernot's departure from the Democrats is a devastating blow to the party .... In a similar manner for a, apart from one annotator targeting a story four days later,[15] all chose the same day's news, with eight selecting an article directly reporting Dutton's win against Kernot, three targeting an article on changed voting patterns in the broader region, and one targeting an article focusing on Kernot's departure. Thus where multiple stories report perspectives on an event, annotators may disagree on – or not make an effort to ensure – the distinction of the target.

While a and b appear similar on the surface, linking the latter is more ambiguous and prone to technical error. Table 5.3 shows the distribution of annotations among six articles: 3 annotators choose articles prior to the election, another two select targets well after the election, with most annotators are split between an article published shortly before and one shortly after Kernot announces her win.

Unlike Dutton, Kernot is mentioned many times before and after her electoral win. The sheer frequency of Kernot's appearances in 1998 news makes the annotation more technically

[12]This includes six performed with the pilot schema, and eight with the final schema. We do not consider the changes to substantively affect this example.

[13]For unknown reason, despite the textual similarity, the annotator missing a marked b and vice-versa. Neither of these annotators contributed to the final corpus.

[14]This seems to be a spurious error: the annotator in question made a number of searches with the incorrect constraint that the event was reported in 1998. After viewing two other candidates from 1997, the annotator viewed the canonical target before selecting *not found*.

[15]This article lists Dutton among debutant members of parliament.

Freq.	Date	Target headline or event
1	1997-12-06	*Lees rids herself of Kernot's style*
2	1998-09-19	*Where is Kernot? In winner's tent*
	1998-10-03	Polling for election
–	1998-10-05	*Outburst by Kernot 'intemperate'*
–	1998-10-06	*Kernot needs miracle to save political career*
–	1998-10-07	*Kernot says sorry, but she'll quit if she loses*
–	1998-10-09	*Kernot edges towards Parliament*
–	1998-10-10	*Kernot edges further ahead*
–	1998-10-12	*Kernot lags, but pins hope on a recount*
–	1998-10-13	*Women the losers on ALP front bench*
–	1998-10-14	*Kernot almost home, but just where is it?*
–	1998-10-15	*Lib claims muckraking, says Kernot*
4	1998-10-16	*Accusations continue as Kernot firms*
	1998-10-17	Kernot announces electoral victory
3	1998-10-19	*Opponents join to take shine off Kernot's win*
1	1998-10-21	*Playing the diplomat may be Kernot's hardest task in her new mega-job*
1	1998-10-31	*Poll proves high profiles count*
1	Target not found	
1	Event unmarked	

Table 5.3: Linking Kernot won the seat: the distribution of fourteen annotations and the timeline from election to declared victory, with a representative article per day where Kernot is mentioned.

involved: it either requires manually examining many candidates, or limiting them through fine-grained keyword searches or date constraints. The date constraint explicit in the article (in 1998) does not provide a tight bound, but if the annotator appreciates that the target must not precede the Australian federal election, they can use its date – easily found from Wikipedia – as a constraint, but still need to look through many articles between this date and Kernot's victory. Of the three annotators that did not ensure the link target followed the election, two were deceived by the headline, *Where is Kernot? In the winner's tent*, which elides the fact that this is merely a pre-election prediction. Thus a naïve perusal of search results may also not suffice to select the correct target. Alternatively, adding keywords may vastly reduce the number of candidates, but miss the canonical target. Consider, for example, including won as well as Kernot and Dickson in a query: of the four articles annotators chose following the election, all contain the term won except for *Opponents join to take shine off Kernot's win*, which we argue is the correct target; in this particular instance, the problem is also one of insufficient term normalisation on the part of our search engine,[16] where annotators might expect the level of query processing applied in mainstream web search engines that would match win – and perhaps even victory used in the article body – for a query containing won. These technical problems result in part from the random-access nature of the task: news as a genre is designed to be read shortly after publication, and to some extent each day's news supplants the previous; for annotators divorced of that synchronous knowledge-building experience, a lot of work may be necessary to accurately pinpoint a particular story.

Yet the predominance of error must also be accounted for by lack of clarity of the event reference or misunderstandings within the task. The certainty of Kernot's win fluctuates in the two weeks following the election. By the report on the 16th of October, Ms Cheryl Kernot appears to have won the seat of Dickson, with further recounting yesterday, although she does not claim victory until the weekend of the 17th, which is reported on the 19th. The ambiguity between these two articles may be from the semantics of Kernot won: is the winning an automatic result of the poll, or subject to her claiming victory? Is it then subject to the absence of later court rulings invalidating (and then upholding on appeal, etc.) that victory? Had the source article instead used the paraphrase Kernot was elected to the seat, would that change the link target? We are again struck by the effect of lexical choice in identifying a precise referent, and the ability to read many event references with narrow or wide interpretations.[17] Alternatively, the ambiguity may stem from interpreting *the article first reporting that event as having happened/begun* in our schema (appendix Section B.2): the article on the 16th reports that the event in question *appears* to have happened, while the later article is more assertive[18]; the requirement to identify the *first* article reporting the

[16]Since the Solr search engine we employ applies Porter2 stemming, the problem also lies in won's morphological irregularity; had the article used elected, the problem would be a different one.

[17]One might argue that elected, or even won, represents a *compound* event in some uses, and that it is ambiguous in the context of Example 47.

[18]Our annotation schemas do not explicitly prescribe that the author of the targeted article must assert

event often contradicts the desire to find an article where the event's occurrence is asserted as certain. However, rejecting the article of the 16th cannot merely be because a *better* candidate exists on the 19th: had the publication not chosen to report the claimed victory, a consistent schema must still reject the alternative, and prescribe *not found* as the correct annotation. This indicates that our schema needs further explication in order to ensure such consistency.

Annotators (at least at this stage in training) also violate the schema in choosing articles not reporting the event, selecting an article predicting Kernot's win (on 1998-09-19) and later analysis that presumes the event but does not report it (1998-10-21,31). Indeed, this example illustrates the necessity of linking to an article only describing the event as having happened. Several articles report that Kernot will win before she claims victory, yet identifying one as a canonical representative of the event for linking is problematic; notably, many future-tense references to Kernot's win suggest that the event will not happen.

The three event references in the Dickson example illustrate a number of sources of annotator disagreement including the difficulty of identifying non-salient references, misunderstandings and underspecifications of the annotation schema, technical limitations in searching through archival news, and divergent readings of referential semantics. All of these are reflected in the aggregate agreement scores reported below.

5.4.2 Quantitative inter-annotator agreement

For a quantitative evaluation of corpus quality and task difficulty, we seek to determine the extent to which annotators agree with each other and with the final, adjudicated corpus. At the token level, we are guaranteed at most one annotation per annotator; when considering a span of tokens, a sentence, or a document, we may compare the set (or bag) of annotations. Hence we consider a number of inter-annotator decision comparisons:

Binary such as *is token x linked?*

kb node such as *what* KB *entry is token x linked to, given that it is linked?*

Multi-valued such as *what* KB *entries are linked to from document x?*

In all cases we apply F measure (or Dice's coefficient) which is defined as the harmonic mean of precision and recall:

$$P_{AB} = R_{BA} = \frac{|A \cap B|}{|A|} \qquad F_1(A, B) = \frac{2P_{AB}R_{AB}}{P_{AB} + R_{AB}} = \frac{2|A \cap B|}{|A| + |B|}$$

Here, A and B represent two annotations, defined as sets of (u, l) pairs for unit u annotated with label l. Chance-corrected metrics like Cohen's κ do not apply to linking where random

the event's occurrence, in part because events are often reported according to the information of some other source, and in part because of the need to keep the schema brief for low-skill annotators.

Unit	Decision	AB			AC		
		P	R	F	P	R	F
Tokens	Is marked	0.56	0.26	0.35	0.56	0.32	0.40
Tokens	Is *linkable*	0.55	0.22	0.32	0.56	0.26	0.35
Tokens	Is linked	0.56	0.18	0.27	0.48	0.14	0.21
Linked tokens	Target KB node	0.5	0.5	0.5	0.7	0.7	0.7
Linked tokens	Target date	0.6	0.6	0.6	0.8	0.8	0.8
Documents	Target KB nodes	0.5	0.2	0.3	0.5	0.3	0.4
Documents	Target dates	0.6	0.3	0.4	0.5	0.4	0.4

Table 5.4: Inter-annotator agreement over selected units and decisions. For each annotator pair (AB and AC), A's annotations are considered ground truth and the others' are compared to calculate precision (P), recall (R) and F measure (F). Annotator training documents are removed.

chance of selecting a particular target is minuscule; F measure effectively accounts for chance in binary decisions with a vast majority negative class.

Marking tokens Inter-annotator agreement at the token level is poor; Table 5.4 indicates under 40% F measure agreement as to whether a token is *linkable*.[19] Most disagreement lies in the decision to mark a particular token as a newsworthy, past event reference. Some of this reflects individual annotator biases: in their shared portion, annotator A marked 3.7 times as many tokens as B. Hence while B and C recalled fewer than 23% of the tokens marked by A, 56% of the tokens marked by B and C were also marked by A.

The schema underspecifies definitions of 'event' and 'newsworthiness', accounting for some of this token-level disagreement, but not directly affecting the task of linking a specified mention to the archive. For example, an adjectival mention such as Apple's *new* CEO is easy to miss and questionable as an explicit past event reference. Events are also confused with facts and abstract entities, such as bans, plans, reports and laws; unlike many other facts, events can be grounded to a particular time of occurrence. Nominal event references such as graft, e-mail or fire may also ambiguously refer to an event or the theme of that event.[20]

Annotators may also select different tokens for the same event reference, such as in the Black Saturday fires burned or another acquisition. The low per-token agreement is therefore a result of the schema's loose prescriptions and requirement of a single token per reference,

[19] For the moment we ignore adjudication. Full token-level agreement statistics are tabulated in Appendix Section C.1.

[20] Other ambiguous references include impressionistic language such as scandal, tragedy and troubles. In one instance, carrot was found to refer to a government's offer of incentive! Negated event references such as missed out and overlooked also present a problem. The schema asks annotators to focus on explicit event references, but this too could be more explicit.

while highlighting the general difficulty of newsworthy event identification and anchoring.

Categorising event tokens The top section of Table 5.4 also shows agreement decreasing with increasingly fine-grained annotation decisions, such that annotators B and C respectively recall only 18 and 14% of tokens successfully linked by A, with precision around 50%.[21] We provide raw pairwise agreement data for token-level annotation in Appendix C.1.

Considering only tokens marked by both annotators in a pair, the most confused token label is *compound*. For every 10 tokens in which annotator pairs agree on *compound*, there are 28 where they disagree, with one choosing *compound*, and the other usually (25 of 28 times) choosing *linkable*.

Among difficult *compound-linkable* ambiguities are bureaucratic and legal processes, such as large business transactions and changes in law. One example in our corpus states that the Carr government loosened restrictions .... The government's loosening initially consists of their presenting a bill to parliament, but is not concluded until two houses of parliament vote in its favour, and the bill receives vice-regal approbation (which, as a formality, usually goes unnoticed in news). Generally there would be a further delay before the loosening comes into effect. So the referent event space is technically *compound*. Yet given another similar reference, a parliamentary victory might be the unambiguous referent.

There is also frequent ambiguity among *compound*, *multiple* and *aggregate*, suggesting that these are not natural delineations of event reference. For example, does food riots in 30 countries [over a short period] constitute reference to a single event reported through its sub-events, a collection of distinct events, or an emergent aggregate? In an earlier schema (in Appendix B.1) these categories were conflated as *plural*, which is too broad for annotators to work with. These delineations were therefore intended to help annotators decide what is not *linkable*, but they do not affect the present event linking task.

Identifying link targets To assess how often annotators agree on a canonical link target, we firstly consider only those tokens a pair of annotators both successfully linked. For these units, there is reasonable (31 out of 64) agreement between A and B and high but statistically weak (11 out of 15) agreement between A and C. Since annotators may mark different tokens for the same event reference, or may mark different numbers of references with the same link target, we also compare the set of distinct link targets identified within each document. Annotators B and C respectively recall 22 and 34% of A's link targets with about 50% precision. Overall, this level of agreement suggests the feasibility of the event linking task, when ignoring complexities introduced by exhaustive and archive-internal linking.

In some cases, there may be multiple articles published on the same day that describe the event in question from different angles; Table 5.4 shows agreement increase substantially

[21] Equivalently, A recalls 50% of the tokens linked by B and C.

Unit	Decision	JA			JB			JC		
		P	R	F	P	R	F	P	R	F
Tokens	Is marked	0.58	0.76	0.66	0.77	0.44	0.56	0.76	0.60	0.67
Tokens	Is *linkable*	0.54	0.73	0.62	0.69	0.39	0.50	0.72	0.46	0.57
Tokens	Is linked	0.55	0.69	0.61	0.70	0.30	0.42	0.61	0.24	0.34
Linked tokens	Target KB node	0.84	0.84	0.84	0.8	0.8	0.8	0.7	0.7	0.7
Linked tokens	Target date	0.87	0.87	0.87	0.9	0.9	0.9	0.9	0.9	0.9
Documents	Target KB nodes	0.68	0.69	0.69	0.8	0.4	0.5	0.6	0.4	0.4
Documents	Target dates	0.72	0.71	0.71	0.8	0.4	0.5	0.7	0.5	0.6

Table 5.5: Adjudicator-annotator agreement over selected units and decisions, as per Table 5.4. The adjudicator J's annotations are considered ground truth for calculation of P, R and F. From J's perspective, P and R are rates of *acceptance* and *contribution*.

when relaxed to accept date agreement. Where a definitive link target is not available, an annotator may erroneously select another candidate: an opinion article describing the event, an article where the event is mentioned as background, or an article anticipating the event. One annotator linked the reference the survivors were flown to an article where the survivors were to be flown, which implies the event in question is uncertain and either imminent or happening at present.

Determining whether a particular archival story reports an event is difficult, as suggested by high confusion between *not found* and *reported here* annotations. For every 10 tokens where annotators agree on *not found*, there are 10 cases of *not found-reported here* confusion, and 6 cases of *not found-linked* confusion. Some confusion results from cases where the SMH only belatedly reports an event, either because it was not sufficiently newsworthy at the time, or because the event's occurrence only later became public knowledge. The disagreements otherwise indicate a lack of clear discourse features for annotators to discern whether a story reports or merely mentions an event.

The task is complicated by changed perspective between an event's first report and its later reference. Can overpayed link to what had been acquired? Can 10 died be linked to a story s where only nine are confirmed dead? For this example, if the tenth death occurred in the same event as the first nine (unbeknownst to the reporter of s), its mention is strictly a reference to the event reported in s. If instead the 10th death occurred as a result of the same event as the first nine, its mention may be better considered an *aggregate*. For the application of adding hyperlinks to news, such a link might still be beneficial; such are the challenges in determining appropriate event link targets.

Adjudication Agreement statistics for adjudication are shown in Table 5.5. For all annotation decisions listed, each annotator achieves high precision against the adjudicator; that is, the adjudicator accepts most input annotations. Since annotators rarely agreed among themselves, this shows that each annotator is likely to miss event references in the exhaustive linking task.

The lowest precision (54%) is reported in annotator A's marking of tokens, suggesting A over-generates annotations or chooses different anchor tokens to J. A's thoroughness is apparent in her recall (or high contribution) of link targets per document, which is substantially higher than B and C.

In all, we find that the primary disagreements in the annotation task regard whether to mark a particular token and whether it can be linked. We have seen similar recall problems in other fine-grained event reference annotation; linking requires a further sustained effort to examine candidates and refine queries, such that some annotators make the effort to identify many more links than others. An exhaustive annotation therefore requires redundant annotations to be merged. Regarding the selection of link target, there is relatively little dispute, suggesting that the event linking task is feasible. Yet agreement statistics also suggest identifying a link target is not trivial, a result which is supported by further analysis of the resulting corpus; in particular, the relationship between link source and target.

5.4.3 Corpus analysis

Where inter-annotator agreement measures the propensity of humans to the task, we must also consider features of the task that make it feasible, while not trivial, for a computerised event linking system. Topic Detection and Tracking's coreference is generally achieved by measures of term overlap together with temporal proximity (Allan et al., 1998). We therefore examine the relationship between the source and target document of each event link, considering:

Textual similarity To what extent does event reference relate to repeated text, or copy the language of its referent representative? We represent each document as a vector of its body text words, weighted using the classical tf.idf formula, and calculate the cosine similarity of source and target document vectors.[22]

Publication date difference How is event reference affected by recency of the event's being first reported? We calculate the number of days between the target and source's publication.

Figure 5.5 scatters links with respect to these axes, labelling points with the approximate genre classification of the source document. At the same time, we would like to suggest

[22]Stop words and punctuation are removed according to Solr's default text analyser, and Porter2 stemming is applied to reduce dimensionality. We calculate idf with respect to the SMH archive.

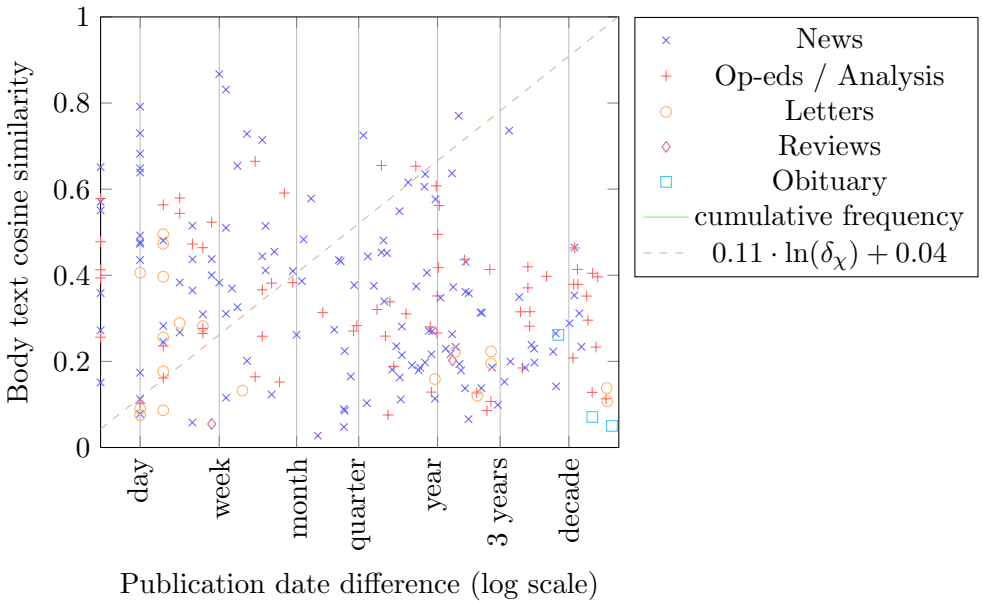


Figure 5.5: Scatter plot of textual similarity (tf.idf-cosine) against publication date difference between source and target documents for each link in our corpus. Link source documents are distinguished according to their approximate genre classification (see Figure 5.3). For each date difference δ_χ, the overlaid green curve indicates the proportion of source-link document pairs with date difference $\leq \delta_\chi$; its line of best fit is shown dashed.

that the task is not only feasible and non-trivial, but interesting, particularly in the way it coindexes references to different aspects of the same event among co-reported events.

Source-target textual similarity The average cosine similarity between source and target document is 0.34 with sample standard deviation 0.18: most document pairs do not have very high overlap. The highest similarity score is 0.88, where an article reporting the second hearing in a court case refers to the lawsuit and first hearing, a week before. While there is some overlap in rewritten content, the outlying similarity score is largely due to the very high tf.idf of the plaintiff's name. In a number of cases, high overlap corresponds with news articles referring to events reported the day before, and high overlap generally only occurs where the link source is a news report. The large proportion of low document similarities suggest that many event link targets are unlikely to be found with trivial bag-of-words retrieval solutions.

Source-target publication date difference Links mostly point to recent articles. Our data shows an additional peak for links to articles around a year prior, which in at least one instance corresponds to an annual event.

We approximate the empirical cumulative distribution of date differences shown in Figure 5.5 by a straight line. This indicates that a link target's likelihood is roughly inversely

Link anchors	Link target title
attack, attacked, detonated, occupied, opened fire, siege	*Terrorists lay siege to police academy*
arrested, assault, caught, charges, driving, failed, resisting	*Lawyer on police assault charges*
blow, incident, inflicting, smashed, strike, suffered	*Olympian faces sack after bar punch-up*
beat, final, winning, win	*At long last, the Dogs have their day*
appointment, dumped, elevation	*The 'puppet' Premier*
ceremony, hitched, married	*Battle of the Nile continues*
bought, overpaying	*Rio bids $44b to win Alcan*

Table 5.6: Examples of diverse reference tokens sharing a link target.

proportional to the time elapsed before the link source. This accords with an intuition expressed in prior work in topic detection, such as Yang et al.'s (2009) assumption that news events generally build on others that recently precede them. However, the distribution is surprisingly heavy-tailed: the median date difference is 139 days, such that half of all link targets precede their sources by 20 weeks or more. Links within obituaries in our corpus exclusively have targets over 8 years before the subject's death. Hence, in contrast with the assumptions of topic detection systems, exhaustively seeking past event references shows that many older events are mentioned, especially in biographical, opinion and analysis articles.

Reference anchor text and co-reporting We hypothesise that co-reporting (see Section 4.1.3) is a useful approximation to coreference. This might be identified from the set of references linked to a single target. Table 5.6 samples some link targets where the set of link anchor texts illustrate diverse events being co-reported. So while win refers to a strict sub-event of final, the two share a canonical link. We think this reflects a news reader's intuitive granularity of event better than strict coreference.

5.5 Discussion

Our annotated corpus goes some way to validating the event linking task. There are however many alternative approaches to annotating event link instances. We therefore discuss some of the decisions we have made in designing the annotation task, and future possibilities.

In particular, the number of decisions inherent in the present annotation task makes it difficult to produce a large, high-agreement corpus. This in turn makes the acquired annotations an infeasible target for a multi-site technology evaluation. While out of scope in this work, an alternative approach that limits annotation decisions may lead to a more robust corpus.

Chronological annotation Event linking is designed after the assumption that a frequent reader of news sees an event referred to over time. By randomly sampling documents in our

task, an annotator is left to understand reports in isolation, to intuit whether a mentioned event is newsworthy and how it might have been reported. Searching for and reviewing candidates in the present approach is an arduous way to learn an event's context and identify its target.

Annotating every document published by a newspaper over a substantial period of time is infeasible, so a chronological annotation requires considering one topical cluster at a time. We therefore attempted the annotation of stories from a TDT topic cluster, selecting a single news source. Documents were exhaustively annotated in chronological order, with links only allowed to previously-annotated documents.

The resulting annotations had some similar characteristics to our corpus, such as evidence of co-reporting. However, faced with an event reference, especially of an event peripheral to the selected topic, the annotator could not know whether the canonical link target lay in the news archive but outside the cluster. Furthermore, remembering all the event content within a large topic cluster and where it is first reported is challenging, especially when it is read in immediate sequence rather than spread out over time. Hence this too was a poor approximation of our idealised reader; and due to the construction of the clusters, the links could not span a long time. Future work might improve the chronological annotation task to avoid these problems.

Links from outside the archive By annotating only documents within the link-target archive, we skew the event-referential language: a particular news-source has both overt and subconscious regional, political and stylistic biases. Annotating documents outside the target archive may also avoid the difficulty of identifying *reported here* references.

Exclusion of difficult genres We opt to only annotate documents from the *News and Features* and *Business* sections of the corpus. We especially exclude the *Sports* section of the corpus, having found that the sports reporting and current affairs genres differ substantially. Excluding *Sports* hence allows annotators to focus on the news genre, but its also has features that make event linking difficult: sports news frequently conflates new event reporting, speculation, gossip and commentary, making them troublesome link targets; in move-by-move descriptions of matches, as well as gossip, determining the newsworthiness of event references may be difficult; and in many cases, no article directly reports an event (e.g. a tennis match) having happened, as interested readers are assumed to be aware of the event from more immediate sources, e.g. having watched a game on television.

Separate unitising and linking Most inter-annotator disagreement in our task stems from identifying and categorising event references, although event linking as defined in Section 4.1 assumes a reference is given (as opposed to exhaustive linking). We could consider

TimeML event references		All references			*linkable* references		
Class	Tense	P	R	F	P	R	F
Any	Any	0.16	0.63	0.25	0.11	0.69	0.20
In top three	Any	0.16	0.63	0.26	0.12	0.68	0.20
Occurrence	Any	0.17	0.50	0.25	0.11	0.51	0.19
Any	Non-future	0.16	0.63	0.26	0.12	0.69	0.20
Any	*Past*, *Present*, *None*	0.19	0.55	0.28	0.14	0.61	0.23
In top three	Non-future	0.17	0.62	0.26	0.12	0.68	0.20
In top three	*Past*, *Present*, *None*	0.19	0.54	0.29	0.14	0.60	0.23

Table 5.7: Predicted TimeML event references in comparison to those marked in our corpus. We show precision (P), recall (R) and F measure (F) of event tokens marked by the system described in Bethard and Martin (2006), filtering by class and tense. The top three classes in descending F measure are *Occurrence*, *I_Action* and *Reporting*. The top three tenses are *Past*, *Present* and *None*, while other non-future tenses are *PastPart*, *PresPart*, *Infinitive*.

annotating the entire corpus in two stages: (a) identifying and marking *linkable* event reference; and then (b) attempting to find a canonical link target for each item marked in (a). However, we expect that non-expert annotators may only gain an intuition of newsworthiness within the particular archive by linking to it.

Another possibility is to automatically mark event references and have annotators select from among them. A state-of-the-art TimeBank-trained event detector (Bethard and Martin, 2006) recovers almost 70% of our *linkable* mentions,[23] but with precision of only 11%. We also consider the most predictive TimeML classes and tenses on our dataset, but this only increases F measure marginally. In pilot annotations where documents were marked with such predicted event references for annotators to select from, we found this vast over-generation distracting, while some clearly *linkable* references were missed.

In a similar vein, a simpler, high-agreement annotation could result from requiring annotators only to validate (or perhaps amend) existing links as valid event links, rather than producing them from a clean slate. A prior corpus of links could be produced either from the predictions of a system built on the basis of this exhaustive annotation, or by acquiring a corpus of appropriate hyperlinks or citations, as otherwise exploited in Section 7.1. Each of these approaches would introduce their own biases to the derived annotated corpus, but in producing a large corpus of true event links with relative inexpense, may be the best way to further develop the task.

[23]TimeBank does not mark so-called generic events, some of which are labelled *multiple* in our schema. We have not specified words to mark identically or as precisely as the TimeML specifications. When we consider the number of annotated *linkable* references with nearby TimeML event predictions, recall increases by 3% and 10% (absolutely) when allowing windows of 1 and 2 tokens (respectively) on either side of the marked token.

5.5.1 Conclusion

We have presented an approach to event linking annotation: sampling documents from the target news archive, and exhaustively annotating their past, newsworthy event references. This introduces additional complications atop the event linking task, mostly associated with recognising references to past events, determining which past event references are newsworthy, and which of those are *linkable*, being where the referent is reported in a single article.

Outsourcing annotation work to non-experts from an online freelancer marketplace resulted in substantial effort spent on training annotators who did not contribute to the final corpus. The annotation task is onerous as it involves searching for and considering numerous candidate articles removed from their temporal context; even adjudication decisions can take substantial effort, limiting the size of our corpus. With adjudication, we produce a corpus of 150 documents, containing 330 distinct event links – or 229 excluding references to events reported in the source document – to targets within a within a 24-year SMH archive.

While annotators tend to recall complementary event references whose union needs to be taken, we find moderate to high agreement – particularly between the adjudicator and each first-pass annotator – on the identification of a canonical event link target. Future exhaustive annotation attempts might benefit from separating the task of unitising (identifying mentions to link) from the linking itself. The cost of producing annotated data also motivates us to seek other sources of event linking knowledge, as presented below in Section 7.1.

While one might presume that similar documents refer to similar events, and that documents in a news corpus usually refer to recent events, we find that that there is a long tail of event links where the target document is not very similar in text to the link source, nor very recent. This highlights event linking's focus on explicit reference, rather than broad notions of topicality, in understanding the use of textual reference to news events. These properties, together with the diversity of referential language for a given link target, present distinct challenges for building event linking systems. We use the new corpus as the primary initial benchmark for the event linking task, reporting in Chapter 7 the effectiveness of a system described in Chapter 6.

Chapter 6

A retrieval approach to event linking

In order to assess the feasibility of the event linking task, we design and evaluate an initial event linking system. It frames the task as information retrieval (IR) in which each query is derived from an event reference and has at most one relevant target document.

We anticipate that the task ultimately involves deep linguistic processing to disambiguate and identify the correct target, since our manual annotation often involved reviewing multiple candidates and reference materials before identifying the link target. Nonetheless, this fine disambiguation is presumed to only be tractable in a system solution where the number of candidates to consider is small. As such, for this initial investigation, we set a bag-of-words model as our baseline, comparing words surrounding the reference anchor with those in the target document, and focus on improved retrieval strategies.

We have also noted the importance of temporal information in event link identification. Applying the date constraints used by annotators recalls a further 20% of targets by rank 50 than when querying with annotators' selected keywords alone (Nothman et al., 2012). This motivates an emphasis on using the temporal structure of the target archive, and temporal knowledge in the source reference.

This chapter proposes scoring candidates according to a combination of source-candidate similarity scores accounting for various content, discourse and temporal features in a vector space model, detailed in Section 6.2. This is moderated by a prior weight over candidate publication dates given a query as described in Section 6.3. Both of these weights are dependent on parameters which distinguish the contribution of various features. Section 6.4 describes a method for learning these empirically from known event links. Initially, we motivate and overview the system.

6.1 System overview

As a retrieval task, event linking differs from ad-hoc retrieval in a number of ways:

Query structure Ad-hoc retrieval generally considers a user-generated query with little context. Input to event linking is a span of text corresponding to an event reference, within a document context. Hence query terms must be inferred from the context, and may incorporate various linguistic structures, rather than mere phrases.

Target structure and single purpose Event linking's retrieval candidates are all news articles. Both the temporal aspects of these articles and their common discourse structure may therefore be exploited. Event linking queries also have a single purpose, such that the retrieval index may store specialised fields related to events.

Zero or one correct response There may be many documents relevant to a retrieval query, but for a linking query, exactly zero or one entry is the correct response. To achieve this, candidates must be evaluated against each other, as well as against the hypothesis that no correct candidate may be found.

In these features, event linking resembles named entity linking and question answering, to different extents. Such tasks tend to be solved by broadly tripartite systems (Harabagiu et al., 2003; Ji and Grishman, 2011):

(Re)formulate query Given a reference to link (or similar), this component produces a query or queries for an IR engine.

Generate candidates This component identifies (and often scores) a limited set of top candidates matching the query within the target KB.

Disambiguate From a set of candidates, this component selects one or none ($\varnothing$). Where multiple answers make sense for a task, it ranks or scores candidates as well as $\varnothing$.

Disambiguation may be a complex process, potentially involving structural matching to identify compatibilities and mismatches between the query and each candidate, pairwise comparison of candidates, or the incorporation of broader contextual knowledge. With effective indexing and scoring of relevant content, the candidate generation step is designed to efficiently identify a small set of likely candidates, allowing disambiguation to be much more sophisticated and thus computationally intensive.

Given the enormity of news archives targeted by event linking and the potential for simple overlap measures to produce many spurious candidates, we focus in this work on the problem of candidate generation. This process centers on a function σ that scores each candidate c given a reference r, such that the chosen event link target for r is:

$$\Lambda(r) = \begin{cases} \arg\max_c \ \sigma(c, r) & \text{if } \max_c \ \sigma(c, r) \geq \theta_{\varnothing}(c) \\ \varnothing & \text{otherwise} \end{cases}$$

In the present work we ignore the case of tuning the nil threshold, and set $\theta_\varnothing$ to 0. We construct the score function as the product of a term overlap score and a temporal weight (both strictly non-negative), each of which depends on parameters $\mathbf{w}$:[1]

$$\sigma(c, r; \mathbf{w}_{\text{terms}}, \mathbf{w}_{\text{time}}) = \sigma_{\text{terms}}(c, r; \mathbf{w}_{\text{terms}}) \cdot \sigma_{\text{time}}(c, r; \mathbf{w}_{\text{time}})$$

Term overlap assigns high scores to documents whose content is close to an extracted vector of weighted query terms, using cosine similarity to measure their overlap as detailed below. We calculate term overlap features that combine with $\mathbf{w}_{\text{terms}}$ to produce σ_{terms}. Each term overlap feature is determined by a coordinate of the following axes:

Target zones Information retrieval strategies may benefit from assigning different weight to content matched in differing portions of the target document (Manning et al., 2008). Such zoning may emphasise sections of the document likely to contain newly reported events.

Term extraction Traditional IR employs a bag of words for document comparison. Since the identity of an event tends to focus on its location, time and participants, it may be useful to distinguish special terms such as named entities.

Term weighting The vector space model usually weights terms in proportion to their frequency in the target document (term frequency) and in inverse relation to the term's frequency across the targeted collection (document frequency). We introduce alternative document frequency weightings in an attempt to reward the first reporting of an event.

Query formulation Given an anchored reference to an event, relevant search terms are likely to appear nearby in surface text. Yet it may be appropriate to weight terms differently according to their surface or syntactic distance from the anchor, or by generating and weighting related terms.

This produces a vector of overlap scores – one score for each combination of zone, term type, term weighting and query formulation settings – which are then combined to produce a term overlap score for each candidate.

Our term overlap features already incorporate a focus on time by indexing the dates mentioned in a candidate and in weighting terms to emphasise those that are infrequent until an article's publication. Temporal weighting accounts for two further assumptions:

1. An event that is recent to the reference is likely to be more salient and thus more referred to than events further in the past.

[1]The relative effect of these two terms could conceivably be altered by raising one to a power, but in the present work we take their raw product.

2. An event is likely to be reported shortly after the time that it occurs, which may be explicitly mentioned in the context of a reference to it.

We therefore introduce features that decay in their distance from periods when the link target is likely published. A linear combination of these features penalises the term overlap score multiplicatively. This ensures that high-scoring candidates match both the content of our query and our expectations of the target timestamp.

6.2 Scoring candidates by term overlap

A term overlap score is determined as the linear combination of similarity scores under different term weighting methods. These differ with regards to the zone matched within the target document, the selection of query text to represent the reference, the type of term to compare, and a scheme for weighting different terms according to their frequency in the target archive.

More precisely, a term vector $\mathbf{v}_{z,e,\omega}(c)$ for a candidate c is constructed as follows:

1. A zone extractor z excerpts portions of the text of c or auxiliary text.
2. Within this excerpt, a term extraction function e identifies a bag of terms.
3. A term weighting function ω weights each term according to its occurrence in the archive and the temporal position of c within it.

A query term vector $\mathbf{u}_{q,e}(r)$ is similarly constructed given an event reference r to be linked:

1. A query formulation function q actualises the reference as fragments of text, weighted according to their location in a source document or auxiliary text.
2. Within these fragments, a term extraction function e identifies a bag of terms, such as stemmed words or mentioned named entities.
3. The weights determined by a term's location are then penalised according to the term's smoothed collection-wide inverse document frequency.[2]

Thus, given z, e, ω, q, we calculate the term overlap for candidate c given reference r using cosine similarity:

$$f_{z,e,\omega,q}(c, r) = \frac{\mathbf{u}_{q,e}(r) \cdot \mathbf{v}_{z,e,\omega}(c)}{\|\mathbf{u}_{q,e}(r)\| \|\mathbf{v}_{z,e,\omega}(c)\|}$$

This score is bounded within $[-1, 1]$ and is non-negative as long as the term weights are non-negative. Different combinations of z, e, ω, q yield different overlap measures between r

[2] We use $\log \frac{|D|+1}{|\{d \in D : w \in d\}|+1}$ given a set of documents D and a term w.

and c, which we combine linearly with weights $\mathbf{w}$, discarding any negative scores:

$$\sigma_{\text{terms}}(c, r; \mathbf{w}) = \max\left\{0, \sum_{z,e,\omega,q} f_{z,e,\omega,q}(c, r) \times w_{z,e,\omega,q}\right\}$$
$$=: \max\left\{0, \mathbf{f}_{\text{terms}}(c, r)\mathbf{w}^\top\right\}$$

We proceed to describe the values of z, e, ω and q that may highlight aspects of the query and candidate that are salient for the event linking task.

6.2.1 Zoning to highlight reported content

A news story may be the canonical event link target only where it reports the event. Naïve document retrieval does not distinguish between content that is newly reported, and that which constitutes background or other detail, except insofar as inverse document frequency down-weights material that is frequently mentioned. In IR, a document may be partitioned into *zones* that are weighted differently to reflect differences in their expected salience. Similarly, a NEL system may score a candidate better for the query name appearing as the title of the candidate's Wikipedia page, than matching the anchor text of a link to that page.

The distinction between reported and background content in a story is illustrated in Figure 6.1. In the news genre, the novelty of content is indicated implicitly by typical discourse characteristics, such as new information generally preceding background in the narrative, if not explicit indications that an event recently took place. We heuristically extract a number of zones that may focus on newly reported content, with examples marked in Figure 6.1:

Full text While we hope to capture particularly salient content through more specific zoning, some query terms will match other sections of the body, which may also provide a better weighting of term frequency.

Headline As in Figure 6.1 the headline of a news article may summarise its new content succinctly. As well as summary, however, headlines serve to entice readers, and may incorporate literary devices such as punning that reduce their lexical similarity to ordinary event reference.

Opening body portions Stereotypical news reporting is characterised by an "inverted pyramid" style in which the most central details of the news are described early in the article, with expansion as the story progresses. We therefore consider zones of the first sentence, and first three paragraphs of a story.

Sentences referring to dates In general, the news concerns recent events, and when the time or date of the event's occurrence is explicitly stated, as in Figure 6.1, it tends to be close to the date of publication. One zone therefore consists of sentences with

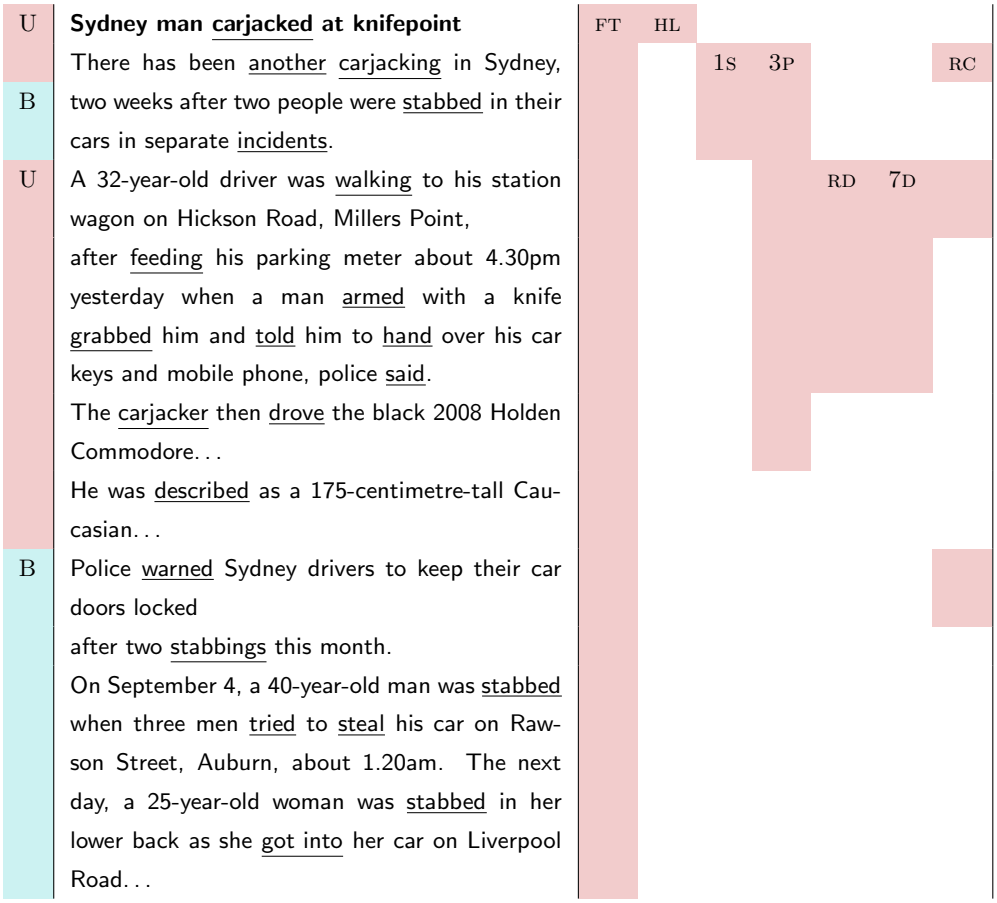


Figure 6.1: Possible event references marked in an article excerpted from smh.com.au (21 Sept. 2011), segmented into update (U) and background (B) event portions. The zones we heuristically extract are marked to the right: full text (FT), headline (HL), first sentence (1S), three paragraphs (3P), most recent date (RD), date within a week of publication (7D) and relativising clause (RC).

Term type	Examples
Stem	commodor; inform; carjack; steal; road; men
NE mention	millers point; holden commodore; auburn; nsw
Location mention	millers point; sydney
Miscellaneous mention	holden commodore; commodore
Entity	*Millers Point, New South Wales*; *Sydney*
Location entity	*Auburn, New South Wales*; *Sydney*
Miscellaneous entity	*Holden Commodore*
Day	*2011-9-20*; *2011-9-4*
Month	*2011-9*
Year	*2011*

Table 6.1: Examples of terms by type extracted from the story in Figure 6.1.

the most recent date reference,[3] but we note that in many cases the date of the event is not explicitly stated, as in the hijacking-release story, where the most recent time expression is to a background event. Another zone consists of sentences referring to dates within the d days prior to publication.

Clauses with relativised background Particular expressions are used to position some content as foreground and other as background. These often imply temporal as well as focal ordering, as with *after* in the first sentence of Figure 6.1: There has been another carjacking in Sydney is news, while [after] two people were stabbed in their cars in separate incidents. is background. We therefore consider a zone consisting of portions of a sentence that precede one of the words *after*, *since* and *following* in lowercase.

6.2.2 Term extraction

While traditional IR treats documents as a bag of words – stemmed tokens excluding very frequent vocabulary and punctuation – in the context of event reference, it seems necessary to specifically assign weight to matched participating entities and spatio-temporal information (the event's *who*, *when* and *where*), as well as semantic content (*what* and *how*) represented in broader vocabulary. We therefore compare term vectors representing the frequency of the following (see examples in Table 6.1):

Word stems the set of stemmed alphanumeric tokens, excluding stop-words.[4] Porter2 stemming is applied to match terms independent of morphological variation, and to reduce the vector dimensionality.

[3] As recognised and normalised by Heideltime (Strötgen and Gertz, 2013). We use the version available for public download as at late 2011.

[4] We apply the English stop-list from Apache Solr and exclude words shorter than 3 characters.

Named entity mentions the set of strings mentioning people, organisations, locations or miscellaneous entities by name. These are identified by the C&C named entity tagger (Curran and Clark, 2003) trained on Sydney Morning Herald news reports with CoNLL-03-style (Tjong Kim Sang and De Meulder, 2003) annotations. We construct a term vector for all named entity strings, and another grouped by type.

Named entities the set of referent concepts identified in Wikipedia for those mentions, where a match is found by the Radford et al. (2012) NEL system. We construct a term vector for all named entities, and another grouped by CoNLL-03 type.

Dates the set of referent dates, truncated to their year, month and day in separate term vectors (where specified). Temporal references are recognised and normalised by HeidelTime (Strötgen and Gertz, 2013).

This set of terms performs little normalisation or interpretation of the language, limiting the approach to source-target pairs with a relatively high degree of word overlap. More nuanced terms within this framework might include disambiguated word senses, syntactic or semantic dependency arcs, or entities related to those that are directly mentioned to account for spatial containment (see e.g. Roberts et al., 2012) and comparable relations among organisations. Additional features might also pursue a less discrete approach to matches within space, time and activity, as Roberts and Harabagiu (2011) applies to first story detection. We leave these considerations to future work.

6.2.3 Temporal term weighting

The vector space model of IR generally weights a term in proportion to its frequency of occurrence in the document at hand, and in inverse relation to the term's overall frequency in the corpus, a weighting scheme known as tf.idf. With some variation (Salton and Buckley, 1988; Manning et al., 2008), a term w within a document d and a corpus D is weighted according to the product of these measures:

$$\mathrm{tf}(w, d) = \text{frequency of } w \text{ in } d$$
$$\mathrm{idf}(w, D) = \log \frac{|D|}{|\{d \in D : w \in d\}|}$$

The inverse document frequency (idf) is designed to give weight to terms that are infrequent across a corpus, and hence likely to be salient in a match.

The temporal nature of event linking leads to us consider variants of df that account for terms being infrequent *during a particular time period*. Partitioning the corpus by date of publication, such that $D = \bigcup_{t=t_-}^{t_+} D_t$ for dates t from t_- to t_+, we propose an alternative inverse date frequency to substitute for idf being the number of days in a window of Δt days

up to date t in which the term w appears:

$$\mathrm{idf}_{\Delta t}(w, t, D) = \log \frac{\min\{\Delta t,\ t - t_- + 1\}}{|\{D_{t'} : \exists\, d \in D_t \text{ s.t. } w \in d;\ t - \Delta t < t' \leq t\}|}$$

Where there is insufficient history of publication before date t, this weight becomes uninformative as the numerator is small and the denominator is allowed little variance; notably, the first appearance of a term gets zero weight. Other work incorporates document frequency knowledge from an additional corpus to get around this problem (Yang et al., 1998); in the present work, we do not handle it. In the case where $\Delta t = \infty$, this variant is similar to a cumulative or online inverse document frequency at time of publication, yet counting days rather than documents. For small Δt, the term's frequency is calculated over a sliding window of days. We count days rather than documents to reduce the effect of topical spikes in multiple documents published on one day. We experiment with both these settings for Δt, in addition to static tf.idf.

In streaming news processing, such as Online First Story Detection, use of tf.idf may implicitly involve updating the document frequency weightings as news documents are available. idf_∞ is an *incremental* idf (Yang et al., 1998), although in some work (e.g. Brants et al., 2003), when comparing a new term vector to previous documents, the weights for the previous documents are updated given the incremented document frequencies. While idf_∞ is similar, we compare queries to weights based on each term's frequency at the time of the candidate's publication, emphasising the novel content of the report in context, and allowing a day's documents to be indexed with static weights. Another temporal variant of idf is suggested by Leibscher (2004), who partitions a corpus into equal-length non-overlapping periods of time in order to analyse lexical change and improve document categorisation. Like $\mathrm{idf}_{\Delta t<\infty}$, this accounts for periodic variation in lexical use, but ignores the position of a document within the windowed period. Kleinberg's method (2002) for bursty term identification builds a state model on the basis of fluctuation in the periods *between* successive occurrences of a term within a time series. Burst information may also benefit our task; using gap length may be more robust to variation between terms in contrast with calculating $\mathrm{idf}_{\Delta t}$ over a finite history. Better exploitation of such temporal fluctuation is an avenue for future work.

Table 6.2 shows that there is a low correlation between traditional $\mathrm{idf}_{\mathrm{static}}$ and $\mathrm{idf}_{\mathrm{inc}}$ ($= \mathrm{idf}_\infty$) weightings. This is despite $\mathrm{idf}_{\mathrm{inc}}$ converging to i·day·f – the log-inverse proportion of days on which a term appears – towards the end of the indexed archive, while the latter is highly correlated with $\mathrm{idf}_{\mathrm{static}}$ ($\rho > .9$). $\mathrm{idf}_{\mathrm{inc}}$ is somewhat correlated with the sliding window variants, which are similar to each other for small Δt.

In Figure 6.2 we plot normalised values of a single term vector under the different weighting schemes – that of word stems from the document in Figure 6.1. We find that the prominent term carjack (tf = 3) is downplayed by $\mathrm{idf}_{\mathrm{static}}$ since it occurs 417 times in the corpus, although only on 7 of the 30 days up to and including the article's publication date. In general,

Weight	$\text{idf}_{\text{static}}$	i·day·f	idf_{inc}	idf_{90}	idf_{30}
i·day·f	0.970				
idf_{inc}	0.192	0.186			
idf_{90}	0.039	0.038	0.670		
idf_{30}	0.039	0.038	0.653	0.977	
idf_{10}	0.038	0.037	0.627	0.928	0.956

Table 6.2: Pearson correlations between weights in term vectors under different document frequency-like schemes, calculated over 168M term-document pairs in an archive of indexed Fairfax Digital stories from 1999 to early 2013.

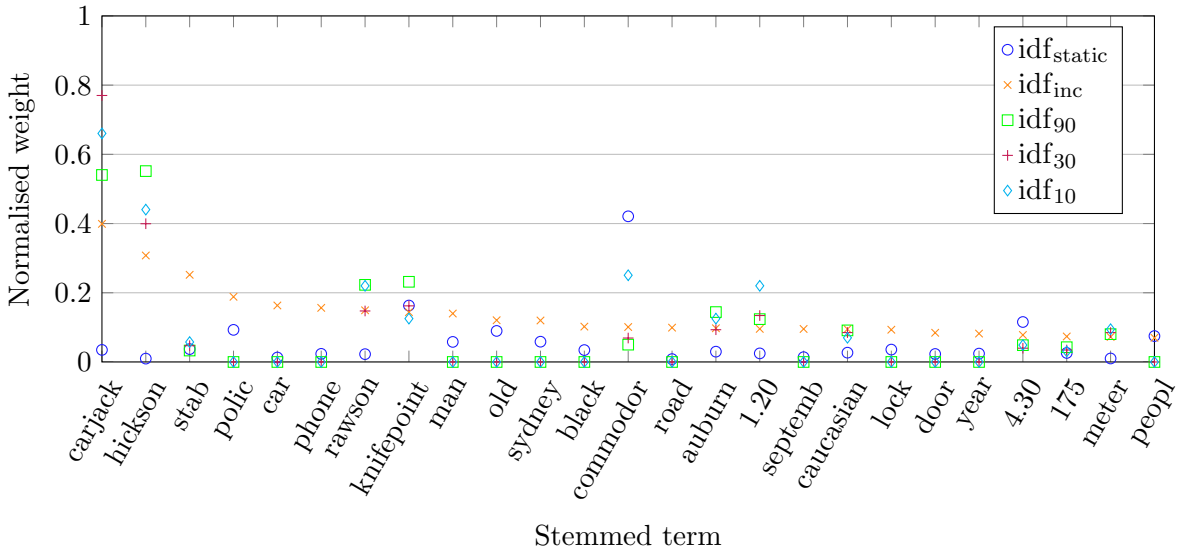


Figure 6.2: The 25 word stems with the highest $\text{tf.idf}_{\text{inc}}$ weighting in the story of Figure 6.1, and their ℓ_2-normed weights under various schemes.

$\text{idf}_{\text{static}}$ – and idf_{inc} with sufficient history – produces a relatively smooth set of non-zero weights for all terms in the document, while a small Δt leads to weights sparsely assigned only to terms that appear infrequently in the sliding window, such as carjack, hickson (tf = 2), rawson, knifepoint and auburn (tf = 1) in the example. While these terms are semantically similar to those we expect to be important for event linking, two pertain to an event that is background to the example news story.

Although we expect temporal variants of idf to better select salient terms and thus documents in many cases, the frequencies of individual terms may not be indicative of the novelty of a specific event, nor that the article reports it. We expect these weighting schemas may be more informative when combined with appropriate novelty zoning techniques.

6.2.4 Query formulation

Our system is supplied an event reference anchored to a short span of text – from a word to a sentence – within a document. We presently assume query terms are derived from the locality of the anchor, rather than expanded with reference to external sources.[5] As with target zoning, the selection of local terms for querying may be modeled in a supervised manner, but we leave this for future work. Rather, we consider term weight as a function of distance from the reference anchor (within its containing document).

Preliminary experiments with our event linking corpus evaluated a bag-of-stems query with different weighting schemes for each token given its distance δ from the centre of the anchor:[6]

$$\begin{aligned}
\text{qw}_{\text{discrete}}(\delta; k) &:= [\delta < k] \\
\text{qw}_{\text{linear}}(\delta; k) &:= \max\left\{0, \frac{k-\delta}{k}\right\} \\
\text{qw}_{\text{inverse}}(\delta; \alpha) &:= (\delta + 1)^{-\alpha} \\
\text{qw}_{\text{exp}}(\delta; \gamma) &:= \gamma^{-\delta}
\end{aligned}$$

We also considered structural distance: equal weight for all terms within the same paragraph as the anchor and 0 for others; and terms from the whole source document body. Querying with the bag of words contained in the anchor's paragraph performs similarly to discretely weighting a window of fixed size ($\text{qw}_{\text{discrete}}$ for $k \geq 16$), but since paragraph size is variable and highly genre-dependent, we avoid using them as window boundaries. Apart from being slow to query, whole documents also introduce excessive noise and far underperform other discrete weighting approaches, as shown in Nothman et al. (2012). We find that $\text{qw}_{\text{inverse}}$

[5] Some of the query terms extracted from this excerpt, however, may be determined using external knowledge, such as through named entity linking.

[6] Note that only the relative value of these weights is important as they are multiplied by idf and normalised for cosine similarity.

and $\mathrm{qw}_{\mathrm{exp}}$ may degrade performance but certainly do not improve it for parameters tested, so our present experiments work with $\mathrm{qw}_{\mathrm{discrete}}$ and $\mathrm{qw}_{\mathrm{linear}}$.

The present work adopts a single query formulation approach for each model, rather than learning to weight overlap features under several query formulations. Since such weighting is redundant, this allows us to focus on other variables in the system.

We also note that the query may be enhanced in other ways. For greater precision, we might consider the syntactic locale instead of the surface distance, as do Arapakis et al. (2014) in selecting a verb and its adjacent noun phrase chunks to search for a hyperlink target. Similarly, discourse analysis might provide more justifiable boundaries than a discrete window of fixed size in an instance-dependent manner. For greater recall of terms, we could expand local anchor context by identifying event or entity coreference within the document and distributing some of the query weight to these other referential forms.

6.3 Scoring candidates by publication time

Time is key in distinguishing events from many other facts, and was found invaluable for annotators seeking the target of an event link (Nothman et al., 2012). While the hard constraints as used by annotators may be inappropriate in an automated retrieval setting, we seek to model assumptions about the publication date of the event link target that may be inferred from a query.

The publication timestamp of the canonical target in general differs from the time the event occurred (or began), but the requirement that linked events have happened or are happening means it can be no earlier. Together with corpus characteristics identified in Section 5.4.3, this presents competing constraints on the publication date:

1. it is no earlier than the onset of the referent event;
2. it is earlier than any other article reporting the event after its onset; and
3. it is no later than, and likely close to, the publication date of the source document.

This suggests the following strategy:

- where the source document is timestamped, exclude any targets published after it;
- where the date of the event is known, prefer targets close to but not preceding that date; and
- where the source document is timestamped and the date of the event is not known, prefer to link to documents close to the source timestamp.

We therefore assign each date in the target archive, and thus candidate c published on that date t_c, a score for a given reference r at date t_r:

$$\sigma_{\text{time}}(c, r; \mathbf{w}) = \begin{cases} 0 & \text{if } t_r < t_c \text{ or } t_r \text{ unknown} \\ \max\left\{0, \mathbf{f}_{\text{time}}(t_c, r)\mathbf{w}^\top\right\} & \text{otherwise} \end{cases}$$

for a vector of weights $\mathbf{w}$ and a vector of real-valued feature functions $\mathbf{f}_{\text{time}}$ described in the following paragraphs. Since matching a target is dependent on both matching time and content, each candidate is scored by the product of this value with the term overlap score described in the previous section.

Bias term These temporal weights are merely heuristic, so we are required to assign some weight to candidates that have term overlap but no temporal weight assigned. Thus one element of $\mathbf{f}_{\text{time}}$ takes on a constant value of 1, independent of its arguments.

Temporal grounding heuristics Where the time of the event's occurrence can be identified, we assign weight to the period in which it occurs and decaying weight following that period. Recent benchmarks of temporal relation extraction (UzZaman et al., 2013) suggest it remains difficult for automatic systems to accurately temporally ground an event to a timestamp. We may therefore assign weight to multiple contradictory timestamps associated with an event reference. We moreover apply the simplistic assumption that all references to dates within the same sentence as an event reference anchor may be indicative of its occurring in that period.

Letting D_r denote the set of days mentioned in the same sentence as r, we have features of the form

$$f_{\text{grounding}}(t_c, r) = \begin{cases} \max_{d \in D_r} g(t_c - d) & \text{if } D_r \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

where g is a decay function for which $g(x) = 0$ for $x < 0$, $g(0) = 1$, and $g(x) \in [0, 1]$ for $x > 0$. This feature function may apply where a period longer than a day (e.g. a year) is mentioned, by simply including in D_r all days in that period. In practice, for references to individual days, we include both that day and the following day in the period, to account for a daily publication cycle.

We use four distinct features, respectively for references to individual days, weeks, months and years.[7] This allows for our intuition that a more precise reference should be awarded higher weight.

(Conditional) recency heuristics Where t_r is known, we default to assuming the target event is recent. To do so, we apply a feature $f_{\text{recency}}(t_c, r) = g(t_r - t_c)$. However, we expect this

[7] Parts of years larger than a month, such as a quarter, are rounded up to a year, etc.

should obtain a higher weight when there is no specification of the event's date of occurrence. We therefore add a further feature:

$$f_{\text{cond.recency}}(t_c, r) = \begin{cases} g(t_r - t_c) & \text{if } D_r = \varnothing \\ 0 & \text{otherwise} \end{cases}$$

In the present work we apply the same decay function for all features, being:

$$g(x) = \begin{cases} \frac{1}{x+1} & \text{if } x \text{ ¿= } 0 \\ 0 & \text{otherwise} \end{cases}$$

where x is counted in days. This reflects the near-logarithmic cumulative frequency of targets with respect to their temporal displacement from the source in our corpus annotation (see Figure 5.5). A shallower decay such as $\frac{1}{\sqrt{x}}$ might better suit the relationship between the date of event occurrence and reporting, but we leave its consideration to future work.

Together these seven features attempt to account for knowledge and prior assumptions of when an event link target is likely to be published, which our annotation effort suggests is an important component of event link retrieval.

6.4 Supervised learning of parameters

Given a collection of event links, we describe a procedure to determine estimates of $\mathbf{w}_{\text{terms}}$ and $\mathbf{w}_{\text{time}}$. Practically, this consists of the following steps:

1. Select a single feature and use it to score candidates for all training instances.
2. Multiply the scores of the top s candidates for each instance by their temporal features, and learn the optimal weights $\mathbf{w}_{\text{time}}$ for their linear combination.
3. Use the initial feature multiplied by the learnt temporal weighting to score candidates for all training instances.
4. Multiply the scores of the top s candidates for each instance by their term overlap features, and learn the optimal weights $\mathbf{w}_{\text{terms}}$ for their linear combination.

This can be seen as a single iteration of a process in which each weight vector is updated in turn using linear modelling while fixing the other, as shown in Algorithm 1. This approach allows us to use standard linear modelling despite learning the product of two linear spaces; we also benefit from drawing a new set of top candidates given a partial model. The models may be optimised in a classification or structured learning paradigm.

In a classification approach, each input reference is represented by a single positive instance, corresponding to the true target, and a sample of negative instances corresponding to alternative candidates. Training adjusts the weights according to loss over the predicted

Given: training references R, corresponding targets Y, candidates C, batch loss function L, maximum iterations $t_{\max}$, $\mathbf{w}^0_{\text{terms}} \neq \mathbf{0}$

1: $t \leftarrow 0$
2: **repeat**
3: $\quad \mathbf{w}^{t+1}_{\text{time}} \leftarrow \arg\min_{\mathbf{w}} L\left(\sigma(C, R; \mathbf{w}^t_{\text{terms}}, \mathbf{w}), Y\right)$
4: $\quad \mathbf{w}^{t+1}_{\text{terms}} \leftarrow \arg\min_{\mathbf{w}} L\left(\sigma(C, R; \mathbf{w}, \mathbf{w}^{t+1}_{\text{time}}), Y\right)$
5: $\quad t \leftarrow t + 1$
6: **until** convergence or $t = t_{\max}$
7: **return** $\mathbf{w}^t_{\text{terms}}, \mathbf{w}^t_{\text{time}}$

Algorithm 1: Learning weights for time and terms features

classification of all sampled instances. In contrast, a structured learning approach only considers the loss with respect to the single best candidate per reference given each setting of $\mathbf{w}$. Preliminary experiments with structured learning under a zero-one loss substantially underperformed the classification approach, which we pursue further.[8]

A few practical considerations pertain to treating the learning task as binary classification and the broader iterative update approach:

Candidate sampling We sample a fixed s candidates for each reference r, or as many as have non-zero scores if it is fewer. We select the s candidates c with the highest $\sigma(c, r)$ given the current model parameters (i.e. $\sigma(c, r; \mathbf{w}^t_{\text{terms}}, \mathbf{w}^t_{\text{time}})$ for line 3 and $\sigma(c, r; \mathbf{w}^t_{\text{terms}}, \mathbf{w}^{t+1}_{\text{time}})$ for line 4 of Algorithm 1), and force the inclusion of the true candidate. Where the true candidate has a score of 0, we discard the instance and its candidates during training, but account for the instance in evaluation. The data is highly skewed to the negative class, yet we find that undersampling degrades performance.

Loss function and optimisation We employ logistic regression to optimise $\mathbf{w}$, although we use these weights in a linear scoring function, rather than for log-linear prediction.[9] Each candidate is represented as a feature vector $\mathbf{x}_i$, assigning $y_i = 1$ to the true link target and $y_i = 0$ otherwise, and we minimise the regularised logistic loss:[10]

$$\mathbf{w}^*_{\lambda,\ell,\beta} = \arg\min_{\mathbf{w}} \min_{w_\beta} \sum_i \log\left\{1 + e^{-y_i\left(\mathbf{x}_i \mathbf{w}^\top + \beta w_\beta\right)}\right\} + \lambda \left(|w_\beta|^\ell + \sum_j |w_j|^\ell\right)^{\frac{1}{\ell}}$$

[8] Exact inference of the best candidate given each $\mathbf{w}$ may be expensive, so in these experiments we employed a similar approximation to the classification task, sampling a fixed number of alternative candidates to compare to the true candidate. We evaluated the structured perceptron and subgradient structured SVM. While other loss functions (perhaps based on candidate similarity) may be more appropriate, the outcomes of our structured prediction and undersampling experiments accord in highlighting the importance of negative examples.

[9] In early experiments we empirically determine that this finds better solutions in terms of our task objective than ℓ_2-regularised least-squares regression, and marginally outperforms a linear SVM model.

[10] This optimisation assumes the role of L in Algorithm 1, such that $\mathbf{x}_i \mathbf{w}^\top$ corresponds to $\sigma(C, R; \mathbf{w})$ but we elide a precise formulation of their relationship.

We use $\ell = 2$ regularisation since our feature space is dense and not large, and select the regularisation coefficient λ through cross-validated grid search for each model. The logistic regression implementation we use from LIBLINEAR (Fan et al., 2008) regularises the bias (or intercept) term w_β, corresponding to an additional feature of fixed value β for all instances; we find $\beta = 10$ is suitable for our data. The optimisation problem is solved using their trust region Newton method (Lin et al., 2008).

Initial weights and iteration Algorithm 1 at first updates $\mathbf{w}_{\text{time}}$ given an initial $\mathbf{w}^0_{\text{terms}}$. Although the opposite is possible, we choose this ordering since a large number of candidates have $\mathbf{f}_{\text{terms}}(c, r) = \mathbf{0}$ (the same being untrue of $\mathbf{f}_{\text{time}}$) eliminating many candidates from the sampling procedure. The initial term feature weights are set to 0, except for a single predetermined feature, which is set to 1. We select this feature to maximise average recall of the correct target among the top s candidates. Assigning non-zero weights to multiple term features makes inferring the top candidates much more costly; in this vein, we set the maximum number of iterations $t_{\max} = 1$ and thus update $\mathbf{w}_{\text{time}}$ and $\mathbf{w}_{\text{terms}}$ once each, leaving an investigation of iterative updates to future work.

We apply this estimation technique to evaluate our model in the next chapter.

6.5 Conclusion

We have described a preliminary system intended to perform event linking as a retrieval task. The system is intended to generate the most likely candidates given an event reference, among which a more precise disambiguation process may select a final target. It scores each candidate according to its time of publication and term overlap with the query text, with particular components to focus on three key aspects of the event linking task:

Entities and event description A target candidate should be preferred if it mentions the entities, location, time and general description of the event that are indicated in the input reference. We thus extract and differentially weight different types of terms in the reference context and candidate story (Section 6.2.2).

News discourse structure Not all mentions of an event in a candidate news story are being reported there, but some constructs within the text, such as the opening sentence, are likely to be indicative of novelty. This leads to the use of weighted zones identified in the candidate documents (Section 6.2.1).

Temporality Since event linking targets the article that first reports an event, we prefer documents that introduce new content (Section 6.2.3). The system also prefers candidates published shortly after the likely date of an event's occurrence, or assumes that

recently-reported events are salient, thus preferring candidates recent to the reference time (Section 6.3).

By and large, our system takes naïve approaches to these components. This allows the system to be easily replicated, providing a benchmark for the task and evaluating its feasibility. As a framework, the system is also extensible to introducing, for example, different query formulation methods, leaving many open directions for future improvement and nuance.

We have described a method for estimating the system's parameters, $\mathbf{w}_{\text{time}}$ and $\mathbf{w}_{\text{terms}}$, from annotated event links. However, our manual annotations are likely too few to learn an accurate estimate. In the following chapter we instead propose inferring parameters from noisy hyperlink data – avoiding the cost of high-effort event linking corpus annotation – and evaluate this system on our gold-standard event links.

Chapter 7

Evaluating event linking with noisy training

We set out to determine to what extent the system described in Chapter 6 effectively performs the event linking task. However, this entails determining its parameters from training data, for which the manually annotated event link corpus may not suffice.

In general, linguistic and media expertise is scarce which makes producing a statistically-sufficient annotated corpus costly. This is particularly applicable to the annotation task described in Chapter 5 which is very time consuming and therefore resulted in only 229 distinct event links from 150 document, which is insufficient to train the model of the last chapter. In previous work we have exploited hyperlinks in Wikipedia to automatically generate training data for multilingual named entity recognition (Nothman et al., 2013). Similarly, we assume that some portion of hyperlinks on the world wide web must correspond to event links: an author may link to a reporting news article when referring to an event. If such a subset is identified, it may be used to train an event linker; in general we believe hyperlinks are under-exploited by the NLP community as indicators of event coreference. This goes hand in hand with the suggestion in Section 4.2 that the output of an event linking system might be used for hypertext construction.

In this work, we go further to suggest learning from noisy, "silver standard" training data. Such data are automatically sampled, not manually verified as valid event links. Thus we experiment with two corpora of hyperlinks to online news with minimal filtering, under the assumption that links to news are often event-oriented. We quantify this assumption with respect to the set of hyperlinks within online content from the same publisher as our manually annotated corpus. The second corpus consists of citations in English Wikipedia targeting that same online news archive.

After further detailing the extraction of such hyperlink corpora in the following section, this chapter performs an event linking system evaluation with the following purposes:

1. to validate the event linking task and establish its difficulty through a performance benchmark;
2. to ascertain whether knowledge from noisy hyperlink data can assist event linking;
3. to identify aspects of Chapter 6's system that are most effective; and
4. to diagnose aspects of the task and manually annotated data of Chapter 5 for which the system and its noisy training is not sufficient.

The metrics for quantitative evaluation and system optimisation are outlined in Section 7.2 before detailing experimental results of development and testing in Section 7.3. We then analyse portions of our annotated corpus that our system is most and least successful at replicating (Section 7.4) before concluding with a discussion of areas where our system and training methods might be enhanced (Section 7.5).

7.1 Learning from hyperlink corpora

Hyperlinks reflect many purposes when connecting content in the world wide web, and the semantics of these links are generally not specified.[1] We might expect that hyperlinks often indicate shared topic, and thus the link graph can be exploited for text categorisation (Chakrabarti et al., 1998) or term disambiguation (Yang et al., 2006). As opposed to the mere network of inter-linked documents, here we focus on *in-text* hyperlinks: those appearing in a discourse context, rather than as isolated navigation elements, thus acting as indicators of reference. In such cases, hyperlinks may function to provide further detail regarding a discourse entity. This is often the case with links to Wikipedia articles, or citation of facts within the blogosphere. We assume that some portion of hyperlinks targeting news perform a similar function, when an author seeks to provide authority or reference about a past event. We consider the extent to which hyperlinks to online news may substitute for manual event link annotations when learning a model.

We presently consider two sources of hyperlinks to online news: those from news sources, comparable to our archive-internal annotation (see Chapter 5); and those from Wikipedia. These sources are often focussed on events, and hence their links are likely to obtain the expected function. In-text hyperlinks to news frequently also appear in the blogosphere and online forums, where the news is the subject of commentary. Our work avoids this additional source of hyperlinked reference, as it carries with it diverse boilerplate content, language style, and rhetorical intent; with the former sources we may assume some amount of stylistic consistency when sampling hyperlinks for training. The following subsections analyse our

[1] The HTML specification allows for a `rel` attribute of hyperlinks that is primarily used to indicate web site structure such as sequence and hierarchy. Hyperlinks in rhetorical contexts tend not to attract similar annotations.

assumptions with respect to these two sources of hyperlinks, before detailing the derived training corpora that enable the event linking evaluation in the remainder of this chapter.

7.1.1 Hyperlinks within online news

The use of in-text hyperlinks by online news sources is increasing, where it had formerly been a distinguishing property of the blogosphere (Coddington, 2014). Sampling articles from news and blog sites shows that 91% of links from news sites are publication-internal, with over half of all links targeting reference material and 30% targeting news reports (Coddington, 2012). This increased presence of hyperlinks from and to news reports presents a resource to cross-document event coreference identification, given the centrality of event reference to news. At the same time, news archives may benefit from this technology which could augment new and archival content with such links.[2] This section assesses the event orientation of existing hyperlinks internal to an online news source, with an eye to their use in learning event linking.

Using links from and to the same archive minimises the work involved in acquiring and processing data. In selecting a corpus of news-internal hyperlinks to exploit, we require:

- a quantity of such hyperlinks sufficient to learn a model from a sub-sample;
- a complete archive spanning multiple years, accounting for reference to long-past events; and
- the ability to identify an archival story given its URL;
- accessible meta-data such as the date of publication for each archival story.

The inclusion of URL information for each article in the New York Times Annotated Corpus (Sandhaus, 2008) makes `nytimes.com` suitable, although we have not found it to mark in-text hyperlink anchors. WikiNews[3] is an option attractive because of its free availability and culture of hyperlinking common to WikiMedia projects, yet we find few links targeting its own articles are in-text, and meta-data such as time of publication was not consistently encoded at the time of investigation. We elect to obtain an archive of Fairfax Digital (FD) content spanning from late 2008 to early 2013, incorporating almost 1.05M articles from a number of Australian news web sites.[4] For the period covered, this collection is the online parallel of the print archive used for annotation in Chapter 5, incorporating some additional content from mastheads belonging to localities other than Sydney.

To analyse the relevance of such hyperlinks to event linking, we label a sample of 350 instances from articles published in 2011 – including news, features, review, etc. – after first

[2] A number of sites utilise automatic insertion of hyperlinks (Coddington, 2014), albeit to reference portals on specific topics rather than individual articles.

[3] `http://en.wikinews.org`

[4] This includes `http://www.{smh,theage,brisbanetimes,watoday,canberratimes}.com.au` which all provide access to the same set of assets.

1. ✔ The body . . . was *found* under a plastic sheet . . .
2. ✔ . . . that the global hacking *incident* won't affect them.
3. ✔ . . . 29-year-old Andy Marshall *died* this week after . . .
4. ✔ Ellison was last year *named* best-paid executive
5. ✖ . . . the Mitsubishi i-MiEV, which is now on sale . . .
6. ✖ . . . according to the QS World University Rankings.
7. ✖ Read our previous Brisbane's Best: CBD lunches for $10 | Chips | Bakery | Mexican

Figure 7.1: Examples of hyperlinks (underlined) in context from the FD corpus: those marked with ✔ are annotated as event references anchored at the word in italics; those with ✖ are marked for discarding.

removing duplicate targets and source sentences. Without confirming that the link targets are canonical with respect to the event linking definition, we aim to identify clearly negative instances, being links that are:

- not within prose (e.g. among a list of links);
- instructive in their anchor text (e.g. click here);
- references to the target document, rather than to the situation it reports (e.g. This paper revealed on Monday . . . ; similarly); or
- based on references to non-events and non-*linkable* events (e.g. a link to a review of the mentioned entity).

The annotator is presented with each randomly-sampled hyperlink within a sentence of context,[5] and decides whether the text and its relationship with its target approximates an event link. If so, the annotator also selects a single word as the event link anchor.

About half of the sample is labelled positively; of the others, 30% are not within prose, 20% are instructional or a direct reference to the target document, and the remainder involve an inappropriate referent, often a non-event or *aggregate* fact.[6]

Examples of annotated hyperlinks are shown in Figure 7.1. The listed positive examples illustrate diverse types of event, and links spanning periods from one day (2) to eleven months (4). The hyperlink is often anchored to a phrase including an event predicate, but may merely span the name of a focal event participant as in 3; in 76 (43%) of the positive instances we marked an anchor word within the given hyperlink.

The negative examples shown target: a review of the mentioned product (5); a mentioned publication available as an article on the FD site (6); and related content in a navigation

[5]By default the annotator does not see the full source or target articles, but may open them separately.
[6]These proportions are estimated from a manual grouping of 50 negative instances.

feature (7). It may therefore be hard to distinguish hyperlinked entity names that are events from those that are not, while we expect they are less harmful in learning an event linking model than repetitive, non-syntactic navigation links like (7).

Other instances break our intuition that an event reference will be linked to an article reporting, or at least mentioning, that event having happened:

(48) a. Mr Price, fellow reporter Melissa Mallett and producer Aaron Wakeley were dismissed by the television network in August after it was revealed the network faked two live helicopter crosses to the Sunshine Coast where police and volunteers were searching for the remains of Daniel Morcombe.

b. The mobilisation against her was reminiscent of the controversy generated after Adshel pulled down Rip'n'Roll safe-sex ads from Brisbane bus shelters in response to complaints.

c. Work on redevelopment at the site began in March , with five World War II air raid shelters planned to have been restored along with the timber wharf

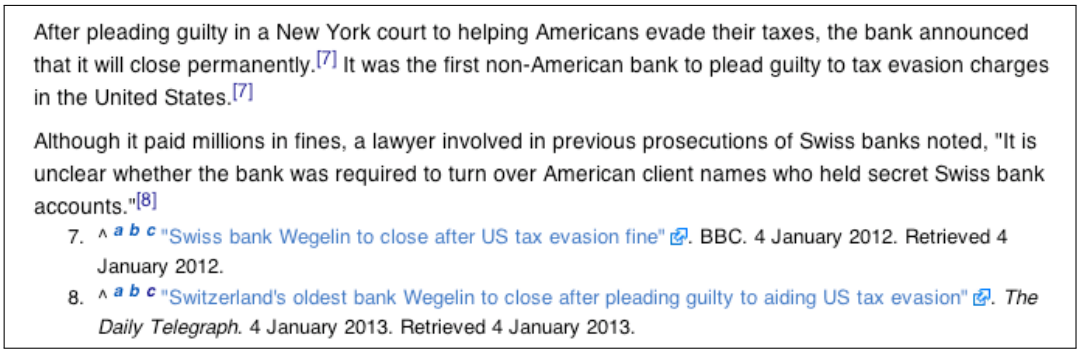
The hyperlink in Example 48a points to an article that is published prior to the mentioned dismissal, reporting the initial controversy. In 48b, the target reports the advertisements being reinstated after their removal generated controversy, rather than focally reporting the referenced event; yet it also happens to be the first article to reports the advertisement's removal. While event linking requires the event to be reported as having happened or happening, the target of Example 48c reports that the work "will begin today". Thus we find that hyperlinks referencing events display a variety of relationships between the mentioned event and the target news report.

Although we have considered using machine learning to distinguish the two classes of hyperlink, we have not yet surpassed the performance of a rule-based approach to removing non-in-text links. Using the rules below, we eliminate 60% of the negative instances in held-out data without sacrificing any recall of positive instances.

7.1.2 Citation in Wikipedia

As an encyclopædia, Wikipedia shares with the news genre a focus on event narrative for many of its topics, citing online news sources to support the factuality of its statements. As at August 2013, over 3,300 Wikipedia articles employ the `cite news` template in over 400,000 references, a common way to make these references, although they may include offline sources; others may use the more frequent and general `cite web` template. Figure 7.2 illustrates the multiple purposes of such citation:

- the encyclopædia may mention an event – here a bank's closure announcement – which is further detailed in a news report;


After pleading guilty in a New York court to helping Americans evade their taxes, the bank announced that it will close permanently.[7] It was the first non-American bank to plead guilty to tax evasion charges in the United States.[7]

Although it paid millions in fines, a lawyer involved in previous prosecutions of Swiss banks noted, "It is unclear whether the bank was required to turn over American client names who held secret Swiss bank accounts."[8]

7. ^ a b c "Swiss bank Wegelin to close after US tax evasion fine". BBC. 4 January 2012. Retrieved 4 January 2012.
8. ^ a b c "Switzerland's oldest bank Wegelin to close after pleading guilty to aiding US tax evasion". *The Daily Telegraph*. 4 January 2013. Retrieved 4 January 2013.

Figure 7.2: An excerpt from English Wikipedia that cites news to support an event reference, an analytic fact and a quotation or point of view. From *Wegelin & Co.* in Wikipedia, The Free Encyclopedia, last revised 2013-01-07 05:15UTC.

- it may mention some other fact – that the bank was the first to plead guilty to tax evasion – for which the news report is considered an authoritative source; or
- it may replicate a quote from the cited report.

The present work does not attempt to distinguish between these purposes.

Citation is possibly more similar to exhaustive event linking annotation than are hyperlinks in online news: the latter are only inserted when the journalist or editor feels it is helpful (or lucrative); in applicable genres, the former are required for any statement of knowledge. Hence news text may be biased towards hyperlinking obscure events, providing more information to the unfamiliar reader. Conversely, reference to events that require no further information may be absent from a news-internal link corpus, while Wikipedia citations' primary bias is towards notable events relevant to notable topics.

While hyperlinks on the broader world wide web often span a relevant fragment of text, Wikipedia references mark a point in the text, as in academic citation. This leaves the scope of the citation ambiguous within the portion of text preceding the citation point. To use our event linking system in these cases involves selecting an anchor point closer to the focus of the event reference, rather than its end.

Because we require a complete target archive for event linking training, we only consider links to the archive used for internal links in the previous section.

7.1.3 Derived corpora

We extract all hyperlinked text from FD articles, which appear in content published from 2008 to 2013. These are then filtered to retain only links that:

- target the indexed content management system (CMS);[7]

[7] These URLs have the following form: `http://<domain>.com.au/<category>[/<category>]/<slug>-<date>-<base36id>.html`, for example `http://www.smh.com.au/national/police-considering-semiautomatic-20091014-gwoh.html`.

- target news media where this can be identified easily from the URL, excluding photo galleries, blogs, polls and travel reviews;
- target Fairfax Digital domains that primarily deal with news to exclude the likes of `essentialbaby.com.au`;[8]
- have non-empty anchor text;
- are obviously not in-text references to the target subject matter, in that they:
  - do not include in their anchor text: the words click or here or terms and conditions; or a FD publication name, as in see more at brisbanetimes.com.au;
  - do not constitute a complete HTML block element or a whole sentence, as these are often indicative of non-textual references, such as lists of links or reference by title to related content;
  - do not neighbour another link with only white-space or punctuation between them, for similar reasons;
- have a target timestamp on the same day or later than the source document, which may be violated due to editorial modifications and inaccurate data;
- are indexed in our archive, thus not having having been revoked, etc.;
- do not have the same target as a previously-processed link, or the same set of words in the linking sentence, in pursuit of sample independence for training and cross validation.

This results in 20,923 hyperlinks for training. As in our manual annotation, we exclude the reference source document as its candidate link target. For the purpose of query formulation, we consider the event reference to be anchored in the middle token of the linked text.

We derive another corpus of links from English Wikipedia of April 2012, retaining only links that:

- are used for article endnote-style citation;
- target the indexed CMS on the `smh.com.au` domain and are indexed in our archive;[9]

[8] While these domains will often provide re-branded access to the same set of assets, the explicit choice of domain may indicate a non-news target. We accept links to the following http domains: business.brisbanetimes.com.au, news.brisbanetimes.com.au, www.brisbanetimes.com.au, www.businessday.com.au, www.canberratimes.com.au, brisbanetimes.domain.com.au, news.domain.com.au, smh.domain.com.au, theage.domain.com.au, watoday.domain.com.au, brisbanetimes.drive.com.au, news.drive.com.au, smh.drive.com.au, theage.drive.com.au, www.nationaltimes.com.au, business.smh.com.au, news.smh.com.au, www.smh.com.au, business.theage.com.au, news.theage.com.au, www.theage.com.au, business.watoday.com.au and www.watoday.com.au.

[9] In future work we would like to expand this to all domains indexed in our archive, as well as links to the legacy content management system, as their absence depletes the corpus substantially.

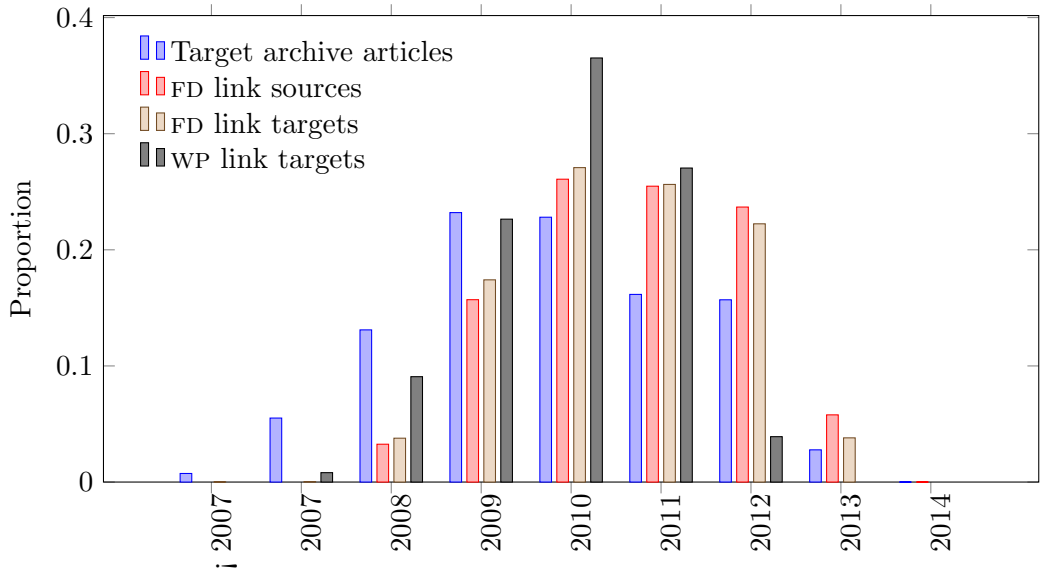


Figure 7.3: Distribution of the target archive and hyperlink sources and targets by year

- are the first link per source article with a particular target, in pursuit of sample independence;
- are the only such link at a particular reference site, assuming that multiple targets cannot all be canonical reports of an event.

This results in 2,226 hyperlinks for training. Because references tend to be attached at a point after the relevant statement, such as a sentence boundary, we estimate the reference anchor as a token halfway between the citation point and the start of the sentence.

Figure 7.3 shows the temporal distribution of these two hyperlink corpora in comparison to the target archive. Firstly, we note some dirtiness in the target archive that is to be expected of non-curated data. A small number of articles (7342) are timestamped prior to 2007. Much of this appears to be archival news imported into the CMS, while some of it is likely incorrectly timestamped, as is the case with content marked 2014 and late 2013. The spike in 2007 (and early 2008) represents content imported into the new CMS, including static content such as travel information, but also some archival news incorrectly assigned the date of import rather than creation. There are fortunately very few hyperlinks to this data, although it will likely cause an artificial inflation of the recency feature for FD, which the following analysis suggests is already subject to skew.

37.8% of hyperlinks in FD target articles published on the same day, which is much more frequent than in our event links (5%), and 69.8% target articles in the past week (vs. 28% of manual event links); the long tail is much lighter (cf. Figure 5.5), with only 3.4% of links (vs. 33%) spanning over a year and 0.9% (vs. 23%) over two. The indexed CMS covers relatively few years compared to the archive for our manual annotation, so we cannot expect FD links to

span over many years. We note a substantial increase in the number of hyperlinks relative to articles over time, reflecting the trend described from the perspective of institutional analysis in Coddington (2014).

We have not identified a certain cause for the spike in WP link targets in 2010, although the dataset is not large and this is the only Australian federal election year and change of prime minister fully covered by the collection.

These differences between these hyperlink corpora are further discussed with respect to development results in Section 7.3, below.

7.1.4 Learning procedure

When learning the weights of the temporal and term overlap features, we must also set the learning algorithm's regularisation coefficient, λ. This is selected by grid search to maximise the average of the task metric (MRR) across n folds, ensuring that the candidates of each query are contained in the same fold since they are interdependent. For the FD data we use $n = 3$, and $n = 10$ for the smaller WP sample.

We find that many negative candidates help in learning these weights, and that this has greater effect than utilising more training instances. Therefore up to 10,000 training hyperlinks are used in each fold of cross-validation, although the discarded instances are retained in evaluation, with $s = 2{,}000$ candidates per hyperlink.

7.2 Metrics

Where an event linking system outputs a single link target for each input instance, its responses may be evaluated using precision and recall. The link target space being sparse, and this work presenting an early evaluation of the task, we consider such an evaluation too strict. Rather, when evaluating two systems on a single instance, we should award the system that places the correct response at a lower rank in a returned list of candidates. This is essential since our baseline systems have very low recall at rank 1; but this evaluation also corresponds well to settings where a human is required to verify or select among the top candidates returned by a system. As such, we use mean reciprocal rank as our primary metric during system development, with recall at rank r used to illustrate the distribution of correct responses within ranked outputs. Both measure the response to a set of queries as a value from 0 to 1 with 1 indicating the correct target always appears as the first response. To convert from scores to ranks, we assign the correct target the highest (worst) ranking among ties in order to penalise ambivalence.

Significance testing During development, Wilcoxon's signed-rank test (1945) is used to test the hypothesis that a system improves upon the baseline, but WP alone has sufficient

folds to apply it. For final testing, we use approximate randomisation (Noreen, 1989) over 50,000 trials. In both cases, we reject the null hypothesis where $p < 0.05$.

7.3 Results

With the goal of establishing the task, the system, and the use of noisy training data, we obtain quantitative results to answer the following questions:

- What query formulation and term weighting establish a high-recall baseline?
- Are event-like hyperlinks predictable? How does this differ between FD when filtered and unfiltered? How do those compare to Wikipedia news citations?
- How does weighting candidates by their publication date affect learning this prediction?
- How do zoning, term weighting and term type affect learning this prediction from term overlap, and how do their effects interact?
- Are the effects of term overlap and temporal weighting cumulative?
- Do our hyperlink prediction results carry over to event linking? How is this affected by choice of training corpus?

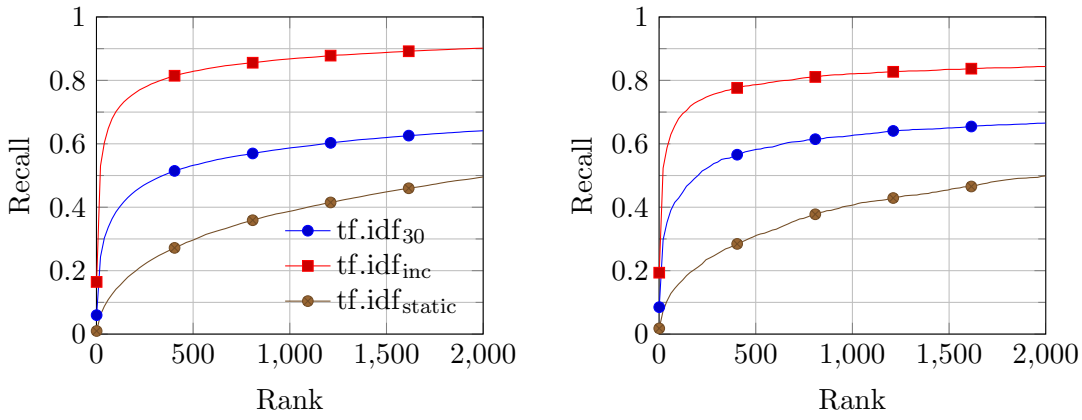
7.3.1 Development results

During training and development, our system is only exposed to and evaluated on hyperlinks. Our development results are therefore indicative of how predictable the selected hyperlinks are, and the effectiveness of our system to make those predictions.

The system depends on an initial ranking of candidates determined by a single term overlap feature with high recall. We begin by identifying this parameter, before examining the impact of term overlap features and temporal weighting.

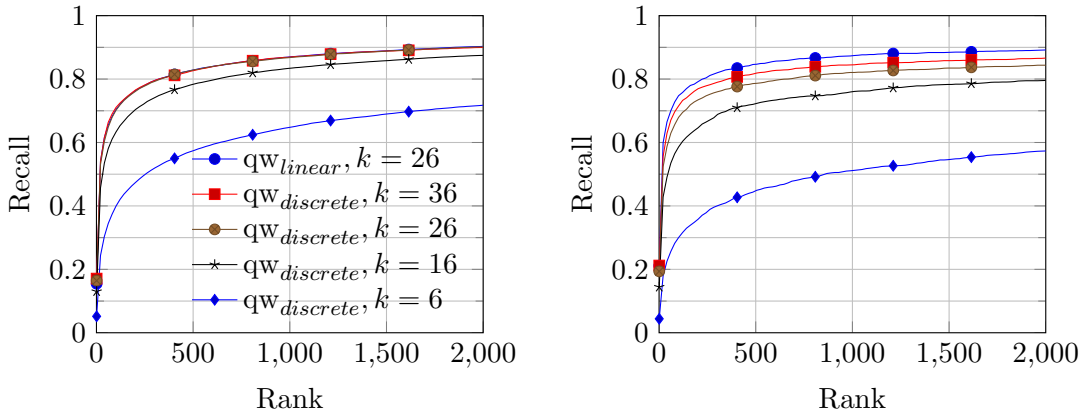
7.3.1.1 Query formulation and term weighting for high recall

Intuitively, high recall (at some rank) is obtained in similarity-based retrieval by using a broad query with as many terms as possible, while avoiding the introduction of excessive noise. Thus we find that comparison over the bag of word stems retrieves more correct targets in the top 10, 100, or 1000 candidates than more precise terms such as named entities; matching against the whole candidate text is similarly most effective, although matching only terms appearing in the first three paragraphs is not far behind and may produce a higher R@1 or MRR depending on how the query is formulated. We therefore investigate term weighting and query formulation for an initial candidate ranking.



(a) FD hyperlinks (b) WP hyperlinks

Figure 7.4: Recall of hyperlink targets with varying term weighting schemes. All queries consists of stems with $\text{qw}_{\text{discrete}}, k = 26$ matched against the full candidate text.



(a) FD hyperlinks (b) WP hyperlinks

Figure 7.5: Recall of hyperlink targets with varying query formulations. All queries consists of stems matched against the full candidate text with $\text{tf.idf}_{\text{inc}}$ weighting.

At all finite ranks, traditional $\text{tf.idf}_{\text{static}}$ recalls substantially fewer of the correct link targets than $\text{tf.idf}_{\text{inc}}$, as shown in Figure 7.4. This is presumably because the latter scheme emphasises earlier – and therefore often seminal – publications using some terminology, while $\text{tf.idf}_{\text{static}}$ down-weights early mentions of a term if they later become frequent. We also experiment with tf.idf_{10}, tf.idf_{30} and tf.idf_{90}, which perform similarly to each other, between $\text{tf.idf}_{\text{static}}$ and $\text{tf.idf}_{\text{inc}}$.

Figure 7.5 illustrates recall in terms of query formulation approaches. We see a dramatic increase in performance from using a window of 11 tokens ($k = 6$) to 31 ($k = 16$), and further gain when each query draws on 51 tokens ($k = 26$). There is no discernible improvement in FD performance with a query expanded by another 20 tokens, while this has some positive effect for WP; to avoid increased query time, we opt to use $k = 26$.

Model	FD	WP
Single term overlap feature	0.257 ±0.006	0.309 ±0.011
All term overlap features	0.339 ±0.010	0.324 ±0.024
Single term feature with temporal weight	0.409 ±0.003	0.377 ±0.030
All term features with temporal weight	**0.444**±0.008	**0.387**±0.030

Table 7.1: Baseline, partial and full development results: mean and standard deviation of MRR under cross-validation of two hyperlink corpora. The single feature used is stem overlap with $\text{tf.idf}_{\text{inc}}$ weighting. FD experiments use $\text{qw}_{\text{discrete}}, k = 26$ for query formulation; WP uses $\text{qw}_{\text{linear}}, k = 26$.

Incorporating decay into query weighting, we note further differences between FD and WP. For FD, there is little difference in recall under a single feature model between $\text{qw}_{\text{discrete}}$ and $\text{qw}_{\text{linear}}$ (Figure 7.5a) while the linear decay model provides a substantial gain on the WP data (Figure 7.5b). When we incorporate a learnt temporal model, we find that discrete weighting outperforms linear for FD (MRR = 0.41 rather than 0.35), but remains behind for WP (MRR = 0.31 rather than 0.38). The effect of the decayed model for WP may be surprising, since links are heuristically anchored halfway between the citation point and the start the sentence. Further consideration suggests that this preference for weight is due to the density of disparate events (and facts) in Wikipedia: in an encyclopædia, a displacement of 30 words is likely to refer to an entirely different event within a chronological narrative; the same distance in news text is likely to be topically relevant, at a minimum.[10] We continue to report performance with $\text{qw}_{\text{linear}}, k = 26$ for WP and the discrete equivalent for FD, establishing these single-feature results as our baselines.

7.3.1.2 Overall development performance

Reporting mean MRR of cross-validation with the best λ parameters, Table 7.1 compares baseline performance to three models:

1. Learning $\mathbf{w}_{\text{terms}}$ for 315 term overlap features without temporal weighting;[11]
2. Learning $\mathbf{w}_{\text{time}}$ for seven temporal weight features, with the baseline term overlap feature; and
3. Learning $\mathbf{w}_{\text{terms}}$ with predetermined $\mathbf{w}_{\text{time}}$.

Each of these models improves upon the previous in both FD and WP training. WP has a higher baseline performance than FD, and all of its gains are more modest, with the temporal

[10] Yet this explanation may not be sufficient, since we have already noted that additional context ($k = 36$) helps WP recall where it does not for news.

[11] Strictly, this means $\mathbf{w}_{\text{time}}$ only includes non-zero weight for the bias feature.

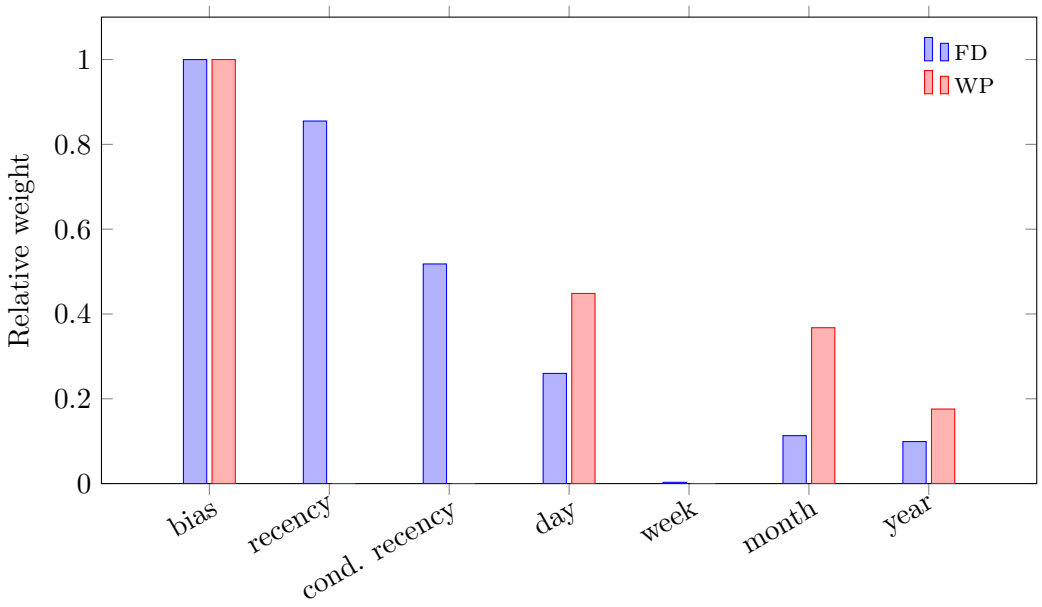


Figure 7.6: Learnt values of $\mathbf{w}_{\text{time}}$ as a proportion of the bias weight.

features alone improving the FD model by 0.15 MRR, and almost half that in WP. Despite the term overlap features providing significant gains off the baseline model for WP, their impact is not significant when augmenting a model with temporal weighting. For FD, adding term features to the temporally-weighted model provides an even smaller gain relative to adding term features on the baseline. This suggests that most of the information obtained from term overlap features is redundantly provided by temporal weighting.

These results show that our system, and particularly its temporal model, is able to to improve the prediction of hyperlink targets. The effectiveness of temporal weighting, especially in WP where recency features do not apply, suggests that a substantial portion of the corpora indeed correspond to event reference. We investigate the contributions of individual features in the following subsections.

7.3.1.3 The effect of temporal weighting

After cross-validation, each models is re-fit on its complete training corpus. Figure 7.6 illustrates the resultant weights for temporal features, recalling from Section 6.3 that these features take the value 1 in the targeted period (same day of publication for recency features, and within the year, day, etc. otherwise) and values between 0 and 0.5 otherwise. It shows that recency is a very informative feature for FD hyperlink detection, with a weight 1.3x bias where conditional recency applies, i.e. for sentences that do not refer to dates. Performance decreases if the model is re-estimated with the unconditional recency feature removed, by 0.027, but is unaffected by the removal of conditional recency (a drop of 0.003), reflecting the limited information used to apply its heuristic. Yet the model does not rely on recency

alone; with only recency features for temporal weighting, performance drops by 0.010.

For both WP and FD models, information from local temporal reference can be utilised positively. For features based on temporal references in the source sentence, we expect weight to have an inverse relationship with the length of the period referenced (i.e. $w_{\text{year}} << w_{\text{month}} << w_{\text{day}}$), since spurious candidates are assigned weight in proportion to the period length. While we find that approximate ordering among the weights, we note some difference from our expectations:

- References to weeks are highly uninformative, despite having a frequency in FD 76% of that of years.[12] This may be due to greater ambiguity – thus incorrect normalisation – in reference to weeks than to other temporal units:[13] expressions such as last week or three weeks prior may or may not be a reference to a seven-day period beginning Sunday, and can be ambiguous when stated on a weekend; on the weekend can refer to a past or a future time.

- Weighting of months is surprisingly close to that of day in WP and to year in FD. We have not yet identified a cause for this difference and may investigate it in future work.

These results suggest that all temporal features but mentioned weeks and perhaps conditional recency are informative where available.

7.3.1.4 The effect of term overlap features

Unlike the temporal features, our term overlap features are highly redundant and interact with each other when learning a logistic regression model. Since their their values are neither normalised nor binary, directly inspecting the feature vector is not very uninformative. Instead, we consider the effects of term overlap features by learning models with a subset of features. We perform cross validation for variants over each axis, otherwise fixing the use of a learnt temporal model, stems, $\text{tf.idf}_{\text{inc}}$ and full target text.

Term extraction Supplementing the baseline with other term extractors, we find named entities by type (a separate feature for each of locations, organisations, people and miscellaneous) are the most effective features for FD, alone accounting for the 0.04 increase in the final model.

The WP model benefits most often by matching temporal references (gain of 0.013), and this improvement is significant, whereas adding all zone-extractor-weight combination features is not, suggesting that the latter is damaged by feature noise. Adding in other term features then marginally (by 0.003) improves the model. Temporal reference overlap is most useful when weighting by publication date is not informative, suggesting that some of the

[12] They are rare in WP, 0.5% of all references considered vs 16.8% in FD.

[13] We have not managed to locate an error analysis for Heideltime by unit.

Model	FD	WP
stem	0.409 ±0.003	0.377 ±0.030
stem, NE mention	0.410 ±0.003	0.380 ±0.029
stem, mention by type	0.411 ±0.004	0.383 ±0.030
stem, NE	0.412 ±0.003	0.385 ±0.028
stem, NE by type	**0.413**±0.004	0.382 ±0.031
stem, mention tokens	0.409 ±0.003	0.377 ±0.031
stem, day	0.409 ±0.004	0.385 ±0.032
stem, day, month, year	0.411 ±0.002	0.390 ±0.030
all	**0.413**±0.007	**0.393**±0.026

Table 7.2: Performance with selected term extraction: mean and standard deviation of MRR under cross-validation of two hyperlink corpora. These models only use $\text{tf.idf}_{\text{inc}}$ features matching the full text, and build on top of temporal weighting.

Model	FD	WP
$\text{idf}_{\text{static}}$	0.180 ±0.001	0.143 ±0.018
idf_{inc}	0.409 ±0.003	0.377 ±0.030
idf_{30}	0.205 ±0.003	0.234 ±0.020
$\text{idf}, \text{idf}_{\text{inc}}$	0.423 ±0.003	**0.395**±0.024
all	**0.425**±0.004	0.382 ±0.026

Table 7.3: Performance under variant term weighting: mean and standard deviation of MRR under cross-validation of two hyperlink corpora. These models only use stem features matching the full text, and build on top of temporal weighting.

Wikipedia citation targets mention events far preceding or following the target's publication date. For example, where Wikipedia states the date of an election, it is unlikely to cite the article where the election outcome is reported, rather one where the election date is set. Similarly, Wikipedia may cite a report on court proceedings to support its narrative of a much earlier crime. Thus the strong positive response of WP to temporal mention features despite temporal weighting may be an indicator of its mismatch to event linking.

Term weighting The high performance of $\text{tf.idf}_{\text{inc}}$ in both models (see Table 7.3) relative to other weighting schemas is not surprising given that it determines the set of candidates for learning. Nonetheless, utilising both $\text{tf.idf}_{\text{static}}$ and $\text{tf.idf}_{\text{inc}}$ features improves MRR significantly by 0.018 for WP and by 0.014 in FD. The tf.idf_{30} features are then significantly detrimental to WP performance – perhaps Wikipedia references articles only after their topic has been in the news for some time – while providing a marginal benefit to the FD model.

Model	FD	WP
full text	0.409 ±0.003	0.377 ±0.030
full text, headline	0.408 ±0.004	**0.379**±0.032
full text, first sent.	**0.414**±0.004	0.378 ±0.030
full text, three paras.	**0.414**±0.004	**0.379**±0.029
full text, most recent timex	0.408 ±0.003	0.377 ±0.028
full text, last week timex	0.409 ±0.003	0.377 ±0.030
full text, relativised	0.409 ±0.003	0.376 ±0.032
all	**0.414**±0.005	0.376 ±0.031

Table 7.4: Performance with additional zones: mean and standard deviation of MRR under cross-validation of two hyperlink corpora. These models only use stem features weighted by $\text{tf.idf}_{\text{inc}}$, and build on top of temporal weighting.

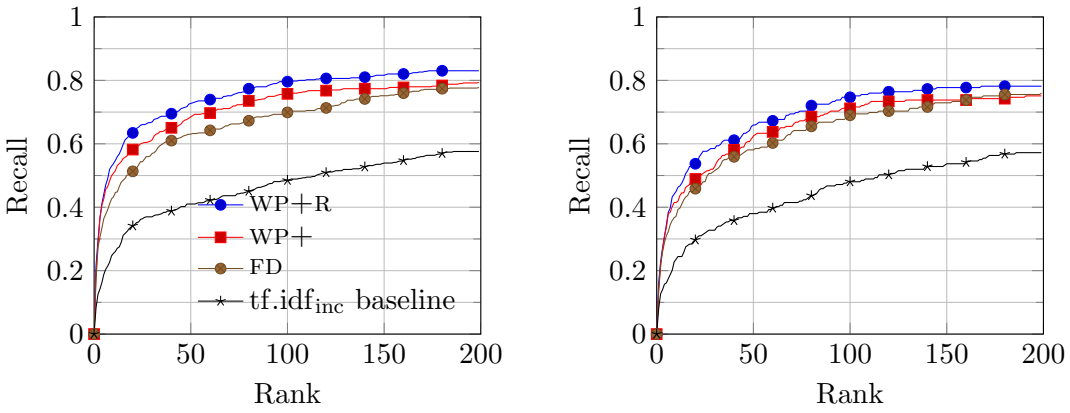
Target zoning Our results suggest that using explicit temporal cues (sentences containing the most recent temporal reference; sentences referring within the past week; fragments before after, since or following) to identify portions of the document that report new content is ineffective, if not noisy, as shown in Table 7.4. Using standard discourse properties – the content of the headline and lead sections of an article – has some effect, albeit marginal in WP and small in FD. A more robust model of zoning might be achieved by learning weights for sections of document as suggested below (Section 7.5.1).

As in the case of term weighting, incorporating all features is detrimental to the WP model but does not affect FD, which may merely be due to WP's much smaller training data.

A selective model, WP+ Considering that some features are found detrimental to the WP model, we re-learn its $\mathbf{w}_{\text{terms}}$, forcing those feature families to zero. This model's term overlap features are thus composed of:

- Full text, headline, first sentence and first three paragraph zones;
- All term extractors;
- $\text{tf.idf}_{\text{static}}$ and $\text{tf.idf}_{\text{inc}}$ weighting; and
- $\text{qw}_{\text{linear}}, k = 26$ query formulation.

This produces a result of MRR = 0.409, which is a significant improvement over all WP models reported so far. We title this improved model WP+ and use it to evaluate Wikipedia citations' utility as event linking training data.



(a) All linked mentions

(b) First linked mention per target

Figure 7.7: Recall of event links by systems trained on Fairfax and Wikipedia hyperlinks compared to a bag-of-words cosine similarity baseline.

Model	MRR	R@1	R@10	R@100	R@1000
$\text{tf.idf}_{\text{inc}}$	0.13	0.08	0.24	0.48	0.75
FD	0.22	0.14	0.37	0.69	0.85
WP+	0.23	0.14	0.41	0.71	0.89
WP+R	0.25	0.16	0.45	0.75	0.90

Table 7.5: MRR and recall of event links by systems trained on Fairfax and Wikipedia hyperlinks compared to cosine similarity, evaluating on the first linked mention per target.

Finally, since our gold-standard event links are built on an archive-internal model, we expect them to respond to recency features that are not applicable to Wikipedia.[14] We therefore construct a hybrid $\mathbf{w}_{\text{time}}$: we first scale the weights from the FD and WP models relative to their bias features, and then adopt the recency and conditional recency feature weights from FD, producing a variant of WP+ that we call WP+R.

7.3.2 Results on event links

Despite their noisy training, our systems far outperform cosine similarity baselines in predicting manual event links.[15] The instances evaluated here include multiple co-reported references within a single document, so we report similar results in Figure 7.7b and Table 7.5 considering only the first reference to each correct target. All trained models significantly differ from the baseline in MRR and R@1. The absolute scores appear low compared to development results, suggesting that manual event links may be more difficult to predict than hyperlinks, towards

[14] Certainly, one could propose using the edit history to determine recency, but this is well out of scope for our work in which we treat Wikipedia as static.

[15] Here we consider all *linkable* references in the corpus where annotators identify a target, and not *reported here*.

which our analysis below provides some insight.

Wikipedia training seems to be a better match for our event links, although the difference in MRR is not significant, perhaps because citation better corresponds to the function of event linking than does the range of hyperlink meanings in FD. Furthermore, the high performance of WP+, which lacks a recency feature, shows that mere textual and temporal similarity between the source and target are not sufficient for the task.

We note that these results differ somewhat from those reported in Nothman et al. (2012), where a simple stem overlap score was reported to perform as well as some of the best systems here. We are only able to account for this as a function of relying on Apache Lucene and its similarity score (Apache Software Foundation, 2012) in that work. That measure has four major differences from tf.idf cosine similarity – intended for efficiency and effectiveness – that we believe are responsible for the difference: (a) it multiplies the overlap score by a "coordination factor", being the proportion of query terms matched in the candidate, making it less susceptible to the influence of a single query term; (b) it does not use the Euclidian norm for the candidate's term vector, instead using a normalising factor of $\sqrt{n}$ for document length n, such that the score is affected by the number, but not the weight, of terms not matched by the query; (c) it reduces the skewing impact of raw term frequency by using its square-root; and (d) its formulation of idf adds a constant value of 1 to the log-inverse frequency, therefore spreading weight more uniformly among a term vector. The combination of (b) and (c) result in scores that do not range from 0 to 1 and can widely vary for different queries. This led us to abandon its scoring function, since such values seemed inappropriate for use as features when learning a weighted model. However, early experiments suggested that temporal weighting and zoning could produce substantial gains in the Lucene framework also. This leads us to believe that our results are accurate to suggest the effectiveness of these features and noisy training, but understate the potential performance for this class of model.

7.4 Analysis

We analyse the output of the WP+R system in order to get a better understanding of the types of links that are readily predicted with this system and those that are not. Among the set of event references correctly linked, of which a sample is shown in Table 7.6, we find various types of event reported up to nine years before the reference (example 5), which appear in textual genres from opinion (4) to obituary (5) as well as the most frequent category of news reports. The system is most successful where the mentioned terms and entities select distinctive topics (e.g. 2, 3, 5), or if the topic is more general but a specific date is given (4), or the reference is to the most recent news on that topic (7).

In some cases where the linker nearly makes the correct decision, assigning it rank 2, the predicted candidate story follows up on the previous day's news that should be targeted,

#		Date	Reference or target headline
(1)		2009-01-02	struggled until the Gates *donation* arrived ,
	→	2005-06-29	*Mozzies' days numbered if Gates grant does trick*
(2)		2009-01-30	Beyinhar project it *acquired* through a scrip acquisition of Golden China
	→	2007-10-17	*Sino Gold seizes the moment*
(3)		2009-01-30	, Sino has *found* Beyinhar does not contain
	→	2008-10-23	*Sino looks for other opportunities in China*
(4)		2009-02-09	distortion Rudd had *uttered* in the same
	→	2008-02-14	*Together we'll build a truly great nation*
(5)		2009-02-28	Pequot Tribal Nation in America *commissioned* a new study
	→	2000-11-02	*'Hanged' man strangled, experts fear*
(6)		2009-04-01	siege , the academy was *retaken* by Pakistani commandos
	→	2009-03-31	*Suspected terrorists killed after Lahore siege*
(7)		2009-05-14	national teachers union has *threatened* to boycott
	→	2009-05-12	*Teachers will boycott standardised schools tables*
(8)		2009-06-26	The corporate regulator *moved* to freeze Mr Groves
	→	2009-06-25	*Watchdog moves to freeze ABC founder's assets*
(9)		2009-07-01	$ 210,100 over the *death* of Nathan Francis
	→	2007-05-16	*Army ration suspected in boy's death*
(10)		2009-07-14	practised by the *collapsed* lender Opes Prime
	→	2008-03-29	*Opes collapse could reveal a sordid tale ...*
(11)		2009-09-14	to make Enough Rope and *poaching* Margaret Pomeranz and ...
	→	2004-04-06	*SBS left reeling as film critics give it the flick*

Table 7.6: Examples of event links correctly identified by WP+R

under the influence of the recency weight; and others where a news item briefly anticipates an upcoming event, and is preferred to the correct target since cosine similarity prefers a small text with the same key terms. Other instances exhibit uncertainty between multiple articles with related coverage of an event, which also occurred in our annotation. However, rank 2 results also frequently exhibit a practical problem of working with news corpora: while our archive is presumed to consist of a newspaper's edited print content, it contains duplicate and near-duplicate articles on occasion.

Apart from successes and near misses, we note the following patterns:

On-topic ranks before canonical Retrieving related content tends to require much less inference than retrieving a canonical representative, suggesting a key difference between this task and topical news clustering. Thus a reference to the Hawke government's *decision* in 1990 to combine the two public monopolies correctly targets news from 1990, but ranks a collection of news, analysis and opinion that prominently mention the monopolies by name well ahead of the canonical target in rank 168. This may be influenced by our noisy source of training, where the link target need not be canonical.

Reference language differs from the target A canonical target may fall behind other candidates if it uses different language, but sometimes this is related to the passage of time or change of perspective. Hence the canonical target of Laura Andrassy, the woman whom Norman *dumped* mentions Laura Norman but never Andrassy, a name she used after their divorce; a report about the arrest of a suspected terror cell does not mention the individuals' names although later references do.

Reference is terse or indirect Events that are particularly familiar in popular discourse can be understood in terse or implicit forms, but the system lacks the relevant background knowledge. In 2009, the Prime Minister's national *apology* to the stolen generations of the previous year requires no further detail or temporal grounding. Note also that despite popular familiarity with this event and its name-like reference, its coverage in Wikipedia is subsumed in related articles, without one of its own.

Context is misleading The window of terms that constitutes a query may include misleading noise. This often occurs when on event is discussed in relation to another, such that the woman whom Norman *dumped* ... to be with Evert scores articles about Evert highly, even though she wasn't relevant when Andrassy was dumped. This problem may also be genre-related: we find letters to the editor can refer densely to many topics. This problem can also provide incorrect temporal information, as with Groves *stepped down* as chief executive of ABC Learning just weeks before it collapsed last November which incorrectly weights on-topic content from November 2008 higher than the report Groves' departure a month earlier. Precise temporal relation extraction could resolve

the latter problem, while the former may benefit from syntactic and discourse-based query formulation.

Similar events are confused Where a number of similar events have been reported in the archive, our system does not ensure that disambiguating information from the reference is harnessed to discriminate between them.

Weighting is too aggressive Several instances show that our model is excessively biased towards term frequency, such that the first candidate for the Groves example is a long analysis piece mentioning ABC Learning, while another highly-ranked candidate is unrelated, but contains the word creditor 11 times and liquidator 9 times. Elsewhere, recency weight biases against the correct target.

These errors suggest that while term overlap and especially temporal weighting provide an effective foundation for event linking, more sensitive weighting, deeper linguistic disambiguation, and incorporation of external knowledge would all help the system and differentiate it further from topic tracking.

7.5 Discussion and future work

The models evaluated here are intended as an initial exploration of event linking candidate generation that are able to incorporate a number of our beliefs about the event linking task. Although fine disambiguation processes will ultimately be necessary to solve the task, the evaluation has also highlighted limitations of the current model. Without proposing an intricate disambiguation method, we consider possible enhancements to an IR approach. In addition, although noisy training data successfully supervised our models, we briefly note possible improvements in that area. Finally we consider other aspects of the event linking task before concluding the evaluation.

7.5.1 Enhancing the system

One area to improve the current framework is in better interpretation of the input reference, by harnessing local context, world knowledge, and initial query results. The query can be enhanced through identifying possible coreference locally and elsewhere, just as our annotators make use of Wikipedia or a selection of initial query results to determine the date of an event. As we've shown, temporal information is very useful in this task; more precise linguistic processing of the query and related content could similarly give a better bound (both before and after) on the expected publication time, while a more approximate approach could seek dates mentioned in related documents, essentially a temporal version of relevance feedback.

The archive, too, could be enhanced: indexed news from multiple sources could be clustered to enrich an event's representation, as could later references to the event, thus tracking

developments in the language of its reference.

While we find the most effective markers of newly reported content to be positional features, we believe it possible to identify more precisely portions of a well-structured news story that are reporting content. This could be driven through a first story detection model that considers a story in comparison to earlier content; or through local syntactic and discourse indicators of new content identified by comparing reference and target text.

More specifically, our method of linearly interpolating cosine similarities might be improved upon[16] by directly learning the weighting of terms for cosine overlap as in Yih (2009). This allows us to utilise similar information (textual position, document frequency variants and term burstiness, etc.) in a more principled manner. Yet we could also incorporate local features that are difficult to represent in the current model, such as whether a stem is realised as a participle preceded by has, a construction very frequently used to introduce news reported content (e.g. ... have *released* the ship); our analysis highlights the need for a focus on canonical reports of events. It also has the potential to learn negative indicators of event links, avoiding the interference of superficially similar candidates, where the current model relies on similarities and so cannot. A single term vector for each document would also avoid a weakness of the current system, wherein cosine similarity may assign undue weight to very small term vectors (e.g. for a specific zone and extractor) through normalisation. Directly learning term weights would obviate this problem while effectively learning zoning and query formulation from data.

7.5.2 Enhancing the training

Our evaluation demonstrates that some hyperlinks to online news are sufficiently like event links to train a model. How best to sample a corpus of hyperlinks and adapt the resulting model(s) to the target task are both open questions. These correspond to the fields of outlier detection and domain adaptation in machine learning, which we do not delve into.

However, we note that manually classifying hyperlinks as approximating to event links is – while by no means trivial – a far less intensive task than searching for event link targets, let alone also identifying references to link in an exhaustive annotation. It has the potential to be compromised by bias towards the existing links, but the alternative is compromised by the limits of annotators' ability, ingenuity and patience in searching. This may therefore present a fruitful and relatively inexpensive approach to producing event link corpora, which could further be enhanced by output of a preliminary system such as presented here. In future work we will consider such data collection for links to a readily available and globally notable target archive such as the New York Times.

[16] Our use of logistic regression is not especially configured for this type of feature either; a non-negativity constraint on the model coefficients and feature scaling may improve machine learning over non-negative cosine similarity features.

It may yet be worthwhile to produce further manual event link annotations, not biased by the distribution of reference style, authorial perspective and event type represented in acquirable hyperlink corpora. However, we found in Chapter 5 that exhaustive event linking annotation could be very time-intensive. A system trained on hyperlinks as presented here is already able to accelerate the annotation task by reducing human effort in searching for candidates. As more data is created, the system can be retrained with a larger proportion of manually-annotated training data, serving to iteratively improve annotation efficiency.

7.5.3 Completing and extending event linking

Our evaluation has only considered successfully linked event references from the gold standard annotation. Out of scope of the present work is the identification of *linkable* event references. Even among *linkable* events, many appear to be newsworthy but may not be reported in a particular archive, so a linking system must also consider outputting ∅. In retrieval approaches, a threshold on the score may suffice, though we note little difference between the distributions of top scores for *not found* and *linked* references from our corpus, in part because the distinction between these cases can be subtle. Hyperlink data sources notably do not include training knowledge for *not found*; yet we conceive of deriving some by limiting the target archive to only its most notable entries, such as front page content. This relates to a popular approach in NEL that empirically outperforms thresholding in Text Analysis Conference evaluations, wherein the target KB represents only a subset of Wikipedia's entities, so matches to excluded entities are indicative of ∅ (Hachey et al., 2013).

The present work also does not handle *reported here*, a category necessary for the archive-internal, exhaustive annotation setting. Our scoring model very frequently ranks the source article in the top position, and we explicitly exclude its candidature. Identifying that a reference in news is reporting the event can often be discerned from typical discourse constructs, as intended with the article zoning used rudimentarily in our system.

We are led to consider extensions to the linking task, not merely its solution, by extending the scope beyond currently *linkable* events or indeed, beyond events. The use of citation as training, and the similar function exhibited by hyperlinks in Fairfax data, suggests that many notable *facts*, not limited to events, could be grounded canonically in a news archive.

The *linkable*-or-not distinction was introduced in Section 5.1 to focus on references for which there is a single and hence canonical report. Yet the retrieval approach suggests that a reference to a set of *linkable* events could be linked to a set of canonical articles. This still needs to be highly specified, as the exact constituents of a *compound* event such as an election, or an *aggregate* such as an economic downturn remain indeterminate.

7.6 Conclusion

This chapter has applied quantitative and qualitative evaluation to assess the feasibility of the event linking task, the extent to which a specialised retrieval system suffices for its solution, and the applicability of noisy-but-easily-acquired training data to approximate event links in parametrising that system.

We take naïve approaches to acquiring corpora of hyperlinks to learn the model. Underlying this is the assumption that a substantial proportion of hyperlinks to online news correspond to event reference and thus approximate event links. We find this is true of about half the site-internal hyperlinks sampled from article bodies of one online news provider; simple rule-based filters remove a large portion of inappropriate links, mostly site navigation hyperlinks that are not truly anchored in body text. This leads us to extract a collection of hyperlinks (FD) almost two orders of magnitude larger than our manual annotation when only considering links with distinct targets. We obtain a smaller collection (WP) in which Wikipedia cites the same news archive and compare their utility for training an event linker.

Both sets of hyperlinks are predicted better when incorporating temporal information, distinguishing this news-oriented task from ad-hoc retrieval. When dates are mentioned in the context of a link anchor, they are often an effective indicator of the target's publication date, which again confirms that a number of hyperlinks reflect event reference, or at least temporally grounded knowledge. For the FD data where the link source is also assigned a publication date, we find a strong bias towards recent articles, which reflects a real bias in actual event reference, but seem stronger since they may be used to encourage a news reader to explore the news. Non-trivial term overlap measures also help predict these hyperlinks, although less so after temporal weighting is accounted for. The most effective terms are mentioned named entities resolved to Wikipedia and normalised date references; matching content in the lead sections of a candidate – where new material appears most prominently – improves upon a full text match. Ultimately, both temporal and term overlap features are effective in improving upon a simple similarity measure for predicting these hyperlinks.

Moreover, we find these gains transfer to event links. A WP-trained model augmented with recency features greatly outperforms a word stem overlap baseline, almost doubling its mean reciprocal rank. Although the system correctly retrieves 40% of link targets before other candidates, we find that it leaves much room for improvement: it retrieves on-topic articles without much preference for a canonical target; it poorly handles changes in referential perspective over time and terse references; and it fails to distinguish between event references in the source text, and to disambiguate referents that are textually similar. Hence while we verify that noisy hyperlink data can train a simple event linking model, our evaluation suggests that an event linker needs to go beyond traditional IR methods in order to achieve high performance.

Chapter 8

Conclusion

Although references to events are common in text and public discourse, they are difficult to characterise computationally, just as they are difficult for humans to consistently recognise and schematise. This thesis has explored existing and novel approaches to event semantics and reference. Its central contribution is a new proposal to canonically ground notable event references in a news archive, a task we call *event linking*. This is based on our understanding of news and related media's role in facilitating public discourse: through publishing knowledge of events they provide a shared foundation for effective communication. The focus on reference as grounding is also inspired by a recent development in the processing of named entity reference: the popular named entity linking task focuses on disambiguation with respect to a collaboratively-edited knowledge base of notable entities, leading our curiosity towards a parallel model for grounding events.

We have demonstrated the feasibility of event linking through a corpus annotation; modelling it through a retrieval system underlines the importance of temporal knowledge to the task, while highlighting a number of challenges that set it apart from the related task of topic tracking. We simultaneously illustrate that event linking knowledge is readily acquirable in the form of hyperlinks to online news, which are able to significantly improve an event linking model even with only minimal refinement.

Although it addresses different aspects of event understanding, the proposed task is informed by our analysis of the literature and resources in the area of event characterisation, and by our own experiments in annotating aspects of event reference. To highlight the relationship among these disparate areas of exploration, we provide a lateral review of this thesis's contributions and some of the directions they provide for further consideration.

8.1 Event typology

Information extraction has sought to extract language content selectively, grouping referents into discrete types, such as references to people, organisations, locations and management

succession events in MUC-6 (Sundheim, 1995). Previous work largely applies event typology to domain-specific corpora, while we explore its application to the set of events reported in broad-coverage news, contributing new analyses of existing resources and new experiments in type annotation.

Although we have exploited Wikipedia for the categorisation of other named entities in Nothman et al. (2013), an analysis of event topics in a sample of Wikipedia articles shows they are unrepresentative of the events covered in news (see Section 3.1), which prefers complex, structured events such as wars, and which is biased in quantity towards events that are notable but enumerable, such as sports contests. The ACE05 corpus attempts to annotate references to a variety of event types; yet the infrequency of many types, the diversity of events within them, and variability across types (see Section 2.5.1) limits their utility.

We attempt two annotation experiments with different approaches to typology: one with a pre-specified, flat typology, similar to, but coarser and broader than, ACE05's (see Section 3.2); and a novel approach employing a type hierarchy, which is constructed as annotation progresses, and which is abstracted from an orthogonal annotation of thematic domain (see Section 3.3). While the latter approach improves upon the former in terms of both coverage and type purity, event features for determining a hierarchy are far from definitive. Event linking allows us to characterise event references without types, although type by whatever definition might be inferred from the targeted news story.

Transforming the space of event types into a discrete, compact, broad-coverage and meaningful representation remains a challenge. In future work we intend to assess the applicability of existing lexical-semantic and ontological resources to describing the space of news events.

8.2 The relationship between news, events and notability

Both our story-driven annotation (Section 3.3) and event linking (Chapter 4) characterise each news report as a reference to one or more update events with respect to a topic, a novel extension to ideas behind Topic Detection and Tracking (Allan, 2002). This approach focusses on notable events, since varying perceptions of salience often causes disagreement in annotation at the (sub-)sentence level; we analyse this problem in the ACE05 corpus in Section 2.4, and in our own annotation in Section 3.2.

We observe in our story-driven annotation that a news report's topic is often ill-defined, and yet a journalist often explicitly refers to past events, connecting update content to existing public knowledge; event linking approximates grounding such references as well as references from non-news text. It treats a news source as a proxy for reported events, establishing a data-driven notion of event notability.

This work presents several directions for future exploration. Event linking annotation where references are randomly sampled from historical texts does not reflect the idea that

a news archive provides shared event knowledge; having existing readers annotate contemporary news would correspond better to this theoretical model. We also touched upon the relationship between topic and update events, where a systematic consideration might improve on existing approaches to event interrelation and script extraction by providing an explicit connection to discourse structure. Lastly, news updates often consist of non-event facts, for which a more general variant of event linking (i.e. *fact linking*) should be evaluated.

8.3 Approximation of event reference

Section 2.6 underscores the impracticality of event coreference identification, particularly between documents that present different perspectives on the same occurrence. Event linking considers the update content of a news report as the minimal unit of a grounded event; under the assumption that co-reported events – including their sub-events – are closely related, multiple expressions linked to the same article are approximately coreferent. We believe that this relationship is often more reflective of our perception of news events than strict identity alone accounts for.

While this novel coreference approach may often constitute a coarser approximation than referential ambiguity necessitates, its advantage lies in being driven by the news archive without depending on a particular analysis of its text. Yet it also fails to represent referents that are not a subset of a single news report's update events. Complete coverage of event references and the ability to specify a more precise relationship between a referent and a link target article are among possible enhancements to event linking's model of reference.

8.4 Eschewing experts for annotation

Expert linguistic annotation can be costly; exhaustive identification of references and full-text searching make event linking annotation in the manner of Chapter 5 particularly time-intensive. In addressing this problem, our work contributes new investigations into outsourced annotation and the use of free hyperlink corpora as training data for disambiguation.

While recent work has employed crowdsourcing for low-cost annotation, we consider an alternative outsourcing approach that allows for prolonged tasks and direct interaction with individual employees. This use of an online freelancing marketplace, while more appropriate for other tasks, is not well-suited to event linking annotation, since annotators can often afford only intermittent attention to the work. Since we do not know of reported linguistic annotation under this model, we contribute an initial discussion based on our experience, compared to traditional expert annotation and crowdsourcing. An empirical comparison of these approaches, and an evaluation of non-expert freelancers working with existing linguistic annotation schemas of varying complexity, seem worthwhile for future consideration.

The effort of our event linking annotation resulted in only a small corpus: suitable for evaluation, but insufficient for statistical modelling. Nonetheless the insight that some portion of hyperlinks targeting online news correspond to event links makes acquiring vast quantities of event link training data, albeit noisy, feasible. We find that – after minimal filtering based on a sample analysis – hyperlinks within an online news web-site and citations from Wikipedia to news are both able to train event linking models that significantly outperform the baseline. Future work will consider the effect of training noise levels on model quality, comparing manual and automatic verification as the means of preparing such corpora.

8.5 Conclusion

Our work proposes and analyses new models of event reference characterisation. Following on from insights into event salience, typology and identity, we establish *event linking* as a general representation of reference to newsworthy events and a news-oriented model of their grounding. Although we have shown that easily-acquired corpora of hyperlinks can be harnessed to train an event linking system, accurate resolution of event reference remains a distinctly challenging task.

Appendix A

Schemas and typologies for exploratory annotations

A.1 Type-driven event annotation guidelines

By Joel Nothman and Matthew Honnibal, version 2011-02-03.

These guidelines apply to the task described in Section 3.2.

A.1.1 Purpose

We would like to be able to identify when certain types of events occur, so that, for instance, a property investor can track events such as new building developments.

Previous attempts at event annotation have produced low inter-annotator agreement, and we intend to rectify this by increasing the granularity of the task.

A.1.2 What is an event?

Philosophers, linguists, etc. don't really know. We therefore adapt Nadeau's (2008) definition of NER:

> The words recognized as events are any that realise an event type in our scheme

Annotators are therefore instructed to pay close attention to the event types in this scheme, and then find and annotate any expressions that refer to an event covered by the scheme.

Events tend to be identified by occurring over a particular time and place. An event can be referred to multiple times in a document and in many ways, just as entities can, and in some cases events too are named.

Events require a change of state. Descriptions of an entity *remaining* in its state over some period of time, or being passively affected, are not considered events.

A.1.3 Textual granularity

We intend to evaluate event detection on a per-sentence basis. Nonetheless, annotations should try to pinpoint the word (often a verb or a nominalisation of a verb) or phrase that best indicates that the event has occurred, facilitating an eventually more nuanced discrimination of events.

However, if one sentence or phrase or word denotes multiple events, it should be tagged with all of them. For example, as *murder* indicates both an attack and a death, one could mark e.g. "mur-" as one event and "-der" as another! Our systems will understand this as tagging the same word.

A.1.4 Event types

We are using broad event types that are almost thematic. But be sure to label according to the event occurring, and not merely by the article's theme. For example, a footballer contracted onto a team is being employed, it is not a sports event.

Lifecycle inclues:

- Birth, death
- Marriage, divorce
- Illness, injury, surgery

Organisation lifecycle includes:

- Establishment
- Mergers, acquisitions, splits
- Restructures
- Privatisation
- Bankruptcy
- Liquidation
- Name change
- Significant expansion or reducing of operations

Employment/award includes:

- Hiring, contract offer
- Nomination, election (but not the voting, just the result)

- Titling, awards
- Firing, resigning, retiring

Conflict includes:

- Attack / battle (personal, or in war)
- Police raid
- Demonstration rallies
- Boycott
- Surrender
- Crime
- Lawsuits and class actions (but not their hearings) (!)

Justice includes:

- Arrest, jail, charge, trial
- Convict, acquit, sentence, appeal, pardon
- Execute, extradite, fine
- Parole
- Subpoena, warrant issuance

Applies to any jurisdiction including sports leagues.

Excludes lawsuit which falls under Conflict, although the *hearing* of civil suit cases is a Justice event

Governance includes:

- Legislation / ban
- Regulation

Applies to any jurisdiction including sports leagues.

The focus here is on the creation of legislation and related executive plans / commissions. This does not the *enforcement* of legislation which is **justice**.

Disaster includes:

- Epidemic
- Car crash

- Financial crisis
- Work accident
- Natural disaster
- Personal disappearance

This category, like conflict, consists of events that affect people and infrastructure, but unlike conflict do not have an intentional cause

Sports includes:

- Match (coreferent with win, lose, draw), rematch

Finance includes:

- Stock price / currency rise/fall
- Interest rate change
- Shareholder relations: Share issuances, dividend, stock split

Significant changes from within an organisation, even if they only affect the financial structure of the org (e.g. demutualisation, privatisation, initial public listing), and not its internal structure, are to be considered Organisation Lifecycle events. Apart from shareholder relations, Finance events affect organisations from the outside.

Real estate / development includes:

- Construction
- Demolition
- Design
- Restoration

New release includes:

- Product release
- Movie, book, album, TV release
- Technology & scientific innovation
- Study results, inquiry reports
- Excavation finds

Note that the *release* or *publication* of these items is the event we are interested in. Statements describing their contents are not events.

Excludes press releases or company announcements.

Correspondence includes:

- Meeting (esp. between delegations)
- Writing and phoning (esp. official correspondence)

Transaction includes:

- Payment (or a bid to pay) in exchange for goods
- Goods in exchange for goods
- Bidding or agreeing on a contractual relationship (excluding hiring personnel)
- Agreement to share assets
- Donations

Excludes unilateral change of ownership, such as *capture* or *theft*

A.1.5 What counts as an event

The notion of event is affected by tense, aspect and figurative language

Please tag all mentions of *specific events*, but mark as "realised" those that the article indicates have happened or are happening.

We ignore **generic** mentions of would-be events, like "many terrorist *attacks*" or "banks have always been reticent about *cutting dividends*". These examples discuss a class of event without referring to any instance of the class. Note that an event doesn't have to have happened or be going to happen to be marked, so long as an event instance is referred to. For instance, "the bank has decided not to cut its dividends", or even "ANZ is discussing cutting its dividends".

Figurative language should not be taken by its literal sense. We are interested in the event referred to by the expression.

A.1.5.1 Substitution tests for aspect and figurative language

Given the sentence:

> Can AMP survive as an independent group or will it **become a victim** in the inevitable rationalisation that will sweep the insurance industry?

it is difficult to decide whether *become a victim* should be marked as an Organisation Lifecycle event (i.e. a takeover). If we substitute less figurative language, we get:

> Can AMP survive as an independent group or will it be **taken over** in the inevitable rationalisation that will sweep the insurance industry?

it is still a little unclear whether this is a specific event because of the rhetorical interrogative context. (It is clear the event is not *realised*.) If we alter the sentence to a declarative form:

> AMP cannot survive as an independent group and will **become a victim** in the inevitable rationalisation that will sweep the insurance industry.

or further, into a sense where the event has been realised:

> AMP did not survive as an independent group and **became a victim** in the inevitable rationalisation that swept the insurance industry.

it now seems clear that the event should be marked.

A.1.6 Coreference and part-of

If two mentions refer to the same event, they should be marked as part of the same chain. To help identify events that are coreferent, consider whether they occur at the same time and place, and with the same participants (e.g. the same victim, weapon, attacker).

We also allow you to mark events that are part of other events. Note that this does **not** include events *caused* by other events, but for instance might be used for an attack (Conflict) event that consists of a verbal insult (Conflict) event and a physical assault (Conflict) event. Similarly, one Sports match event may be part of a larger Sports match, i.e. a tournament.

A.1.7 Which articles do we annotate?

We are interested in only news-type genre, particularly when written in the inverted pyramid style. We are not interested, for instance, in product reviews or sports betting summaries.

We have automatically excluded articles from sections [‘Metro’, ‘Domain’, ‘Spectrum’, ‘Travel’, ‘Traveller’, ‘The Guide’, ‘My Career’, ‘Drive’, ‘Good Weekend’, ‘Money’, ‘Good Living’].

A.2 Story-driven event annotation typologies

The following are dynamic typologies constructed for event domain (Figure A.2) and type (Figure A.1) for and during the story-driven event annotation task (Section 3.3). Note that these are produced organically and would in practice be refined after the annotation as we have done for named entity types (Nothman et al., 2013).

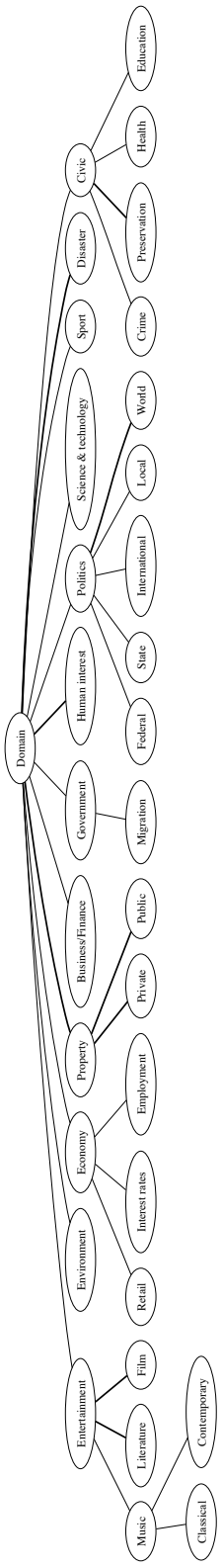


Figure A.1: Hierarchy of event domains. Bold arcs were added during the annotation; others were initially provided.

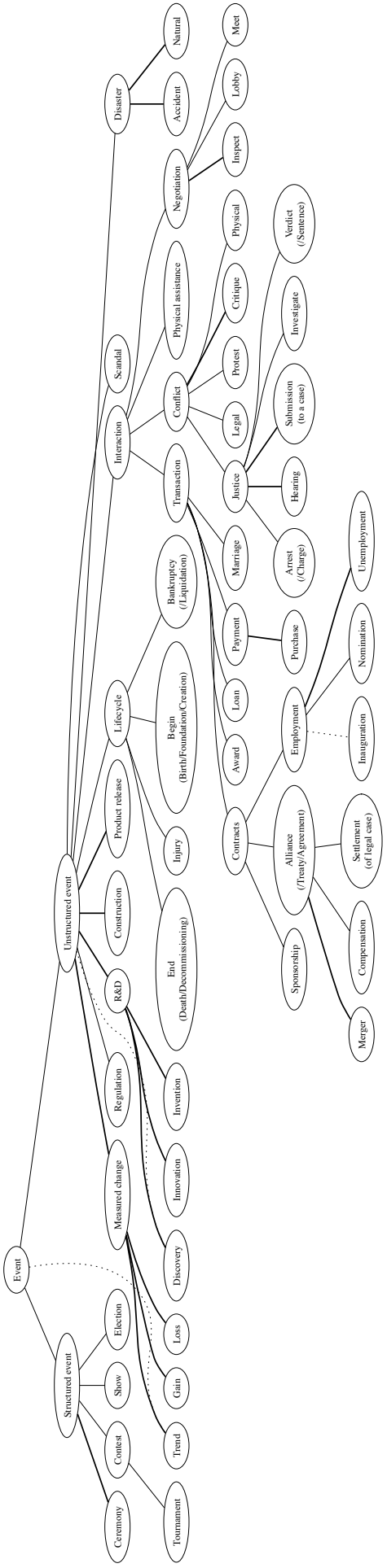


Figure A.2: Hierarchy of event types. Bold arcs were added during the annotation; dotted arcs were deleted from the initial hierarchy; others remained throughout the annotation.

Appendix B

Annotation schemas for event linking

This appendix reproduces the annotation schemas for our pilot (Section B.1) and final (Section B.2) event linking annotations, as well as the worked example (Section B.3) employed for annotator training (refer to Chapter 5).

We note some differences in category labels between the schema and the body of the thesis in Table B.1.

In schema	In body
Basic event	*linkable*
Complex event	*compound*
Trend or measured change	*aggregate*
Many / multiple / generic	*multiple*

Table B.1: Equivalence between labels used in the schema and the main work

B.1 Pilot annotation schema for event linking

Task: *link each expression which refers to a newsworthy event that has happened (from the perspective of the expression's author) – and which the informed reader would know of from previous news – to an article representing that event.*

The idea is to decode language which refers to events as a pointer to something in the reader's knowledge/memory of events, assuming that the reader has read and remembered any event in the archive that someone might later refer to.

What to mark for linking

Ideally, one is to mark all expressions that refer to newsworthy events the reader is likely to have known about (i.e. from previous reporting of that event). But defining what constitutes an event for this purpose, both textually and semantically, is difficult.

Events that are first reported as news in the article being annotated should not be marked. (Even though they are referring expressions, we will concern ourselves only with the communicative pragmatics of referring to a previously known event.)

Extent

We annotate expressions which denote a (newsworthy) event that happened. We are particularly interested in expressions the author uses to refer to events that a regular reader would know about from prior news.

Generally, annotation spans will be:

- a *single* head verb or noun
- which *predicates* the event / bears the event-denoting *content* (not merely tense, etc.)
- for rigid designators and proper nouns, annotate only the head

Auxiliaries (e.g. has/had in *X* attacked) should generally not be marked.
The following should not be marked:

Implicit references

Quasi-events Expressions referring to change in some measured value (e.g. prices rose)

Introduced here Expressions denoting events which are being reported first in the current article (i.e. they are not being used to refer back to existing knowledge). If at first you aren't sure if an event was introduced in the present article, mark it, and if you conclude that it was, indicate such under "why no link?".

The following should be marked, but not linked (see below):

Plural Expressions which refer to many event instances that would have been reported individually, such as the recent wave of attacks... attacked on Monday and again on Tuesday, or his spate of injuries. While these may refer to a specific array of past events, we have no way to encapsulate them in our knowledge base. (Nonetheless, Sunday's demonstrations may be linked despite events occurring over a multitude of locations.) This also includes another in residents fear *another* attack (implies a reference to a past event in addition to hypothetical events), and complex events which would be difficult to pinpoint to a particular article, such as an election or sports season.

Semantic class

A relatively precise discussion of event-referring language is given in the TimeML event annotation guidelines (Saurí et al., 2009) which divides all verbs and other event predicates into:

REPORTING say, according to ...

PERCEPTION see, hear ...

ASPECTUAL begin, complete, (dis)continue ...

I-ACTION try, cancel, investigate, order, nominate ...

I-STATE feel, desire, fear, require ...

STATE have ...

OCCURRENCE arrive, explode, distribute ...

Generally we are interested in events that TimeML would class as Occurrence and Aspectual. So for our purposes, generally:

- someone saying something is a non-event
- someone hearing or seeing something is a non-event
- someone having something is a non-event
- someone feeling or wanting something is a non-event
- someone proposing something we still haven't decided on...
- someone continuing to do something is a non-event, despite being driven by an aspectual event

There may be exceptional cases, where one of these activities was in itself a newsworthy event, in which case it *should* be marked, and a note made to that effect in the comments.

Nonetheless, we are less concerned with these *semantics* than with the *pragmatics* of how the expression is used to refer to a known and newsworthy past event.

Coreferring expressions

Mark all expressions in *separate sentences* which refer to the same event as coreferent. However, unlike in other event annotations, coreference does not need to be precise, as long as the events have the same link.

What to link to

Treating the archive as a knowledge base

We are only interested in linking to stories that introduce (to the reader's knowledge) one or more events. As such, link targets should always be *news articles* designed to report events (and the event reported should exist outside of the article).

We only consider the first article which refers to the event *having happened*. Articles which posit the event's future occurrence are not canonical representatives of that event in the knowledge base.

Not found in the archive (kb)

An event reference may not refer to an article in our knowledge base, either because it precedes our archive, or no appropriate article exists.

Mark such expressions, and instead of linking, select the appropriate response from the "why no link?" drop-down.

"Why no link?"

There are some cases when expressions should be marked but not linked, some of which have been discussed above. We provide a widget to indicate why no link is given:

Plural reference the expression can't be pinpointed to a single reporting

Not found the event should probably be in the archive, given its time and topical scope, but was not found.

Precedes the archive the SMH archive only goes back to 1986, so nothing can be linked before then

Introduced here indicates that you had thought the expression might refer back to past news, but in fact you found it was first reported in the present article.

B.2 Final annotation schema for event linking

We are developing a system which:

- identifies when news refers to background events that a regular reader would already know about; and
- finds the article which originally reported that event happening.

To develop this system, we need many example documents marked up by humans. (Note, however, that the task we set out below is not identical to what the system described above would produce.)

Events as seen by news

We are interested in the sorts of events that make news. Throughout this document, we talk about *newsworthy events*. These are, roughly:

> Things that happen and directly trigger news.

News articles are often directly triggered by one or more related events, which they *report* or *break*. In discussing those trigger events, they also refer to *background events*, which a regular reader of the news may know about (because there was an earlier article reporting it).

Sometimes news refers to events that are hypothetical or yet to happen. As far as this work is concerned, *an event is not real until it happens*.

We also categorise references to newsworthy events, based on how they are reported: whether in a single day's news (*basic*); across multiple articles, but altogether one event (*complex*); across multiple articles about disparate events (*many/generic*); or events which only 'happen' by calculating or aggregating some value over time, such as stock price changes or fashion revivals (*trend/measured change*).

Using this document

This document describes our approach to annotating references to past events in news. First we describe the annotation procedurally as a list of steps; following, we give descriptions and/or examples of the cases described (and linked-to) in the procedure.

Procedure

By carefully reading through a document, you mark words referring to events, classify each into one of a few categories, and find an article to link to when appropriate.

1. Find an event-denoting expression
2. *Ignore it* if it is:
    - Future hypothetical or uncertain
    - Not newsworthy
3. Otherwise:
    - *Select a single word expressing the event*
    - If you have already marked another mention for the same event (or a closely related event first reported in the same article), mark the new mention as part of the same event.
    - Otherwise, mark the new mention as a new event.
4. *Select a category for the event*:
    - Basic event – probably first reported in one news article
    - Complex event – likely to have multiple articles; often a named event
    - Trend or measured change (*Trend/change* in the annotation tool)
    - Many events or a generic reference (*Many/generic* in the annotation tool)
5. If a *basic event*:
    - Try to link to the article first reporting that event as having happened/begun
    - Or mark as:
        - First reported here
        - Precedes 1986
        - Not found, which includes: No mention in archive, Not reported in archive, Not reported after occurrence
6. If a *complex event*:
    - Try to link to a Wikipedia article specifically about the event
    - Or mark as Not found.

Key points to remember

- Mark and categorise *every* explicit reference to a past or ongoing newsworthy event.
- Label a basic event as *reported here* if the article being annotated is reporting that the event has happened or is ongoing.
- Try to link each other basic event to the first article where it was reported as ongoing or happened.
- Link complex events to Wikipedia only when there is an article directly on that topic.
- If you're uncertain, just choose what you think is best, and leave a comment!

Seek event-denoting expressions

These may be nouns (attack), verbs (attacked), proper names (World War II), or occasionally other forms. For various examples, read this document!

Many concepts have aspects that are event-like, but themselves are not events:

- References to publications including legislation or articles should not be considered events, despite a date being specified, as in Patents Act 1990 or Column 8, last Tuesday or Manning and Schtze (2002). However, the text may refer to markable events related to legislation being debated, passed or amended, and publications being released, etc.
- A court case, however, is an event.
- Some named events, such as Olympic Games or Wimbeldon, can refer to both a group of periodic events, and to the particular events (e.g. the 2000 Olympic Games). Only mark expressions which refer to particular events.

Sometimes, an event will be implicit, but unless there is an expression indicating that an event happened, it should not be marked. For example, in A wind farm big enough to power 25,000 homes will be built, the news is implicitly reporting that a decision has been made to build a wind farm; do not mark an expression to represent this implicit decision event.

Which events should be ignored

Future, hypothetical or uncertain

Only mark events that have happened or are ongoing. Do not mark future events, or ones which uncertainly happened (i.e. are not reported as fact). Examples:

- A wind farm big enough to power 25,000 homes will be *built* near Glen Innes in the New England tablelands, Do no mark built, because the building is happening in the future.
- the first of several huge renewable energy plants *planned* for the region. Mark planned, because the planning happened in the past
- Police will investigate if a Mercedes-Benz stolen from the jockey Nash Rawiller was *used* as the getaway car after a cash-delivery van was rammed and shot at outside Sydney Airport. Do not mark used, because the if means we do not know whether the car was used; in contrast, we know that a cash-delivery van was rammed and shot at, and these should be marked.
- Kiesha Weippeart was thrown against a wall or the corner of a bed before she died, police believe. Mark died, because it certainly happened, but not thrown because it is alleged, not certain.
- . . . renewed hope for an *agreement* in Washington to *raise* the countrys debt ceiling. Do not mark raise, because the sentence expresses *hope* that the event will occur, but does not state that it has occurred.

Not newsworthy

We are interested only in newsworthy events like purchase, arrival, explosion, eruption, beginning and cancellation (though other word forms like arrived, exploded, erupted, began, cancelled

are just as good), whereas someone feeling, saying, seeing, having, continuing or suggesting something is usually not a newsworthy event, though there are exceptional cases.

Some examples and counter-examples:

- Although He *told* the Herald he had already... should not be marked, the same verb could be newsworthy when it deals with important subjects, as in Packer *told* Leckie to "f— off". This is more clear in its context because the same article describes Packer's action as an attack. Similarly, when a hearing is *told* some alarming or new information, this may be newsworthy in itself.
- Although Iran *bought* warheads from Russia is newsworthy, my wife *bought* a new car is not the sort of event we expect to be reported in the news, and should not be marked.

Coreference

When marking a mention, the annotator can either create a new event [chain] or add the mention to an existing chain. We say that the mentions together in an event [chain] are *coreferent*.

Mentions should be marked coreferent if they refer to the same or closely-related events (often happening at the same place and time, or one causing the other) *and* if their attributes (event category, linked article, etc.) would have the same value.

Select a single word expressing the event

When selecting a single word to mark, start by looking for the word that indicates that the event happened. Often this is an obvious verb (e.g. attacked, began), or a noun (the *attack*, 2000 Olympic *Games*, *it*). When it is not obvious, consider the question: if you had to choose one word to abbreviate the whole event, which would you choose?

For example:

Event expression	Best word
spent two years sailing	sailing
achieved a legal victory	victory
copped a bollocking	bollocking
made his opinion known	opinion

Most of these are cases of *light verbs*, words like make, get, take, give, have, achieve, seem, which get a lot of their meaning from their context, and would rarely by themselves describe an event of interest.

However, you should also take care to avoid ambiguity: try not to select a word that could be misunderstood as referring to a different event:

Event expression	Explanation
she was put on the job of ...	put describes the employment event; don't mark job because it might be misunderstood as a different event.
he began working on...	began describes a commencement event; you might mark working as a separate event reference if it is a newsworthy event.

Two closely related events mentioned as separate event expressions should both be marked as part of the same event (see Coreference), e.g. The Swifts have made the *coup* of the off-season, with the *signing* of England co-captain Sonia Mkoloma.

In the following examples, words that do not explicitly describe the event are the best candidates:

- In recent years, NSW have *lost* defenders Liz Ellis (*retired* 2007), Selina Gilsenan (*retired* 2008) and Mo'onia Gerrard (*Adelaide*).: The word lost incorporates 3 events, all of which should be marked. (Lost refers to many specific events, namely Ellis and Gilsenan's retirements and Gerrard's relocation to Adelaide, each of which are basic events.)
- . . . a few months ago in South *Africa*, Johnson seemed invincible: There is an implied reference to a sporting event in South Africa.

Try to label all appropriate expressions, but *Basic event* and *Complex event* expressions are most important to us.

Event categories

We make a distinction between a few categories of event, based on how they are reported.

Basic event

A *basic event* is the sort of event that would be first reported in one news article (although sometimes covered in multiple stories on the same day).

A sports match is a usually basic event, while a sporting series/season/tournament is complex.

Some examples:

- She was *arrested*. . .
- Last year they *entered* into a joint venture. . .
- He held the *winning* bid. . .
- She was *hired* in 2000 to. . .
- The government's *introduction* of the policy. . .

Complex event

A *complex event* consists of many smaller basic events, which would usually reported over many articles over many days, but which still can be referred to as one event. For example:

- the Montreal wipeout led to the *creation* of the Australian Institute of Sport: the creation of an organisation involves a number of different events, especially when it is created by an act of government.
- Johnson's Ashes tour should be understood as including a reference to that year's Ashes (cricket tournament) series.
- among the hardest hit in the *financial crisis*

Named events are often complex events (although the names are generally invented long after the events), such as:

- elections (Australian 2007 federal *election*)

- multi-day sports events and conferences (Sydney 2000 Summer *Olympics*)
- scandals (*Watergate*)
- periods of economic activity (Great *Depression*, Gold *Rush*).
- court cases

However, some names refer to periodical events, and should only be marked as a complex event when they refer to one specific event. For example, in won the Super Bowl five times, Super Bowl refers to multiple events, while in won the Super Bowl last year it refers to a single complex event.

Sub-events of complex events may also be mentioned, so Kevin Rudd *won* the 2007 federal *election* includes a basic won event, and a complex election event (which might include campaigning, polling and results determined over time).

Trend or measured change

News often refers to the change in measurements over time. While these are facts that can be referred to, rises and falls in values are not clearly events, but a reference to some metric over a period. They may describe a trend that can be summed from the events reported in numerous articles (such as records in sport), or could refer more to external records of metrics (such as weather gagues or stock tickers).

The following should be marked as trend/change:

- Shares *rallied* today, with strong *gains* across the board *adding* $23 billion to the market's value (all marked coreferent)
- The 30-year-old goal defence, ... will boost the Swifts ... after *finishing* ninth this year
- ... during his career, where he *took* 21 wickets in five Tests ...
- BHP said output of steel-making or coking coal *jumped* 19 per cent from the previous quarter to 7.9 million tonnes
- The US economy*grew* a modest two percent in the third quarter.
- It is estimated that more than $200 billion worth of projects in Dubai have *sunk* beneath the sand
- ... during the folk music *revival* in the 1950s
- a *doubling* of terror attacks (note: attacks should be marked separately)

Many events or a generic reference

These are expressions which refer to many events in many articles, or some unspecified past newsworthy event. These could not be pinned down to one article (except by picking a prominent example), and as opposed to complex events, the many events do not together constitute one event. For example:

- The drone *attacks* on Pakistani territory have continued since Barack Obama became president in January.
- ... during the Depression and *the world wars*.

- As coach of the women's team that *won* back-to-back gold medals . . .
- One mitigating factor is the continued weak *demand* in the aviation sector
- . . . how *spending* on Olympic sports has contributed to reductions. . .
- . . . financial counsellors were *given* mental health *training* . . .

Link to the archive

The article you select to link to *must* ***report*** *the event*, it must be *after* the event has begun/occurred, and it should be the *first time* that event is reported as having begun/occurred in the archive. By *report*, we mean that the event is described like it is new information (i.e. we seek the article that *breaks news of the event*). Please do not link to documents that are opinion pieces (letters, editorials, op-eds), feature articles or reviews. However, there are sometimes multiple articles published *on the same date* which all seem to report a particular event: choose whichever you think is most relevant.

We provide a tool to search through the news archive from one news source. We are only interested in linking to news articles within that news source. You can type search queries just as in search engines like Google (though our system may not be as clever as Google in interpreting your query). *Hit enter* to search.

Query keywords As in other search engines:

- generally, all the keywords you enter need to appear in the articles
- a query like `dogs OR cats OR mice` will instead search for any of the words {`dogs`, `cats`, `mice`}.
- a query like `cats -dogs` will find all articles containing `cats`, but not dogs
- a query like `"cat sat"` will require the words `cat` and `sat` to appear next to each other as one phrase

Date constraints You are also able to constrain the results by the *date* when they were published. You may require the result to be *in* a particular year, month or day; *in or following* (`>=`) a particular year, month or day; or *in or preceding* (`<=`) a particular year, month or day. If you know when the event happened, and constrain the date, *it makes finding the correct article much faster.* When reading the article, you should consider how you might constrain the date:

- Look for absolute date expressions, like *in 2007.*
- Look for relative date expressions, like *one year ago*, and work out the date given the publication date of the article you are annotating, shown at the top of the annotation tool.
- Search for the event in Google/Bing/Yahoo/etc. or Wikipedia. It is often easy to find out when an event occurred by using these external search engines.

Note that we keep a record of the queries you enter (into our private archive search tool only) because we think they might be useful in training an artificial intelligence system to search for event articles. For this reason, if you find an article in Google that you expect to be in our news archive, please *do not* just copy its headline and search for it in our in-built archive search engine. The article may also have a different headline in our archived version to what is found online.

Result order You can also *order* search results so that you see the oldest or newest archive articles first:

- *oldest* is useful because we require that you select the first article reporting an event.
- *newest* is useful when you expect there to be lots of search results for your query terms, and for the relevant article to be among the most recent.

First reported here

Events that are breaking news in this article should be marked, but not linked.

Example: Leighton announced yesterday that partners Al Habtoor, Murray & Roberts and Takenaka had mutually *agreed* with the Dubai Department of Civil Aviation to *withdraw* from the airport project due to the parties' "inability to conclude an acceptable contract".

If an event is mentioned as if it's background to the present story, do NOT mark it *Reported here*. If it is the first time you can find it mentioned in the archive, instead mark it *Not found*.

Precedes 1986

Our Sydney Morning Herald archive only goes back to 1986. Any basic events that precede that date can't be linked.

Not found in archive

This applies to events that you expected might be reported, but can't find an appropriate article, for one of these reasons:

No mention in archive The event is not at all mentioned in this news archive, but the present article treats it like it's background.

Not reported in archive The event is mentioned in the news archive after it happened, but only in feature articles, opinion pieces, etc., not being reported as news.

Not reported after occurrence The event is mentioned in the news archive, but only before it actually happened.

Link to Wikipedia

Use our linking search to find a Wikipedia article whose topic is *specifically* the event in question. When there is no article *specifically* about the event, mark it *Not Found*.

The event should be found by searching the provided copy of Wikipedia, which is now a few years outdated by the current Wikipedia. However, clicking on a search result to view the Wikipedia article will show the current version of the article, which may have been deleted, changed name, or merged with another article. Try to consider only the older version supplied.

When linking a reference to the Sydney 2000 Olympics, link to *2000 Summer Olympics*, not to *Summer Olympic Games* or *Olympic Games*. In some cases, this means that an event which is mentioned in Wikipedia but which does not have a dedicated article will have to be marked *not found*.

Not found in Wikipedia

Many complex events will not be mentioned in Wikipedia; in other cases, there will not be an article dedicated to the event. In these cases you should indicate that the complex event was *not found*.

For example:

- the Montreal wipeout led to the *creation* of the Australian Institute of Sport: while Wikipedia includes an article on the Institute, which refers to its establishment, it is not an article specifically about the creation of the institute.

B.3 Worked example for event linking annotation

While familiarising yourself with the Event Linking annotation task, it may be worthwhile following this broad approach:

1. Skim-read the article, getting a general sense of the events it discusses
2. Work out what is being newly reported, and what is background information (if it is not a news article, there should be nothing reported)
3. Work out which of the events discussed would probably be newsworthy (i.e. they might trigger news articles)
4. Ignore any events which haven't happened at the time of writing
5. Mark each event expression, either joining an existing event, or deciding whether it is basic/complex/trend/many
6. Look up any dates of useful events, and link, using those dates to help search

We illustrate this briefly with the *Deadly virus breakthrough* article[1] as an example.

What is the breaking news?

In this article, the new event being reported is the breakthrough being published in Science (i.e. the successful results of a research project).

More generally, it might be useful to put the groups of events mentioned in the article in chronological order:

1. the research was conceived/proposed/begun (A)
2. funding by the Gates foundation (B)
3. funding by the NHMRC (C)
4. the research was conducted/continued (D)
5. recent cases of dengue fever (E)
6. a breakthrough, and a publication of results (F)
7. some future halting of dengue fever (G)

[1] At `http://newsstore.smh.com.au/apps/viewDocument.ac?docID=SMH090102C067A1H0JKI`

What's likely to trigger an article?

The research proposal (A) might have triggered an article, as might its funding (B, C). The research process itself (D) is not really an event that would trigger news. The recent cases of dengue fever (E) may have triggered many articles. The breakthrough (F) has triggered the present article. And the halting of dengue fever (G) hasn't happened yet, and so should be ignored.

Marking words and categorising them

I would choose the following words in the given paragraph numbers, and mark them with the event corresponding to the given letter:

- (1) mark breakthrough as F (category: Basic Event, Reported Here)
- (3) mark grant as B (category: Basic Event)
- (3) mark used as F
- (4) mark gave as B
- (4) mark cases as E (category: Trend/Change)
- (5) mark published, breakthrough as F
- (6) mark began as A
- (8) mark proposed as A
- (9) mark donation as B
- (9) mark followed (or from) as C (category: Basic Event)
- (12) mark spending as B

Linking the events

We label F as *Reported Here* and proceed to try link to articles reporting events A, B and C.

It is very useful to know the date an event occurred when searching for the article that reported it. Event A happened ten years ago. Since the present article was written in 2009 (see the top of the screen), we assume any article reporting the proposal of this work was no earlier than 1999. Event B happened in 2005, and event C happened after that. (It is common for an article to refer to an event without a clear indication of when it happened; in such cases, it may be worthwhile to seek the event date using Wikipedia or Google.)

If we search for `"dengue fever" o'neill` $\geq$ `1998` we only find one article, *Mozzies' days numbered if Gates grant does trick*.[2] This article is clearly a good match for event B, but is unlikely to be breaking news of A, given that it happened 6 years before. It is worthwhile to search for other terms in place of `O'Neill`, like `"university of queensland"`, but no earlier or later article is found discussing this research (though other dengue fever research projects are discussed). So we link event B and mark events A and C as not found (noting that the latter article doesn't report event C either).

[2] At `http://newsstore.smh.com.au/apps/viewDocument.ac?docID=SMH050629E21EU6BUS95`

Appendix C

Detailed event linking corpus statistics

C.1 Inter-annotator confusion over token categories

This section provides pairwise inter-annotator confusion statistics in double-annotated portions of our annotated event linking corpus. Here we do not account for (dis)agreement on the target of any links identified. These statistics exclude annotator training documents.

The categories available to label each token are are described in Section 5.1, but are abbreviated for economy of space according to following legend.

Abbreviation	Annotation category
—	Unmarked
L:RH	*linkable*:*reported here*
L:NF	*linkable*:*not found*
L:P	*linkable*:*precedes archive*
L:L	*linkable*:*linked*
C:NF	*compound*:*not found*
C:L	*compound*:*linked*
M	*multiple*
A	*aggregate*

Table C.1: Legend of annotation category abbreviations

A \ B	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**39833**	86	23	10	39	16	16	55	20	40098
L:RH	379	**70**	5	0	7	4	0	5	5	475
L:NF	75	12	**16**	0	3	3	0	2	3	114
L:P	5	0	0	**12**	0	0	0	0	0	17
L:L	245	16	8	0	**64**	5	4	14	1	357
C:NF	24	4	1	0	1	**3**	0	0	1	34
C:L	74	1	2	1	1	1	**7**	0	0	87
M	144	6	0	0	0	0	0	**22**	0	172
A	35	0	0	0	0	0	1	1	**24**	61
Total	40814	195	55	23	115	32	28	99	54	41415

Table C.2: Token-level inter-annotator confusion between A and B

A \ C	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**14825**	33	8	0	13	38	10	16	2	14945
L:RH	103	**38**	1	0	2	6	0	0	0	150
L:NF	16	5	**7**	0	1	9	0	0	0	38
L:P	2	0	0	**0**	0	0	0	0	0	2
L:L	61	6	2	0	**15**	18	4	3	0	109
C:NF	14	3	0	0	0	**4**	0	0	0	21
C:L	37	0	0	0	0	3	**5**	0	1	46
M	84	1	0	0	0	5	0	**2**	1	93
A	11	2	0	0	0	0	0	0	**8**	21
Total	15153	88	18	0	31	83	19	21	12	15425

Table C.3: Token-level inter-annotator confusion between A and C

B \ C	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**4080**	16	7	0	3	13	4	0	0	4123
L:RH	24	**6**	2	0	1	7	0	0	0	40
L:NF	2	1	**2**	0	0	4	0	0	0	9
L:P	0	0	0	**0**	0	0	0	0	0	0
L:L	2	0	0	0	**3**	0	0	0	0	5
C:NF	1	0	0	0	0	**4**	0	0	0	5
C:L	1	0	0	0	0	0	**4**	0	0	5
M	4	1	0	0	0	1	0	**1**	0	7
A	4	4	0	0	0	0	0	0	**3**	11
Total	4118	28	11	0	7	29	8	1	3	4205

Table C.4: Token-level inter-annotator confusion between B and C

J \ A	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**81117**	454	68	22	166	40	101	235	19	82222
L:RH	127	**442**	37	0	10	3	4	23	1	647
L:NF	53	8	**86**	0	24	4	5	4	0	184
L:P	16	0	1	**19**	0	1	2	0	0	39
L:L	110	8	16	0	**322**	0	4	9	1	470
C:NF	33	6	11	1	18	**24**	29	2	0	124
C:L	25	0	0	0	16	0	**41**	0	1	83
M	91	8	8	1	27	2	10	**161**	5	313
A	39	12	3	0	6	7	7	13	**88**	175
Total	81611	938	230	43	589	81	203	447	115	84257

Table C.5: Token-level adjudicator-annotator confusion between J and A

J \ B	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**40235**	55	16	3	21	3	8	22	11	40374
L:RH	189	**104**	7	0	2	1	0	4	2	309
L:NF	53	11	**23**	0	2	0	0	0	0	89
L:P	4	0	0	**19**	0	0	1	0	0	24
L:L	154	15	5	0	**81**	1	1	9	1	267
C:NF	31	0	3	1	2	**19**	0	0	0	56
C:L	19	1	1	0	1	3	**14**	0	0	39
M	83	7	0	0	3	5	1	**56**	1	156
A	46	2	0	0	3	0	3	8	**39**	101
Total	40814	195	55	23	115	32	28	99	54	41415

Table C.6: Token-level adjudicator-annotator confusion between J and B

J \ C	—	L:RH	L:NF	L:P	L:L	C:NF	C:L	M	A	Total
—	**15012**	19	5	0	6	23	2	9	0	15076
L:RH	40	**54**	3	0	2	6	0	0	0	105
L:NF	7	0	**7**	0	3	9	0	0	0	26
L:P	1	0	1	**0**	0	0	0	0	0	2
L:L	35	9	1	0	**19**	14	1	1	0	80
C:NF	12	1	0	0	1	**20**	0	0	0	34
C:L	12	0	0	0	0	2	**15**	0	0	29
M	22	1	0	0	0	8	1	**11**	1	44
A	12	4	1	0	0	1	0	0	**11**	29
Total	15153	88	18	0	31	83	19	21	12	15425

Table C.7: Token-level adjudicator-annotator confusion between J and C

Bibliography

Kareem S. Aggour, John Interrante, and Ibrahim Gokcen. 2006. Integrating techniques for event-based business intelligence gathering. In *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pages 30–35.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

David Ahn, Steven Schockaert, Martine de Cock, and Etienne Kerre. 2006. Supporting temporal question answering: strategies for offline data collection. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*, pages 127–132.

Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1253.

James Allan, editor. 2002. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Boston, Massachusetts, USA.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

James Allan, Stephen Harding, David Fisher, Alvaro Bolivar, Sergio Guzman-Lara, and Peter Amstutz. 2005. Taking topic detection from evaluation to practice. In *Proceedings of the Thirty-Eighth Annual Hawaii International Conference on System Sciences*, pages 101–110.

James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 374–381.

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2007. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41. New York, New York, USA.

Omar Alonso, Jannik Strötgen, Ricardo Baeza-Yates, and Michael Gertz. 2011. Temporal information retrieval: Challenges and opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop*, pages 1–8.

Pascal Amsili and Corinne Rossari. 1998. Tense and connective constraints on the expression of causality. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 48–54.

Chinatsu Aone and Mila Ramos-Santacruz. 2000. REES: a large-scale relation and event extraction system. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 76–83.

Apache Software Foundation. 2012. *Class org.apache.lucene.search.Similarity.* `http://lucene.apache.org/core/3_6_1/api/core/org/apache/lucene/search/Similarity.html`. Accessed 20 December.

Ioannis Arapakis, Mounia Lalmas, Hakan Ceylan, and Pinar Donmez. 2014. Automatically embedding newsworthy links to articles: From implementation to evaluation. *Journal of the American Society for Information Science and Technology*, 65(1):129–145.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21. American Medical Informatics Association.

Naveen Ashish, Doug Appelt, Dayne Freitag, and Dmitry Zelenko, editors. 2006. *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis.*

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 79–85.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments and observations. In *Proceedings of the Workshop on Coreference and Its Applications*, pages 1–8.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational linguistics*, pages 86–90.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence.*

BBN Technologies. 2011. *Co-reference Guidelines for English OntoNotes.* LDC2011T03, Linguistic Data Consortium. Version 6.0.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422.

Cosmin Adrian Bejan. 2008. Unsupervised discovery of event scenarios from texts. In *Proceedings of the 21st Florida Artificial Intelligence Research Society International Conference.*

Cosmin Adrian Bejan and Sanda Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.*

Jonathan Bennett. 1998. *Events and their names.* Hackett Publishing Company, Indianapolis, Indiana, USA.

Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. 2010. Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27.

Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating story timelines as temporal dependency structure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2721–2726.

Steven Bethard and James H. Martin. 2006. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 807–815.

David Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Liliana Bounegru and Eric Karstens. 2012. *Blog source validation report.* Technical Report FP7–231854, D7.3.3, Sync3: Synergetic Content Creation & Communication.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 330–337.

Bertram C. Bruce. 1972. A model for temporal references and its application in a question answering program. *Artificial Intelligence*, 3:1–25. Elsevier.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16. Trento, Italy.

Chris Callison-Burch and Mark Dredze, editors. 2010. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning*, pages 89–97.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164.

Soumen Chakrabarti, Byron Dom, and Piotr Indyk. 1998. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 307–318.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of I-Semantics*, pages 42–49.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 976–986.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 173–176.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, pages 380–388.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. 2011. A unified event coreference resolution by integrating multiple resolvers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 102–110.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 54–57.

Zheng Chen, Heng Ji, and Robert Haralick. 2009. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 17–22.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–432.

Nancy Chinchor. 1998a. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). *Computational Linguistics*, 19(3):409–449.

Nancy Chinchor and Patricia Robinson. 1998. MUC-7 named entity task definition (version 3.5). In *Proceedings of the 7th Message Understanding Conference*.

Nancy A. Chinchor. 1998b. MUC/MET evaluation trends. In *Proceedings of the TIPSTER Text Program: Phase III*, pages 235–239.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. SystemT: An algebraic approach to declarative information extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 128–137.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, pages 113–120.

Herbert H. Clark and Adrian Bangerter. 2004. Changing conceptions of reference. In Ira A. Noveck and Dan Sperber, editors, *Experimental Pragmatics*, pages 25–49. Palgrave Macmillan, Basingstoke, UK.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149.

Mark Coddington. 2012. Building frames link by link: The linking practices of blogs and news sites. *International Journal of Communication*, 6:2007–2026.

Mark Coddington. 2014. Normalizing the hyperlink: How bloggers, professional journalists, and institutions shape linking values. *Digital Journalism*, 2:140–155. Published online April 2013.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 164–167.

James R. Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of the First Text Analysis Conference*.

Naomi Daniel, Dragomir Radev, and Timothy Allison. 2003. Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop*, pages 9–16.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1435–1444.

Donald Davidson. 1969. The individuation of events. In Nicholas Rescher, editor, *Essays in Honor of Carl G. Hempel*, volume 24 of *Synthesise Library*, pages 216–234. Springer. Republished in Donaldson's *Essays on Actions and Events*, Oxford, 1980.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program: Tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 837–840.

Keith S. Donnellan. 1966. Reference and definite descriptions. *Philosophical Review*, 75(3):281–304.

Doug Downey, Oren Etzioni, and Stephen Soderland. 2005. A probabilistic model of redundancy in information extraction. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1034–1041.

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, and Jonathan Wright. 2012. Linguistic resources for 2012 Knowledge Base Population evaluations. In *Proceedings of the Fifth Text Analysis Conference*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chi-Jen Lin. 2008. LIBLINEAR: A library for large linear classication. *Journal of Machine Learning Research*, 9:1871–1874.

Ao Feng and James Allan. 2005. *Hierarchical Topic Detection in TDT-2004*. IR 389, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts Amherst.

Ao Feng and James Allan. 2007. Finding and linking incidents in news. In *Proceedings of the ACM Sixteenth Conference on Information and Knowledge Management*, pages 821–829.

Ao Feng and James Allan. 2009. Incident threading for news passages. In *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, pages 1307–1316.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL–Shared Task*, pages 41–48.

Elena Filatova and Vasileios Hatzivassiloglou. 2003. Domain-independent detection, extraction, and labeling of atomic events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 104–111.

Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2006. Automatic creation of domain templates. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 207–214.

Elena Filatova and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In *Proceedings of the workshop on Temporal and Spatial Information Processing*.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Charles J. Fillmore, Srini Narayanan, and Collin F. Baker. 2006. What can linguistics contribute to event extraction? In *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pages 18–23.

Radu Florian, Hany Hassan, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos. 2004. A statistical model for multilingual entity detection and tracking. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1–8.

Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. GAF: A grounded annotation framework for events. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 11–20.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237.

Guillermo Garrido, Anselmo Peñas, Bernardo Cabaleiro, and Álvaro Rodrigo. 2012. Temporally anchored relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 107–116.

Julien Gaugaz, Patrick Siehndel, Gianluca Demartini, Tereza Iofciu, Mihai Georgescu, and Nicola Henze. 2012. Predicting the future impact of news events. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, pages 50–62.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 929–936.

Oren Glickman and Rosie Jones. 1999. Examining machine learning for adaptable end-to-end information extraction systems. In *Proceedings of AAAI 1999 Workshop on Machine Learning for Information Extraction.*

Chung Heong Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 9–16.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002a. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002b. Real-time event extraction for infectious disease outbreaks. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 366–369. Morgan Kaufmann Publishers Inc.

Ralph Grishman and Tomasz Kslezyk. 1990. Causal and temporal text analysis: The role of the domain model. In *Proceedings of the 13th Conference on Computational linguistics–Volume 1*, pages 126–131.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference – 6: A brief history. In *Proceedings of the 16th Conference on Computational linguistics–Volume 1.*

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. *NYU's English ACE 2005 System Description.* Technical report, Department of Computer Science, New York University.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 239–249.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 145–154.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194:130–150. Elsevier.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Sanda Harabagiu and Cosmin Adrian Bejan. 2006. An answer bank for temporal inference. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 741–746.

Sanda M. Harabagiu, Steven J. Maiorano, and Marius A. Paşca. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(3):1–38.

Lisa Harper, Inderjeet Mani, and Beth Sundheim, editors. 2001. *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing.*

Zelig Harris. 1954. Distributional structure. *WORD*, 10(23):146–162. International Linguistic Association.

Laura Hasler and Constantin Orăsan. 2009. Do coreferential arguments make event mentions coreferential? In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, pages 151–163.

Laura Hasler, Constantin Orvasan, and Karin Naumann. 2006. NPs for events: Experiments in coreference annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 1167–1172.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295.

Matthew Honnibal and Tobias Hawker. 2005. Identifying FrameNet frames for verbs from a real-text corpus. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 200–206.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013a. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013b. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194(0):2–27. Elsevier.

Ruihong Huang and Ellen Riloff. 2012. Bootstrapped training of event extraction classifiers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–295.

Ruihong Huang and Ellen Riloff. 2013. Multi-faceted event recognition with bootstrapped dictionaries. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 41–51.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts: Proceedings of a workshop sponsored by the ACL*, pages 75–81.

Seohyun Im and James Pustejovsky. 2010. Annotating lexically entailed subevents for textual inference tasks. In *Proceedings of the 23rd Florida Artificial Intelligence Research Society International Conference*, pages 204–209.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158.

Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *Proceedings of Recent Advances in Natural Language Processing*.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC2011 knowledge base population track. In *Proceedings of the Fourth Text Analysis Conference*.

Heng Ji, Xiang Li, Angelo Lucia, and Jianting Zhang. 2010. Annotating event chains for carbon sequestration literature. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

Hyuckchul Jung and Amanda Stent. 2013. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 20–24.

Christopher S.G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186. Oxford University Press.

Jaegwon Kim. 1973. Causation, nomic subsumption and the concept of event. *Journal of Philosophy*, 70(8):217–36. The Journal of Philosophy, Inc.

Jim-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9.

Paul Kingsbury and Martha Palmer. 2003. PropBank: The next level of TreeBank. In *Proceedings of Treebanks and Lexical Theories*.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 91–101.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Newbury Park, California, USA.

Amichai Kronfeld. 1990. *Reference and Computation: An Essay in Applied Philosophy of Language*. Studies in Natural Language Processing. Cambridge University Press.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304.

Erdal Kuzey and Gerhard Weikum. 2012. Extraction of temporal facts and events from Wikipedia. In *Proceedings of the 2nd Temporal Web Analytics Workshop*, pages 25–32.

LDC. 2003. *Annotation Guidelines for Event Detection and Characterization (EDC)*. Linguistic Data Consortium, for ACE. Version 2.0.

LDC. 2004. *TDT 2004: Annotation Manual, Version 1.2*. LDC2006T19, Linguistic Data Consortium.

LDC. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. Linguistic Data Consortium. Version 5.4.3.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500.

Robert Leibscher. 2004. Temporal context: Applications and implications for computational linguistics. In *Proceedings of ACL 2004: Student Research Workshop*, pages 19–24.

Jochen L. Leidner. 2004. Toponym resolution in text: "Which Sheffield is it?". In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 602–602.

Hao Li, Xiang Li, Heng Ji, and Yuval Marton. 2010. Domain-independent novel event discovery and semi-automatic event annotation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 233–242.

Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li, and Heng Ji. 2011. Joint inference for cross-document information extraction. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, pages 2225–2228.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. 2006. Extractive summarization using inter- and intra- event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 369–376.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797.

Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. 2008. Trust region newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9:627–650. JMLR Inc.

Ken Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12.

Maofu Liu, Wenjie Li, Mingli Wu, and Qin Lu. 2007. Extractive summarization based on event term clustering. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 185–188.

Hector Llorens, Estela Saquete, and Borja Navarro-Colorado. 2013. Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, 49(1):179–197.

Hector Llorens, Naushad Uzzaman, and James Allan. 2012. Merging temporal annotations. In *Proceedings of the 19th International Symposium on Temporal Representation and Reasoning (TIME)*, pages 107–113. IEEE.

Juha Makkonen and Helena Ahonen-Myka. 2003. Topic detection and tracking with spatio-temporal evidence. In *Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003)*, pages 251–265.

Inderjeet Mani, James Pustejovsky, and Robert Gaizauskas, editors. 2005. *The Language of Time: A Reader*. Oxford University Press.

Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 69–76.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Lluís Màrquez, Lluis Villarejo, M. A. Martí, and Mariona Taulé. 2007. SemEval-2007 task 09: Multilevel semantic annotation of catalan and spanish. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 42–47.

Nancy McCracken, Necati Ercan Ozgencil, and Svetlana Symonenko. 2006. Combining techniques for event extraction in summary reports. In *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pages 7–11.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 396–404.

Olena Medelyan and Catherine Legg. 2008. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. 2004. Event threading within news topics. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*.

Martina Naughton, Nicola Stokes, and Joseph Carthy. 2010. Sentence-level event classification in unstructured texts. *Information Retrieval*, 13(2):132–156.

NIST. 2005. *The ACE 2005 (ACE05) Evaluation Plan*. `http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf`.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Joel Nothman. 2008. *Learning Named Entity Recognition from Wikipedia*. Honours thesis, School of IT, University of Sydney, Sydney, Australia.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. In *Proceedings of the Australian Language Technology Workshop*, pages 124–132.

Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232.

Joel Nothman, Tara Murphy, and James R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 612–620.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175. Elsevier.

Timothy O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Miles Osborne, Saša Petrović, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2012. Bieber no more: First story detection using Twitter and Wikipedia. In *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access*.

Alexis Palmer and Caroline Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Coling 2010: Posters*, pages 928–936.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. LDC2011T07, Linguistic Data Consortium. June 17.

Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu, and Sibel Yaman. 2009. Who, what, when, where, why? comparing multiple approaches to the cross-lingual 5W task. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 423–431.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 717–727.

Philip L. Peterson. 1997. *Fact Proposition Event*, volume 66 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189.

Jakub Piskorski, Hristo Tanev, Martin Atkinson, and Erik Van Der Goot. 2008. Cluster-centric approach to news event extraction. In *Proceeding of the 2008 Conference on New Trends in Multimedia and Network Information Systems*, pages 276–290. IOS Press.

Simone Paolo Ponzetto and Michael Strube. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10):1737–1756. Elsevier, Essex, UK.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007a. SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL–Shared Task*, pages 1–40.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007b. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing*, pages 446–453.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003a. The TimeBank corpus. In *Proceedings of Corpus Linguistics*, pages 647–656.

James Pustejovsky, Jessica Littman, Roser Saurí, and Marc Verhagen. 2006. *TimeBank 1.2 Documentation.*

James Pustejovsky, José Casta no, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003b. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics.*

James Pustejovsky, Anna Rumshisky, Alex Plotnick, Elisabetta Jezek, Olga Batiukova, and Valeria Quochi. 2010. SemEval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 27–32.

Dragomir Radev, Jahna Otterbacher, and Zhu Zhang. 2004a. CST Bank: A corpus for the study of cross-document structural relationships. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation.*

Dragomir R. Radev. 2000. A common theory of information fusion from multiple text sources. In *Proceedings of the 1st ACL SIGdial Workshop on Discourse and Dialogue.*

Dragomir R. Radev, Hongyan Jing, Magorzata Styś, and Daniel Tam. 2004b. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938. Elsevier.

Will Radford. 2014. *Linking Named Entities to Wikipedia.* Ph.D. thesis, School of IT, University of Sydney.

Will Radford, Will Cannings, Andrew Naoum, Joel Nothman, Glen Pink, Daniel Tse, and James R. Curran. 2012. (Almost) Total Recall – SYDNEY_CMCRC at TAC 2012. In *Proceedings of the Text Analysis Conference*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the International World Wide Web Conference*, pages 909–918.

Marta Recasens, Eduard Hovy, and M. Antnia Mart. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Marta Recasens, Eduard Hovy, and M. Antnia Mart. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Marta Recasens, Maria Antònia Martí, and Constantin Orvasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 165–172.

Marta Recasens and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

Roi Reichart and Regina Barzilay. 2012. Multi-event extraction guided by global constraints. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 70–79. Montréal, Canada.

Reuters OpenCalais. 2009. *Entity/Fact/Event Index and Definitions*. `http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions`, editions in the Internet Archive from March 2009.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049.

Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 49–56.

Kirk Roberts, Travis Goodwin, and Sanda M. Harabagiu. 2012. Annotating spatial containment relations between events. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3052–3059.

Kirk Roberts and Sanda Harabagiu. 2010. UTDMet: Combining WordNet and corpus data for argument coercion detection. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 252–255.

Kirk Roberts and Sanda M. Harabagiu. 2011. Detecting new and emerging events in streaming news documents. *International Journal of Semantic Computing*, 5(4):407–431.

Kirk Roberts, Michael A. Skinner, and Sanda M. Harabagiu. 2013. Recognizing spatial containment relations between event mentions. In *Proceedings of the 10th International Conference on Computational Semantics – Long Papers*, pages 216–227.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of The First Joint Conference on Lexical and Computational Semantics*, pages 218–227. Montréal, Canada.

Josef Ruppenhofer, Jonas Sunde, and Manfred Pinkal. 2010. Generating FrameNets of various granularities: The FrameNet transformer. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. LDC2008T19, Linguistic Data Consortium. October 17.

Estela Saquete, Jose L. Vicedo, Patrico Martínez-Barco, Rafael Mu noz, and Hector Llorens. 2009. Enhancing QA systems with complex temporal question processing capabilities. *Journal of Artificial Intelligence Research*, 35:775–811.

Roser Saurí, Lotus Goldberg, Marc Verhagen, and James Pustejovsky. 2009. *Annotating Events in English: TimeML Annotation Guidelines*. Technical report, Brandeis University. Version TempEval-2010.

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: A robust event recognizer for QA systems. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 700–707.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, And Understanding: An Inquiry Into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale, New Jersey, USA.

Steven Schockaert, David Ahn, Martine De Cock, and Etienne E. Kerre. 2006. Question answering with imperfect temporal information. In *Proceedings of the 7th International Conference on Flexible Query Answering Systems*, volume LNAI 4207, pages 647–658.

Satoshi Sekine. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 731–738.

Andrea Setzer and Robert Gaizauskas. 2000. Annotating events and temporal information in newswire texts. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000) Workshop on Information Extraction Meets Corpus Linguistics*, pages 9–14.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 100–111.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 65–71.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 304–311.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628.

Heather Simpson, Stephanie Strassel, Robert Parker, and Paul McNamee. 2010. Wikipedia and the web of confusable entities: Experience from entity linking query creation for TAC 2009 knowledge base population. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Ian Soboroff and Donna Harman. 2005. Novelty detection: The TREC experience. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Rohini K. Srihari, Wei Li, Cheng Niu, and Thomas Cornell. 2003. InfoXtract: A customizable intermediate level information extraction engine. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*.

Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 713–722.

Stephanie Strassel, Christopher Walker, and Alexis Mitchell. 2004. Annotation consistency study. In *Proceedings of the ACE Mid-Course Correction Workshop*. Linguistic Data Consortium.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298. Springer.

Beth M. Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 13–32.

Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, and Christopher D. Manning. 2011. Stanford's distantly-supervised slot-filling system. In *Proceedings of the Fourth Text Analysis Conference*.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56. ACM.

Shohei Tanaka, Naoaki Okazaki, and Mitsuru Ishizuka. 2012. Acquiring and generalizing causal inference rules from deverbal noun constructions. In *Proceedings of COLING 2012: Posters*, pages 1209–1218.

Sam Tardif, James R. Curran, and Tara Murphy. 2009. Improved text categorisation for Wikipedia named entities. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 104–108.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–4.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179.

Naushad UzZaman and James Allen. 2010. TRIOS-TimeBank Corpus: Extended TimeBank corpus with help of deep understanding of text. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Teun A. Van Dijk. 1988. *News as discourse*. Communication. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Willem Robert van Hage, Vronique Malais, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the simple event model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

Christopher Walker, Zhiyi Song, Ramez Zakhary, Alexis Mitchell, and Stephanie Strassel. 2005. The event annotation task: Overview, inventory and open issues. In *Proceedings of the ACE Mid-Course Correction Workshop*. Linguistic Data Consortium.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. *ACE 2005 Multilingual Training Corpus*. LDC2006T06, Linguistic Data Consortium. February 15 edition.

Rui Wang and Yajing Zhang. 2008. Recognizing textual entailment with temporal expressions in natural language texts. In *Proceedings of the IEEE International Workshop on Semantic Computing and Applications*.

Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. 2010. Timely YAGO: Harvesting, querying, and visualizing temporal knowledge from Wikipedia. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 697–700.

Bonnie Lynn Webber. 1987. Position paper: Event reference. In *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing*, pages 158–163.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Craig A. Will. 1993. Comparing human and machine performance for natural language information extraction: Results from the Tipster text evaluation. In *Proceedings of the TIPSTER Text Program: Phase I*, pages 179–193.

Travis Wolfe, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, and Xuchen Yao. 2013. PARMA: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68.

Fei Wu, Raphael Hoffmann, and Daniel S. Weld. 2008. Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th International Conference on Knowledge Discovery & Data Mining*.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying Wikipedia. In *Proceedings of the 16th Conference on Information and Knowledge Management*.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 433–443.

Christopher C. Yang, Xiaodong Shi, and Chih-Ping Wei. 2009. Discovering event evolution graphs from news corpora. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 34(4):850–863.

Kai-Hsiang Yang, Kun-Yan Chiou, Hahn-Ming Lee, and Jan-Ming Ho. 2006. Web appearance disambiguation of personal names based on network motif. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 386–389.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 688–693.

Roman Yangarber. 2006. Verification of facts across document boundaries. In *Proceedings of the International Workshop on Intelligent Information Access*.

Roman Yangarber and Ralph Grishman. 1997. Customization of information extraction systems. In *Proceedings of International Workshop on Lexically-Driven Information Extraction*, pages 1–11. Università di Roma.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 940–946.

Wen-tau Yih. 2009. Learning term-weighting functions for similarity measures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 793–802.