Antecedent Selection Techniques for High-Recall Coreference Resolution

Yannick Versley
versley@sfs.uni-tuebingen.de
SFB 441 / Seminar für Sprachwissenschaft
Universität Tübingen

Abstract

We investigate methods to improve the recall in coreference resolution by also trying to resolve those definite descriptions where no earlier mention of the referent shares the same lexical head (coreferent bridging). The problem, which is notably harder than identifying coreference relations among mentions which have the same lexical head, has been tackled with several rather different approaches, and we attempt to provide a meaningful classification along with a quantitative comparison. Based on the different merits of the methods, we discuss possibilities to improve them and show how they can be effectively combined.

1 Introduction

Coreference resolution, the task of grouping mentions in a text that refer to the same referent in the real world, has been shown to be beneficial for a number of higher-level tasks such as information extraction (McCarthy and Lehnert, 1995), question answering (Morton, 2000) and summarisation (Steinberger et al., 2005).

While the resolution of pronominal anaphora and tracking of named entities is possible with good accuracy, the resolution of definite NPs (having a common noun as their head) is usually limited to the cases that Vieira and Poesio (2000) call direct coreference, where both coreferent mentions have the same head. The other cases, called coreferent bridging by Vieira and Poesio[1], are notably harder because the number of potential candidates is much larger when it is no longer possible to rely on surface similarity.

To overcome the limit of recall that is encountered when only relying on surface features, newer systems for coreference resolutions (Daumé III and Marcu, 2005; Ponzetto and Strube, 2006; Versley, 2006; Ng, 2007, *inter alia*) use lexical semantic information as an indication for semantic compatibility in the absence of head equality. Most current systems integrate the identification of discourse-new definites (i.e., cases like *"the sun"* or *"the man that Ben met yesterday"*, which are definite, but not anaphoric) with the antecedent selection proper, which implies that the gain obtained for new features is dependent on the feature's usefulness both in finding semantically related mentions and for the use in detecting discourse-new definites.

One goal of this paper is to provide a better understanding of these information sources by comparing proposed (and partly new) approaches for resolving coreferent bridging by separately considering the task of antecedent selection (i.e., presupposing that discourse-new markables have been identified beforehand). Although state of the art methods for modular discourse-new detection (Uryupina, 2003; Poesio et al., 2005) do not achieve near-perfect accuracy for discourse-new detection, the results we give for antecedent selection represent an upper bound on recall and precision for the full coreference task, and we think that this upper bound will be useful for

[1] Because bridging (in the sense of Clark, 1975, or Asher and Lascarides, 1998) is a much broader concept, the term 'coreferent bridging' is potentially confusing, as many cases are examples of perfectly well-behaved anaphoric definite noun phrases. Because we want to emphasise the important difference to the more easily resolved cases of same-head coreference, we will stick with 'coreferent bridging' as the only term that has been established for this in the literature.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 496–505, Prague, June 2007. ©2007 Association for Computational Linguistics

the design of features in both systems using a modular approach, such as (Poesio et al., 2005), where the decision on discourse-newness is taken beforehand, and those that integrate discourse-new classification with the actual resolution of coreferent bridging cases. In contrast to earlier investigations (Markert and Nissim, 2005; Garera and Yarowsky, 2006), we provide a more extensive overview on features and also discuss properties that influence their combinability.

Several approaches have been proposed for the treatment of coreferent bridging. Poesio et al. (1997) use WordNet, looking for a synonymy or hypernymy relation (additionally, for coordinate sisters in WordNet). The system of Cardie and Wagstaff (1999) uses the node distance in WordNet (with an upper limit of 4) as one component in the distance measure that guides their clustering algorithm. Harabagiu et al. (2001) use paths through Wordnet, using not only synonym and is-a relations, but also parts, morphological derivations, gloss texts and polysemy, which are weighted with a measure based on the relation types and number of path elements. Other approaches use large corpora to get an indication for bridging relations: Poesio et al. (1998) use a general word association metric based on common terms occuring in a fixed-width window, Gasperin and Vieira (2004) use syntactic contexts of words in a large corpus to induce a semantic similarity measure (similar to the one introduced by Lin, 1998), and then use lists of the n nouns that are (globally) most similar to a given noun. Markert and Nissim (2005) mine the World Wide Web for shallow patterns like "*China* and other *countries*", indicating an is-a relationship. Finally, Garera and Yarowsky (2006) propose an association-based approach using nouns that occur in a 2-sentence window before a definite description that has no same-head antecedent.

1.1 Lexical vs. Referential Relations

One important property of these information sources is the kind of lexical relations that they detect. The lexical relations that we expect in coreferent bridging cases are:

- instance: The antecedent is an instance of the concept denoted by the anaphor
  Corsica . . . the island

- synonymy: The antecedent and the anaphor are synonyms
  the automobile . . . the car

- hyperonymy: The anaphor is a strict generalisation of the antecedent
  the murderer . . . the man

- near-synonymy: The anaphor and antecedent are semantically related but not synonyms in the strict sense
  the CD . . . the album

Of course, not all cases of coreferent bridging realise such a lexical relation, as sometimes the anaphor takes up information introduced elsewhere than in the lexical noun phrase head (Peter was found dead in his flat . . . the deceased), or the coreference relation is forced by the discourse structure, without the items being lexically related.

As an illustrating example, in

(1) John walked towards [1 the house].

(2) a. [1 The building] was illuminated.
b. [1 The manor] was guarded by dogs.
c. [2 The door] was open.

Typical cases of coreference include cases like 1,2a (hypernym) or 1,2b (compatible but non-synonymous term). The discourse in 1,2c is an example of associative bridging between the NP *"the door"* and its antecedent to *"the house"*; it is inferred that the door must be part of the house mentioned earlier (since doors are typically part of a house), which is *not* compatible with coreferent bridging, but is also ranked highly by association measures.

While hypernym relations (as found by hypernym lookup in WordNet, or patterns indicating such relations in unannotated texts) are usually a strong indicator of coreference, they can only cover some of the cases, while the near-synonymous cases are left undiscovered. Similarity and association measures can help for the cases of near-synonymy. However, while similarity measures (such as WordNet distance or Lin's similarity metric) only detect cases of semantic similarity, association measures (such as the ones used by Poesio et al., or by Garera and Yarowsky) also find cases of associative bridg-

Lin98	RFF	TheY	TheY:G^2	PL03
		Land ***(country/state/land)***		
Staat *state*	Staat *state*	Kemalismus *Kemalism*	Regierung *government*	Kontinent *continent*
Stadt *city*	Stadt *city*	Bauernfamilie *agricultural family*	Präsident *president*	Region *region*
Region *region*	Landesregierung *country government*	Bankgesellschaft *banking corporation*	Dollar *dollar*	Stadt *city*
Bundesrepublik *federal republic*	Bundesregierung *federal government*	Baht *Baht*	Albanien *Albania*	Staat *state*
Republik *republic*	Gewerkschaft *trade union*	Gasag *(a gas company)*	Hauptstadt *capital*	Bundesland *state*
		Medikament ***(medical drug)***		
Arzneimittel *pharmaceutical*	Pille *pill*	RU *(a drug[*])*	Patient *patient*	Arzneimittel *pharmaceutical*
Präparat *preparation*	Droge *drug (non-medical)*	Abtreibungspille *abortion pill*	Arzt *doctor*	Lebensmittel *foodstuff*
Pille *pill*	Präparat *preparation*	Viagra *Viagra*	Pille *pill*	Präparat *preparation*
Hormon *hormone*	Pestizid *pesticide*	Pharmakonzern *pharmaceutical company*	Behandlung *treatment*	Behandlung *treatment*
Lebensmittel *foodstuff*	Lebensmittel *foodstuff*	Präparat *preparation*	Abtreibungspille *abortion pill*	Arznei *drug*

highest ranked words, with very rare words removed

[*]: RU 486, an abortifacient drug

Lin98: Lin's distributional similarity measure (Lin, 1998)

RFF: Geffet and Dagan's *Relative Feature Focus* measure (Geffet and Dagan, 2004)

TheY: association measure introduced by Garera and Yarowsky (2006)

TheY:G^2: similar method using a log-likelihood-based statistic (see Dunning 1993)
this statistic has a preference for higher-frequency terms

PL03: semantic space association measure proposed by Padó and Lapata (2003)

Table 1: Similarity and association measures: most similar items

ing like 1a,b; the result of this can be seen in table (2): while the similarity measures (Lin98, RFF) list substitutable terms (which behave like synonyms in many contexts), the association measures (Garera and Yarowsky's TheY measure, Padó and Lapata's association measure) also find non-compatible associations such as *country–capital* or *drug–treatment*, which is why they are commonly called *relation-free*. For the purpose of coreference resolution, however we do *not* want to resolve *"the door"* to the antecedent *"the house"* as the two descriptions do not corefer, and it may be useful to filter out non-similar associations.

1.2 Information Sources

Different resources may be differently suited for the recognition of the various relations. Generally, it would be expected that using a wordnet is the best solution if we are interested in an isa-like relation between two words. On the other hand, wordnets usually have limited coverage both in terms of lexical items and in terms of relations encoded (as their construction is necessarily labor-intensive), and – as Markert and Nissim remark – they do not (and arguably should not) contain context-dependent relations that do not hold generally but only in some rather specific context, for example *steel* being anaphorically described as a *commodity* in a financial text. Context-dependent relations, Markert and Nissim argue, can be found using shallow patterns (for example, *steel and other commodities*), since a use in such a context would mean that the idiosyncratic conceptual relation holds in that context. Wordnets also have usually have poor (or non-existant) coverage of named entities, which are especially relevant for instance relations; this kind of instance relations can often be found in large text corpora. The high-precision patterns that Markert and Nissim use only occur infrequently, but the approach using shallow patterns allows to perform the search of the World Wide Web, which somewhat alleviates the sparse data problem.

While some near-synonyms can be found by looking at the distance in a wordnet, they may be far apart from each other because of ontological modeling decisions, or lexical items not covered by the wordnet. Similarity and association measures can provide greater coverage for these near-synonym relations.

The measures both of Lin (1998) and of Padó and Lapata (2003, 2007) are distributional methods; for each word, they create a distribution of the contexts they occur in, and similarity between two words is calculated as the similarity of these distributions.[2] The difference in these two methods is the representation of the contexts. While Lin uses contexts that are expected to determine semantic preferences (like being in the direct object position of one verb), Padó and Lapata only use the co-occuring words, weighted by syntax-based distance. For example, in

(3) Peter $\overset{subj}{\rightarrow}$ likes $\overset{dobj}{\leftarrow}$ ice-cream.

Lin's approach would yield $\uparrow subj{:}\texttt{like}$ for `Peter` and $\uparrow dobj{:}\texttt{like}$ for `ice-cream`, while Padó and Lapata's approach would yield the contexts `like` (with a weight of 1.0) and `ice-cream` (with a weight of 0.5) for `Peter`. As a consequence, Padó and Lapata's measure is more robust against data sparseness but also finds related non-similar terms (which are ultimately unwanted for coreference resolution). Padó and Lapata show their dependency-based measure to perform better in a word sense disambiguation task than the measure of Lund et al. (1995), on which Poesio et al. (1998) based their experiments and which is based on the surface distance of words.

We also reimplemented the approach of Garera and Yarowsky (2006), who extract potential anaphor-antecedent pairs from unlabeled texts and rank these potentially related pairs by the mutual information statistic. As an example, in a text like

(4) Peter likes ice-cream.
The boy devours tons of it.

we would extract the pairs $\langle \texttt{boy}, \texttt{(person)} \rangle$ and $\langle \texttt{boy}, \texttt{ice-cream} \rangle$, in the hope that the former pair occurs comparatively more often and gets a higher mutual information value.

2 Experiments on Antecedent Selection

In a setting similar to Markert and Nissim (2005), we evaluate the precision (proportion of correct cases in the resolved cases) and recall (correct cases to all cases) for the resolution of discourse-old definite noun phrases. Before trying to resolve coreferent bridging cases, we look for compatible antecedent candidates with the same lexical head and resolve to the nearest such candidate if there is one.

For our experiments, we used the first 125 articles of the coreferentially annotated TüBa-D/Z corpus of written newspaper text (Hinrichs et al., 2005), totalling 2239 sentences with 633 discourse-old definite descriptions, and the latest release of GermaNet (Kunze and Lemnitzer, 2002), which is the German-language part of EuroWordNet.

Unlike Markert and Nissim, we did not limit the evaluation to discourse-old noun phrases where an antecedent is in the 4 preceding sentences, but also included cases where the antecedent is further away. As a real coreference resolution system would have to either resolve them correctly or leave them unresolved, we feel that this is less unrealistic and thus preferable even when it gives less optimistic evaluation results. Because overall precision is a mixture of the precision of the same-head resolver and the precision of the resolution for coreferent bridging, which is lower than that for same-head cases, we forcibly get less precision if we resolve more coreferent bridging cases. As it is always possible to improve overall precision by resolving fewer cases of coreferent bridging, we separately mention the precision for coreferent bridging cases alone (i.e., number of correct coreferent bridging cases by all resolved coreferent bridging cases), which we deem more informative.

In our evaluation, we included hypernymy search and a simple edge-based distance based on GermaNet, as well as a baseline using semantic classes (automatically determined by a combination of simple named entity classification and GermaNet subsumption), as well as an evolved version of Markert

[2] Both measures use a weighted Jaccard metric on mutual information vectors to calculate the similarity. See Weeds and Weir (2005) for an overview of other measures.

	Prec	Recl	$\mathrm{F}_{\beta=1}$	Prec.NSH
same-head	0.87	0.50	0.63	—
nearest$^{(1)}$ (only number check)	0.57	0.55	0.56	0.12
semantic class+gender check$^{(1)}$	0.68	0.61	0.64	0.35
semantic class+gender check$^{(2)}$	0.67	0.62	0.65	0.36
GermaNet, hypernymy lookup	**0.83**	0.58	**0.68**	**0.67**
GermaNet, node distance$^{(1)}$	0.71	0.61	0.65	0.39
single pattern: "Y wie X"$^{(1)}$	0.83	0.54	0.66	0.55
TheY$^{(1)}$ (only number checking)	0.66	0.59	0.62	0.29
TheY$^{(2)}$ (only number checking)	0.66	0.60	0.63	0.31
Lin$^{(1)}$ (only number checking)	0.66	0.60	0.63	0.30
Lin$^{(2)}$ (only number checking)	0.69	0.64	0.66	0.39
PL03$^{(1)}$ (only number checking)	0.68	0.63	0.65	0.38
PL03$^{(2)}$ (only number checking)	0.70	**0.64**	0.65	0.42
15-most-similar$^{(1)}$	0.82	0.54	0.65	0.50
100-most-similar$^{(2,3)}$	0.73	0.60	0.66	0.42

Prec.NSH: precision for coreferent bridging cases

$^{(1)}$: consider candidates in the 4 preceding sentences

$^{(2)}$: consider candidates in the 16 preceding sentences

$^{(3)}$: also try candidates such that the anaphor is in the antecedent's similarity list

Table 2: Baseline results

and Nissim's approach, which is presented in (Versley, 2007). For the methods based on similarity and association measures, we implemented a simple ranking by the respective similarity or relatedness value. Additionally, we included an approach due to Gasperin and Vieira (2004), who tackle the problem of similarity by using lists of most similar words to a certain word, based on a similarity measure closely related to Lin's. They allow resolution if either (i) the candidate is among the words most similar to the anaphor, (ii) the anaphor is among the words most similar to the candidate, (iii) the similarity lists of anaphor and candidate share a common item. We tried out several variations in the length of the similar words list (Gasperin and Vieira used 15, we also tried lists with 25, 50 and 100 items). The third possibility that Gasperin and Vieira mention (a common item in the similarity lists of both anaphor and antecedent) resolves some correct cases, but leads to a much larger number of false positives, which is why we did not include it in our evaluation.

To induce the similarity and association measures presented earlier, we used texts from the German newspaper *die tageszeitung*, comprising about 11M sentences. For the extraction of anaphor-antecedent candidates, we used a chunked version of the corpus (Müller and Ule, 2002). The identification of grammatical relations, was carried out on a subset of all sentences (those with length ≤ 30), with an unlexicalised PCFG parser and subsequent extraction of dependency relations (Versley, 2005). For the last approach, where dependency relations were needed but labeling accuracy was not as important, we used a deterministic shift-reduce parser that Foth and Menzel (2006) used as input source in hybrid dependency parsing.[3]

For all three approaches, we lemmatised the words by using a combination of SMOR (Schmid et al., 2004), a derivational finite-state morphology for German, and lexical information derived from the lexicon of a German dependency parser (Foth and Menzel, 2006). We mitigated the problem of vocabulary growth in the lexicon, due to German synthetic compounds, by using a frequency-sensitive unsupervised compound splitting technique, and (for semantic similarity) normalised common person and location names to '(person)' and '(location)', respectively.

Same-head resolution (including a check for modifier compatibility) allows to correctly resolve 49.8% of all cases, with a precision of 86.5%. The most simple approach for coreferent bridging, just resolving coreferent bridging cases to the nearest possible antecedent (only checking for number agreement), yields very poor precision (12% for the coreferent bridging cases), and as a result, the recall gain is very limited. If we use semantic classes (based on both GermaNet and a simple classification for named entities) to constrain the candidates and then use the nearest number- and gender-compatible antecedent[4], we get a much better precision (35% for coreferent bridging cases), and a much better recall of 61.1%. Hyponymy lookup in GermaNet, without a limit on sentence distance, achieves a recall of 57.5% (with a precision of 67% for the resolved coreferent bridging cases), whereas using the best single pattern (Y wie X, which corresponds to

[3] Arguably, it would have been more convenient to use a single parser for all three approaches, but differing tradeoffs between speed on one hand and accuracy for relevant information and/or fitness of representation on the other hand made the respective parser or chunker a compelling choice.

[4] In German, grammatical gender is not as predictive as in English as it does not reproduce ontological distinctions. For persons, grammatical and natural gender almost always coincide, and we check gender equality iff the anaphor is a person.

the English Ys such as X), with a distance limit of 4 sentences[5], on the Web only improves the recall to 54.3% (with a lower precision of 55% for coreferent bridging cases). This is in contrast to the results of Markert and Nissim, who found that Web pattern search performs better than wordnet lookup; see (Versley, 2007) for a discussion. Ranking all candidates that are within a distance of 4 hyper-/hyponymy edges in GermaNet by their edge distance, we get a relatively good recall of 60.5%, but the precision (for the coreferent bridging cases) is only at 39%, which is quite poor in comparison.

The results for Garera and Yarowsky's TheY algorithm are quite disconcerting – recall and the precision on coreferent bridging cases are lower than the respective baseline using (wordnet-based) semantic class information or Padó and Lapata's association measure. The technique based on Lin's similarity measure does outperform the baseline, but still suffers from bad precision, along with Padó and Lapata's association measure. In other words, the similarity and association measures seem to be too noisy to be used directly for ranking antecedents. The approach of Gasperin and Vieira performs comparably to the approach using Web-based pattern search (although the precision is poorer than for the best-performing pattern for German, "X wie Y" – X such as Y, it is comparable to that of other patterns).

2.1 Improving Distributional Similarity?

While it would be naïve to think that the methods purely based on statistical similarity measures could reach the accuracy that can be achieved with a hand-constructed lexicalised ontology, it would of course be nice if we could improve the quality of the semantic similarity measure used in ranking and the most-similar-word lists.

Geffet and Dagan (2004) propose an approach to improve the quality of the feature vectors used in distributional similarity measures: instead of weighting features using the mutual information value between the word and the feature, they propose to use a measure they call *Relative Feature Focus*: the sum of the similarities to the (globally) most similar words that share this feature.

By replacing mutual information values with RFF values in Lin's association measure, Geffet and Dagan were able to significantly improve the proportion of substitutable words in the list of the most similar words. In our experiments, however, using the RFF-based similarity measure did not improve the similarity-list-based resolution or the simple ranking, to the contrary, both recall and precision are less than for the Weighted Jaccard measure that we used originally.[6]

We attribute this to two factors: Firstly, Geffet and Dagan's evaluation emphasises the precision in terms of *types*, whereas the use in resolving coreferent bridging does not punish unrelated rare words being ranked high – since these are rare, the likelihood that they occur together, changing a resolution decision, is quite low, whereas rare related words that are ranked high can allow a correct resolution. Secondly, Geffet and Dagan focus on high-frequency words, which makes sense in the context of ontology learning, but the applicability for tasks like coreference resolution (directly or in the approach of Gasperin and Vieira) also depends on a sensible treatment of lower-frequency words.

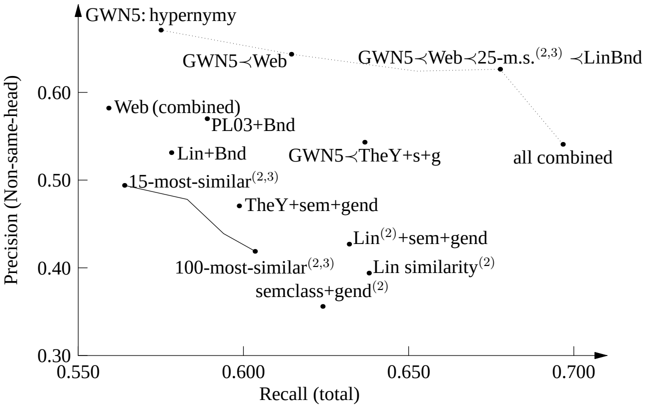
Using the framework of Weeds et al. (2004), we found that the bias of lower frequency words for preferring high-frequency neighbours was higher for RFF (0.58 against 0.35 for Lin's measure). Weeds and Weir (2005) discuss the influence of bias towards high- or low-frequency items for different tasks (correlation with WordNet-derived neighbour sets and pseudoword disambiguation), and it would not be surprising if the different high-frequency bias were leading to different results.

2.2 Combining Information Sources

The information sources that we presented earlier and the corpus-based methods based on similarity or association measures draw from different kinds of evidence and thus should be rather complementary. To put it another way, it should be possible to get the best from all methods, achieving the recall of the high-recall methods (like using semantic class in-

[5] There is a degradation in precision for the pattern-based approach, but not for the GermaNet-based approach, which is why we do not use a distance limit for the GermaNet-based approach.

[6] Simple ranking with RFF gives a precision of 33% for coreferent bridging cases, against 39% for Lin's original measure; for an approach based on similarity lists, we get 39% against 44%.



	Prec	Recl	$F_{\beta=1}$	Prec.NSH
sem. class+gender checking	0.68	0.61	0.64	0.35
GermaNet, hypernymy lookup	0.83	0.57	0.68	0.67
GermaNet ≺ “Y wie X”	0.81	0.60	0.69	0.63
GermaNet ≺ all patterns	0.81	0.61	0.70	0.64
TheY$^{(2)}$+semclass+gender	0.76	0.60	0.67	0.47
TheY+sem+gend+Bnd	0.78	0.59	0.67	0.50
Lin$^{(2)}$+semclass+gender	0.71	0.63	0.67	0.43
Lin+sem+gend+Bnd	0.80	0.58	0.67	0.53
PL03$^{(2)}$+semclass+gender	0.72	0.64	0.68	0.45
PL03+sem+gend+Bnd	0.80	0.59	0.68	0.57
GermaNet ≺ all patterns	0.81	0.62	0.70	0.64
≺ 25-most-similar$^{(2,3)}$	0.79	0.65	0.72	0.62
≺ LinBnd	0.79	0.68	**0.73**	0.63
≺ Lin ≺ TheY+sem+gend	0.74	**0.70**	0.72	0.54

$^{(2)}$: consider candidates in the 16 preceding sentences
$^{(3)}$: also try candidates such that the anaphor is in the antecedent’s similarity list

Table 3: Combination-based approaches

formation, or similarity and association measures), with a precision closer to the most precise method using GermaNet. In the case of web-based patterns, Versley (2007) combines several pattern searches on the web and uses the combined positive and negative evidence to compute a composite score – with a suitably chosen cutoff, it outperforms all single patterns both in terms of precision and recall. First resolving via hyponymy in GermaNet and then using the pattern-combination approach outperforms the semantic class-based baseline in terms of recall and is reasonably close to the GermaNet-based approach in terms of precision (i.e., much better than the approach based only on the semantic class).

As a first step to improve the precision of the corpus-based approaches, we added filtering based on automatically assigned semantic classes (persons, organisations, events, other countable objects, and everything else). Very surprisingly, Garera and Yarowsky’s TheY approach, despite starting out at a lower precision (31%, against 39% for Lin and 42% for PL03), profits much more from the semantic filter and reaches the best precision (47%), whereas Lin’s semantic similarity measure profits the least.

Since limiting the distance to the 4 previous sentences had quite a devastating effect for the approach based on Lin’s similarity measure (which achieves 39% precision when all the candidates are available and 30% precision if it choses the most semantically similar out of the candidates that are in the last 4 sentences), we also wanted to try and apply the distance-based filtering after finding semantically related candidates.

The approach we tried was as follows: we rank all candidates using the similarity function, and keep only the 3 top-rated candidates. From these 3 top-rated candidates, we keep only those within the last 4 sentences. Without filtering by semantic class, this improves the precision to 41% (from 30% for limiting the distance beforehand, or 39% without limiting the distance). Adding filtering based on semantic classes to this (only keeping those from the 3 top-rated candidates which have a compatible semantic class and are within the last 4 sentences), we get a much better precision of 53%, with a recall that can still be seen as good (57.8%). In comparison with the similarity-list-based approach, we get a much better precision than we would get for methods with comparable recall (the version with the 100 most similar items has 44% precision, the version with 50 most similar items and matching both ways has 46% precision).

Applying this distance-bounding method to Garera and Yarowsky’s association measure still leads to an improvement over the case with only semantic and gender checking, but the improvement (from 47% to 50%) is not as large as with the semantic similarity measure or Padó and Lapata’s association measure (from 45% to 57%).

For the final system, we back off from the most precise information sources to the less precise. Starting with the combination of GermaNet and pattern-based search on the World Wide Web, we begin by adding the distance-bounded semantic similarity-based resolver (LinBnd) and resolution based on the list of 25 most similar words (following the approach of Gasperin and Vieira 2004). This results in visibly improved recall (from 62% to 68%), while the precision for coreferent bridging cases does not suffer much. Adding resolution based on Lin's semantic similarity measure and Garera and Yarowsky's TheY value leads to a further improvement in recall to 69.7%, but also leads to a larger loss in precision.

3 Conclusion

In this paper, we compared several approaches to resolve cases of coreferent bridging in open-domain newspaper text. While none of the information sources can match the precision of the hypernymy information encoded in GermaNet, or that of using a combination of high-precision patterns with the World Wide Web as a very large corpus, it is possible to achieve a considerable improvement in terms of recall without sacrificing too much precision by combining these methods.

Very interestingly, the distributional methods based on intra-sentence relations (Lin, 1998; Padó and Lapata, 2003) outperformed Garera and Yarowsky's (2006) association measure when used for ranking, which may due to sparse data problems or simply too much noise for the latter. For the association measures, the fact that they are relation-free also means that they can profit from added semantic filtering.

The novel distance-bounded semantic similarity method (where we use the most similar words in the previous discourse together with a semantic class-based filter and a distance limit) comes near the precision of using surface patterns, and offers better accuracy than Gasperin and Vieira's method of using the globally most similar words.

By combining existing higher-precision information sources such as hypernym search in GermaNet and the Web-based approach presented in (Versley, 2007) together with similarity- and association-based resolution, it is possible to get a large improvement in recall even compared to the combined GermaNet+Web approach or an approach combining GermaNet with a semantically filtered version of Garera and Yarowsky's TheY approach.

In independent research, Goecke et al. (2006) combined the original LSA-based method of Lund et al. (1995) with wordnet relations and pattern search on a fixed-size corpus.[7] However, they evaluate only on a small subset of discourse-old definite descriptions (those where a wordnet-compatible semantic relation was identified and which were reasonably close to their antecedent), and they did not distinguish coreferent from associative bridging antecedents. Although the different evaluation method disallows a meaningful comparison, we think that the more evolved information sources we use (Padó and Lapata's association measure instead of Lund et al's, combined pattern search on the World Wide Web instead of search for patterns in a fixed-size corpus), as well as the additional information based on semantic similarity, lead to superior results when evaluated in a comparable task.

3.1 Ongoing and Future Work

Both the distributional similarity statistics and the association measure can profit from more training data, something which is bound by availability of similar text (Gasperin et al., 2004 point out that using texts from a different genre strongly limits the usefulness of the learned semantic similarity measure), and by processing costs (which are more serious for distributional similarity measures than for non-grammar-related association measures, as the former necessitate parsed input).

Based on existing results for named entity coreference, a hypothetical coreference resolver combining our information sources with a perfect detector for discourse-new mentions would be able to achieve a precision of 88% and a recall of 83% considering all full noun phrases (i.e., including names, but not pronouns). This is both much higher than state-of-the art results for the same data set (Versley, 2006, gets 62% precision and 70% recall), but such accuracy may be very difficult to achieve in practice, as perfect (or even near-perfect) discourse-new detection does not seem to achievable in the near future. Preliminary experiments show that the integration of pattern-based information leads to an increase in recall of 0.6% for the whole system (or 46% more coreferent bridging cases), but the integration of distributional similarity (loosely based on the approach by Gasperin and Vieira) does not lead

[7] Thanks to Tonio Wandmacher for pointing this out to me at GLDV'07.

to a noticeable improvement over GermaNet alone; in isolation, the distributional similarity information did improve the recall, albeit less than information from GermaNet did.

The fact that only a small fraction of the achievable recall gain is currently attained seems to suggest that better identification of discourse-old mentions could potentially lead to larger improvements. It also seems that firstly, it makes more sense to combine information sources that cover different relations (e.g. GermaNet for hypernymy and synonymy and the pattern-based approach for instance relations) than those that yield independent evidence for the same relation(s), as GermaNet and the Gasperin and Vieira approach do for (near-)synonymy; and secondly, that good precision is especially important in the context of integrating antecedent selection and discourse-new identification, which means that the finer view that we get using antecedent selection experiments (compared to direct use in a coreference resolver) is indeed helpful.

Acknowledgements I am very grateful to Sabine Schulte im Walde, Piklu Gupta and Sandra Kübler for useful criticism of an earlier version, and to Simone Ponzetto and Michael Strube for feedback on a talk related to this paper. The research reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (Sonderforschungsbereich) 441 “Linguistic Data Structures”.

References

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, 15(1):83–113.

Cardie, C. and Wagstaff, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 1999)*, pages 82–89.

Clark, H. H. (1975). Bridging. In Schank, R. C. and Nash-Webber, B. L., editors, *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174, Cambridge, MA. Association for Computing Machinery.

Daumé III, H. and Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT/EMNLP'05*, pages 97–104.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Foth, K. and Menzel, W. (2006). Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *ACL 2006*.

Garera, N. and Yarowsky, D. (2006). Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *CoNLL 2006*.

Gasperin, C., Salmon-Alt, S., and Vieira, R. (2004). How useful are similarity word lists for indirect anaphora resolution? In *Proc. DAARC 2004*.

Gasperin, C. and Vieira, R. (2004). Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Geffet, M. and Dagan, I. (2004). Feature vector quality and distributional similarity. In *CoLing 2004*.

Goecke, D., Stührenberg, M., and Wandmacher, T. (2006). Extraction and representation of semantic relations for resolving definite descriptions. In *Workshop on Ontologies in Text Technology (OTT 2006)*. extended abstract.

Harabagiu, S., Bunescu, R., and Maiorano, S. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*.

Hinrichs, E., Kübler, S., and Naumann, K. (2005). A unified representation for morphological, syntactic, semantic and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor.

Kunze, C. and Lemnitzer, L. (2002). Germanet – representation, visualization, application. In *Proceedings of LREC 2002*.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proc. CoLing/ACL 1998*.

Lund, K., Atchley, R. A., and Burgess, C. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *IJCAI 1995*, pages 1050–1055.

Morton, T. S. (2000). Coreference for NLP applications. In *ACL-2000*.

Müller, F. H. and Ule, T. (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*.

Ng, V. (2007). Shallow semantics for coreference resolution. In *IJCAI 2007*, pages 1689–1694.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of ACL 2003*.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, to appear.

Poesio, M., Alexandrov-Kabadjov, M., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*.

Poesio, M., Vieira, R., and Teufel, S. (1997). Resolving bridging descriptions in unrestricted text. In *ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts*.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *HLT-NAACL 2006*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A german computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*.

Steinberger, J., Kabadjov, M., Poesio, M., and Sanchez-Graillet, O. (2005). Improving LSA-based summarization with anaphora resolution. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*.

Versley, Y. (2005). Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.

Versley, Y. (2006). A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.

Versley, Y. (2007). Using the Web to resolve coreferent bridging in German newspaper text. In *Proceedings of GLDV-Frühjahrstagung 2007*, Tübingen. Narr.

Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Weeds, J. and Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterizing measures of lexical distributional similarity. In *CoLing 2004*.