Entity-centric document understanding

Entity aspects and salience

Wu, C.

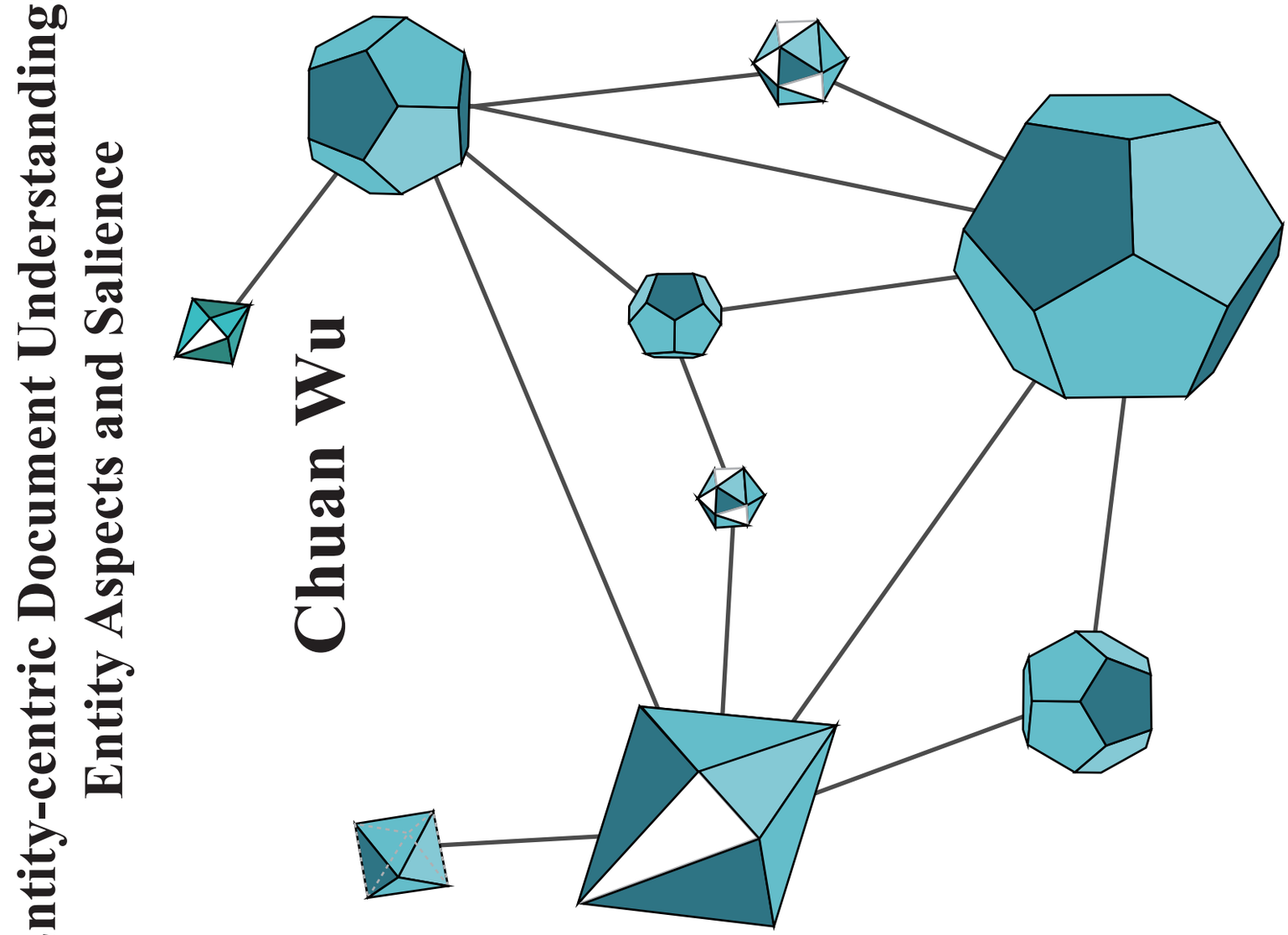
Publication date
2020

Document Version
Final published version

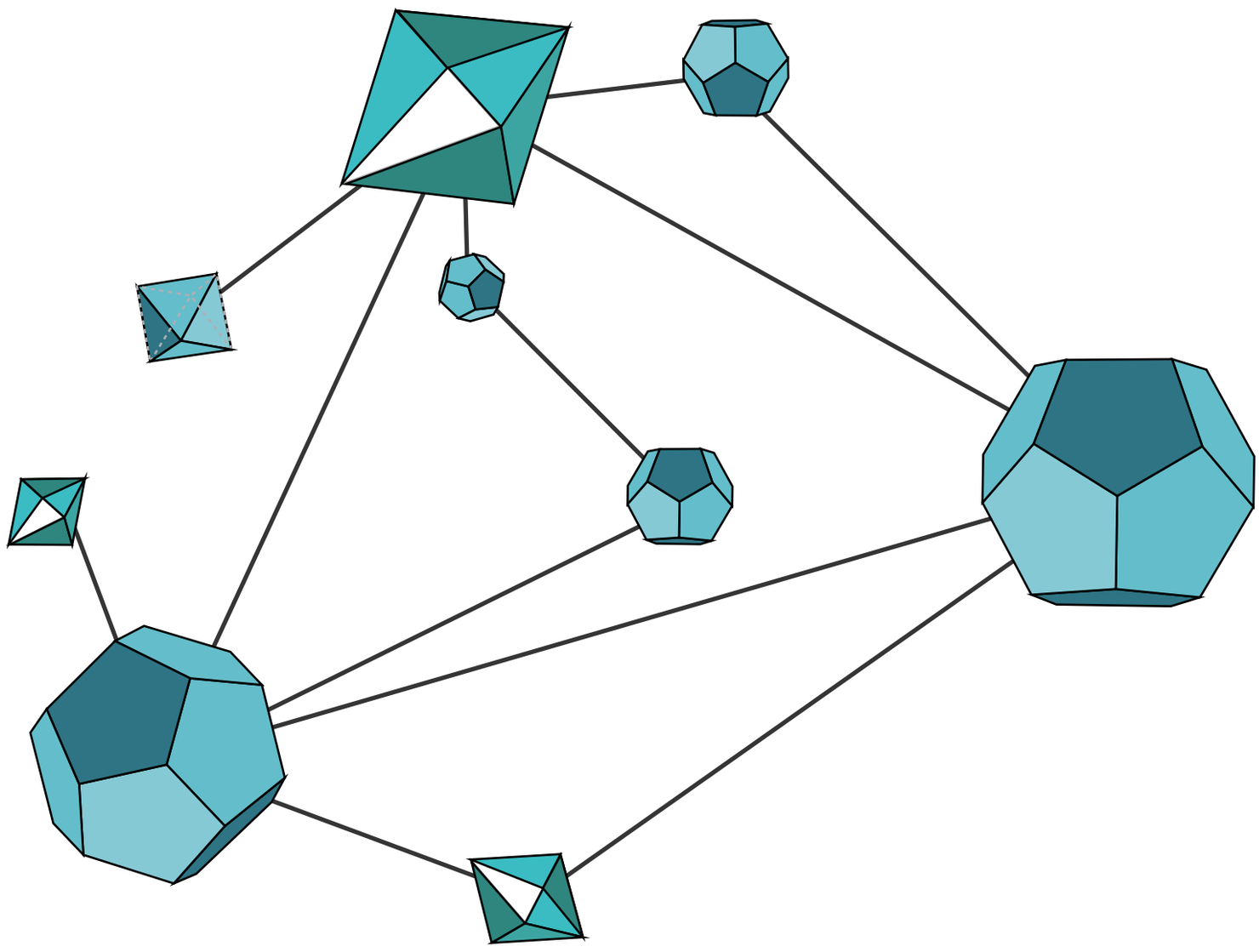
License
Other

Entity-centric Document Understanding: Entity Aspects and Salience

Chuan Wu



Entity-centric Document Understanding: Entity Aspects and Salience Chuan Wu 2020



Entity-centric Document Understanding: Entity Aspects and Salience

Chuan Wu

Entity-centric Document Understanding: Entity Aspects and Salience

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op dinsdag 3 november 2020, te 12:00 uur

door

Chuan Wu

geboren te Hubei

Promotiecommissie

Promotor:		
	Prof. dr. E. Kanoulas	Universiteit van Amsterdam
	Prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:		
	Prof. dr. W. Lu	Wuhan University
Overige leden:		
	Prof. dr. P. Groth	Universiteit van Amsterdam
	Prof. dr. H. Haned	Universiteit van Amsterdam
	Dr. C. Monz	Universiteit van Amsterdam
	Dr. R. Reinanda	Bloomberg
	Prof. dr. Z. Ren	Shandong University

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was supported by the China Scholarship Council.

Cover by Chuan Wu
Printed by Off Page, Amsterdam, The Netherlands

ISBN: 978-94-93197-21-3

Acknowledgement

It has been a long journey since I started my PhD program. I cannot say that I enjoy every second of my PhD life. But it is definitely true that I enjoy many parts of this time period of my life. It is partially because that it is a pleasure to learn new things and keep improving myself. More importantly, it is because of all people that have been around me along the process. You all together make my PhD life a memorable experience.

First of all, I would like to thank my main promotor Prof. dr. Evangelos Kanoulas for his supervision during the past five years. Thank you so much for your patience, guidance and support during all stages of my doctoral journey. You gave me the freedom to explore whatever I was interested at and shared insights on my immature work. It was your encouragement that helped me get through research challenges. Besides work, you also helped me a lot with other things such as sharing advice on getting life work balance. Your help is invaluable and I really appreciate it.

Then, I want to thank my promotor Prof. dr. Maarten de Rijke for his support during my PhD. Thank you for granting me a chance to be in the Netherlands. I benefit a lot from your help and insightful guidance. I was so lucky to be able to join such a fantastic research group and I believe that this is the best place to do a PhD.

I also want to thank my co-promotor Prof. dr. Wei Lu for his support across my PhD life. Thank you for your supervision and support in both research and life. Your patience and guidance helped me get through many challenges. The research group you lead is a great one and I enjoyed my time there a lot.

I am honoured to have Paul, Hinda, Christof, Ridho and Zhaochun to be my committee members. Thanks a lot for agreeing to join my PhD committee.

It was great to work with a group of very talented and interesting people in and around ILPS. I enjoyed spending time with you over tea breaks, our social events, volleyball and so on. The Friday afternoon drinks was a very nice time to share whatever we have in mind. Thanks a lot: Adith, Aldo, Aleksandr, Alexey, Ana, Anna, Anne, Arianna, Artem, Bob, Boris, Chang, Christof, Christophe, Cristina, Daan, Dan, Dat, David, David, Dilek, Eva, Evangelos, Evgeny, Fei, Hamid, Harrie, Isaac, Ilya, Ivan, Jiahuan, Jie, Julia, Kaspar, Katya, Ke, Maarten, Maarten, Maartje, Manos, Marlies, Marzieh, Mostafa, Nikos, Petra, Praveen, Ridho, Tobias, Tom, Xiaojuan, Xinyi and Ziming. Special thanks to Zhaochun and Xinyi for your help on starting my life in the Netherlands and to Ziming and Shuai for lots of helpful discussion on both research and life. Thank Petra for helping me take care of many details about my PhD program.

I benefited a lot from my friends during my doctoral studies. I thank all my friends in Amsterdam and Wuhan for sharing time with me. I thank Bin, Biwen, Huiqin, Jian, Jie, Jiming, Jinglan, Muhe, Que, Songyu, Wei, Wenyang, Xiaojuan, Yiwen, Yang, Yang, Yipeng, Zhiyong for our leisure time together. I thank Heng, Kun, Mengrui, Pengcheng, Qikai, Qiuzi, Shengwei, Sisi, Xiangbin, Xiaohua, Yong for sharing time at Wuhan University and our collaboration in research projects.

I would like to acknowledge the China Scholarship Council (CSC) for the financial support that funded parts of the research discussed in this dissertation. I also appreciate the support from Wuhan University when I was studying there.

Last but not least, I would like to thank my parents for their love, encouragement

and support, especially the effort on providing the best education environment they can for me. Thank my brother for his accompany along growing process and his advice on reading and thinking. Special thanks go to my girlfriend, Ying, for her accompany, understanding, encouragement and love.

Contents

1 Introduction

In the early 1990s, the World Wide Web was born. Content creators were few and the vast majority of users simply acted as consumers of content [17]. With the development of technology, the World Wide Web gradually moved to the era of Web 2.0, which was characterized by a rich user experience and user participation [17]. Lots of sites, including social networking sites or social media, blogs, and wikis, allowed users to create content (user-generated content). Since then, the volume of information has increased dramatically and has led to the well-known crisis in the modern information age: information overload.

The mismatch between the capability of information processing and the amount of information to be processed creates significant cognitive challenges. Since it is impossible and probably undesirable to slow down the process of generating new information, the only choice is to improve information processing. For example, in library science, subject indexing is the process used for describing the subject matter of documents [69]. It involves identifying terms to represent what documents are about. Subject indexing can be done manually when newly incoming documents are limited in number. However, this became expensive and inefficient soon after the arrival of the information age. To alleviate this issue, a task called keyword extraction [39], was proposed by researchers to automatically identify terms that best describe the subject of a document.

Automated information processing techniques make it possible to efficiently deal with large amounts of documents in ways that enhance our understanding of these documents. In addition to keyword extraction, there are many other automated information processing tasks. We categorize them into two categories: *coarse-grained* tasks and *fine-grained* tasks. Coarse-grained tasks focus on the understanding of a document as a whole, while fine-grained tasks deal with elements at a finer level of granularity, such as words, sentences and paragraphs. Examples of coarse-grained tasks are text classification and document representation learning. Examples of fine-grained tasks include spelling correction, part-of-speech tagging, named entity recognition, and sentiment analysis.

In the early days of computing over natural language data, the representation of text was more a design choice. Bag-of-words (BoW) was a simple document representation method that is widely used in natural language processing and information retrieval, representing a document as the set of its words, disregarding grammar and word order

but keeping multiplicity[1]. In 2003, latent Dirichlet allocation [11] was proposed based on probabilistic latent semantic analysis (PLSA) [40], to better discover the abstract topics that occur in a collection of documents. As a byproduct, documents can be represented by a multinomial distribution over learned topics (topic distribution). Since then, it has become popular to represent documents using topic distributions for text processing tasks, such as document classification. In recent years, the prevalence of neural networks has also led to new document representation methods, including doc2vec [57] and BERT [23].

Most document representation approaches are based on words in text. However, sometimes words are not just strings. Instead, words might refer to real world things (entities). *Named-entity recognition* is the research field which seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc. [76]. Entities are not of the same importance in documents. In other words, some entities are more important than others in a document. This observation has led to research on entity salience detection [26, 32, 88], which aims at identifying the *salient entities* in a document among all entities mentioned.

Capturing *document aboutness* has been a key research focus throughout the history of automated information processing [81]. Document aboutness aims at creating a succinct representation of a document's subject matter, such as keywords, named entities, concepts, and sentences [13]. People have been working on keywords and keyphrase extraction to reflect the central theme of documents. In recent years, more work is centered around entities rather than keywords or keyphrases. Advances in named entity recognition make it possible to enhance document understanding on the basis of entities. As a result, the necessity of identifying entity salience has been recognized. As an example, Gamon et al. [33] suggest that identifying salient entities should be the first step towards understanding document aboutness.

In fine-grained information processing tasks, named-entity recognition has attracted much interest [76]. Named-entity recognition helps us differentiate entities from words in documents. However, when a named-entity appears in a different document, it is also important to understand which precisely real-world thing it refers to.

Researchers have developed entity repositories so that one not only recognizes entities but also links the occurrences of entities to their entries in the repository. In this way, entities in different documents are semantically connected and documents are connected and enhanced by the attributes of these entities which enhances the ability to better understand documents.

Wikipedia[2], born in 2001, potentially serve as such an entity repository. Wikipedia was designed as a free online encyclopedia, where everyone can contribute by writing and editing. The number of articles in Wikipedia is always increasing and articles get updated frequently by active users. Articles in Wikipedia can be viewed as entities and contents of these articles act as representations of these entities. A new task, i.e., entity linking [102], has been proposed to recognize text fragments (entity mentions) that represent entities and link them to corresponding entries in an entity repository, such as Wikipedia and YAGO[3].

[1] https://en.wikipedia.org/wiki/Bag-of-words_model

[2] https://en.wikipedia.org

[3] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-

Entity linking approaches usually match between the contents of documents to that of entity representations, e.g., Wikipedia articles. Documents are usually related to one or many topics, while entities can also include multiple aspects. A document might be related to one aspect of an entity in the document, rather than all aspects of the entity. It make sense to link an entity mention of an entity to a particular aspect of the entity. Entity aspect linking is proposed by Nanni et al. [77] and defined as follows: given an entity-mention in a specific context (e.g., in a tweet, a sentence or a paragraph), our goal is to link it to one from a set of predefined aspects that captures the addressed topic. In their setting, it is assumed that each of these entity-aspects is accompanied by a textual description and a heading. Even though Nanni et al. [77] claim that their approach is general and applies to any source of predefined aspects, from biomedical catalogs (e.g. MeSH[4]) to historical knowledge resources [77], in their work they exclusively focus on predefined aspects extracted from Wikipedia. The task can be viewed as an application of leveraging entity aspects to enhance the semantics of documents.

In this thesis, we explore entity aspects and entity salience based solutions to enhance document understanding. Specifically:

(1) First, we study how to learn entity-centric document representations by introducing entity aspects. In particular, we model each entity using multiple aspects, where each entity aspect is represented as a mixture of latent topics. A novel graphical model is proposed in order to learn entity-centric document representations, entity facets, and latent topics.

(2) Second, we propose an attentive interaction based convolutional neural network for entity aspect linking.

(3) Third, we propose a novel topic model that takes salient entities into consideration in the document generation process. Specifically, we model salient entities as a source of topics used to generate words in documents, in addition to the topic distribution of documents used in traditional topic models.

(4) Fourth, we present a new dataset, which can be used to benchmark tasks such as entity salience detection and salient entity linking.

1.1 Research Outline and Questions

The main themes of this thesis concern entity aspects and entity salience. We aim to address the following broad question: *how can we enhance document understanding by entity aspect and entity salience information?* To address this question, we consider the following directions: document representation, semantic annotation for documents, and document generative process. Specifically, the thesis covers four research themes:

(1) Learning entity-centric document representations (Chapter 2).

(2) Improving entity aspect linking using a neural network based approach (Chapter 3).

naga/yago/

[4]https://www.nlm.nih.gov/mesh/

(3) Incorporating entity salience information into topic modeling (Chapter 4).

(4) Extracting entity salience annotations from WikiNews (Chapter 5).

1.1.1 Learning entity-centric document representations

Document representation is a central concern for document understanding tasks, such as text classification. Traditionally, documents are represented as bag-of-words [97], topic distributions [11], and so on. With advances in named entity recognition and entity linking [102], it is possible to take entities into consideration when representing documents. Our first study focuses on entity related document representation. Existing work on entity related document representation mostly considers shallow representations, which treat entities as special tokens different from words [91, 122]. However, entities are not just another type of token for documents. Importantly, documents are usually stories that are centred around a set of entities.

We begin our investigation by considering entities as sources of information for documents. The motivation is that documents that are relevant to entities usually reflect particular aspects of these entities. Inspired by earlier work on the author topic model and entity aspects, we formulate the task of learning entity-centric document representations, which represents documents from the perspective of entities. We assume that entities are multi-faceted and can be modeled by multiple aspects. Each aspect of an entity is considered as a particular topic specific to the entity. Then we ask the following research question:

RQ1 Can we learn entity-centric document representations by modeling entities with multiple aspects?

1.1.2 Improving entity aspect linking using a neural network based approach

In our second study, we move in a different direction: semantic annotation for documents, which aims at enriching documents with semantic information [2, 6, 52, 75, 102]. For example, Al-Bukhitan et al. [2] construct a collection of full-text biomedical journal articles to facilitate research on the construction of ontology for semantic annotation of biomedical documents. A research direction close to our study is entity linking, which links entity mentions (text fragments that represent entities) in text to their corresponding entry in a knowledge base [102]. However, entities are usually multi-faceted and one document related to an entity is likely to be topically relevant to one aspect rather all aspects of the entity. Ideally, we expect to link not only to a particular entity but also to a specific aspect of the entity.

We study entity aspect linking in this chapter and view it as a pairwise semantic matching problem. We assume that matching signals are encoded in interactions between entity contexts and candidate aspects. Besides, convolutional neural networks have proved to be useful for identifying local matching patterns. In this study, we ask the following research question:

RQ2 Can we improve entity aspect linking using a convolutional neural network based approach?

1.1.3 Incorporating entity salience information into the document generative process

We move on to the second theme: leveraging entity salience for document understanding. We focus on the fact that the importance of different entities in a document varies across entities. In other words, some (salient) entities in a document are more important for the document than other (non-salient) entities. Intuitively, this leads to two questions: (1) how can we identify salient entities out of all entities? (2) how can we make use of the entity salience information to enhance document understanding? In our third study, we attempt to leverage salient entities in documents to improve the modeling of document generative process. Here, instead of considering how to identify salient entities among all entities, we assume binary entity salience annotations. That is to say, we know in advance whether a given entity in a document is salient or not.

We study the problem of incorporating entity salience information into the document generative process to improve topic modeling. We assume that stories in documents are built upon a story line (topic) and a set of main characters in the story (salient entities). As an example, imagine that a news reporter is writing a news article about a specific story. The primary point under consideration is what the document is about. The second point is about which entities are salient entities in the story described in the document. And finally, other words and entities are added to the document to complete the story. In this study, we ask the following research question:

RQ3 Can we improve our understanding of the document generative process by incorporating entity salience information into topic modeling?

1.1.4 Extracting entity salience annotations from WikiNews

Driven by the need to support research related to entity salience, such as entity salience detection and salient entity linking, we set out our last study. We focus on the extraction of entity salience annotations in documents. Specifically, we aim to extract entity salience annotations in the context of WikiNews. We begin our investigation by considering structural information associated with documents within WikiNews, which can be helpful for deciding the salience of entities in news articles.

Traditionally, entity salience annotations are obtained by crowdsourcing [24, 109], which is costly and time-consuming. To address the limitations of previous approaches, we examine the structure of WikiNews. In each article, text fragments referring to entities are linked by the article authors to Wikipedia pages corresponding to the respective entity or WikiNews categories. Inspired by the author guidelines and the organization structure of WikiNews, we view the category annotations made by writers of news articles as an important signal that indicates entity salience. We ask the following research question:

RQ4 Can we automatically extract entity salience information from WikiNews?

1.2 Main Contributions

In this section, we list the theoretical, algorithmic and empirical contributions of this thesis. For each contribution, we list the chapter from which it originates.

1.2.1 Theoretical contributions

(1) We formulate the task of learning entity-centric document representations (Chapter 2).

(2) We propose a novel pooling strategy (q-singular pooling) to extract features from outputs of a convolutional neural network (Chapter 3).

(3) We propose a strategy and process of automatically extracting entity salience annotations from WikiNews (Chapter 5).

1.2.2 Algorithmic contributions

(4) We propose a novel topic model, the Entity Facet Topic Model (EFTM) to learn entity-centric document representations. We derive a Gibbs sampling algorithm for parameter estimation (Chapter 2).

(5) We propose a novel convolutional neural network based model for entity aspect linking, the Multi-Interaction based Convolutional Matching Network (MICMN) to perform semantic matching between entity context and candidate aspect (Chapter 3).

(6) We propose a novel topic model, the Salient Entity Topic Model (SETM), which takes entity salience into consideration in the document generative process. We derive a Gibbs sampling algorithm for parameter estimation (Chapter 4).

1.2.3 Empirical contributions

(7) We confirm our hypothesis regarding the existence of multiple facets of an entity by analyzing the learned entity facets using qualitative and quantitative analysis, and identify an effective number of facets per entity (Chapter 2).

(8) We demonstrate the effectiveness of EFTM in downstream applications using a multi-label classification task (Chapter 2).

(9) We demonstrate the effectiveness of MICMN for entity aspect linking. The superiority of q-singular pooling is also demonstrated by an ablation study (Chapter 3).

(10) We demonstrate the effectiveness of SETM to model text by performing both a qualitative and a quantitative analysis (Chapter 4).

(11) We analyze the entity salience dataset we built and compare it with previous entity salience datasets. We conduct experiments to demonstrate the utility of the dataset (Chapter 5).

1.2.4 Resource contributions

(12) We construct a new dataset with entity salience annotations (Chapter 5).

(13) We make the implementation of our model, the Salient Entity Topic Model (SETM), publicly available at here[5] (Chapter 4).

1.3 Thesis Overview

This thesis starts with an introduction. We then turn to four main research chapters that are divided into two parts as described in the previous section. Finally, the thesis closes off with the conclusions and bibliography.

In this thesis, we investigate enhancing document understanding with entity aspect and entity salience. In particular, in Chapter 2 we learn entity-centric document representation using entity facet topic model. Then in Chapter 3 we propose a convolutional neural network based approach for entity aspect linking. In Chapter 4, we propose a novel topic model that incorporates entity salience information into document generative process. In Chapter 5, we propose an automated approach to extract entity salience annotations from WikiNews. Finally, in Chapter 6 we conclude the thesis and discuss limitations and future directions.

1.4 Origins

In this section, we list the publications that form the basis of this thesis. Each research chapter is based on a paper. We provide references to these publications and explain the roles of the co-authors.

Chapter 2 is based on C. Wu, E. Kanoulas, and M. de Rijke. Learning entity-centric document representations using an entity facet topic model. *Information Processing & Management*, 57(3), 2020. CW designed the algorithm, ran the experiments and did most of the writing; EK helped with writing, algorithm design; MdR contributed to the writing.

Chapter 3 is based on C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. A multi-interaction based convolutional matching network for entity aspect linking. In *COLING*, 2020 (under review). CW designed the algorithm, ran the experiments and did most of the writing; EK helped with writing, algorithm design; WL contributed to the writing; MdR contributed to the writing.

Chapter 4 is based on C. Wu, E. Kanoulas, and M. de Rijke. It all starts with entities: A salient entity topic model. *Natural Language Engineering*, pages 1–19, 2019. CW designed the algorithm, ran the experiments and did most of the writing; EK helped with writing, algorithm design; WL contributed to the writing; MdR contributed to the writing.

[5] https://github.com/setm2nle/salient-entity-topic-model

Chapter 5 is based on C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. WN-Salience: a corpus of news articles with entity salience annotations. In *LREC 2020*, pages 1–8. LREC, 2020. CW built the dataset, ran the experiments and did most of the writing; EK helped with experiment design and writing; WL contributed to the writing; MdR contributed to the writing.

2

Learning Entity-Centric Document Representations using An Entity Facet Topic Model

In the previous chapter, we have introduced the background material for this thesis. Starting from this chapter, we begin our research and answer the research questions listed in Chapter 1. In this chapter, we address RQ1, which is concerned with learning entity-centric document representations.

2.1 Introduction

Understanding the content of documents by learning semantic representations can benefit various downstream applications, such as information retrieval [19] and text classification [100]. Existing document representation methods include (1) traditional bag of words (BoW), which represent documents using term frequencies; (2) topic distributions [11], which represent documents using mixed distributions of latent topics; and (3) dense vector representations [57], which represent documents as points in a low dimensional space.

Existing document representation methods assume a representation in terms of the semantic topics discussed in the document. However, when it comes to understanding a document from the perspective of entities, it is natural to think of a document in terms of the facets of the different entities it relates to. In other words, in this chapter we hypothesize that entities are not monolithic concepts; instead they have multiple facets, and different documents may be discussing different facets of a given entity. Given that, we argue that from an entity-centric point of view, a document related to multiple entities shall be (a) represented differently for different entities (multiple entity-centric representations), and (b) each entity-centric representation should reflect the specific facets of the entity discussed in the document. Let us illustrate this hypothesis with an example. Political world leaders, as entities, may have different facets of them described in news articles, such as family, foreign policy, campaigning, economy, etc.

This chapter was published as C. Wu, E. Kanoulas, and M. de Rijke. Learning entity-centric document representations using an entity facet topic model. *Information Processing & Management*, 57(3), 2020.

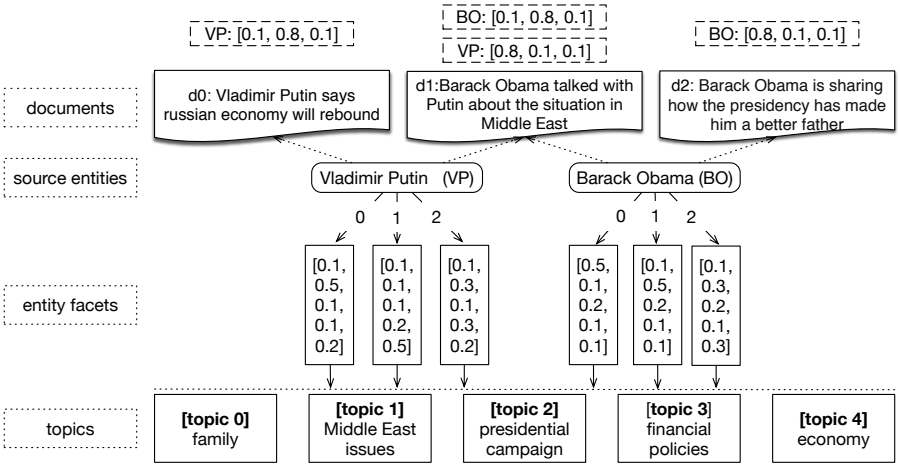


Figure 2.1: Illustration of an entity-centric document representation. Documents d_0, d_1 and d_2 are associated with the sets of source entities {VP}, {BO, VP}, and {BO}, respectively. Each source entity has three facets. Each entity facet is a distribution over five latent topics, where each element of the distribution indicates the relatedness between the facet and corresponding topic. The entity-centric document representations are mixed proportions of facets shown above the corresponding documents.

A document related to political world leaders is expected to reflect only specific facets of these leaders. For example, a document discussing a meeting between the presidents of the USA and Russia in 2016 might mention multiple entities, such as Obama and Putin. Economic and foreign policy facets of these entities are probably reflected in the document, but it is unlikely that the presidential campaign facet or family facet are discussed.

The main objective of our work is to explore representing documents based on entity facets. In particular, to accurately represent a document, we propose to use entity specific topics that reflect the facet of the entity discussed in a document, which we call *entity facets*. Further, we define an *entity-centric document representation* as a distribution over entity facets. Each entity facet is further defined as a distribution over latent topics. Continuing our world leader example, Fig. 2.1 shows an example of an entity-centric document representation involving the two leaders. Document d_1 is associated with two entities, while d_0 and d_2 are each associated with a single entity. The entity-centric representation of each document is shown within the dashed rectangles above the documents; for every associated entity, a distribution over its facets is learned for that entity. As we can see, facet 1 of Barack Obama (BO) and facet 0 of Vladimir Putin (VP) are reflected in d_1, while for d_0 and d_2, the entity-centric representation also indicates the facet reflected for the corresponding source entity, i.e., VP and BO, respectively.

We propose a new task, that of entity-centric document representation learning. For a document associated with multiple entities, multiple facet distributions are learned for the document. To understand our modeling decisions and the contribution that we

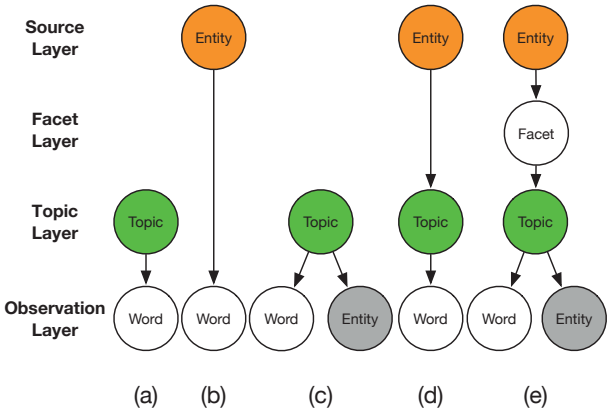


Figure 2.2: Different dependencies among words, entities, topics, facets, sources. The simplified plate representations of the models are: (a) LDA; (b) Author Model (AM); (c) Link-LDA; (d) Author Topic Model (ATM); (e) Entity Facet Topic Model (EFTM).

make, we list different topic models, with their prime ingredients: words, entities, topics, and facets in Fig. 2.2. Previously proposed topic models have considered words and topics (e.g., LDA, Fig. 2.2(a)); words and entities-as-sources-of-information (the Author Model, Fig. 2.2(b)); words, entities-as-observables and topics (Link-LDA, Fig. 2.2(c)); and words, topics, and entities-as-sources (the Author Topic Model, Fig. 2.2(d)). We propose the *Entity Facet Topic Model* (EFTM) (Fig. 2.2(e)) as a model for learning entity-centric document representations. In a generative perspective on representation learning, we model entities as a source of information (*source layer*); such source entities are assumed to be given and could be labels or metadata assigned to a document, or in-text entities; they consist of multiple entity facets (*facet layer*). Each facet is built upon general latent topics (*topic layer*), which are further used to generate observed variables (*observation layer*) in documents. Since different types of observed variables (e.g., words and entities) might appear in documents [27] (Fig. 2.2(c)), we consider both words and entities. To differentiate between entities in the source layer and the observation layer, we refer to the former as source entities, and the latter as document entities.

To illustrate how our work can be used in practical applications, we present the following examples.

Example 1. Knowledge Base Construction/Population. Fetahu et al. [30] propose to populate Wikipedia entity pages by automated news article suggestion. Their approach consists of two steps, identifying articles relevant to entities and then connect relevant articles to particular sections. By learning entity-centric document representations using news articles labeled by their related entities (source entities), one might be able to cluster news articles into smaller groups on the per-entity basis. If groups of articles match contents of particular sections of a Wikipedia page, we might update the section. If it does not match, we might update the Wikipedia page by adding new

sections using groups of articles identified as references.

Example 2. Entity Salience Detection. Dunietz and Gillick [26] propose a new entity salience task called entity salience detection, which aim at identifying whether an entity is salient in a document it appears in. If a particular facet of an entity is reflected in a document, it is likely to play an important role in the document, and thus to be considered as salient for the document. By learning an entity-centric document representation of the document using a model trained by available datasets, we can perform binary classification from the perspective of the entity and tell whether the document matches any facet of the entity.

Example 3: Personalized News and Blog Recommendation. Kazai et al. [50] present a prototype mobile app that provides personalized content recommendations to its users by combining various user signals. In the application, individual models are built for each user. Similarly, our model learns individual facets for each user from news and blogs that each user is interested at. Then, entity(user)-centric representations of documents are inferred and judged whether it should be recommended to each user.

Summarizing, EFTM models entities as sources of information along with their multi-faceted properties. At the same time, EFTM represents documents on the basis of the facet distributions of the source entities, which provides the desired entity-centric representations for documents. Our work aims at facilitating information processing applications, in which entities are central towards understanding, and modeling text. Examples include multi-authored article collections, multi-labeled textual collections and so on.

The main contributions of this chapter are as follows:

1. We propose the task of entity-centric document representation learning.
2. We propose a novel Entity Facet Topic Model (EFTM) to learn entity-centric document representations.
3. We confirm our hypothesis regarding the existence of multiple facets of an entity by analyzing the learned entity facets using qualitative and quantitative analysis, and identify a effective number of facets per entity.
4. We demonstrate the effectiveness of EFTM in downstream applications using a multi-label classification task.

2.2 Related Work

We discuss three lines of related work: document representation, topic modeling, and entity facet mining.

2.2.1 Document representation

Research on document representation dates back (at least) to the early days of the vector space model, which represents documents as vectors of index terms. Index terms are weighted to capture the importance of a term in describing the content of a

particular document (e.g., term frequency) and the discriminative power of a term (e.g., inverse document frequency). Many term weighting methods have been proposed to achieve better document representations, including the widely used TF-IDF method. Bouadjenek et al. [12] propose to integrate social information of users in the index structure of an IR system. The index model provides a Personalized Social Document Representation of each document per user based on his/her activities in a social tagging system. Similar with our work, they start from the intuition that each user has his/her own understanding and point of view of a given document. However, our work differs from their work in that our entity-centric document representations are based on latent entity facets learned from document collections, while theirs are based on associated social annotations of users (entities).

In contrast to representing documents using weighted terms, dense vector representations became popular since the prevalence of LDA [11]. In LDA, a document is represented by a mixture of latent topics, which are further represented by multinomial distributions of words and shared across all documents. By using topic distributions to represent documents, the topical difference between documents is captured. Extensions of LDA usually model different document generative processes, while using topic distributions to represent documents. Traditionally, a single topic distribution is learned for each document and applied whenever document representations are used. Our work aligns with this line of research and differs from existing work in that we propose to learn multiple entity-centric representations for each document, where each representation corresponds to an entity and is a mixture of facets of the entity.

Today, latent document representations such as doc2vec [57] are often inferred using neural networks. Doc2vec is an extension of word2vec [73] that learns document-level embedding to predict the next word given contexts sampled from a document. Van Gysel et al. [114] propose a neural vector space model (NVSM), which learns document representations directly by gradient descent from sampled n-gram/document pairs extracted from a given corpus. Compared to our proposed representation, doc2vec and NVSM consider neither entities in documents nor entities as sources of information. Recent work on contextualized word representations, such as BERT [23] and ELMo [84], is becoming increasingly popular. The major difference between our work and theirs is that we consider document representations with respect to entities, while they infer word representations with respect to contexts around words.

Recent work has extended the embedding paradigm to entities [112, 113]; however, their focus is restricted to the semantics of entities, whereas we focus on the relation between entities, entity facets, topics and documents. Xiong et al. [123] propose a word entity duet representation for ad-hoc retrieval, which uses entity based representation of documents. Dai et al. [20] propose to learn an entity mention aware document representation, which learns representations of documents from semantics between not only document-word pairs, but also document-entity pairs and entity-entity pairs. Raviv et al. [91] devise an entity based language model which takes into account both the uncertainty inherent in the entity-markup process and the balance between entity-based and term-based information. Their work differs from ours in that they are using entities in documents as a bag-of-entities representation, while our work considers learning multiple document representations, each of which corresponds to an entity semantically relevant to the document.

2.2.2 Topic models

LDA [11] models a document as a mixture of topics, and automatically generates summaries of topics in terms of a multinomial distribution over words. Many extensions based on it have been proposed to learn topics by assuming an alternative document generative process, thus leading to alternative semantics of learned document representations. E.g., Link-LDA [27] (Fig. 2.2(c)) extends LDA by modeling words and references (viewed as entities) of an article separately.

In recent years, research has focused on extending topic models to address specific tasks, such as short text modeling [9, 65, 131], user clustering [87], dataless text classification [59], and opinion mining [68]. In contrast to designing topic models for specific tasks, our work aims at mining entity facets so as to learn entity-centric document representations, which could be used in downstream applications. Therefore, our work is particularly related to topic models that either considers entities in documents or external labels associated to documents.

The Author Model (AM) [71] (Fig. 2.2(b)) and Author Topic Model (ATM) [95] (Fig. 2.2(d)) have been proposed for multi-labeled documents, where each label (author) is viewed as an entity. In AM, words are generated by first selecting an author and then sampling from an author-specific multinomial distribution over words. ATM extends AM by introducing topics between authors associated with documents and words in documents. In particular, ATM chooses a latent topic from an entity-specific multinomial distribution over topics; a word is then drawn from a topic-specific multinomial distribution. Similar to AM and ATM, we consider labels of documents as source entities. However, ATM focus on learning representation for entities (authors) and does not learn document representations, while our model aims at learning document representations. There are other entity topic models, such as an entity topic model for entity linking [38], and a hierarchical entity topic model designed for streaming data [43]. Newman et al. [79] propose CorrLDA2 to learn the relationship between topics discussed in news articles and entities mentioned in articles. Chang et al. [14] propose a topic model that analyzes free text to extract descriptions of relationships between entities. These models differ from our work in that they aim at resolving particular tasks, while we focus on modeling entities to learn better document representations.

The Entity Topic Model (ETM) [51] represents entities in the same way as latent topics. For each document, a topic distribution is drawn from a Dirichlet prior and a joint multinomial distribution over words Φ is obtained by linearly combining entities and topics of a document. To generate a word, a topic is sampled from Φ and a word is sampled from the topic word distribution. Though ETM seems to be a valid baseline for our work, it is not applicable because of scalability issues. In particular, given N words, E entities, F facets and K topics, the number of parameters of ETM is $K \times N + E \times N + E \times K \times N$, which represents latent topics, entity topics and entity-topic pairs, while that of our model is $K \times N + E \times F \times K$, which represents latent topics and entity facets. The number of parameters of ETM increases fast with increasing numbers of entities because of the component $E \times K \times N$. For example, given 10 entities, 100 topics and 10,000 words, the number of parameters becomes 10 million. In comparison, $E \times F \times K$ is much smaller than $E \times K \times N$ due to the fact that $F \ll N$.

A different line of extensions to LDA focuses on leveraging supervised labels. The supervised topic model in [70] is designed for single-labeled documents, while our model is mainly designed for multi-labeled documents. Though also designed for multi-labeled documents, the hierarchically supervised topic model proposed in [83] considers a scenario where labels of documents form a hierarchy, while our models do not consider a label hierarchy. One work considering both supervised and flat labels is Labeled LDA [89]; it focuses more on credit attribution within tagged documents or visual analysis, instead of learning a better document representation. The constraint that one label corresponds to one topic helps in their task but limits the representation ability of Labeled LDA in that a label is a high granularity semantic unit that might have various facets. Compared to Labeled LDA, we consider the labels of a document as source entities, which themselves consist of smaller semantic units, i.e., entity facets. The advantage of our model is that different facets related to labels are captured by entity facets, instead of being mixed under the same topic. Other topic models that are relevant to our work include DFLDA [64] and CPTM [63]. Since these models target the task of multi-label classification and make use of global prior information, such as label frequency, we do not consider them as baselines.

Another family of topic models that look similar to our model are topic models with a hierarchy of topics. To generate a document using the hierarchical topic model [34], a path with L nodes from the root node of a tree to a leaf is selected and a vector of topic proportions θ is drawn from an L-dimensional Dirichlet distribution. Then, words in the document are drawn from a mixture of the topics along the path with mixing proportions θ. Since different nodes in the topic hierarchy represent different topics, the semantics of the topic proportions representation of different documents are different. While somewhat similar to a two-layer hierarchy of topics, our model is different in that the second layer of topics in our model are entity facets (entity specific topics), which makes it specific to entities compared to general topics defined in a two layer hierarchical topic model.

2.2.3 Entity facet mining

With the growing importance of semantic search and knowledge graphs in recent years, mining and leveraging entity information has received considerable attention [7, 8]. Among various categories of information of entities, such as facts and entity relations, entity facets are also considered useful for entity related tasks; e.g., Reinanda et al. [93] use entity aspect similarity as a feature to help filtering documents for long-tail entities.

Significant work in entity facet mining has been conducted in relation to product facets and online reviews. Applications include product related QA [129], online review mining [3, 25, 120], review summarization [67]. Titov and McDonald [107] propose to first extract facets of objects from online user reviews and then cluster them into coherent topics. Yu et al. [128] automatically identify important product (entity) aspects from online consumer reviews. Sikchi et al. [104] propose an aspect based product comparator to help consumers in purchasing decision making. Yu and Lam [129] propose to learn aspects of product categories to predict answers for product-related questions. Li et al. [60] develop an event-aspect topic model to cluster sentences into aspects for events. The related work above focuses on a particular category of entities,

while our work is targeted on mining facets for general purpose entities related to textual documents.

In addition to focusing on particular categories of entities, there is also work on mining facets of general purpose entities from various sources, such as Wikipedia [30, 77], query logs [92] and microblog posts [105]. Spina et al. [105] propose to identify entity aspects from social web streams (such as tweets) in the field of online reputation management. Given all queries containing an entity, Reinanda et al. [92] propose to obtain *query contexts* by removing mentions of the entity in queries. Then all query contexts are clustered and entity facets are identified as clusters which includes similar query contexts. Our work differs from existing work in that we identify entity facets from document collections where documents are associated to entities. The major advantage is that information related to entities can be widely and mostly found in textual collections.

On the other hand, our work differs from existing work in terms of facet representation. In existing work, entity facets are represented by a bag of text segments [92, 105], textual description [77], or sentence patterns [61]. Spina et al. [105] consider terms as aspects and try to rank aspects that are being discussed with respect to a given company. Li et al. [61] propose a model to perform clustering, and use the clustered sentences and words as aspects. Nanni et al. [77] directly use sections in Wikipedia pages as aspects of corresponding entities. In comparison, our work represents entity facets as a mixture of latent topics and uses it as the basis of entity-centric representations. The facet representations are useful in existing work with regard to their end goal. However, our choice is advantageous in our case for the following reasons. First, we do not target particular categories of entities, which makes it impossible to simply find sentence patterns by clustering. Second, our end goal is to learn entity-centric representations, which are based on entity facets. By representing entity facets as a mixture of topics, we can jointly learn entity-centric representations and entity facets using our proposed topic model.

Overall, we extend the state-of-the-art in three ways: a new task (learning entity-centric document representations), a new topic model to address this task (the entity facet topic model), and a new way of capturing and reasoning about entity-related facet information.

2.3 Research Objectives

The key objective of this chapter is to model documents with respect to the specific facets of the entities that are discussed in each document. In particular, our work derives from the hypothesis that entities are not monolithic concepts, but instead have different facets, and documents associated with certain entities discuss these entities from a specific facet perspective. Our goal is to automatically identify entity facets from documents and derive multi-faceted entity-centric representations of the documents in a collection.

We set forth the following research objectives:

RO1.1 *Modeling entity facets, as a mixture of latent topics, and learning them from documents associated to these entities.*

Our first research objective aims to set up a theoretical definition of entity facets based on latent topics. By defining entity facets as a mixture of latent topics, we connect the specificity of entities (entity facets) to the generality of documents (topics in documents).

RO1.2 *Learning multiple entity-centric document representations based on entity facets.*

The focus of our second research objective is to model the generative process of documents as a joint effort of particular facets of entities. In this way, we attempt to learn both facets and representations of documents which are based on these facets.

RO1.3 *Confirming that considering entity facets has practical implications in downstream applications.*

The focus of the third objective is to understand whether considering entities as multi-faceted concepts makes a difference when it comes to downstream applications, i.e., whether denoising entity and document representation by focusing on specific facets discussed in a document can help in applications such as text classification.

RO1.4 *Identifying by means of predictive modeling what should one consider to be an effective number of facets an average entity has, and how many topics, in traditional terms, are good enough to effectively define these facets.*

The focus of the last objective is to gain a better understanding of how many facets an average entity has within a collection of documents. Clearly, different entities may have different numbers of facets, but in this work, we focus on what is the effective number of them on average. Further, one could explore different ways to validate the number of facets. In this work we focus on the predictive power of the multi-faceted entity-centric document representation to fulfil this objective.

To address the above objectives, we propose to learn entity facets and entity-centric document representations, where each representation corresponds to an entity and is a mixture of entity facets of the entity.

2.4 Problem Formulation

The task of *entity-centric representation learning* is formulated as follows. Given a collection of documents D, in which each document d consists of a bag of words $\mathbf{w}_d$ and a bag of entities $\mathbf{e}_d$, and is associated with a set of source entities $\mathbf{S}_d$, our goal is to learn entity-centric document representations for all documents in D. The elements in $\mathbf{w}_d$, $\mathbf{e}_d$, and $\mathbf{S}_d$ belong to a word vocabulary V_W, an entity vocabulary V_E, and a source entity set S, respectively. The association between source entities and documents is assumed to be predefined. Table 2.1 lists the main notation we use.

Here, we define entities as unique identifiers, such as tags of pictures and entities in documents as represented by identifiers in a collaborative knowledge base (e.g., machine IDs in Freebase). Source entities are entities that satisfy the following two conditions:

Table 2.1: Notation.

Symbol	Description
D	document collection
V_W	word vocabulary
V_E	document entity set
S	source entity set
$\mathbf{w}_d$	bag of words in document d
$\mathbf{e}_d$	bag of entities in document d
$\mathbf{S}_d$	set of source entities associated with document d
s_d	s-th source entity in d
F	number of entity facets
K	number of topics
f_s	facet f of source entity s
z_s	topic selected from f_s
ϕ_k	word distribution of topic k
ψ_k	document entity distribution of topic k
Φ^k	topic token distribution of topic k
η_s	weight of ϕ_s when doing linear combination
$\rho_{f,s}$	facet topic distribution of f_s
θ_s	multinomial distribution of facets of s
$w_{d,i}$	the i-th word in document d
$e_{d,j}$	the j-th entity in document d

(1) source entities are topically multi-faceted; (2) each source entity is associated with a group of documents. For example, in Example 1 in section 2.1, entities can be viewed as source entities because entities usually have multiple facets (multi-faceted) and each entity is related to many documents that are centered around them. Note that source entities are usually different from document entities. For example, tags in news articles, or authors of papers are considered as source entities, while mentions of people or location entities in news articles are document entities. However, the sets of source entities and document entities could also overlap or be identical. For example, if someone wants to learn facet information of Freebase entities in an entity-annotated document collection, they can define the source entities to be the salient entities of a document.

To make matters concrete, Fig. 2.1 provides an example of the entity-centric representation learning task. Given three documents, d_0, d_1 and d_2, which are associated with source entities {VP}, {BO, VP} and {BO}, respectively, our goal is to learn a document representation of d_0 for VP, of d_1 for VP and BO, and of d_2 for BO, as shown in the figure.

2.5 Method

In this section, we introduce our method for learning entity-centric document representations. We first provide an overview of how we define entity-centric document representations. Then we propose a novel topic model to model the process of generating documents, which is followed by a Gibbs sampling-based learning algorithm.

2.5.1 Overview

To define an entity-centric document representation, we first introduce the concept of an entity facet. An *entity facet* is a latent aspect of a specific entity. Each facet is represented by a mixed proportion of latent topics. Unlike LDA's topics that are defined as probability distributions over words, each topic in our model is defined as two separate probability distributions, one over words and one over document entities, respectively, which helps to account for the different observed variables.

An entity-centric document representation is defined as a mixed proportion of entity facets, called (entity) facet distributions. Given a document d associated with a set of source entities $\mathbf{S}_d$, the entity-centric representation of d is a set of facet distributions $\{\theta_s \mid s \in \mathbf{S}_d\}$. The goal of our model is to learn $\{\theta_s \mid s \in \mathbf{S}_d\}$ for all d in the document collection D. As part of the model, we also learn: (1) facet topic distributions ρ_f, i.e., a multinomial distribution over topics; (2) topic word distributions ϕ, and (3) topic (document) entity representations ψ, i.e., multinomial distributions over words and document entities.

2.5.2 Entity facet topic model

A graphical representation of the entity facet topic model (EFTM) is shown in Fig. 2.3; the generative process underlying EFTM is given in Algorithm 1, while detailed explanations are given below.

Generative process

During model initialization, several multinomial distributions are drawn from Dirichlet priors. For each topic, a topic word distribution ϕ and topic entity distribution ψ are generated using Dirichlet priors α and β. For facet f of source entity s, the corresponding facet topic distribution $\rho_{f,s}$ is drawn from a Dirichlet prior τ.

In the generative process, given a document d associated with a set of source entities $\mathbf{S}_d$, a set of multinomial distributions $\{\theta_s \mid s \in \mathbf{S}_d\}$ is generated using a Dirichlet prior μ, where θ_s is a distribution over entity facets of source entity s. To generate a token (word or entity), we iterate over each source entity s in $\mathbf{S}_d$, and draw a facet f_s from θ_s, a topic z_s from $\rho_{f,s}$, and a weight η_s from $B(\sigma)$. Then, for each topic z_s, its topic word distribution and topic entity distribution are first weighted using η_s and then concatenated to obtain a new multinomial distribution Φ_z^s, which is referred to as the *topic token distribution*. In this way, words and entities under the same topic are correlated. Finally, the final topic token distribution Φ is obtained as an equally weighted combination of a set of topic token distributions $\{\Phi_z^s \mid s \in S_d\}$, which is then

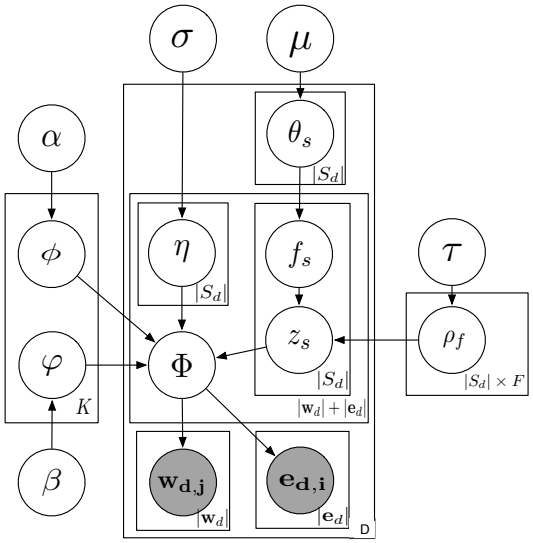


Figure 2.3: A graphical representation of the entity facet topic model (EFTM). A concise overview of parameters is as follows: τ, α, β and μ are Dirichlet priors used to generate corresponding multinomial distributions; η is a Beta prior used to generate Binomial distribution; ρ_f is the facet topic distribution of the f-th facet of an entity; ϕ and ψ are topic word distribution and topic (document) entity distribution respectively.

used to generate a token, i.e., either a word or entity.

Joint facets

In existing topic models, a document is usually represented by one topic distribution, which is assumed to be used to generate the document. To generate a word, a topic is selected from the topic distribution of the document, and a word is sampled from the topic word distribution of the selected topic. For our model, in order to learn multiple representations for a document, we assume that all source entities contribute to the generation of words and entities in documents. In particular, to generate a word or an entity, a facet is sampled from the facet distribution of each source entity, and a topic is sampled from the facet topic distribution of selected facets. Then, all selected topics are merged by a weighted combination of the corresponding topic word distribution and topic entity distribution, and the resulting topic token distribution is used to sample a word or token. The intuition behind this is that the facet distribution of all source entities of a document should contribute to the generation of words and entities in the document. For example, given a document with three source entities: *USA*, *Barack Obama*, and *Mitt Romney*, the representation of this document could be dependent on a facet that is semantically closest to the "presidential campaign."

Algorithm 1 Generative Process of the Entity Facet Topic Model.

for each topic z **do**
  Draw $\phi \sim Dir(\alpha)$
  Draw $\psi \sim Dir(\beta)$
end for
for each source entity $s \in S$ **do**
  Draw $\theta_s \sim Dir(\mu)$
  for each facet of source entity f_s **do**
    Draw $\rho_{f,s} \sim Dir(\tau)$
  end for
end for
for each document d **do**
  for each token t_x **do**
    for each source entity $s \in S_d$ **do**
      Draw $f_s \sim \theta_s$
      Draw $z_s \sim \rho_{f,s}$
      Draw $\eta_s \sim B(\sigma)$
      $\Phi_t^s = \eta_s \phi_{z_s} \oplus (1 - \eta_s)\psi_{z_s}$
    end for
    $\Phi = \frac{1}{|S_d|} \sum_{s \in S_d} \Phi_t^s$
    Draw $t_x \sim \Phi$, $t_x \in V_W \cup V_E$
  end for
end for

Parameters

The number of parameters to be estimated is $K \times (|V_W| + |V_E|) + |S| \times F \times K$, where K is the number of topics, $|S|$ is the size of the source entity set, and F is the number of facets. We use symmetric Dirichlet priors to generate multinomial distributions.

Model advantage

The advantage of our model over existing document representation methods is grounded in two things. First, the multi-faceted nature of source entities is incorporated into our model, which captures the rich semantics captured by different facets of source entities. Second, entity specific facets and general aspects (topics) are semantically connected, thus bridging the gap between entities and topics.

2.6 Inference for EFTM

In this section, we present our Gibbs sampling algorithm for EFTM. Two key latent variables used to generate a word w_i are estimated, i.e., z_s^i and f_s^i. The former is the topic sampled from a given facet f_s, while the latter is the facet selected from a given source entity s. The parameter estimation process is described as follows.

The conditional posterior distribution for z_s^i, the topic from the f-th facet of source entity s, to generate word w_i is

$$P(z_s^i = t \mid \mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}, \mathbf{w}, \mathbf{e}) \propto P(w_i \mid z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) \cdot P(z_s^i = t \mid \mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}), \tag{2.1}$$

where $\mathbf{z}_{-s}^i$ is the assignment of all z related to word w_i except for z_s^i, $\mathbf{z}_{(\cdot)}^{-i}$ is the assignment of all z for all words and entities except for w_i, $\mathbf{f}$ is the assignment of all facets for all words and entities, $\mathbf{w}_{-i}$ are all words except word w_i, $\mathbf{e}$ are all entities.

For the first item on the right hand side of Eq. 2.1, we have

$$P(w_i \mid z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) = \int P(w_i \mid z_s^i = t, \mathbf{z}_{-s}^i, \Phi) P(\Phi \mid \mathbf{z}_{(\cdot)}^{-i}, \mathbf{w}_{-i}, \mathbf{e}) d\Phi. \tag{2.2}$$

Here, Φ is the joint word and entity distribution over all facets of source entities. In particular, given a $|V_W|$-dimensional word distribution and a $|V_E|$-dimensional entity distribution under topic t, a $|V_W| + |V_E|$-dimensional joint multinomial distribution over words and entities Φ_t can be obtained as a weighted concatenation. For the topic t sampled from a facet of source entity s, the topic token distribution Φ_t^s is obtained as a weighted concatenation of ϕ_t and ψ_t with weights η_s and $(1 - \eta_s)$, respectively. Since different source entities are assumed to contribute equally, Φ is obtained as follows:

$$\Phi = \frac{1}{|S_d|} \sum_{s \in [1, |S_d|]} \Phi_t^s = \frac{1}{|S_d|} \sum_{s \in [1, |S_d|]} (\eta_s \phi_t \oplus (1 - \eta_s) \psi_t), \tag{2.3}$$

where $\oplus$ means concatenation of two distributions.

Based on how we obtain Φ, we rewrite Eq. 2.2 as follows:

$$P(w_i \mid z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) = \frac{1}{|S_d|} \left(P(w_i \mid z_s^i = t, \mathbf{w}_{-i}, \mathbf{e}) + \sum_{s' \in S_d \setminus \{s\}} P(w_i \mid z_{s'}^i) \right), \tag{2.4}$$

where $P(w_i \mid z_{s'}^i)$ is $n_{i,s'}^{w_i}$, the number of instances of w_i under topic $z_{s'}^i$. For the first item in Eq. 2.4, we have:

$$P(w_i \mid z_s^i = t, \mathbf{w}_{-i}, \mathbf{e}) = \int P(w_i \mid z_s^i = t, \Phi_t) P(\Phi_t \mid \mathbf{w}_{-i}, \mathbf{e}) d\Phi_t. \tag{2.5}$$

For the second item on the right-hand side of Eq. 2.5, we have:

$$P(\Phi_t \mid \mathbf{w}_{-i}, \mathbf{e}) \propto P(\mathbf{w}_{-i}, \mathbf{e} \mid \Phi_t) P(\Phi_t). \tag{2.6}$$

Since $P(\phi_t)$ and $P(\psi_t)$ are Dirichlet priors $Dir(\alpha)$ and $Dir(\beta)$, and $P(\eta_s)$ is $Beta(\sigma)$, the prior distribution $P(\Phi_t)$ is:

$$\frac{\sigma_1}{\sigma_1 + \sigma_2} \alpha + \frac{\sigma_2}{\sigma_1 + \sigma_2} \beta,$$

where $\sigma = [\sigma_1, \sigma_2]$ is a Beta prior. Since Φ_t is conjugate to the likelihood function (the first item in Eq. 2.5), the posterior distribution in Eq. 2.5 is obtained by putting the multinomial likelihood into the Dirichlet prior:

$$Dir\left(\frac{\sigma_1}{\sigma_1+\sigma_2}\alpha + \frac{\sigma_2}{\sigma_1+\sigma_2}\beta + n_{-i,t}^{w_i}\right),$$

where $n_{-i,t}^{w_i}$ is the number of words and entities that is assigned to Φ_t. Combining the previous equations, we have:

$$P(w_i \mid\sim z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{f}, \mathbf{w}_{-i}, \mathbf{e}) = \frac{1}{S_{|d|}}\left(\frac{\sum_{s'\in S_d\setminus\{s\}} n_{i,s'}^{w_i}}{\sum_{s'\in S_d\setminus\{s\}} n_{i,s'}^{(\cdot)}} + \frac{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{w_i} + \sigma_1\alpha + \sigma_2\beta}{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{(\cdot)} + (\sigma_1\alpha+\sigma_2\beta)\cdot(|V_E|+|V_W|)}\right). \tag{2.7}$$

This addresses the first term on the right-hand side of Eq. 2.1.

For the second term on the right-hand side of Eq. 2.1, we have:

$$P(z_s^i = t \mid \mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}) = \int P(z_s^i = t \mid f_s^i, \rho_{f,s}) P(\rho_{f,s} \mid \mathbf{z}_s^{-i}, \mathbf{f}_s^{-i}) d\rho_{f,s} \tag{2.8}$$

The second item on the right-hand of Eq. 2.8 is a posterior as follows:

$$P(\rho_{f,s} \mid \mathbf{z}_s^{-i}, \mathbf{f}_s^{-i}) \propto P(\mathbf{z}_s^{-i} \mid \rho_{f,s}, \mathbf{f}_s^{-i}) P(\rho_{f,s}). \tag{2.9}$$

Here, $P(\rho_{f,s})$ is a Dirichlet prior $Dir(\tau)$ and $P(\mathbf{z}_s^{-i} \mid \rho_{f,s}, \mathbf{f}_s^{-i})$ is the number of all words and entities that are assigned to topic t because of source entity s except w_i, denoted as $n_{s,t}^{-i}$. Then, Eq. 2.8 can be written as:

$$P(z_s^i = t \mid \mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}) = \frac{n_{s,t}^{-i} + \tau}{n_{s,\cdot}^{-i} + K\tau}. \tag{2.10}$$

The conditional posterior distribution for f_s^i, the facet chosen to first generate a topic and then generate w_i, is:

$$P(f_s^i = f \mid \mathbf{f}_{-s}^i, \mathbf{f}_{(\cdot)}^{-i}, \mathbf{z}) \propto P(z_s^i \mid f_s^i = f, \mathbf{z}_s^{-i}) P(f_s^i = f \mid \mathbf{f}_s^{-i}). \tag{2.11}$$

For the first item on the right-hand side of Eq. 2.11, we have:

$$P(z_s^i \mid f_s^i = f, \mathbf{z}_s^{-i}) = \int P(z_s^i \mid f_s^i = f, \rho_{f,s}) P(\rho_{f,s} \mid \mathbf{z}_s^{-i}) d\rho_{f,s}. \tag{2.12}$$

The second item on the right hand side of Eq. 2.12 is:

$$P(\rho_{f,s} \mid \mathbf{z}_s^{-i}) \propto P(\mathbf{z}_s^{-i} \mid \rho_{f,s}) P(\rho_{f,s}). \tag{2.13}$$

Since $P(z_s^{-i} \mid \rho_{f,s})$ is a likelihood function and $P(\rho_{f,s})$ is a Dirichlet prior $Dir(\tau)$, we have $P(f_s)$ as $Dir(\tau + n_{f,s}^{-i})$, where $n_{f,s}^{-i}$ is the number of topics assigned to f_s except for the current word.

We combine all equations and obtain:

$$P(z_s^i \mid f_s^i = f, \mathbf{z}_s^{-i}) = \frac{n_{f,s}^i + \tau}{n_{\cdot,s}^i + K\tau}. \tag{2.14}$$

The second term in Eq. 2.11 is obtained as follows:

$$P(f_s^i = f \mid \mathbf{f}_{-s}^{(\cdot)}) = \int P(f_s^i = f \mid \theta_s) P(\theta_s \mid \mathbf{f}_s^{-i}) d\theta_s. \tag{2.15}$$

The second term on the right-hand side of Eq. 2.15 is as follows:

$$P(\theta_s \mid \mathbf{f}_s^{-i}) \propto P(\mathbf{f}_s^{-i} \mid \theta_s) P(\theta_s). \tag{2.16}$$

As a result, we have Eq. 2.15 as follows:

$$P(f_s^i = f \mid \mathbf{f}_{-s}^{(\cdot)}) = \frac{n_f^{-i} + \mu}{n_{(\cdot)}^{-i} + F\mu}. \tag{2.17}$$

Finally, by combining Eqs. 2.1, 2.7, 2.10 and 2.11, 2.14, 2.17, we obtain the desired estimates of posterior distributions of z_s^i and f_s^i, respectively:

$$P(z_s^i = t \mid \mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}, \mathbf{w}, \mathbf{e}) \propto \frac{1}{S_{|d|}} \left(\frac{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{w_i}}{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{(\cdot)}} + \frac{(\sigma_1 + \sigma_2) \cdot n_{-i,t}^{w_i} + \sigma_1 \alpha + \sigma_2 \beta}{(\sigma_1 + \sigma_2) \cdot n_{-i,t}^{(\cdot)} + (\sigma_1 \alpha + \sigma_2 \beta) \cdot (|V_E| + |V_W|)} \right) \cdot \frac{n_{s,t}^{-i} + \tau}{n_{s,\cdot}^{-i} + K\tau}. \tag{2.18}$$

$$P(f_s^i = f \mid \mathbf{f}_{-s}^i, \mathbf{f}_{(\cdot)}^{-i}, \mathbf{z}) \propto \frac{n_{f,s}^i + \tau}{n_{\cdot,s}^i + K\tau} \cdot \frac{n_f^{-i} + \mu}{n_{(\cdot)}^{-i} + F\mu}. \tag{2.19}$$

2.7 Experimental Setup

We set up experiments to address the following research questions:

RQ1.1 Can we confirm that entities have multiple aspects, with different aspects reflected in different documents?

RQ1.2 Can we learn a representation of entity aspects from a collection of documents, and a representation of documents based on multiple entities and their aspects as reflected in the documents?

RQ1.3 Does this novel representation improve algorithm performance in downstream applications?

RQ1.4 What is a reasonable number of aspects per entity?

Table 2.2: Statistics of our two datasets. Columns 2 and column 3 are the number of documents in the training and test set; column 4 is the number of labels in the dataset; columns 5 and 6 are the label cardinality (LC) and label density (LD).

Dataset	# training	# test	# Labels	LD	LC
1	6,377	540	56	0.038	2.139
2	34,976	3,146	33	0.077	2.533

2.7.1 Datasets

Two datasets are used in our experiments. Both of them are extracted from the New York Times Corpus [98]. Since we consider both words and entities in documents, we use the entity annotations of New York Times articles from 2003–2007 provided by Google [26]. Documents from 2003–2006 are used as a training set, while documents in 2007 are used as the test set. Articles in the New York Times Corpus are all associated with multiple labels, called *descriptors*. Since a source entity is a general concept referring to any semantic unit representing a source of information, descriptors are considered as source entities in EFTM.

To examine the performance of EFTM on datasets of different sizes, we extract two datasets according to the following procedure: We first count the frequency of the descriptors, and then select target descriptors based on their frequency. The statistics about descriptors are presented in Fig. 2.4. We do not consider descriptors with either a high or low number of associated documents, so that our extracted datasets are of moderate size. Articles associated to at least two target descriptors are selected to be part of the output dataset. For the first dataset ("dataset 1" in Table 2.2), we select descriptors whose frequency ranges from 500 to 1,000, which leads to a set of 56 descriptors. For the second dataset ("dataset 2" in Table 2.2), we choose descriptors whose frequency is higher than 2,000, with 33 descriptors being selected.

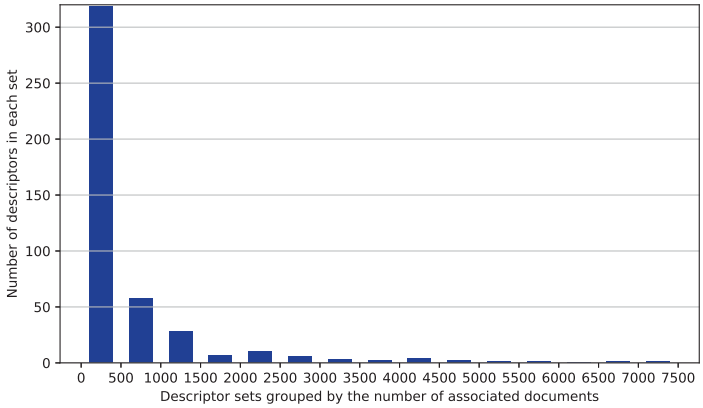


Figure 2.4: Statistics of descriptors in NYT Corpus. The x-axis are descriptor sets grouped by the number of associated documents; the y-axis is the number of descriptors in each set.

We count the descriptive statistics of our two datasets and present them in Table 2.2. The definitions of label cardinality and label density [111] are as follows. Let D be a multi-label dataset consisting of D multi-label examples $(x_i, Y_i), i = 1, \cdots, D$. The *label cardinality* of D is defined as the average number of labels of the examples in D:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|.$$

The *label density* of D is defined as the average number of labels of the examples in D divided by L:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|/|L|,$$

where L is the number of all labels.

2.7.2 Experimental design

Modeling and recovering entity facets

To answer RQ1.1 we conduct a qualitative analysis that seeks to uncover the correlation between the facets of source entities. Our hypothesis is that in a set of documents associated with a pair of source entities, the facets of these entities that are mainly reflected in the documents will be represented by similar topic distributions. If this is true, it would mean that our model is able to identify matching facets of different entities in order to semantically represent a document and it will confirm that indeed entities are not single-faceted. For example, given a set of documents associated with both *Finance* and *Education and Schools*, we expect to identify the funding (education investment) facet of *Education and Schools* and *Finance*. Therefore, given two source entities, first the documents associated to both entities are selected. For each source entity, the corresponding entity-centric representation is computed as the mean of the entity representations of all shared documents. The resulting facet distribution can be considered as a centroid across these documents. Then we examine whether certain facets have a high probability in the facet distribution.

Generative capability of entity facet topic model

To answer RQ1.2 we perform a quantitative analysis, using perplexity as the evaluation measure. Perplexity is a standard measure for estimating the performance of a probabilistic model. We evaluate EFTM by estimating the perplexity of unseen held-out documents given training documents. A better model will have a lower perplexity of held-out documents on average. We follow the perplexity definition in [51], which is defined as follows:

$$\exp\left(-\frac{\log P(D^{test} \mid D^{train})}{\sum_{d \in D^{test}} N_d}\right).$$

Let Φ denote the set of all parameters in a topic model; then

$$P(D^{test} \mid D^{train}) = \int P(D^{test} \mid \Phi) P(\Phi \mid D^{train}) d\Phi.$$

This integral can be approximated by averaging $P(D^{test} \mid \Phi)$ under samples from $P(\Phi \mid D^{train})$. Note that EFTM can be seen as a generative process that generates words and entities for a given set of source entities. Thus, $P(D^{test} \mid \Phi)$ is defined as follows:

$$P(D^{test} \mid \Phi) = \Pi_{d \in D^{test}} P(w_d, e_d \mid S_d, \Phi).$$

If our model demonstrates lower perplexity it will confirm that indeed a document is better represented and understood by considering the specific facets of an entity discussed in this document.

Multi-label text classification

To answer RQ1.3 we perform an extrinsic evaluation of EFTM, for which we consider a multi-label classification task. We assume that a good document representation model should encode distinctive information about whether a document should be associated to a label or not. Therefore, better document representations should yield a better performance in multi-label classification.

A simple and straightforward method for multi-class classification is to train a binary classifier for each label using the representation of a document as features. We adopt this setting and use SVM as the binary classifier in our experiments. We have experimented with a number of different classifiers and all of them support our conclusions. The goal of this work is to compare representations and not classification algorithms, we do not report the performance using other classifiers, such as Random Forests, or Naive Bayes. The features are the elements of the vector representations, and the number of features is the dimensionality of the vectors.

Given a label (i.e., a source entity), EFTM provides an entity centric representation of the document corresponding to that label. However, for the purpose of training we also need an entity centric representation of documents that are not associated to this label, to be used as negative instances. *Pseudo inference* [90] is performed on the document and the label, by assuming that the label is the only source entity associated to the given document, hence inferring the entity-centric document representation using EFTM.

For each label s, we collect N^s_{train} documents associated to the label as positive instances for training. Then we randomly sample N^s_{train} negative instances from documents not associated to this label for training. During testing, for each label and test instance, we perform pseudo inference on the test instance, and input the obtained representation to the binary classifier corresponding to the label. Then, the binary classifier outputs 1 if the test instance is considered to be associated to the label, and 0 otherwise.

To assess the performance of multi-label classification we use three evaluation measures [10, 111]: *multi-label accuracy*, *macro F1*, and *micro F1*.[1] We conduct a statistical significance test in our experiments via a paired t-test. Our experiments

[1] Bielza et al. [10] also define Mean Accuracy as an effectiveness measure. However, our data is rather skewed with respect to each label, i.e., for each label there is a small fraction of documents that are associated with it. This allows a naive classifier that predicts each instance not being associated with a label to achieve high performance when measured by Mean Accuracy. For instance, such a classifier, when applied to "Dataset 1" achieves a Mean Accuracy of 0.9629. Multi-label Accuracy, macro and micro F1 avoid this bias.

consist of two steps, i.e., model training and multiple binary classifications. Since we focus on comparing representations, it is the model training step that is considered as the source of uncertainty. We train each model 5 times, and repeat the classification steps to obtain multiple results. All significance tests are performed at $\alpha = .05$ level.

If our model demonstrates a better performance than baseline methods, it will confirm that multi-faceted entity-centric representation of documents has a positive effect not only towards better understanding documents but also in downstream applications.

Number of entity facets and topics in a collection

To answer RQ1.4 we perform a quantitative analysis similar to the previous section, altering the number of topics and the number of entity facets in our model and quantifying the effect in terms of the performance of the model in the multi-label classification task.

2.7.3 Baselines and model variations

Table 2.3 lists the document representation methods considered in our experiments. In both our intrinsic and extrinsic evaluation we seek to assess the quality of document representations, and not the ability of a machine learning algorithm to succeed in multi-label classification. Our baselines include the traditional bag of words with TF-IDF weights representation (BoW) [97]; the LDA [11], LLDA [27] and Labeled-LDA [89] topic model document representations, that allow us to directly assess the quality of EFTM; and Mean Word Embedding (MWE) [53] and Doc2vec [57] as state-of-the-art dense vector representations, similar with the setup in [114]. The word embeddings used here are 50d GloVE vectors [82] pre-trained using Wikipedia and Gigaword 5.[2] We use a pre-trained BERT model [23] with a hidden layer size of 768, 12 Transformer blocks and 12 self-attention heads. Since the representations learned by baseline methods are not optimised for multi-label classification, we use bert-as-service[3] to get the representation of an input sequence without fine-tuning with respect to multi-label classification to make results comparable to other baseline methods, as well as our work. We truncate the input sequences to 512 tokens to meet the restriction of maximum sequence length required by BERT. To get a feeling of how BERT performs on multi-label classification when fine-tuning, we also fine-tune the same pre-trained BERT model for the task of multi-label classification, with a batch size of 32, max sequence length of 512, learning rate of 5e-5. The maximum number of training epochs is set to 3 according to the parameter setup in [23].

Note that even though ETM [51] is related to our work, it is not a valid baseline, as explained in Section 2.2. For baseline models that do not distinguish entities from words, we feed unique identifiers of entities together with words, so that our model does not benefit from being fed more data than baselines.

For EFTM, we consider different source entity and observed variables settings. In terms of source entity, we consider a multi-source (MS) setting, where each document is associated with multiple source entities, and a single-source (SS) setting, where all documents are assumed to be associated with one universal source entity. In terms of

[2] https://catalog.ldc.upenn.edu/LDC2011T07

[3] https://github.com/hanxiao/bert-as-service

Table 2.3: Methods and baselines used for comparison.

Acronym	Description	Reference
EFTMWO-SS	EFTM with a single and same source entity, and words as the only observed variables.	§2.5.2
EFTMWO-MS	EFTM with multiple source entities, and words as the only observed variables.	§2.5.2
EFTMWE-SS	EFTM with a single and same source entity, and considers two kinds of observed variables.	§2.5.2
EFTMWE-MS	EFTM with multiple source entities, and considers two kinds of observed variables.	§2.5.2
BoW	Bag of words weighted by TF-IDF; a traditional document representation.	[97]
LDA	Latent Dirichlet Allocation, which learns a latent topic distribution to represent documents with; a widely used document representation method.	[11]
LLDA	Mixed membership of topics, which is similar to LDA, except that it considers words and entities separately.	[27]
Labeled LDA	A supervised topic model, which extends LDA to leverage supervised labels of documents.	[89]
Doc2vec	Dense vector representation, which is the state-of-the-art neural method for learning document representations.	[57]
BERT	A state-of-the-art vector representation method.	[23]
MWE	Mean word embedding, which is a strong state-of-the-art neural method for representing documents.	[16, 57]
Most-Frequent	A naive baseline which always predict the most frequent label in train set.	–

observed variables, we consider a word-only (WO) setting, which is a simplified version of EFTM in which we only use words as observed variables, and a words and entities (WE) setting, which is the full EFTM model. We write EFTMWO-MS, EFTMWO-SS, EFTMWE-SS and EFTMWE-MS, respectively, to denote these variants. See Table 2.3 for a summary.

2.7.4 Parameter settings

Following standard practice [51], we set the hyperparameters of the baseline methods and EFTM to pre-defined values. In LDA, LLDA and EFTM, we set both α and β as 0.1. In LLDA and our model, σ is set to 0.5, which means no prior information is known. Note that our model might perform better if σ is set to a value that corresponds to the frequency of words vs. entities. We use uninformative prior 0.5 and leave the impact of σ as future work. In EFTM, we set τ and μ to 0.1. The number of iterations of Gibbs Sampling is set to 2,000 for all models. We set the number of topics to $\{10, 15, 20, 30, 40, 50, 80, 100\}$, and the number of facets to $\{5, 10\}$, so as to be able to

Table 2.4: The number of hours used to train our models under different setups.

	EFTMWO-SS	EFTMWO-MS	EFTMWE-SS	EFTMWE-MS
Dataset 1	4.0 hours	9.2 hours	3.9 hours	9.0 hours
Dataset 2	13.2 hours	42.9 hours	12.6 hours	38.9 hours

compare the effectiveness under different parameter settings.

In this work we make the assumption that all entities have an equal number of facets. Clearly this can not the case with the number of facets most likely changing across entities. Nonparametric Bayesian models could capture this and we leave it for future work. Instead, we fixed the number of facets to 5 (or 10). Remember that the number of facets is a modeling choice and should be decided by empirical evidence. Our choice is derived from the fact that the median number of sections of English Wikipedia articles (which can be viewed as entity facets [77]) is 4 for the entire collection and 7 for a high quality sub-collection [85]. Also note that the number of facets does not have to be accurate for all entities, as long as learned facets could be helpful for downstream applications, e.g., classification, as demonstrated in our experiments. A similar scenario with the number of facets for our model is the number of topics for topic models, where the number of topics can be predefined according to empirical evidence (rather than accurate estimation), as long as the learned topics are useful for inferring meaningful topic distributions of documents.

Regarding runtime, we run our model on single core Intel Xeon CPU with 256GB RAM. We run our models under different setups and against two datasets and the training time are presented in Table 2.4.

2.8 Results

In this section, we present the outcomes of our experiments and provide answers for our research questions.

2.8.1 Modeling and recovering entity facets

To answer RQ1.1, we perform a qualitative analysis of the outcomes of the EFTMWE-MS model trained on Dataset 2 with the number of facets set to 5 and the number of topics set to 50. The setup of the qualitative analysis can be found in Section 2.7.2. We choose three pairs of source entities for analysis, i.e., *Finances*, *Law and Legislation*, and *Education and Schools*, and the results are shown in Fig. 2.5. As shown in Fig. 2.5(c), facet 1 of *Education and Schools* has a higher probability, indicating that facet 1 is related to the law facet of *Education and Schools*, while facet 3 of *Law and Legislation* is related to its education facets. Further, we see that in Fig. 2.5(a) and 2.5(b), while there is no particular facet with a strong presence in these documents, there is a reverse relation between facet 1 and 2 of *Finances*. When documents are associated with *Education and Schools*, facet 2 of *Finances* has a higher presence than 1, while the opposite is true when documents are associated with *Law and Legislation*. Hence, different facets of source entities are captured by the entity facets proposed in EFTM.

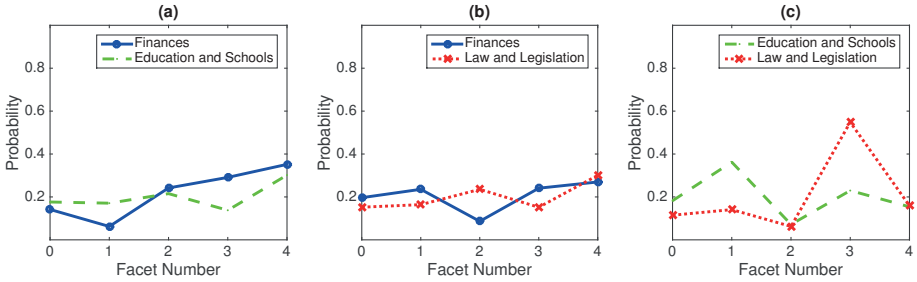


Figure 2.5: Analysis on three sets of documents, where each set of documents is associated with the same pair of source entities. Each line is the averaged facet distribution of a specific source entity in the corresponding documents. The x-axis is the facet number, while the y-axis is the value of elements in the facet distributions. The number of facets is 5.

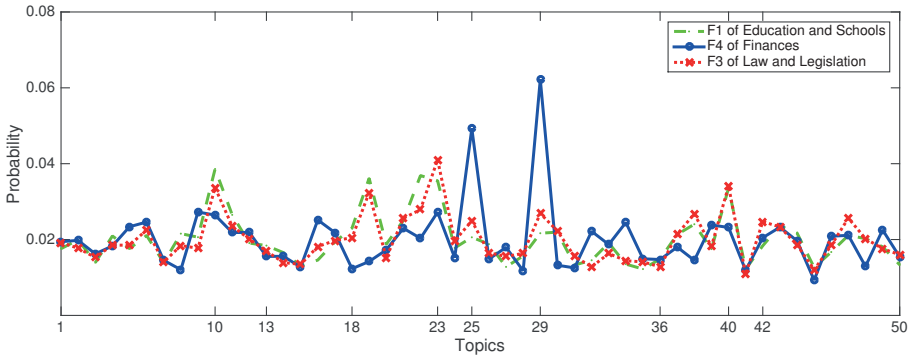


Figure 2.6: Facet topic distribution of facet number 1, 4, 3 of *Education and Schools*, *Finances* and *Law and Legislation*, respectively.

To further investigate whether these facets indeed capture similar information, we present the facet topic distribution of those facets, i.e., facet 4 of *Finances*, facet 1 of *Education and Schools*, facet 3 of *Law and Legislation* in Fig. 2.6. The topic distributions of all the facets we consider are similar, demonstrating that similar information is captured by entity facets and understood from the perspective of source entities.

Therefore, the document representations we propose are able to discover and model facets of different entities within a document that semantically align with each other and the theme of the document, while at the same time disagree with the facets learnt by documents of different theme.

Finding 1. Entities are associated with documents by means of specific facets of them discussed in these documents, while different documents may focus on different facets of an entity; this confirms our hypothesis that entities should be considered multi-faceted concepts.

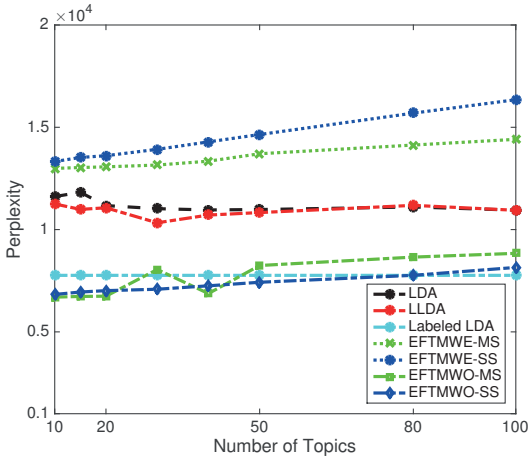


Figure 2.7: Perplexity values on Dataset 1 with different numbers of topics.

2.8.2 Generative capability of entity facet topic model

To answer RQ1.2, we compare the perplexity values obtained by the different topic models. The perplexity scores over the two datasets are shown in Fig. 2.7 and 2.8, respectively. Note that the number of topics of Labeled-LDA is decided by the number of labels, thus its perplexity value is fixed. EFTMWO-MS and EFTMWO-SS perform better than the baseline methods, whereas EFTMWE-MS and EFTMWE-SS perform worse. The difference lies in the representation of topics. In the WE-version of EFTM, topics are represented by two distributions, i.e., a topic-word distribution and topic-(document) entity distribution. The prior of σ is set to 0.5, which means that words and entities have the same chance of being selected to generate a document. However, in our dataset, there are much more words than entities, which leads to a lower probability of generating the documents and higher perplexity values. By introducing entity facets, we achieve better generative capability for unseen documents.

Finding 2. The generative capability of a model that considers entities as multi-faceted concepts is better compared to models that do not.

2.8.3 Multi-label text classification

To answer RQ1.3, we present the performance of different document representation methods on the multi-label classification task. Results over two datasets are shown in Table 2.5 and 2.6.

As we can see, semantic representation methods, such as LDA, MWE, D2V and BERT, perform better than traditional BoW representations. When comparing semantic representation methods, BERT is better than LDA but not as good as advanced topic models, such as LLDA, indicating that advanced topic modeling has the potential to learn good representations. As mentioned earlier we also fine-tuned BERT on the classification task. We did that for the small Dataset 1, since fine-tuning for the large dataset (Dataset 2) proved to be a rather computationally expensive task (the task was

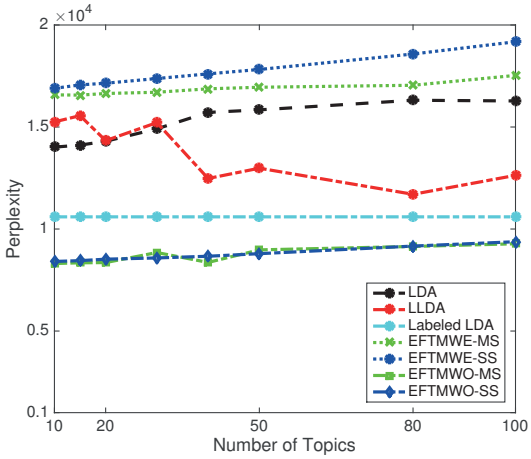


Figure 2.8: Perplexity values on Dataset 2 with different numbers of topics.

abandoned after 8 days of training). The BERT-fine-tuned model obtained a ml-Acc of 0.0593, Micro F1 of 0.1082, and Macro F1 of 0.0163, better than many unsupervised representation learning methods, but still not better than the unsupervised representation learning of our advanced topic modeling. Note that the naive baseline *Most-Frequent* is a strong baseline in terms of multi-label accuracy. This is because if we predict the most frequent label, it is likely to get at least one right label for each instance. Together with the fact that the label cardinality is around 2, there should be quite a few instances that get almost correctly predicted.

On both datasets, EFTMWO-SS outperforms LDA, indicating the effectiveness of introducing entity facets and representing documents using facet distributions, especially given that the number of features of LDA in multi-label classification (i.e., the number of topics, 50) is much bigger than that of EFTMWO-SS (number of facets, 5). Hence, representing documents in an entity-centric fashion gives an explicit way to facilitate downstream entity related tasks, such as judging whether a document should be associated with the entity (label) here.

We also study whether it is helpful to distinguish different kinds of observed variables, such as words and entities. The performance of LLDA is better than LDA, showing the effectiveness of distinguishing different kinds of observations in relatively simple topic models. As to variants of EFTM, the performance of EFTMWE-SS/EFTMWE-MS are consistently better than that of EFTMWO-SS/EFTMWO-MS, which confirmed the superiority of considering both observed variables. Note that sometimes macro F1 could be zero because of skewed performance across different labels. Specifically, the performance of many labels is close to zero.

To study the impact of multiple sources, we consider two pairs for comparison, i.e., EFTMWE-MS vs. EFTMWE-SS and EFTMWO-MS vs. EFTMWO-SS. On Dataset 1, EFTM with multi-sources (MS) is consistently better than the single-source (SS) version of EFTM, while EFTMWE-SS is better than EFTMWE-MS on Dataset 2. This appears to be related to the size of the set of labels. The number of labels in Dataset 2 is smaller than that of Dataset 1, which lessens the impact of modeling multiple sources.

Table 2.5: Comparing the performance of document representation methods on the task of multi-label classification on Dataset 1. The number of facets of EFTM is set to 5, and the number of topics/dimensions is set to 50. We test the significance of results of our models compared to baseline methods. All results of our model are significant compared to baseline methods at $\alpha = .05$ level.

	ml-Acc	**Micro F1**	**Macro F1**
Most-Frequent	0.0398	0.0571	0.0029
BoW	0.0135	0.0267	0.0272
D2V	0.0047	0.0091	0.0090
MWE	0.0047	0.0094	0.0097
BERT	0.0054	0.0106	0.0109
LDA	0.0064	0.0123	0.0128
LLDA	0.0358	0.0690	0.0632
Labeled LDA	0.0000	0.0033	0.0034
EFTMWO-SS	0.0377	0.0717	0.0710
EFTMWO-MS	0.0392	0.0745	0.0778
EFTMWE-SS	0.0390	0.0738	0.0000
EFTMWE-MS	**0.0399**	**0.0758**	**0.0788**

In sum, EFTMWE-MS and EFTMWE-SS are the preferred choices for the multi-label classification task, with a slight preference for EFTMWE-MS in case where a dataset has many labels (source entities) and for EFTMWE-SS in case the dataset has fewer labels.

Finding 3. By fixing supervised learning algorithm and using different document representations for multi-label classification, we demonstrate that the proposed multi-faceted entity-centric representation outperforms state-of-the-art representations. The smaller the dataset the more important it is to use a multi-source representation.

2.8.4 Number of entity facets and topics in a collection

To answer RQ1.4, we explore different parameter settings for EFTM. In particular, we consider two cases: (1) varying the number of topics (30, 50, 80, 100) under a fixed number of facets (5); (2) varying the number of facets (5, 10) under a fixed number of topics (50). The results under a fixed number of facets and a varying number of facets are shown in Table 2.7 and 2.8, respectively.

As we can see in Table 2.7, the full model (EFTMWE-MS) performs best on Dataset 1 and 2 when the number of topics is set to 30 and 80, respectively. Dataset 2 is a much bigger dataset compared to Dataset 1 and 30 topics appears to be enough to capture the topical patterns of the smaller dataset but is insufficient for the bigger dataset. In terms of the number of facets, we can see in Table 2.8 that the performance with just five facets is consistently better than the performance with 10 facets, which indicates that a small number of facets is probably enough to capture entity specific topics and a big number of facets might make things complicated. Overall, we can see

Table 2.6: Comparing the performance of document representation methods on the task of multi-label classification on Dataset 2. The number of facets of EFTM is set to 5, and the number of topics/dimensions is set to 50. We test the significance of results of our models compared to baseline methods. Results of our model are significant compared to baseline methods at $\alpha = .05$ level.

	ml-Acc	**Micro F1**	**Macro F1**
Most-Frequent	0.0809	0.0983	0.0085
BoW	0.0234	0.0454	0.0478
D2V	0.0088	0.0173	0.0176
MWE	0.0082	0.0161	0.0171
BERT	0.0065	0.0125	0.0133
LDA	0.0082	0.0158	0.0170
LLDA	0.0620	0.1158	0.0948
Labeled LDA	0.0035	0.0069	0.0073
EFTMWO-SS	0.0684	0.1235	0.1185
EFTMWO-MS	0.0658	0.1216	0.0000
EFTMWE-SS	**0.0758**	**0.1387**	**0.1375**
EFTMWE-MS	0.0658	0.1207	0.1214

that the performance varies considerably under different parameter settings, indicating a space of improvements by tuning parameters. We leave the tasks of finding the optimal number of facets and topics as future work.

Finding 4. A limited number of facets is enough to capture the different facets of an entity, on average.

2.9 Conclusions

In this chapter, we answer RQ1. We propose a model and an algorithm to learn entity-centric document representations, where a document associated with multiple entities is represented by multiple representations and each representation is built on the basis of entity facets. We demonstrate the effectiveness of our model, EFTM, by comparing it against state-of-the-art document representation methods on the task of multi-label classification, confirming that multi-faceted entity-centric modeling of documents has an effect in downstream applications. Although we evaluate our method on multi-label classification, our method is more broadly applicable to other multi-labeled settings where documents are associated with multiple source entities, such as tag analysis [62] and tag suggestion [48]. We further investigated the notion of facets we learn by performing both an intrinsic and an extrinsic evaluation and confirmed that learned facets can capture semantically similar facets of different entities.

The theoretical implication of this chapter is that entities should not be considered and modeled as monolithic concepts, with a single representation for every document

Table 2.7: Comparing the performance of document representation methods on the task of multi-label classification, using Dataset 1 and Dataset 2. The evaluation metric is multi-label accuracy (ml-Acc). For our models the number of facets is set to 5. The number of topics of our model and other models considered are 30, 50, 80, and 100, as indicated in row two.

	Number of Topics	30	50	80	100
Dataset 1	EFTMWO-SS	0.0377	0.0378	0.0303	0.0302
	EFTMWO-MS	0.0455	**0.0453**	**0.0437**	**0.0422**
	EFTMWE-SS	0.0356	0.0381	0.0389	0.0362
	EFTMWE-MS	**0.0488**	0.0447	0.0402	0.0414
Dataset 2	EFTMWO-SS	0.0641	0.0681	0.0713	0.0568
	EFTMWO-MS	**0.0779**	0.0737	0.0695	0.0659
	EFTMWE-SS	0.0698	**0.0767**	**0.0806**	**0.0781**
	EFTMWE-MS	0.0728	0.0717	0.0736	0.0661

associated with them, but instead be thought of as multi-faceted concepts, with different facets discussed in different documents. Gaining a deeper understanding of what these facets precisely represent, how many facets each specific entity has, and whether these facets can be mapped to explicit categories is left for future work that can enable interesting methods for analyzing and visualizing document collections from the perspective of an entity, but also analyzing how entities are presented in a document corpus. In practice, our work demonstrates that such a multi-faceted entity consideration can have an impact in downstream applications.

Table 2.8: Comparing the performance of document representation methods on the task of multi-label classification, using Dataset 1 and Dataset 2. The values 5 and 10 in columns 3–6 indicate the number of facets.

		Micro F1		**Macro F1**	
		5	10	5	10
Dataset 1	EFTMWO-SS	0.0710	0.0679	0.0702	0.0677
	EFTMWO-MS	**0.0857**	0.0661	**0.0885**	**0.0718**
	EFTMWE-SS	0.0722	**0.0680**	0.0740	0.0685
	EFTMWE-MS	0.0843	0.0618	0.0836	0.0657
Dataset 2	EFTMWO-SS	0.1217	0.1109	0.1182	0.1141
	EFTMWO-MS	0.1361	0.1127	0.1331	0.1123
	EFTMWE-SS	**0.1395**	**0.1393**	**0.1366**	**0.1386**
	EFTMWE-MS	0.1323	0.1193	0.1332	0.1173

3

A Multi-Interaction based Convolutional Matching Network for Entity Aspect Linking

In the previous chapter, we studied the task of learning entity-centric document representation by modeling entities as multiple aspects and each aspect as an entity-specific topic. In this chapter, we continue our exploration of entity aspects for document understanding. Specifically, we study the problem of entity aspect linking.

3.1 Introduction

Understanding the semantics of a collection of textual documents is important for many applications, including text classification and information retrieval. Departing from bag-of-words representation, research has moved towards treating semantically meaningful units, such as entities, separately from words. One of the representative techniques to perform semantic annotation over text collections is entity linking, which links entity mentions in text to entities in knowledge bases [102]. It is well recognized [29, 66, 92] that entities may have multiple aspects, which stand for very different topics related to these entities. As an example, in the Wikipedia page of *Barack Obama*, there are multiple aspects, such as *early life and career* and *post-presidency (2017–present)*. When performing entity linking over a mention of Barack Obama in a news article that discusses his early life, ideally, we would like to know both which entity the mention refers to and which aspect of the entity relates to the article.

Recently, a task called *entity aspect linking* [77] was proposed to address this need. The goal is to link entity mentions to aspects of entities rather than entities. Compared to entity linking, entity aspect linking is aimed at semantic annotation at a finer granularity. Existing approaches for entity aspect linking are mostly related to traditional matching methods, such as BM25 and learning to rank. Nanni et al. [77] propose to use hand-crafted features and learning-to-rank algorithms to rank candidate aspects for entities

This chapter was published as C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. A multi-interaction based convolutional matching network for entity aspect linking. In *COLING*, 2020 (under review).

using contexts around entity mentions. We take a different approach, and view entity aspect linking as a pairwise semantic matching problem.

In this chapter, we present the Multi-Interaction based Convolutional Matching Network (MICMN) for entity aspect linking, which combines interactions from multiple perspectives and extracts features using a convolutional neural network. To be specific, given an entity in context (referred to as entity context) and a candidate aspect, we first represent words using pre-trained embeddings, and then use them as inputs to construct interactions between the entity context and the candidate aspect. Here, an interaction is pairwise word similarity between words in entity contexts and candidate aspects. We consider interactions from multiple perspectives: exact match interactions, soft match interactions, and self-attention weighted soft match interactions. Then, a convolutional layer is applied to each interaction matrix to extract convolved features. In our task, entity contexts are usually much shorter than candidate aspects. Intuitively, this means that it is easier to identify matching patterns in candidate aspects since it is longer in text. On the other hand, matching signals in entity contexts is likely to be important since it has limited information. To reflect this intuition, we propose a novel q-singular pooling strategy to extract useful features, which emphasizes matching on entity contexts. Finally, a multi-layer perceptron is used to compute a matching score.

The advantage of our model is three fold. First, our model does not rely on hand-crafted features and can be trained end-to-end with input sequences, which makes it easy to use. Second, our model combines multiple interactions, so that relevance signals can come from different perspectives. Third, we present a simple yet effective pooling strategy for extracting features from the outputs of convolutional layers, motivated by the characteristics of the task of entity aspect linking.

To verify the effectiveness of our model, we compare our model with several baselines. Our experiments on four entity aspect linking datasets show the advantages of our model. Our model outperforms state-of-the-art neural ranking methods. The ablation study we conducted reveals that the key to MICMN's advantages is the fact that different interactions complement each other and the q-singular pooling strategy works for entity aspect linking.

In summary, our contributions are as follows. (1) A novel convolutional neural network based approach and a novel pooling strategy for entity aspect linking. We combine a convolutional neural network with soft match and self attention to find the most appropriate candidate entity aspect. (2) An ablation study on how different interactions and pooling strategies affect the performance on entity aspect linking. (3) A parameter analysis that shows how the length of entity contexts and candidate aspects affects entity aspect linking performance of our model.

3.2 Related Work

Entity aspect linking, which can be viewed as fine-grained entity linking, has recently been proposed by Nanni et al. [77]. They propose a feature engineering approach that combines multiple hand-crafted features and use a learning to rank approach to identify the best matching candidate aspect. Naive features are used and proved to be effective. For example, one of the features measures the content overlap between entity context

and candidate aspect. Traditional information retrieval features, such as tf-idf and BM25, are used by viewing the entity context as a query and each candidate aspect as a document. In contrast to devising effective features for training classifiers, we resort to neural network based approaches. Specifically, we view entity aspect linking as a pairwise semantic matching task and devise a supervised neural network approach to identify matching signals between entity context and candidate aspect.

Semantic matching identifies the semantic meaning and infers the semantic relation between two pieces of text [36]. One way to categorize semantic matching tasks is to divide them into two categories: symmetrical problems and asymmetrical problems [106]. Examples of symmetrical problems include paraphrase identification [127] and semantic textual similarity [1], while other tasks such as question answering (QA) and ad-hoc retrieval are considered as asymmetrical problems. Our task aligns with the second category, since contexts of entity mentions are usually shorter than textual description of candidate aspects.

Existing approaches for semantic matching can be classified into two categories: representation-based methods and interaction-based methods [36]. Representation-based methods [42, 103] leverage deep semantic representation learning models to embed input sequences into continuous vectors for further computation of pairwise similarity. Most recent studies are focusing on interaction-based methods which learn matching patterns based on interactions between two input sequences. Xiong et al. [124] utilize a translation matrix to model word-level similarities via word embeddings, where a new kernel-pooling technique is proposed to extract multi-level soft match features, and a learning-to-rank layer is combined with those features to obtain the final ranking score. Zhang et al. [130] design an attentive interactive neural network to focus on text segments that are useful to answer selection. Our approach is similar to the latter. Compared with previous approaches, we combine interactions from multiple perspectives and show that different interactions complement each other.

Convolutional neural networks are widely used in semantic matching, either to learn a better representation or to extract useful features from pairwise interactions [42, 80]. Pang et al. [80] leverage a matrix to model word-pair interactions between questions and answers for matching, and a hierarchical convolutional model then operates on this single interaction matrix to compute the final score. Hu et al. [42] propose two convolutional neural network models to capture hierarchical structures of sentences and learns fixed length vector representation for sentences. Shen et al. [103] use convolutional neural network to model local contextual information and combine salient local features to form a global vector representation of queries and documents. In this chapter we use one convolutional module for each interaction so that information from different interactions can be fully utilized.

3.3 A Multi-Interaction based Convolutional Matching Network

In this section we introduce a novel neural matching network specifically designed for entity aspect linking. As discussed in the introduction to chapter, our model, MICMN (Multiple Interaction based Convolutional Matching Network) consists of four layers. It

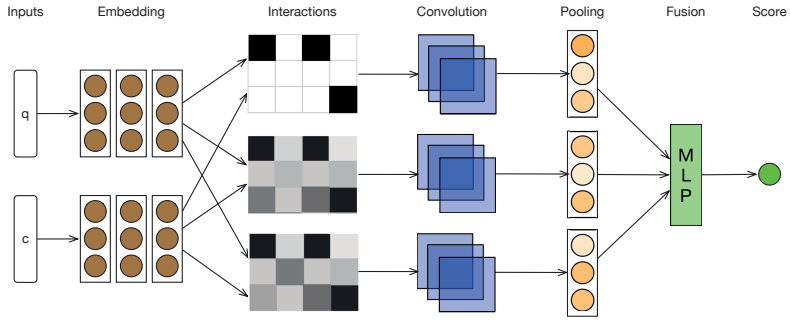


Figure 3.1: Overview of the Multi-Interaction Convolutional Matching Network.

first accepts textual sequences as inputs. Then, given their input embeddings, it models four kinds of interactions as matching patterns from different perspectives. Furthermore, these interactions are passed through convolutional layers and a novel pooling strategy is used to extract local features. Finally, a fusion layer is used to combine all matching signals to produce a score for ranking. Our overall model architecture is shown in Fig. 3.1 and each of the above layers is described in detail below.

3.3.1 Input

The inputs to our model are an entity context q and a candidate aspect document c, which consists of context terms $\{w_1^q, w_2^q, \ldots, w_m^q\}$ and document terms $\{w_1^c, w_2^c, \ldots, w_n^c\}$ respectively. m and n are the number of terms in q and c. We employ an embedding layer to convert each term into a L-dimensional vector representation, generating matrix representation $Q \in R^{m \times L}$ and $C \in D^{n \times L}$ for q and c respectively.

3.3.2 Interactions from multiple perspective

Given Q and C, we construct an interaction matrix M, where each element M_{ij} stands for the basic interaction, i.e., similarity between word w_i^q and w_j^c (see Equation 3.1) and the operator $\otimes$ stands for an operator to obtain the similarity.

$$M_{ij} = w_i^q \otimes w_j^c. \tag{3.1}$$

The way the interaction matrix is built is similar with that in [80]. We adopt different kinds of $\otimes$ operator to model the interactions between two words.

Exact match interaction

The first operator is the ***indicator function***, which accepts token inputs and produces 1 to indicate that two words are identical and 0 otherwise. The element at i-th row and j-th column in exact match interaction matrix M^{em} is shown in Equation 3.2:

$$M_{ij}^{em} = I_{w_i^q = w_j^c} = \begin{cases} 1, & \text{if } \sim w_i^q = w_j^c \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

This is similar to the indicator matching matrix used in [74, 80]. The indicator function can capture exact match signals, which has proved to be a good baseline for entity aspect linking, as shown by the content overlap method in [77]. However, one of the limitations of the indicator function is that it cannot capture matching signals between two semantically similar words. To address this issue, we resort to soft match interaction, self-attentive interaction and self-attention weighted soft match, which are capable of capturing semantically similar words.

Soft match interaction

Soft match interaction is used in various semantic matching tasks, such as paraphrase identification [80], ad-hoc retrieval [21, 124], and short text matching [15]. In our model, we use ***cosine similarity*** as the interaction operator and compute the semantic similarity between words using pre-trained embeddings. The embedding of word w_i^q and w_j^c are denoted as v_i and v_j respectively. The element at i-th row and j-th column in soft match interaction matrix M^{sm} is shown in Equation 3.3; it may be viewed as a soft indicator function:

$$M_{ij}^{sm} = \frac{v_i v_j}{\mid v_i \mid \ldots \mid v_j \mid}. \tag{3.3}$$

Self-attentive interaction

The third type of interaction we consider is self-attentive interaction, which applies self attention over q and c respectively and then combines the weights together. Specifically, it takes as an input the word embeddings of the sequence q, i.e., Q, and outputs a vector of self-attentive weights a_q:

$$a_q = \mathrm{softmax}(w_{s2} \tanh(W_{s1} Q^T)). \tag{3.4}$$

Here, W_{s1} is a weight matrix with a shape of $u \times L$, and w_{s2} is a vector of parameters with size u, where u is a hyperparameter we can set arbitrarily. Since Q is a matrix of $m \times L$, the weight vector a_q will have size m. The $\mathrm{softmax}(\cdot)$ ensures that all the computed weights sum up to 1. The weights of the sequence c, i.e., a_c, can be computed in the same way. The size of a_c is n.

After obtaining the weights of input sequences, i.e. a_q and a_c, we compute the self-attentive weights for soft match interaction W_{sa} as follows:

$$W^{sa} = a_q (a_c)^T. \tag{3.5}$$

The size of W^{sa} is $m \times n$, since a_q and a_c are vectors of size m and n respectively. Then we apply self attention weights to soft match interaction as follows:

$$M^{sw} = M^{sm} \cdot W^{sa}. \tag{3.6}$$

3.3.3 Convolutional matching

Given the aforementioned multi-perspective interactions, we apply a separate convolutional layer to each interaction matrix to extract convolved features. Since the operation

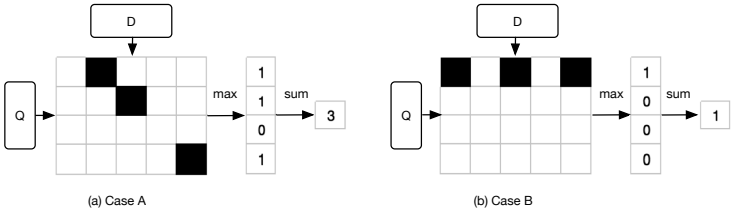


Figure 3.2: Illustration of q-singular pooling.

over all interactions is similar, we take M^{sm} as an example. Given M^{sm}, the convolutional layer applies convolutional filters to the matrix to capture the compositional matching patterns of consecutive terms. F different filters with square kernel are used to extract features, each describing the n-gram local matching in a different perspective. The kernel size is set to k. Then we add a bias and apply a non-linear activation function and obtain convolved features CF for each interaction.

3.3.4 Q-singular pooling

Given the convolved features, we want to extract features using the pooling operation. Column-wise max-pooling is proposed in [99], which operates over M to generate the vectors $g_q \in R^m$ and $g_c \in R^n$ respectively. Formally, the j-th element of g_q and g_c are computed as follows:

$$[g^q]_j = \max_{1<u<m} [G_{j,u}] \tag{3.7}$$

$$[g^c]_j = \max_{1<v<n} [G_{v,j}]. \tag{3.8}$$

The j-th element of vector g_c can be viewed as an importance score for the context around the j-th word in the candidate aspect c with respect to the entity context q. Similarly, the j-th element of vector g_q can be viewed as the importance score for the context around the j-th word in q with regard to c. This operation is adopted in [130] and applied over both rows and columns over 3D tensors.

Inspired by the row-wise and column-wise pooling in [99], we propose *q-singular max-sum pooling* (q-singular pooling) over convolved features. Specifically, given convolved feature matrix G, we first perform row-wise max pooling over G to get a scalar for each row. And then, scalars in each row are summed up as the final feature value. The q-singular pooling operation applied over two convolved feature matrices is illustrated in Fig. 3.2.

As an example, consider the first case where most entity context words are matched in candidate aspect, as shown in Fig. 3.2 (a). In contrast, the second case is shown in Fig. 3.2 (b), where one word in the entity context is matched in candidate aspect multiple times. Though the number of exact match counts is the same in both cases, we consider the first case as better matching between entity context and candidate aspect. Formally, the q-singular max-sum pooling is computed as follows:

$$\text{qs-pooling}(G) = \sum_{[1,m]} \max_{j \in [1,n]} [G_j]. \tag{3.9}$$

Q-singular pooling is applied to each convolved feature matrix and a scalar feature value is used.

3.3.5 Score

The convolutional matching and pooling operations make a convolutional block and the number of blocks used for each interaction is B. The feature vectors produced by convolutional blocks are concatenated to one feature vector h. We use one full-connected layers to produce the final matching score, which is obtained by Eq. 3.10:

$$score = h^T W + b. \tag{3.10}$$

where W and b are the weight and bias. The matching score is activated by a $\tanh$ activation function such that the matching score is between -1 and 1.

We adopt a pair-wise learning to rank objective to learn the parameters for predicting the final matching score:

$$L = \sum_{q,c+,c-} \max((1 - score(q, c+) + score(q, c-)), 0). \tag{3.11}$$

3.4 Experimental Setup

In the remainder of the chapter we address the following research questions:

RQ2.1 How does MICMN compare with state-of-the-art approaches for entity aspect linking?

RQ2.2 How do different components of MICMN affect the performance?

RQ2.3 What is the impact of parameters on MICMN?

3.4.1 Dataset

We evaluate our method using four datasets. The first dataset used in our experiments is EAL18 [77], which consists of 201 instances. Entity contexts are provided at three levels of granularity, i.e., sentence context, paragraph context and section context. Candidate aspects are represented by a header, a content with one or more passages and a list of entities that appears in content. The other three datasets, i.e. EAL-DB, EAL-DO and EAL-DS, are subsets of EAL19-D [78]. Compared with EAL18, candidate aspects is represented in a similar way with EAL18, while sentence context is the only entity context available. Statistics of the four datasets are shown in Table 3.1.

There are two major differences between EAL18 and the other three datasets. First, in addition to sentence context, there are also paragraph context and section context available in EAL18. Second, entity annotations are available for entity context in EAL18, while they are not available in EAL-DB, EAL-DO and EAL-DS. To keep the consistency of our experiments across datasets, we use the sentence context only as our entity context. Besides, we do not make use of entity annotations in either entity

Table 3.1: The minimum/average/maximum number of candidate aspects, length of aspect content and length of entity (sentence) context.

Name	Size	Entity Context Length	# Aspects	Aspect Length
EAL18	201	4/15.1/68	2/6.3/29	2/334/21756
EAL-DB	1067	3/17.9/70	2/9.4/140	1/210.7/13487
EAL-DO	13111	2/16.2/74	2/7.9/140	1/209.3/13770
EAL-DS	8338	2/16.7/67	2/7.4/140	1/223.8/13770

context or candidate aspects since we do not have information about entities in the entity context.

We use NLTK tokenizer[1] for preprocessing. Since only the sentence context is available in the three subsets of EAL-DB, EAL-DO and EAL-DS, we only use sentence context in EAL18 as the entity context. We split each dataset into 5 folds and conduct 5-fold cross validation. We use 70% of the data for training, 10% for validation and the remaining 20% for evaluation.

3.4.2 Baselines

In order to demonstrate the effectiveness of our proposed model, we employ the following baselines.

Naive baselines. *Random* ranks candidate aspects randomly. *Size* ranks aspects by the length of aspects. *Content overlap* ranks aspects by the number of overlapping tokens in q and c.

IR baselines. *tf-idf* and *BM25* are traditional information retrieval baselines.

Neural baselines. *K-NRM* [124] uses a translation matrix that models word-level similarities via word embeddings, a kernel-pooling technique that uses kernels to extract multi-level soft match features, and a learning-to-rank layer that combines those features into the final ranking score. *Conv-KNRM* [21] uses convolutional neural networks to represent q and c, performs soft matches between them and utlize kernel pooling and learning-to-rank layers for producing the final ranking score. Conv-KNRM is a state-of-the-art approach.

3.4.3 Evaluation metrics and parameter setting

Following [77], we employ Precision@1(P@1) and Mean Average Precision (MAP) to evaluate the experimental results. P@1 is the average number of times that the best answer is ranked first among all candidate aspects. MAP is the average of the average precision value for a set of queries.

To enable fair comparisons with the baselines, we adopt the same training setup in all experiments wherever possible, including embeddings, optimizer and hyper-parameters. We use the fixed Glove embeddings [82] with a dimension size of 50. Tokens that did not appear in the pre-trained word embeddings are replaced with a special token,

[1] http://www.nltk.org

of which the embedding is initialized randomly. We train all models with the Adam optimizer with an learning rate of 3×10^{-4}. The L2 regularization is set to 10^{-6} and a dropout of $d = 0.8$ is applied to all layers (except the embedding layer). The batch size is set to 64. The window size and number of feature map of CNN is 3 and 128.

The number of kernels in the kernel pooling layers in K-NRM and Conv-KNRM is 11 kernels/bins. The first one is the exact match kernel $\mu = 1, \sigma = 10^{-3}$, or bin $[1, 1]$. The other 10 kernels/bins equally split the cosine range $[-1, 1]$: the μ or bin centers are: $\mu_1 = 0.9, \mu_2 = 0.7, \ldots, \mu_{10} = -0.9$. The σ of the soft match bins was set to be 0.1 [124]. We implement our model using the MatchZoo library [37] and use its implementation of K-NRM and Conv-KNRM for our experiments.

3.5 Results

3.5.1 Quantitative results

The quantitative results on entity aspect linking are shown in Table 3.2. Each row of the table represents a method for entity aspect linking. As shown by results over EAL18 in Table 3.2, MICMN is significantly better than neural baselines,[2] while worse than content overlap and IR baselines on small datasets. Since the EAL18 dataset has only 201 instances, it shows that the IR baselines are the best choice when there is little training data.

On the other hand, when we turn to datasets with a large size, content overlap is the best choice. As we can see, content overlap achieves the best performing results on EAL-DO and EAL-DS, and is also competitive with tf-idf on EAL-DB. Our model is competitive with IR baselines and content overlap. The reason why we are not performing better might result from the fact that we are using limited information. We restrict the maximum length of the entity context and candidate aspect to 30 and 300 respectively for neural models. To show the impact of this restriction, we apply this restriction to content overlap and give the results indicated by *content overlap (lim)*. As shown by the third and fourth row, the limitation leads to a drop in terms of both P@1 and MAP, and the results are not as good as our model. The difference is significant on all datasets except EAL-DB. We consider it as an implication that when entity context and candidate aspects are not limited in length, MICMN is likely to perform better than naive baselines and IR baselines.

The performance of neural baselines is not as good as naive baselines and IR baselines, showing that existing neural ranking models are not naturally adaptable to entity aspect linking. The performance of MICMN is consistently better than neural baselines, which shows the superiority of MICMN for entity aspect linking.

3.5.2 Ablation study

An ablation study is performed to better understand the contribution of each module in our proposed model. By removing one component at a time from the full system

[2] We test for statistical significance using a paired t-test. All significance tests are performed at the $\alpha = .05$ level.

Table 3.2: Entity aspect linking results on four datasets. The abbreviation *cont. ov.* means content overlap.

Method	EAL18		EAL-DB		EAL-DO		EAL-DS	
	P@1	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**
random	21.89	45.32	17.13	39.62	17.02	39.88	18.52	41.70
size	40.40	61.14	44.42	61.50	43.91	62.65	43.47	62.33
cont. ov.	56.22	72.15	53.98	69.21	**52.75**	**69.26**	**55.77**	**71.50**
tf-idf	**59.70**	**74.52**	**56.89**	**70.76**	47.56	64.82	53.51	68.79
BM25	56.22	71.76	54.45	69.05	50.99	66.25	50.00	66.39
cont. ov. (lim)	57.21	72.71	48.73	65.53	43.41	63.57	50.49	67.88
K-NRM	27.82	50.18	24.74	46.19	34.07	54.57	26.90	49.31
Conv-KNRM	32.00	54.27	44.17	60.78	44.52	62.16	44.50	62.53
MICMN	49.28	68.01	56.32	70.42	50.48	66.37	53.69	69.21

Table 3.3: Results of MICMN with different components removed. EM means exact match; SM means soft match; SA-SM means self-attentive soft match.

Method	EAL18		EAL-DB		EAL-DO		EAL-DS	
	P@1	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**
MICMN	**49.28**	68.01	56.32	70.42	50.48	66.37	53.69	69.21
– EM	33.82	56.39	43.86	61.20	48.04	65.03	50.32	66.98
– SM	**49.28**	**68.32**	56.61	70.50	**51.50**	**67.46**	**54.81**	**69.84**
– SA-SM	46.29	66.16	**57.26**	**70.77**	50.37	66.54	54.56	69.78

and performing re-training and re-testing, we are able to study the effectiveness of each module. Here, we first study how exact match, soft match and self-attentive soft match contribute to the effectiveness of the model. Results on each dataset are shown in Table 3.3, with each row denoting the removal of a particular module. For example, the row "– exact match" represents removing the exact match module.

From the first two rows "– exact match", we can see that removing exact match leads to a significant effectiveness drop. This confirms that exact match is always an important indicator of relevance in semantic matching tasks. When the soft match component is removed, the results is better than that the case of self-attentive soft match being removed. This shows that weighted soft match can help improve the performance compared to naive soft match.

After studying the effectiveness of different interactions, we compare the proposed q-singular pooling with existing pooling methods. Results are shown in Table 3.4. Here we compare with max pooling and mean pooling. Comparing the first and the last column, we can see large drops when using max pooling compared to that of Q-singular pooling, indicating that Q-singular pooling is a better pooling strategy for entity aspect linking.

Table 3.4: Results of MICMN with different pooling strategies. QSP means q-singular pooling; MaP means max pooling; MeP means mean pooling.

Method	EAL18		EAL-DB		EAL-DO		EAL-DS	
	P@1	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**	**P@1**	**MAP**
QSP	**49.28**	**68.01**	**56.32**	**70.42**	**50.48**	**66.37**	**53.69**	**69.21**
MaP	42.82	62.55	53.32	67.14	46.60	65.03	49.30	65.74
MeP	40.84	61.81	51.73	66.41	48.02	64.76	49.76	65.94

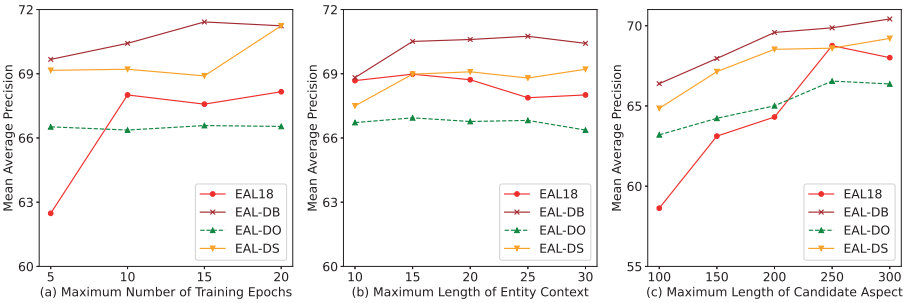


Figure 3.3: MAP scores with different maximum number of training epochs and maximum length of entity context and candidate aspect.

3.5.3 Parameter analysis

Results with different number of epochs. We vary the number of maximum training epochs max_epochs from 5 to 20 and report the results on the four datasets in Fig. 3.3 (a). With the increase of max_epochs, some improvements can be observed on EAL18, and not much difference is observed with other datasets. Actually, the performance becomes stable for bigger dataset when max_epochs is above 10. The results show that our model might converge after training for around 10 epochs.

Results with different maximum entity context length. We also explore the impact of the maximum length of entity contexts and candidate aspects. We first vary the maximum entity context length from 10 to 30 and fix candidate aspect length to 300. The result is shown in Fig. 3.3 (b). The figure shows that longer entity contexts slightly improve the performance. As we can see from Table 3.1, the length of most entity contexts is around 15 and covering instances with longer entity contexts is not that helpful for the overall performance.

Results with different maximum aspect length. Next we vary the maximum candidate aspect length from 100 to 300 and fix the maximum entity context length to 30. The result is shown in Fig. 3.3 (c). The results show that with the increase in candidate aspect length, the overall performance increases. This shows that longer candidate aspects are helpful. However, we also observe that the performance improvements are

getting smaller with the increase in aspect length. Since increasing aspect length means a higher computational cost, in a practical setting the costs and benefits need to be weighted carefully.

3.6 Conclusions

In this chapter, we answer RQ2 and have proposed a multi-interaction based convolutional matching network for entity aspect linking. We adopt exact match, soft match and self-attentive weighted soft match interactions between entity contexts around entity mentions and candidate aspects to extract matching patterns. We propose a novel pooling strategy to extract features from convolved features. Experimental results on four publicly available datasets have showed that our method is competitive compared with state-of-the-art baselines. An ablation study and parameter analysis are performed to study the impact of different components in our model and different parameter setup. For future work, we will explore more effective ways to model the interactions between entity contexts and candidate aspects.

4

It All Starts with Entities: A Salient Entity Topic Model

In the previous chapter, we conducted research on the task of entity aspect linking by modeling multiple interactions between query contexts of entity mentions and candidate aspects. In this chapter, we address RQ3, which is concerned with incorporating entity salience into document generative process.

4.1 Introduction

The importance of entities has been well recognized in domains as diverse as data mining [101], knowledge representation [121], language technology [58], and information retrieval [7]. Downstream applications in the aforementioned domains have benefited from modeling entities as vital sources of information in the generative process of documents. This has led to the development of a range of entity topic models, with entities either treated as external labels of documents [95] or observed variables [27]. For example, the Author Topic Model (95) assumes a topic distribution for each author, representing the research interest of authors. To generate a word in a document, an author is selected and a topic is sampled from the topic distribution of the author, before sampling a word from the topic distribution of the selected topic. In contrast, entities can be viewed as observed variables different from words in documents. For example, Link-LDA (27) models references of papers as observed variables to model the generation of academic articles.

One limitation of existing entity topic models is that none of them takes the salience of entities into account. Entity salience reflects the importance of an entity for a particular document. Entity salience can be characterized by local scoping and invariable perception [32]. Research about entity salience in web pages shows that fewer than 5% of the entities mentioned on a web page are salient for the page [32]. Intuitively, salient entities should play a more important role in the process of generating documents than non-salient entities.

This chapter was published as C. Wu, E. Kanoulas, and M. de Rijke. It all starts with entities: A salient entity topic model. *Natural Language Engineering*, pages 1--19, 2019.

In this work, we propose a novel topic model, *Salient Entity Topic Model* (SETM), that models salient entities in the generative process of documents. We model the generative process as a three-step procedure:

(1) sample a topic distribution for a document from a Dirichlet prior;

(2) sample salient entities using the topic distribution of the document; and

(3) sample words from the joint topic distribution combined from document topic distribution and salient entity topic distributions.

The advantage of SETM is that it models the mutual reinforcement between topics and entity salience. For example, if an entity is likely to be salient under a given topic, it will not only have higher probability to show up in documents around this topic, but also have higher probability to be generated as a salient entity. Another advantage of SETM is that if an entity e_a is salient in document d_a and document d_b is semantically similar to d_a, then an entity e_b (in d_b) that is similar to e_a is likely to be salient in d_b.

The assumption behind our model is that stories are built upon a story line (topic) and a set of main characters in the story (salient entities). Imagine that a news reporter is writing a news article about a specific story. The primary thing under consideration is what the document is about, which is modeled by the topic distribution of the document. The second thing is which entities are salient entities in the story described in the document. And finally, other words and entities are added to the document to complete the story.

Experiments on a publicly available dataset show that our model better explains and models the generative process of documents, outperforming state of the art methods. Both a qualitative and quantitative analysis is performed to demonstrate that by taking salient entities of documents into consideration, our model is better than similar models. The code of our work is published at Github.[1]

The main contributions of this work are as follows:

1. We propose a novel Salient Entity Topic Model (SETM) to model the generation of documents.
2. We derive a Gibbs sampling algorithm for parameter estimation.
3. We demonstrate the effectiveness of SETM to model text by performing both a qualitative and a quantitative analysis.

4.2 Related Work

Topic models have been widely used in text analysis.)) Latent Dirichlet Allocation (LDA) [11] models each document as a mixture of topics, and automatically generates summaries of topics in terms of a multinomial distribution over words. The original LDA has been extended in a wide variety of directions. Recent topic model extensions are either designed for specific tasks, such as multi-label classification [63, 64] and opinion mining [115], or particular kinds of texts, such as short texts [9, 65, 87, 131].

On the other hand, the notion of entity salience is attracting more attention [28, 32, 108, 125]. Gamon et al. [32] propose the task of identifying salient entities on web

[1] https://github.com/setm2nle/salient-entity-topic-model

pages. Tran et al. [108] take entity salience into consideration in ranking entities for summarization of high-impact events. Escoter et al. [28] group business news stories based on the salience of named entities. Xiong et al. [125] propose a Kernel Entity Salience Model to better estimating entity salience in documents so as to improve text understanding and retrieval.

In this work we extend on previous work by considering salient entities in modeling the generative process of documents. In this context, there are three branches of closely related work, i.e., entities as sources, entities as observed variables, and entities as entity topics. Below, we first summarize these three types of related work and clarify the differences between our model and previous work. Then, we discuss related work on topic labeling and clarify their difference with our work.

4.2.1 Entities as a source of information

In some scenarios, entities represent an external source of information that generates documents. For example, the Author Model (AM) [71] models document content and its authors' interests, where each author (that corresponds to an external entity corresponds to one topic. To generate a word, an author z is sampled uniformly from a set of authors of the document, and then a word w is generated by sampling from an author-word multinomial distribution. The Author Topic Model (ATM) [95] extends AM by introducing a topical layer between authors and words. An author's interests are modeled with a mixture of topics. Each document is associated to a set of observed authors. To generate a word, an author x is chosen uniformly from this set, then a topic is selected from the topic distribution of author x, and then a word is generated by sampling from a topic-specific multinomial distribution over words. The Author Recipient Topic Model (ARTM) [72] takes the recipient of messages into account. In ARTM, recipients of a message are also considered as authors of the message, and contribute to the generation of a particular message.

In all previous models, authors/entities are external labels, such as senders or recipients of messages. Similarly, we also consider salient entities as a source of information to generate documents. The distinguishing feature of our model is that we use entities that are both *observed* and *salient* in documents to model the sources of information. This distinction is important because unlike authors as external labels of documents, salient entities can serve as external labels and as representations of the content of documents at the same time. Hence, we hypothesize that salient entities capture more of the available information.

4.2.2 Entities as observed variables

Entities are different semantic units from words, hence they should be modeled as a special kind of observed variable. Link-LDA (LLDA) [27] models the generation of academic articles. In academic articles, references of papers can be viewed as entities. In the document generation process of LLDA, a topic distribution is sampled from a Dirichlet prior in the same way as in LDA. Then, a topic is sampled from the topic distribution of the document, and a word or entity is sampled from the topic-word or topic-entity distribution. To better model the correlation between words and entities,

CorrLDA2 [79] models word topics and entity topics separately, where word (entity) topics are used to generate words (entities). In the generative process, words are generated first, and then entity topics are sampled uniformly from all sampled word topics. Some authors propose an entity topic model for entity linking [38]; though it also considers entities in topic modeling, it is designed for entity linking, thus not directly comparable with our model.

In our work, we propose two variants of a topic model: one models words and entities with a single observed variable, while the other uses two observed variables to distinguish entities from words. The advantage of our model lies in the fact that we do not only consider entities as part of observed variables, but also incorporate entity salience information in the document generation process. In this way, our model can make the best use of available entity (salience) information.

4.2.3 Entities as entity topics

Entities can also be treated as special topics and contribute to the generation of documents together with general topics. For example, the Entity Topic Model (ETM) [51] learns the topical nature of entities. Similar to topics, entities are represented by a multinomial distribution over words. For each document, a topic distribution is drawn from a Dirichlet prior and a joint multinomial distribution over words Φ is obtained by linearly combining entities and topics of a document. To generate a word, a topic is sampled from Φ and a word is sampled from the topic word distribution. Though ETM seems to be a valid baseline for our work, it is not applicable because of scalability issues. It is applicable to short texts with few entities, such as abstracts of academic papers or small collections of news articles but not for long web documents. In contrast, the models that we propose do scale to large documents.

Another disadvantage of ETM is that it treats all entities equally, while in reality, salient entities are more important than non-salient entities. Compared to ETM, our model only introduces salient entities into the document generation process, which is more realistic.

4.2.4 Topic modeling vs. topic labeling

Existing work on topic modeling can be roughly classified into two categories. The first category proposes novel topic models for resolving particular applications, such as document classification [96], entity linking [38, 49] and question answering [47].

The second category focuses on improving topic modeling by incorporating new information. Kim et al. [51] propose an entity topic model for mining documents associated with entities. Xu et al. [126] incorporate wikipedia concepts and categories into topic models. Andrzejewski et al. [5] incorporate domain knowledge into topic modeling and conducts qualitative analysis on both synthetic and real world dataset. This work explores a new paradigm of improving over existing topic models, rather than solving a particular important downstream application. Our work aligns to this category.

Topic labeling, on the other hand, is to make topic representations learned by topic models more interpretable. Topics are conventionally represented by their top N

Table 4.1: Notations.

Symbol	Description
D	document collection
S_d	bag of salient entities in document d
E_d	bag of entities in document d
N_d	bag of words in document d
K	number of topics
θ_d	topic distribution of document d
ϕ_k	multinomial distribution over salient entities of topic k
ψ_e	multinomial distribution over entities of topic k
ψ_k	multinomial distribution over words of topic k
ρ_s	multinomial distribution over topics of entity $s \in S_d$
Φ	multinomial distribution over topics

words or terms [11, 35]. Recent work on topic labeling proposes to label topics using phrases [55], structured knowledge base data [44], entities [56], and even images [4]. Compared to topic labeling, which label topics mined by topic models, our work is focused on improving topic model itself by incorporating entity saliency.

4.3 Salient Entity Topic Model

4.3.1 Overview

We present two variants of the Salient Entity Topic Model (SETM), i.e., SETM-WO and SETM-WE. SETM-WO is a simplified version of SETM-WE, where documents are represented by a bag of words, while in SETM-WE, documents are represented by a bag of words and a bag of entities. The reason to have two variants is two-fold. First, we want to understand the effect of differentiating between words and entities as observed variables, if any. Second, there may be situations that such a separation provides flexibility; for instance, in academic articles, references can be viewed as entities, and hence considered separately from words, while in news articles, words and entities can be mixed together because they appear in the same context.

In what follows we focus on SETM-WE; SETM-WO can be considered to be a simplified version of SETM-WE and it is described only when this simplification affects the proposed algorithms.

The input used to train SETM-WE is a collection of documents, which consist of a bag of words and a bag of entities. An entity is a real world thing that has a corresponding entry in a knowledge base and is represented by a unique identifier. An entity can have multiple surface forms, which could be a unigram or n-gram. Entities are recognised by entity linking tools and this preprocessing step is not considered in our work. In other words, we take recognised entities as inputs. Each entity in a document has a binary label indicating whether the entity is salient or not. Formally, a document d is represented by a word vector N_d, where each $w_{d,i} \in N_d$ is chosen from

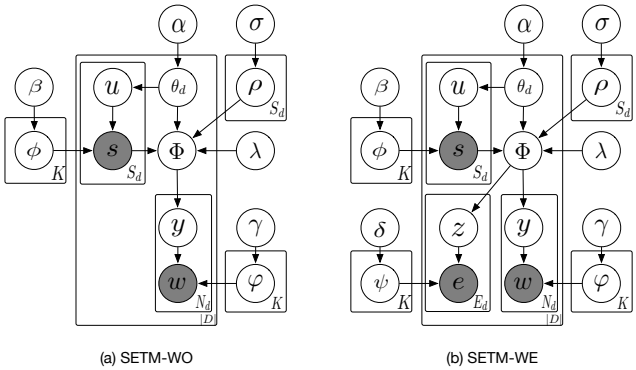


Figure 4.1: A graphical representation of the Salient Entity Topic Model (SETM).

the vocabulary of words W, and an entity vector E_d, where each $e_{d,j}$ is chosen from the vocabulary of entities E. Since we have salience labels for each entity, we have the set of salient entities in d, denoted as S_d. The goal is to discover word patterns of topics, and learn topic distributions of documents and entities. The notation used in the chapter is summarized in Table 4.1.

After model training, we need to infer the topic distribution of a newly incoming document using SETM. However, we might not have salience labels for new-coming documents. This is similar to the scenario considered by Labeled LDA (89). We adopt their strategy and perform inference by assuming that no entity is salient in the document.

4.3.2 The SETM Model

Graphical representations of SETM-WO and SETM-WE are shown in Fig. 4.1 (a) and Fig. 4.1 (b), respectively. A detailed explanation follows.

Hypotheses

The main hypotheses of our model are:

(1) salient entities are derived from the topics of a document; and

(2) salient entities themselves affect the generation of words and other entities in a document.

The intuition behind the first hypothesis is that the topics of a document are decided before choosing salient entities. When composing a story in an article, one first has some abstract story-line indicating the main theme of the story. For example, to write a news report on a football game, one first decides the topics, e.g., sports, and then adds teams, players and their interactions. The second hypothesis comes from the fact that non-salient entities may have some connection to salient entities, but they are loosely related to the theme of the document. For example, in the news item *Liberia: Former*

football striker George Weah wins presidential election[2], football club *Manchester City* is mentioned because the person of interest used to play for the club, though the club is not very important for this particular news article.

Algorithm 2 Generative Process of the SETM-WE Model.

for each topic k **do**
  Draw $\phi_k \sim Dir(\beta)$
  Draw $\psi_k \sim Dir(\gamma)$
  Draw $\psi_k \sim Dir(\delta)$
end for
for each entity e **do**
  Draw $\rho_e \sim Dir(\sigma)$
end for
for each document d **do**
  Draw $\theta_d \sim Dir(\alpha)$
  for each salient entity s **do**
    Draw topic $u \sim \theta_d$
    Draw salient entity $s \sim \phi_u$
  end for
  Obtain $\Phi_d = \lambda\theta_d + (1-\lambda)\frac{1}{|S_d|}\sum_{s \in S_d} \rho_s$
  for each entity e **do**
    Draw topic $z \sim \Phi_d$
    Draw entity $e \sim \psi_z$
  end for
  for each word w **do**
    Draw topic $y \sim \Phi_d$
    Draw word $w \sim \psi_y$
  end for
end for

Generative process

The generative process is shown in Algorithm 2. For each topic k, a topic salient-entity distribution ϕ_k, a topic entity distribution ψ_k and topic word distribution ψ_k are drawn from a Dirichlet prior with parameters β, δ and γ, respectively. For each document d, a multinomial distribution θ_d over topics is drawn from a Dirichlet prior with parameter α. Then, each salient entity $s \in S_d$ in the document is generated by first sampling a topic u from θ and then drawn from the topic-salient-entity distribution ϕ_u. To generate words and observed entities in document d, a joint topic distribution Φ is obtained by combining θ_d and the topic distribution of all salient entities of the document ρ_s ($s \in S_d$). Finally, words (or entities) are generated by first sampling a topic y (z), and then sampling a word (or an entity) from the topic word (or entity) distribution ψ (ψ).

[2] https://en.wikinews.org/wiki/Liberia:_Former_football_striker_George_Weah_wins_presidential_election

Note that ϕ and ρ are obtained from the same matrix, while from different perspective. ϕ is a matrix with size $K \times V_E$. From a row viewpoint, it is a a list of topics (ϕ_k), with each topic represented by a multinomial distribution over entities. When viewed from a column perspective, it is a list of entities, with each entity represented by a topic distribution (ρ_e).

One could also assume a switch distribution after Φ is derived, which is used to generate either words or entities. A similar switch distribution can be found in Switch LDA [79], as illustrated by the $Binomial(\psi_{z_i})$. However, we do not consider switch distribution in our model for the following reasons. First, we want to keep the flexibility of our model, so that it is still valid in cases where there is no direct connection between words and entities. For example, when analysing scientific publications, documents (papers) are represented by bag of words in abstracts and list of references of papers. In this case, it is inappropriate to have the switch probability since words in abstracts are very different from references. Second, given our model and the extension from CI-LDA to Switch LDA, we consider it straightforward to extend our model by taking the switch distribution into account when necessary.

4.4 Model Inference

Gibbs sampling is used for parameter estimation. Specifically, we repeatedly sample the topic assigned to each salient entity, word and entity in the document collection, given the topic assignment of the remaining salient entities, words and entities, as well as the priors. The inference process for SETM is detailed first, followed by a clarification of the difference between the inference process of SETM-WE and SETM-WO.

4.4.1 Inference of SETM-WE

Sampling salient entity topics s

The conditional posterior of assignment u_i to the i-th salient entity in document d is:

$$P(u_i = j|\mathbf{u}_{-i}, \mathbf{s}) \propto P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i})P(u_i = j|\mathbf{u}_{-i}), \tag{4.1}$$

where $\mathbf{u}_{-i}$ is the topic assignments of all salient entities except the i-th one. The first item on the right hand side is a likelihood and the second is a prior.

For the first term in Eq. 4.1, we have

$$\begin{aligned} &P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i}) \propto \\ &\quad \int P(s_i|u_i = j, \phi^{(j)})P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i})d\phi^{(j)}, \end{aligned} \tag{4.2}$$

where $\phi^{(j)}$ is the multinomial distribution over salient entities associated with topic j, and the integral is over all such distributions. We can obtain the rightmost item from Bayes's rule

$$P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i}) \propto P(\mathbf{s}_{-i}|\phi^{(j)}, \mathbf{u}_{-i})P(\phi_{(j)}). \tag{4.3}$$

Since $P(\phi^{(j)})$ is Dirichlet(β) and conjugate to $P(\mathbf{s}_{-i}|\phi^{(j)}, \mathbf{u}_{-i})$, the posterior distribution $P(\phi^{(j)}|\mathbf{u}_{-i}, \mathbf{s}_{-i})$ will be Dirichlet($\beta + n_{-i,j}^{(s_i)}$), where $n_{-i,j}^{(s_i)}$ is the number of instances of salient entity s assigned to topic j, not including the current salient entity.

Since the first term on the right hand side of Eq. 4.2 is just $\phi_{s_i}^{(j)}$, we can complete the integral to obtain

$$P(s_i|u_i = j, \mathbf{u}_{-i}, \mathbf{s}_{-i}) = \frac{n_{-i,j}^{(s_i)} + \beta}{n_{-i,j}^{(\cdot)} + V_S\beta}, \tag{4.4}$$

where $n_{-i,j}^{(\cdot)}$ is the total number of salient entities assigned to topic j, not including the current one.

For the second item in Eq. 4.1, we have

$$\begin{aligned} P(u_i = j|\mathbf{u}_{-i}) &= \int P(u_i = j|\theta_d)P(\theta_d|\mathbf{u}_{-i})d\Phi_d \\ &= \frac{n_{-i,j}^{(d_i,s_i)} + \alpha}{n_{-i,\cdot}^{(d_i,s_i)} + K\alpha}, \end{aligned} \tag{4.5}$$

where θ_d is the topic distribution of document d, $n_{-i,j}^{(d_i,s_i)}$ is the number of times salient entities from document d_i assigned to topic j except the current salient entity, and $n_{-i,\cdot}^{(d_i,s_i)}$ is the total number of salient entities in document d_i except the current one.

Putting together the results in Eq. 4.4 and Eq. 4.5, we obtain the conditional probability

$$P(u_i = j|\mathbf{u}_{-i}, \mathbf{s}) \propto \frac{n_{-i,j}^{(s_i)} + \beta}{n_{-i,j}^{(\cdot)} + V_S\beta} \frac{n_{-i,j}^{(d_i,s_i)} + \alpha}{n_{-i,\cdot}^{(d_i,s_i)} + K\alpha}. \tag{4.6}$$

Sampling word topics y

The conditional posterior of assignment y_i to the i-th word in document d is:

$$P(y_i = j|\mathbf{y}_{-i}, \mathbf{w}) \propto P(w_i|y_i = j, \mathbf{y}_{-i}, \mathbf{w}_{-i})P(y_i = j|\mathbf{y}_{-i}), \tag{4.7}$$

where $\mathbf{y}_{-i}$ is the topic assignments of all words except the i-th one. The first item on the right hand side is a likelihood and the second is a prior. By following a similar line of reasoning as from Eq. 4.2 to Eq. 4.4, we have

$$P(w_i|y_i = j, \mathbf{y}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \gamma}{n_{-i,j}^{(\cdot)} + V_W\gamma}. \tag{4.8}$$

For the second item in Eq. 4.7, by integrating over the multinomial distribution over topics for the document from which w_i is drawn, specified by Φ_d, we obtain

$$P(y_i = j|\mathbf{y}_{-i}) = \int P(y_i = j|\Phi_d)P(\Phi_d|\mathbf{y}_{-i})d\Phi_d, \tag{4.9}$$

where Φ_d is a combination of the document and salient entities in the document. In particular, the influence of the topic distribution of the document is weighted by λ compared with the influence from salient entities, and the topic distribution of salient entities are equally weighted. Finally, we have Φ_d represented as:

$$\Phi_d = \lambda\theta_d + (1-\lambda)\frac{1}{|S_d|}\sum_{s\in S_d}\rho_s.$$

Since $P(\theta_d)$ and $P(\rho_s)$ are Dirichlet priors Dir(α) and Dir(σ), the prior distribution $P(\Phi_t)$ is $\lambda\alpha + (1-\lambda)\sigma$. Since Φ_d is conjugate to the likelihood function (the first item in Eq. 4.9), the posterior distribution in Eq. 4.9 is as follows:

$$Dir(\lambda\alpha + (1-\lambda)\sigma + \lambda n_{-i,j}^{(d_i,w_i)} + (1-\lambda)\frac{1}{|S_d|}\sum_{s\in S_d} n_{-i,j}^{s}),$$

where $n_{-i,j}^{(d_i,w)}$ is the number of words assigned to topic j in document d_i except the current instance, and $n_{-i,j}^{s}$ is the number of instances of salient entity s assigned to topic j, except the current instance. Then by Dirichlet-multinomial conjugate, we have

$$P(y_i = j|\mathbf{y}_{-i}) = \frac{\lambda\alpha + (1-\lambda)\sigma + \lambda n_{-i,j}^{(d_i,w_i)} + (1-\lambda)(\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,j}^{s})}{K(\lambda\alpha + (1-\lambda)\sigma) + \lambda n_{-i,\cdot}^{(d_i,w_i)} + (1-\lambda)\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,\cdot}^{s}}. \tag{4.10}$$

Sampling entity topics z

The conditional posterior of assignment z_i to the i-th entity in document d is:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{e}) \propto P(e_i|z_i = j, \mathbf{z}_{-i}, \mathbf{e}_{-i})P(z_i = j|\mathbf{z}_{-i}), \tag{4.11}$$

where $\mathbf{z}_{-i}$ is the topic assignments of all entities except the i-th one. The first item on the right hand side is a likelihood and the second is a prior.

By following a similar line of reasoning as from Eq. 4.2 to Eq. 4.4, we have

$$P(e_i|z_i = j, \mathbf{z}_{-i}, \mathbf{e}_{-i}) = \frac{n_{-i,j}^{(e_i)} + \delta}{n_{-i,j}^{(\cdot)} + V_E\delta}, \tag{4.12}$$

where $n_{-i,j}^{(\cdot)}$ is the total number of entities assigned to topic j, not including the current one.

By following the steps we followed to derive Eq. 4.10 from Eq. 4.9, we have

$$P(z_i = j|\mathbf{z}_{-i}) = \frac{\lambda\alpha + (1-\lambda)\sigma + \lambda n_{-i,j}^{(d_i,e_i)} + (1-\lambda)(\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,j}^{s})}{K(\lambda\alpha + (1-\lambda)\sigma) + \lambda n_{-i,\cdot}^{(d_i,e_i)} + (1-\lambda)\frac{1}{|S_{d_i}|}\sum_{s\in S_{d_i}} n_{-i,\cdot}^{s}}. \tag{4.13}$$

4.4.2 Inference of SETM-WO

The Gibbs sampling process for SETM-WO is similar to SETM-WE, except that there is no sampling process for entity topic assignments in SETM-WO. In other words, the process of sampling entity topics does not exist in the inference process for SETM-WO because entities are not distinguished from non-entity words in SETM-WO.

4.5 Experimental Setup

In the remainder of the chapter we address the following research questions:

RQ3.1 How does SETM compare to state-of-the-art ETMs in terms of perplexity?

RQ3.2 How does SETM perform in the task of entity salience detection?

RQ3.3 Why can SETM achieve better performance in distinguishing salient entities from non-salient entities?

4.5.1 Datasets

The dataset used in our experiments at answering our research questions is the New York Times corpus, with salience annotations provided by Dunietz and Gillick [26]. We refer to this dataset as the NYT-Sal dataset. Annotations were automatically generated by aligning the entities in the abstract and the document and assuming that every entity occurring in the abstract is salient. The New York Times dataset consists of two partitions. Documents from 2003 to 2006 are used as the training set, while documents in 2007 are used as the test set. The number of documents in the training set and test set are $80,667$ and $9,706$, respectively. We further split the training set into a smaller training set (80%) for model training and a validation set (20%) for parameter selection. The size of the word vocabulary is $621,724$, including $189,480$ entities.

To analyze the performance on different types of entities, we categorize entities based on their document frequency and salience. In particular, we define the salience percentage of an entity e, $sp_e = \frac{SDF_e}{DF_e}$, as the percentage of the documents in which entities appear and are labeled as salient, where SDF_e is the number of documents in which entity e is salient. The salience percentage (SP) and the log of document frequency (DF) for each entity in the collection are shown as a scatter plot in Fig. 4.2. We choose two threshold values to define high and low salience entities and high and low frequency entities. The lower and upper salience thresholds are set to 0.05 and 0.5 respectively, indicated by the red solid line $y = 0.05$ and the red dashed line $y = 0.05$. We define entities whose document frequency higher than 400 (approximately 5% of all training documents) or lower than 5 as head and tail entities respectively. The thresholds are indicated by the solid blue line $x = 1.6$ ($\ln(5) = 1.6$) and the dashed blue line $x = 6$ ($\ln(400) = 6$).

We consider entities that satisfy the following conditions as torso entities: (1) entities for which salience percentage is above 0.05 and below 0.5; (2) entities for which document frequency is above 400 and below 5. In other words, torso entities fall into the square formed by the lines in Fig. 4.2.

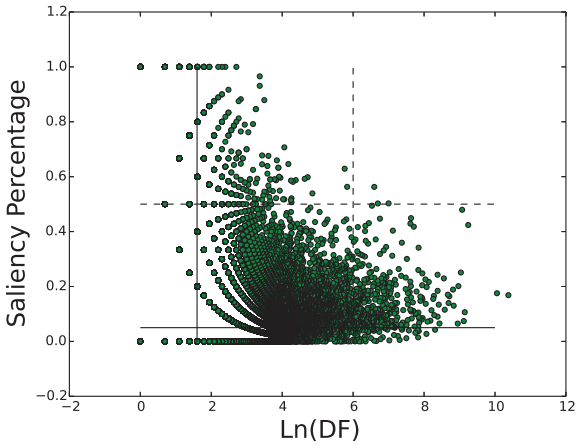


Figure 4.2: Scatter plot of log DF and SP.

4.5.2 Intrinsic evaluation

The first type of evaluation we conduct is an intrinsic evaluation. We quantify the ability of the SETM to represent entities and documents better than baseline entity topic models by computing the similarity between topically similar entities, and the similarity between topically similar documents. We further quantify the ability of the SETM to generate new documents by computing the perplexity of our model. Instead of analyzing all entities, we focus on entities that are neither highly frequent (HDF) nor rare (LDF). This way, we avoid any possible bias introduced by head or tail entities. We want to perform analysis on entities with neither abundant nor limited information.

Entity-to-entity topical similarity

First, we test the ability of our topic model to produce an effective representation of entities compared to the baseline models. We make the assumption that two ("torso") entities are topically similar if both entities are salient in more than 50% of the documents they co-occur. Out of all entity pairs, 141 fulfill this condition. We test our model against baseline models by computing the cosine similarity of these entity pairs; the higher the computed similarity is the better the topic model.

Document-to-document topical similarity

Second, we test the ability of our topic model to produce an effective representation of documents compared to baseline models. Given an entity e, we denote with D_e^s the set of documents where entity e is salient, and with D_e^{ns} the set of documents where entity e is not salient. To measure the topical coherence of a set of documents, we follow the definition of coherence score due to Kulkarni et al. [54], and define the topical

coherence of a set of documents D related to e as

$$topical\text{-}coherence(e, D) = \sum_{m=2}^{D} \frac{1}{m-1} \sum_{l=1}^{m-1} \cos(d_m, d_l).$$

Our hypothesis is that the topical coherence calculated by using the document representations learned by the SETM will be higher than baseline models, which means that our learned document representations are better in capturing topical similarity. We use the set of 567 "torso" entities.

Model perplexity.

Perplexity is a standard measure for estimating the performance of a probabilistic model. We evaluate SETM by estimating the perplexity of unseen held-out documents given a set of training documents. A better model will have a lower perplexity of held-out documents on average. We follow the perplexity definition used by Blei et al. [11]. For a test set of M documents, perplexity is defined as follows

$$perplexity(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \right\}. \tag{4.14}$$

4.5.3 Extrinsic evaluation

The second type of evaluation we conduct is an extrinsic evaluation. We first quantify the usefulness of the document representations learned by the SETM by the task of entity salience detection. For each document, we measure the similarity between the document and its set of salient entities, and that of its set of non-salient entities. We further measure the divergence between these two similarities to identify the capability of our model in capturing the topical differences.

Entity salience detection

To evaluate the learned topic distributions of entities, we test our model on the task of entity salience detection. The goal of entity salience detection is to iterate over each entity in a document and identify whether the entity is salient or not.

Our classification setup is as follows. First of all, we train a SETM model using the training set and the information about the salience of entities in that set. Then, for each training instance (an entity document pair), the topic distribution representations of the entity and the document are used as features to train a classifier. For each test entity document pair, we infer the topic distribution of the document and make predictions about whether the entity is salient or not in the document. Since entity saliency information is document specific, we have no prior information about the saliency of an entity in the test documents during classification.

We assume that if a model learns better entity and document representations, it should achieve higher classification performance. It is important to note that in this work we do not compare our proposed method with the current state-of-the-art entity saliency systems, such as SWAT [86]. This is due to the fact that the focus of this work

is to model text in a more faithful way, around topics and salient entities, use the task of salient entity detection as a way to compare the learned topic distributions of our model with that of baseline topic models, rather than improving the state-of-the-art performance over entity salience detection (which is approximately around 0.56 F1 score for part of our dataset).

Following Dunietz and Gillick [26], we use a set of standard binary classification metrics, i.e., recall, precision and F1, to quantify the classification performance. Note that since the majority of entities are non-salient our metrics are calculated only over the positive class, i.e., salient entities. Statistical significance of the observed differences between the performance of two methods is tested using a two-tailed paired t-test and is denoted by $^{\blacktriangle}$ for strong significance for $\alpha = 0.01$, and $^{\triangle}$ for weak significance for $\alpha = 0.05$. In our experiments, all models are tested for significance against the best performing baseline, CorrLDA2. In addition to evaluate the performance on all entities, we also analyze over head and tail entities as defined in Fig. 4.2.

Document-entity similarity divergence analysis

To intuitively understand the performance, we analyze the topical similarity between salient entities and non-salient entities within individual documents.

The reason to perform an analysis on the basis of individual documents is that entity saliency is document specific. In other words, an entity could be salient in one document while not salient in another, which makes analysing salient entities across document impossible. Ideally, we expect that the similarity between salient entities and documents is higher than that of non-salient entities and documents. By visualising the divergence between these two similarities for each document we can see how close we are to the ideal situation compared to baseline models.

Given a document d, we denote with E_s the set of salient entities, and with E_{ns} the set of non-salient entities. We calculate the similarity between each salient entity $s \in E_s$ and document d, and obtain the average similarity $avg\text{-}sim(E_s, d)$ across all salient entities and the document. We do the same for E_{ns} and obtain $avg\text{-}sim(E_{ns}, d)$. The assumption is that the better a model is the larger the difference between $avg\text{-}sim(E_s, d)$ and $avg\text{-}sim(E_{ns}, d)$. Then, we calculate the *se-ne-divergence* as $avg\text{-}sim(E_s, d) - avg\text{-}sim(E_{ns}, d)$, and rank documents based on the divergence (which ranges from 1 to -1) in descenting order. The higher the divergence value is, the better the model.

Entity topic analysis

Given an entity e, we have a collection of documents D_s where e is salient in $d \in D_s$ and another collection of documents D_n where e appears in $d \in D_n$ and is not salient. We first compute the average topic distribution of documents in D_s and D_n respectively to find topics that are most relevant with e. Then we present the top words under those relevant topics to see their relevance with the given entity. We choose entity *New York Jets*, a professional American football team located in New York, as an example. The size of D_s and D_n is 407 and 403, respectively.

Table 4.2: Methods and baselines used for comparison.

Acronym	Description	Ref.
LDA	Latent Dirichlet Allocation, which use latent topic distributions to represent documents.	[11]
LLDA	Link-LDA, similar with LDA, except that it considers words and entities in documents separately.	[27]
CorrLDA2	Correlated topic model, which models the correlation between word topics and entity topics.	[79]
SETM-WO	Our proposed model with only one observed variables, i.e., words.	This chapter
SETM-WE	Our proposed model with two observed variables, i.e., words and entities.	This chapter

4.5.4 Baselines and parameter settings

Table 4.2 lists the entity salience detection methods considered in our experiments. Since our goal is to evaluate the effectiveness of our topic model, we compare with existing topic models, such as LDA [11], LLDA [27], CorrLDA2 [79]. LDA is used as a simple baseline to showcase how a standard model without considering entities works in our setting.

Beyond the baselines mentioned, there is a growing body of work on topic models that involve entities [46]. However, their focus is on sequential topic flows of entities and entity groups in a single document [46] or on dynamic topic hierarchies and timeliness of news data [43]. Our task and our focus is not on the dynamics of topics. Therefore, such methods are not included as baselines.

Last, there is work in the literature that explicitly focuses on entity salience detection, such as [26]. This work is not included in our comparison since they target developing discriminative models with a specific focus on entity salience detection. Our goal is different, that is, to evaluate topic distributions learned by topic models. A comparison with such algorithms is beyond of the scope of this work.

Following standard practice [51], we set the hyperparameters of the baseline methods and our models to pre-defined values. In LDA, LLDA, CorrLDA2, and our models, we set both α and β as 0.1. The number of iterations of Gibbs Sampling is set to 1,000 for all topic models. For perplexity analysis, we set the number of topics to 5, 10, 15, 20, 30, 40, 50, 80, 100. For model analysis and extrinsic evaluation, we use the corresponding model trained with the number of topics set to 100.

4.6 Results

4.6.1 Intrinsic evaluation

Entity-to-entity similarity

The results on the entity representation analysis are presented in Table 4.3. The average

Table 4.3: Entity representation analysis.

Model	Similarity
LDA	0.6960
LLDA	0.7240
CorrLDA2	0.1392
SETM-WO	0.7271
SETM-WE	0.7336

Table 4.4: Document representation analysis.

Model	Similarity SD	Similarity NSD	Ratio of difference
LDA	0.6585	0.4943	1.3322
LLDA	0.6906	0.5405	1.2778
CorrLDA2	0.9293	0.9301	0.9991
SETM-WO	0.6722	0.5141	1.3075
SETM-WE	0.6641	0.4916	1.3509

similarity of LLDA is higher than LDA, indicating that by distinguishing words from entities as observed variables we obtain better entity representations. This is also demonstrated by the comparison between SETM-WO and SETM-WE. Further, we can observe that SETM-WO outperforms LDA and that SETM-WE outperforms LLDA. This demonstrates that incorporating entity salience information into the topic models can be helpful in learning good entity representations, regardless of the setting of observed variables in topic models. Here the entity representation learned by CorrLDA2 is not performing well. The reason might be that the entity topics are forced to align with word topics in documents, which makes entity representations meaningless.

Document-to-document similarity

The results on the document representation analysis are presented in Table 4.4. We expect the documents in D_e^s to be topically coherent, while documents in D_e^{ns} not. Therefore, the higher the value of *Similarity SD* the better, while the lower the value of *Similarity NSD* the better. To combine these two metrics, we calculate the ratio between them, and the higher the ratio the better. As we can see in the results, the ratio achieved by SETM-WE is the highest, which means that by considering entity salience information, our learned document representations can actually capture the similarity between similar documents better, and make dissimilar documents more distinguishable. The results of CorrLDA2 is below 1.0, which indicates that the topic coherence of D_e^s is even lower than that of D_e^{ns}. This partially demonstrates that the topic distributions learned by CorrLDA2 are not as good as other topic models.

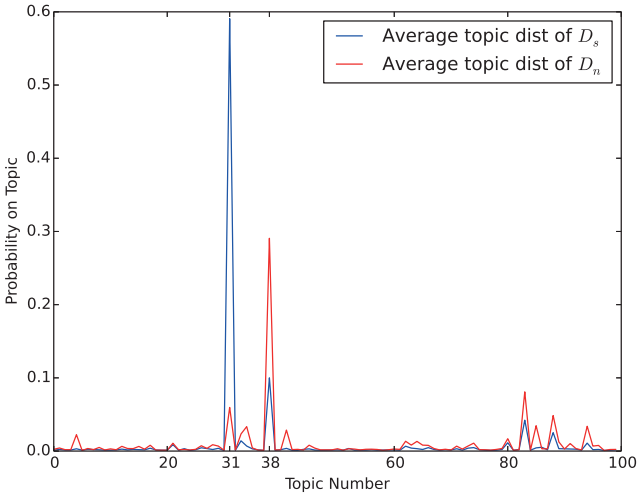


Figure 4.3: Average topic distribution between documents where entity *New York Jets* is salient and documents where it is not.

Entity topic analysis

We first present the average topic distribution between D_s and D_n in Figure 4.3. As we can see, topic 31 stands out in the blue line, indicating the relevance between entity *New York Jets* and the collection of documents where it is salient. Similarly, topic 38 is the most relevant topic in the red line. Note that the probability of topic 31 is close to 0.6 and much higher than that of topic 38, which indicates higher coherence within the salient documents of *New York Jets*.

We present the top 10 words under topic 31 and 38 in Table 4.5. It is obvious that both topics are closely related to sports and American football. The difference is that topic 38 is a more general topic about National Football League (NFL), where words such as "super", "bowl", and "season" appear frequently. On the other hand, topic 31 is more relevant to entity *New York Jets*. "Jets" is one word in the name of the team, while "edwards" is the surname of a professional player of the team[3]. By analysing on the basis of individual entity, we find that it is possible to explain the learned topics. Therefore, we consider it helpful to take entity salience into account in topic modeling whenever possible.

Perplexity

Fig. 4.4 shows the perplexity values of our models and the baselines under different number of topics. Since the baseline models do not have entity salience information in their models, they cannot take advantage of salience labels. As we can see in Fig. 4.4 (a), our models and Link-LDA outperform LDA and CorrLDA2. For Link-LDA, the

[3]https://en.wikipedia.org/wiki/Lac_Edwards

Table 4.5: Top 10 words under topic 31 and topic 38 in a SETM model trained on the NYT-Sal dataset.

Topic 31	**Topic 38**
jets	giants
west	football
team	game
edwards	season
stadium	bowl
club	coach
diamond	team
south	nfl
east	super
game	players

reason is that it distinguishes entities from words when learning topic distributions in documents. For the case of our models, it is better because the entity salience information is incorporated into the generative process of documents. Link-LDA performs slightly better than our models. This might be because during inference we assume no entity salience information, which has a negative impact on the inferred topic distributions of documents.

To study the perplexity of different observed variables, we present the perplexity of words, salient entities, and non-salient entities in Fig. 4.4 (b), (c), and (d), respectively. For LDA, the perplexity is lower for words, while much higher for salient or non-salient entities. This is not surprising since the number of words is larger than the number of entities in documents and LDA is biased to be better at generating words than entities. For LLDA, CorrLDA2 and SETM-WE, the perplexity of entities is obviously lower than that of words, demonstrating the effective of distinguishing entities from words. Both of our model variants are better than the baseline models, showing that our model incorporates entity salience information into a topic model in an effective manner.

4.6.2 Extrinsic evaluation: Entity salience detection

The overall results on the entity salience detection tasks are shown in Table 4.6. As we can see, the performance of our models on all entities is better than other methods in terms of F1. This demonstrates the effectiveness of our model by learning better topic distributions for entities and documents. Our model has the highest precision, but lower recall, which means that our model makes fewer positive predictions. This makes sense, since the dataset is biased to negative instances. Note that we are not comparing our work with the work by Dunietz and Gillick [26] because their goal is to optimize for the task of entity salience detection, while our goal is to compare the entity and document representations.

Results on seen, head and tail entities are also shown in Table 4.6. As we expect, the overall performance on seen entities is better for all models. Compared to baseline models, the recall of our models is higher while sharing similar precision. The precision

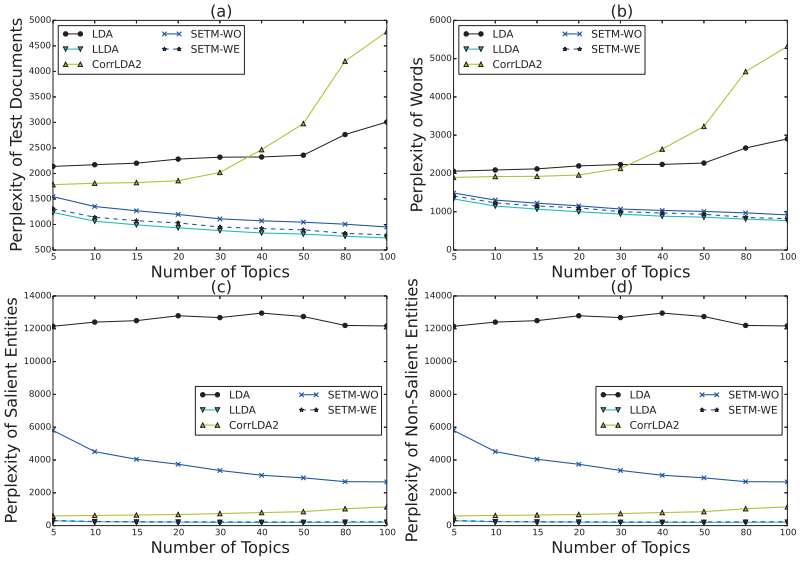


Figure 4.4: Perplexity of (a) Document, (b) Words, (c) Salient Entities, and (d) Non-Salient Entities.

of head entities is significantly better than baseline models. The reason is that we have more training examples on positive and negative examples on entity salience for head entities. This demonstrates that with more training examples, our model learns the salience of entities better by showing better capability at predicting entity salience. For tail entities, the performance of all models are similar. This is because little information is available for tail entities, and the strength of our models can not be leveraged by tail entities.

The result of topical similarity analysis within individual documents is shown in Fig. 4.5. Ideally, we expect that all lines are above zero and as close to $y = 1$ as possible, indicating that for each document, the average similarity between salient entities in the document and the document is higher than that of non-salient entities. We can observe that the lines of our models: (1) are higher than baseline models, especially in the beginning; (2) cross the $y = 0$ line later than baseline models. This demonstrates that our models are more capable in distinguishing salient entities from non-salient entities. As we can see, CorrLDA2 shows relatively consistent behavior across documents. Together with the results of LLDA, they demonstrate that modeling entities in topic models might not help leraning the salience of entities.

4.7 Conclusions

In this chapter, we answer RQ3 and have proposed to incorporate entity salience information into topic models. A novel Salient Entity Topic Model (SETM) is proposed that can explicitly model the generation of documents with salient entities under considera-

Table 4.6: Performance of entity salience detection methods on the NYT-Sal dataset.

	P	R	F1	P	R	F1
	All entities			**Seen entities**		
LDA	0.1362▲	0.8875▲	0.2361▲	0.1372▲	0.8334△	0.2348▲
LLDA	0.1606	0.5896	0.2493	0.1718▲	0.4673▲	0.2509
CorrLDA2	0.1544	0.6664	0.2507	0.1551	0.6659	0.2516
SETM-WO	0.1700▲	0.5184▲	0.2560▲	0.1717▲	0.5256▲	0.2589▲
SETM-WE	0.1718▲	0.5038▲	0.2562△	0.1736▲	0.5046▲	0.2583▲
	Head entities			**Tail entities**		
LDA	0.1598▲	0.8860▲	0.2708▲	0.1221▲	0.9067▲	0.2152▲
LLDA	0.1998	0.6273	0.3005	0.1294	0.4680	0.1990
CorrLDA2	0.1854	0.7484	0.2972	0.1269	0.5787	0.2081
SETM-WO	0.2348▲	0.5123▲	0.3220▲	0.1340▲	0.5261▲	0.2136
SETM-WE	0.2372▲	0.4878▲	0.3192▲	0.1347▲	0.4967▲	0.2120

tion. A Gibbs sampling-based algorithm is proposed for the parameter estimation of the model. We compare our model with several state-of-the-art baselines in terms of the generative capability. The evaluation shows that our model is better than the baselines, which demonstrates the effectiveness of incorporating entity salience information into document generative process. We also evaluate the learned document representations and entity representations by the task of entity salience detection. The results show that the representations of document and entities using our model can better distinguish salient entities out of non-salient entities compared to baseline representations.

Our model can be used for topic analysis with the increasingly available entity salience information, either extracted from web log [32] or news corpus [26]. As a potential application, by performing clustering on documents where a particular entity is salient, we might find different aspects of the entity by detecting the difference in learned topic distributions of documents.

One of the limitations of our model lies in the fact that training our model requires large scale and high quality labels of entity salience. However, this can be approximated by automatically mining salience information from existing data, such as the soft labeling approach introduced by Gamon et al. [32], which we leave as future work. On the other hand, we now assume binary entity salience annotations. However, one salient entity might be more important than another salient entity. It might be better to assign weights ranging from 0 to 1 to salient entities so that different levels of importance can be reflected.

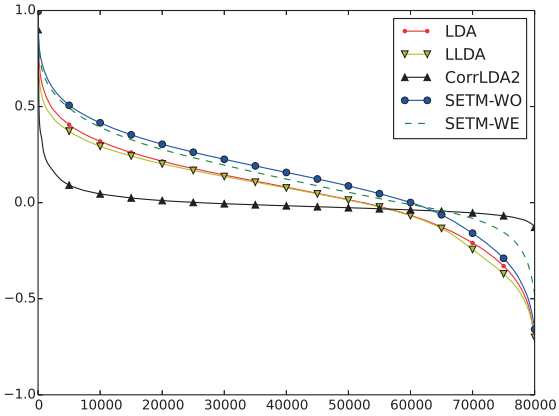


Figure 4.5: Topical similarity analysis on documents in training set. X-axis is the number of documents and y-axis is the *se-ne-divergence* of a document, as described in Section 4.5.3. Documents are ranked by their divergence values in descent order.

5

WN-Salience: A Corpus of News Articles with Entity Salience Annotations

In the previous chapter, we work on incorporating entity salience into document generative process given binary entity salience information. In this chapter, we address RQ4, which is concerned with automatically extracting entity salience annotations from WikiNews.

5.1 Introduction

Text modeling has traditionally made no distinction between different terms in the text. Examples include bag of words representation, language models, and term weighting methods. Research on knowledge extraction and text semantics has shifted some of the attention towards utterances that represent real world entities, while recent work on entity linking [102] has made it possible to take entities into consideration in various downstream applications, such as information retrieval [22, 91].

Various corpora annotated with entities have been built for entity related research, such as FACC1 [31]. However, these corpora make no distinction between salient and non-salient entities, despite the fact that only few entities are central to a document. For instance, in the Web domain, fewer than 5% of the entities on a web page are salient to the page [32]. Many existing publications have recognized the importance of understanding entity salience [30, 86, 108, 116, 125]. For example, automatically suggesting news pages for populating Wikipedia requires determining whether a news article should be referenced by an entity, considering several aspects of the article, including entity salience, relative authority, and novelty of the article [30]. In general, there is a growing interest in understanding entity salience, demonstrated by research on entity salience detection [26, 32]. Therefore, it is very important to be able to quantify the salience of an entity.

To facilitate research involving entity salience, datasets with both entity annotations and salience labels are necessary. Ideally, one would like to have human annotators labeling salient entities in documents. Unfortunately, this is not scalable due to the high

This chapter was published as C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. WN-Salience: a corpus of news articles with entity salience annotations. In *LREC 2020*, pages 1--8. LREC, 2020.

volume of documents that need to be annotated and the cost of human labor. At the same time, with the rise of deep learning algorithms datasets should consist of tens of thousands of annotations to allow effective learning.

A small number of datasets [26, 32] have been developed, to facilitate research on entity salience. However, existing datasets suffer from several limitations: (1) computational errors in entity annotations, (2) strong assumptions in collecting entity salience labels, and (3) noise in entity salience labeling. For example, in the NYT-salience dataset [32], entities in documents are identified by applying an NP extractor, a co-reference resolver, and an entity resolver, which might propagate mistakes to the final annotations. Gamon et al. [32] assume a soft labeling approach: if users click on a web page link after they issue an entity query, the entity is likely to be salient in the web page. It is also believed that heuristic design is a difficult proposition [32].

To address the aforementioned limitations, we propose a method to extract a new dataset by collecting news articles from WikiNews,[1] and build a new dataset referred to as *WN-Salience*. WikiNews is a free-content news source wiki, where anyone can write news articles. In each article, text fragments referring to entities are linked by the article authors to Wikipedia pages corresponding to the respective entity or WikiNews categories. Though WikiNews itself is multi-lingual, without loss of generality, we focus on English language news articles only, given the popularity and the number of articles in the language. We believe that our method can be applied to other languages as well.

Our method is based on the following observation. Authors are highly advised to link news articles to WikiNews categories, to allow effective information organization in WikiNews, and do so only when a category is strongly related to the written article. Therefore, the categories can be viewed as salience annotations and entities corresponding to these categories as salient entities.

To illustrate the utility of the developed WN-Salience dataset, we conduct experiments on entity salience detection. By applying simple algorithms, we confirm the effectiveness of positional features in entity salience detection found in [26], but also demonstrate the inferiority of other hand crafted features found discriminative in the literature, which shows that this dataset is challenging and likely orthogonal in some aspects to existing datasets. The dataset is available on GitHub.[2] We follow the license policy of WikiNews and publish the dataset under a free license.[3]

The main contributions of this work are summarized as follows:

1. We propose a method for extracting human-annotated entity salience labels using WikiNews categories and in-text entity annotations.

2. We develop a new dataset for research around entity salience.

3. We analyze our dataset and compare it with previous datasets.

4. We conduct experiments to demonstrate the utility of the dataset.

[1] https://en.wikinews.org/wiki/Main_Page
[2] https://github.com/researchdatasets/wn-salience-dataset
[3] https://creativecommons.org/licenses/by/2.5/

5.2 Related Work

5.2.1 Notion of salience

A recent definition of entity salience is given in [32]. Gamon et al. [32] first declare that a thing that has a Wikipedia page associated with it to be an entity and then present a notion of entity salience using two assumptions, i.e., local scoping and invariable perception. Local scoping indicates that the salience of an entity in a document can be solely determined by the document itself, while invariable perception means that entity salience can be assessed independently from the interests of readers, and independently from the prior importance of the entity as it exists outside of the document. Another notion of entity salience is more empirical: salient entities are those that human readers deem most relevant to the document [26].

Even though they are reasonable, the two assumptions above are not easy to handle in practice. In this work, we adopt an assumption similar to the empirical definition of entity salience: salient entities are those that authors of articles deem most relevant to the document. Given an article, there might be hundreds or thousands readers, while there can only be one or few writers. Instead of considering salience from the perspective of readers, we adopt the opinion of writers. Two advantages of the assumption are the following: first, the potential inconsistency between different readers is avoided; and second, it is easier to capture authors' opinion on salience than that of readers, which makes it more convenient to collect explicit salience labels.

Table 5.1: Comparison of existing datasets on entity salience.

Dataset	**Entity Annotations**	**Salience Labels**	**Size**
MDA dataset	proprietary NER pipeline	soft labeling	∼50,000
nyt-salience	proprietary NLP pipeline	heuristic rules	100,976
Reuters-128	human labeling	crowdsourcing	128
WikiNews	human labeling	crowdsourcing	604
WN-Salience	human labeling	automatic deriviation	∼7,000

5.2.2 Existing datasets

Gamon et al. [32] propose to identify salient entities in web pages by using a soft labeling approach based on behavioral signals from web users as a proxy for salience. The assumption is that when a user issues an entity query and clicks on an URL on the returned results page, the entity is salient in the corresponding web page. For pages that receive enough traffic, reliable user click statistics can be obtained and used to derive entity salience labels. As a result, a dataset called Microsoft Document Aboutness (MDA), was constructed. A major limitation of the dataset is that it is not publicly available. Furthermore, it is also hard to reproduce a similar dataset without access to large scale web search log data. Another limitation of the approach is that the assumption relies on the behavior of web users, which is known t come with bias, e.g.,

position bias [18] or domain bias [45].

The New York Times salience (NYT-Salience) benchmark collection introduced by Dunietz and Gillick [26] is built on top of the New York Times corpus [98]. To build the NYT-Salience corpus, two steps were taken, recognizing entities and assigning salience labels. Given a document and its abstract, a standard NLP pipeline was first run to identify entities both in the abstract and in the text of the news article; then, entities in the abstract were aligned with entities in the document. Entities in the document that also appear in the abstract are considered salient. Two limitations lie in NYT-Salience. First, entities are identified by a multi-step NLP pipeline, which might lead to errors in entity annotations. Second, the dataset is only partially available. The NYT-Salience dataset does not provide the underlying textual content along with the annotations due to copyright restrictions.

The Reuters-128 Salience dataset is a corpus built on top of Reuters-128 [94], an English corpus built for evaluating NER systems, which contains 128 news articles in economy. The entity salience labels are obtained by crowdsourcing [24]. The key limitation of the dataset is its small size, which does not allow for the development of supervised learning algorithms. In addition, the entity annotation process used might suffer from errors introduced by entity linking tools. Finally, entities in the dataset are uniquely identified by Wikipedia titles, DBpedia urls and others. Ideally, it is expected that all entities come from the same knowledge base. If entities are identified by entities in different knowledge bases, then many additional processing steps are needed whenever it is necessary to refer to information in knowledge bases.

The WikiNews dataset [110] is constructed for salient entity linking, which combines the task of entity linking and entity salience detection. Since WikiNews is a collection of news articles with entity annotations, the creators created entity salience labels and used them for salient entity linking. The entity salience labels are collected using a crowdsourcing platform. The dataset creators define entity salience using a 4-grade metric, i.e., *top relevant*, *highly relevant*, *partially relevant* and *not relevant*. To deal with subjectivity in the assignment of salience scores, the salience scores from multiple annotators are averaged. Though also extracted from WikiNews, this dataset is different from our dataset. First, Trani et al. [110] use graded scores to measure salience. Second, we exploit the category information to induct entity salience labels automatically, while they rely on annotators from crowdsourcing platform.

5.2.3 Summary

Here, we summarize existing datasets involving entity salience and present the comparison in Table 5.1. In terms of entity annotation, manual entity annotation is preferred over entities tagged by entity recognition pipelines. For salience labels, human annotated salience labels are considered to be more reliable. However, human annotated salience labels rely on crowdsourcing, which is usually very expensive. Therefore, we prefer to derive salience labels using automated methods.

As we can see, existing datasets suffer from either less preferred entity annotations (MDA dataset and NYT-Salience) or the limitation of expensive salience label collection method (Reuters-128 and WikiNews). By making use of entity annotations in WikiNews articles and categories assigned to articles by writers, our dataset is able to use human

annotated entity annotations and collect salience labels using automated methods. In this way, we avoid either limitation. As for the size of corpus, our dataset is of moderate size compared to existing datasets.

5.3 WikiNews and Annotations

WikiNews is a Wikipedia project, the mission of which is to present reliable, unbiased and relevant news.[4] News articles in WikiNews are written by volunteers, who can write or edit a page by expanding it, correcting facts and so on. There are various types of article in WikiNews, such as original reporting,[5] interviews,[6] daily summaries[7] and so on. For example, interview articles usually start with background descriptions of interviews, followed by conversations between interviewers and the interviewees.

In this work, we mainly focus on two types of article in WikiNews, i.e., synthesis articles and original reporting. Synthesis articles are written by collecting media reports from many other sources (always fully cited), synthesizing them into a single article. Bias is stripped out and a neutral point of view is presented. Original reporting articles are first-hand news reports written by WikiNews contributors on-the-spot of news events.[8] The reason we only focus on these two types of article is that they are usually the most typical and popular types, that are also frequently observed on the web.

A typical WikiNews article consists of a title, body content with **in-text annotations**, related news, sources, and **WikiNews categories**. In the rest of the work, we will use the example WikiNews article, entitled "*Koreas hold joint training session for Olympics*."[9] Among all the elements of a WikiNews article, WikiNews categories and in-text annotations within the body content are the important ones for constructing our dataset; they are introduced below.

5.3.1 WikiNews categories

In WikiNews, every article needs to be listed under one or more categories, so that articles under a particular category can be easily found. The process of selecting appropriate categories is guided by the following principle provided by WikiNews: "*Typically, both a "location" category (where did the news event take place?) and a "topic" category (what is the event about?) is required*."[10] For example, an article about a computer science conference in Brussels might have the following categories: *Computer Science*, *Brussels*, and *Belgium*. Such a set of categories can be seen at the bottom of every WikiNews article.

[4] https://en.wikinews.org/wiki/Main_Page
[5] https://en.wikinews.org/wiki/Wikinews:Original_reporting
[6] https://en.wikinews.org/wiki/Category:Interview
[7] https://en.wikinews.org/wiki/Category:Wikinews_Shorts
[8] https://en.wikinews.org/wiki/Wikinews:Introduction
[9] https://en.wikinews.org/wiki/Koreas_hold_joint_training_session_for_Olympics?dpl_id=2833718
[10] https://en.wikinews.org/wiki/Wikinews:Writing_an_article

5.3.2 In-text annotations

WikiNews encourages authors to add wikilinks when textual fragments (i.e., entity mentions) are referring to entries in other Wiki sites, such as categories in WikiNews, and article pages in Wikipedia. These wikilinks are considered in-text annotations.

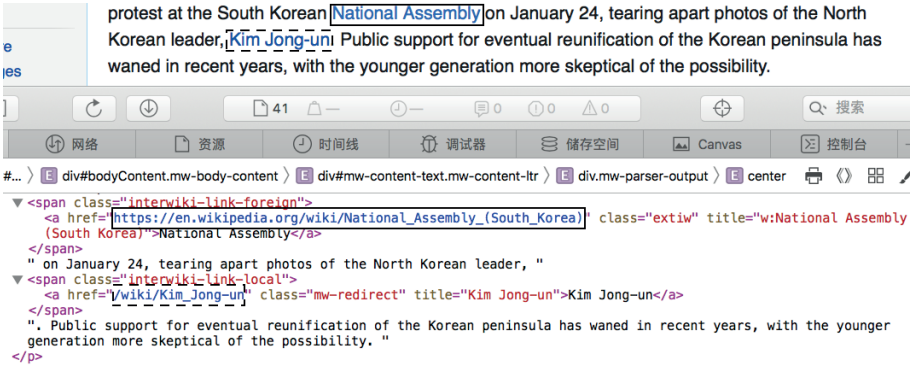


Figure 5.1: Examples of WikiNews category annotation (dash line box) and Wikipedia entity annotation (solid line box).

WikiNews articles typically contain two types of in-text annotation, WikiNews category annotations, and Wikipedia entity annotations, as shown in Fig. 5.1. WikiNews category annotations are links to WikiNews categories. For example, in the example article, the entity mention *Kim Jong-un* is representing an entity and has corresponding WikiNews category *Kim Jong-un*.[11] As a result, a wikilink is added to refer to the WikiNews category *Kim Jong-un*. Wikipedia entity annotations are links to Wikipedia entities. For example, the text fragment *National Assembly* in the example article can be linked to the corresponding Wikipedia page *National_Assembly_(South_Korea)*.[12]

We observe that even though many WikiNews categories correspond to Wikipedia entities, authors annotate entity mentions by WikiNews categories first, and by Wikipedia pages only when WikiNews categories are not available.

5.4 Entity Salience Hypothesis

In this section, we present our entity salience hypothesis, which is used to induce salience labels in our datasets. Based on how WikiNews categories are annotated and how WikiNews category pages are organized, we propose the following hypothesis: *an entity is salient if the WikiNews category that corresponds to the entity is also labeled as a category of the article*. In contrast, if an entity in an article is labeled as a category that is not included in the set of the article categories, or if it is labeled as a Wikipedia page, it is not salient in the article.

To illustrate the above hypothesis, we examine the example article mentioned in Section 5.3. In the example article, categories such as *North Korea*, *South Korea*,

[11] https://en.wikinews.org/wiki/Category:Kim_Jong-un
[12] https://en.wikipedia.org/wiki/National_Assembly_(South_Korea)

Olympics, *Ice Hockey*, *Kim Jong-un*, and *Moon Jae-in* are labeled as categories by the author of the article. Based on the main content of this article, we can observe that the two countries and the two presidents represent the "*main characters*" of the story presented, while *Olympics* and *Ice Hockey* serve as the topic explaining the reason why the characters connect with each other in this article. And it is clear that the category entities labeled here are salient entities in the article.

On the other hand, we can see that category entities that are not annotated as a category of the article are not salient entities. For example, categories such as *Seoul* are not labeled as a category of the example article. *Seoul* appears when the article mentions the historical fact that the 1988 Summer Olympics happens in Seoul, and this fact is not related to the main story of article. Therefore, it is not a salient entity of the article, and is not labeled as a category of the article.

Note that some categories of articles might not appear in the body content of articles. Since our focus is the salience of entities in documents, we do not consider entities that do not appear in documents, even though they might be helpful for document understanding. We preserve all categories of articles in our dataset, including the categories that are simple dates.

5.5 The WN Salience Dataset

In this section, we first describe the dataset extraction process, including the categories collection and the articles collection process. Then, we show some basic statistics of the dataset, and analyze entity salience within and across documents.

5.5.1 Dataset collection

We collect raw web pages from WikiNews, and parse them using jsoup.[13] Given the elements in WikiNews articles, we extract the following fields: *title*, *date*, *body content*, *categories*. Note that we keep the paragraph structure of articles to facilitate possible scenarios where paragraph information is needed. For each paragraph, we extract the main text and the annotations. The information in each annotation includes mention text, the corresponding entity (Wikipedia title or WikiNews category), position in the paragraph (begin offset and end offset).

On the basis of our aforementioned entity salience hypothesis, we include in each annotation a binary entity salience label (1 for salient entities, 0 otherwise). Since our focus is to extract a dataset for entity salience related tasks, we focus on articles that have at least one salient entity. The collection process consists of two steps, i.e., collecting categories and collecting articles under selected categories.

Collecting categories. In WikiNews, categories are organized in a hierarchy, where each category belongs to at least one parent category. The root category of the WikiNews category hierarchy is *Internal WikiNews organization*, which belongs to itself. If we start from *Internal WikiNews organization*, and iterate over subcategories of each category, we are able to iterate over all categories.

[13] https://jsoup.org

Table 5.2: Statistics of WN Salience. The numbers on the lower part are document-wise. Document length and paragraph length are counted in terms of words.

	Train set	Test set
# of articles	5928	1040
Avg. doc length	335	679
Avg. paragraph length	50	78
Avg. # of paragraphs	6.7	8.7
Avg. # of unique entities	12.5	14.2
Avg. # of annotations	13.0	19.2
Avg. # of categories	11.9	15.0

Instead of iterating over all categories and parsing all articles, we consider a category as a target category if it satisfies the following criterion: the WikiNews category has a corresponding Wikipedia page. The reason for this is that we want to have a unified representation of salient entities, the Wikipedia entity unique identifier. Imagine an extreme case, where the only salient entity is a WikiNews category and the WikiNews category does not have corresponding Wikipedia entity. Then the salient entity would be just a unique identifier and does not have connection to any knowledge base. This is undesirable because: (1) there is no guarantee that all WikiNews categories are entities; (2) in existing datasets involving entity salience, all (salient) entities are knowledge base entities, either Freebase entities or Wikipedia entities; and (3) it would prevent research that involves entity salience and knowledge bases.

Note that there are also categories that are irrelevant to our purpose. For example, news articles whose titles start with *WikiNews interviews* are very different documents compared to ordinary news report. Other examples include *WikiNews Shorts*, *Original reporting*, *Translated news*, *Photo essays*, *Published*, *Archived* and so on. These categories are not meaningful categories in terms of representing some real world entity. Instead, they are either for the purpose of website organization (e.g., Published and Archived), or for the purpose of guiding the writing of authors (e.g., Photo essays). However, no filtering is needed for these categories because they usually do not have corresponding Wikipedia pages. In the end, 4,214 categories are found, out of which 1,813 categories have corresponding Wikipedia pages.

Collecting articles. We iterate over the collected category pages and obtain the articles within each category. We iterate over all articles in all categories and obtain 11,005 articles. Then we select articles that have at least one salient entity, which means that at least one category of an article is an entity that appears in article body. In the end, we obtain 6,968 articles, which constitute the WN-Salience dataset.

5.5.2 Dataset statistics

To facilitate supervised methods, we divide all articles into a training set and a test set. Temporal splitting is an intuitive way to construct a training and a test set. In previous work, temporal splitting by year was used. However, we observe that basic statistics

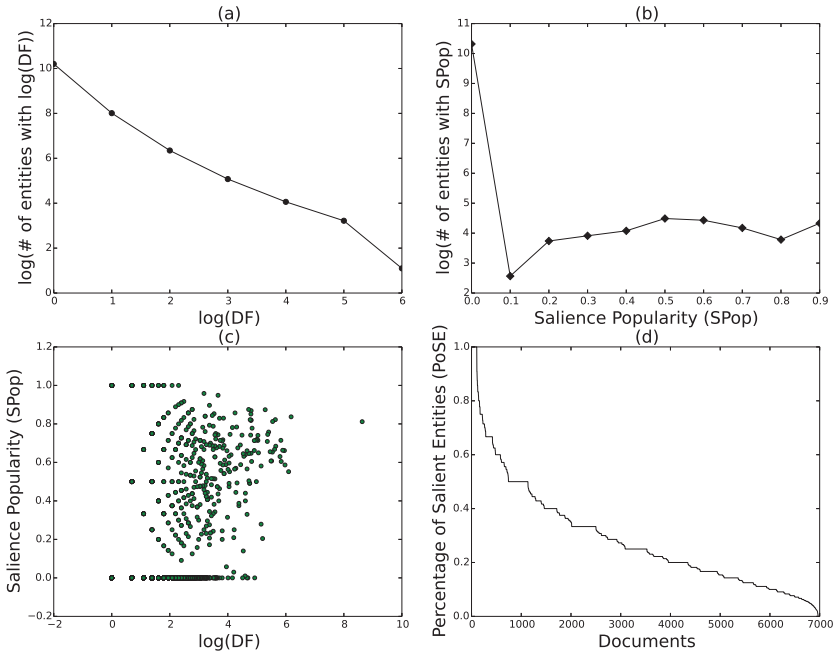


Figure 5.2: Analysis of the WN-Salience dataset.

show major differences between news articles in different years. Therefore, we choose to split the dataset on a monthly basis, i.e., all articles up to a threshold month are placed in the training set, while the remaining articles are placed in the test set. We set the threshold month to September. Basic statistics of our dataset are shown in Table 5.2.

5.5.3 Dataset analysis

In order to have an intuitive understanding of the statistics of our dataset (WN-Salience), we perform an analysis of how document frequency and salience popularity of entities are distributed. For the purpose of comparison, we also present a similar analysis results of the NYT Salience dataset.

Entity document frequency (DF). We present the distribution of log document frequency of entities in Fig. 5.2 (a) and Fig. 5.3 (a). Since the document frequency of entities varies a lot from high frequency entities to low frequency entities (power law distribution), we focus on the scale of the document frequency of entities. Specifically, we put entities whose log document frequency under the same scale into the same group, and present the log of the number of entities in each group. As shown in the results, the statistics of WN-Salience are similar to those of NYT Salience.

Entity salience popularity (SPop). The salience popularity of entity e is defined as SDF_e/DF_e, where SDF_e is the number of documents where e is salient and $DF(e)$

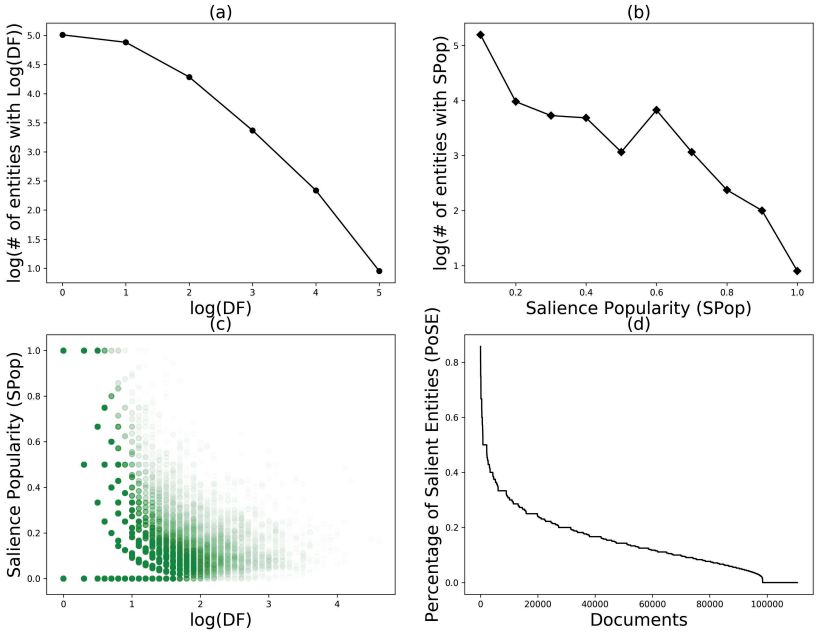


Figure 5.3: Analysis results of NYT-Salience dataset.

is the document frequency of e. We count the log number of entities whose salience popularity range from $[sp, sp + 0.1]$, where $sp \in [0, 0.9]$. The results are shown in Fig. 5.2 (b) and Fig. 5.3 (b). In both datasets, the SPop of many entities are zero, which indicates that entity salience is skewed towards few entities. More entities in NYT Salience dataset shows moderate salience percentage (0.3 to 0.7) compared to that of WN-Salience. This indicates that it might be more difficult to identify salient entities in WN-Salience compared to NYT Salience.

DF vs. SPop. To see how document frequency and salience percentage of entities correlate with each other in our dataset, we represent each entity as a two dimensional point in a figure, where the two dimension are its DF and SPop. The results are shown in Fig. 5.2 (c) and Fig. 5.3 (c). Entities tend to be evenly distributed in WN-Salience and skewed towards bottom-left in NYT Salience. This shows that with the increase of document frequency, the SPop of entities in NYT Salience is very likely to decrease, while that in WN-Salience can still remain high.

Percentage of salient entities (PoSE) of documents. The percentage of salient entities (PoSE) in document d is defined as S_d/E_d, where S_d is the number of salient entities in d, while E_d is the total number of entities in d. We count PoSE in each entity and rank them in descent order. The results are shown in Fig. 5.2 (d) and Fig. 5.3 (d). As we can see, the PoSE of most entities is lower than 5%, which conforms with the observation in [32] that fewer than 5% entities on a web page are salient to the web

page.

5.6 Experiments

5.6.1 Research questions

We address the following research questions:

RQ4.1 How consistent is salience annotation between our dataset and the WikiNews dataset proposed in [110]?

RQ4.2 Does the small number of existing WikiNews categories affect the quality of salience labels?

RQ4.3 How do baseline methods on entity salience detection perform on our dataset?

5.6.2 Comparative analysis between datasets

An existing dataset with salience labels proposed by [110], referred to as the *SEL-WikiNews* dataset, has also been extracted from WikiNews. Given the same origin, we are able to perform a comparative analysis between SELWikiNews and our dataset. To make the comparison possible, article matching and entity alignment are necessary. In particular, we first identify a common set of articles by title matching, i.e., only articles with the exactly same title are selected. Then, we match the entities across the datasets. Entities in SELWikiNews dataset are represented as entity id in Wikipedia, while in our dataset, entities are represented by their Wikipedia title. We process the 2018.07.20 Wikipedia dump to extract the mapping from entity id to its Wikipedia title, so that we can match entities between the two datasets.

After extracting a common set of articles and making entities comparable, we perform salience label matching to validate annotation consistency. The salience score in SELWikiNews ranges from 0.0 to 3.0, while in our datasets, we have binary salience labels, indicating whether an entity is salient or not. We propose to use simple rules to flatten the salience scores in SELWikiNews to binary labels: if the salience score of an entity is above a predefined threshold value, the entity is salient and it is not salient otherwise. Then, we use the salience labels derived from SELWikiNews as ground truth, and those in our datasets as predictions. We choose binary evaluation metrics over the salience labels, including precision, recall, F1, and accuracy in our experiments. We use three threshold values, i.e., 1.0, 2.0 and 3.0, to see the results for different levels of saliency.

The article title matching identified 243 articles that exist in both datasets. The results for different thresholds are shown in Table 5.3. Since the individual salience score given by annotators in SELWikiNews range from 0.0 to 3.0, and the final score is the average score of multiple annotators, we consider 2.0 as a reasonable threshold for the flattening process. The results for the other two thresholds are given for comparison.

As we can see, our dataset has a reasonable accuracy, which is around 0.6. The high recall and moderate precision indicate the fact that we are more aggressive at assigning salience labels to entities. This can be either due to the fact that (1) human

Table 5.3: The results of comparing salience annotations in WN-Salience dataset against that in SELWikiNews. Each row presents the results under different threshold of salience score.

	WN-Salience			
Threshold	P	R	F1	Acc.
3.0	0.0433	0.8750	0.0825	0.4166
2.0	0.4031	0.8556	0.5480	0.5971
1.0	0.9784	0.6079	0.7499	0.6046

Table 5.4: In-text annotation statistics. CE stands for category entity.

# of Wikipedia entity annotation	49,556
# of WikiNews category annotation	19,534
# of CE as Wikipedia entity annotation	2,002
# of CE as WikiNews category annotation	15,968
# of other annotations	3,086

annotators who created SELWikiNews are more cautious in annotating salient entities (low precision), or (2) article writers tend to annotate more salient entities (high recall). Therefore, we consider our dataset as complementary to existing datasets given its different method of salience annotation.

5.6.3 Risk of missing salient entities

Table 5.4 provides statistics about the in-text annotation of entity mentions. We define as category entities, those entity mentions that have both a corresponding WikiNews category and a Wikipedia page. As one can observe from this table, when an entity mention is a category entity, then the chance of the writer annotating it as a WikiNews category is about 89%, while the chance of annotating it as a Wikipedia page is 11%. This is rather important, given that only entities annotated as WikiNews categories can be considered for salience. What is worrying, however, is that if we consider all annotations, 70% of those are Wikipedia page annotations. This means that there is a large number of entity mentions for which there is no corresponding WikiNews category, and hence they are annotated as Wikipedia pages. This also means that these 49,556 entity mentions will never be considered for salience.

Since not all entities have corresponding WikiNews categories, there might be a risk of missing salient entities. We refer to this risk as *low recall risk* (LRR), since it might lead to lower recall than it should be. We investigate this issue by measuring the impact of LRR. In particular, we extract subsets with decreasing LRR and present ESD results of the subsets. In principle, if all entities are category entities, LRR does not exist, since all entities will be considered for salience. The higher the ratio of category entity in articles is, the lower LRR is. To measure LRR, we define the ratio of category entity in an article as follows:

$$ce\text{-}ratio = \frac{N_{ce}}{N_{ce} + N_{nc} + N_{pe}},$$

Table 5.5: WN-Salience subsets with different levels of ce-ratio, and their comparison against SELWikiNews. The results of WN Salience is using threshold 2.0 to convert graded scores in SELWikiNews to binary labels.

ce-ratio	# of docs	P	R	F1
0.5	111	0.4021	0.8519	0.5463
0.6	69	0.3974	0.8564	0.5429
0.7	39	0.3808	0.8505	0.5260
0.8	19	0.4091	0.8333	0.5488
WN Salience	243	0.4031	0.8556	0.5480

where $N_{i,ce}$, $N_{i,nc}$ and $N_{i,pe}$ represents the number of category entity annotations, WN category annotations and WP entity annotations in the i-th article. Note that WN category annotations represent categories that does not have a corresponding Wikipedia page. We extract subsets of WN Salience by specifying $ce\text{-}ratio$ ranging from 0.5 to 0.8 and compare against SELWikiNews dataset.

After extracting WN-Salience subsets under different $ce\text{-}ratio$, we compare each subset against SELWikiNews as was done in Section 5.6.2. The results are shown in Table 5.5. As we can see, the value of all metrics of these subsets are quite close and there is no clear winner between subsets under different levels of $ce\text{-}ratio$. Therefore, we assume that the LRR risk can be neglected for our dataset.

5.6.4 Application: Entity salience detection

Since the focus of this work is to introduce a new dataset for tasks involving entity salience, we run simple algorithms to showcase the use of our dataset. We choose to evaluate on the task of entity salience detection over WN-Salience.

We follow the work of Dunietz and Gillick [26]. In particular, we use some hand-crafted features to train a binary classifier to identify whether an entity is salient in a document. Because of the difference between our dataset and their dataset (NYT-Salience), we do not follow all their implementation in complete detail. We use Naive Bayes as our classifier.

We consider three kinds of feature, i.e., positional features, count features and entity centrality features. Positional features are investigated here because they achieve reasonable performance. Since count of head word is actually ambiguous, we use entity frequency in articles as count features. Following [26], we also apply the function $f(x) = \text{round}(\log(k(x+1)))$ to count features, and k is set to 10 in our experiments.

We use precision, recall, and F1 on salient entities as our evaluation metrics. In all experiments, a classification threshold of 0.5 is used by default, since in each case it is close to threshold that maximized F1.

Table 5.6 shows experimental results on two datasets on the task of entity salience detection. As we can see in the results, positional features achieve reasonable performance, which conforms with the results in [22]. Adding the first location of an entity does not help much. The reason is that they are both positional features and thus indicate similar information.

Table 5.6: The results of entity salience detection over two datasets.

	NYT-Salience			WN-Salience		
Features	P	R	F1	P	R	F1
positional baseline	0.5598	0.4095	0.4730	0.4794	0.5322	0.5044
head count	0.3346	0.5221	0.4078	0.2422	0.2138	0.2271
mentions	0.4198	0.4167	0.4182	0.2422	0.2138	0.2271
1st-loc	0.1901	0.4133	0.2604	0.2908	0.7890	0.4250
+ head count	0.3206	0.7079	0.4413	0.2643	0.8124	0.3988
+ mentions	0.3919	0.5970	0.4732	0.2920	0.4806	0.3633
+ centrality	0.3506	0.6554	0.4568	0.2921	0.4850	0.3646

To our surprise, features that are used to approximate entity frequency, i.e., head counts and mentions, have a negative impact on the performance. As also observed by [22], the precision decreases on both datasets compared to the positional baseline. However, the recall shows different trends (increasing on NYT Salience, decreasing on WN-Salience). This might come from the fact that documents in WN-Salience are not very long and entities might not appear in documents many times, which makes entity frequency less meaningful as a feature.

The effectiveness of using entity centrality feature is not as good as expected. Comparing the performance in two datasets, it works better in NYT Salience. The recall decreases a lot after using the centrality feature in WN-Salience, which means that the salience of entities is less sensitive to centrality rank, compared to NYT Salience.

5.7 Conclusions and Future Work

In this chapter, we answer RQ4 and uncover entity salience information in WikiNews website. Based on our observations, we propose an automated method to extract datasets with entity salience annotations, which leverages the category annotations in WikiNews news articles. Our extracted dataset, WN-Salience is presented. Experiments are performed to validate our proposed assumptions, measure the consistency between our dataset and an existing dataset and set a benchmark for evaluating on the task of entity salience detection. We believe that WN-Salience will stimulate the development of more advanced method for entity salience detection and salient entity linking. Here we focus on English language only. Our method for extracting a similar dataset in other languages is possible.

This chapter serves as an endeavour to promote research related to entity salience detection. Since the focus of all previous chapters is not entity salience detection, we did not use WN-Salience in the thesis. If we can effectively identify salient entities in documents, it is possible to enhance document understanding by treating salient entities and non-salient entities in different ways, e.g., our work in chapter 4. On the other hand, this chapter is also related to entity aspects. Imagine the case when an entity is salient for a document, it is likely that the document reflects a particular aspect of the entity.

If we want to mine entity aspects from documents, it would be much more efficient to focus on documents where the entity is salient, rather than all documents that the entity appears in. We posit that advances on entity salience can potentially promote research on entity aspects.

6 Conclusions

In this concluding chapter, we first revisit our research questions introduced in Chapter 1 and summarize the main findings and implications of our research in Section 6.1. Then, in Section 6.2, we describe the main limitations of our studies and the possible future directions.

6.1 Main Findings

6.1.1 Learning entity-centric document representations

We started with the task of learning entity-centric document representation by modeling entities with multiple aspects. In Chapter 2, we asked the following question:

RQ1 Can we learn entity-centric document representations by modeling entities with multiple aspects?

We refined RQ1 into the following questions:

RQ1.1 Can we confirm that entities have multiple aspects, with different aspects reflected in different documents?

RQ1.2 Can we learn a representation of entity aspects from a collection of documents, and a representation of documents based on multiple entities and their aspects as reflected in the documents?

RQ1.3 Does this novel representation improve algorithm performance in downstream applications?

RQ1.4 What is a reasonable number of aspects per entity?

We put forward the hypothesis that entities are not monolithic concepts; instead, they have multiple aspects, and different documents may be discussing different aspects of a given entity. Given that, we argue that from an entity-centric point of view, (a) a document related to multiple entities should be represented differently for different entities (multiple entity-centric representations), and (b) each entity-centric representation should reflect the specific aspects of the entity discussed in the document. We model each entity using multiple aspects (facets), where each entity facet is represented as a

mixture of latent topics. Then, given a document associated with multiple entities, we assume multiple entity-centric representations, where each entity-centric representation is a mixture of entity facets for each entity. Finally, a novel graphical model, the Entity Facet Topic Model (EFTM), is proposed in order to learn entity-centric document representations, entity facets, and latent topics.

We confirmed that entities are multi-faceted concepts that we can model and learn. We show that a multi-faceted entity-centric modeling of documents can lead to effective representations. Through experimentation we confirm that such a representation can have an impact in downstream applications, and considering a small number of facets is effective enough. In particular, we visualize entity facets within a set of documents, and demonstrate that, indeed, different sets of documents reflect different facets of entities. Further, we demonstrate that the proposed entity facet topic model generates better document representations in terms of perplexity, compared to state-of-the-art document representation methods. Moreover, we show that the proposed model outperforms baseline methods in the application of multi-label classification. Finally, we study the impact of EFTMs parameters and find that a small number of facets better captures entity specific topics, which confirms the intuition that on average an entity has a small number of facets reflected in documents.

6.1.2 Improving entity aspect linking using a neural network based approach

We work on improving entity aspect linking by using multiple interaction-based convolutional neural networks and a novel pooling strategy. In Chapter 3, we asked the following question:

RQ2 Can we learn entity-centric document representations by modeling entities with multiple aspects?

We refined RQ2 into the following questions:

RQ2.1 How does MICMN compare with state-of-the-art approaches for entity aspect linking?

RQ2.2 How do different components of MICMN affect the performance?

RQ2.3 What is the impact of parameters on MICMN?

We proposed a multi-interaction based convolutional matching network for entity aspect linking. Our approach is an interaction-based approach which uses pre-trained embeddings to represent text sequences. Given the context of input entity and candidate aspects, we construct multiple interactions, i.e., exact match, soft match and self-attentive weighted soft match between them. These interactions are passed to convolutional neural network to generate convolved features. A novel pooling strategy is devised to extract features from convolved features.

Experimental results on four publicly available datasets have showed that our method is competitive compared to state-of-the-art baselines. An ablation study is performed to demonstrate the effectiveness of our proposed pooling strategy. A parameter analysis

shows that our model can converge after training for around 10 epochs. In addition, we show how the length of input entity contexts and candidate aspects affect the overall performance of entity aspect linking.

6.1.3 Incorporating entity salience information into topic modeling

We started with the task of learning entity-centric document representation by modeling entities with multiple aspects. In Chapter 4, we asked the following question:

RQ3 Can we improve entity aspect linking using a convolutional neural network based approach?

We refined RQ3 into the following questions:

RQ3.1 How does SETM compare to state-of-the-art ETMs in terms of perplexity?

RQ3.2 How does SETM perform in the task of entity salience detection?

RQ3.3 Why can SETM achieve better performance in distinguishing salient entities from non-salient entities?

We proposed to incorporate entity salience information into topic models. A novel Salient Entity Topic Model (SETM) is proposed that can explicitly model the generation of documents with salient entities under consideration. A Gibbs sampling-based algorithm is proposed for the parameter estimation of the model.

We compare our model with several state-of-the-art baselines in terms of the generative capability. The evaluation shows that SETM is better than the baselines, which demonstrates the effectiveness of incorporating entity salience information into document generative process. We also evaluate the learned document representations and entity representations by the task of entity salience detection. The results show that the representations of document and entities using our model can better distinguish salient entities out of non-salient entities compared to baseline representations.

6.1.4 Extracting entity salience annotations from WikiNews

We started with the task of learning entity-centric document representation by modeling entities with multiple aspects. In Chapter 5, we asked the following question:

RQ4 Can we automatically extract entity salience information from WikiNews?

We refined RQ4 into the following questions:

RQ4.1 How consistent is salience annotation between our dataset and the WikiNews dataset proposed in [110]?

RQ4.2 Does the small number of existing WikiNews categories affect the quality of salience labels?

RQ4.3 How do baseline methods on entity salience detection perform on our dataset?

We uncover entity salience information in WikiNews website. Based on our observations, we propose an automated method to extract datasets with entity salience annotations, which leverages the category annotations in WikiNews news articles. Our extracted dataset, WN-Salience is presented.

Experiments are performed to validate our proposed assumptions, measure the consistency between our dataset and an existing dataset and set a benchmark for evaluating on the task of entity salience detection.

6.2 Future Work

In this section, we discuss limitations of the research discussed in this thesis and list possible directions for future work.

Varying the number of aspects. In Chapter 2, we proposed entity facet topic models which model entities as mixtures of aspects. We assume a fixed number of entity aspects for all entities. However, real world entities usually have different numbers of aspects. As we can imagine, a popular entity that can be widely seen on the web should have more aspects than other types of entities, such as long tail entities. It should be better to automatically decide the number of entity aspect of a given entity based on both the number of documents associated with the entity and the strength of the semantic connection between documents and the entity. To achieve this goal, non-parametric Bayesian methods might helpful. Thus, using non-parametric Bayesian methods might help in improving our model.

Temporal entity aspects. We view entity aspects as fixed entity specific topics in Chapter 2 and each such topic is represented by a topic distribution. However, entity aspects might evolve over time. New information might be added to an aspect, while outdated information could be removed. This is especially true when we want to mine entity aspect information from fast changing social media, such as microblog posts [105]. Existing temporal topic models [41, 132] take the temporal aspect into consideration. It should be useful to learn from existing temporal topic models and improve our model when working with social media data.

Entity document associations. We assume *entity-document associations* in Chapter 2. Specifically, given a document, we assume that several entities are semantically associated to the document and particular aspects of these entities contribute to the generation of the document. However, this entity-document association might not always exist. For example, given an entity in a document, how can we tell whether the entity is semantically associated with the document? Entity salience detection might be helpful in this case. For example, we might identify salient entities of documents using entity salience detection approaches and assume the entity document association between a document and the salient entities of the document. One research direction is to apply entity salience detection techniques to identify entity-document associations so as to promote the learning of entity-centric document representations.

Knowledge base population. We work on entity aspect linking in Chapter 3, which matches between the query context of an entity mention and candidate aspects of the referent entity. If we can effectively establish such links, it should be helpful for knowledge base population. In particular, if an entity in an article is linked to its aspect in a knowledge base, the context in which the entity appears might contain information related to the aspect. Existing work has explored the direction of recommending news articles for entities [30]. With entity aspect links, we can achieve a similar yet more fine-grained goal: recommending news articles for entity aspects. This should be especially helpful if we can measure the information overlap between information in existing knowledge base and in documents, so that we can better discover documents that are able to provide new and useful information for enriching knowledge bases.

Entity salience weights. In Chapter 4, we assume the availability of binary entity salience information. We consider salient entities as the starting point for generating documents. As for salient entities themselves, we assume equal importance. However, it is likely that some salient entities are more important than some other salient entities. It should improve SETM if we can add weights to salient entities.

Entity salience detection. We constructed a dataset for entity salience related tasks in Chapter 5. One future direction is to run and compare state-of-the-art algorithms on entity salience detection and develop new algorithms. For example, we can use information in Wikipedia to identify matching signals for entity salience detection. Given an entity in a WikiNews article, if one aspect (section of Wikipedia page, as assumed in datasets used in Chapter 3) of the entity is semantically highly related to the article, we should have a higher confidence that the entity is a salient entity in the article. Such signals might lead to effective approaches for entity salience detection.

Bibliography

[1] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval 2012*, pages 385–393, 2012. (Cited on page 41.)

[2] S. Al-Bukhitan, T. Helmy, and M. Al-Mulhem. Semantic annotation tool for annotating arabic web documents. *Procedia Computer Science*, 32:429 – 436, 2014. (Cited on page 4.)

[3] M. H. Alam, W.-J. Ryu, and S. Lee. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223, 2016. (Cited on page 15.)

[4] N. Aletras and A. Mittal. Labeling topics with images using a neural network. In *ECIR 2017*, pages 500–505. Springer, 2017. (Cited on page 55.)

[5] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, pages 25–32. Association for Computing Machinery (ACM), 2009. (Cited on page 54.)

[6] M. Bada and L. Hunter. Desiderata for ontologies to be used in semantic annotation of biomedical documents. *Journal of Biomedical Informatics*, 44(1):94 – 101, 2011. (Cited on page 4.)

[7] K. Balog. *Entity-Oriented Search*. Springer, 2018. (Cited on pages 15 and 51.)

[8] H. Bast, B. Buchhold, and E. Haussmann. Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval*, 10(2–3):119–271, 2016. (Cited on page 15.)

[9] P. Bicalho, M. Pita, G. Pedrosa, A. Lacerda, and G. L. Pappa. A general framework to expand short text for topic modeling. *Information Sciences*, 393:66–81, 2017. (Cited on pages 14 and 52.)

[10] C. Bielza, G. Li, and P. Larranaga. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011. (Cited on page 27.)

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. (Cited on pages 2, 4, 9, 13, 14, 28, 29, 52, 55, 63, and 65.)

[12] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and A. Vakali. Persador: personalized social document representation for improving web search. *Information Sciences*, 369:614–633, 2016. (Cited on page 13.)

[13] P. Bruza and T. W. Huibers. A study of aboutness in information retrieval. *Artificial Intelligence Review*, 10(5-6):381–407, 1996. (Cited on page 2.)

[14] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *KDD*, pages 169–178. ACM, 2009. (Cited on page 14.)

[15] H. Chen, F. X. Han, D. Niu, D. Liu, K. Lai, C. Wu, and Y. Xu. Mix: Multi-channel information crossing for text matching. In *KDD*, pages 110–119, 2018. (Cited on page 43.)

[16] M. Chen. Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*, 2017. (Cited on page 29.)

[17] G. Cormode and B. Krishnamurthy. Key differences between web 1.0 and web 2.0. *First Monday*, 2008. (Cited on page 1.)

[18] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM*, pages 87–94. ACM, 2008. (Cited on page 76.)

[19] W. B. Croft. Document representation in probabilistic models of information retrieval. *J. American Society for Information Science*, 32(6):451–457, 1981. (Cited on page 9.)

[20] H. Dai, S. Tang, F. Wu, and Y. Zhuang. Entity mention aware document representation. *Information Sciences*, 430:216–227, 2018. (Cited on page 13.)

[21] Z. Dai, C. Xiong, J. Callan, and Z. Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *WSDM*, pages 126–134. ACM, 2018. (Cited on pages 43 and 46.)

[22] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *SIGIR*, pages 365–374. ACM, 2014. (Cited on pages 73, 85, and 86.)

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. (Cited on pages 2, 13, 28, and 29.)

[24] M. Dojchinovski, D. Reddy, T. Kliegr, T. Vitvar, and H. Sack. Crowdsourced corpus with entity salience annotations. In *LREC*, Paris, France, 2016. European Language Resources Association (ELRA). (Cited on pages 5 and 76.)

[25] M. Dragoni, M. Federici, and A. Rexha. An unsupervised aspect extraction strategy for monitoring real-time reviews stream. *Information Processing & Management*, 2018. (Cited on page 15.)

[26] J. Dunietz and D. Gillick. A new entity salience task with millions of training examples. In *EACL*, volume 14, page 205, 2014. (Cited on pages 2, 12, 25, 61, 64, 65, 68, 70, 73, 74, 75, 76, and 85.)

[27] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227, 2004. (Cited on pages 11,

14, 28, 29, 51, 53, and 65.)
[28] L. Escoter, L. Pivovarova, M. Du, A. Katinskaia, and R. Yangarber. Grouping business news stories based on salience of named entities. In *EACL*. ACL, 2017. (Cited on pages 52 and 53.)
[29] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and G. Muhammad. Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media. *IEEE Transactions on Multimedia*, 17(12): 2281–2296, 2015. (Cited on page 39.)
[30] B. Fetahu, K. Markert, and A. Anand. Automated news suggestions for populating Wikipedia entity pages. In *CIKM*, pages 323–332. ACM, 2015. (Cited on pages 11, 16, 73, and 93.)
[31] E. Gabrilovich, M. Ringgaard, and A. Subramanya. FACC1: Freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). http://lemurproject.org/clueweb09/FACC1/, 2013. (Cited on page 73.)
[32] M. Gamon, T. Yano, X. Song, J. Apacible, and P. Pantel. Identifying salient entities in web pages. In *CIKM*, pages 2375–80. Association for Computing Machinery (ACM), 2013. (Cited on pages 2, 51, 52, 70, 73, 74, 75, and 82.)
[33] M. Gamon, T. Yano, X. Song, J. Apacible, and P. Pantel. Understanding document aboutness-step one: Identifying salient entities. *Microsoft Research*, 2013. (Cited on page 2.)
[34] D. Griffiths and M. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *NIPS*, 16:17, 2004. (Cited on page 15.)
[35] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004. (Cited on page 55.)
[36] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A deep relevance matching model for ad-hoc retrieval. In *CIKM*, pages 55–64. ACM, 2016. (Cited on page 41.)
[37] J. Guo, Y. Fan, X. Ji, and X. Cheng. MatchZoo: A learning, practicing, and developing system for neural text matching. In *SIGIR*, pages 1297–1300. ACM, 2019. (Cited on page 47.)
[38] X. Han and L. Sun. An entity-topic model for entity linking. In *EMNLP*, pages 105–115. ACL, 2012. (Cited on pages 14 and 54.)
[39] K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey of the state of the art. In *ACL (1)*, pages 1262–1273, 2014. (Cited on page 1.)
[40] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999. (Cited on page 2.)
[41] L. Hong, D. Yin, J. Guo, and B. D. Davison. Tracking trends: incorporating term volume into temporal topic models. In *SIGKDD*, pages 484–492, 2011. (Cited on page 92.)
[42] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *NIPS*, pages 2042–2050, 2014. (Cited on page 41.)
[43] L. Hu, J. Li, J. Zhang, and C. Shao. o-HETM: an online hierarchical entity topic model for news streams. In *PAKDD*, pages 696–707. Springer, 2015. (Cited on pages 14 and 65.)
[44] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised graph-based topic labelling using dbpedia. In *WSDM*, pages 465–474. Association for Computing Machinery (ACM), 2013. (Cited on page 55.)
[45] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain bias in web search. In *WSDM*, pages 413–422. ACM, 2012. (Cited on page 76.)
[46] Y.-S. Jeong and H.-J. Choi. Sequential entity group topic model for getting topic flows of entity groups within one document. In *PAKDD*, pages 366–78. Springer, 2012. (Cited on page 65.)
[47] Z. Ji, F. Xu, B. Wang, and B. He. Question-answer topic model for question retrieval in community question answering. In *CIKM*, pages 2471–2474. Association for Computing Machinery (ACM), 2012. (Cited on page 54.)
[48] I. Katakis, G. Tsoumakas, and I. Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, pages 75–83, 2008. (Cited on page 35.)
[49] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *SIGKDD*, pages 1037–1045. Association for Computing Machinery (ACM), 2011. (Cited on page 54.)
[50] G. Kazai, I. Yusof, and D. Clarke. Personalised news and blog recommendations based on user location, facebook and twitter user profiling. In *SIGIR*, pages 1129–1132. ACM, 2016. (Cited on page 12.)
[51] H. Kim, Y. Sun, J. Hockenmaier, and J. Han. Etm: Entity topic models for mining documents associated with entities. In *ICDM*, pages 349–358. IEEE, 2012. (Cited on pages 14, 26, 28, 29, 54, and 65.)
[52] N. Kiyavitskaya, N. Zeni, J. R. Cordy, L. Mich, and J. Mylopoulos. Cerno: Light-weight tool support for semantic annotation of textual documents. *Data & Knowledge Engineering*, 68(12):1470–1492,

2009. (Cited on page 4.)
[53] P. Kraft, H. Jain, and A. M. Rush. An embedding model for predicting roll-call votes. In *EMNLP*, pages 2066–2070, 2016. (Cited on page 28.)
[54] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of Wikipedia entities in web text. In *CIKM*, pages 457–66. Association for Computing Machinery (ACM), 2009. (Cited on page 62.)
[55] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *ACL*, pages 1536–1545. Association for Computational Linguistics (ACL), 2011. (Cited on page 55.)
[56] A. Lauscher, F. Nanni, P. Ruiz Fabo, and S. P. Ponzetto. Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability. *IJCol-Italian journal of computational linguistics*, 2(2):67–88, 2016. (Cited on page 55.)
[57] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014. (Cited on pages 2, 9, 13, 28, and 29.)
[58] M. Levit, S. Parthasarathy, S. Chang, A. Stolcke, and B. Dumoulin. Word-phrase-entity language models: Getting more mileage out of n-grams. In *INTERSPEECH*, 2014. (Cited on page 51.)
[59] C. Li, J. Xing, A. Sun, and Z. Ma. Effective document labeling with very few seed words: A topic model approach. In *CIKM*, pages 85–94. ACM, 2016. (Cited on page 14.)
[60] P. Li, Y. Wang, W. Gao, and J. Jiang. Generating aspect-oriented multi-document summarization with event-aspect model. In *EMNLP*, pages 1137–1146. ACL, 2011. (Cited on page 15.)
[61] P. Li, Y. Wang, and J. Jiang. Automatically building templates for entity summary construction. *Information Processing & Management*, 49(1):330–340, 2013. (Cited on page 16.)
[62] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW*, pages 675–684. ACM, 2008. (Cited on page 35.)
[63] X. Li, J. Ouyang, and X. Zhou. Centroid prior topic model for multi-label classification. *Pattern Recognition Letters*, 62:8–13, 2015. (Cited on pages 15 and 52.)
[64] X. Li, J. Ouyang, and X. Zhou. Supervised topic models for multi-label classification. *Neurocomputing*, 149:811–819, 2015. (Cited on pages 15 and 52.)
[65] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, and J. Ouyang. Filtering out the noise in short text topic modeling. *Information Sciences*, 456:83–96, 2018. (Cited on pages 14 and 52.)
[66] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015. (Cited on page 39.)
[67] M. Liu, Y. Fang, A. G. Choulos, D. H. Park, and X. Hu. Product review summarization through question retrieval and diversification. *Information Retrieval Journal*, 20(6):575–605, 2017. (Cited on page 15.)
[68] B. Ma, N. Zhang, G. Liu, L. Li, and H. Yuan. Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management*, 52(3):430–445, 2016. (Cited on page 14.)
[69] J.-E. Mai. Semiotics and indexing: an analysis of the subject indexing process. *Journal of documentation*, 57(5):591–622, 2001. (Cited on page 1.)
[70] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *NIPS*, pages 121–128. Curran Associates, Inc., 2008. (Cited on page 15.)
[71] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI Workshop on Text Learning*, pages 1–7. AAAI, 1999. (Cited on pages 14 and 53.)
[72] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, page 33, 2005. (Cited on page 53.)
[73] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119. Curran Associates Inc., 2013. (Cited on page 13.)
[74] B. Mitra, F. Diaz, and N. Craswell. Learning to match using local and distributed representations of text for web search. In *WWW*, pages 1291–1299. WWW, 2017. (Cited on page 43.)
[75] F. Moscato, B. D. Martino, S. Venticinque, and A. Martone. Overfa: A collaborative framework for the semantic annotation of documents and websites. *Int. J. Web Grid Serv.*, 5(1):3045, 2009. (Cited on page 4.)
[76] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. (Cited on page 2.)
[77] F. Nanni, S. P. Ponzetto, and L. Dietz. Entity-aspect linking: Providing fine-grained semantics of entities in context. In *JCDL*, pages 49–58. ACM, 2018. (Cited on pages 3, 16, 30, 39, 40, 43, 45,

and 46.)
[78] F. Nanni, J. Zhang, B. Ferdinand, and K. Gashteovski. Eal: A toolkit and dataset for entity-aspect linking. In *JCDL*. ACM, 2019. (Cited on page 45.)
[79] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD*, pages 680–686. ACM, 2006. (Cited on pages 14, 54, 58, and 65.)
[80] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. Text matching as image recognition. In *AAAI*, pages 2793–2799, 2016. (Cited on pages 41, 42, and 43.)
[81] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *CIKM*, pages 365–74. Association for Computing Machinery (ACM), 2009. (Cited on page 2.)
[82] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. (Cited on pages 28 and 46.)
[83] A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett. Hierarchically supervised latent dirichlet allocation. In *NIPS*, pages 2609–2617, 2011. (Cited on page 15.)
[84] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. (Cited on page 13.)
[85] T. Piccardi, M. Catasta, L. Zia, and R. West. Structuring Wikipedia articles with section recommendations. In *SIGIR*, pages 665–674. ACM, 2018. (Cited on page 30.)
[86] M. Ponza, P. Ferragina, and F. Piccinno. Swat: A system for detecting salient Wikipedia entities in texts. *Computational Intelligence*, 35(4):858–890, 2019. (Cited on pages 63 and 73.)
[87] Z. Qiu and H. Shen. User clustering in a dynamic social network topic model for short text streams. *Information Sciences*, 414:102–116, 2017. (Cited on pages 14 and 52.)
[88] P. Radhakrishnan, G. Jawahar, M. Gupta, and V. Varma. Sneit: Salient named entity identification in tweets. *Computación y Sistemas*, 21(4), 2017. (Cited on page 2.)
[89] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. ACL, 2009. (Cited on pages 15, 28, 29, and 56.)
[90] J. Rao and C. Wu. Pseudo–empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105(492):1494–1503, 2010. (Cited on page 27.)
[91] H. Raviv, O. Kurland, and D. Carmel. Document retrieval using entity-based language models. In *SIGIR*, pages 65–74. ACM, 2016. (Cited on pages 4, 13, and 73.)
[92] R. Reinanda, E. Meij, and M. de Rijke. Mining, ranking and recommending entity aspects. In *SIGIR*, pages 263–272. ACM, 2015. (Cited on pages 16 and 39.)
[93] R. Reinanda, E. Meij, and M. de Rijke. Document filtering for long-tail entities. In *CIKM*, pages 771–780. ACM, 2016. (Cited on page 15.)
[94] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. N^3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *LREC 2014*, pages 3529–33. European Language Resources Association (ELRA), 2014. (Cited on page 76.)
[95] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494. AUAI Press, 2004. (Cited on pages 14, 51, and 53.)
[96] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012. (Cited on page 54.)
[97] G. Salton, C.-S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *J. American society for Information Science*, 26(1):33–44, 1975. (Cited on pages 4, 28, and 29.)
[98] E. Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12): e26752, 2008. (Cited on pages 25 and 76.)
[99] C. d. Santos, M. Tan, B. Xiang, and B. Zhou. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*, 2016. (Cited on page 44.)
[100] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1–47, 2002. (Cited on page 9.)
[101] W. Shen, J. Wang, P. Luo, and M. Wang. Linking named entities in tweets with knowledge base via user interest modeling. In *KDD*, pages 68–76. ACM, 2013. (Cited on page 51.)
[102] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015. (Cited on pages 2, 4, 39, and 73.)
[103] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *WWW*, pages 373–374. ACM, 2014. (Cited on page 41.)
[104] A. Sikchi, P. Goyal, and S. Datta. Peq: An explainable, specification-based, aspect-oriented product comparator for e-commerce. In *CIKM*, pages 2029–2032. ACM, 2016. (Cited on page 15.)

[105] D. Spina, E. Meij, M. de Rijke, A. Oghina, M. T. Bui, and M. Breuss. Identifying entity aspects in microblog posts. In *SIGIR*, pages 1089–1090. ACM, 2012. (Cited on pages 16 and 92.)
[106] Y. Tay, A. T. Luu, and S. C. Hui. Hermitian co-attention networks for text matching in asymmetrical domains. In *IJCAI*, pages 4425–4431, 2018. (Cited on page 41.)
[107] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120. ACM, 2008. (Cited on page 15.)
[108] T. A. Tran, C. Niederée, N. Kanhabua, U. Gadiraju, and A. Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *CIKM*, pages 1201–1210. ACM, 2015. (Cited on pages 52, 53, and 73.)
[109] S. Trani, D. Ceccarelli, C. Lucchese, S. Orlando, and R. Perego. Sel: A unified algorithm for entity linking and saliency detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 85–94. ACM, 2016. (Cited on page 5.)
[110] S. Trani, C. Lucchese, R. Perego, D. E. Losada, D. Ceccarelli, and S. Orlando. SEL: A unified algorithm for salient entity linking. *Computational Intelligence*, 2018. (Cited on pages 76, 83, and 91.)
[111] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007. (Cited on pages 26 and 27.)
[112] C. Van Gysel, M. de Rijke, and E. Kanoulas. Learning latent vector spaces for product search. In *CIKM*, pages 165–174. ACM, 2016. (Cited on page 13.)
[113] C. Van Gysel, M. de Rijke, and M. Worring. Unsupervised, efficient and semantic expertise retrieval. In *WWW*, pages 1069–1079. ACM, 2016. (Cited on page 13.)
[114] C. Van Gysel, M. de Rijke, and E. Kanoulas. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems*, 36(4):Article 38, June 2018. (Cited on pages 13 and 28.)
[115] S. Wang, Z. Chen, and B. Liu. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web*, pages 167–76. International World Wide Web Conference Committee, 2016. (Cited on page 52.)
[116] C. Wu, E. Kanoulas, and M. de Rijke. It all starts with entities: A salient entity topic model. *Natural Language Engineering*, pages 1–19, 2019. (Cited on page 73.)
[117] C. Wu, E. Kanoulas, and M. de Rijke. Learning entity-centric document representations using an entity facet topic model. *Information Processing & Management*, 57(3), 2020.
[118] C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. A multi-interaction based convolutional matching network for entity aspect linking. In *COLING*, 2020.
[119] C. Wu, E. Kanoulas, M. de Rijke, and W. Lu. WN-Salience: a corpus of news articles with entity salience annotations. In *LREC 2020*, pages 1–8. LREC, 2020.
[120] D. Xiao, Y. Ji, Y. Li, F. Zhuang, and C. Shi. Coupled matrix factorization and topic modeling for aspect mining. *Information Processing & Management*, 54(6):861–873, 2018. (Cited on page 15.)
[121] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, pages 2659–2665, 2016. (Cited on page 51.)
[122] C. Xiong, J. Callan, and T.-Y. Liu. Bag-of-entities representation for ranking. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, page 181184. Association for Computing Machinery, 2016. (Cited on page 4.)
[123] C. Xiong, J. Callan, and T.-Y. Liu. Word-entity duet representations for document ranking. In *SIGIR*, pages 763–772. ACM, 2017. (Cited on page 13.)
[124] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power. End-to-end neural ad-hoc ranking with kernel pooling. In *SIGIR*, pages 55–64. ACM, 2017. (Cited on pages 41, 43, 46, and 47.)
[125] C. Xiong, Z. Liu, J. Callan, and T.-Y. Liu. Towards better text understanding and retrieval through kernel entity salience modeling. In *SIGIR*, pages 575–584. ACM, 2018. (Cited on pages 52, 53, and 73.)
[126] K. Xu, G. Qi, J. Huang, and T. Wu. Incorporating Wikipedia concepts and categories as prior knowledge into topic models. *Intelligent Data Analysis*, 21(2):443–461, 2017. (Cited on page 54.)
[127] W. Yin and H. Schütze. Convolutional neural network for paraphrase identification. In *NAACL*, pages 901–911, 2015. (Cited on page 41.)
[128] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua. Aspect ranking: Identifying important product aspects from online consumer reviews. In *HLT*, pages 1496–1505. ACL, 2011. (Cited on page 15.)
[129] Q. Yu and W. Lam. Review-aware answer prediction for product-related questions incorporating aspects. In *WSDM*, pages 691–699. ACM, 2018. (Cited on page 15.)
[130] X. Zhang, S. Li, L. Sha, and H. Wang. Attentive interactive neural networks for answer selection in community question answering. In *AAAI*, pages 3525–3531, 2017. (Cited on pages 41 and 44.)

[131] Y. Zhang, W. Mao, and D. Zeng. A non-parametric topic model for short texts incorporating word coherence knowledge. In *CIKM*, pages 2017–2020. ACM, 2016. (Cited on pages 14 and 52.)

[132] N. Zheng, S. Song, and H. Bao. A temporal-topic model for friend recommendations in Chinese microblogging systems. *IEEE transactions on systems, man, and cybernetics: systems*, 45(9):1245–1253, 2015. (Cited on page 92.)

Summary

The amount of information available for consumption has been overwhelming since the end of the 20th century, leading to information overload. Automated information processing techniques make it possible to process and organize large volumes of information. Textual document is one category of information that is widely available with entities playing a key role in automatically understanding the semantics of documents. In this thesis, we aim at enhancing document understanding by using entity aspects and entity salience information.

First, we hypothesize that entities are not monolithic concepts; instead they have multiple aspects, and different documents may be discussing different aspects of a given entity. Given that, we argue that from an entity-centric point of view, a document related to multiple entities shall be (a) represented differently for different entities (multiple entity-centric representations), and (b) each entity-centric representation should reflect the specific aspects of the entity discussed in the document. We show that entities are multi-faceted concepts which we can model and learn. We find that a multi-faceted entity-centric modeling of documents can lead to effective representations. Then we study entity aspect linking, which links text fragments (entity mentions) to particular aspects of entities. We view entity aspect linking as a pairwise semantic matching problem and propose a multi-interaction based convolutional matching network (MICMN) to solve the task. Specifically, we first construct multiple interactions for entity context and candidate aspects, including exact match, soft match, and self-attentive interactions. Then, we pass the interactions to convolutional layers to identify matching patterns and apply a novel method (q-singular pooling) to extract features. Finally, a multi-layer perceptron is used to output a matching score. We show the effectiveness of MICMN on four datasets.

Then we move to the direction of enhancing document understanding using entity salience information. We assume that in long textual documents, not all entities are equally important: some are salient and others are not. We propose a novel entity topic model, i.e., Salient Entity Topic Model (SETM), to take salient entities into consideration in the document generation process. In particular, we model salient entities as a source of topics used to generate words in documents, in addition to the topic distribution of documents used in traditional topic models. We conduct qualitative and quantitative analysis on the proposed model to show the effectiveness of SETM. Application to entity salience detection demonstrates the effectiveness of our model compared to state-of-the-art topic model baselines. Then we present a new dataset, the WikiNews Salience dataset (WN-Salience), to support research on entity salience related tasks such as entity salience detection and salient entity linking. WN-Salience is built on top of Wikinews, a Wikimedia project whose mission is to present reliable news articles. Entities in Wikinews articles are identified by the authors of the articles and are linked to Wikinews categories when they are salient or to Wikipedia pages otherwise. We compare the WN-Salience dataset against existing datasets on the task and analyze their differences. Furthermore, we conduct experiments on entity salience detection; the results demonstrate that WN-Salience is a challenging testbed that is complementary to existing ones.

Samenvatting

De hoeveelheid beschikbare informatie die sinds het eind van de 20e eeuw beschikbaar is gekomen is overweldigend. Technieken om informatie automatisch te verwerken maken het mogelijk om wijs te worden uit deze overvloed. Tekstdocumenten zijn algemeen toegankelijk. Voor het begrip van deze documenten spelen entiteiten een sleutelrol. In dit proefschrift willen we het automatisch begrijpen van documenten verbeteren door gebruik te maken van verschillende aspecten van entiteiten en informatie over hoe saillant een entiteit is.

Ten eerste veronderstellen we dat entiteiten geen monolithische concepten zijn; in plaats daarvan hebben ze meerdere aspecten, en verschillende documenten kunnen verschillende aspecten van een bepaalde entiteit belichten. Op basis hiervan stellen we dat, vanuit een oogpunt van de entiteit, een document dat betrekking heeft op meerdere entiteiten (a) verschillend zal worden weergegeven voor verschillende entiteiten (meerdere representaties van de entiteit), en (b) elke representatie van de entiteit de specifieke aspecten van de entiteit die in het document worden besproken weerspiegelt. We laten zien dat entiteiten veelzijdige concepten zijn die we kunnen modelleren en leren. We laten zien dat een veelzijdige, op entiteiten gerichte modellering van documenten kan leiden tot effectieve representaties. Vervolgens bestuderen we het koppelen van aspecten van entiteiten, waarbij tekstfragmenten (vermeldingen van entiteit) gekoppeld worden aan bepaalde aspecten van entiteiten. We modelleren het koppelen van aspecten van entiteiten als een paarsgewijs semantisch koppelingsprobleem en stellen een Multi-Interaction Based Convolutional Matching Network (MICMN) voor om de taak op te lossen. Meer specifiek construeren we eerst meerdere interacties voor de context van een entiteit en kandidaat-aspecten, waaronder exact match, soft match en self-attentive interacties. Vervolgens geven we de interacties door aan convolutional layers om overeenkomende patronen te identificeren en passen we een nieuwe methode (q-singular pooling) toe om features te extraheren. Ten slotte wordt een multi-layer perceptron gebruikt om een matching score te berekenen. We laten de effectiviteit van MICMN zien op vier datasets.

In het volgende deel van het proefschrift gaan behandelen we het verbeteren van het begrip van documenten met behulp van de saillantie-informatie van de entiteit. We gaan ervan uit dat in lange tekstdocumenten niet alle entiteiten even belangrijk zijn: sommige zijn saillant en andere niet. We stellen een nieuw entiteits-topic model voor, Salient Entity Topic Model (SETM), dat rekening houdt met saillante van entiteiten bij het genereren van documenten. In het bijzonder modelleren we saillante entiteiten als een bron van onderwerpen die worden gebruikt om woorden in documenten te genereren, naast de topic distributie van documenten die gebruikt wordt in traditionele topic models. We voeren kwalitatieve en kwantitatieve analyses uit op het voorgestelde model om de effectiviteit van SETM aan te tonen. We tonen de effectiviteit van ons model aan, vergeleken met de modernste baselines van topic models, op de taak van het detecteren van entiteit-saillantie. Vervolgens presenteren we een nieuwe dataset, de WikiNews Salience-dataset (WN-Salience), ter ondersteuning van onderzoek naar taken gerelateerd aan entiteits-saillantie zoals het detecteren van de saillantie van entiteiten en het koppelen van saillante entiteiten. WN-Salience is gebouwd op basis van Wikinews, een Wikimedia-project dat betrouwbare nieuwsartikelen presenteert.

Entiteiten in Wikinews-artikelen worden gemarkeerd door de auteurs van de artikelen en zijn gekoppeld aan Wikinews-categorien als ze saillant zijn, of ze worden gelinkt aan Wikipedia pagina's. We vergelijken de WN-Salience-dataset met bestaande datasets voor deze taak en analyseren hun verschillen. Verder voeren we experimenten uit met het detecteren van de saillantie van entiteiten; de resultaten tonen aan dat WN-Salience een uitdagende dataset, complementair is aan de bestaande datasets.