Time-aware Topic-based Contextualization

Nam Khanh Tran
supervised by Prof. Wolfgang Nejdl
and Dr. Claudia Niederée
L3S Research Center & University of Hannover
Appelstrasse 9a, 30167 Hannover, Germany
ntran@L3S.de

ABSTRACT

In the past, various studies have been proposed to acquire the capacity to perceive and comprehend language in articles or human communications. Recently, researchers focus on higher semantic levels to what human would need to understand the contents of articles. While human can smoothly interpret documents when they have knowledge of the context of documents, they have difficulty with those as their context is lost or changes. In this PhD proposal, we address three novel research questions: detecting uninterpretable pieces in documents, retrieving contextual information and constructing compact context for the documents, then propose approaches to these tasks, and discuss related issues.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Temporal Semantic Linking, Contextualization, Document Modeling, Document Understanding

1. INTRODUCTION

"What is Camel? What are the reasons behind the fact that Doctors did smoke Camels more than others? Why was the article accepted to publish?" These are some examples of typical questions that are likely to be asked by people when reading the article "More Doctors smoke Camels than any other cigarette" (see Figure 1) published in the 1950s. Providing the readers information to answer these questions will help them to obtain better understanding of the article. Unfortunately, documents are typically created with some context behind that is assumed to be known by creators, consequently the readers who are not aware these context will find difficult to understand the documents. When such context is provided, in other words the creators and readers have common knowledge, the readers would be able to form correct and coherent interpretations of the documents.

Example 1. Related to the article "More Doctors smoke Camels than any other cigarette", if we look back at the old commercials, it is extremely ironic that cigarette companies' key spokesmen were Doctors that you would never see today. In addition, in the 1950s, although there was evidence being gathered about the ill effects of smoking these are widely accepted and popular. Evidently, if we are aware this fact when reading the article, we can understand it better.

Example 2. Twitter provides its users facilities to share short text messages, comprising a maximum of 140 characters. The property usually make the tweets difficult to fully interpret. For example, considering the tweet "36 years ago this week, one of the greatest soccer players ever made his debut", even though we know it is about a soccer player, we still can not understand it. But with acquiring that the player mentioned in the tweet is Maradona and he made his great international debut for Argentina 36 years ago, we can perceive the tweet without any problem.

In this research proposal, we aim at acquiring such context to help humans in understanding the contents of documents. By context, we refer to any information *beyond documents* that help to interpret them. Context can be texts or images that explain either difficult items or topics in documents (see Figure 1). We call them *linking context* and *topical context*, respectively. There have been extensive studies on linking texts to concepts in Wikipedia, a collection of human common knowledge [24, 25, 21, 1] (e.g. (1) in Figure 1). Most, if not all, of the linking methods attempt to seek Wikipedia articles that explain specific wording in documents, little tackles the problem of retrieving topical context for topics discussed in documents. In addition, whilst current approaches disregard temporal and relational dimensions of context, we consider them as crucial components. For instance, in the example 1, the word "*Camel*" and the main topic "*Doctor smoke Camels*" must be considered in the 1950s as they might not exist today anymore or express another meaning, and the related topics like "*Camels advertisement industry*", "*Doctor spokenman cigarette*" have to be considered because they are factors to understand the main topic.

Given the above examples and motivations, we address the following research questions: (1) How to identify the parts of documents that need additional context to be interpreted; (2) How to obtain context with taking into account temporal and relational dimensions; (3) How to convey the context to humans, either making links to the external link-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2567957.

Camel is a brand of cigarettes that was introduced by American company R.J. Reynolds Tobacco in the summer of 1913. (1)

Prior to 1964, many of the cigarette companies advertised their brand by falsely claiming that their product did not have serious health risks. A couple of examples would be "More doctors smoke Camels". Such claims were made both to increase the sales of their product and to combat the increasing public knowledge of smoking's negative health effects (2)

Figure 1: Example of an article published in the 1950s and its context

ing data source (e.g. Wikipedia) as traditional approaches or condensing all context into a cohesive summary.

The main contributions of my Ph.D. thesis will be as follows: (i) improve state of the art in semantic linking by taking temporal and relational dimensions into account; (ii) a general contextualizing framework will be created, which allows the detecting, retrieving and summarizing of context of documents; (iii) novel algorithms for solving each research questions.

2. STATE OF THE ART

The targeted research goal relates to several areas namely semantic linking, topic modeling, temporal information retrieval and multi-document summarization. In this section, we will give an overview over the most recent work in each area and discuss their limits in our task.

2.1 Semantic Linking

Links to a knowledge structure are considered as a natural way of adding semantics to digital items which has received considerable attention from research community [24, 25, 21, 1]. The Wikify! system [24], for example, detects potential anchor texts from a given document based on term statistics derived from Wikipedia links. Further, knowledge-based and machine learning-based approaches are exploited for identifying the corresponding Wikipedia concepts. In the same vein, Milne and Witten [25] solved the problem with machine learning techniques in which contextual information in the source text was used to detect the best related Wikipedia concepts, which in turn served as features for anchor text detection. The method yields more accurate results and greatly improves the performance in terms of precision and recall over [24].

There have been also some studies for adding semantics to microblog posts. Whilst Abel et al. [1] analysed methods for mapping Twitter posts with related news articles in order to contextualize Twitter activities, Meij et al. [21] added semantics to tweets by identifying related Wikipedia concepts that are semantically related to it and generating links to the corresponding Wikipedia articles. Apart from other work, Bron et al. [8] attempted to link items with a rich textual representation in a news archive to items with sparse annotations in a multimedia archive if they describe the same or related event.

These studies focus on seeking linking context for utterances in documents while we aim at higher semantic levels (e.g. topics discussed in documents). In addition, most of the studies ignore the temporal and relational dimensions of context whilst we consider them as crucial components.

2.2 Topic Modeling

In the PhD work, we will work on both linking and topical context but focus more on the latter that helps readers to understand topics discussed in documents. Hence, in this section we will review some related work in the topic modeling community. In recent years, topic modeling is an area in machine learning that discovers the latent "topic" (represented by a group of works - textual or visual) implied by a collection of documents. The two most popular techniques are probabilistic Latent Semantic Analysis (pLSA) [15] and latent Dirichlet Allocation (LDA) [7]. Both consider documents as a mixture of topics, and use topic distribution to characterize the document-topic relationships, in which LDA is claimed to outperform pLSA in generalization ability. My PhD work will follow the generative idea of LDA, but will have to tackle its several limits. First, LDA is not able to model topic correlations, making the model unrobust in the contextualization task. This issue was first investigated by Blei et al. 2006 [6] where the authors model topic correlations via the logistic normal distribution. In addition, LDA takes no notice of temporal information which was later considered by [3, 32]. In this work, we aim at taking into account both temporal and relational information.

2.3 Temporal Information Retrieval

As mentioned in Section 1, one of our tasks is to retrieve context for some parts of documents with time-aware consideration. There is a bunch of studies on this area recently [5, 20, 22, 18]. Li and Croft [20] experimented with time-based language models by assigning a document prior using an exponential decay function of its creation date. Berberich et al. [5] integrated temporal expressions into query-likelihood language modeling, which considers uncertainty inherent to temporal expressions in a query and in documents. Kanhabua et al. [18] is technically the closet to our work. They attempted to retrieve and rank sentences that relates to future events whilst our aim is to find information to help users to interpret documents better.

2.4 Multi-Document Summarization

We aim at constructing a concise context for a given document from the context that are retrieved for each part of the document. In order to do that, multi-document summarization techniques have to be studied. As one of the most popular extractive systems, the centroid-based multi-document summarizer (MEAD) [27] generates summaries by using information from a set of words that are statistically important to a cluster of documents for selecting sentences. Erkan and Padev [11] presented LexRank approach to rank sentences by weighting each vote so that the vote coming from a more prestigious sentence has a greater value in the centrality of a sentence. There is also a couple of studies on supervised learning, e.g. [26, 23]. The work proposed by Štajner et al. [31] is the closet to what we want to obtain where they formulate the summarization problem as an optimization problem.

3. PROPOSED RESEARCH

Since Schilit et al. [28] introduced the term *context-aware computing*, some definitions of the term *context* has been proposed [34, 13] but for the use of interpretation, there is no agreed upon definition. Hence, in this research, we propose the definition of context as follows: *Context of a document is any information beyond the document that help to interpret the document.* We now discuss in details the three research tasks in Section 1. For each task, we identify the challenges and propose possible methods based on current state of the art, as well as elicit their potential issues.

3.1 Contextual Document Modeling

In this work, we will consider both linking and topical context but focus more on the latter one. For the linking context we need to identify words or phrases to link them to external sources (e.g. Wikipedia). Seeking topical context is a more challenging problem because we first have to detect themes or topics discussed in documents. To handle this issue, we can represent documents as a bag of words [14], a mixture of topics [7] or through their readers [10]. The challenges here are how to integrate temporal and relational information into each representation.

For the topic case, following the idea of LDA, we will model documents as a mixture of topics. In contrast to previous work, in our model, each topic has two attributes: relations with other topics and time value. For detecting topic relations, one of the baseline methods is to follow the correlated topic model proposed by Blei et al. [6], where the authors model topic correlations via logistic normal distribution. In addition, we consider two methods for estimating the probability of a topic given a time value $P(t|y)$. The first method is to follow the previous work [32] where each topic is associated with a continuous distribution over timestamps and the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. The second method is to do post-processing for that estimation

$$P(t|y) = \frac{1}{|W_t|} \sum_{w \in W_t} P(w|y)$$

where W_t is the set of words representing the topic t, $P(w|y)$ is estimated by counting how many times the word occurs in the sentences that mention time y. To identify both relational and temporal information for topics, we plan to use correlated topic model [6] to detect topic correlations, then perform post-processing to estimate temporal probability.

3.2 Temporal Semantic Linking

After identifying the parts of documents (e.g. topics), we tackle the problem of retrieving context from the linking data source based on their complementary relations to one or several topics of interest. We identify the following challenges for this task: 1) how to formalize the semantic of "complementary relation"; and 2) how to integrate the temporal information into the ranking model.

Humans can perceptually recognize the pieces of information that appears complementary to each other. But unlike the pure relations, similarity and contrast, complementarity is rather broad and subjective in a sense that it is something in-between and becomes kind of imprecise. Thus, it would be difficult to define and measure accurately. The first study that attempts to tackle this problem was proposed by Gao et al. [12]. Given two pieces of information p_i and p_j, the complementarity measure is defined as:

$$I_{comp} = \begin{cases} \frac{I_{comm}}{I_{diff}}, & if\ I_{comm} < I_{diff}; \\ \frac{I_{diff}}{I_{comm}}, & otherwise \end{cases}$$

where I_{comm} and I_{diff} are the strength of their commonality and difference of p_i and p_j, respectively. The tricky point here is how to integrate the temporal information into the complementarity measure either as a weighting score or as an independent measure.

We plan to formalize the task as the problem of temporal information retrieval. Inspired by previous work [30, 18], we propose using a learning to rank approach based on topic-based similarities and the above complementarity measure. To incorporate the temporal feature in ranking, we will first employ two features proposed in previous studies (TSU [19] and FS [17]). We then exploit some learned ranking algorithms such as RankSVM [16], SGD-SVM [33] and PA-Perceptron [9] for learning the ranking model.

3.3 Contextual Summarization

My PhD work aims at constructing a compact context for a given document from the retrieved context of each topic in the document.

Problem statement. Given a document d represented by a mixture of topics T, each topic t in T has a number of context C_t, we seek a subset $C \subseteq \cap C_t$ of context which are most informative and cohesive.

Proposed approach. Since each topic in our model has relational information, we first want to follow LexRank approach [11], the current graph-based model on multi-document summarization. The approach rank sentences by weighting each vote so that the vote coming from a more prestigious sentence has a greater value in the centrality of a sentence. Alternatively, in spired by the previous work [31], we can also see the problem of context selection as an optimization problem. An utility function can be computed for the context at topic-level and document-level. The solution to the selection problem is then to find a subset C^* that maximizes the objective function $g(C)$

$$g(C) = \lambda \sum_{c \in C_t} r(c) + (1 - \lambda) H(C)$$

where $r(c)$ represents the utility score of context c at topic-level and $H(C)$ is that in document-level. The issues here are to find good indicators and a efficient approximation algorithm to estimate the utility functions.

4. METHODOLOGY

In this PhD work, we will solve the three research questions stated previously, and will systematically evaluate the proposed approaches. This section outlines the design principles of my work. It describes datasets selected for the experiments and discusses evaluation methodology.

Datasets. My PhD work will use two types of datasets: *primary sources* and *linking data sources*. The primary sources are temporal document collections, in our case, New York Times Corpus and Twitter based corpora (TREC Tweet2011 or crawled tweet), which need to be contextualized. Flickr images is another interesting dataset which can be used to study for the multimedia case. For the linking data sources, we currently refer to different versions of the Wikipedia dumps, collections of human common knowledge.

Preliminary experiments. We first want to conduct some preliminary experiments to evaluate performance of the current approaches. In the first experiment, we want to see how well Wikify! [24] and Wikipedia Miner systems [25] can link NYT articles to Wikipedia. Then, we will integrate temporal information into these systems, for example by attaching time values into each detected anchor texts in documents. In addition, we plan to ask users for what they often find difficult to understand a document, for example, either difficult terms or topics that will guide us how to model documents.

Evaluation. Evaluating contextualization systems is a extremely challenging problem. To the best of our knowledge, there is no *standard test set* for the assessment although there are some manual valuable datasets for microblog posts, e.g. [21]. Hence, we propose creating a evaluation dataset manually. For example, we plan to give documents to annotators for selecting the parts that they find difficult to understand. The parts here can be terms, sentences or a set of words representing topics, called anchors. For each anchor, the annotators use Wikipedia's search engine to find the most appropriate sentences or paragraphs that help them to understand the anchor. Evidently, each anchor can be assigned to several Wikipedia elements since each annotator has different background knowledge. To evaluate the performances of systems, we can use the traditional measures (*precision (P)*, *recall (R)* and *F-measure (F)*). The tricky point here is how to determine whether the context detected by the systems are relevant to those annotated by the annotators, either on surface level or semantic level.

5. PRELIMINARY RESULT

This PhD work is still in early stage, hence, in this section we present some early preliminary results.

5.1 Topic Cropping

In this work, we address the problem of characterizing documents in small corpora with topic models [29]. A "topic" consists of a cluster of words that frequently occur together. As the limited size of the corpora leads to poor quality topic models that make human difficult to interpret, higher quality topic models can be learned by incorporating additional domain-specific documents with similar topical content. This, however, requires finding or even manually composing such corpora, requiring considerable effort. For solving this problem, we developed a fully automated adaptable process of *topic cropping*:

- Analyzing corpus coverage by selecting characteristic terms which reflect the contents of documents. Starting from the original documents and a random subset of pages selected from Wikipedia, in order to do that we used the metric of Mutual Information.
- Tailoring a cropping corpus by collecting relevant documents. We used a general Web search engine to identify the set of highest ranked Wikipedia pages for each of the representative terms.
- Learning a topic model from the cropping corpus using latent Dirichlet allocation (LDA [7]). Then, applying topic inference to the original corpus.

We judge the quality of the automatically detected topics by measuring *topic diversity*, *topic coherence* and *topic relevance*. Our experiments showed substantial improvements in diversity as well as in internal coherence of inferred topics compared to a naive approach using the limited size corpora exclusively (more details refers to [29]). In this way documents are characterized by topics learned from the external source, which in turn can provide context to understand those documents. By either choosing different linking sources, different periods or including time values to each representative terms, we can take the temporal aspect into account and learn more useful topics depending on the time of source documents.

5.2 Tweet Contextualization

This section describes our initial work [2] for the Tweet Contextualization track at INEX 2013.[1] Given a new tweet and a recent dump of the Wikipedia, the system is required to provide some context about the subject of the tweet in order to help the reader to understand it.

Our preliminary results. We conducted a pipeline system based on the following process: tweet analysis, context retrieval and construction of the answer. Firstly, we detect and extract key phrases that are more informative than the others in the tweet. We used ArkTweet toolkit[2] to tokenize the tweet content and annotate each token with an adjusted part-of-speech tags. After that, we employed several heuristics to detect the key phrases as overlapping consecutive tokens. These phrases are then posed as queries to the index of the Wikipedia powered by Indri[3]. We made use of MEAD toolkit[4] to construct the answer from the retrieved passages. The system was evaluated based on informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is. Our initial approach did not obtain a very good performance (see [4]) which motivates our proposal that considers higher semantic levels (topics, events) and temporal information.

[1] https://inex.mmci.uni-saarland.de/tracks/qa/
[2] http://www.ark.cs.cmu.edu/TweetNLP/
[3] http://www.lemurproject.org/
[4] http://www.summarization.com/mead/

6. CONCLUSION

In this PhD proposal, we address the problem of acquiring context to fully interpret document contents. Semantic linking is an active research area over the past decade, and we conceive the introduction of temporal aspect as an important extension. We also discuss the three novel issues on contextual modeling, temporal semantic linking and summarization and propose solutions to these tasks. We believe the thesis outcome can benefit the research in related areas.

7. ACKNOWLEDGEMENTS

The work was supported by the project *"Gute Arbeit" nach dem Boom (Re-SozIT)* (01UG1249C) funded by the German Federal Ministry of Education and Research (BMBF) and by the European project ForgetIT (GA600826).

8. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of ESWC '2011*, 2011.

[2] K. Ansary, A. T. Tran, and N. K. Tran. A pipeline tweet contextualization system at inex 2013. Technical report, CLEF 2013 Evaluation Labs and Workshop, 2013.

[3] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of CIKM '2011*, 2011.

[4] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, and X. Tannier. Overview of inex tweet contextualization 2013 track. Technical report, CLEF, 2013.

[5] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *Proceedings of ECIR '2010*, 2010.

[6] D. Blei and J. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 2006.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003.

[8] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *Proceedings of TPDL '2011*, 2011.

[9] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 2006.

[10] K. El-Arini, M. Xu, E. B. Fox, and C. Guestrin. Representing documents through their readers. In *Proceedings of KDD '2013*, 2013.

[11] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 2004.

[12] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of CIKM '2012*, 2012.

[13] T. Gross and M. Specht. Awareness in context-aware information systems. In *Mensch and Computer*, 2001.

[14] Z. Harris. Distributional structure. *Structural and Transformational Linguistics*, 1981.

[15] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR '1999*, 1999.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD '2002*, 2002.

[17] P. J. Kalczynski and A. Chou. Temporal document retrieval model for business news archives. *Information Processing and Management*, 2005.

[18] N. Kanhabua, R. Blanco, and M. Matthews. Ranking related news predictions. In *Proceedings of SIGIR '2011*, 2011.

[19] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *Proceedings of ECDL '2010*, 2010.

[20] X. Li and W. B. Croft. Time-based language models. In *Proceedings of CIKM '2003*, 2003.

[21] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of WSDM '2012*, 2012.

[22] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In *Proceedings of SIGIR '2009*, 2009.

[23] D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *Proceedings of SIGIR Learning to Rank Workshop*, 2008.

[24] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of CIKM '2007*, 2007.

[25] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of CIKM '2008*, 2008.

[26] Y. Ouyang, S. Li, and W. Li. Developing learning strategies for topic-based summarization. In *Proceedings of CIKM '2007*, 2007.

[27] D. R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 2004.

[28] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, 1994.

[29] N. K. Tran, S. Zerr, K. Bischoff, C. Niederée, and R. Krestel. Topic cropping: Leveraging latent topics for the analysis of small corpora. In *Proceedings of TPDL '2013*, 2013.

[30] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *Proceedings of WSDM '2011*, 2011.

[31] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes. Automatic selection of social media responses to news. In *Proceedings of KDD '2013*, 2013.

[32] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of KDD '2006*, 2006.

[33] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of ICML '2004*, 2004.

[34] A. Zimmermann, A. Lorenz, and R. Oppermann. An operational definition of context. In *Proceedings of CONTEXT '2007*, 2007.