Time-Aware Entity Linking

Renato Stoffalette João

L3S Research Center, Leibniz University of Hannover,
Appelstraße 9A, Hannover, 30167, Germany,
joao@L3S.de

Abstract. Entity Linking is the task of automatically identifying entity mentions in a piece of text and linking them to their corresponding entries in a reference knowledge base like Wikipedia. Although there is a plethora of works on entity linking, existing state-of-the-art approaches do not explicitly consider the time aspect and specifically the temporality of an entity's prior probability (popularity) and embedding (semantic network). Consequently, they show limited performance in annotating old documents like news or web archives, while the problem is bigger in cases of short texts with limited context, such as archives of social media posts and query logs. This thesis focuses on this problem and proposes a modeling that leverages time-aware prior probabilities and word embeddings in the entity linking task.

1 Introduction and Problem Statement

One way to enhance machine understandability of natural language documents is by adding semantics to documents. The Linked Open Data and Semantic Web communities have played a major role and gained a lot of attention over the past years. Researchers have produced a great deal of methods for publishing structured data on the web and interlinking related concepts coming from different sources. A diverse number of applications can benefit from these initiatives, for example, query expansion and auto completion, knowledge base population, ranking results from search engines, among others.

This thesis focuses on the entity linking problem [20], which basically consists in automating the process of entities extraction from natural language documents and linking them to the correct concepts in a reference knowledge base, for instance Wikipedia [13], YAGO [11] and DBPedia [12]. Although there is a plethora of works trying to solve this problem, existing state-of-the-art approaches do not explicitly consider the time aspect in their modeling and thus work well mainly when dealing with documents published close to the model creation and training times. Nevertheless, the increasing number of digital archives worldwide, including news, web, and social media archives, and the need for their effective analysis and exploration, requires the use of effective, time-aware entity annotations methods [7].

Consider for example the text snippet "*Ronaldo scored and Real Madrid won*". If this text belongs to an article from 2017, the mention *Ronaldo* probably refers to the portuguese soccer player *Cristiano Ronaldo*. However, if the document is dated from 2002, then it probably refers to the brazilian soccer player *Ronaldo Luís Nazário de Lima*. For handling such cases, existing entity linking approaches should be trained again using data and knowledge from older versions of the reference knowledge base(s), which can be very laborious and also infeasible since such training information is probably neither sufficient nor available (especially for older time periods). The problem is even bigger when trying to annotate short texts with limited context, like old social media posts or query logs. For example, the query "*Germany Brazil*" submitted in July 2014 probably refers to the football match of the 2014 FIFA World Cup, and thus the mentions should be linked to the football teams, not the countries.

An entity's *prior probability* (i.e its popularity) is an important component in most entity linking approaches and is considered a strong indicator (as well as a baseline) to select the correct entity for a given mention [6]. In addition, the *context* of an entity (i.e its semantic network) is another important characteristic that is exploited by state-of-the-art entity linking approaches [17, 11]. In this thesis, we consider that both the prior probability of an entity and the context of an entity mention are temporal in nature, and thereby the time aspect should be explicitly considered in entity linking. Towards this objective, we introduce and formulate the problem of *"Time-Aware Entity Linking"*. Given a document and a time period (e.g., publication date or an estimation of publication date), our approach links entities in the document to a contemporary knowledge base (Wikipedia) by considering time-aware prior probabilities and word embeddings.

Below, we motivate the problem, discuss related works, pose the main research questions and hypotheses, and describe the methodology that we follow as well as our evaluation plan.

2 Motivation

Easy access to historical Web information becomes more and more important as significant parts of our cultural heritage are produced and consumed online. National and international initiatives have recognized this need and started to collect and preserve parts of the Web. The most prominent one being the Internet Archive[1], has collected more than 2.5 Petabyte of Web content since 1996. In the same direction, efforts have emerged to collect and preserve social media archives, like the Twitter Archive at the Library of Congress [24].

Despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning them into usable and useful information sources [2, 4].

Entity linking can play an important role in assisting to add semantics to archived documents, which in turn will enable their effective analysis and exploration [7]. However, from our experience in trying to annotate a variety of historical documents in the context of the Alexandria project[2] (like news

[1] https://archive.org/

[2] ERC Advanced Grant, Nr. 339233, http://alexandria-project.eu/

articles, web pages, query logs, social media posts), existing state-of-the-art entity linking approaches show limited performance, producing a reasonable number of false annotations due to their time-agnostic approach. We believe that making entity linking time-aware can increase the quality of annotations, especially for the case of historical Web content.

3 Related work

Entity linking requires a knowledge base containing the entities to which entity mentions can be linked. In the open domain text, one popular approach is to construct a dictionary of entities from online encyclopedias such as Wikipedia or from semantic networks such as DBPedia, Babelnet and YAGO.

Bunescu et al. [3] were the first to propose a solution for the entity linking problem using Wikipedia, but only in the work done by Mihalcea and Csomai [13] the term *Wikification* was formally introduced. *Wikification*, sometimes also called Disambiguation to Wikipedia (D2W), consists basically in extracting the most important concepts in a document and identifying for each of these concepts an appropriate link to a Wikipedia article.

Both works are considered local approaches, because only one mention is disambiguated at a time and the methods are focused on well written documents. The main drawback of disambiguating one mention at a time is the fact that there is no assumption of relatedness among the entities. Thus, in order to overcome such limitation, Cucerzan [5] introduced a global approach in which the disambiguation process is performed for all the mentions at the same time. In his work, he recognizes coherence and a general interdependence (i.e a semantic relation) among mentions in the same document.

Milne and Witten [15] proposed a machine learning approach that combined an entity's prior probability with its relatedness to the surrounding context, but their approach relied strongly on the unambiguous mentions to create context and disambiguate the ambiguous ones, and hence it was not very efficient on fragments of texts that did not have at least one single unambiguous mention.

Other authors have tried to solve entity linking problems using different techniques and also link to entities derived from other knowledge bases or semantic networks. DBPedia Spotlight [12] for example, links entities to DBPedia by basically calculating cosine similarity between the context of the entity mention and the context of each candidate entity using the Bag-of-Words approach. Yet, more robust approaches such as AIDA [11] which links entities to YAGO and Wikipedia, Babelfy [17] which links entities to Babelnet or even Tagme [9] and WAT [19] which link only entities in Wikipedia, employ graph based algorithms and try to find densest sub graphs as a solution, but finding densest sub graph is computationally *NP-hard*, therefore each one adopts a different heuristic to find an approximation algorithm and solve the entity ambiguity.

More recent approaches, such as the ones proposed by Blanco et al. [1], Pappu et al. [18] and Moreno et al. [16] employ neural network models, and the reason for their success is due to the lower computational complexity of neural models and the possibility to calculate very accurate high dimensional word vectors.

Despite the number of contributions and the variety of techniques applied to solve entity linking problems, none of the previous works have incorporated the time aspect. The work which is more related to our objectives is the one proposed by Fang and Chang [8] in which spatiotemporal signals are modeled for solving the entity ambiguity in microblogs.

Our preliminary results validate our hypothesis that entities are temporal in nature, and thus we believe that the time aspect should be taken into account in the entity linking task.

4 Research question(s)

This thesis aims at addressing the following research question:

- (Q1) *Can we improve entity linking on archived content by considering the time aspect as well as characteristics of the underlying corpus?*

Besides, we also focus on the following two sub-questions:

- (Q2) *How can we evaluate such entity linking approaches?*
- (Q3) *How can we efficiently annotate large collections of documents?*

Addressing the main research question (Q1) allows someone to provide high-quality annotations for a variety of archived content, like documents in web or news archives, social media posts, query logs, etc. Since existing benchmarks and ground truths do not consider the time aspect, Q2 focuses on providing the means to evaluate the effectiveness of time-aware entity linking approaches and compare it with baseline and time-agnostic approaches. Finally, since web and social media archives are usually huge in size, Q3 aims at providing efficient and scalable approaches.

5 Hypotheses

Our main research question (Q1) is based on the following two hypotheses:

- (H1) The *prior probability* of an *entity* is temporal in nature.
- (H2) The *context* of an *entity mention* is temporal in nature.

In simple terms, H1 considers that the popularity of an −ambiguous in our case− entity changes over time, while H2 considers that the semantic network (strongly connected words) of an ambiguous entity mention also changes over time. Consider for instance the example in Section 1. The brazilian soccer player Ronaldo was very popular in 2002, while the portuguese player Cristiano Ronaldo is very popular nowadays but not in 2002. For the same reason, the context of the word "ronaldo" has changed over time. In 2002, "ronaldo" was probably co-occurring with words like "brazil", "inter" and "zidane", while nowadays it co-occurs with "madrid", "portugal", "messi", etc.

Both hypotheses have been partially validated by Fang and Chang [8] and Hamilton et al. [10]. The former have shown that entities' prior probabilities often change across time and domains, while the latter have studied words semantic evolution and shown that frequent words change more slowly while polysemous words change more quickly. Moreover, Zhang et al. [23] have demonstrated that entity recommendations for keyword queries change over time, while Tran et al. [21] noticed that entity relatedness is both time and context dependent. These works confirm our assumption that entities are dynamic in nature.

6 Approach

The main idea behind our approach is the use of time and corpus-dependent word embeddings.

6.1 Modeling and Problem Definition

Let D be a corpus of documents, for example a set of news articles, covering the time period $T_D = [\tau_s, \tau_e]$ (where τ_s, τ_e are two different time points with $\tau_s < \tau_e$). Consider also a contemporary reference knowledge base K, for instance Wikipedia, containing information for a set of entities E. Given a piece of text s extracted from a document $d \in D$ published during the time period $T_d \subseteq T_D$, the output of the *"time-aware entity linking"* task is a set of mappings (m, e) where m is a word or phrase from s, and $e \in E$ is an entity from K that determines the identity of m based on both the context of m (s, d, and D) and the time period T_d.

For example, given i) the sentence *"Ronaldo scored and Real Madrid won"* extracted from a 2002 article of a news archive, and ii) Wikipedia as the reference knowledge base, a correct mapping is the following: (Ronaldo, `https://en.wikipedia.org/wiki/Ronaldo_(Brazilian_footballer)`).

6.2 Approach Overview

The main components in the proposed method are depicted in Figure 1 and further described below.

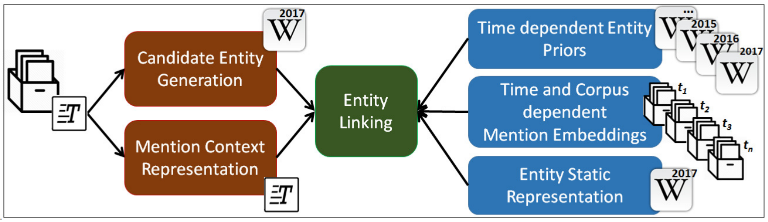
2017
Candidate Entity Generation
Mention Context Representation
Entity Linking
Time dependent Entity Priors
2015
2016
2017
Time and Corpus dependent Mention Embeddings
t_1
t_2
t_3
t_n
Entity Static Representation
2017

Fig. 1: The main components in our approach.

Candidate Entity Generation and Mention Context Representation. For an input document d, we first generate a list of pairs (m, E_m) where each mention m has a list of candidate entities $E_m \in E$. For this we exploit the anchor texts from Wikipedia articles and we create a dictionary of mentions and entities. Thus, for every mention m we select as candidate entities those that appear as link destinations for m. Each mention has also a *context representation* that considers its surrounding text and mentions. For this we intend to investigate other variations, such as combining topical coherence or exploring different windows sizes.

Time dependent Entity Priors. We compute time-dependent entity prior probabilities for all entities in E by exploiting different Wikipedia editions. There are two major challenges in this step. The first one is how to deal with long-tail entities for which limited or no content exists in old Wikipedia versions. The second one is how to deal with older time periods for which there is no Wikipedia version.

Time and Corpus dependent Mention Embeddings. We create different mention embedding models for different time periods (e.g., month-wise and year-wise) by exploiting the documents in D. For this, we use the *word2vec* algorithm proposed by Mikolov et al. [14] and also adopt standard techniques from vector quantization and signal compression methods to quantize the entries of the word vectors and encode the quantization results [1]. This enables storage and retrieval in a space and time efficient way.

Entity Static Representation. An entity $e \in E$ has a static and up-to-date (independent of T_d) representation, built by exploiting encyclopedic-like text describing information about e (i.e its Wikipedia page). A challenging issue here is how to represent long-tail entities for which limited information is provided.

7 Evaluation plan

Existing gold standard data sets used for entity linking methods assessment do not take into account the time dimension, making it difficult to compare between our proposed approach and time-agnostic methods. Therefore we need to create new ground truth data sets for a variety of archived corpora, including news and web archives, social media archives and query logs. Currently we are in the process of annotating news documents extracted from the New York Times corpus with AIDA, Babelfy and Tagme. Our next steps include analyzing their overall agreement and crowd source corrections of eventual mistakes. We plan to evaluate the effectiveness of our approach for both old and new documents and within several time granularities.

We also intend to conduct experiments and comparisons with entity linking approaches integrated into the Gerbil framework [22] as it already offers a web-based platform for comparison of tools using multiple data sets and uniform measuring approaches.

8 Preliminary results and Reflections

Since the computation of prior probabilities is typically done over knowledge bases such as Wikipedia, we confirmed H1 by analyzing Wikipedia editions in two different time periods (2006 and 2016). For each edition, we collected all links in Wikipedia articles and created a reference knowledge base of mentions and their referring entities. The probability that a mention m links to the entity

e is given by the number of times m links to e over the number of times that m occurs in the whole corpus.

Initially we were only concerned with the top ranked candidate entity for each one of the 31,123 mentions that commonly occurred in between the two Wikipedia editions. When considering both ambiguous and unambiguous mentions, in 9.44% of the cases the mention changed its top ranked candidate entity, whilst when removing the unambiguous mentions this number increased to 15.36%. This is mainly due to the fact that most of the unambiguous mentions kept the same entity mappings.

Moreover, we examined the changes in the top-5 ranked positions of the candidate entities. For that we computed the entities rank correlation by treating an element i which appears in list L_1 and not in L_2, at position $|L_2| + 1$. We computed the rank correlation for 18,727 mentions, since this is the number of mentions that are ambiguous and appear both in the 2006 and 2016 Wikipedia corpus. We noticed that in 71.98% of the cases the rank correlation values were greater than 0.5, which tells us there is some significant number of changes in the candidate entities' rank positions. Table 1 shows an example of a mention and its top-5 candidate entities for different Wikipedia editions.

Table 1: Mention example and its top-5 ranked candidate entities in two different Wikipedia editions (i.e 2006 and 2016).

Mention	***Entity***	***Prior***	***Year***
Watson	Doctor Watson	0.146	2006
	James D. Watson	0.130	
	Watson, Australian Capital Territory	0.115	
	Division of Watson	0.076	
	Watson	0.061	
Watson	Watson (computer)	0.068	2016
	Ben Watson (footballer, born July 1985)	0.054	
	Je-Vaughn Watson	0.050	
	Jamie Watson (soccer)	0.047	
	Arthur Watson (footballer, born 1870)	0.043	

In Zhang et. al [23] the authors created a dataset with 22 queries to enable evaluations for time-aware entity recommendation. We used this dataset to perform some preliminary evaluations of our approach. We created month-wise word embeddings using the provided news corpus which spans over 7 months and contains about 7 million articles. We evaluated a total of 26 ambiguous mentions (with $>$ 1 candidates) without considering context information (for queries such as "*Tour de France Nibali*" we counted two ambiguous mentions, being "*Tour de France*" and "*Nibali*"). For each mention, we collected its candidate entities and evaluated if our approach could identify the correct Wikipedia entity in a reference knowledge base created from a 2017 Wikipedia edition. Our approach reaches a promising accurary of about 73% on **only ambiguous** mentions (19/26 cases of correct entity linking).

9 Conclusion

This thesis introduces and formalizes the problem of time-aware entity linking which can be particularly useful for annotating archived collections of documents, such as news or web archives. We validated our hypothesis that the prior probability of an entity is temporal in nature and we presented an entity linking modeling that incorporates time-aware entity priors and word embeddings. In the future, we intend to extensively evaluate variations of our model for different archived corpora and time periods, using ground truth data sets created specifically for time-aware entity linking.

Acknowledgments

I am thankful to my advisors Dr. Pavlos Fafalios and Prof. Dr. Wolfgang Nejdl who provided expertise and helped me structuring this work.

References

1. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 179–188. ACM (2015)
2. Bruns, A., Weller, K.: Twitter as a first draft of the present: and the challenges of preserving it for the future. In: Proceedings of the 8th ACM Conference on Web Science. ACM (2016)
3. Bunescu, R.C., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Eacl. vol. 6, pp. 9–16 (2006)
4. Calhoun, K.: Exploring digital libraries: foundations, practice, prospects. Facet Publishing (2014)
5. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic. pp. 708–716 (2007)
6. Fader, A., Soderland, S., Etzioni, O.: Scaling wikipedia-based named entity disambiguation to arbitrary web text. In: IN PROC. OF WIKIAI (2009)
7. Fafalios, P., Holzmann, H., Kasturia, V., Nejdl, W.: Building and Querying Semantic Layers for Web Archives. In: ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'17) (2017)
8. Fang, Y., Chang, M.W.: Entity linking on microblogs with spatial and temporal signals. Transactions of the Association for Computational Linguistics 2 (2014)
9. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1625–1628. ACM (2010)
10. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096 (2016)
11. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 782–792. Association for Computational Linguistics (2011)
12. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics) (2011)

13. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 233–242. ACM (2007)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
15. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM (2008)
16. Moreno, J.G., Besançon, R., Beaumont, R., Dhondt, E., Ligozat, A.L., Rosset, S., Tannier, X., Grau, B.: Combining word and entity embeddings for entity linking. In: European Semantic Web Conference. pp. 337–352. Springer (2017)
17. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics 2, 231–244 (2014)
18. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 365–374. ACM (2017)
19. Piccinno, F., Ferragina, P.: From tagme to wat: a new entity annotator. In: Proceedings of the first international workshop on Entity recognition & disambiguation. pp. 55–62. ACM (2014)
20. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering 27(2), 443–460 (2015)
21. Tran, N.K., Tran, T., Niederée, C.: Beyond time: Dynamic context-aware entity recommendation. In: European Semantic Web Conference. Springer (2017)
22. Usbeck, R., Röder, M., Ngonga Ngomo, A.C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al.: Gerbil: general entity annotator benchmarking framework. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1133–1143. ACM (2015)
23. Zhang, L., Rettinger, A., Zhang, J.: A probabilistic model for time-aware entity recommendation. In: International Semantic Web Conference. pp. 598–614. Springer (2016)
24. Zimmer, M.: The twitter archive at the library of congress: Challenges for information practice and information policy. First Monday 20(7) (2015)