Linking Named Entities across Languages using Multilingual Word Embeddings

Elvys Linhares Pontes, Antoine Doucet, Jose G. Moreno

Linking Named Entities across Languages using Multilingual Word Embeddings

Elvys Linhares Pontes
elvys.linhares_pontes@univ-lr.fr
University of La Rochelle
La Rochelle, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle
La Rochelle, France

José G. Moreno
jose.moreno@irit.fr
University of Toulouse
Toulouse, France

ABSTRACT

Digital libraries are online collections of digital objects that can include text, images, audio, or videos in several languages. It has long been observed that named entities (NEs) are key to the access to digital library portals as they are contained in most user queries. However, NEs can have different spellings for each language which reduces the performance of user queries to retrieve documents across languages. Cross-lingual named entity linking (XEL) connects NEs from documents in a source language to external knowledge bases in another (target) language. The XEL task is especially challenging due to the diversity of NEs across languages and contexts. This paper describes a XEL system applied and evaluated with several languages pairs including English and various low-resourced languages of different linguistic families such as Croatian, Finnish, Estonian and Slovenian. We tested this approach to analyze documents and NEs in low-resourced languages and link them to the English version of Wikipedia. We present the resulting study of this analysis and the challenges involved in the case of degraded documents from digital libraries. Further works will make an extensive analysis of the impact of our approach on the XEL task with OCRed documents.

KEYWORDS

Cross-Lingual Named Entity Linking, Multilingual Word Embeddings, Digital Library, Indexing

ACM Reference Format:
Elvys Linhares Pontes, Antoine Doucet, and José G. Moreno. 2020. Linking Named Entities across Languages using Multilingual Word Embeddings. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital libraries are composed of a large number of digital contents (e.g. journals, books, magazines, videos, and so on) in several languages about diverse subjects (e.g. history, languages, politics, sciences, philosophy, and so on). Named entities have been demonstrated to be essential to digital library access as they are included in a majority of the search queries submitted to digital library portals [2]. However, the spelling of an entity is language-dependent which impacts the performance of search engines when trying to retrieve all relevant documents with respect to a query. For instance, the entity "United States" has different spelling in other languages: "*Estados Unidos*" (in Portuguese and Spanish) and "*États-Unis*" (in French).

Moreover, data from different sources can contain ambiguous, complementary, and duplicate information about named entities. Therefore, they are often not distinctive since one single name may correspond to multiple entities. A disambiguation process is thus essential to distinguish the correct named entities to be indexed in digital libraries. In this case, a monolingual disambiguation analysis cannot disambiguate these entities in several languages for a common knowledge base.

Named Entity Linking (NEL) aims to recognize mentions in a document and link them to their corresponding entries in a Knowledge Base (KB), such as Wikipedia[1], DBpedia[2], and Freebase[3]. Additionally, Cross-Lingual Named Entity Linking (XEL) considers documents that are written in a source language that is different from the target language of the KB [18]. In addition to the challenges of NEL such as multiple surface forms of a named entity [16], XEL disambiguates mentions in several languages by analyzing different spellings and contexts related to each language.

Digital libraries often contain the digitised version of old documents that are degraded due to storage conditions, handling of users and inherent vice of the material (e.g. paper naturally deteriorates over time). These problems cause numerous errors at the character and word levels in the OCR of these documents [10]. Linhares Pontes et al. [10] analyzed the impact of OCR quality on the NEL task and achieved satisfying results for NEL. They provided recommendations on the OCR quality that is required for a given level of expected NEL performance. However, their approach is monolingual, restricting the analysis and linking of entities to knowledge bases that are in the same language (in this case, English).

The XEL task is especially challenging due to the diversity of NEs across languages and contexts. This paper describes a XEL system applied and evaluated with several languages pairs including English and various low-resourced languages of different linguistic families such as Croatian, Finnish, Estonian and Slovenian. We tested this approach to analyze documents and NEs in low-resourced languages and link them to the English version of Wikipedia. We present the resulting study of this analysis and the challenges involved in the case of degraded documents from digital libraries.

The remainder of the paper is organized as follows : Section 2 makes a brief overview of the most recent and available NEL and XEL approaches in the state of the art. Section 3 details our approach to extend a monolingual NEL system for the XEL task by using multilingual word embeddings. Then, the experimental setup and

ACM/IEEE Joint Conference on Digital Libraries, 2020, China
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] http://www.wikipedia.org/
[2] https://wiki.dbpedia.org/
[3] http://www.freebase.com/

evaluation are respectively described in Sections 4 and 5. Finally, conclusions and future works are set out in Section 6.

2 AN OVERVIEW OF (CROSS-LINGUAL) NAMED ENTITY LINKING

Given a set of documents $D = \{d_1, d_2, \ldots, d_l\}$, a set of detected mentions $M^j = \{m_1^j, m_2^j, \ldots, m_n^j\}$ in the document d_j for $\forall j \in [1, l]$, and a knowledge base $KB = \{e_1, e_2, \ldots, e_s\}$, Named Entity Linking (NEL) aims to map each mention m_i^j with its corresponding entity e_k in the KB [16]. NEL approaches can be divided into two classes: disambiguation (they use M^j as an input) and end-to-end approaches (they do not use M^j as an input, but calculate it). While end-to-end approaches extract candidate entities from documents and then disambiguate them to the correct entries in a given KB [8], disambiguation approaches only disambiguate entities already recognized from documents [5, 9, 14].

Among the only disambiguation approaches, Ganea and Hofmann [5] built a deep learning model for joint document-level entity disambiguation. They embed entities and words in a common vector space and use a neural attention mechanism to select words that are informative for the disambiguation decision. Then, their model collectively disambiguates the mentions in a document (more details in Section 3.1). Motivated by Ganea and Hofmann's approach, Le and Titov [9] analyzed relations between mentions as latent variables in their neural NEL model. They rely on representation learning and learn embeddings of mentions, contexts, and relations to reduce the amount of human expertise required to construct the system and make the analysis more portable across languages and domains.

In the class of end-to-end approaches, Raiman and Raiman [14] developed a system for integrating symbolic knowledge into the reasoning process of a neural network through a type system. They constrained the behavior to respect the desired symbolic structure, and automatically design the type system without human effort. Their model first uses heuristic search or stochastic optimization over discrete variables that define a type system informed by an Oracle and a learnability heuristic. Based on a joint analysis of the named entity recognition and linking tasks, Kolitsas et al. [8] proposed an end-to-end NEL system that jointly discovers and links entities in a document. They generate all possible spans (mentions) that have at least one possible entity candidate. Then, each mention-candidate pair receives a context-aware compatibility score based on word and entity embeddings [5] coupled with neural attention and a global voting mechanism.

Extending this monolingual analysis, Cross-Lingual Named Entity Linking (XEL) analyzes documents and named entities that are in a different language than that used for the content of the knowledge base. In this context, McNamee et al. [11] proposed an XEL approach and examined the importance of transliteration, the utility of cross-language information retrieval, and the potential benefit of multilingual named entity recognition on the XEL task.

Zhou et al. [18] extensively evaluated the effect of resource restrictions on existing XEL methods in low-resource settings. They investigated a hybrid candidate generation method, combining existing lookup-based and neural candidate generation methods and proposed a set of entity disambiguation features that are entirely language-agnostic. Finally, they designed a non-linear feature combination method, which makes it possible to combine features in a more flexible way.

3 OUR CONTRIBUTION

This section describes our contribution to adapt Ganea and Hofmann's approach for the XEL task. We make a short description of Ganea and Hofmann's approach (Section 3.1), and then we detail how we extended this approach for the XEL task by using multilingual word embeddings (Section 3.2).

3.1 Ganea and Hofmann's approach

Entity Disambiguation (ED) approaches consider having already identified the named entities in the documents. In this case, these approaches aim to analyse the context of these entities to disambiguate them in a KB. In this context, Ganea and Hofmann [5] (GH) proposed a deep learning model for joint document-level entity disambiguation[4].

They project entities and words in a common vector space, which avoids hand-engineered features, multiple disambiguation steps, or the need for additional ad-hoc heuristics when solving the ED task. Entities for each mention are locally scored based on cosine similarity with the respective document embedding. Combined with these embeddings, they proposed an attention mechanism over local context windows to select words that are informative for the disambiguation decision. The final local scores are based on the combination of the resulting context-based entity scores and a mention-entity prior. This mention-entity prior ($p(e|m)$) is a conditional distribution of the co-occurrence of the mention m with the entity e. In this case, GH collected mention-entity co-occurrence counts from Wikipedia to calculate this distribution.

Finally, mentions in a document are resolved jointly by using a conditional random field in conjunction with an inference scheme.

3.2 Cross-lingual extension

Most data sets for NEL are available only in English. Among them, the AIDA data set is the main data used to train NEL system on the state of the art. Unfortunately, there are few data sets for low-resourced languages, with the notable exception of the WikiANN corpora.

In order to extend GH's system to a cross-lingual setting, we made a number of modifications to their approach. Instead of using the Word2Vec embeddings, we used the pre-trained multilingual MUSE embeddings[5] [3]. These embeddings are available in 30 languages (including Croatian, Estonian, Finnish, Slovenian, to mention a few) and they are aligned in a single vector space. Therefore, words like "house" and "talo" ("house" in Finnish) have similar word representations. One of the main goals of using these embeddings is to generate multilingual entity embeddings that can provide entity representations for mentions in several languages. Then, GH's approach will be able to analyse documents in the languages of these embeddings and link them to an English KB. Therefore, we generate the entity embeddings using the English version of Wikipedia and train this system on the AIDA data set using the MUSE embeddings.

[4]The code is publicly available: https://github.com/dalab/deep-ed
[5]The MUSE embeddings are available at: https://github.com/facebookresearch/MUSE

In this scenario, GH's approach analyses English documents and links their mentions to an English KB.

Moreover, we extend the training process for some low-resourced languages by using the previous English model and continue the training process with data on other languages. This tuning procedure optimises our model to analyse better the documents on low-resourced languages and link their mention to an English KB. More precisely, we initialized the weights of the neural network model with the weights of the English model, and we reduced the learning rate to tune our model for the target languages. This process enables our model to adapt the analysis of words and their context for each language (e.g. the order of words and how they are combined to express a same idea in different languages).

4 EXPERIMENTAL SETUP

In order to analyse the impact of using multilingual embeddings on the representation of entity embeddings, we used the entity relatedness data set of Ceccarelli et al. [1] to compare the quality of entity embeddings produced by the Word2Vec and multilingual embeddings. This data set contains 3319 and 3673 queries for the test and validation sets. Each query consists of one target entity and up to 100 candidate entities with gold standard binary labels indicating if the two entities are related or not. The associated task requires ranking of related candidate entities higher than unrelated ones. Following GH's work, we used the normalised discounted cumulative gain (NDCG) and mean average precision (MAP) measures to evaluate them. We also performed candidate ranking based on cosine similarity of entity pairs.

We then trained and tested GH's approach with the following benchmarks: AIDA-CoNLL [7], AQUAINT [6], ACE2004 [6, 15], WikiANN [13], CWEB [4] and WIKI [15].

The WikiANN data set was split into 2 separate data sets, 70% of the corpus for training and 30% for testing. For the training process, we use AIDA data set to train the NEL system for English using the MUSE embeddings. Then, we use the WikiANN training data set to optimise the English model for each low-resourced language. Finally, we tested our model on the WikiANN test data sets.

Following previous works, we evaluate the performance of our approach by analyzing the precision, recall and F1-measure. Precision is the fraction of correctly linked entity mentions that are generated by a system. Recall considers all entity mentions that should be linked and determines how correct linked entity mentions are concerning the total entity mentions that should be linked. Finally, the F1-measure is defined as the harmonic mean of precision and recall.

Since knowledge bases contain millions of entities, only mentions that contain a valid ground-truth entry in the KB are analysed. For mentions without corresponding entries in the KB, NEL systems have to provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB.

5 EXPERIMENTAL ASSESSMENT

Entity embeddings performance: Table 1 shows the entity relatedness results using Word2Vec and MUSE embeddings for the English data set [1]. Both embeddings have the same dimensional space (300 dimensions) but different vocabulary sizes: Word2Vec (3 million tokens) and MUSE (200 thousand tokens). This large difference helps Word2Vec to achieve the best results for all entity related measures. More precisely, the Word2Vec embeddings provide a better analysis of the Wikipedia documents because it has less out-of-vocabulary words than the MUSE embeddings and can represent better the meaning of sentences and entities. Despite this performance drop, GH's approach using MUSE embeddings achieved better results than [17] and [12] for all metrics.

Table 1: Entity relatedness quality for English.

Embeddings	NDCG1	NDCG5	NDCG10	MAP
Ganea and Hofmann (Word2Vec)	**0.632**	**0.609**	**0.641**	**0.578**
Ganea and Hofmann (MUSE)	0.613	0.568	0.592	0.536
Yamada et al. [17]	0.59	0.56	0.59	0.52
Milne and Witten [12]	0.54	0.52	0.55	0.48

NEL analysis for mono- and multilingual embeddings: Advancing our analysis of GH's system, we compared the F1-measure results for this system on English corpora using the Word2Vec and MUSE embeddings (Table 2). As expected, the small vocabulary and lower performance in the entity relatedness measures reduced the performance of GH's system in the NEL task. These factors reduced the quality of the attention and the context embeddings, and prioritised the relevance of entity priors ($\log\ p(e|m)$) to disambiguate the mentions in a document. Despite this drop, GH's system using MUSE achieved identical or very close perfomance for most data sets.

Table 2: F1-measure results for Ganea and Hofmann's approach on English corpora.

Embed.	AIDA	ACE2004	AQUAINT	CLUEWEB	WIKI
Word2Vec	**92.2**	**88.5**	**88.5**	**77.9**	**77.5**
MUSE	86.6	**88.5**	87.5	74.9	74.2

XEL analysis: Table 3 presents the F1-measure results for the NEL on four languages of the WikiANN corpora. We tested the NEL system using only the AIDA training data set to train GH's model in order to link mentions to the English version of the Wikipedia; and using the AIDA training data set in a first step and, then, the WikiANN training data set for each language (second line of Table 3). The tuning process on the WikiANN data set improved the performance of GH's for the WikiANN test data sets. Unfortunately, the WikiANN data set is composed of short sentences with little contextual information. This characteristic makes the context analysis of GH's system less relevant and implies that the disambiguation process mainly consists in pairwise matching between mentions and entities using $\log\ p(m|e)$. Another limiting factor is the small MUSE vocabulary. Finally, the English version of Wikipedia does not have all entities listed on the Croatian, Estonian, Finnish, and Slovenian Wikipedia versions, which reduces the number of entities that can be linked to the KB.

Table 3: F1-measure results for Ganea and Hofmann's models on the test WikiANN corpora (Croatian, Estonian, Finnish, and Slovenian languages only).

Models	hr	et	fi	sl
AIDA data set (using MUSE)	60.97	57.82	62.51	69.78
pre-trained model on AIDA data set + tuning on WikiANN data set (using MUSE)	**61.53**	**58.47**	**63.04**	**70.31**

XEL is a fundamental tool for search engines in digital libraries to retrieve documents where their contents (including named entities) are in different languages and contexts. Linhares Pontes et al. [10] showed an analysis of the impact of problems detected in these libraries using Ganea and Hofmann's and Le and Titov's systems. In this analysis, these systems had a small reduction in NEL performance despite the errors caused by the deterioration and conservation problems in libraries. In this work, we showed that Ganea and Hofmann's system using multilingual embeddings achieved satisfactory results for the English NEL task (maximal F1-measure drop of 5.6%). Additionally, the tuning procedure improved the results for XEL in the low-resourced languages. We assume our approach will perform similarly for the XEL task in OCRed documents, but additional experiments are needed to validate this assumption.

6 CONCLUSION

This paper is the first step to analyze the impact of multilingual embeddings to extend monolingual NEL to XEL. The next step is to investigate the impact of degraded documents on this cross-lingual task.

Despite the small multilingual vocabulary on the word embeddings and the poor context quality of training data sets for low-resourced languages, our experiments showed a worst drop of 5.6% on F1-measure on the English test data set (and the same performance of monolingual embeddings in the best case) and a small improvement with the tuning procedure on low-resourced languages. Therefore, we intend to build training data sets on the target languages that are composed of long sentences with rich context information to improve our XEL model.

Further work is under progress to develop and analyze the performance of end-to-end XEL systems on OCRed data sets. More precisely, we want to extend the analysis of multilingual embeddings with language-agnostic features and relations between entities to provide correct predictions in different languages and overcome the problem of OCR degradation. We also intend to analyze and test the performance of these systems using real data in other languages (e.g. Spanish and Chinese) including other low-resourced languages.

ACKNOWLEDGMENTS

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

REFERENCES

[1] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Learning Relatedness Measures for Entity Linking. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 139–148.

[2] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: towards a better access to information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 249–252.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *International Conference on Learning Representations (ICLR '18)*.

[4] Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0).

[5] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2619–2629.

[6] Zhaochen Guo and Denilson Barbosa. 2014. Robust Entity Linking via Random Walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 499–508.

[7] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 782–792.

[8] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 519–529.

[9] Phong Le and Ivan Titov. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1595–1604.

[10] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. 2019. Impact of OCR Quality on Named Entity Linking. In *Digital Libraries at the Crossroads of Digital Information for the Future*, Adam Jatowt, Akira Maeda, and Sue Yeon Syn (Eds.). Springer International Publishing, Cham, 102–115.

[11] Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-Language Entity Linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 255–263.

[12] David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*. ACM, New York, NY, USA, 509–518.

[13] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1946–1958.

[14] Jonathan Raiman and Olivier Raiman. 2018. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 5406–5413.

[15] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1375–1384.

[16] W. Shen, J. Wang, and J. Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (Feb 2015), 443–460.

[17] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Berlin, Germany, 250–259.

[18] Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. Towards Zero-resource Cross-lingual Entity Linking. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Hong Kong, China, 243–252.