LSI_UNED at M-WePNaD: Embeddings for Person Name Disambiguation

Andres Duque, Lourdes Araujo, and Juan Martinez-Romo

Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain
aduque@lsi.uned.es, lurdes@lsi.uned.es, juaner@lsi.uned.es

Abstract. In this paper we describe the participation of the LSI_UNED team in the multilingual web person name disambiguation (M-WePNaD) task of the IberEval 2017 competition. Our proposal is based on the use of word embeddings for representing the documents related to individuals sharing the same person name. This may lead to an improvement of the clustering process that aims to eventually separate those documents depending on the individual they refer to. This is one of the first approximations to the use of techniques based on embeddings for this kind of tasks. Our preliminary experiments show that a system using a representation setting based on word embeddings is able to obtain promising results on the addressed task, overcoming the proposed baselines in all the tested configurations.

Keywords: Person Name Disambiguation, word embeddings, document representation, clustering

1 Introduction

Person Name Disambiguation (PND or PNaD) is the task that addresses to disambiguate proper names belonging to different people that can be found on the Web, using a search engine. That is, given a specific person name (e.g., John Smith) and the results offered by a search engine when that name is introduced, the final aim of the task is to cluster the webpages offered by the engine as a result. All the webpages contained in a particular cluster should ideally refer to the same person. It is a well known area of research in both the Natural Language Processing (NLP) and Information Retrieval (IR) communities, and its difficulty comes from the high ambiguity that can be found in person names, considered as named entities, as well as from the heterogeneus results that can be offered by the search engine (professional pages, blogs, social media links and many more) [8].

First works on PND presented a set of unsupervised clustering algorithms for testing how automatically extracted relevant features (proper nouns and most relevant words) and biographical information (birth place and year, occupation) could be used for improving unsupervised clustering [14]. However, PND tasks were widely popularized thanks to the WePS (Web People Search) campaigns

(WePS-1, WePS-2 and WePS-3) [3, 4, 1], which presented a standardized framework for evaluating systems addressing this kind of tasks.

The PND task can be divided into two different parts: first, each of the documents (web pages) retrieved by a search engine when looking for a particular name have to be transformed into a structured representation. This structure could be subsequently used by a clustering algorithm in what is considered to be the second part of the process. For the first subtask (representing the documents that refer to different people), a key aspect shared by most of the systems performing PND is the extraction of feature sets able to characterize the documents. These features will eventually represent the discriminators for determining which cluster should contain which document after the clustering process. Amongst these features, bag-of-words and named entities are normally used by almost all the systems presenting state-of-the-art results [7, 18]. Those basic features are then usually enriched with others such as Wikipedia concepts [13] or hostnames and page URLs [10].

Regarding the clustering algorithms, Hierarchical Agglomerative Clustering (HAC) seems to be able to offer the best results in this task, so many of the systems able to offer the best results in the campaings make use of that algorithm or different versions of it [10, 12]. However, we can find other works in the literature achieving state-of-the-art results through the use of novel clustering algorithms, such as the proposed in [9], based on adaptative thresholds which circumvent the problem of depending on training data for determining the thresholds used by HAC.

The main objective in the development of our system is to introduce the use of word embeddings to a specific PND task, more particularly to the first subtask, related to the representation of the documents used for performing the clustering process. Word embeddings were introduced in [5] as a distributed representation of words in which the dimensionality of a particular vocabulary was reduced to a much smaller, fixed size. This way, each word in the vocabulary is associated to a point in a vector space. Although there are some works in the literature that make use of embeddings for document representation in named entity disambiguation [11, 6], we have not found studies regarding their use in the specific Person Name Disambiguation task. Hence, our aim is to propose a system that uses these embeddings for eventually generate a vector which represents the whole document. This vector will be then used by the clustering algorithm for separate the documents belonging to different people bearing the same name.

The rest of the paper is organized as follows: Section 2 introduces the task and the characteristics of the corpus used in it. The description of the system is detailed in Section 3, while the results obtained are presented in Section 4. Finally, Section 5 offers some conclusions and future lines of work.

2 The Task: Multilingual Web Person Name Disambiguation (M-WePNaD)

As it has been stated before, the Person Name Disambiguation task consists in clustering the different webpages offered by a search engine when a particular person name is introduced as query, for distinguishing the different real individuals associated with that name. The specific task presented within the context of the IberEval 2017 competition represents a difference with common PND tasks in the sense that it is assessed in a multilingual setting. That is, the documents offered by the search engine, and considered for performing the final clustering, can be written in different languages.

2.1 Resources

The main resource used for this task is the multilingual corpus MC4WePS, developed by the organizers of the task [16]. The corpus, which aims to become a reference resource for this task, was built by extracting information from two search engines (Yahoo and Google), regarding 100 different person names, under two main criteria: ambiguity and multilingualism. The first criterion aims to obtain non-ambiguous, ambiguous and highly ambiguous names (related to 1, to between 2 to 9, and to more than 9 different people, respectively). The multilingualism criterion is satisfied by offering both monolingual or multilingual pages (that is, pages written in only one, or more than one, languages), and also by considering cases in which there exists both monolingual and multilingual pages that refer to the same individual.

The corpus is split in two different parts: training and testing. Participants in the task were given the training part of the corpus, composed of the results related to 65 person names, together with the Gold Standard containing the different clusters for each person name and the correspondence between each single document and the cluster it belongs to, amongst those associated to a particular person name. The testing part of the corpus, containing the results from the remaining 35 person names, was released to the participants at the end of the training phase, and in this case no Gold Standard was provided. Participants had to run their systems and assign, for each document belonging to a person name, the identifier of the cluster to which it belonged.

For each of the web pages retrieved by the search engine when looking for a specific name, the HTML document was obtained and transformed into plain text, using Apache Tika [1]. Hence, the corpus contains, for each result, the HTML document, the associated text document, and a XML file containing metadata such as the URL of the search result, ISO 639-1 codes for the languages the page is written in, the download date and the name of the annotator.

[1] https://tika.apache.org/

2.2 Evaluation

The evaluation metrics used for this task take into account the possibility of presenting overlapping clusters, that is, a document can belong to two or more different clusters at the same time. The metrics are: Reliability (R), Sensitivity (S), and their harmonic mean $F_{0.5}$ [2]. Given that the Gold Standard takes into account web pages which are not considered to be related to any individual bearing the ambiguous name, the evaluation is performed under two different settings: results considering only related web pages and results considering all web pages.

Two different baselines are offered for comparison, both in the training phase and in the results of the testing phase:

- **ALL-IN-ONE**: All the documents related to the same person name are gathered in a single cluster.
- **ONE-IN-ONE**: A different cluster is proposed for each single document related to the same person name.

3 System Description

The main objective of the system proposed for this task is to explore the use of word embeddings for eventually generating a vector representation of the documents from the corpus. Hence, we will be able to apply a clustering algorithm over those vectors for determining the different clusters for each person name proposed in the task.

3.1 Preprocessing

An initial preprocessing step is needed for preparing the documents retrieved by the search engine. As we stated before, the plain text from each document is provided within the corpus, and that is the main source of information that we will use in our process. We are interested in only considering named entities within the documents for representing them, that is, we consider that named entities are the most representative elements in the documents for this particular task. Hence, from the text documents, we perform a removal of the stopwords and we extract the named entities that can be found, through the use of the Stanford Named Entity Recognizer [2]. This way, we can transform our text documents into a "bag of named entities", which will be used for building the vectors representing the documents. Although we can find text written in different languages in the documents, we run the Stanford NER as if the document was always written in English (standard configuration).

In addition to this, as we will explain later on, we also maintain a version of each text document in which we retain all the words but the stopwords, without performing named entity recognition, in order to develop an experiment considering all the words in the documents, and not only named entities.

[2] https://nlp.stanford.edu/software/CRF-NER.shtml

3.2 Document vectors

The first step for building the vectors which represent the documents is to transform each word in the preprocessed documents (bags of named entities) to a specific word vector. For this purpose, we use pre-trained word embeddings, more particularly a collection of word vectors generated from Wikipedia concepts, publicly available for research purposes [3]. This collection presents around 1.7 million vectors representing Wikipedia concepts, and more than 1.5 million regular words [17]. The word vectors, with 300 dimensions, are built following the Skip-gram model used in Word2Vec [15], which has proved to present better overall results than the CBOW model presented in the same work.

Although we considered the possibility of building our own word vectors using the provided corpus, the preliminary experiments that we conducted did not offer promising results. This may be due to the reduced size of the corpus, which leads to a poor vector representation of the words.

Using the pre-trained collection, we transform each word in our preprocessed document into a word vector, and then we calculate the average vector of all the vectors related to the words in the document. This way, we generate a document vector of 300 dimensions which represents each particular document in the corpus.

3.3 Clustering algorithm

Once that we have transformed each document related to a particular person into a vector, we should be able to apply a clustering algorithm over the vectors of each specific person name, in order to separate those clusters related to different individuals sharing that name. We performed a set of preliminary experiments testing different clustering algorithms directly over the document vectors, but the obtained results were not successful enough (when compared with the baselines in the training dataset). Because of that, we focused on developing our own algorithm, based on the characteristics of the training corpus provided by the organizers. We observed that, for most of the person names in the corpus, there usually existed a big cluster, related to one individual, and gathering most of the documents of that person name, and then many small clusters, each of them related to a different individual also sharing that name. That is, the corpus seems to be somehow biased towards person names for which most of the results offered by the search engine refer to the same individual (the most "important" one), and then a reduced percentage of results related to other people.

Following this intuition, we have developed our clustering algorithm adapted to the characteristics of the corpus. The first step would be to select those documents related to the "important" individual for a person name. For this purpose, we need to calculate the similarity between each document and the rest of documents related to the same person name. Thanks to the conversion from text documents to vectors, we can easily compute this calculation by using

[3] https://github.com/ehsansherkat/ConVec

cosine similarity. The similarity weight associated to each document will be the average of the similarity between that document, and the rest of documents related to the same person name:

$$w_i = \frac{\sum_{\substack{k=1 \\ k \neq i}}^{n} cos(d_i, d_k)}{n-1}, \quad (1)$$

where d_i is the vector representing document i, d_k the vector representing document k and $cos(d_i, d_k)$ is the cosine similarity between d_i and d_k. The total number of documents related to a particular person name is n.

Once we have this similarity weight w_i for document i, we can perform a first pruning step, in which we will consider that all the documents with a similarity weight above a specific threshold γ should be gathered in the same initial cluster. For this cluster to be seen as representing the "important" individual for that person name, we should select a high threshold. This way most of the documents will be assigned to that cluster. After that, we will generate a different cluster for each of the remaining documents related to that person name, in order to follow the intuitions that we explained before. The final output of the system for each person name will be a list containing the document identifiers, followed by the identifier of the cluster to which each document has been assigned.

Through the experiments performed using the training dataset provided by the organizers, we have observed that the best value for the threshold is $\gamma = 0.75$, that is, considering all the documents with a similarity weight $w_i \geq 0.75$ to belong the same big cluster. Then, there is a cluster of size 1 for each of the remaining documents. However, for our experiments with the test dataset, and considering that we are allowed to propose up to 5 runs of our system, we will generate different runs by slightly varying this threshold around this value of $\gamma = 0.75$, as we will explain in the results section.

4 Results

As we said before, the evaluation is conducted using the measures of reliability and sensitivity, and their harmonic mean $F_{0.5}$. Tables 1 and 2 show the results obtained by the different runs of our system, for the testing dataset of the M-WePNaD task. Run 1 corresponds to a threshold of $\gamma = 0.70$, Run 2 to $\gamma = 0.75$, Run 3 to $\gamma = 0.80$ and Run 4 to $\gamma = 0.85$. The last run, Run 5, is a slightly different configuration of the system in which we consider all the words in the documents (except stopwords) to be representative of them, and not only named entities. The rest of the process remains the same (extraction of word vectors and construction of document vectors, and clustering algorithm). The threshold selected for this last run is $\gamma = 0.75$.

As we can observe, results are consistent in relation to the best run of our system, Run 3. That is, for the testing dataset, a threshold of $\gamma = 0.80$ is able to achieve the best results, although the differences with the other runs are small.

Table 1. Results considering only related web pages

System	R	S	$F_{0.5}$
Run 3	0.59	0.85	**0.61**
Run 4	0.74	0.71	**0.61**
Run 2	0.52	0.93	0.58
Run 5	0.52	0.92	0.57
Run 1	0.49	0.97	0.56
Baseline - ALL-IN-ONE	0.47	0.99	0.54
Baseline - ONE-IN-ONE	1.00	0.32	0.42

Table 2. Results considering all web pages

System	R	S	$F_{0.5}$
Run 3	0.59	0.81	**0.60**
Run 2	0.52	0.92	0.59
Run 5	0.52	0.90	0.59
Run 1	0.49	0.97	0.58
Run 4	0.74	0.66	0.58
Baseline - ALL-IN-ONE	0.47	1.00	0.56
Baseline - ONE-IN-ONE	1.00	0.25	0.36

We can also observe how using all the words in the documents (Run 5) is always under the run that only considers named entities and uses the same threshold (Run 2, with $\gamma = 0.75$). This implies that the addition of all the possible words in the documents introduces more noise than valuable information for the final disambiguation. In general, for the runs that make use of named entities, we can see how as we increase the threshold (that is, as the "important" cluster contains less documents), the reliability increases while the sensitivity decreases. That is, with small values of γ the results are closer to the "ALL-IN-ONE" baseline (all the documents in the same cluster), and as we increase γ we get closer to the "ONE-IN-ONE" baseline. However, those baselines are always overcome by all the runs of our system, which indicates that the strategy adopted for performing the clustering is valid for this particular task.

The differences between the two tables are quite small. In general, considering all web pages can be seen as a slightly harder task, since it expects the systems to determine those unrelated results. However, the number of unrelated web pages in the corpus is small and hence the results are quite similar between the two settings. We can observe that the order of the runs for our system is almost the same in both cases, except for run 4 (the highest value of gamma for the configuration that only takes named entities into account), which performs worse when considering all web pages.

5 Conclusions and Future Work

In this paper we have described our participation in the multilingual web person name disambiguation (M-WePNaD) task of the IberEval 2017 competition. The main contribution of our work is the application of embedding techniques for representing text documents related to different individuals. We have shown how word vectors extracted from pre-trained collections offer interesting possibilities for creating document vectors, that is, vectors representing whole documents, which can then be used for performing the final clustering process in the task. Results presented in Section 4 indicate the appropriateness of our proposal, from the point of view of overcoming the baselines proposed by the organizers of the task. It is important to remark that our system uses word embeddings in a very preliminary way, that is, we do not perform any other preprocessing apart from the extraction of named entities, and we have even tested the system without extracting them, only removing stopwords. This indicates that additional processing of the texts with other techniques (analysis of languages, type of web page, URL, etc.), will probably offer interesting improvements over the proposed technique. Also, the ad-hoc creation of word vectors whose contexts are closer to the task might also improve the results presented in this work, which have been obtained with pre-trained embeddings. Finally, the exploration of different clustering techniques is an important factor for the improvement of the system.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects EXTRECM (TIN2013-46616-C2-2-R) and PROSA-MED (TIN2016-77820-C3-2-R), as well as by the Universidad Nacional de Educación a Distancia (UNED) through the FPI-UNED 2013 grant.

References

1. Amigó, E., Artiles, J., Gonzalo, J., Spina, D., Liu, B., Corujo, A.: Weps-3 evaluation campaign: Overview of the online reputation management task. In: CLEF 2010 (Notebook Papers/LABs/Workshops) (2010)
2. Amigó, E., Gonzalo, J., Verdejo, F.: A general evaluation measure for document organization tasks. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 643–652. SIGIR '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2484028.2484081
3. Artiles, J., Gonzalo, J., Sekine, S.: The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 64–69. Association for Computational Linguistics (2007)
4. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: overview of the web people search clustering task. In: 2nd web people search evaluation workshop (WePS 2009), 18th www conference. vol. 9 (2009)

5. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research 3(Feb), 1137–1155 (2003)
6. Cai, R., Wang, H., Zhang, J.: Learning entity representation for named entity disambiguation. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 267–278. Springer (2015)
7. Chen, Y., Martin, J.: CU-COMSEM: Exploring rich features for unsupervised web personal name disambiguation. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 125–128. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1621474.1621498
8. Delgado, A.D., Martínez, R., Fresno, V., Montalvo, S.: A data driven approach for person name disambiguation in web search results. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 301–310. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (August 2014), http://www.aclweb.org/anthology/C14-1030
9. Delgado, A.D., Martnez, R., Montalvo, S., Fresno, V.: Person name disambiguation in the web using adaptive threshold clustering. Journal of the Association for Information Science and Technology pp. n/a–n/a, http://dx.doi.org/10.1002/asi.23810
10. Elmacioglu, E., Tan, Y.F., Yan, S., Kan, M.Y., Lee, D.: Psnus: Web people name disambiguation by simple clustering with rich features. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 268–271. Association for Computational Linguistics (2007)
11. Fang, W., Zhang, J., Wang, D., Chen, Z., Li, M.: Entity disambiguation by knowledge and text jointly embedding. CoNLL 2016 p. 260 (2016)
12. Liu, Z., Lu, Q., Xu, J.: High performance clustering for web person name disambiguation using topic capturing. In: Proceedings of The First International Workshop on Entity-Oriented Search (EOS). pp. 1–6. ACM, New York, NY, USA (2011), http://research.microsoft.com/en-us/um/beijing/events/eos2011/9.pdf
13. Long, C., Shi, L.: Web person name disambiguation by relevance weighting of extended feature sets. In: CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010), http://ceur-ws.org/Vol-1176/CLEF2010wn-WePS-LongEt2010.pdf
14. Mann, G.S., Yarowsky, D.: Unsupervised personal name disambiguation. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 33–40. Association for Computational Linguistics (2003)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
16. Montalvo, S., Martínez, R., Campillos, L., Delgado, A.D., Fresno, V., Verdejo, F.: Mc4weps: a multilingual corpus for web people search disambiguation. Language Resources and Evaluation pp. 1–28
17. Sherkat, E., Milios, E.: Vector embedding of wikipedia concepts and entities. arXiv preprint arXiv:1702.03470 (2017)
18. Yoshida, M., Ikeda, M., Ono, S., Sato, I., Nakagawa, H.: Person name disambiguation by bootstrapping. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 10–17. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835454