UNED PanMorCrepsTeam at M-WePNaD

Pablo Panero, Manuel Moreno[1], Tomás Crespo[2],
Jorge Carrillo-de-Albornoz[3], and Enrique Amigó[3]

[1] Junta de Andalucía, mmorenomaldonado@gmail.com
[2] Salenda Software Factory, t.crespo.g@outlook.com
[3] NLP&IR Group, UNED, {jcalbornoz,enrique}@lsi.uned.es

Abstract. This paper describes the participation of the PanMorCresp team in the Multilingual Web Person Name Disambiguation task of IberEval 2017. The solutions consisted of different variants of the traditional hierarchical agglomerative clustering algorithm. The four approaches have been defined and implemented independently by three Master's students over the same vocabulary generation software. The purpose of this is to analyse to what extent the HAC design and implementation can affect the effectiveness of clustering. Using a simplistic approach based on hierarchical agglomerative clustering method, considering just word occurrence, is able to achieve relatively good results regarding the rest of systems presented in the campaign.

1 Introduction

This work proposes a new people name disambiguation system used in the Multilingual Web Person Name Disambiguation task of IberEval 2017. In this task we receive a set of training web pages and an associated gold standard with the grouping of these web pages according to the different individuals they refer to. The goal is to group a new set of web pages belonging to a test set data, where no information about the correct grouping is provided.

It is usual to search information on the Web about people, where the query that expresses the information need is a person name. Because different people in the world share the same name, the results returned by a search engine can contain web pages related to several persons, not only for the searched individual. For this reason this task is really interesting, especially because of the multilingual nature of the Web. Despite this, the previous campaigns dedicated to this task focus only on corpora with web pages in a single language, for instance, the WePS campaigns [4], [3] and [2] in English, and a Chinese campaign [5]. The objective of the MWepDNaD task is providing a chance to develop person name disambiguation systems, with the additional challenge that results for a query, as well as each individual, can be written in multiple languages.

In this work, using the same vector generation software, three Master's students have implemented independently the HAC agglomerative clustering methods taking different decisions about the vocabulary size, linkage and stop criterion. The purpose of this is to check to what extent the implementation of

the HAC algorithm and the related decisions can affect the effectiveness of approaches.

The rest of the paper is organised as follows: our proposed methods to disambiguate person names are described in Section 2. Section 3 present the results obtained by our proposals, and the analysis and discussion of them can be found in Section 4. Finally, Section 5 presents the conclusions.

2 Methods

In a first step we transform each document into a vector of values which is used as input for the hierarchical agglomerative clustering algorithm. To this aim each document is divided in tokens by just splitting the text by blank characters. After this, each token is transformed into a lowercase representation in order to avoid ambiguity and decrease the number of words in the dictionary for the vector representation. The vocabulary is generated independently for each person name (query). Finally, all words with frequency one in the corpus for the corresponding query were removed. Due to computational constrains, for each entity only the most frequent n words in the dictionary generated in the previous step were selected. The experiments include variants for several n values. In all runs, we have used the presence of words as projection function.

The four approaches have been defined and implemented independently by three Master's students over the same vocabulary generation software. The evaluated approaches in the competition are:

- **PanMorCresp_Team - run 1:** The vocabulary contains the 4000 most frequent terms.The feature projection is the word occurrence. The HAC algorithm works under the complete linkage (maximum distance between items from both clusters), and the used similarity criterion is the cosine. As stop criterion, it considers a similarity threshold, which is adapted for each clustering case. That is the average similarity between documents divided by n. Several n values have been checked over the training corpus. Finally, we set the n parameter at two.
- **PanMorCresp_Team - run 2:** The vocabulary generation criterion is the same as in the previous approach. In this case, the employed linkage is the average similarity between documents in both clusters. The similarity threshold was tunned over the training corpus.
- **PanMorCresp_Team - run 3:** As well as in the previous approaches, it uses cosine similarity. For this run, we have eliminated stopwords and puntctuation marks. The vocabulary contains the 7500 most frequent terms in the collection. It uses the single linkage and the stop criterion is based on similarity (0.65). The similarity threshold was tunned over the training corpus.
- **PanMorCresp_Team - run 4:** This approach is analogous than the previous one, but using 9 clusters as stop criterion.

3 Results

We have submitted 4 runs and the results are shown in Tables 1 and 2. We report the following metrics: Reliability (R), Sensibility (S) and their harmonic mean $F_{0.5}(R, S)$ [1]. The final value of the evaluation will be the average of $F_{0.5}(R, S)$ in all person names evaluated.

Table 1 shows the results achieved by our methods considering in the evaluation only related web pages, and Table 2 shows the results considering all web pages. In addition, both tables show the result of the two baselines provided by the organizers: *One-in-one*, where every Web page is assigned to a different cluster, and *All-in-one*, where all Web pages are assigned to a single cluster.

Table 1. Results for the clustering task considering only related web pages. The run name is the name in official evaluation results.

Run	**R**	**S**	$F_{0.5}(R, S)$
ALL-IN-ONE	0.47	0.99	0.54
PanMorCresp_Team - run 1	0.80	0.51	0.43
PanMorCresp_Team - run 2	0.50	0.65	0.41
PanMorCresp_Team - run 3	0.53	0.82	0.47
PanMorCresp_Team - run 4	0.53	0.87	0.57
ONE-IN-ONE	1.0	0.32	0.42

Table 2. Results for the clustering task considering all web pages. The run name is the name in official evaluation results.

Run	**R**	**S**	$F_{0.5}(R, S)$
ALL-IN-ONE	0.47	1.0	0.56
PanMor_Team - run 1	0.79	0.46	0.40
PanMor_Team - run 2	0.49	0.62	0.43
PanMor_Team - run 3	0.53	0.81	0.5
PanMor_Team - run 4	0.52	0.86	0.58
ONE-IN-ONE	1.0	0.25	0.36

4 Discussion

The results suggest that increasing the vocabulary (from 4000 to 7500 words) increases substantially the effectiveness of the algorithm, as well as using single linkage instead of other approaches such as the average linkage or complete linkage (runs 3 and 4 vs. runs 1 and 2). However, notice that the first run

(using complete linkage and the average cost function value as stop criterion) achieves a high Reliability (precision) value. In fact, there exists a strong trade off (Reliability vs. Sensitivity) between the first run and the rest. Therefore, the results cannot be compared objectively. They depend to a great extent on the relative weight of Reliability and Sensitivity in the F measure.

On the other hand, Run 4 outperforms substantially the third run: from 47 to 57 when considering only related documents and from 0.5 to 0.58 in F when considering all documents. This improvement is mainly due to an increase in Sensitivity (recall). That is, using 9 clusters as stop criterion captures more relationships than using a similarity threshold without penalising the precision (Reliability).

5 Conclusions

We have presented in this paper the evaluation of four different runs in the Web Person Name Disambiguation task of IberEval 2017. The most remarkable result is that using a simplistic method (word occurrences, HAC, single linkage an number of clusters as stop criterion, is able to achieve an effectiveness which is (relatively) comparable with the best approach presented in the campaign. In fact, other more sophisticated approaches produce lower evaluation results.

The approaches based on HAC have been designed and implemented independenly by different students. This experiment also suggests that, considering the HAC algorithm, its effectiveness is highly sensitive to the decisions about linkage, vocabulary size and stop criterion, as well as the relative weight of the complementary evaluation metrics (reliability and sensitivity).

References

1. Emrique Amigó & Julio Gonzalo & Felisa Verdejo. A General Evaluation Measure for Document Organization Tasks. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), pp. 643-652. (2013)
2. Javier Artiles & Andrew Borthwick & Julio Gonzalo & Satoshi Sekine & Enrique Amigó. WePS-3 Evaluation Campaign: Overview of theWeb People Search Clustering and Attribute Extraction Tasks. In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010 (2010).
3. Javier Artiles & Julio Gonzalo & Satoshi Sekine. Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.
4. Javier Artiles & Julio Gonzalo & Satoshi Sekine. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 6469, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

5. Ying Chen & Peng Jin & Wenjie Li & Chu-Ren Huang. The Chinese Persons Name Disambiguation Evaluation: Exploration of Personal Name Disambiguation in Chinese News. In CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 346-352. (2010)
6. Soto Montalvo & Raquel Martínez & Leonardo Campillos & Agustín D. Delgado & Víctor Fresno & Felisa Verdejo. MC4WePS: a multilingual corpus for web people search disambiguation, Language Resources and Evaluation. (2016)