UNED Loz_Team at M-WePNaD

Lucía Lozano[1], Jorge Carrillo-de-Albornoz[2], and Enrique Amigó[2]

[1] VASS Consultora de Sistemas, lucia.lozano90@gmail.com
[2] NLP&IR Group, UNED, {jcalbornoz,enrique}@lsi.uned.es

Abstract. This paper describes the participation of the Loz team in the Multilingual Web Person Name Disambiguation task of IberEval 2017. The solutions consist of diferent variants of the traditional hierarchical agglomerative clustering algorithm. Tha analysis of results gives information about the relative effectiveness of considering different feature projections (word presence, term frequency and if.idf).

1 Introduction

Users search information on the Web using search engines, and frequently the need of information is about people. As different people share the name, the problem of disambiguating people names consists in grouping the results of a web search engine according to the different individuals they refer to. This problem has been studied extensively, but usually in a monolingual context, with all the web pages results written in the same language. Previously, WePS (Web People Search) evaluation campaigns proposed this task in a web searching scenario providing several corpora for evaluating the results of their participants, particularly WePS-1 [4], WePS-2 [3] and WePS-3 [2] campaigns, but always these corpora are monolingual. However, when a user gives a query consisting of a person name to a search engine, it returns web pages in different languages. For this reason the Multilingual Web Person Name Disambiguation task (M-WePNaD) provides the participants with a multilingual corpus (MC4WePS [5]) of web pages in order for them to develop new systems for disambiguation of person names in a multilingual context In this work we propose a basic strategy consisting of an agglomerative clustering under different feature projections: occurrence, tf and tf.idf.

The rest of the paper is organized as follows: Section 2 describes our proposed methods to disambiguate person names. The results obtained are presented in Section 4 and their analysis and discussion are in Section 5. Finally, Section 6 draws the conclusions of the work.

2 Methods

2.1 Feature Extraction

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 into tokens by just splitting the text by spaces. 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. Finally, all words with a frequency of 1 were removed. Due to computational constraints, for each entity only the 2000 most frequent words in the dictionary generated in the previous step were selected. In order to transform each document into a vector of values we have proposed three state of the art approaches based on term frequency.

- Presence: for each document the resulting vector contains a value of 0 if the document does not contain the word of the dictionary in the position i of the vector, and a value of 1 if the document contains the word.
- Frequency: for each document the resulting vector contains a value of 0 if the document does not contain the word of the dictionary in the position i of the vector, and the frequency of the word in the document if the document contains the word.
- tf/idf: for each document the resulting vector contains a value of 0 if the document does not contain the word of the dictionary in the position i of the vector, and the tf/idf value for each word if the document contains the word. The if.idf weight is computed regarding the full set of documents associated to the corresponding person name.

2.2 Similarity Measure

Using the training data set, we have compared the clustering results using euclidean distance vs. cosine similarity. The second one is the most recommended approach in the literature. The advantage of using angular distances (cosine) is that frequent words in documents are not overweighted. However, our preliminary experiments over the training data set show that the euclidean distance gives better results. It can be due to the fact that term weighting in the representation step is computed over the person name collection of document. Therefore, the normalization of frequencies in the person name domain is implicitely solved in the previous step.

2.3 Linkage and Stop Criterion

Finally, we have implemented the traditional Hierarchical Aglommerative clustering algorithm. HAC required to define the criterion to determine what pair of clusters are joined in each step (linkage) and in what moment the clustering process stops. We have experimented with single linkage (minimum distance between items from both clusters), and complete linkage (maximum distance between items from both clusters). However, the results showed that single linkage is more suitable. This could be due to the heterogeneity of documents that refer to the same person.

The most basic alternatives for the stopping criterion is a certain proximity and stating a predefined amount of final clusters (k value). We have used the second approach, testing the results with k=5 and k=15.

3 Runs

We have submitted 5 runs, with different linkages, stopping criteria and feature projections as we described in the following table:

Run	k (stop criterion)	Feature Weighting
Run 1:	k=5	Word frequency (tf)
Run 2:	k=15	Word presence
Run 3:	k=15	tf.idf
Run 4:	k=5	Word presence
Run 5:	k=5	tf.idf

In addition, the organization of the task provided two baselines:

- **One-in-one**. The baseline method where every Web page is assigned to a different cluster.
- **All-in-one**. The baseline method where all Web pages are assigned to a single cluster.

4 Results

This section reports the results of the experiments that we have performed to disambiguate person names. The metrics for evaluating the results are: Reliability (R), Sensibility (S) and their harmonic mean $F_{0.5}(R, S)$ [1]. In this task the final value of the evaluation will be the average of $F_{0.5}(R, S)$ in all person names.

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.

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
Loz_Team - run 1	0.50	0.76	0.46
Loz_Team - run 2	0.55	0.65	0.50
Loz_Team - run 3	0.57	0.71	0.52
Loz_Team - run 4	0.50	0.81	0.50
Loz_Team - run 5	0.51	0.83	0.52
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
Loz_Team - run 1	0.49	0.73	0.58
Loz_Team - run 2	0.54	0.61	0.50
Loz_Team - run 3	0.56	0.66	0.53
Loz_Team - run 4	0.49	0.78	0.52
Loz_Team - run 5	0.50	0.80	0.54
ONE-IN-ONE	1.0	0.25	0.36

5 Discussion

Interestingly, for both k=5 and k=15, we have found the same relative behavior of feature projections (presence, term frequency, tf.idf). The presence of words outperforms term frequency. This result suggests considering the frequency of terms in documents can overscore high frequent terms, even when the cosine distance mtigate this effect. These improvements across feature projections are robust across component evaluation metrics. Both Reliability and Sensitivity are improved simultaneously.

The second remarkable conclusion is that considering five clusters as stop criterion instead of 15, increases substantially the Sensitivity (recall) at the cost of a relatively small decrease in reliability. However, there is a clear trade off anyway, so the solutions are not comparable and could be highly affected by the weight of each metrics in F.

6 Conclusions

This paper describes the participation of the Loz team in the Multilingual Web Person Name Disambiguation task of IberEval 2017. The solutions consist of diferent variants of the traditional hierarchical agglomerative clustering algorithm. The results have reported information about the relative effectiveness of considering different feature projections (word presence, term frequency and if.idf). In addition, they have corroborated the sensitivity of the k value (stop criterion) to the trade of between recall and precision oriented evaluation metrics.

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. 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).