ARC '16

مؤتمر مؤسسة قطر
السنوي للبحوث
QATAR FOUNDATION
ANNUAL RESEARCH
CONFERENCE

Towards World-class
Research and Innovation

مؤسسة قطر
Qatar Foundation
لإطلاق قدرات الإنسان
Unlocking human potential

Information Communications Technology Pillar

http://dx.doi.org/10.5339/qfarc.2016.ICTPP3064

Named Entity Disambiguation using Hierarchical Text Categorization

Abdelaali Hassaine, Jameela Al Otaibi, Ali Jaoua

Qatar University, QA

Email: hassaine@qu.edu.qa

Named entity extraction is an important step in natural language processing. It aims at finding the entities which are present in text such as organizations, places or persons. Named entities extraction is of a paramount importance when it comes to automatic translation as different named entities are translated differently. Named entities are also very useful for advanced search engines which aim at searching for a detailed information regarding a specific entity. Named entity extraction is a difficult problem as it usually requires a disambiguation step as the same word might belong to different named entities depending on the context.

This work has been conducted on the ANERCorp named entities database. This Arabic database contains four different named entities: person, organization, location and miscellaneous. The database contains 6099 sentences, out of which 60% are used for training 20% for validation and 20% for testing.

Our method for named entity extraction contains two main steps: the first step predicts the list of named entities which are present at the sentence level. The second step predicts the named entity of each word of the sentence.

The prediction of the list of named entities at the sentence level is done through separating the document into sentences using punctuation marks. Subsequently, a binary relation between the set of sentences (x) and the set of words (y) is created from the obtained list of sentences. A relation exists between the sentence (x) and the word (y) if, and only if, (x) contains (y). A binary relation is created for each category of named entities (person, organization, location and miscellaneous). If a sentence contains several named entities, it is duplicated in the relation corresponding to each one of them. Our method then extracts keywords from the obtained binary relations using the hyper concept method [1]. This method decomposes the original relation into non-overlapping rectangles and highlights for each rectangle the most representative keyword. The output is a list of keywords sorted in a hierarchical ordering of importance. The obtained keyword list associated with each category of named entities are fed into a random forest classifier of 10000 random trees in order to predict the list of named entities associated with each sentence. The random forest classifier produces for each sentence the list of

Cite this article as: Hassaine A, Al Otaibi J, Jaoua A. (2016). Named Entity Disambiguation using Hierarchical Text Categorization. Qatar Foundation Annual Research Conference Proceedings 2016: ICTPP3064 http://dx.doi.org/10.5339/qfarc.2016.ICTPP3064.

probabilities corresponding to the existence of each category of named entities within the sentence.

Random Forest [sentence(i)]=(P(Person),P(Organization),P(Location),P(miscellaneous)).

Subsequently, the sentence is associated with the named entities for which the corresponding probability is larger than a threshold set empirically on the validation set.

In the second step, we create a lookup table associating to each word in the database, the list of named entities to which it corresponds in the training set.

For unseen sentences of the test set, the list of named entities predicted at the sentence level is produced, and for each word, the list of predicted named entities is also produced using the lookup table previously built. Ultimately, for each word, the intersection between the two predicted lists of named entities (at the sentence and the word level) will give the final predicted named entity. In the case where more than one named entity is produced at this stage, the one with the maximum probability is kept.

We obtained an accuracy of 76.58% when only considering lookup tables of named entities produced at the word level. When performing the intersection with the list produced at the sentence level the accuracy reaches 77.96%.

In conclusion, the hierarchical named entity extraction leads to improved results over direct extraction. Future work includes the use of other linguist features and larger lookup table in order to improve the results. Validation on other state of the art databases is also considered.

Acknowledgements

This contribution was made possible by NPRP grant #06-1220-1-233 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Reference

[1] A. Hassaine, S. Mecheter, and A. Jaoua. “Text Categorization Using Hyper Rectangular Keyword Extraction: Application to News Articles Classification”. Relational and Algebraic Methods in Computer Science. Springer International Publishing, 2015. 312–325.