Entity Recognition and Language Identification with FELTS

Pierre Jourlin

Laboratoire d'Informatique, Université d'Avignon, 84911 Avignon, France
Pierre.Jourlin@univ-avignon.fr

1 Introduction

This working notes describe the experiments we conducted in the *Microblog Cultural Contextualization Lab* [2] of CLEF 2017 [3]. The microblog data is composed of very short texts, with very heterogeneous styles. Some of them are written in more than one language. We decided to takle the entity recognition problem by using a non-statistical, dictionary-based, multiword term extractor. On the other hand, our participation in the language identification task is based on word and character *uni-gram* probabilities.

2 Task 1.5: Entity Recognition

In order to address the entity recognition problem, we make use of a free software that we developed in 2012 : FELTS (for Fast Extractor for Large Term Sets)[1]. It was designed to support very large multi-word term dictionaries such as the list of Wikipedia page titles. Using the Wikipedia's database dumps of march 1st 2017, we were able to provide FELTS with a corpus of :

- 14,971,916 distinct terms, containing 4,811,345 distinct words for English.
- 3,384,979 distinct terms, containing 1,390,569 distinct words for French.
- 2,910,899 distinct terms, containing 978,297 distinct words for Spanish.
- 1,787,280 distinct terms, containing 800,612 distinct words for Portuguese.

In order to obtain a good level of efficiency, our approach is based on *Minimal Perfect Hash Function*, more specifically, the *Compress, Hash and Displace* algorithm[1], as it was implemented in the *C Minimal Perfect Hashing Library (CMPH) V2.0*[2] in 2012.

We processed the 63,192,980 micro-blog messages of task 1 with a 64 bits personal computer equipped with a Intel Core i7-2600 (an octo-core CPU running at 3.40GHz) and 7,8 Gb of RAM. The English term corpus and the associated hash function needed 3.6Gb of RAM. It took less that half a second to extract the 20,665 English terms contained in the 1095 task 1 "topics" and less than

[1] https://github.com/jourlin/FELTS
[2] http://cmph.sourceforge.net/

8 hours to extract the 1,2 billion English terms contained in the 63 millions of micro-blog messages.

With such an approach, we found it difficult to choose a relevance score for each entity-language pair. Our system simply finds or does not find a Wikipedia entity in a text. However, we believe longer entities are more likely to indicate narrower senses and more relevant topics than shorter entities. We thus decided to simply score the multi-word terms with their character length. For each text of the test data, and each of the 4 languages, we provided the assessors with the 10 longuest extracted entities, ranked by decreasing character length. At the time this paper was written, we were not provided with relevance scores.

3 Task 1.2: Language identification

As an exploratory approach, we used a proven technique for language identification on long text : probabilistic decision based on word uni-grams. The probabilities of a language given a word were computed on two distinct corpora : The Wikipedia full text articles in all the 281 available languages (1st run) and the 63 millions of micro-blogs messages for task 1 (2nd run). Both corpora are very large but they both carry specific issues : high size disparities, distant language levels, multi-byte character encoding, lack of word boundaries, erroneous language identification, untranslated terms, multi-language texts, etc. As we realised that a word-based approach was bound to fail on languages such as Japanese or Korean were word boundaries are not explicit, we submitted a 3rd run based on character uni-gram probabilities.

We were provided with a partial manual evaluation of our first run. For only 121 out of 1095 microblog messages, our firt run identified a different language than the *locale* configuration of its author. For 90 of these 121 messages, the language identified by our 1st run was evaluated as correct. 11 of the remaining 31 erroneous identifications occurred on Japanese or Korean texts mixed with english multiword names. The last 20 erroneous indentifications are rather difficult to analyse and various causes such as the co-occurence of original and tanslated named entities can be suspected. Our third run (character uni-grams) seems to be slighlty better for Japanese or Korean but it still mostly fails on multi-language messages and is very weak at classifying languages that shares a same root. Our second run (word uni-grams according the author's locale configuration) found the correct language for 6 of 31 messages for which our first run failed. However, it is overall weaker than our first run.

4 Conclusion

The language identification results look very promising. However, we believe that there is still room for improvement and that a combination of several methods, and a specific processing of named entities could help.

References

[1] Botelho, F. C., Belazzougui, D. and Dietzfelbinger, M. Compress, hash and displace. In Proceedings of the 17th European Symposium on Algorithms (ESA2009) Springer LNCS 5757, 682-693 (2009)

[2] Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.-Y., and SanJuan, E. CLEF 2017 Microblog Cultural Contextualization Lab Overview International Conference of the Cross-Language Evaluation Forum for European Languages Proceedings Springer LNCS volume, Springer, CLEF 2017, Dublin.

[3] Jones, G. J. F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings. Springer LNCS 10456.