Mining Maximal Frequent Sequences for Multilingual Multi-Word Expression Extraction*

Antoine Doucet
University of La Rochelle
L3i Laboratory
17042 La Rochelle, France
firstname.lastname@univ-lr.fr

1 Introduction

An interesting approach to the problem of the automatic extraction of multilingual multi-word expression (MWE) resides in procedures that do not take language specifics into account. This paper summarizes techniques for the extraction and selection of maximal frequent sequences (MFS) from text and summarizes the current results. Its conclusion outlines unexploited hybridization potential.

2 Maximal Frequent Sequences

Definition. Maximal frequent sequences [1], as their name suggests, are sequences of items that occur more often than a fixed threshold of occurence σ. An MFS of length n is deemed maximal if there is no way to expand it to an $(n+1)$-sized sequence of frequency at least equal to σ. This property implies that MFSs form compact document descriptors.

It is important to underline the key difference between MFS and other comparable descriptors. A sequence being an ordered set of items, it can be gapped: other words may occur between the constituents of a multi-word compound. This implies that "set on fire" can be extracted from the sentence "The farmer set the impressive haystack on fire", without any linguistic analysis.

The latest algorithm for the extraction of MFS from text [2] is proven to be able to deal with document collections of any size, thanks to a divide and conquer approach.

A statistical selection procedure was designed, so as to efficiently compute the expected frequency of a sequence of items, assuming random distribution. By comparing expected and observed frequencies, we could evaluate the statistical interestingness of sequences and rank them accordingly. The most meaningful word compounds came up on top of the ranking [3]. An interesting property of this technique is its ability to interrank sequences of different

*Main working group concerned: "Statistical, Hybrid and Multilingual Processing of MWEs" (WG3)

lengths, whereas most probability-based measures strictly favor longer sequences, regardless of their content.

The MFS extraction algorithm was evaluated both directly and indirectly (through an information retrieval task) on distinct languages such as French, English, Japanese, Korean, and simplified Chinese [4]. For practical reasons, up to this day, there were unfortunately no experiments of hybridization with state-of-the-art approaches in MWE extraction.

3 Conclusion and Future Work

Maximal Frequent Sequences have proven their potential for the purpose of MWE extraction and selection, in a multilingual (or rather, *alingual*) context. This constant and resource-free approach has been shown to work for languages with different alphabets, from agglutinating and isolating linguistic families, both for technical and non-technical text.

We believe that there is consistent and vastly unexplored potential for the hybridization of such language-agnostic mining techniques with state-of-the-art approaches in (language-specific) MWE extraction. Mining MFS may for instance allow to filter out poor MWE candidates without requiring a prior (and often, language-specific) linguistic analysis, which would allow to process text more efficiently. Another sample approach is the combination of MFS with POS patterns, which shall allow to greatly improve the precision of MWE extraction using the MFS framework. Finer hybridization approaches shall be designed in relation with state-of-the-art techniques, following a case by case analysis.

References

[1] Helena Ahonen-Myka. Mining all maximal frequent word sequences in a set of sentences. In *Proceedings of the 2005 International Conference on Information and Knowledge Management, poster session, October 31 - November 5, 2005, Bremen, Germany*, pages 255–256. ACM, 2005.

[2] Antoine Doucet and Helena Ahonen-Myka. Fast extraction of discontiguous sequences in text: a new approach based on maximal frequent sequences. In *Proceedings of IS-LTC 2006, Information Society - Language Technologies Conference*, pages 186–191, Ljubljana, Slovenia, October 2006.

[3] Antoine Doucet and Helena Ahonen-Myka. Probability and expected document frequency of discontinued word sequences, an efficient method for their exact computation. *Traitement Automatique des Langues (TAL)*, 46(2):13–37, 2006.

[4] Antoine Doucet and Helena Ahonen-Myka. An efficient any language approach for the integration of phrases in document retrieval. *International Journal of Language Resources and Evaluation*, 44(1-2):159–180, 2010.