Mining Maximal Frequent Sequences for Multilingual Multi-Word Expression Extraction

Antoine Doucet
L3i Laboratory
University of La Rochelle
F-17042 La Rochelle, France

Maximal Frequent Sequences

Maximal frequent sequences, as their name suggests, are **sequences** of items that occur more **frequently** 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. The maximality property implies that MFSs form compact document descriptors.

Example

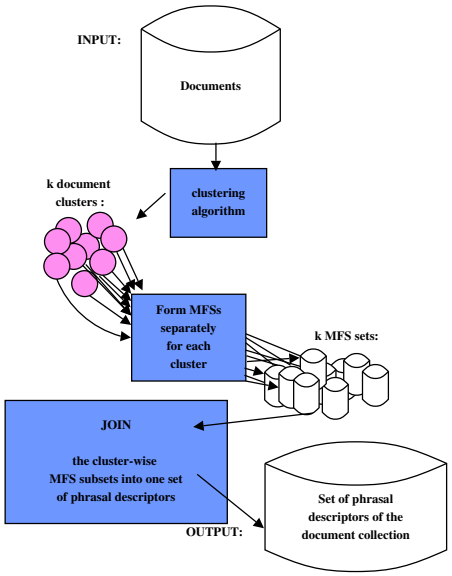
A key difference between MFS and other comparable descriptors is their sequential nature. Since sequence being an ordered set of items, it can be **gapped**: other words may occur between the constituents of a multi-word compound.

- The local farmers **set** a big haystack **on fire**.
- You don't have to **set** the world **on fire**, just do a good job.
- Antoine Doucet's poster didn't exactly **set** me **on fire**, but it's a good summary of his project.

→ a maximal frequent sequence "**set on fire**" is extracted, **without any linguistic knowledge**.

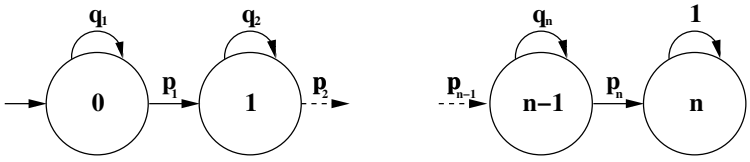
Extraction

The latest algorithm for the extraction of MFS from text [1] can deal with document collections of any size, thanks to a divide and conquer approach that divides document collections into disjoint homogeneous subsets. This process was proven to maintain the informativeness of the resulting sequence sets.



Selection

Naturally, a lot of noise stems from the extraction step. We designed a statistical selection procedure, so as to rank the extracted sequences by their (statistical) interestingness. We developed a technique to efficiently compute the probability of occurrence of word sequences, using a Markov model where states correspond to the number of words already seen in a sequence of length n, and transitions are achieved following the probability of the next expected word.



By comparing the expected and the observed frequencies of sequences, we could evaluate the statistical interestingness of sequences and rank them accordingly. The most meaningful word compounds came up on top of the ranking [2].
An interesting property of this technique is its ability to **interrank sequences of different lengths**, whereas most probability-based measures strictly favor longer sequences, regardless of their content.

Sample extracted MFSs

With the Reuters 21578 corpus, which contains about 19,000 non-empty documents, we split the collection in 106,325 sentences, and with a minimal sentence frequency σ=10, we obtained 4,855 MFSs, distributed in 4,038 2-grams, 604 3-grams, 141 4-grams, and so on. The longest sequences contained 10 words. The scores and the ranking were computed in 31.425 seconds on a laptop with a 1.40 Ghz processor and 512Mb of RAM.

rank	MFS	score
1	los(127) angeles(109)	.0311
2	kiichi(88) miyazawa(184)	.0282
3	kidder(91) peabody(94)	.0274
5	latin(246) america(458)	.0249
...	...	...
10	chancellor(120) exchequer(100) nigel(72) lawson(227)	.0232

Table: Best-ranked MFSs

While most of the descriptors are bigrams (4,038 out of 4,855), the 604 trigrams are ranked between the 38^{th} and $3{,}721^{st}$ overall positions. For the 141 4-grams, the position range is 10–3,508. This confirms the interranking of sequences of different lengths.

Applications

The MFS extraction algorithm was evaluated both directly and indirectly (through information retrieval tasks) on distinct languages such as French, English, Japanese, Korean, and simplified Chinese [3].

Summary and hybridization potential

Maximal Frequent Sequences have proven their potential for the purpose of MWE extraction and selection, in a multilingual (or rather, *alingual*) context. This resource-free approach has been tested with different alphabets, for agglutinating and isolating languages, 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. In our opinion, MFS can be used in at least two ways:

1. **As a prior**, so as to efficiently discard sentences that are unlikely to contain MWEs. The MFS extraction and selection shall save the execution of language-specific state-of-art techniques and save costly linguistic processing.
2. **As an integrated layer** of the extraction process, for instance, by POS-tagging terms before the MFS extraction step, and use syntactic patterns for filtering.

Finer hybridization shall be designed in relation with state-of-the-art techniques, **following a case by case analysis**. Feel free to get in touch!

Further details

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

[2] 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** (2006), no. 2, 13–37.

[3] 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** (2010), no. 1-2, 159–180.