Local function words discovery in raw corpora of unknown languages, without any resource

… (concealed for blind reviewing)
…
…
…
…
…
…

ABSTRACT

Our present research is in the field of exploring NLP methods which use no other resource than the text to analyse itself. This drives us to analysis methods which use very general linguistic properties, as for instance differences of length and frequencies of words. To illustrate our approach, we present in this paper a local computation method to discover function words from raw corpora. This method can be used to extract term candidates or to index raw texts in unidentified alphabetic natural languages.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**] Information Search and Retrieval.

I.2.6 [**Artificial Intelligence**] Learning – *Induction, Language acquisition*

I.2.7 [**Artificial Intelligence**] Natural Language Processing – *Text analysis*

General Terms

Algorithms, Experimentation, Human Factors, Languages.

Keywords

Multilingual Retrieval, Machine Learning for IR, Text Data Mining, Natural Language Processing for IR, Text Indexing, Multilingual NLP, Natural Language Learning, Function Words Discovery.

INTRODUCTION

This work takes place in the frame of exploring linguistic processing without any other resource than the text to analyse itself. In this case, the absence of resource allow to envisage that the same parser will be able to process texts in different languages, without any language identification. Of course such a parser has to exploit very general linguistic properties, and not properties which are local to one language, as a monolingual dictionary. Therefore, this exploration into processing goes with an exploration into common linguistic properties to a group of languages.

In this paper, we present a method for discovering function words, which can take place for instance within a task of automatically indexing a document, or for extracting candidate terms from a corpus [14]. Such tasks need to locate frequent nominal expressions. Usually, a word tagging is done, or a morpho-syntactic parsing [2], or a list of function words is used, to select frequent segments which are **not** function words [10, 11, 1]. In both cases, the corpus is supposed to be monolingual and the language to be identified. We present a method without any linguistic resource, working on monolingual or multilingual written raw corpora, in non identified alphabetic languages.

The task that we propose is to take such a corpus, and to discover function words which are **present in this corpus**. The more general task of discovering syntactic structures from raw corpora has already been explored (and partially solved) by Hervé Déjean [4], and this work now is continued in the trend of "Grammatical Induction of languages" (or Grammar Induction, Grammatical Inference, or Grammar Inference[1] [6]).

FUNCTION WORDS, STOPLIST AND STOPWORDS

In French, the terms: "*mot vide*" ("empty" word) and "*mot plein*" ("full" word) are often used for "function word" and "content word", following Lucien Tesnière ([13], p. 53) and Fathi Debili [3].

In the French information retrieval tradition, the definition is more extensive: a "*mot vide*" is a word which has not to be indexed, whatever it is a function word or a non discriminating content word (main topic in a document data base). Function words are often placed in an "*anti-dictionnaire*" or a "stoplist" or a list of "stopwords". It is generally accepted that these very frequent words (about half of the occurrences of a text) are not to be indexed, because they are not informative, and they tremendously increase the index size in case of a fulltext indexing, which is the common case. On the contrary, some authors advocate indexing function words, as they can be informative, prepositions for instance [8].

In the literature, we have not found any discovering method of function words in raw corpora (except the one of Hervé Déjean, founded on discovering morphemes and their positional statistical study). But, there are some works where a list of function words is placed as input of a process:

- For Wilbur and Sirotkin, in their paper "*The automatic identification of stop words*" [15], a "stop word" is any non informative word in a collection of documents. Their

[1] See the "Grammatical Induction Community website": `http://eurise.univ-st-etienne.fr/gi/`

algorithm uses a "*standard stopword list*" as resource. Yiming Yang [16] describes their method: "*In contrast to using generic stop words, Wilbur and Sirotkin developed a novel stopword identification method which allows a far more aggressive removal of words from documents without losing retrieval accuracy*."

- Tin Kam Ho, in his paper "*Fast Identification of Stop Words for Font Learning and Keyword Spotting*" [7], has to main aim to improve font learning in an OCR system, while concentrating the beginning of this learning process on function words, and while trying to match written forms of function words placed as input, with images of short words delimited in the document image. Let us notice that the author uses the length of words images ("*word width*") as discriminating criterion between function and content words, but without citing Zipf.
- Andrew Roberts, in "*Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words*" [9] suggests "*a method which can automatically infer word classification*" from a list of function words placed as input. He mentions that he could obtain them automatically, and by the way, he cites Eliott's method [5], of the same laboratory: "*Elliott has found that function words can be obtained by combining relatively small samples from at least 3 sources*". The (surprising) aim of the authors is to detect a linguistic material into signals coming from space, and, for them, this task includes to discover function words; to do that, they suggest to use the "*least inter-text variation*" of function words, from the (non validated) hypothesis that function words are more stable than content words in thematically different corpora.

LINGUISTIC PROPERTIES

In this work, we use very general linguistic properties to categorise words as function words or content words: properties of word length and word frequency, and properties of differences of these criteria between two contiguous words:

- A Zipf's fundamental observation is that content words are rare and long and that function words are **more frequent and shorter**; what is frequent is short: it is the law of the least effort in the use of a code, characterised by Zipf [17, 18], and it is also observable in programming languages (let us notice that the "Zipf law", still now very frequent in the literature, is usually presented as a law on word frequencies only, and that statistical properties of word lengths are more rarely invoked).
- While applying the propriety that Saussure stated: "*dans la langue, il n'y a que des differences*" ([12] p.166), we shall base our computation upon **local differences** of lengths and frequencies of contiguous words.

Another very general linguistic property drives us to find a method which does not use a "stoplist": the same written form can code two different **homograph** words in different contexts, and these two words can have to be categorised one function word, and the other content word. Such written forms are frequent, for instance *like* in English, or *car*, *or*, *la*, *son*, *une* (in our news websites corpus), *pendant*, *avions* in French. If we trust the absolute values of written forms of a stoplist, these words will not be indexed, hence a systematic silence on these content words. On the contrary, we must trust the context while making a **local computation**, which can give different results for the same written form, according to the context.

DISCOVERING FUNCTION WORDS WITH A LOCAL PROCESS

A preliminary step computes the frequency (in the corpus) of every written form of the corpus. Then, the corpus is divided into segments which do not contain any punctuation.

Every segment is processed the following way:

- it is divided into words;
- then a **paving** of the segment is found with the following patterns: `CfffC`, `CffC`, `CfC`, `fCf`, `fCCf`, `fCCC]`, `ffC]`, `fCC]`, `[fffC`, `[ffC`, `[fC`, where `f` = **f**unction word, `C` = **C**ontent word, `]` = at the end of a segment, `[` = at the beginning of a segment.

To get a linear complexity, we have chosen a deterministic paving: several pavings would be possible, but the first valid pattern is accepted, and this only solution is computed. Patterns are tried in the order above. Patterns can overlap on 1 or 2 words. If a pattern applies, words it contains take the corresponding category of the pattern. If no pattern applies, words it contains stay indeterminate.

A pattern is validated by the following rule: in the pattern, the set of function words must be different from the set of content words; it means that for every criterion (length in number of letters and frequency), the geometric average of minimal and maximal values of the criterion separates function words from content words in the pattern.

Example of validating a pattern at the beginning of a segment:

```
0  2-189 le
1  6-41  nombre   (for every word: length-frequency)
2  3-384 des
```

- the pattern `CffC` is not valid because there is no difference content word - function word on the words `0-1`, because we have: 2<6 (on lengths)

- the pattern `CfC` is not valid for the same reason

- validation of the pattern `fCf`:

  . there is a cut between content and function words according to frequencies:

  geometric average (41; 384) = **125.46**
  `41 (C) < 125.46 < 189 (f) 384 (f)`

  . there is a cut between content and function words according to lengths:

  geometric average (2; 6) = **3.46**
  `6 (C) > 3.46 > 3 (f) 2 (f)`

Therefore, the pattern `fCf` is valid, and the 3 occurrences are categorised function - Content - function words:

```
0   f   2-189   le
1   C   6-41    nombre
2   f   3-384   des
```

Here is an example of result for a segment:

```
---CC-C    C is imposed  by the written form (length > 6)
fCf              validated
.CfC                patterns
..fCCf                              while paving
....CfC
fCfCCfC   paving result => local category
                                 for every word occurrence
```

```
0   f   2-189   le
1   C   6-41    nombre       (length-frequency)
2   f   3-384   des
3   C   7-10    lycéens
4   C   9-5     préparant
5   f   2-75    un
6   C   12-21   baccalauréat
```

The complexity of the algorithm is linear in time according to the number of words of the corpus.

RESULTS AND EVALUATION

Here are examples of results on three monolingual corpora (English, French and German) of about the same sizes (in Kb) and of the same genre (texts of webpages, automatically downloaded by newswebsites crawling). We present a comparative evaluation on the three corpora, and we give some examples of homograph words of different categories.

1.1 English monolingual corpus

In these two segments, "like" is a function word or a content word:

```
0   f   4-27   like   (here, "like" is a function word)
1   C   6-1    bamboo
2   C   6-1    shoots
3   f   5-11   after
4   f   1-252  a
5   C   6-1    spring
6   C   4-1    rain
```

```
0   f   3-33   But
1   f   2-37   we
2   C   4-27   like   (here, "like" is a content word)
3   f   2-289  to
4   C   3-4    buy
5   -   5-16   those   (indeterminate occurrence)
6   C   10-8   businesses
7   f   2-249  in
8   f   1-252  a
9   C   10-1   contrarian
10  C   7-1    fashion
```

1.2 French monolingual corpus

In these two segments, "une" is a content word or a function word:

```
0   C   10-12   ACTUALITES
1   f   1-201   à
2   f   2-323   la
3   C   3-119   une   (here, "une" is a content word)
```

```
0   f   3-273   les
1   C   8-3     Français
2   C   11-1    rencontrent
3   f   6-15    samedi     (noise on function words)
4   f   3-119   une    (here, "une" is a function word)
5   C   6-1     équipe
6   f   2-140   du
7   C   6-3     Canada
8   f   1-117   a
9   C   6-1     priori
10  f   1-201   à
11  f   2-16    sa
12  C   6-2     portée
13  f   2-196   et
14  C   8-1     composée
15  f   2-143   en
16  C   8-12    majorité
17  f   2-531   de
18  C   7-3     joueurs
19  C   5-3     ayant      (silence on function words)
20  f   4-22    fait       (noise on function words)
21  f   2-206   le
22  C   5-2     choix
```

1.3 German monolingual corpus

```
0   f   3-144   Die
1   C   6-2     zweite
2   C   8-1     Vorrunde
3   f   3-234   der
4   C   19-1    Ausscheidungsrennen
5   f   3-20    zum
6   C   12-1    diesjährigen
7   C   7-3     America
8   f   1-13    s
9   C   3-4     Cup
10  f   3-39    ist
11  f   2-66    zu
12  C   4-7     Ende
13  C   8-3     gegangen
```

1.4 Comparative evaluation on the three English, French and German corpora

Let us define the global rate of determination: 1 - (number of occurrences of non categorised words / total number of occurrences).

The evaluation is done from the point of view of the extraction of function words. So, we will define a local occurrence of a function word the following way: a function word is not a content word; a content word is a noun, a non numeral adjective, a non auxiliary and non modal verb, or an adverb.

As the same written form can be differently categorised in different contexts, only the evaluation on numbers of occurrences has a meaning (and not the evaluation on numbers of different written forms):

precision = 1 - noise = 1 - (number of occurrences of content word categorised function words / number of occurrences of extracted function words)

recall = 1 - silence = 1 - (number of occurrences of function word not categorised function words / total number of occurrences of function words).

Table 1. Comparative evaluation on the three English, French and German corpora

language	news website (23/11/02)	size	determination	precision	recall	lengths	
						average	standard deviation
English	The International Herald Tribune	76 Ko, 12 501 words	90,8%	92,6%	78,2%	4,51	2,73
French	Le Monde	82 Ko, 12 348 words	94,2%	96,3%	85,8%	4,96	3,14
German	Der Spiegel	81 Ko, 9 897 words	92,8%	90,8%	88,7%	5,99	3,71

In the English corpus, we observe that words are shorter and their lengths are not very dispersed, while in the German corpus, words are longer, and lengths more dispersed (compound words). We also can observe a correlation between recalls and length standard deviation. Hypothesis: recall is better if lengths are more dispersed, as it allows a better contrast in the validation of a paving pattern. That could explain why recall is better in German.

DISCUSSION

Let us wonder about both used criteria: frequency and length. They are global features of the written form, which have the same value for every occurrence of this written form. It is only the local sequence of some occurrences of written forms which allows to affect a local value, function or content word, to an occurrence, by a local computation on global values. About word length, Zipf [17] chooses the syllable or the phoneme as unit of the metrics, without motivating his choice; we wanted to choose a unit which is independent from languages, therefore the character, that is directly the written form, without the computation (dependent on the language) that a syllable segmentation or a phonetisation would have required. We made tests with the syllable length unit, for French, with analogous results. We also made tests with only frequency or only length, with correct but lower results, with more noise, but less silence (on function words).

To obtain a (local) detection of a function word between two content words or vice-versa, the differences must be **sufficiently contrasted**. This condition is not satisfied for about 10% of occurrences of a function word (it is the main cause of silence), when a function word is long and/or rare, or when a contiguous content word is short and/or frequent. In an subsequent step of the analysis (not described in this paper, but in progress), it is possible to locally apply that has been discovered elsewhere. It is the problem of how to globally apply some local deductions; we have to satisfy two contradictory constraints: the same form can get different categories, and be a member of different paradigms of function words (mainly if the corpus is multilingual), and it forbids a total generalisation, but it is necessary to make a partial generalisation to lower the silence for function words (this point is in progress).

On which languages this method gives good results ? Tests on other languages have been done, and they drive us to answer: this method gives good results on languages where function words are isolated words (Italian, Spanish, Swedish, Polish, Albanian for instance) and not morphemes attached to content words (agglutinative languages as Turkish, Finnish for instance). In this last case, a morpheme isolating preprocessing would be possible (as Déjean has shown).

CONCLUSION AND PROSPECTS

We presented a local computation method to discover function words in raw corpora in non identified alphabetical languages. Such a method is an example of NL processing without any other resource than the analysed corpus itself. Such a processing, that we could call "alingual", has to exploit very general proprieties of languages, proprieties that we have to explicit and validate. The good detection of function words is a clue of the generality of the linguistic proprieties we exploit. The way toward linguistic processing without resources is a promising way.

Our present work is on the parsing step, and is developed in the following directions: partial generalisation of local deductions, categorising function words, chunking, and distinguishing nominal chunks from verbal chunks. The "alingual" chunking without any resource is now possible. We also have to make tests on more languages, with a native speaker for every language to make the evaluation.

This method can be applied for automatically indexing documents in unidentified alphabetic natural languages, while properly centring every indexed term on one or several content words, without using a stoplist, and without identifying the language of the document; this sort of task will be useful for documents found on the web, where more and more different and new languages can be found.

REFERENCES

[1] Ahonen-Myka Helena. Discovery of frequent word sequences in text. *The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, Imperial College, London, 2002. `www.cs.helsinki.fi/u/hahonen/ahonenmyka_patws02.ps`

[2] Bourigault Didier. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy, 2002, 75-84.

www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc

[3] Debili Fathi. *Analyse syntaxico-sémantique fondée sur une acquisition automatique des relations lexicales-sémantiques*. Thèse de doctorat d'état en sciences informatiques, Université de Paris XI, 1982.

[4] Déjean Hervé. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Spécialité : informatique, thèse de l'Université de Caen, 1998.

[5] Elliott J., Atwell E. and Whyte B.. Language identification in unknown signals. *Proceeding of CoLing'2000, 18th International Conference on Computational Linguistics*, Saarbrücken, 2000, 1021-1026.

[6] Grammatical Induction Community website: http://eurise.univ-st-etienne.fr/gi/

[7] Ho Tin Kam. Fast Identification of Stop Words for Font Learning and Keyword Spotting. *Proceedings of the 5th Int'l Conference on Document Analysis and Recognition*, 1999.

[8] Riloff Ellen. Little Words Can Make a Big Difference for Text Classification. *Proceedings of {SIGIR}-95, 18th {ACM} International Conference on Research and Development in Information Retrieval*, 1995, 130-136.

[9] Roberts Andrew. Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words. School of Computing, University of Leeds, 2002.

[10] Salem André. *Pratique des segments répétés*. Publications de l'INaLF, collection "St.Cloud", Klincksieck, Paris, 1987.

[11] Salton Gerard and Allan James. Selective Text Utilization and Text Traversal. *In UK Conference on Hypertext*, 1993, 131-144.

[12] Saussure F. de. *Cours de Linguistique Générale*. Payot, Paris, (éd. 1974), 1922.

[13] Tesnière Lucien. *Éléments de syntaxe structurale*. Klincksieck (Paris), 1959.

[14] ... (concealed for blind reviewing)

...

...

. . .

. . .

[15] Wilbur JW, Sirotkin K.. The automatic identification of stop words. *Journal of Information Science*, 18(1), 1992, 45-55.

[16] Yang Yiming. Noise Reduction in a Statistical Approach to Text Categorization. *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, 1995.

[17] Zipf George Kingsley. *The Psychobiology of Language, an Introduction to Dynamic Philology*. Houghton Mifflin, Boston, 1935.

[18] Zipf George Kingsley. *Human Behavior and the Principle of Least Effort*. Harper, New York, 1949.