Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource

Jacques Vergne

www.info.unicaen.fr/~jvergne

GREYC – Université de Caen – BP 5186 – F-14032 Caen cedex

Abstract

Our present research is in the field of exploring NLP methods using 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 method of local computation for discovering function words from raw corpora. This method can be used for extracting term candidates or indexing raw texts of unidentified alphabetic natural languages.

Résumé

Nous nous plaçons dans une perspective de traitements linguistiques sans autre ressource que le texte à analyser. Ceci nous conduit à des méthodes d'analyse exploitant des propriétés très générales des langues, comme par exemple les différences de longueur et de fréquence des mots. Pour illustrer notre démarche, nous présentons dans cet article une méthode de découverte des mots vides par un calcul local. Cette méthode peut s'appliquer à l'extraction de candidats termes ou à l'indexation de textes bruts de langues alphabétiques non identifiées.

Mots-clés : traitements multilingues, découverte des mots vides, multilingual NLP, natural language learning, grammar induction, function words discovery.

1. Introduction

Ce travail se situe dans le cadre d'une exploration dans la direction de traitements linguistiques sans aucune autre ressource que le texte à analyser lui-même. Dans ce cas, l'absence de ressources permet d'envisager qu'un même analyseur puisse traiter des textes de langues différentes, sans identification de la langue. Bien évidemment, un tel analyseur doit exploiter des propriétés linguistiques très générales, et non pas des propriétés locales à une langue, telles qu'un lexique monolingue. Cette exploration sur les traitements s'accompagne donc aussi d'une exploration de propriétés linguistiques communes à un groupe de langues.

Dans cet article, nous présentons une méthode de découverte des mots vides, méthode pouvant par exemple prendre place dans le cadre d'une tâche d'extraction de candidats termes (Vergne, 2003), ou d'indexation automatique. De telles tâches nécessitent de repérer les expressions nominales fréquentes. Habituellement, on pratique un étiquetage des mots, ou une analyse morphosyntaxique (Bourigault, 2002), ou bien on dispose d'un «anti-dictionnaire» des mots vides pour pouvoir sélectionner les segments fréquents qui ne sont pas des mots vides (Salem, 1987 ; Salton et al., 1993 ; Ahonen-Myka, 1999). Dans les deux cas, le corpus est supposé monolingue et la langue identifiée. Nous présentons ici une méthode sans ressource linguistique, fonctionnant sur corpus écrits bruts monolingues ou multilingues, de langues alphabétiques non identifiées.

La tâche que nous nous proposons est de prendre un tel corpus, et de découvrir les mots vides **présents dans ce corpus**. La tâche plus générale de découverte des structures syntaxiques à partir de corpus bruts a déjà été explorée (et partiellement résolue) par Hervé Déjean (Déjean, 1998), et ce travail est actuellement poursuivi dans le courant de la «Grammatical Induction of languages» (ou Grammar Induction, Grammatical Inference, ou Grammar Inference[1]).

2. Mots vides, mots grammaticaux, stop-list et stopwords

Nous avons choisi les termes : «mot vide» - «mot plein», synonymes de «mot grammatical» - «mot lexical» et de «function word» - «content word», à la suite de Lucien Tesnière (Tesnière 1969, p. 53) et de Fathi Debili (Debili, 1982).

Dans la tradition de l'informatique documentaire, la définition est plus extensive : un «mot vide» est un mot qui ne doit pas être indexé, qu'il soit mot grammatical ou mot lexical non discriminant (thème commun à une base documentaire). Les mots vides sont alors souvent regroupés dans un «anti-dictionnaire» ou une «stop-list» ou une liste de «stopwords». Il est généralement admis que ces mots très fréquents (environ la moitié des occurrences d'un texte) ne sont pas à indexer, car ils ne sont pas informatifs, et ils augmentent énormément la taille de l'index si l'on réalise une indexation fulltext, ce qui est le cas courant. Des auteurs au contraire prônent l'indexation des mots vides, comme pouvant être informatifs, les prépositions par exemple (Riloff, 1995).

Dans la littérature, nous n'avons pas trouvé de méthode de découverte des mots vides dans les corpus bruts (hormis celle d'Hervé Déjean, fondée sur la découverte des morphèmes et leur étude statistique positionnelle). Par contre, on trouve des travaux où une liste de mots vides placée en entrée constitue l'amorce d'un traitement :

- Pour Wilbur et Sirotkin, dans leur article «*The automatic identification of stop words*» (Wilbur et Sirotkin, 1992), un «stop word» est tout mot non informatif dans une collection de documents. Ils amorcent leur algorithme par une «*standard stopword list*». Yiming Yang (Yang, 1995) décrit leur méthode : «*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, dans son article «*Fast Identification of Stop Words for Font Learning and Keyword Spotting*» (Ho, 1999), a pour premier objectif d'améliorer l'apprentissage des polices dans un système d'OCR, en concentrant le début de cet apprentissage sur les mots vides, et en cherchant à apparier les graphies des mots vides placés en entrée, et les images des mots courts délimités dans l'image du document. Remarquons que l'auteur utilise la longueur des images de mots («*word width*») comme critère discriminant vide - plein, mais sans citer Zipf.
- Andrew Roberts, dans «*Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words*» (Roberts, 2002) propose «*a method which can automatically infer word classification*» à partir d'une liste de mots vides placée en entrée. Il mentionne qu'il aurait pu les obtenir automatiquement, et à ce propos, il cite la méthode d'Eliott (Elliott et al., 2000), du même laboratoire : «*Elliott has found that function words can be obtained by combining relatively small samples from at*

[1] Voir le «Grammatical Induction Community website» : `http://eurise.univ-st-etienne.fr/gi/`

least 3 sources». L'objectif (surprenant) des auteurs est de détecter un matériau linguistique dans des signaux venant de l'espace, et, pour eux, cette tâche inclut la découverte de mots vides ; pour ce faire, ils proposent d'utiliser la «*least inter-text variation*» des mots vides, à partir de l'hypothèse (non explorée, ni validée) que les mots vides sont plus stables que les mots pleins dans des corpus thématiquement différents.

3. Propriétés linguistiques

Des propriétés linguistiques très générales sont utilisées pour catégoriser les mots vides ou pleins : des propriétés de longueur et de fréquence des mots, et de différences de ces critères entre deux mots contigus :

- Une observation fondamentale de Zipf est que les mots pleins sont rares et longs et que les mots vides sont **plus fréquents et plus courts** : ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf (Zipf, 1949), et observable aussi dans les langages de programmation (remarquons que la «loi de Zipf», toujours très présente dans la littérature, est une loi sur les effectifs des mots, et que les propriétés statistiques des longueurs des mots sont plus rarement invoquées).
- En application de la propriété énoncée par Saussure : «*dans la langue, il n'y a que des différences*» (Saussure, 1922, éd. 1974 p.166), nous allons fonder nos calculs sur les **différences locales** de longueur et d'effectif de mots contigus.

Une autre propriété linguistique très générale nous conduit à rechercher une méthode qui n'utilise pas de «stoplist» : une même graphie peut coder deux mots **homographes** différents dans des contextes différents, et ces deux mots peuvent devoir être catégorisés l'un vide, l'autre plein. De telles graphies sont fréquentes, par exemple en français : *car*, *or*, *la*, *son*, *une* (dans notre corpus de sites de presse), *pendant*, *avions*, ou *like* en anglais. Si l'on se fie à la valeur absolue des graphies d'une stoplist, on n'indexe pas ces mots, d'où un silence systématique. On doit au contraire se fier au contexte en faisant un **calcul local**, pouvant donner pour une même graphie des résultats différents selon le contexte.

4. Processus de découverte des mots vides par calcul local

Une étape préliminaire compte le nombre d'occurrences dans le corpus, de chaque graphie du corpus. Puis, le corpus est découpé en segments de corpus ne contenant pas de ponctuation.

Chaque segment est ainsi traité :

- il est d'abord segmenté en mots ;
- puis on recherche un **pavage** du segment avec les motifs suivants : `PvvvP`, `PvvP`, `PvP`, `vPv`, `vPPv`, `vPPPƒ`, `vvPƒ`, `vPPƒ`, `∂vvvP`, `∂vvP`, `∂vP`, où `v` = vide, `P` = plein, `ƒ` = en fin de segment, `∂` = en début de segment.

Pour avoir une complexité linéaire, nous avons choisi un pavage déterministe : plusieurs pavages seraient possibles, mais le premier motif validé est accepté. Les motifs sont essayés dans l'ordre ci-dessus. Les motifs peuvent se recouvrir sur 1 ou 2 mots. Si un motif s'applique, les mots qu'il contient prennent la catégorie correspondante du motif. Si aucun motif ne s'applique, les mots qu'il contient restent indéterminés.

Un motif est validé par la règle suivante : dans le motif, l'ensemble des mots vides doit être différent de l'ensemble des mots pleins ; c'est-à-dire que pour chaque critère (les longueurs en

nombre de lettres et les effectifs), la moyenne géométrique des valeurs minimale et maximale du critère sépare les mots vides des mots pleins du motif.

Exemple de validation d'un motif sur le début d'un segment :

```
0          2-189 le
1          6-41  nombre      pour chaque mot : longueur-effectif
2          3-384 des
```

- le motif `PvvP` est invalidé car on n'a pas la différence plein - vide sur les mots `0 1`, car on a : 2<6 (sur les longueurs)
- le motif `PvP` est invalidé pour la même raison
- validation du motif `vPv` :
  - . on a une coupe entre pleins et vides selon les effectifs :
    moyenne géométrique (41 ; 384) = **125,46**
    `41 (P) < 125.46 < 189 (v) 384 (v)`
  - . on a une coupe entre pleins et vides selon les longueurs :
    moyenne géométrique (2 ; 6) = **3,46**
    `6 (P) > 3.46 > 3 (v) 2 (v)`

le motif `vPv` est donc validé, et les 3 occurrences sont catégorisées vide - Plein - vide :

```
0     v     2-189 le
1     P     6-41  nombre
2     v     3-384 des
```

Voici un exemple de résultat pour un segment :

```
---PP-P     P est imposé par la graphie (si longueur > 6)
vPv         motifs
.PvP            validés
..vPPv              au cours du pavage
....PvP
vPvPPvP     résultat du pavage => catégorie locale pour chaque occurrence de mot
0     v     2-189 le
1     P     6-41  nombre      pour chaque mot : longueur-effectif
2     v     3-384 des
3     P     7-10  lycéens
4     P     9-5   préparant
5     v     2-75  un
6     P     12-21 baccalauréat
```

La complexité de l'algorithme est linéaire en temps selon le nombre de mots du corpus.

5. Résultats et évaluation

Voici des exemples de résultats sur trois corpus monolingues anglais, français et allemand de tailles analogues (en Ko) et de même genre (textes de pages acquises automatiquement par crawling de sites de presse), suivis d'une évaluation comparative sur les trois corpus. Nous donnons quelques exemples de mots homographes de catégories différentes.

5.1. Corpus monolingue anglais

Dans ces deux segments, «`like`» est mot vide ou mot plein :

```
0     v     4-27  like        ici, «like» est un mot vide
1     P     6-1   bamboo
2     P     6-1   shoots
3     v     5-11  after
4     v     1-252 a
5     P     6-1   spring
6     P     4-1   rain

0     v     3-33  But
1     v     2-37  we
2     P     4-27  like        ici, «like» est un mot plein
3     v     2-289 to
4     P     3-4   buy
5     -     5-16  those       occurrence indéterminée
6     P     10-8  businesses
7     v     2-249 in
8     v     1-252 a
9     P     10-1  contrarian
10    P     7-1   fashion
```

5.2. Corpus monolingue français

Dans ces deux segments, «une» est mot plein ou mot vide :

```
0     P     10-12 ACTUALITES
1     v     1-201 à
2     v     2-323 la
3     P     3-119 une         ici, «une» est un mot plein

0     v     3-273 les
1     P     8-3   Français
2     P     11-1  rencontrent
3     v     6-15  samedi      bruit sur les mots vides
4     v     3-119 une         ici, «une» est un mot vide
5     P     6-1   équipe
6     v     2-140 du
7     P     6-3   Canada
8     v     1-117 a
9     P     6-1   priori
10    v     1-201 à
11    v     2-16  sa
12    P     6-2   portée
13    v     2-196 et
14    P     8-1   composée
15    v     2-143 en
16    P     8-12  majorité
17    v     2-531 de
18    P     7-3   joueurs
19    P     5-3   ayant       silence sur les mots vides
20    v     4-22  fait        bruit sur les mots vides
21    v     2-206 le
22    P     5-2   choix
```

5.3. Corpus monolingue allemand

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

5.4. Évaluation comparative sur les 3 corpus anglais, français, allemand

Définissons le taux global de détermination : 1 - (nombre d'occurrences de mots non catégorisées / nombre total d'occurrences).

L'évaluation est faite du point de vue de l'extraction des mots vides. Pour ce faire, nous allons définir une occurrence locale de mot vide de manière opératoire : un mot vide n'est pas un mot plein ; un mot plein est un nom, un adjectif non numéral, un verbe non auxiliaire et non modal, ou un adverbe.

Comme une même graphie peut être catégorisée différemment dans des contextes différents, seule l'évaluation sur les nombres d'occurrences a un sens (et non pas l'évaluation sur les nombres de graphies différentes) :

- précision = 1 - bruit = 1 - (nombre d'occurrences de mot plein catégorisées vides / nombre d'occurrences de mot vide extraites)
- rappel = 1 - silence = 1 - (nombre d'occurrences de mot vide non catégorisées vides / nombre total d'occurrences de mot vide).

langue	**site de presse** (23/11/02)	**taille**	**détermi -nation**	**précision**	**rappel**	**longueurs**	
						moyenne	écart-type
anglais	The International Herald Tribune	76 Ko, 12 501 mots	90,8%	92,6%	78,2%	4,51	2,73
français	Le Monde	82 Ko, 12 348 mots	94,2%	96,3%	85,8%	4,96	3,14
allemand	Der Spiegel	81 Ko, 9 897 mots	92,8%	90,8%	88,7%	5,99	3,71

Dans le corpus anglais, on observe que les mots sont plus petits et leurs longueurs peu dispersées, alors que dans le corpus allemand, les mots sont plus longs et les longueurs plus dispersées (mots composés). On peut observer en outre une corrélation entre rappels et écart-types des longueurs. Hypothèse : le rappel est meilleur si les longueurs sont plus dispersées, ce qui permet un meilleur contraste dans la validation d'un motif du pavage. Ceci pourrait expliquer pourquoi le rappel est meilleur en allemand.

6. Discussion

Interrogeons-nous sur les deux critères utilisés : l'effectif et la longueur. Ce sont des caractéristiques globales de la graphie, de même valeur pour toutes les occurrences de cette graphie. C'est seulement la séquence locale de quelques occurrences de graphies qui permet d'affecter une valeur locale, vide ou plein, à une occurrence, par un calcul local sur des valeurs globales. Au sujet de la longueur d'un mot, Zipf (Zipf, 1935) choisit la syllabe ou le phonème comme unité de la métrique, sans motiver ce choix ; en ce qui nous concerne, nous devons choisir une unité indépendante des langues, donc la lettre, c'est-à-dire directement la forme écrite de la graphie, sans le calcul (dépendant de la langue) qu'aurait nécessité la syllabation ou la phonétisation. Nous avons fait des tests en choisissant la syllabe, pour le français, avec des résultats analogues. Nous avons aussi fait des tests en ne prenant que les effectifs ou que les longueurs, avec des résultats corrects mais inférieurs, marqués par plus de bruit, mais moins de silence (sur les mots vides).

Pour qu'il y ait une détection (locale) d'un mot vide entre deux mots pleins ou inversement, il faut que les différences soient ensemble **suffisamment contrastées**. Cette condition n'est pas satisfaite en moyenne pour environ 10% des occurrences d'un mot vide (c'est la cause principale du silence), quand un mot vide est long et/ou rare, ou quand un mot plein voisin est court et/ou fréquent. Dans une phase ultérieure d'analyse (non décrite ici, mais en cours de réalisation), on peut appliquer localement ce qui a été découvert ailleurs. C'est tout le problème de l'application globale de certaines déductions locales ; on doit satisfaire à deux contraintes contradictoires : une même forme doit pouvoir obtenir plusieurs catégories, et appartenir à plusieurs paradigmes de mots vides (surtout si le corpus est multilingue), ce qui interdit la généralisation totale, mais il est nécessaire de pratiquer une généralisation partielle pour diminuer le silence des mots vides (ce point est à l'étude).

Sur quelles langues (alphabétiques) cette méthode donne-t-elle de bons résultats ? Les tests conduits sur d'autres langues nous conduisent à répondre : sur celles dont les «mots» vides sont des mots isolés (italien, espagnol, suédois par exemple) et non pas des morphèmes accolés à des mots pleins (langues agglutinantes telles que turc, finnois par exemple). Dans ce dernier cas, un prétraitement d'isolation des morphèmes est envisageable (comme l'a montré Hervé Déjean).

7. Conclusion et perspectives

Nous avons présenté une méthode de découverte des mots vides dans des corpus bruts de langues alphabétiques non identifiées, par calcul local. Une telle méthode constitue un exemple de traitement sans autre ressource que le texte analysé. De tels traitements, que l'on pourrait appeler «alingues», doivent exploiter des propriétés très générales des langues, propriétés qu'il s'agit d'expliciter et valider. La bonne détection des mots vides est un indice de la généralité des propriétés linguistiques exploitées. La direction des traitements linguistiques sans ressources est prometteuse.

Nos travaux actuels portent sur la phase d'analyse dans les directions suivantes : généraliser partiellement les déductions locales, catégoriser les mots vides, réaliser le chunking, et distinguer les chunks nominaux des chunks verbaux. Le chunking «alingue» sans ressources est maintenant envisageable. Nous devons aussi augmenter le nombre de langues, ce qui nécessite un locuteur de chaque langue pour l'évaluation.

Références

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

Bourigault Didier (2002). 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, pp. 75-84.
www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc

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

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

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

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

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

Riloff Ellen (1995). 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*, 130--136.

Roberts Andrew (2002). *Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words*. School of Computing, University of Leeds.

Salem André (1987). *Pratique des segments répétés*. Publications de l'INaLF, collection «St.Cloud», Klincksieck, Paris.

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

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

Tesnière Lucien (1982). *Éléments de syntaxe structurale*. Klincksieck (Paris), (1ère édition: 1959)

Vergne Jacques (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN 2003*, tome 2, 139-148.
www.info.unicaen.fr/~jvergne/TALN2003/JVergne-TAL2003multV23.pdf

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

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

Zipf George Kingsley (1935). *The Psychobiology of Language, an Introduction to Dynamic Philology*. Houghton Mifflin, Boston.

Zipf George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Harper, New York, réédition 1966.