Détection de mots-clés par approches au grain caractère et au grain mot

Gaëlle Doualan, Mathieu Boucher, Romain Brixtel, Gaël Lejeune, Gaël Dias
Équipe HULTECH (GREYC, Université de Caen), Bd Maréchal juin, 14032 Caen Cedex
prenom.nom@unicaen.fr

RÉSUMÉ

Nous présentons dans cet article les méthodes utilisées par l'équipe HULTECH pour sa participation au Défi Fouille de Textes 2012 (Deft 2012). La tâche de cette édition du défi consiste à retrouver dans des articles scientifiques, les mots-clés choisis par les auteurs. Nous nous appuyons sur la détection de chaînes répétées maximales ($rstr_{max}$), au grain caractère et au grain mot. La méthode développée est simple et non supervisée. Elle a permis à notre système d'atteindre la 3e place (sur 10 équipes) sur la première piste du défi.

ABSTRACT

Keywords extraction by repeated string analysis

We present here the HULTECH(Human Language Technology) team approach for the Deft 2012 (french text mining challenge). The aim of the challenge is to retrieve the keywords given by the authors of scientific articles. Our method relies on a text algorithmics technic : detection of maximal repeated strings. This technic is applied at character level and word level. We achieved the third rank (over 10) of the first track.

MOTS-CLÉS : Recherche d'information, extraction de mots-clés, algorithmique du texte.

KEYWORDS: Information retrieval, keywords extraction, string algorithmics.

1 Introduction

La tâche proposée dans le cadre du Défi Fouille de Textes 2012 consiste à retrouver dans des articles de sciences humaines les mots-clés proposés par les auteurs. Le corpus de travail est scindé en deux pistes, la première comportant 140 articles et la seconde 141. Une terminologie qui regroupe tous les mots-clés des articles est proposée avec la première piste. Dans cet article nous proposerons deux approches : une basée sur la connaissance de la terminologie, une autre adaptée à l'absence de cette terminologie. Ce sera pour nous l'occasion de comparer les deux approches et leurs résultats. Nos deux approches s'appuient sur un algorithme de recherche de chaînes répétées maximales, ci-après $rstr_{max}$ [1]. Dans la première approche, basée sur la terminologie, nous prenons comme grain d'analyse le caractère. Dans la seconde approche nous prenons comme grain d'analyse le mot graphique, sans appui sur la terminologie ni pour la piste 1 ni pour la piste 2. Dans la section 2, nous procédons à une analyse du corpus qui nous permet d'appréhender le matériau sur lequel nous travaillons. Dans la section 3, nous détaillons

1. L'implantation en python utilisée est disponible à l'url suivante : code.google.com/p/py-rstr-max

JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes, pages 41–48,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

nos deux approches. Ensuite, nous présenterons les résultats dans la section 4et proposons une confrontation de ces deux approches dans la section 5.

2 Description du corpus

Le corpus utilisé comporte des articles de sciences humaines provenant de quatre revues diffusées sur le site Erudit [2]. Nous présenterons ici plus précisément les articles2.1 à traiter et les mots-clés qui leur sont associés2.2.

2.1 Les articles du corpus DEFT 2012

Le corpus DEFT 2012 est constitué de 300 articles répartis sur 4 revues de sciences humaines :
– Anthropologie et Société (AS)
– Revue des Sciences de l'Education (RSE)
– Traduction, Terminologie et Rédaction (TTR)
– Méta : journal des traducteurs (META)

2.1.1 Configuration des articles

Les articles sont au format *xml*. Ils sont constitués d'un identifiant, de la liste des mots-clés fournis par l'auteur, d'un résumé et du corps de l'article lui-même. Le nom de la revue n'apparaît pas dans le fichier *xml* mais dans le nom du fichier. De même, le nom de l'auteur et le titre de l'article ne figurent pas dans le fichier *xml*. Ceci a rendu plus complexe la recherche des mots-clés notamment du fait que le nom de l'auteur figurait systématiquement parmi les mots-clés des articles de la revue Anthropologie et Société.

Nous présentons dans la figure 1 un exemple d'article du corpus afin de montrer sa configuration et sa structure, notons que les titres et sous-titres des articles n'étaient pas disponibles.

2.1.2 Statistiques sur les articles

Nous avons effectué des statistiques sur les articles afin de pouvoir mieux les appréhender (Tableau 2).

	Nombre de documents	Taille moyenne en paragraphes	Taille moyenne en caractères
Piste 1	94	67,8	41235
Piste 2	93	80,2	39153

Tableau 1 – Statistiques sur les documents du corpus d'évaluation

Le nombre moyen de paragraphes ne varie pas particulièrement en fonction de la revue, à l'exception de certains articles de META pour lequel le découpage en paragraphes était mauvais.

2. http://www.erudit.org

```
< ?xml version="1.0" encoding="UTF-8" ?>
-<doc id="0001">
-<motscles>
<nombre>4</nombre>
<mots>Labrecque ;économie politique ;féminisme ;ethnographie</mots>
</motscles>
-<article>
-<resume>
<p>Tout en poursuivant l'objectif de la présentation du numéro,
. . . . . .
la consolidation de la théorie.</p>
</resume>
-<corps>
<p>Qui sape l'ethnographie ébranle la théorie
. . . . . .
d'une anthropologie engagée, d'autre part.</p>
</corps>
</article>
</doc>
```

FIGURE 1 – Un exemple d'article du jeu d'entraînement

2.2 Les mots-clés

Nous avons remarqué que les articles ne comportent pas le même nombre de mots-clés : en moyenne 5,4 95,2 sur la piste 2 et 5,7 sur la piste 1). Mais une grande disparité peut exister d'un texte à l'autre, l'étendue étant de 9 (1 à 10 mots clés par article). Nous avons noté que le premier mot-clé est systématiquement le nom de l'auteur de l'article pour la revue Anthropologie et Société. C'est dans un tel cas que la mention du nom de l'auteur dans le fichier nous aurait été utile.

2.2.1 Nature des mots-clés

– Noms propres : nom de l'auteur (ex : Labrecque), auteur faisant l'objet de l'article (ex : Jack Kerouac), lieu géographique (ex : Japon)
– Noms communs : des noms communs seuls ou parfois accompagnés d'adjectifs, mais jamais de verbes ni d'adverbes (ex : féminisme, économie politique)
– Parfois les noms sont complétés par des compléments du noms, formant des motifs tels que celui-ci : N de art N (ex : traitement de l'information sociale)
– Cas particuliers : des noms coordonnés (ex : traduction scientifique et technique)

Nous avons remarqué que plus les mots-clés étaient longs, moins on avait de chances de les retrouver tels quels dans le texte. Lorsque l'on a la chance de les rencontrer dans le texte, ils y sont peu fréquents. Globalement 79% des mots-clés sont présents tels quels dans le corps du texte, 44,5% dans le résumé et 42% et dans le corps et dans le résumé.

3 Description des approches

Notre première approche basée sur le grain caractère utilise la terminologie afin de s'attaquer à la piste 1. Notre seconde approche n'utilise pas la terminologie et a été utilisée sur les deux pistes.

3.1 Approche au grain caractère

Nous reprenons ici les principes de la méthode utilisée pour le Deft 2011 (Lejeune *et al.*, 2011). On suppose que les segments communs entre le résumé et le reste du texte constituent de bons descripteurs. Pour sélectionner les descripteurs pertinents nous nous fondons sur leur proximité avec des éléments terminologie, technique utilisée dans le domaine de l'Extraction d'Information multilingue (Lejeune *et al.*, 2010).

La méthode $rstr_{max}$ L'analyse au grain caractère est effectué en recherchant des motifs sans trous (ci-après *motifs*) tels que définis par (Ukkonen, 2009). Ces motifs sont des sous-chaînes du texte ayant les caractéristiques suivantes [3] :

répétés : les motifs apparaissent au moins deux fois ;

maximaux : les motifs ne sont pas inclus dans des motifs plus grands et de même effectif

3. Pour une description plus formelle voir `code.google.com/p/py-rstr-max`

Nous comparons les deux segments textuels(résumé et corps) et l'ensemble de la terminologie en une seule opération. Nous conservons les $rstr - max$ apparaissant dans ces deux segments et dans un élément de la terminologie. Seuls les motifs respectant un critère de longueur donné sont considérés comme pertinents. Pour tenir compte des variations morphologiques du français, nous avons fixé la proximité minimale entre un motif trouvé et un élément de la terminologie à 0.9. Autrement dit, un élément t de la terminologie est considéré comme mot-clé du texte s'il existe une chaîne c telle que :

– c est présent dans le résumé et dans le corps de l'article
– c est une sous chaîne de t
– $\frac{len(c)}{len(t)} \geq \frac{9}{10}$ avec len le nombre de caractères dans c et t

Nous n'avons pas appliqué cette méthode à la seconde piste car la sélection de chaînes de caractères adaptées à l'évaluation était malaisée. Il aurait fallu un grand nombre d'heuristiques pour retrouver des mots-clés comparables à la référence. Nous avons préféré garder la "pureté" de cette méthode. En effet le seul pré-traitement effectué est le découpage en deux segments textuels (résumé et corps). Aucun outillage linguistique (lemmatisation, étiquetage...) n'est nécessaire. Par ailleurs, aucun post-traitement n'est effectué.

3.2 Approche au grain mot

Pour notre seconde approche, nous procédons à un découpage plus classique en mots. Cette méthode est conçue pour fonctionner en l'absence de terminologie de référence. Nous appliquons l'algorithme de détection des $rstr_{max}$ (section : 3.1) mais en l'appliquant cette fois sur des mots.

L'algorithme $rstr_{max}$ est appliqué à tout ce qui est compris entre les balises <article> ce qui correspond au résumé et au corps de l'article. Nous considérons le tout comme une chaîne. Nous obtenons ainsi un ensemble de chaînes de mots répétées et maximales. Un grand nombre de motifs sont détectés dont certains sont partiellement redondants. Par exemple, on a les motif $ABCD$ et $BCDF$ et on souhaite souvent ne garder que la partie centrale BCD. Pour améliorer la précision, nous appliquons donc une seconde fois $rstr_{max}$ sur la liste des chaînes obtenues.

3.2.1 IDF

Pour améliorer la précision de nos résultats, nous voulons réduire encore le nombre de chaînes obtenues. Cependant, il nous faut conserver un rappel correct. Pour ce faire nous avons choisi de calculer l'IDF (Inverse Document Frequency) de chaque chaîne. Cette mesure fait ressortir les chaînes spécifiques à un texte par rapport au corpus. L'IDF est l'inverse de la fréquence de la chaîne dans un ensemble de documents. Cette mesure est généralement couplée avec le TF (term frequency ou effectif du mot dans un document) en Voici comment se calcule le $TF \times IDF$ d'une chaîne C dans un document D [4] :

$$TF \times IDF = \frac{freq(C,D)}{t(D)} \times -\log_2 \frac{nd(C)}{N}$$

Avec :

– freq(C,D) le nombre de fois que la chaîne C apparaît dans le document D
– t(D) le nombre de mots du document D
– nd(C) le nombre de documents contenant C dans le corpus
– N la taille du corpus en documents

Cependant, nous ne conservons que l'IDF. Dans notre cas, il n'était pas nécessaire l'appliquer le TF. En effet, grâce à la méthode $rstr_{max}$, nous obtenons les chaînes maximales répétées, ce qui signifie qu'elles ont déjà une certaine fréquence dans le document. Par ailleurs, le TF a tendance à privilégier les chaînes très fréquentes d'un texte, autrement dit des mots vides peu susceptibles d'être des mots-clés.
Pour calculer l'IDF, nous considérons l'ensemble des articles d'une piste. Cela nous permet de caractériser un article par rapport à une piste. Cela se justifie si nous nous replaçons dans le sémantique textuelle de François Rastier : " le texte pour une linguistique évoluée l'unité *minimale*, et le corpus l'ensemble dans lequel cette unité prend son sens "(Rastier, 2002). Ainsi, un article ne prend son sens que dans le corpus de travail si bien que nous devons caractériser ces chaînes et ces mots-clés par rapport à l'ensemble du corpus. Lorsque nous calculons l'IDF des chaînes nous obtenons des résultats compris entre 0 et 5. Nous classons les chaînes en ordre décroissant de leur IDF. Le but étant de réduire le nombre de chaînes, nous ne conservons que celles dans l'IDF est supérieure à 2.

3.2.2 Pondération des chaînes

L'IDF constitue un premier filtrage par pondération mais ce n'est pas suffisant. Nous procédons donc à un second filtrage par pondération en attribuant un poids aux chaînes restantes en fonction des critères suivants :

– IDF
– fréquence de la chaîne dans l'article
– fréquence de la chaîne dans le résumé
– longueur de la chaîne
– présence de la chaîne dans le premier paragraphe (a priori : introduction
– présence de la chaîne dans la dernier paragraphe (a priori : conclusion)

A chacune de ces mesures est attribué un coefficient qui pondère leur importance. Nous avons effectué des statistiques sur le corpus afin d'anticiper les places occupées par les mots-clés dans les articles. Ainsi, si une chaîne est fréquente dans le résumé, elle a davantage de chance d'être un mot-clé qu'une autre chaîne. Nous attribuons donc un certain poids à ces mesures en fonction de leur capacité à traduire le comportement des mots-clés. Notons que l'absence des titres dans les documents analysés rend difficile la détection des segments introductifs et conclusifs . Les chaînes sont rangées en ordre décroissant de poids et nous sélectionnons les 7 premières chaînes en guise de mots-clés. Ce seuil a été fixé à partir des meilleurs résultats obtenus sur le corpus d'entraînement.

4 Résultats

	Résultat piste 1	Résultat piste 2
Approche 1 : $rstr_{max}$ au grain caractère	**0.44**, 3e/10	⊘
Approche 2 : $rstr_{max}$ au grain mot	0,12	**0,13**, 7e/9
Baseline : tf-idf simple	0,08	0,07

Tableau 2 – Résultats et rangs pour nos 2 approches et notre baseline

La première approche donne de bons résultats en raison de l'appui de la terminologie, bien meilleurs qu'avec l'approche par poids. Sans doute ces résultats auraient pu être améliorés avec quelques heuristiques, par exemple : chercher à affecter chaque mot-clé de la terminologie à au moins un document. Mais nous n'avons pas souhaité complexifier la procédure utilisée.

Concernant la seconde approche, elle aurait sans doute eu de meilleurs résultats sur la piste 1 en s'appuyant sur la terminologie mais nous avons souhaité pour les deux pistes conserver l'aspect 'sans ressources externes".

5 Discussion

Nous avons opté pour des méthodes simples à mettre en place et peu coûteuses en temps, peut être au détriment de la qualité des résultats. La première approche se voulait avant tout indépendante de la langue considérée. Travailler sur le grain caractère permet de dépasser les problèmes de découpage des textes en mots. Toutefois pour se conformer aux modalités d'évaluation, le soutien de la terminologie s'est avéré nécessaire. La seconde approche se voulait indépendante de tout support extérieur. En effet, ne pas utiliser la terminologie permet d'extraire des informations nouvelles à partir d'un document brut.

Nos deux approches ont en commun l'utilisation d'une méthode d'algorithmique du texte : $rstr_{max}$. L'algorithme recherche des chaînes répétées maximales, supposées caractéristiques d'un texte. Nos approches diffèrent par le grain d'analyse : caractère pour l'une, mot pour l'autre.

La première méthode présente l'avantage de la simplicité, elle ne nécessite aucun paramètre mais e base sur la terminologie. La seconde méthode ne nécessite pas de terminologie mais impose des traitements supplémentaires.

Nos deux méthodes présentent par ailleurs l'avantage de détecter facilement des unités multi-mots, souvent plus pertinentes pour des tâches d'indexation documentaire et de recherche d'information.

Enfin, nos deux approches sont indépendantes de tout module d'analyse linguistique (lemmatisation, étiquetage...) ce qui les rend a priori moins sensibles à une utilisation sur d'autres langues que le français. Il serait donc intéressant d'expérimenter ces techniques sur des corpus multilingues.

Références

Lejeune, G., Brixtel, R., Giguet, E. et Lucas, N. (2011). Deft2011 : appariement de résumés et d'articles scientifiques fondé sur les chaînes de caractères. *In Défi Fouille de Textes/TALN 2011*, pages 53–64.

Lejeune, G., Doucet, A., Yangarber, R. et Lucas, N. (2010). Filtering news for epidemic surveillance : towards processing more languages with fewer resources. *In 4th Workshop on Cross Lingual Information Access*, pages 3–10.

Rastier, F. (2002). Enjeux épistémologiques de la linguistique de corpus. *In 2ème journées de la linguistique de corpus*.

Ukkonen, E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theor. Comput. Sci.*, 410(43):4341–4349.