Processus d'acquisition d'un dictionnaire de sigles et de leurs définitions à partir d'un corpus

Vladislav Matviico, Nicolas Muret, Mathieu Roche

LIRMM, Université Montpellier 2 - CNRS UMR5506,
mroche@lirmm.fr

Résumé. Le logiciel présenté dans cet article s'appuie sur une approche d'acquisition de sigles à partir de données textuelles.

1 Introduction

De nombreux domaines comme la biologie ou la médecine voient naître chaque jour de nouveaux termes et abréviations, notamment des sigles. Un sigle est un ensemble de lettres initiales servant d'abréviation, par exemple "RATP" peut être associé à la définition (aussi appelée expansion) "Régie Autonome des Transports Parisiens". Nos travaux ont consisté à développer un logiciel afin de faciliter l'acquisition ou l'enrichissement de dictionnaires en extrayant automatiquement, à partir de diverses sources, les sigles et leur(s) définition(s). Une fois ces dictionnaires constitués, l'approche $AcroDef$ que nous avons proposée dans (Roche et Prince (2007)) consiste à établir la définition pertinente d'un sigle présent dans un document. Dans ces documents, la définition n'est pas toujours présente d'où la difficulté du traitement. Dans ce contexte, il est donc essentiel d'avoir à disposition un dictionnaire adapté, ce qui justifie les travaux présentés dans cet article.

De nombreuses méthodes pour extraire les sigles et leur(s) définition(s) ont été développées (Larkey et al. (2000); Okazaki et Ananiadou (2006)). La plupart des approches de détection de sigles dans les textes s'appuient sur l'utilisation de marqueurs spécifiques associés à des heuristiques adaptées. Certains travaux récents (Okazaki et Ananiadou (2006)) consistent à associer ces approches à des mesures statistiques spécifiques pour améliorer la qualité des méthodes d'acquisition de dictionnaires. L'approche que nous avons développée se compose de deux étapes successives qui sont détaillées dans la section 2.

2 Acquisition d'un dictionnaire sigles/définitions

Notre méthode qui consiste à extraire les candidats sigles/définitions s'appuie sur la présence de marqueurs (parenthèses, crochets). Deux situations peuvent alors être considérées :
1. Le sigle se situe avant la définition qui se trouve entre les marqueurs (les parenthèses dans le cas le plus courant). Exemple : "`... S.I.G. (Solde Intermédiaire de Gestion) ...`"
2. La définition se trouve avant le sigle qui se trouve entre les marqueurs. Exemple : "`... les Systèmes d'Informations Géographiques (SIG) ...`". Dans ce cas, la taille de la définition est pour le moment indéterminable. Il est ainsi nécessaire de la définir arbitrairement en fonction du nombre de lettres composant le sigle. Nous avons expérimentalement fixé cette taille à trois fois le nombre de lettres composant le sigle.

La seconde étape de notre application utilise les résultats obtenus lors de la première phase afin de filtrer les candidats pertinents. Les résultats sont triés afin de (1) supprimer les paires sigle/définition non pertinentes, (2) extraire précisément les définitions présentes dans les définitions potentielles (ces dernières pouvant être trop longues puisque coupées arbitrairement lors du second cas de la recherche des candidats). Pour permettre un tel filtrage, nous effectuons un alignement des lettres contenues dans le sigle avec les mots de la définition. Cet alignement consiste à vérifier la correspondance entre les lettres des sigles avec les premières lettres de chacun des mots des définitions. Dans notre méthode, si le premier caractère des mots de la définition candidate ne peut être aligné, les caractères qui suivent au sein des mots sont considérés. Par exemple, cette méthode permet de reconnaître "Extraction Itérative de la Terminologie" comme la définition du sigle EXIT dans lequel la lettre "X" a pu être alignée. Nous présentons ici une évaluation de notre système d'alignement des sigles avec les définitions candidates. Pour cette évaluation, nous nous appuyons sur les données issues du site http ://www.sigles.net/. L'évaluation consiste à extraire aléatoirement de ce site des sigles de 2, 3 et 4 caractères et d'évaluer le taux de réussite de l'alignement (nombre de sigles alignés avec les définitions du site en utilisant la version actuelle de notre logiciel développé en Java - *version 1.0*). Le tableau ci-dessous présente les résultats de plus de 800 alignements qui sont globalement très satisfaisants (taux de réussite de 78% à 98%).

Nb de lettres	Nb de sigles	Nb de définitions	Nb de définitions non reconnues	% de réussite
2	100	616	11	**98.2 %**
3	50	157	10	**93.6 %**
4	20	32	7	**78.1 %**

3 Conclusion et perspectives

L'application présentée dans cet article consiste à acquérir ou enrichir de manière automatique un dictionnaire de sigles/définitions à partir d'un corpus. Notre approche n'utilise aucune connaissance linguistique, elle peut donc s'appliquer à des textes en différentes langues. Dans nos futurs travaux, nous proposons d'associer de manière automatique le domaine de chaque sigle/définition en utilisant des méthodes fondées sur le contexte (Roche et Prince (2007)).

Références

Larkey, L. S., P. Ogilvie, M. A. Price, et B. Tamilio (2000). Acrophile : An automated Acronym Extractor and Server. In *Proceedings of the Int. Conf. on Digital Libraries*, pp. 205–214.

Okazaki, N. et S. Ananiadou (2006). A Term Recognition Approach to Acronym Recognition. In *Proceedings of ACL*, pp. 643–650.

Roche, M. et V. Prince (2007). *AcroDef :* A Quality Measure for Discriminating Expansions of Ambiguous Acronyms. In *Proceedings of CONTEXT, Springer-Verlag, LNCS*, pp. 411–424.

Summary

The software presented in this paper is based on an approach of acronyms extraction in textual data.