Extraction de termes centrée autour de l'expert

Thomas Heitz, Mathieu Roche, Yves Kodratoff

Université Paris-Sud, Bât 490, 91405 Orsay Cedex France,
{heitz, roche, yk}@lri.fr,
http://www.lri.fr/∼{heitz, roche, yk}/

Résumé. Nous développons un logiciel, Exit, capable d'aider un expert à extraire des termes qu'il trouve pertinents dans des textes de spécialité. Tout est mis en place pour faciliter le travail de l'expert afin qu'il puisse consacrer son temps à la seule reconnaissance des termes pertinents. Pour cela, différentes mesures statistiques et de nombreuses options d'extraction sont disponibles dans Exit. Afin d'utiliser au mieux les connaissances de l'expert, notre approche est semi-automatique. De plus, l'expert construit des termes pouvant inclure des termes précédemment extraits ce qui rend itératif et constructif notre processus de formation des termes. Enfin, l'ergonomie du logiciel a profité des enseignements tirés lors de son utilisation pour une compétition internationale d'extraction de connaissances.

1 Introduction

Le logiciel Exit (Roche et al., 2004b), **EX**traction **I**térative de la **T**erminologie permet l'extraction des collocations. Une collocation peut être définie comme *une combinaison de mots dont le sens global est déductible des unités qui la composent, une des unités caractérisant l'autre*. Par exemple : *plante à fleurs*, *à fleurs* caractérisant *plante*.

Toutes les expressions extraites par ce logiciel sont des collocations. Mais ce qui intéresse l'expert, ce sont les **collocations pertinentes** qui sont des expressions ayant un sens unique pour un domaine précis. Nous les appellerons **termes** dans la suite de cet article.

Exit est destiné à des utilisateurs experts d'un domaine. Ceux-ci doivent donc avoir une connaissance approfondie des notions employées dans les textes de spécialité analysés pour pouvoir reconnaître les collocations pertinentes parmi celles extraites par le logiciel.

Ce logiciel est un des modules d'une chaîne de fouille de textes (Mathiak et Eckstein, 2004; Roche et al., 2004a) qui comprend les étapes de normalisation, étiquetage grammatical, construction de la terminologie avec Exit, classification conceptuelle, etc. Ceci permet ensuite de pouvoir traduire, résumer, générer ou interroger des textes. Les entrées d'Exit correspondent à des textes étiquetés grammaticalement, notamment (Brill, 1995), et les sorties correspondent à des listes de termes qui peuvent être associés à des concepts grâce à des systèmes de construction de classifications conceptuelles (Kodratoff, 2004).

Un des points forts d'Exit est son processus itératif qui permet de construire des termes incluant des termes précédemment trouvés. De plus, tout le processus d'extraction est centré autour de l'expert et nous avons fait en sorte qu'il soit aidé au maximum pour qu'il passe le moins de temps possible à cette tâche.

Exit a déjà été utilisé par notre équipe pour la compétition TREC [1] Novelty 2004 (Soboroff et Harman, 2003) qui consistait à trouver des phrases pertinentes et nouvelles dans des articles de journaux américains. Son utilisation a permis d'extraire des termes d'un corpus d'environ 10 Mo et d'obtenir une liste de termes classés afin de construire une terminologie des articles. De nombreuses discussions entre l'expert et le développeur ont permis des améliorations significatives de l'ergonomie du logiciel.

Voyons maintenant quelques procédés existant d'extraction de termes et le positionnement de plusieurs logiciels l'effectuant dont Exit.

2 Procédés d'extraction de termes

Beaucoup d'approches ont été utilisées pour l'extraction automatique de la terminologie. Notamment l'approche **statistique** utilisée en partie dans Xtract (Smadja, 1993), l'ajout d'un filtrage **syntaxique** tel adjectif-nom, la recherche de variations morpho-syntaxiques à partir d'un noyau de termes fournis (Jacquemin, 1995), la pondération par les mots entourant les termes de ATR (Frantzi et Ananiadou, 1997), etc.

Il existe une **approche constructive** de l'extraction de termes (Smadja, 1993) qui consiste à étendre des collocations et une approche **déconstructive** (Frantzi et Ananiadou, 1997) qui consiste à chercher des sous-chaînes d'une chaîne de mots plus grande.

Les systèmes peuvent être **automatiques** comme Lexter (Bourigault, 1993) qui extrait les groupes nominaux maximaux puis les décomposent en termes de *têtes* et *d'expansions* à l'aide de règles grammaticales et Acabit (Daille, 1994) qui extrait des termes principalement nominaux puis les classe selon une mesure statistique.

D'autres systèmes, tel Exit, sont **semi-automatiques** afin de donner à l'expert la possibilité d'intervenir au cours du processus d'extraction de termes.

Notre approche avec Exit est fondée sur une méthode statistique aidée d'informations syntaxiques. Elle a aussi la particularité d'être itérative. C'est-à-dire que l'expert, en réitérant les extractions de termes, peut construire de nouveaux termes en incluant ceux précédemment construits.

Dans le tableau 1, nous avons résumé les caractéristiques des logiciels cités et d'Exit. Un état de l'art plus complet des systèmes d'extraction de la terminologie est présenté dans (Aussenac-Gilles et Bourigault, 2003).

	Lexter	Acabit	Xtract	ATR	Exit
statistique / syntaxique	/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
constructif / déconstructif	/ ✓	/ ✓	✓/	/ ✓	✓/
automatique / semi auto	✓/	✓/	✓/	✓/	/ ✓

Tab. 1 – *Comparaison de quelques procédés d'extraction de termes.*

La pertinence d'un terme extrait dépend de sa fréquence dans le texte, de celle de ses unités constitutives ainsi que d'autres paramètres (Daille, 1994) que reprennent les

1. **T**ext **RE**trieval **C**onference, http://trec.nist.gov/

mesures statistiques utilisées dans ce logiciel et que nous allons évoquer dans la section suivante.

3 Pertinence des termes extraits

Deux principales mesures statistiques sont disponibles dans le logiciel :

- La mesure d'information mutuelle (Church et Hanks, 1990) plutôt utilisée pour trouver les expressions figées telle *humour noir*. Une expression figée étant une *combinaison de mots dont le sens global n'est pas la somme des sens des mots le composant* ;
- La mesure de vraisemblance (Dunning, 1993) utile pour trouver des traces linguistiques de concepts (Kodratoff, 2004) spécifiques aux textes tel *droits de l'homme* dans un texte de droit international.

L'expert, disposant de plusieurs mesures, peut choisir celle qui convient le mieux aux types de termes qu'il veut extraire. Par exemple, il peut être intéressant de commencer par une extraction des termes de type **expressions figées** afin d'en obtenir la liste et de l'utiliser comme prétraitement. Ceci permet ensuite d'utiliser d'autre mesures qui extrairont des termes plus proches des **concepts** sans tenir compte des expressions figées.

De nombreuses options d'extraction, disponibles dans le panneau extraction (figure 1) du logiciel permettent à l'expert d'obtenir les termes qui correspondent à ses attentes. Notamment tous les types de relations grammaticales, comme nom-nom ou adjectif-nom, le nombre de termes présentés à l'expert, le seuil d'élagage, c'est-à-dire la fréquence minimum d'apparition d'un terme dans le texte, etc.

Dans la section suivante, nous expliquons l'un des principes clé de notre approche, à savoir, l'itération des extractions.

4 Principe d'itération des extractions

Ce logiciel regroupe les mots ou unités des collocations reconnues par l'expert comme pertinentes et forme ainsi des termes. L'intérêt de regrouper les unités d'un terme est d'obtenir une nouvelle unité plus riche de sens, c'est-à-dire plus proche d'un concept.

Le fait d'itérer les extractions permet de construire de nouveaux termes incluant les termes précédents. Par exemple, si *droits de l'homme* est sélectionné comme un terme pertinent par l'expert, à l'itération suivante le terme *défense des droits-de-l'homme* pourra apparaître dans les résultats de l'extraction. En effet, les termes acceptés à une itération sont groupés à l'aide de tirets avant l'extraction suivante.

Ainsi, il est possible de **construire des termes en incluant des termes précédemment acceptés**. C'est pourquoi il est important d'utiliser un ordre précis dans le choix des relations extraites afin de tirer parti de cette caractéristique du système.

Par exemple, il est important de commencer par repérer les formes longues de termes du type adjectif-et-adjectif tel *orange-et-bleue* avant de traiter les formes incluant ces

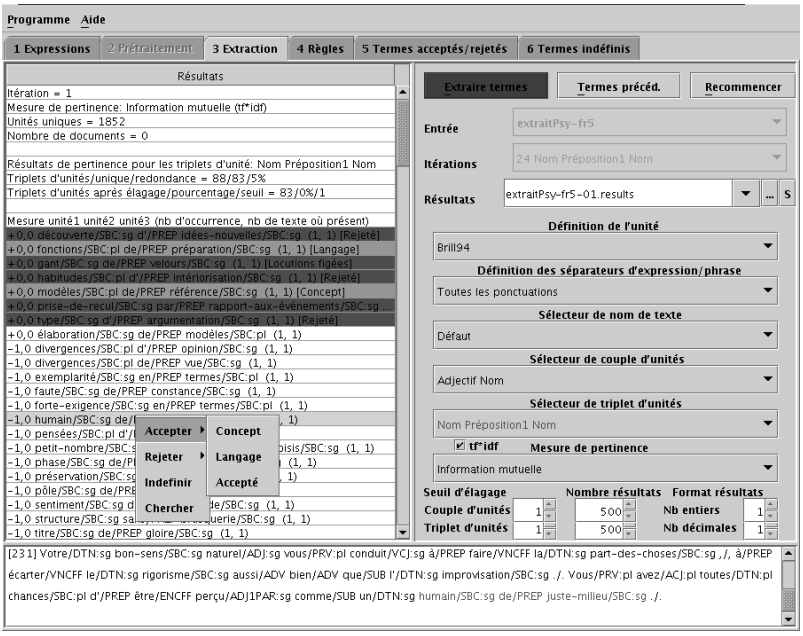

FIG. 1 – *Panneau d'extraction. À gauche se trouve la liste des termes extraits et à droite les options d'extraction. En bas est affiché le contexte du terme sélectionné.*

précédents termes comme nom-adjectif tel *chemise orange-et-bleue*. Ici, *orange-et-bleue* aura été regroupé avec des tirets lors de la première extraction et réétiqueté adjectif. Ce qui rend possible son inclusion dans un nouveau terme de type nom-adjectif à l'extraction suivante.

C'est à l'expert de trouver le bon ordre de construction des termes. Cependant, cela est facilité par l'affichage du contexte du terme dans les phrases qui le contiennent.

La possibilité de revenir en arrière et l'automatisation du classement des termes simplifient le travail de l'expert, comme nous allons l'expliquer. Enfin, nous terminerons sur un aperçu de l'ergonomie générale du logiciel.

5 Retour en arrière, automatisation et ergonomie

Le logiciel sauvegarde chaque extraction, juste avant l'extraction suivante, de façon transparente pour l'utilisateur. Ainsi, il est possible de revenir à n'importe quelle itération avec les options et les règles associées.

Les termes déjà classés peuvent être reclassés par la suite si l'expert découvre une erreur dans ce classement. Il lui suffit d'aller dans les panneaux des termes et de sélectionner les termes à reclasser puis de leur attribuer une nouvelle classe.

Des **règles de classement automatique des termes**, booléennes et pouvant

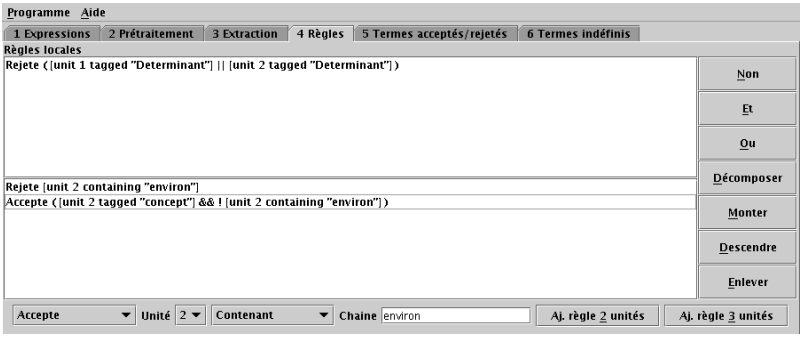

FIG. 2 – *Panneau des règles. Ici sont définies les règles de classement des termes en fonction de leurs mots et étiquettes.*

contenir des expressions régulières, peuvent être très facilement élaborées dans l'interface. Ces règles donnent automatiquement une classe aux termes qui les satisfont dès que l'extraction est terminée. Ensuite, l'expert choisit de conserver ou de modifier ce classement automatique puis traite les termes non classés. Voir la figure 2 pour un exemple de règles qui acceptent ou rejettent des termes en fonction de leurs unités.

Ce logiciel dispose de **deux niveaux d'aide** avec des bulles d'aide et une aide contextuelle complète ainsi que d'un guide de premier démarrage. EXIT est programmé en Java 1.4 et fonctionne donc sur tous les systèmes d'exploitation courants. Il est entièrement commenté en JavaDoc ce qui en facilite les modifications ultérieures. Le manuel complet du logiciel est accessible sur internet [2].

6 Conclusion

Le logiciel EXIT d'extraction de termes s'intègre dans une chaîne de fouille de textes comme module de construction de la terminologie. Son approche itérative, ses différentes mesures statistiques et ses options d'extraction permettent d'obtenir les types de termes que l'expert désire. Son interface classique avec deux niveaux d'aide et la sauvegarde automatique des extractions facilitent la prise en main du logiciel. Enfin, son adaptabilité et son automatisation le rendent capable de traiter des textes de taille importante. Tout est donc mis en place pour faciliter le travail de l'expert afin qu'il puisse consacrer son temps à la seule reconnaissance des termes pertinents.

Références

Aussenac-Gilles, N. et Bourigault, D. (2003). Construction d'ontologies à partir de textes. Dans *Actes de TALN03*, volume 2, pages 27–47.

Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *T.A.L.*, 34(2):105–118.

2. http://www.lri.fr/ia/fdt/exit/aide.html

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Church, K. W. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. Dans *Computational Linguistics*, volume 16, pages 22–29.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7. Thèse de Doctorat en Informatique Fondamentale.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Frantzi, K. et Ananiadou, S. (1997). Automatic term recognition using contextual clues.

Jacquemin, C. (1995). A symbolic and surgical acquisition of terms through variation. Dans *Learning for Natural Language Processing*, pages 425–438.

Kodratoff, Y. (2004). Induction extensionnelle : définition et application l'acquisition de concepts à partir de textes. *Revue RNTI E2, numéro spécial EGC'04*, 1:247–252.

Mathiak, B. et Eckstein, S. (2004). Five steps to text mining in biomedical literature. Dans *Proceedings of « Data Mining and Text Mining for Bioinformatics » workshop of ECML/PKDD Conference*, pages 44–49.

Roche, M., Azé, J., Matte-Tailliez, O., et Kodratoff, Y. (2004a). Mining texts by association rules discovery in a technical corpus. Dans *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pages 89–98.

Roche, M., Heitz, T., Matte-Tailliez, O., et Kodratoff, Y. (2004b). Exit: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. Dans *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, volume 2, pages 946–956.

Smadja, F. A. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Soboroff, I. et Harman, D. (2003). Overview of the trec 2003 novelty track. Dans *NIST Special Publication: SP 500-255 The Twelfth Text Retrieval Conference (TREC 2003)*.

Summary

We present a software, named EXIT, built in order to help a field expert to spot relevant terms in texts of his/her specialty. We aim at helping the expert to focus the attention on pertinence. We provide the expert with various statistical measures, and several mining methods. The bottom-up strategy we use to build the terms enables the expert to watch the growth of the terms as new components are included to the shorter terms obtained at iteration n-1. Participating in the TREC'04 Novelty challenge helped us to improve the relevance of the questions asked to the expert.