TTC TermSuite alignement terminologique à partir de corpus comparables

Béatrice Daille et Rima Harastani
LINA, 44322 Nantes Cedex 03

beatrice.daille@univ-nantes.fr, rima.harastani@univ-nantes.fr

RÉSUMÉ

TermSuite est outil libre multilingue réalisant une extraction terminologique monolingue et une extraction terminologique bilingue à partir de corpus comparables.

ABSTRACT

TTC TermSuite – Terminological Alignment from Comparable Corpora

TermSuite is based on a UIMA framework and performs monolingual and bilingual term extraction from comparable corpora for a range of languages.

MOTS-CLÉS : corpus comparable, extraction terminologique, alignement, UIMA
KEYWORDS : comparable corpora, terminology extraction, terminology alignment, UIMA

Le projet européen TTC[1] s'est intéressé à l'exploitation des corpus comparables de domaines techniques pour l'amélioration des outils informatiques de traduction. La plateforme web TTC[2] permet de compiler des corpus comparables à partir du web, d'en extraire et traduire la terminologie, et d'exporter cette terminologie dans EuroTermBank[3]. TermSuite[4] constitue le cœur de la plateforme web TTC : il réalise l'extraction et l'alignement terminologique dans 7 langues : Anglais, Français, Allemand, Espagnol, Letton, Chinois et Russe. TermSuite adopte la plate-forme Apache UIMA[5] conçue pour faciliter l'assemblage de composants, leur intégration au sein d'une chaîne de traitement ainsi que le passage à l'échelle en contexte industriel.

TermSuite effectue les traitements informatiques en 3 phases :

1. **Analyses linguistiques** : découpage du texte en mots, analyse morphosyntaxique et lemmatisation et conversion au format Multext ;
2. **Extraction terminologique monolingue** : détection d'occurrences de termes simples et complexes, normalisation et regroupement des termes en fonction de leurs variations, filtrage statistique ; listes de termes en format tsv et TBX.
3. **Alignement terminologique bilingue** : plusieurs types d'alignement par paires de langues sont proposés qui adoptent différentes approches : distributionnelle (Fung, 1998), compositionnelle (Grefenstette, 1999), ou mixte (Daille et Morin, 2012). Les approches s'appliquent aux termes simples, aux termes complexes et aux composés savants (Harastani et al., 2012)

[1] http://www.ttc-project.eu
[2] http://ttc.syllabs.com/
[3] http://www.eurotermbank.com/
[4] http://code.google.com/p/ttc-project
[5] http://uima.apache.org

Remerciements

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no 248005.

Références

FUNG, P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. *In* FARWELL, D., GERBER, L. et HOVY, E., éditeurs : *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98),* pages 1–16, Langhorne, PA, USA.

GREFENSTETTE, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *In ASLIB'99 Translating and the Computer 21*, London, UK.

HARASTANI, R., DAILLE, B., MORIN, E. (2012). Neoclassical Compound Alignments from Comparable Corpora. *In Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics* (*CICLING 2012)*, vol. 2, pages 72-82. New Delhi, India.

MORIN, E. et DAILLE, B. (2012). Compositionalité et contexte pour l'extraction de terminologies bilingues à partir de corpus comparables. *In Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Grenoble. ATALA, LIG.