

TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue

Béatrice Daille Christine Jacquin Laura Monceaux Emmanuel Morin Jérôme
Rocheteau

Université de Nantes - LINA – 2 rue de la Houssinière – BP 92208 – 44322 Nantes cedex 3,
France

{prenom.nom}@univ-nantes.fr

Le projet européen TTC¹ vise à exploiter les possibilités offertes par les corpus comparables pour améliorer les performances des outils informatiques de traduction. Il s'agit de traiter des domaines techniques dans un contexte massivement multilingue où il est nécessaire de traduire un même document dans plusieurs langues. TTC TermSuite est un ensemble de composants logiciels pour l'extraction et l'alignement terminologique multilingue à partir de corpus comparables dans 5 langues européennes - Anglais, Français, Allemand, Espagnol et une langue peu dotée, le Letton, ainsi qu'en Chinois et en Russe.

TTC TermSuite adopte la plate-forme Apache UIMA² conçue pour faciliter l'assemblage de composants, leur intégration au sein d'une chaîne de traitement ainsi que le passage à l'échelle dans un contexte industriel.

TTC TermSuite procède à une extraction terminologique monolingue pour les 7 langues, puis à son alignement par paire de langues. En entrée, sont fournis plusieurs corpus comparables dont les documents sont composés de deux types de fichiers : le texte du document et les métadonnées associées au format Dublin Core³. Ces métadonnées recensent la langue, la source du document, la date d'extraction s'il s'agit d'un fichier extrait du web, le format (.txt, .html, .pdf, etc.), le sujet. Seule la langue est une métadonnée obligatoire. En sortie, sont produites des listes terminologiques monolingues et bilingues sous la forme d'un fichier XML au format TermBase eXchange⁴.

TTC TermSuite effectue les traitements informatiques dédiés à l'acquisition terminologique en 4 phases :

Traitements préliminaires : identification et conversion des encodages de caractères, détection de la langue ;

Analyses linguistiques découpage du texte en mots, analyse morphosyntaxique et lemmatisation et conversion au format Multext ;

Extraction terminologique monolingue détection d'occurrences de termes simples et complexes, normalisation et regroupement des termes en fonction de leurs variations, filtrage statistique ;

Alignement terminologique bilingue alignement contextuel par paires de langues.

Chacune des unités fonctionnelles qui composent les 4 phases de cette architecture logicielle est réalisée par un composant UIMA dédié. Chacun de ces composants gère le multilinguisme et, au besoin, répartit le document en cours de traitement à un sous-composant dédié au traitement de la langue de ce document.

TTC TermSuite est librement distribué⁵ accompagné d'une vidéo sur Youtube⁶ expliquant comment l'utiliser.

La démonstration présentera l'extraction et l'alignement des termes simples sur l'Anglais, Français, Allemand, Espagnol, Chinois et Russe, ainsi que l'extraction et alignement de termes complexes sur le Français et l'Anglais.

1. <http://www.ttc-project.eu>

2. <http://uima.apache.org>

3. <http://dublincore.org>

4. <http://www.lisa.org/Term-Base-eXchange.32.0.html>

5. <http://code.google.com/p/ttc-project>

6. <http://www.youtube.com/watch?v=Vi6yoXaFZ44>