Mixer les moyens pour extraire les gloses

Augusta Mela*, Mathieu Roche** et Mohamed el Amine Bekhtaoui**

* Université Montpellier 3, 34199 Montpellier Cedex 5, France
Augusta.Mela@univ-montp3.fr
** LIRMM,CNRS, Université Montpellier 2, 34392 Montpellier Cedex 5, France
Mathieu.Roche@lirmm.fr, a.bekhtaoui@gmail.com

Résumé : Nous proposons d'extraire des connaissances lexicales en exploitant les « gloses » de mot, ces descriptions spontanées de sens, repérables par des marqueurs lexicaux et des configurations morpho-syntaxiques spécifiques. Ainsi dans l'extrait suivant, le mot *testing* est suivi d'une glose en *c'est-à-dire* : « *10 % de ces embauches vont porter sur un métier qui monte : le «testing», c'est-à-dire la maîtrise des méthodologies rigoureuses de test des logiciels*». Cette approche ouvre des perspectives pour l'acquisition lexicale et terminologique, fondamentale pour de nombreuses tâches. Dans cet article, nous comparons deux façons d'extraire les unités en relation de glose : patrons et statistiques d'associations d'unités sur le web, en les évaluant sur des données réelles.

1 Introduction

L'acquisition automatique de connaissances lexicales à partir de textes vise à identifier divers types d'unités lexicales (termes, entités nommées, mots composés, mots nouveaux, mots à sens nouveau) ainsi que leurs propriétés syntaxiques et sémantiques. En contexte multilingue, à partir des corpus bilingues, elle consiste à repérer les traductions de ces unités.

Elle constitue une aide précieuse pour la construction de dictionnaires, thesaurus et terminologies, qu'ils soient de langue générale ou spécialisée. Elle intéresse également la recherche documentaire grâce à l' « expansion de requête » puisqu'elle permet de comparer aux index des documents susceptibles de correspondre à la requête de l'utilisateur, non seulement les mots présents dans la requête mais également leurs synonymes, hyperonymes ou hyponymes, voire leurs traductions dans une autre langue, toutes connaissances obtenues en amont par des procédés d'acquisition lexicale.

Trois objectifs peuvent donc être distingués :

l'*extraction* d'unités, ou comment repérer des unités lexicales spécifiques : termes, entités nommées, mots composés ;

l'*alignement* d'unités, ou comment repérer leurs traductions à partir de corpus bilingues ;

la *structuration* de ces unités, c'est-à-dire les relations de ces unités entre elles.

La section 2 présente les approches généralement utilisées en structuration, la section 3 présente notre approche conceptuelle, la section 4 est réservée au logiciel qui en résulte et à partir duquel sont obtenus les résultats expérimentaux rapportés en section 5. La section 6 dessine quelques perspectives de ce travail.