Sujet de stage de recherche - Master 2 ATAL

Multi-alignement en corpus comparables

Laboratoire d'accueil : LINA – Equipe TALN (Traitement Automatique du Langage Naturel)
Encadrants : Béatrice Daille/Emmanuel Morin
Courriel : beatrice.daille@univ-nantes.fr/emmanuel.morin@univ-nantes.fr
Tél. : 02 51 12 58 54/ 02 51 12 58 39

Mots clefs : TALN, corpus, comparabilité, multilinguisme, alignement, similarité

Contexte du stage

Dans le cadre des recherches en extraction automatique de lexiques bilingues à partir de corpus comparables, la qualité des ressources terminologiques produites dépend beaucoup de la qualité des corpus exploités. Les corpus comparables regroupent des textes dans des langues différentes qui ne sont pas la traduction l'un de l'autre mais qui partagent un certain nombre de traits comme le domaine, le thème, domaine ou encore la période.

En domaine spécialisé, les corpus comparables sont d'une taille modeste (autour de 1 million de mots) ce qui est un frein aux méthodes numériques exploités. Dans ce contexte, différentes solutions sont possibles comme :

- améliorer la qualité des corpus comparables en s'appuyant sur une mesure de comparabilité (Li et Gaussier, 2010)
- utiliser des corpus comparables déséquilibrés lorsque les données disponibles le permettaient (Morin et Hazem, 2014)

Il existe une autre alternative qui n'a pas été abordée dans la littérature et qui consiste à exploiter des corpus comparables regroupant plus de deux langues. L'objectif de ce travail est de proposer plusieurs stratégies permettant d'exploiter des corpus comparables multilingues et *in fine* d'en augmenter la qualité des lexiques bilingues extraits avec une paire de langues. Les pistes à explorer seraient la traduction simultanée d'un lexique disponible dans une langue dans plusieurs langues (multi-alignement) ou la traduction multi-sources (Och and Ney, 2001), c'est-à-dire un lexique à traduire dans une langue mais qui exploite des traductions existantes dans une autre langue.

Plan de travail

Le travail se déroulera en quatre étapes principales :

1) Etat de l'art sur
    - l'alignement à partir de corpus comparables : méthode directe et stratégies de reclassement de candidats
    - le multi-alignement appliqués aux corpus parallèles (Simard, 1999).
2) Définition d'une méthode de référence reposant sur une réitération des alignements obtenus avec la méthode directe sur un corpus comparable bilingue.
3) Proposition d'une méthode de triangulation pour réaliser un multi-alignment ou

une méthode de tuilage pour tirer partie d'un alignement existant.

4) Expérimentation et évaluation sur plusieurs corpus comparables dans le domaine des énergies éoliennes, du cancer du sein, de la vulcanolgie sur le Français/Anglais/Allemand.

Références

Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational LinguisTics* (*COLING*), 644–652.

Emmanuel Morin and Amir Hazem (2014). Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction. In *Proceedings, 52nd Annual Meeting of the Association for Computational Linguistics (ACL*), 1284–1293, Baltimore, Maryland, USA

Franz Josef Och & Hermann Ney (2001). Statistical multi-source translation. *MTSummit* 2001.

Michel Simard, Text-translation alignement (1999). Three languages are better than two. *EMNLP*, 2-11.