Statistical alignment of bitexts: challenges, methods and applications

François Yvon[1]

[1]LIMSI/CNRS et Université de Paris Sud– yvon@limsi.fr

Abstract

Recent years have witnessed a rapid dissemination of multilingual Natural Language Processing technologies, such as online machine translation services and cross-lingual search engines. These tools rely primarily on the statistical analysis of massive quantities of bitexts, i.e. of composite, multilingual documents comprising a source text and its translation(s) in one or several target language(s). As a first step towards turning these documents into useful resources, such as multilingual dictionaries or terminologies, translation memories, or statistical translation models, the bitexts need to be aligned, so as to recover correspondences between small chunks of texts, corresponding to paragraphs, sentences, syntactic phrases, segments or even words. The computation of such correspondences is a computationally challenging task, which needs to be performed efficiently, on very large batches of data.
In this talk, I will review the main families of statistical alignment techniques and illustrate their use in several applications, with a specific emphasis on sampling-based alignment algorithms, which have been recently developed and used to build multilingual dictionaries and machine translation systems.

Résumé

Ces dernières années ont vu l'émergence et la rapide dissémination des technologies du traitement automatique des langues, comme par exemple les services de traduction automatique en ligne, ou encore les moteurs de recherche d'information cross-lingue. Ces outils s'appuient principalement sur l'analyse statistique de larges bitextes, c'est-à-dire de documents multilingues associant un texte source avec sa ou ses traduction(s) en langue(s) cible(s). La première phase du traitement permettant de transformer ces corpus en ressources utilisables, que ce soit sous la forme de dictionnaires ou de terminologies multilingues, de mémoires de traduction ou encore de modèles de traductions pour les systèmes de traduction statistiques consiste à construire un *alignement* de ces bitextes. Cet alignement permet d'identifier des correspondances entre fragments de textes en langue source et cible, fragments qui correspondent à des paragraphes, des phrases, des syntagmes, voire à des segments non interprétables linguistiquement ou à des mots isolés. Le calcul de ces correspondances est une tâche qui pose des problèmes computationnels difficiles, et qui doit néanmoins être accomplie efficacement sur de très gros ensembles de données.
Dans cet exposé, je passerai en revues les principales familles de méthodes statistiques utilisées pour réaliser ces alignements et illustrerai leur utilisation dans plusieurs cadres applicatifs. Je m'attarderai plus particulièrement sur des propositions récentes qui s'appuient sur des techniques d'échantillonnage et des calculs d'associations statistiques très simples et qui ont été utilisées avec succès pour extraire des dictionnaires bilingues et alimenter des systèmes de traduction statistique.

Keywords: bitext alignment, sentence alignment, word alignment, statistical association measures