Extraction des séquences fermées fréquentes à partir de corpus parallèles : Application à la traduction automatique

Chiraz Latiri*, Cyrine Nasri**, Kamel Smaili***, Yahya Slimani**

*Unité de Recherche URPAH, Faculté des Sciences de Tunis, Tunisie
chiraz.latiri@gnet.tn
**Unité de Recherche MOSIC, Faculté des Sciences de Tunis, Tunisie
cyrine.nasri@gmail.com, yahya.slimani@fst.rnu.tn
***LORIA, Groupe PAROLE, Vandoeuvre, France
kamel.smaili@loria.fr

Résumé. Dans cet article, nous abordons la problématique d'extraction de séquences fréquentes à partir de corpus de textes parallèles en prenant en compte l'ordre d'apparition des mots dans une phrase. Notre finalité est d'exploiter ces séquences dans la traduction automatique (TA). Nous introduisons ainsi la notion de règles associatives inter-langues (RAIL) et nous définissons notre modèle de traduction à base de ces associations. Nous décrivons également les différentes expérimentations conduites sur le corpus EUROPARL afin de construire à partir des RAIL une table de traduction bilingue qui est intégrée par la suite dans un processus complet de TA.

1 Introduction

Initialement introduit dans (Srikant et Agrawal, 1996), l'extraction de motifs séquentiels reste intuitivement applicable à tout domaine dans lequel il existe une relation d'ordre entre les éléments. Dans cet article, nous proposons d'aborder la problématique d'extraction de séquences fréquentes, en se plaçant dans le domaine de la fouille de données textuelles (Berry, 2008) et en prenant en compte l'ordre d'apparition des mots dans une phrase, et ce à partir d'un corpus parallèle aligné au niveau de la phrase. Nous nous intéressons particulièrement aux approches basées sur un parcours en largeur et dédiées à l'extraction *des motifs séquentiels fermés fréquents* (Yan et al., 2003; Wang et Han, 2004; Chang, 2004). Ces dernières évoquent le problème de la redondance des sous-séquences extraites et ayant le même support que d'autres super-séquences fréquentes.

Notre finalité est de déployer les séquences fréquentes de mots dans la traduction automatique (TA) (Brown et al., 1993). Notre choix pour la TA est justifié par le fait que des travaux récents en traduction automatique statistique confirment que les modèles fondés sur des séquences de mots (Och et al., 1999; Koehn, 2004) obtiennent des performances significativement meilleures que ceux fondés sur des mots simples (Brown et al., 1993).

La suite de l'article est structuré comme suit : la section 2 présente un bref aperçu du processus de recherche des séquences fermées fréquentes adapté aux corpus textuels, suivie de la section 3 qui décrit l'application à la TA en introduisant la notion de règles associatives