Master 2 ATAL

Alignement de chaînes et de textes

Implémentation de la méthode standard

Étudiante :
Coraline Marie

Encadrant :
Emmanuel Morin

5 décembre 2014

UNIVERSITÉ DE NANTES

Table des matières

Introduction

Principalement utilisée pour la traduction automatique ou pour la recherche d'informations, *l'alignement de chaînes et de textes* est une discipline de traitement des langues, qui permet de mettre en correspondance des unités textuelles, par processus automatiques.

La traduction automatique est un outil aujourd'hui commun, pratique, et facilement accessible sur Internet. Cependant, il reste encore imparfait, car il n'est pas capable de donner de bonnes traductions dans toutes les langues, et pour tous les mots. Les thermes techniques sont notamment difficiles à traduire, car ils sont rarement utilisés, et également peu présents dans les dictionnaires.

La méthode standard de Pascale Fung et de Kathleen McKeown[1] propose un algorithme qui permet de traduire des termes techniques inconnus, à l'aide de corpus comparables. Ce rapport présente donc les résultats de l'implémentation de cette méthode sur deux corpus comparables (français et anglais), ainsi que le travail de prétraitement des corpus, réalisé dans le but d'améliorer les résultats de traduction de la méthode standard.

1 Présentation de la méthode standard

L'algorithme de la méthode standard détaillé dans l'article de Fung et de McKeown[1], se déroule en quatre temps :

1. La première étape consiste à construire une liste bilingue de paires de termes connus. Cette liste servira plus tard de *dictionnaire*, pour la traduction des vecteurs de contextes.
2. Lors de la seconde étape, un vecteur de contexte doit être construit pour chaque terme inconnu (sans traduction) de la langue source. Ces vecteurs sont ensuite traduit dans la langue cible, à l'aide du dictionnaire créé lors de la première étape.
3. Pour la troisième étape, un vecteur de contexte est créé pour chaque terme du corpus de la langue cible. Ils serviront d'éléments de comparaison lors de la quatrième étape.
4. Pour finir, chaque vecteur de contextes traduits est comparé avec les vecteurs de contexte des termes de la langue cible : s'ils sont similaires, cela signifie qu'ils sont traduction l'un de l'autre.

2 Prétraitement des corpus

Avant même de commencer le traitement des données, il faut au préalable nettoyer les corpus. Ces derniers sont souvent bruités et incompatible avec l'algorithme sans prétraitement.

2.1 Le corpus source

Pour ce projet, le corpus source choisi est en français, et traite du cancer du sein. Il est également annoté avec des étiquettes morphosyntaxiques. Ainsi

lors du prétraitement, les étiquettes morpho-syntaxique sont d'abord supprimées pour ne garder que le lemme. Les accents sont ensuite supprimés, et les majusules sont converties en minuscules. Tous les termes contenant des éléments de ponctuations, des symboles ou des chiffres sont également supprimés. Dans ce corpus, il existe également des phrases écrites en anglais (citations, liens ...), qu'il faut supprimer. De plus, afin d'améliorer le temps de traitement, ainsi que les calculs des vecteurs de contextes, les *stopwords* sont supprimés.

2.2 Le corpus cible

suppression des information inutiles suppression des phrases française suppression des majuscules supp des mots outils suppr de la ponctuation et des chiffres

2.3 Le dictionnaire

suppression des information inutiles remplacement des espaces par _

2.4 La liste des mots à traduire

suppression des mot absent du corpus français suppression des trad absent du corpus anglais

3 construction du dictionnaire de cognats

suppression des termes à préfixe inter - semi - intra - anti - poly - post - micro - radio - méta - multi suppression des mots composer - combien de cognats 4 grammes combien de cognats 5 grammes

4 Vecteurs de contextes

construction taille de la fenêtre traduction sans cognat avec cognat mesures de similarité

5 Résultats

méthode standard classique avec pondération normale méthode standard avec table de contingence méthode standard avec cognats

Création des cognats ajout du dictionnaire des cognats

Délimitation de la zone de test top 1 top 5 top 10 agrandissement du dictionnaire avec les synonymes

Amélioration des résultats limitation des vecteurs de contexte dans une phrase +0.5% suppression des mots outils après la construction du vecteur de contextes +0.1% changement de la taille de la taille du vecteur +-0.0 %

Conclusion

Références

[1] Pascale Fung and Kathleen Mckeown. Finding terminology translations from non-parallel corpora, 1997.