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 termes 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, 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 contextes 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 sans prétraitement ils sont incompatible avec l'algorithme.

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 morpho-syntaxiques. Ainsi lors

du prétraitement, les étiquettes morpho-syntaxiques sont d'abord supprimées pour ne garder que le lemme. Les caractères accentués sont remplacés par des caractères classiques, et les majusules sont converties en minuscules. Tous les termes contenant des éléments de ponctuations, des symboles ou des chiffres sont également supprimés. Par ailleurs, il existe dans ce corpus des phrases écrites en anglais (citations, liens, ...), qu'il faut retirer manuellement. De plus, afin d'améliorer le temps de traitement, ainsi que les calculs des vecteurs de contextes, les *stopwords* et les hapax sont également effacés.

2.2 Le corpus cible

Afin d'obtenir des corpus comparables bilingues, le corpus cible choisi est en anglais, et traite également du cancer du sein. Comme le corpus source, il est annoté avec des étiquettes morpho-syntaxiques qu'il faut au préalable retirer, pour ne garder que le lemme. Les majuscules et tous les termes contenant des éléments de ponctuations, des symboles ou des chiffres sont supprimés, tous comme les phrases écrites en français, les *stopwords* et les hapax.

2.3 Le dictionnaire

Le dictionnaire français/anglais utilisé est légèrement bruité, et nécessite quelques modifications. Les étiquettes morpho-syntaxiques sont au préalable supprimées pour ne garder que le lemme des mots sources et des mots cibles. Puis, les espaces séparant les termes des expressions traduites (exemple : "a priori", "trou noir", "get off", ...) sont remplacés par des "_". Ceci est fait dans le but de respecter les conventions d'anotation des corpus source et cible.

2.4 La liste des mots à traduire

L'évaluation de la méthode standard de Fung et de McKeown se fait par l'intermédiaire d'une liste de termes techniques absents du dictionnaire. Cependant, il est nécéssaire de vérifier que ces termes soient présents dans le corpus source, et que leur traduction soit également présente dans le corpus cible. Si ce n'est pas le cas, l'algorithme n'a aucune chance de traduire un terme qu'il ne rencontre pas dans les corpus.

3 Construction du dictionnaire de cognats

L'une des pistes évoquée pour améliorer les résultats de la méthode standard, est l'utilisation d'un dictionnaire de cognats. En effet, un dictionnaire de cognats construit à partir de corpus comparables peut aisément renforcer le dictionnaire de base, en apportant de nouvelles traductions. Cependant, la construction d'un tel dictionnaire nécessite quelques précautions, comme par exemple la suppression des termes préfixés par :

- inter
- semi
- intra
- anti
- poly
- post
- micro
- radio
- méta
- multi

Ainsi, deux dictionnaires de cognats on été construits pour les tests :
- 4-grammes : 32265 termes alignés
- 5-grammes : 15056 termes alignés
- 6-grammes : 7695 termes alignés

Les résultats obtenus en utilisant ces trois dictionnaires seront présenté dans la partie résultats.

4 Vecteurs de contextes

4.1 Construction des vecteurs de contextes

La méthode standard utilise les vecteurs de contextes comme base pour la traduction automatique. Chaque termes à traduire doit donc avoir un vecteur de contextes qui lui est propre. Pour cela, il suffit de parcourir l'intégralité du corpus source, et de récupérer tous les termes qui entourent chaque occurence du mot qu l'on souhaite traduire. Cependant, les termes récupérés doivent être proche du mot à traduire, c'est à dire qu'ils doivent être situés au maximum 3 mots avant ou 3 mots après chaque occurence du terme à traduire.

En ce qui concerne les vecteurs de contexes anglais, la méthode de construction est la même. Cependant il faut construire un vecteur de contexte pour tous les termes présents dans le corpus cible. Ces vecteurs serviront ensuite d'éléments de comparaison pour les termes à traduire, il faut donc qu'ils soient construit de la même manière. Il y a environ 7900 vecteurs de contextes pour une centaine de termes à traduire.

4.2 Traduction des vecteurs de contextes

Lorsque les vecteurs de contextes des termes à traduire sont construits, il est nécessaire de les traduire pour pouvoir les comparer avec ceux du corpus cible. Pour ce projet, la traduction c'est faite de deux manières différentes : avec et sans les dictionnaires de cognats.

Dans la méthode standard, seul le dictionnaire classique est utilisé pour traduire les vecteur de contextes, mais pour ce projet, nous avons décidé d'ajouter les dictionnaires de cognats afin d'améliorer la traduction. En effet, si un terme présent dans un vecteur de contexte ne peut pas être traduit par le dictionnaire classique, on utilise alors les dictionnaires de cognats pour leur trouver une traduction potentielle. Les résultats obtenu par ces différents procédés sont détaillé dans la partie Résultats.

5 Résultats

mesures de similarité

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.