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 utilisé 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 pour tous les mots dans toutes les langues. 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 contexte.
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 traduits 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 contexte traduit est comparé avec les vecteurs de contexte des termes de la langue cible : s'ils sont similaires, cela signifie qu'ils sont traductions 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 incompatibles 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 contexte, 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. A l'instar des phrases écrites en français, des stopwords et des hapax, les majuscules sont converties et les termes contenant des éléments de ponctuations, des symboles ou des chiffres sont supprimés.

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 le caractère "_". Ceci est fait dans le but de respecter les conventions d'annotation 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 nécessaire 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 évoquées 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, trois 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 présentés dans la partie Résultats.

4 Vecteurs de contexte

4.1 Construction

La méthode standard utilise les vecteurs de contexte comme base pour la traduction automatique. Chaque terme à traduire doit donc avoir un vecteur de contexte 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 que 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.

Après avoir construit les vecteurs de contexte des termes techniques, il est nécessaire de les traduire. Cette traduction permettra ensuite de les comparer avec d'autres vecteurs construits à partir du corpus cible. Pour ce projet, la traduction s'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 vecteurs de contexte, mais pour ce projet, il a été décidé d'ajouter des dictionnaires de cognats pour améliorer les résultats. En effet, si un terme présent dans un vecteur de contexte ne peut pas être traduit par le dictionnaire classique, on utilise alors un dictionnaire de cognats pour trouver une traduction alternative. Les résultats obtenus par ces différents procédés seront détaillés dans la partie Résultats.

En ce qui concerne les vecteurs de contexes anglais, la méthode de construction est la même que celle utilisée pour les termes techniques. Cependant il faut construire un vecteur de contexte pour tous les termes présents dans le corpus cible, car ces vecteurs serviront ensuite d'éléments de comparaison pour les termes à traduire. Il y a environ 7900 vecteurs de contexte et une centaine de termes à traduire.

4.2 Comparaisons

Lorsque tous les vecteurs de contexte sont construits, il ne suffit plus que de les comparer. Pour cela, on utilise la fonction *cosinus*, qui permet de mesurer la similarité entre deux vecteurs. Ainsi, chacun des vecteurs représentant un terme à traduire est comparé avec tous les vecteurs représentant un terme du corpus cible. La fonction *cosinus* retourne ensuite un score pour chaque comparaison : plus la valeur de ce score est élevée, plus les vecteurs analysés se ressemblent.

5 Résultats

Afin de mesurer et de comparer les différents résultats obtenus par la méthode standard de Fung et de McKeown, la précision a été mesurée sur plusieurs niveaux :

— le top 1, qui vérifie si la plus forte proposition est également la traduction attendue ;
— le top 5, qui vérifie si la bonne traduction est dans les 5 meilleures propositions ;
— le top 10, qui vérifie si la traduction attendue est dans les 10 meilleures propositions.

	Top 1	**Top 5**	**Top 10**
méthode standard seule	28,24 %	42,35 %	43,53 %
avec cognats 4 grammes	31,76 %	45,88 %	51,76 %
avec cognats 5 grammes	32,94 %	47,06 %	54,12 %
avec cognats 6 grammes	32,94 %	48,24 %	55,29 %

Conclusion

Les résultats précédents démontrent que la méthode standard de Fung et de McKeown reste très limitée. Sur les corpus comparables du cancer du sein, seulement 28 % des meilleures suggestions sont les traductions attendue. Cependant, l'ajout des différents dictionnaires de cognats démontrent que ces résultats peuvent être légèrement améliorés jusqu'à 33 %.

Néanmoins, il existe plusieurs pistes pour améliorer ces résultats. Tout d'abord, les corpus comparables utilisés sont de petite taille. L'utilisation de corpus plus importants augmenterait la précision des vecteurs de contexte, ce qui améliorerait également la précision de l'évaluation. De plus, il a été constaté que certains termes techniques tels que *letrozole* ou *raloxifene* sont des transfuges. Il est donc logique de penser que l'ajout d'un dictionnaire de transfuge en complément du dictionnaire classique augmenterait la précision des résultats.

Références

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

L'intégralité du code source est consultable et téléchargeable sur :
https://github.com/Slayerxoxo/ACT_project_std_method