Correspondance automatique de textes en langues différentes sans connaissances préalables

Bastien DÉCHAMPS - Michaël KARPE

Projet MOPSI 2017-2018 encadré par M. Renaud MARLET (Laboratoire IMAGINE)

École des Ponts ParisTech

Introduction

Dans le domaine de la traductologie, il existe de nombreuses approches et de nombreux algorithmes pour réaliser des traductions d'une langue à une autre. Chaque méthode possède ses avantages et ses inconvénients, et il convient généralement d'utiliser un compromis entre plusieurs méthodes pour obtenir un algorithme de traduction efficace.

Nous nous intéressons ici à un algorithme d'alignement de textes en langues différentes basé sur la déformation (ou dilatation) temporelle dynamique (ou *Dynamic Time Warping, DTW*). Un tel algorithme n'a besoin ni de ressources bilingues préalables, ni de connaissances de similarités entre les langues étudiées.

Il peut ainsi fonctionner sur n'importe quel couple de langues, et notamment sur des langues pour lesquelles on possède peu de ressources linguistiques. En revanche, sa précision dans la traduction est faible ; ainsi nous nous concentrerons principalement sur l'alignement de paragraphes plutôt que de mots.

Une approche avec les mains

L'approche employée dans l'algorithme de Kim Gerdes [1], bien que simple à première vue, n'en est pas moins originale. Elle consiste à travailler sur les occurrences et la position des mots dans les textes considérés en employant le DTW.

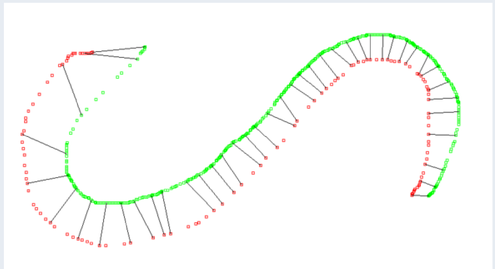

FIGURE 1: Exemple de DTW appliquée à deux séries de points.

Alignement de paragraphes

L'application du DTW dans un système de traduction peut réaliser de mauvaises associations, si l'on cherche à être trop précis, en raison de syntaxes différentes au sein des langues. Il s'agit donc, dans l'algorithme proposé par Gerdes, de se servir de la somme des alignements de mots pour procéder à l'alignement de paragraphes, et ainsi faire disparaître les éventuels signaux parasites dus à des associations erronées de mots (voir FIGURE 4).

L'alignement par longueur est une méthode d'alignement, non basée sur le DTW, qui permet de réaliser un premier alignement de paragraphes uniquement basé sur la longueur des paragraphes.

Alignement par longueur

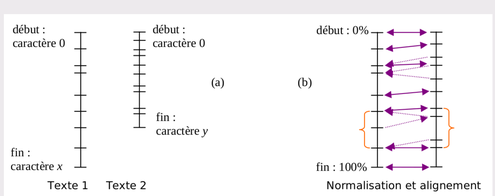

FIGURE 2: Marques de paragraphes, normalisation et alignement.

Distance de Jaro-Winkler

Pour compléter la distance DTW, on utilise la distance syntaxique de Jaro-Winkler entre 2 chaînes de caractères d_w, fonction de la distance de Jaro d_j :

$$d_w = d_j + lp(1 - d_j) \quad \text{où} \quad d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$$

- $|s_i|$ est la longueur de la chaîne de caractères s_i
- m est le nombre de caractères correspondants
- t est le nombre de transpositions
- l est la longueur du préfixe commun (maximum 4 caractères)
- p est un coefficient qui permet de favoriser les chaînes avec un préfixe commun (Winkler propose pour valeur $p = 0.1$)

Dynamic Time Warping (DTW)

De façon générale, les algorithmes DTW "cherchent à trouver l'alignement monotone (sans croisement) optimal de deux séquences de longueur variable." [1] Par monotone, on entend que l'ordre des positions des séquences avant et après DTW est respecté, mais que l'écart entre ces positions peut changer.

Les algorithmes DTW sont des algorithmes de programmation dynamique. Pour notre alignement de paragraphes, il est appliqué sur des vecteurs de récence $(p_1, p_2 - p_1, \ldots, p_n - p_{n-1}, 1 - p_n)$ où p_i est la position de l'occurrence i (en fraction de texte) du mot p considéré.

Algorithme DTW

$$W_{i+1,j+1} = |r_1 - r_2| + \min\left(W_{i,j+1}, W_{i+1,j}, W_{i,j}\right)$$

0	1	1	1	1	1	1
1						
1						
1				W		
1						
1						
1						d

FIGURE 3: Calcul de la distance DTW par chemin optimal.

Intégration dans un système d'alignement

Nous décrivons ici, de façon synthétique, l'algorithme global réalisant la mise en correspondance automatique de textes en langues différentes sans connaissances préalables, incluant l'algorithme DTW expliqué ci-dessus. Cet algorithme, décrit en détails par Kim Gerdes [1], peut être synthétisé en 4 grandes étapes :

1. Lecture et nettoyage des textes : suppression des signaux parasites (accents, espaces en trop, ponctuation...)
2. Construction des tables de hachage associant les mots à leurs différentes caractéristiques (nombre d'occurrences, indices d'apparition...)
3. Calcul pour chacun des textes des cognats internes à la langue (distance de Jaro-Winkler), des mots ou groupes de mots significatifs (fréquents et bien répartis dans le texte)
4. Application de l'algorithme DTW pour l'alignement des mots, des groupes de mots et des paragraphes

Exécution du système d'alignement

Le système d'alignement décrit précédemment a été développé en langage C++ et appliqué sur la Déclaration Universelle des Droits de l'Homme (ou *Universal Declaration of Human Rights, UDHR*) dans les langues suivantes : français, anglais, espagnol, allemand, russe. Nous présentons ici les résultats obtenus pour le couple de langues français-anglais.

Alignement de mots

Français	Anglais	DTW
respect	respect	0.0039
ces	these	0.0046
déclaration	declaration	0.0070
sans	without	0.0107
considérant	whereas	0.0115
nationalité	nationality	0.0116
libertés	freedoms	0.0124
famille	family	0.0150
société	society	0.0152
unies	united	0.0213

Français	Anglais	DTW
nations	nations	0.0228
conscience	conscience	0.0245
dignité	dignity	0.0306
religion	religion	0.0326
éducation	education	0.0375
contre	against	0.0383
article	article	0.0545
droits	rights	0.0714
présente	declaration	0.0849
religion	property	0.0979

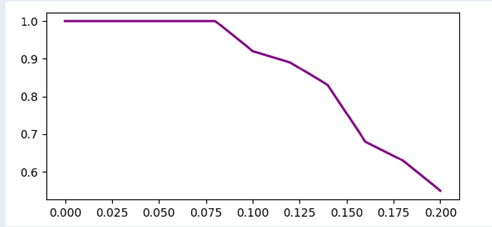

FIGURE 4: Rapport du nombre de traductions correctes sur le nombre de traductions total en fonction du seuil maximal de DTW.

Analyse des résultats

Après application de l'algorithme DTW sur différentes langues, nous avons pu constater que la mise de correspondance de mots était efficace lorsque la valeur de DTW renvoyée pour un couple de mots était inférieure à $0,1$ (celle-ci ayant été normalisée pour être comprise entre 0 et 1).

Pour le couple de langues français et anglais, on constate une seule association erronée lorsque le DTW est utilisé seul, en raison de la syntaxe différente des langues (voir ci-dessous). Il convient donc d'inclure la distance de Jaro-Winkler dans le système d'alignement pour corriger ce problème.

Utilisation de Jaro-Winkler avec DTW

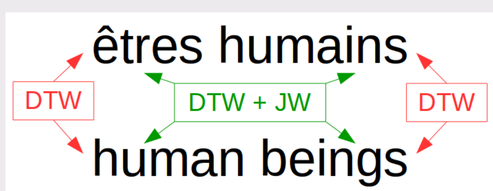

FIGURE 5: Exemple de piège tendu par l'utilisation seule de DTW.

Une fois la distance de Jaro-Winkler incluse dans le système d'alignement (étape 3), nous pouvons combiner ces associations de mots pour procéder à l'association des paragraphes de la DUDH. La DUDH est constituée de 89 paragraphes, pour un total de 2106 mots et 11663 caractères.

Exemple de textes à aligner

FIGURE 6: Déclaration Universelle des Droits de l'Homme.

Alignement de paragraphes

Les associations réciproques de paragraphes (tirets sur les traits verticaux) sont représentées par des flèches pleines, tandis que les flèches en pointillés représentent les associations dans un seul sens.

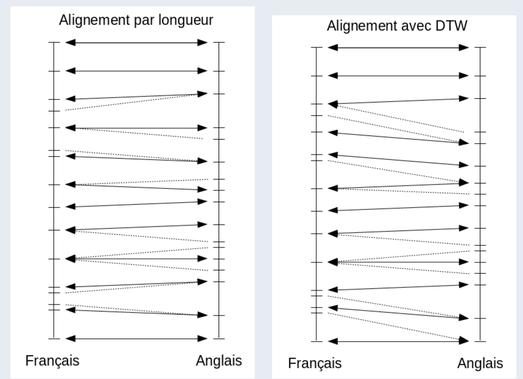

FIGURE 7: Schémas simplifiés des alignements de la DUDH.

Limites du système d'alignement

La limite principale du système d'alignement développé réside dans la précision de l'alignement, limitée à une échelle de paragraphes. L'algorithme DTW seul n'est pas suffisant pour réaliser un alignement efficace, et l'ajout d'une distance syntaxique comme Jaro-Winkler est nécessaire pour améliorer les résultats. Les résultats obtenus sont très satisfaisants malgré tout.

Références

[1] Kim Gerdes.
L'alignement pour les pauvres : Adapter la bonne métrique pour un algorithme dynamique de dilatation temporelle pour l'alignement sans ressources de corpus bilingues.
9e Journées internationales d'Analyse statistique des Données Textuelles, Mars 2008.

[2] Meinard Müller.
Information retrieval for music and motion.
Chapter 4 : Dynamic Time Warping, Septembre 2007.