Alignement non-supervisé de bitextes

Philippe FERREIRA DE SOUSA, Victoriya KASHTANOVA et Nada SOUSSI

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

École des Ponts ParisTech

Objectifs

L'alignement de bitextes consiste à aligner des textes écrits dans deux langues différentes, supposés être des traductions l'un de l'autre.
L'objectif est de faire correspondre les phrases et paragraphes.
Les corpus bilingues alignés sont essentiels dans l'élaboration de ressources bilingues comme dans tout travail traductologique, mais l'alignement nécessite en lui-même des ressources bilingues ou d'importantes interventions de locuteurs bilingues.

Introduction

Ces recherches s'intègrent dans un projet de dévelopement de site internet qui mette à disposition en tant que service les algorithmes conçus pour aligner des textes dans des langues différentes. En vue de déboucher sur un auto-renforcement par génération d'un dictionnaire, nos efforts se concentrerons sur un couple de langue anglais / français, en essayé de perdre le minimum de généralité.

Production

- **Service en ligne** : `https://alignment.fdesousa.fr`
- **Projet opensource** : `https://github.com/PhilippeFerreiraDeSousa/bitext-matching`

Modèle probabiliste conjoint basé sur des phrases

Modèle mathématique

On cherche à réaliser l'alignement statistique des phrases dans un bitexte à travers une compréhension des relation probabilistes entre le language source et le langage cible.
On considère l'ensemble des phrases dans la langue source e et dans la langue cible f *(foreign)*

$$\begin{aligned} e_1^I &= e_1, \ldots, e_i, \ldots, e_I \\ f_1^J &= f_1, \ldots, f_j, \ldots, f_J \end{aligned} \tag{1}$$

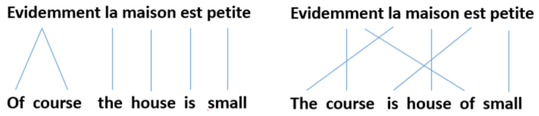
Les phrases de la langue source de la langue cible contiennent un certain nombre -I et J respectivement - de tokens (labels).
On suppose que les tokens peuvent être alignés les uns aux autres et on note l'ensemble des alignements possibles A. Chaque alignement de i à j (langue source vers langue cible) est indiqué par a_i qui contient l'indice du token correspondant j dans de la langue cible.

$$\begin{aligned} &A \subseteq \{(i,j) : i = 1, \ldots, I; j = 1, \ldots, J\} \\ &i \to j = a \\ &j = a_i \end{aligned} \tag{2}$$

- **IBM 1**
  La vraisemblance d'un alignement assignant un certain token dans la phrase f à un token de la phrase e est donnée par les équations :

$$Pr(\ e_1^I \mid f_1^J\ ) = \sum_{a_1^I} Pr(\ e_1^I, a_1^I \mid f_1^J\ )$$

- **IBM 2**
  Le modèle 1 ne permet pas de différencier entre les alignements qui alignent les mots présents sur les extrémités opposées des phrases des alignements qui sont plus proches. Le modèle 2 ajoute cette distinction.



En effet les probabilités des deux alignements sont identiques avec le Modèle 1 qui ne permet donc pas de faire la distinction. Le Modèle 2 corrige ce défaut en distinguant dans l'alignement l'assignation token par token et puis la cohérence de l'alignement selon l'équation :

$$Pr(\ e_1^I, a \mid f_1^J\ ) = \Pi_{j=1}^{l_e} t(\ e_j | \mathrm{v} f_{a_j}) p(a_j \mathrm{v} j, l_e, l_f)$$

t : probabilité de traduction.
p : probabilité d'alignement.
$l_e,\ l_f$: longueur de la phrase source et cible (respectivement).

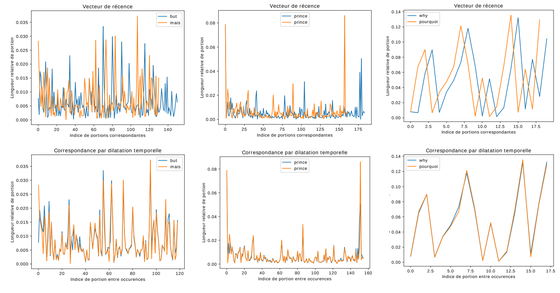
Modèle

- Texte composé de segments séparables
- Les mots ont un nombre fini de formes
- Sous-ensemble lexical non-négligeable à correspondance bijective
- Dilatation sans déchirement (traduction fidèle)

Alignement par dilatation temporelle

Signal d'un mot

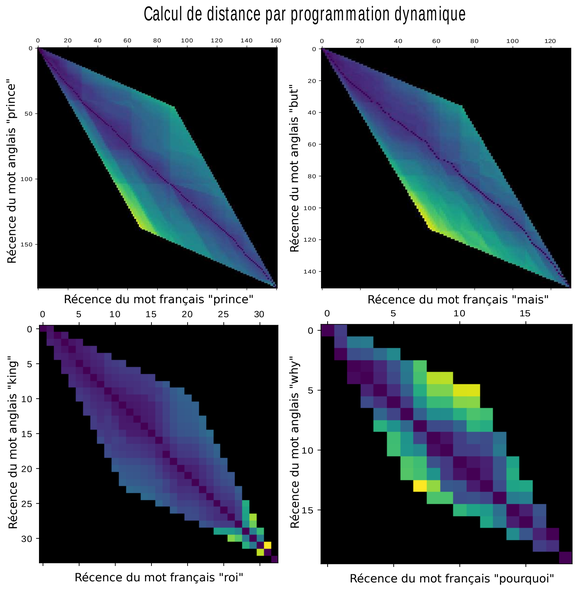
- **Bitexte d'essaie** : Le Petit Prince, Saint Exupéry, 16930 / 14702 mots
- **Mise en correspondance de cognats** (c_{fr}^i), (c_{en}^i) :
  Mots, groupes de mots ou ensemble des inflexions d'un mot.
- **Vecteur de récence** d'un cognat :
  $r_c^i = [p_1^i, p_2^i - p_1^i, \ldots, p_n^i - p_{n-1}^i, 1 - p_n^i]$ où $(p_j^i)_{1 \le j \le n}$ est le vecteur des positions relatives du cognat c^i.



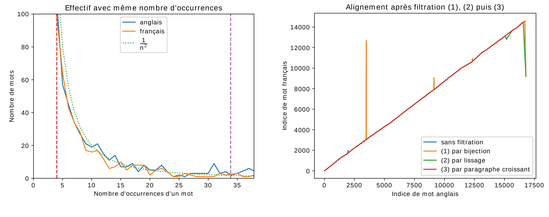
- **Choix d'une distance** d
  Entre 2 cognats de langues différentes c^k et c^l :

$$\begin{cases} l(0,0) &= 0 \\ l(i,j) &= \min(l(i-1,j-1) + |p_i - p_j|, l(i-2,j-1) \\ &\quad + |p_i + p_{i-1} - p_j|, l(i-1,j-2) + |p_j + p_{j-1} - p_i|) \\ d(c^k, c^l) &= \dfrac{l(n,m)}{freq_k^f} \end{cases}$$

On fait correspondre ainsi les portions entre 2 cognats, en s'autorisant à ignorer une occurrence ponctuellement. Lorsque la fréquence du cognat est grande ($> 0.1\%$), on peut ajouter dans le minimum les termes $l(i-1, j-3) + |\Sigma_{t=j-2}^{j} p_t - p_i|)$ et $l(i-3, i-3) + |\Sigma_{t=i-2}^{i} p_t - p_j|)$ pour permettre de sauter ponctuellement 2 occurences à la suite. On peut se restreindre à un sous-ensemble des états à parcourir et ajuster le paramètre t pour favoriser les correspondances de basses ou hautes fréquences.



- **Couples de cognats à tester**
  En dessous de 4 occurrences, il y a beaucoup de candidats (loi de Zipf) et des signaux pauvres, au delà d'une fréquence de 0.002, la valeur sémantique diminue (articles, conjonctions, particules de verbes propositionnels en anglais).
  Ensuite on se restreint à des couples de fréquences proches ($\frac{f_i}{f_j} < 1.25$).
  On estime d'abord un seuil à partir duquel on accepte la mise en correspondance, sinon le backtracking est inutile. On applique l'algorithme dynamique en entier pour les couples dont le mot en anglais apparait 4 fois, et on accepte les $\frac{nb_{mots}}{200}$ meilleures couples. On fixe le seuil de distance en conséquence.
- **Filtration**
  Les points de correspondances entre les deux textes gagnent à être filtrées pour enlever les abérations, les erreurs et rendre l'alignement de ces cognats croissant selon les textes. On peut d'abord pour chaque cognat sélectionné, le faire correspondre avec son cognat le plus proche si plusieurs ont été sélectionnés (hypothèse de lexique bijectif). Puis on peut lisser en bornant la dérivée discrète (si traduction proche), et enfin rendre les correspondances croissantes par index de paragraphe poru pouvoir enfin interpoler l'alignement des paragraphes restants.



- **Renforcement**
  On établit plus de correspondances en mettant en relation les couples de mots qui sont dans le plus de phrases qui ont été alignés ensemble.
- **Réduction**
  Si le texte est très grand, il est judicieux d'effectuer quelques correspondances de cognats, puis d'utiliser le meilleure pour partitionner le bitexte en instances moindres du problème, afin de rendre la méthode viable en temps comme en fiabilité, la partie dynamique étant de complexité quadratique.
- **Enrichissement lexical**
  On peut enrichir le lexique an y ajoutant des groupes de cognats consécutifs récurants et ainsi faire correspondre des mots composés ou expressions.

Apprentissage

Le projet dans sa version non-supervisée et disponible comme service en ligne in-line, peut permettre d'établir un dictionnaire utilisable dans une optique supervisée pour préaligner le texte pour passer directement à l'étape de renforcement ou de réduction du problème de dilatation.

Résultats

Dilatation temporelle
475 correspondances, 360 après filtration, soit 294 phrases alignées sur 1636 (fr) avec 95% de fiabilité avant renforcement et interpolation sur l'ensemble du texte.
Modèle IBM 2
281 correspondances, soit 158 phrases alignées sur 1636 (fr).

Conclusion

Les recherches dans le domaine de l'alignement non-supervisé de bitextes semble être orienté vers les méthodes statistiques. Le Modèle 1 d'IBM est un modèle relativement simple et rapide à entraîner (mots uniquement). Cependant, ces défaillances montrent que l'espace des alignements possède plusieurs maxima locaux, le Modèle 2 permet d'affiner la sélection en introduisant une considération de la position absolue des mots.
On constate que la dilatation temporelle donne de meilleurs résultats avec une marge d'amélioration importante.

Bibliographie

[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.

[2] Daniel Marcu and William Wong.
A phrase-based, joint probability model for statistical machine translation.

[3] Yarin Gal et Phil Blunsom.
A systematic bayesian treatment of the ibm alignment models.

[4] Antoine de St Exupéry.
Le petit prince.

[5] Tomáš Odaha.
The Little Prince.