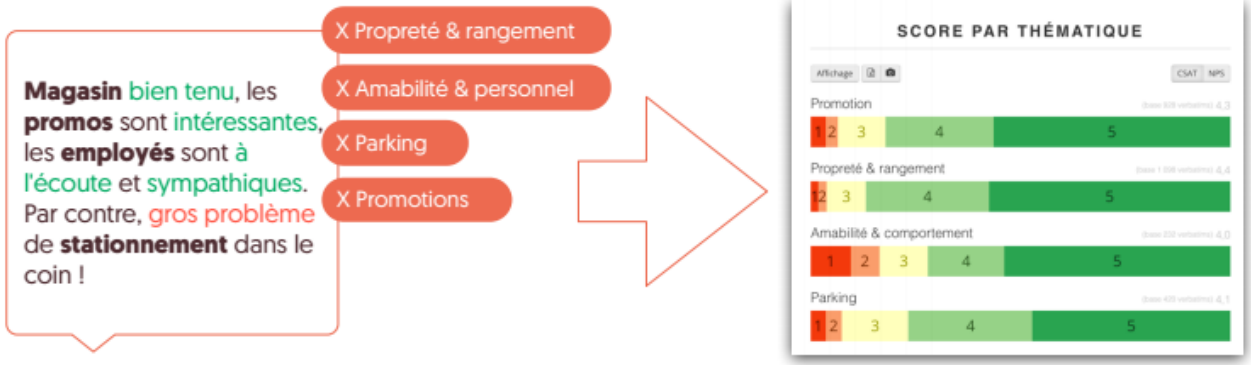
Alignement de termes de longueur variable en corpus comparables spécialisés

Jingshu Liu (LS2N, Dictanova)
Emmanuel Morin (LS2N)
Sebastián Peña S. (Dictanova)

Plan

Introduction



- Contexte du travail : industrie, logiciel d'analyse sémantique.
- Pour quoi alignement ?
  - Faciliter le passage d'une langage à une autre.
  - Permettre une projection des termes de différentes langues dans un espace commun.
- La définition des termes dans notre cas réel est plus "large". (*nombreux choix*)

Introduction

- termes simples, *apple* → *pomme*
- termes complexe compositionel, *repair status* → *état réparation*
- termes complexe non-compositionel, *axillary dissection* ↛ *dissection axillaire*, → *curage axillaire*
- alignment de longueur variable, *ankle boot* → *bottine* , *rotor shaft* → *arbre*, ce phénomène est beaucoup plus fréquent entre certaines paires de langue par exemple en/zh.

Introduction

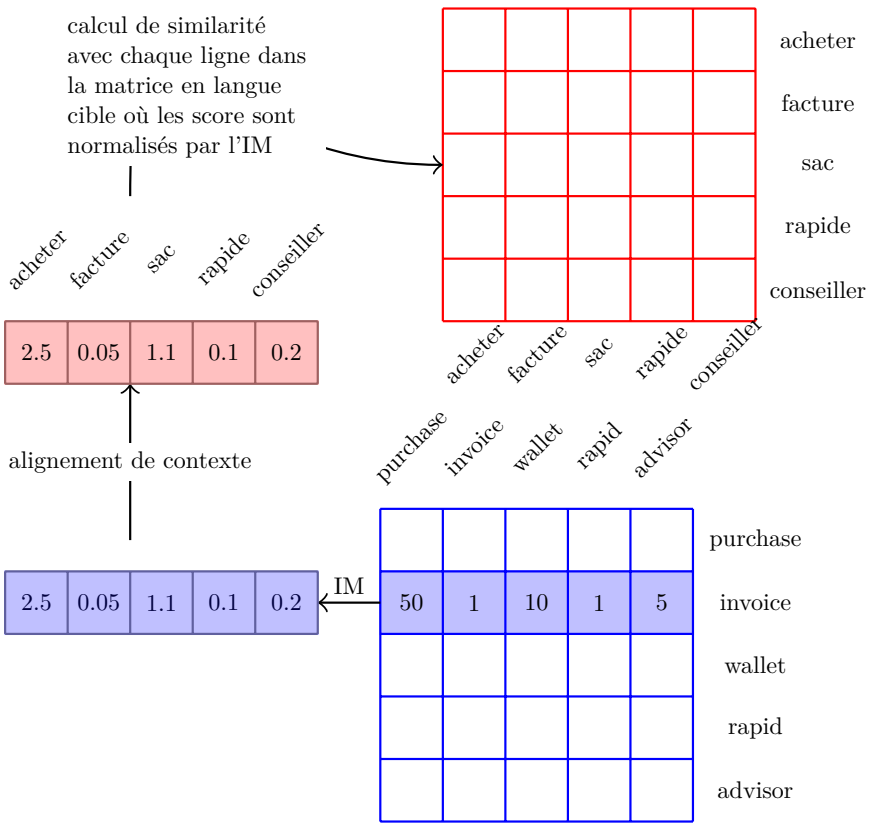
Objectif : un système capable d'aligner un terme en langue source avec ses traduction autorisées indépendamment de longueur.
Concrètement nous proposons une liste de candidats ordonnés par une mesure similarité. Nous avons utilisé le Cosinus parce qu'il permet la parallélisation à l'aide de librairies d'algèbre linéaire.

invoice :	
facture	0.6
facture d'achat	0.5
récipissé d'achat	0.45
achat	0.4

Approche standard (AS)

- Approche distributionnelle, approche par alignement de contexte, approche standard
  1. Construction de matrice de co-occurrence. Les co-occurrences sont ensuite normalisées par une mesure d'association (IM).
  2. Alignement de contexte.
  3. Calcul de similarité (Cosinus).

Approche standard (AS)



Remarques:

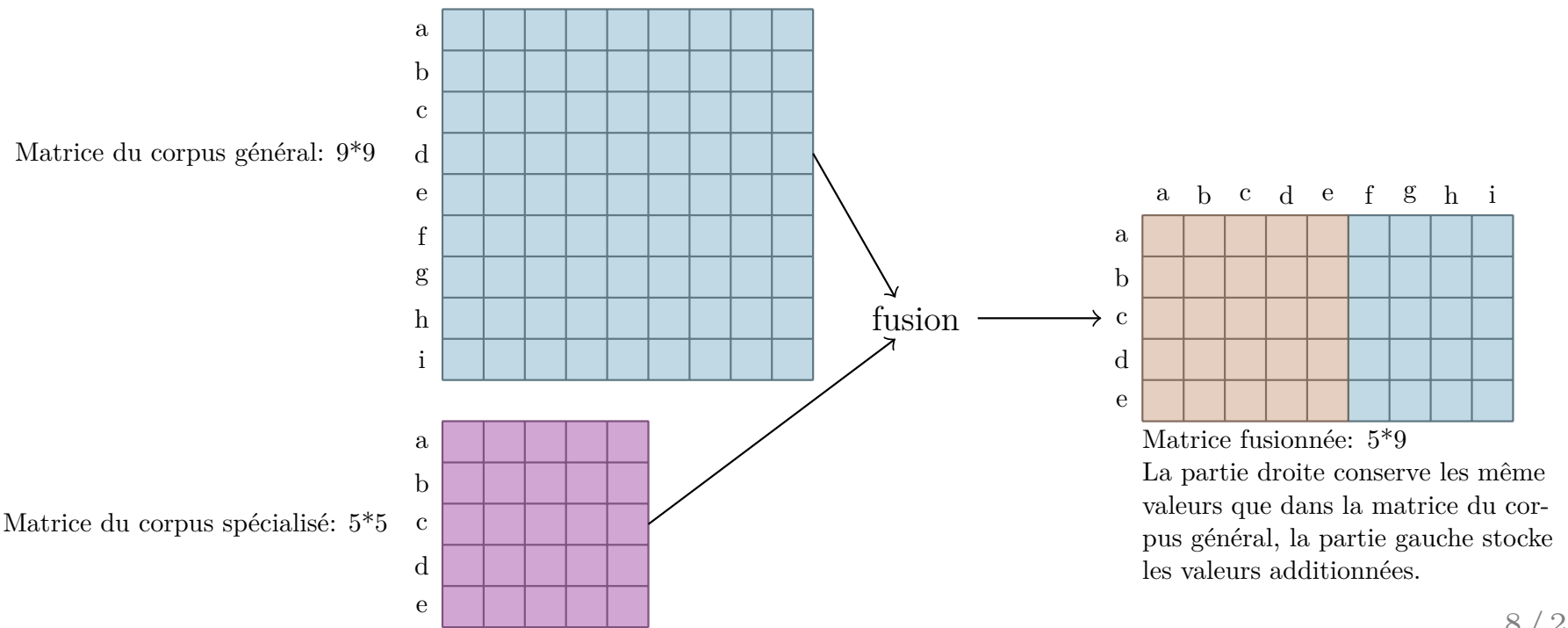
- Représentation vectorielle à partir d'une matrice de cooccurrence
- Lissage par une mesure d'association (IM)
- Traits spécifiques pour les vecteurs de contextes
- Projection des vecteurs par alignement de contextes

Approche standard sélective (ASS)

$$\forall w \in S \cap G, \forall c \in S \cap G,\ \ cooc(w,c) = cooc_S(w,c) + cooc_G(w,c)$$

(Hazem et Morin, 2016)

Pourquoi? Les occurrences des mots ne sont pas statistiquement fiables à cause de la petite taille des corpus spécialisés.



Notre proposition: pondération en fonction de la distance (PFD)

$$PFD(w, c) = g(c|w) \times cooc(w, c)$$
$$\text{où } \; g(c|w) = \Delta(w, c)^{-\lambda}, \; \lambda \in [0, 1]$$

- Pourquoi ? Plus un mot central est éloigné d'un mot du contexte, moins ils sont sémantiquement liés. Par exemple:
  - *vous trouverez les photos demandées ainsi que la facture d'achat suite à un craquement du cuir*
    si $\lambda = 0.25$, $g(achat|facture) = 1$, $g(achat|trouver) = 3^{-0.25} = 0.76$
- Quelle étape ? Au moment de la construction de matrice de co-occurrence.

Notre proposition: information mutuelle pondérée (IMP)

$$IMP(w, c) = f(cooc(w, c)) \times IM(w, c)$$

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, \alpha \in [0, 1], \text{ si } x < x_{max} \\ 1 \text{ sinon} \end{cases}$$ (Pennington et al, 2014)

- Pourquoi ? L'IM surestime les faibles occurrences et sous-estime les hautes occurrences.
- Quelle étape ? Au moment de la normalisation par l'IM.
- Pénalisation pour les coocurrences inférieures à x_{max}.

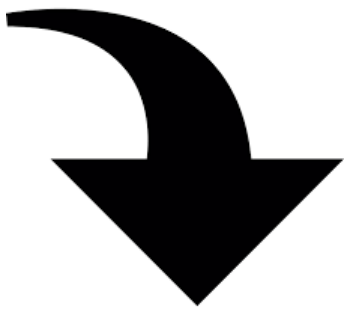
Approche compostionnelle (AC)

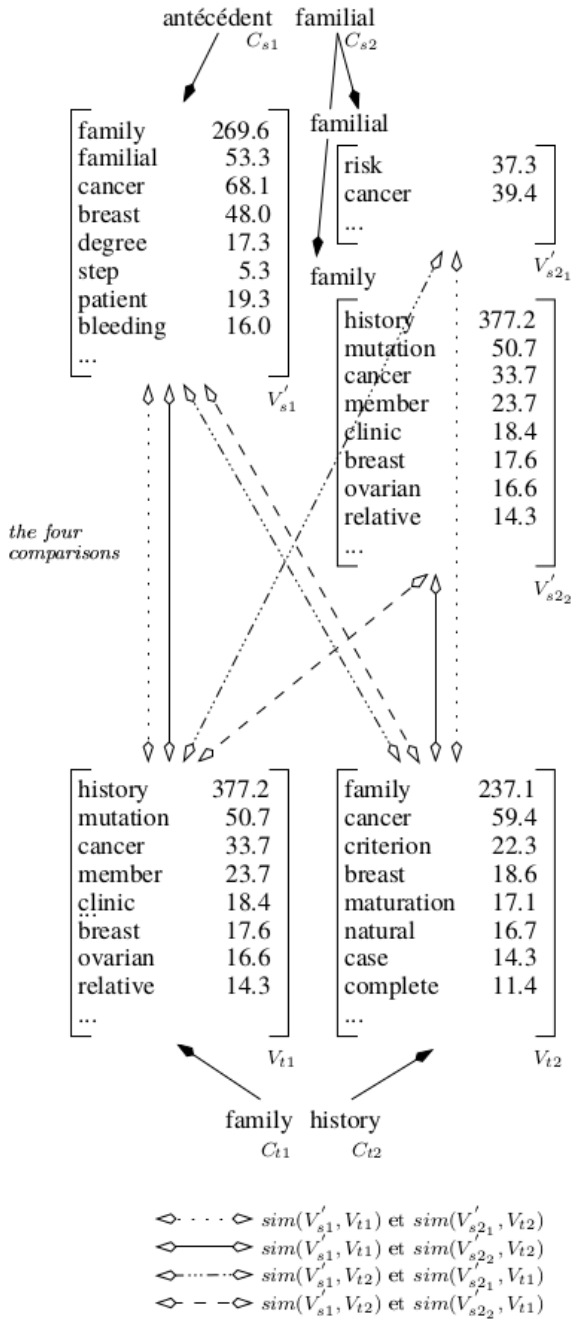
- Traduction de chaque élément d'un terme complexe via un dictionnaire.

  repair status → *statut réparation*, *réparation statut*, *position réparation*, *réparation position*, *statut réparer*, *réparer statut*, *position réparer*, *réparer position*
- Remarques:
  - Incapacité pour aligner les termes dont un mot composant n'est pas dans le dictionnaire.
  - Limite sur la gestion de l'alignement de termes de longueurs variables.
  - Complexité factorielle pour calculer toutes les combinaisons possibles.

Approche compositionnelle étendue

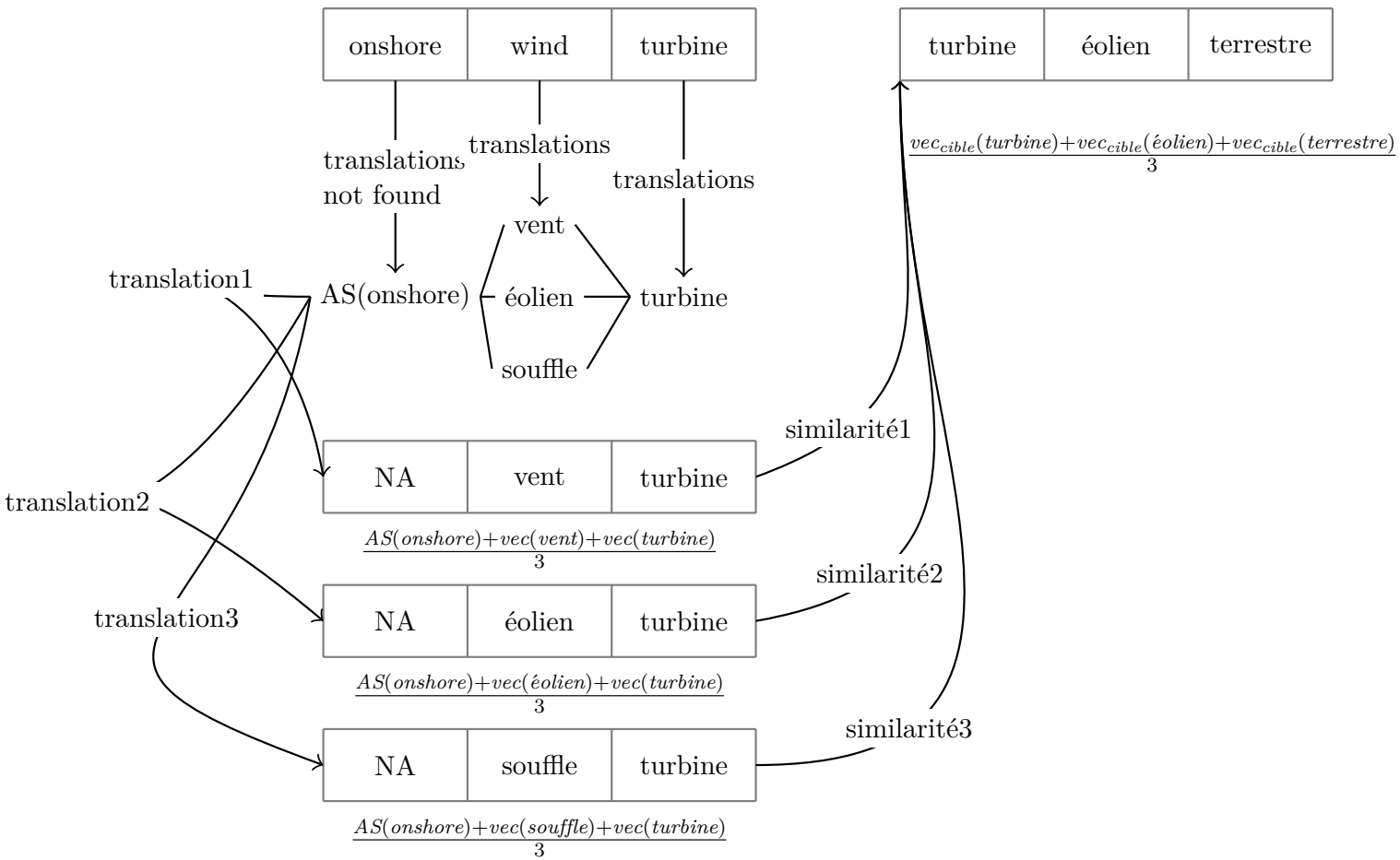
- Morin et Daille, 2012. Idée : traduire chaque mot composant par son vecteur de contexte.
  - Si le mot existe dans le dictionnaire => vecteur de contexte de la traduction obtenu directement par la matrice de cooccurrence en langue cible.
  - Sinon => vecteur de contexte obtenu par l'AS.
  - Un schema d'exemple




antécédent
C_{s1}
familial
C_{s2}
family 269.6
familial 53.3
cancer 68.1
breast 48.0
degree 17.3
step 5.3
patient 19.3
bleeding 16.0
...
V'_{s1}
familial
risk 37.3
cancer 39.4
...
family
V'_{s2_1}
history 377.2
mutation 50.7
cancer 33.7
member 23.7
clinic 18.4
breast 17.6
ovarian 16.6
relative 14.3
...
V'_{s2_2}
the four comparisons
history 377.2
mutation 50.7
cancer 33.7
member 23.7
clinic 18.4
breast 17.6
ovarian 16.6
relative 14.3
...
V_{t1}
family 237.1
cancer 59.4
criterion 22.3
breast 18.6
maturation 17.1
natural 16.7
case 14.3
complete 11.4
...
V_{t2}
family
C_{t1}
history
C_{t2}
$sim(V'_{s1}, V_{t1})$ et $sim(V'_{s2_1}, V_{t2})$
$sim(V'_{s1}, V_{t1})$ et $sim(V'_{s2_2}, V_{t2})$
$sim(V'_{s1}, V_{t2})$ et $sim(V'_{s2_1}, V_{t1})$
$sim(V'_{s1}, V_{t2})$ et $sim(V'_{s2_2}, V_{t1})$

Notre représentation pour les TC

$$vecteur(terme) = \frac{1}{n}\sum_{i}^{n}\frac{vecteur(w_i)}{||vecteur(w_i)||}$$



Données et ressources

- Corpus spécialisés: Breast Cancer(BC), Wind Energy(WE), Luxe, Cosmétique; Corpus général: News Commentary(NC).
- Dictionnaire en/fr de 243 539 entrées
- Système par patrons syntaxiques pour l'extraction terminologique
- Listes de références construites manuellement
- Candidats pour chaque terme en langue source : tous les mots simples dans le vocabulaire du corpus spécifique plus tous les termes complexes extraits par le système d'extraction.

Corpus	Nombre de mots		Taille de vocabulaire		Référence
	FR	EN	FR	EN	
BC	521 262	525 934	6 630	8 821	TS: 248
WE	314 549	313 943	6 038	7 134	TC: 73
Luxe	101 542	139 867	3 064	3 981	TC: 276, TS: 13
Cosmétique	430 106	837 579	3 913	5 592	TC: 185, TS: 14
NC	5,7 M	4,7 M	23 597	29 489	

Les termes complexes choisis sont hors le dictionnaire.

Résultats sur les TS

Sur le corpus BC :

Modèle	P@5	MAP
AS	35,5	25,9
AS + IMP	**37,1**	28,9
AS + PFD	36,3	27,4
AS + IMP + PFD	**37,1**	**29,5**
ASS	62,5	56,5
ASS + IMP	**64,1**	55,0
ASS + PFD	62,9	**57,3**
ASS + IMP + PFD	**64,1**	55,8

Table 1: Précision@k et MAP (%) pour l'alignement des termes simples sur le corpus BC.

- Amélioration homogène de PFD
- PFD et IMP ne sont pas mutuel exclusives.
- La surestimation des petites occurrences est lissée par l'ajout de données exogènes.
- Certains termes à traduire sont assez peu fréquents dans le corpus général, il est possible que pénaliser toutes les petites occurrences réduise des traits discriminants du corpus général.

Résultats sur les TC

Exemples des traductions trouvées dans le top@5 :

Corpus	Alignements (anglais → français)
Luxe	sneaker shoe → sneaker invoice → facture d'achat
Cosmétique	wide variety of choice → nombreux choix very small tight store → petit magasin
WE	greenhouse gas → gaz à effet de serre power system → système éolien de puissance

Résultats

Corpus	Modèle	P@5	MAP
Luxe	AC	25,6	24,7
	ACE + IMP + PFD	**44,6**	**40,1**
	ACE + ASS + IMP + PFD	37,0	31,2

- Meilleur résultat: ACE+IMP+PFD
- Luxe: corpus très bruité avec beaucoup de mots hors le corpus général => ASS ne fonctionne pas bien
- IMP améliore toujours les résultats: le vecteur pour un terme complexe est une moyenne => le risque que chaque élément soit peu fréquent est beaucoup moins important.

Résultats

Corpus	Modèle	P@5	MAP
Cosmétique	AC	**19,1**	15,4
	ACE + IMP + PFD	**19,1**	**16,8**
	ACE + ASS + IMP + PFD	16,6	12,9

- Meilleur résultat : ACE+IMP+PFD
- Cosmétique: corpus très bruité avec beaucoup de mots hors le corpus général => ASS ne fonctionne pas bien
- IMP améliore toujours les résultats: le vecteur pour un terme complexe est une moyenne => le risque que chaque élément soit peu fréquent est beaucoup moins important.

Résultats

Corpus	Modèle	P@5	MAP
WE	AC	68,5	61,5
	ACE + IMP + PFD	80,8	60,0
	ACE + ASS + IMP + PFD	**87,7**	**66,3**

- Meilleur résultat: ACE+ASS+IMP+PFD
- WindEnergie: corpus scientifique avec peu de mots hors le corpus général => ASS fonctionne
- IMP améliore toujours les résultats: le vecteur pour un terme complexe est une moyenne => le risque que chaque élément soit peu fréquent est beaucoup moins important.

Conclusion et perspectives

- Nous avons proposé un système capable d'aligner les termes de longueurs variables, en utilisant les vecteurs de contextes. De plus nos améliorations pour les approches état de l'art sont validés sur les données des travaux précédents et des corpus extraits de l'internet.
- Perspectives : les vecteurs embedding pour remplacer les vecteurs de contexte.
  - Les traits discriminants sont d'une certaine façon "extraits".
  - Les vecteurs denses ont une meilleure efficacité à stocker/calculer.
- Potentiellement une représentation pondérée pour les termes complexes parce qu'il y a des termes complexes dont le sens s ' appuie plus sur un élément que les autres. (sneaker shoe)
- D'autres langues linguistiquement plus éloignées.

Merci pour votre attention!