Data Selection for Bilingual Lexicon Induction from Specialized Comparable Corpora
-
Coling 2020

Martin Laville, Amir Hazem, Emmanuel Morin, Philippe Langlais

L'Extraction de Lexique Bilingue

- Créer des dictionnaires bilingues à partir de corpus de deux langues

Breast Cancer	
oncology	cancérologie
cisplatin	cisplatine
brain	cerveau
mortality	mortalité
breast	sein

Wind Energy	
efficiency	rendement
hinge	articulation
mast	mât
emission	émission
gas	gaz

Général	
revamped	remanié
revamped	réorganisé
fichier	file
ageing	vieillir
trophée	trophy

Les Corpus Comparables

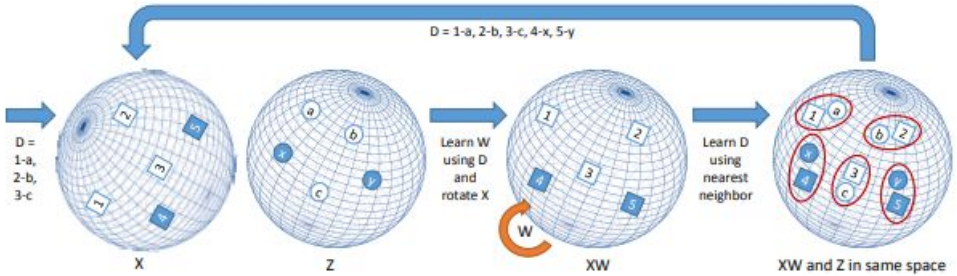
- Taille, registre, époque similaires
- Ici, ils seront au moins bilingues
- Servent à compenser la rareté des corpus parallèles

Les Corpus Spécialisés

- Encore plus durs à trouver en parallèle
- Vocabulaire plus restreint
- Mots souvent avec un seul sens
- *Breast Cancer*, *Wind Energy*...

Etat de l'art

- Embeddings entraînés séparément sur deux langues différentes
- Projection dans un même espace (Mikolov, 2013; Artetxe, 2016)
- Mesure de similarité pour classer les traductions



(Artetxe, 2017)

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

Etat de l'art

- Corpus spécialisés malgré tout de petite taille
    - Peu d'occurrences des mots
    - Représentations vectorielles moyennes
    - Résultats moyens
- On ajoute des données générales (Wikipedia…)
    - Beaucoup plus d'occurrences
    - Représentations vectorielles de bien meilleure qualité
    - Résultats très intéressants
- MAIS :
    - Introduction de polysémie
    - Temps de calcul beaucoup plus élevés

	Breast Cancer	Breast Cancer + Wikipedia
Map Score	50.6	83.9

Résultats avec et sans données générales

Sélection des données

- Peut-on sélectionner les données qui nous intéressent dans notre corpus général ?
  - Tf-Idf : par documents, on construit les vecteurs Tf-Idf de chaque documents
  - Cross Entropy : par phrases, à partir d'un modèle de langue
  - BERT : par phrases, en tant que classifier
  - (Random : par phrases)
- On ajoute au corpus spécialisé les données générales sélectionnées
  - Entraînement des embeddings a partir de ces données
  - Mapping des embeddings

	English		French	
Corpus	# tokens	# types	# tokens	# types
BC	525,934	14,800	521,262	11,746
WE	311,898	15,344	656,178	15,799
JRC	64.2M	229,836	70.3M	231,126
WIKI	300M	3M	300M	3.1M

- Dictionnaire ELRA
- Données d'évaluation :
  - Breast Cancer : 248 paires de mot
  - Wind Energy : 145 paires de mot
- MAP Score :

$$MAP(Ref) = \frac{1}{|Ref|} \sum_{i=1}^{|Ref|} \frac{1}{r_i}$$

Résumé

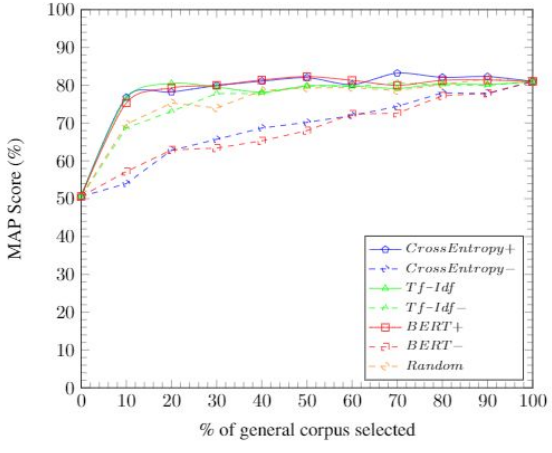
- On compense la petite taille des corpus spécialisés en allant les enrichir de données générales
    - Améliorer la représentation des mots du domaine spécialisé
    - Introduction de polysémie
    - Augmentation des temps de calcul

- Sélection des données générales
    - tfIdf, cross entropy, BERT, Random
- Réduction des temps de calcul
- Diminuer l'impact sur la polysémie

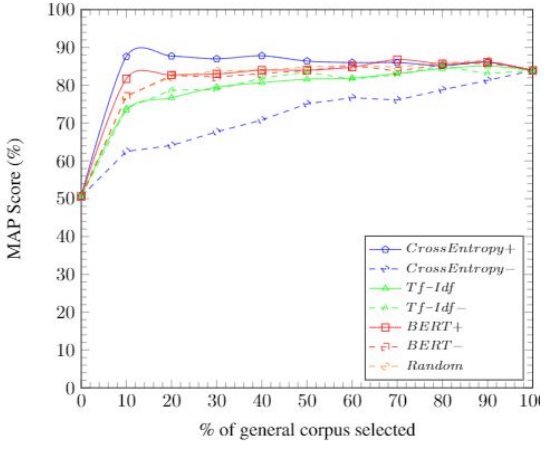
Résultats

- Courbe + : du meilleur vers le moins bon
- Courbe - : du moins bon vers le meilleur

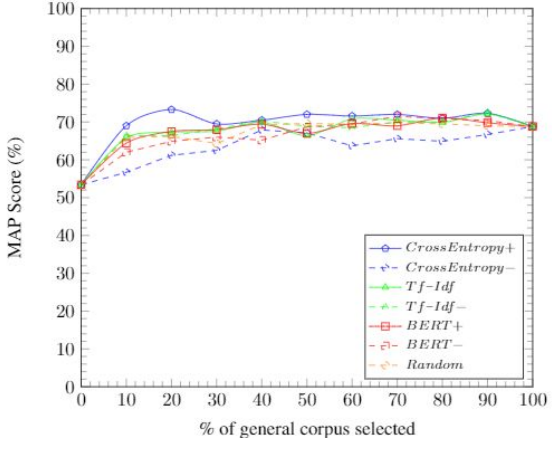
- Cross Entropy intéressante
- Wikipedia plus intéressant : diversité et taille du corpus

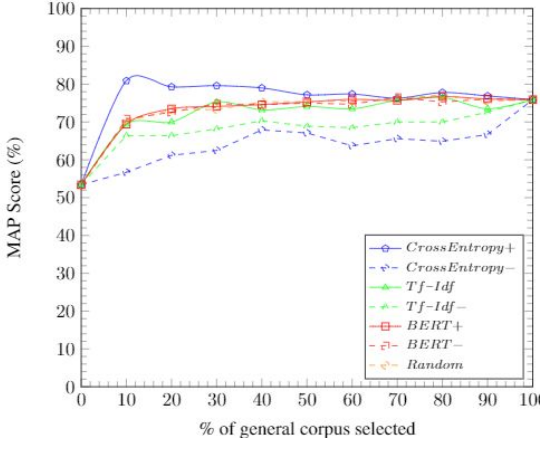


(a) (BC + JRC)



(b) (BC + WIKI)





Résultats

	Specialized Corpus
BC	50.6
WE	53.4

- Avec les corpus spécialisés seuls

Résultats

	Specialized Corpus	JRC
BC	50.6	59.8
WE	53.4	66.4

- Avec le corpus JRC seul
- Corpus de taille “moyenne”

Résultats

	Specialized Corpus	JRC	*Spec.* + *n*% JRC		
			100%	70%+	70%-
BC	50.6	59.8	81.0	83.2	74.4
WE	53.4	66.4	68.8	72.0	65.6

- Combinaison corpus spécialisé et général
- Pic en MAP avec 70%

Résultats

	Specialized Corpus	JRC	*Spec.* + *n*% JRC			WIKI
			100%	70%+	70%-	
BC	50.6	59.8	81.0	83.2	74.4	82.7
WE	53.4	66.4	68.8	72.0	65.6	69.7

- Avec le corpus Wiki seul

Résultats

	Specialized Corpus	JRC	*Spec.* + *n*% JRC			WIKI	*Spec.* + *n*% WIKI		
			100%	70%+	70%-		100%	10%+	10%-
BC	50.6	59.8	81.0	83.2	74.4	82.7	83.9	**87.6**	62.5
WE	53.4	66.4	68.8	72.0	65.6	69.7	75.9	**80.9**	55.5

- Combinaison corpus spécialisé et général
- Pic en MAP avec seulement 10%

Résultats

- Wiki : 3M de mots = 1%
- JRC : 3M de mots = 5%

- 1% suffisent à égaler 100%

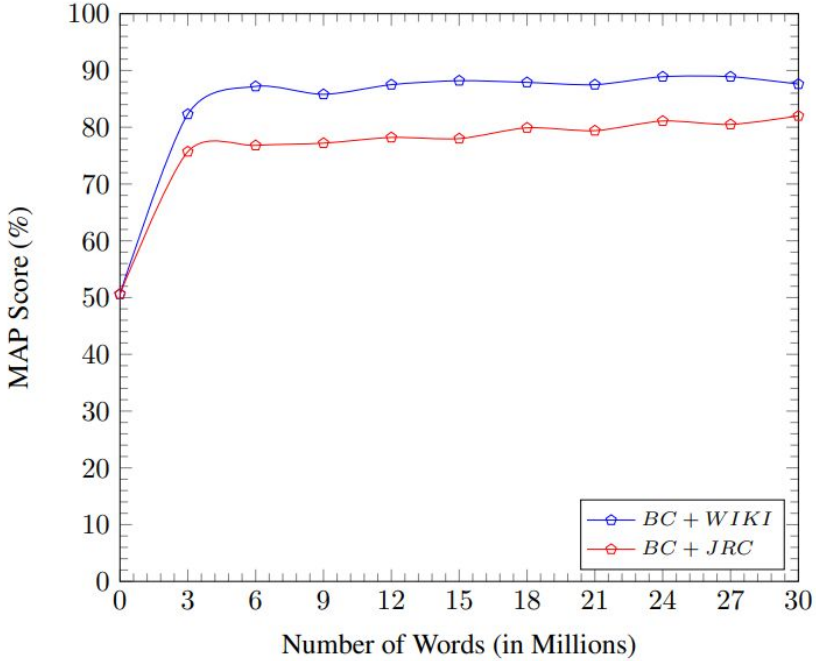


Figure 2: Results on $CrossEntropy$, on lower percentage of the general corpus.

Résultats

- Problème de polysémie en partie réglé

BC + n% WIKI	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
En Occ (Breast)	$3.6k$	$6.5k$	$7.3k$	$7.8k$	$8.2k$	$8.5k$	$8.7k$	$8.8k$	$8.9k$	$8.9k$	$9.0k$
Fr Occ (Sein)	$2.9k$	$7.3k$	$12.3k$	$17.8k$	$23.8k$	$30.2k$	$37.1k$	$44.3k$	$51.5k$	$57.0k$	$59.4k$
Rank	2	3	12	604	399	933	≥ 1000	≥ 1000	≥ 1000	≥ 1000	≥ 1000
En Occ (Calcium)	23	$1.8k$	$2.2k$	$2.3k$	$2.4k$	$2.5k$	$2.6k$	$2.6k$	$2.7k$	$2.7k$	$2.7k$
Fr Occ (Calcium)	14	983	$1.3k$	$1.5k$	$1.7k$	$1.8k$	$1.8k$	$1.9k$	$1.9k$	$2.1k$	$2.2k$
Rank	140	1	1	1	1	1	1	1	1	1	1
En Occ (Back)	27	$7.3k$	$19.5k$	$33.8k$	$49.7k$	$66.6k$	$84.0k$	$102.1k$	$120.2k$	$139.8k$	$153.9k$
Fr Occ (Dos)	7	883	$2.0k$	$3.3k$	$4.6k$	$6.0k$	$7.4k$	$8.8k$	$10.3k$	$11.9k$	$14.0k$
Rank	≥ 1000	37	12	4	5	9	14	50	43	27	99
En Occ (Lymphoscintigraphy)	20	20	20	20	20	20	20	20	20	20	20
Fr Occ (Lymphoscintigraphie)	27	28	28	28	28	28	28	28	28	28	28
Rank	4	1	1	1	1	1	1	1	1	1	1

Table 4: 4 interesting translation pairs over several data selection percentages from $CrossEntropy+$. The optimal selection is for 10% as seen in the previous section.

Conclusion

- La sélection de données présente des améliorations intéressantes
    - Problème de polysémie
    - Temps de calcul

Merci de m'avoir écouté,

Y'a-t-il des questions ?