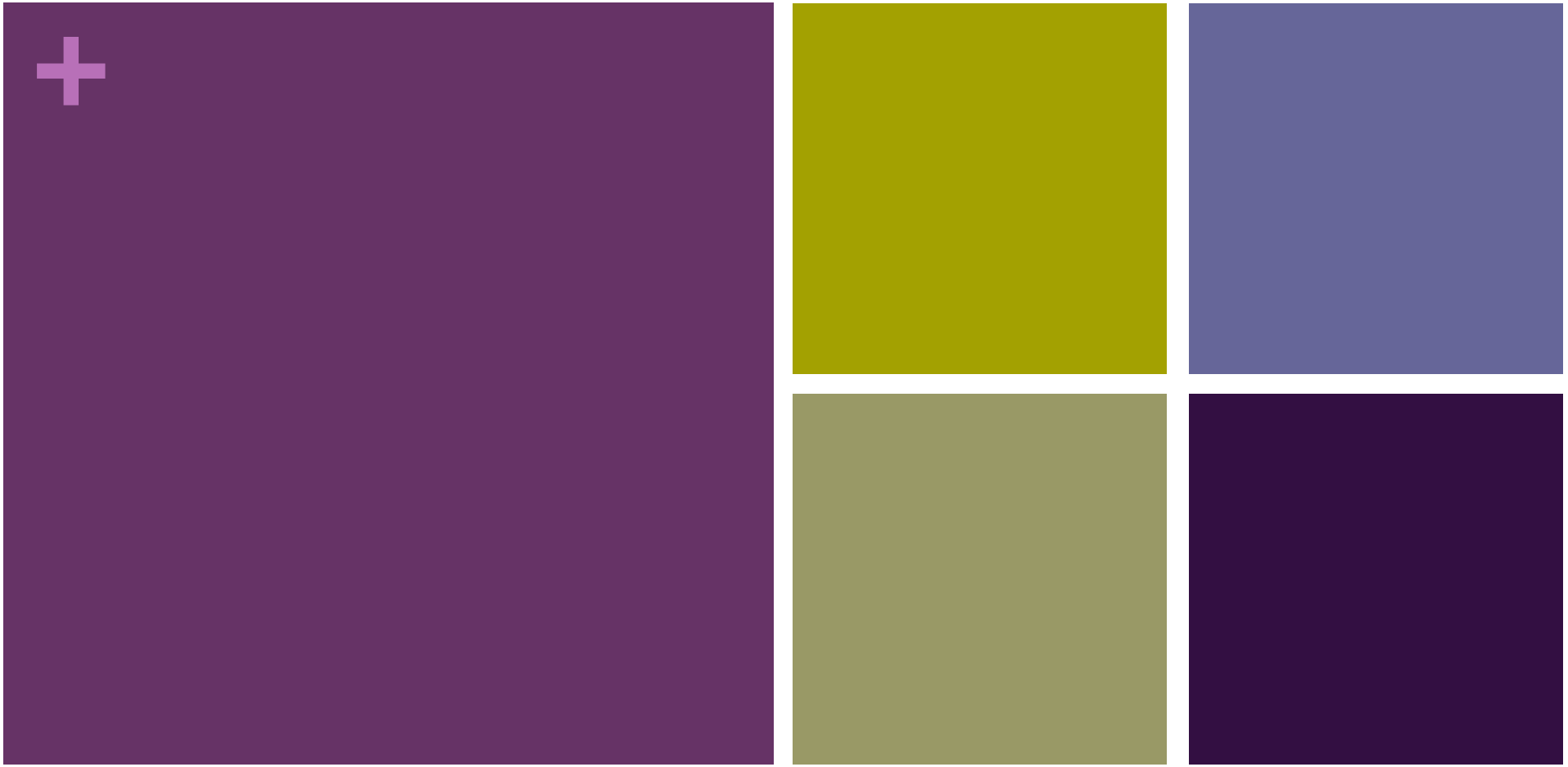


Extraction de lexiques bilingues à partir de corpus comparables spécialisés : la langue générale au secours de la langue de spécialité

Emmanuel Morin & Amir Hazem – LS2N
LIG – 22 mars 2018

Motivation

- Corpus comparables : « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres* » (Bowker et Pearson, 2002, p. 93) mais qui partagent des caractéristiques communes : domaine, période, média, auteur, genre, type de discours…

⇒ leur statut de production monolingue garantit la qualité du vocabulaire spécialisé

⇒ l'aspect multilingue du web garantit une disponibilité de ressources textuelles dans un grand nombre de langues et pour de nombreux domaines de spécialités

Motivation

- Corpus comparables exploités dans différentes tâches du TALN :
  - Traduction automatique (Munteanu and Marcu, 2005)
  - Recherche d'information interlangue (Nie, 2010)
  - Assistance à la traduction (Delpech, 2014)

Contexte

- Extraction de lexiques bilingues à partir de corpus comparables :
  - pour un terme à traduire dans une langue source proposer une liste de traductions candidates dans une langue cible
- En domaine de spécialité :
  - ressources textuelles réduites en comparaison à la langue générale
  - certaine proportion de vocabulaire spécifique absent des dictionnaires monolingues ou bilingues de langue générale
- Vocabulaire relevant d'un domaine de spécialité recense :
  - des termes simples : cancer
  - des termes complexes : cancer du sein

Données et ressources (1/2)

- Corpus comparables Français/Anglais
  - Cancer du sein (BC) : corpus spécialisé relevant du domaine médical et réduit à la thématique du cancer du sein composé d'articles pour lesquels le titre ou les mots clés *comportent breast cancer* en Anglais et *cancer du sein* en Français (1 M de mots)
  - JRC acquis (JRC) : corpus général qui rassemble des textes législatifs de la communauté européenne (132,7 M de mots)
  - Common crawl (CC) : corpus général composé de données collectées durant 7 ans (172,4 M de mots)
- Dictionnaire bilingue Français/Anglais
  - ELRA-M0033 (243,539 traductions)

Données et ressources (1/2)

- Liste de référence :
  - En domaine général, liste qui est une sous-partie du dictionnaire bilingue (Gaussier et al., 2004; Jakubina et Langlais, 2017)
  - En domaine de spécialité, liste généralement composée d'une centaine de mots qui reflètent la terminologie de corpus comparable (p. ex. 95 pour Chiao et Zweigenbaum (2002), 125 et 79 mots simples pour Bouamor et al. (2013))

  ⇒ 248 termes issus du méta-thesaurus de l'UMLS

- Evaluation des méthodes en termes de MAP (Mean Average Precision) (Manning and Schutze, 2008) :

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{Rank_i}$$

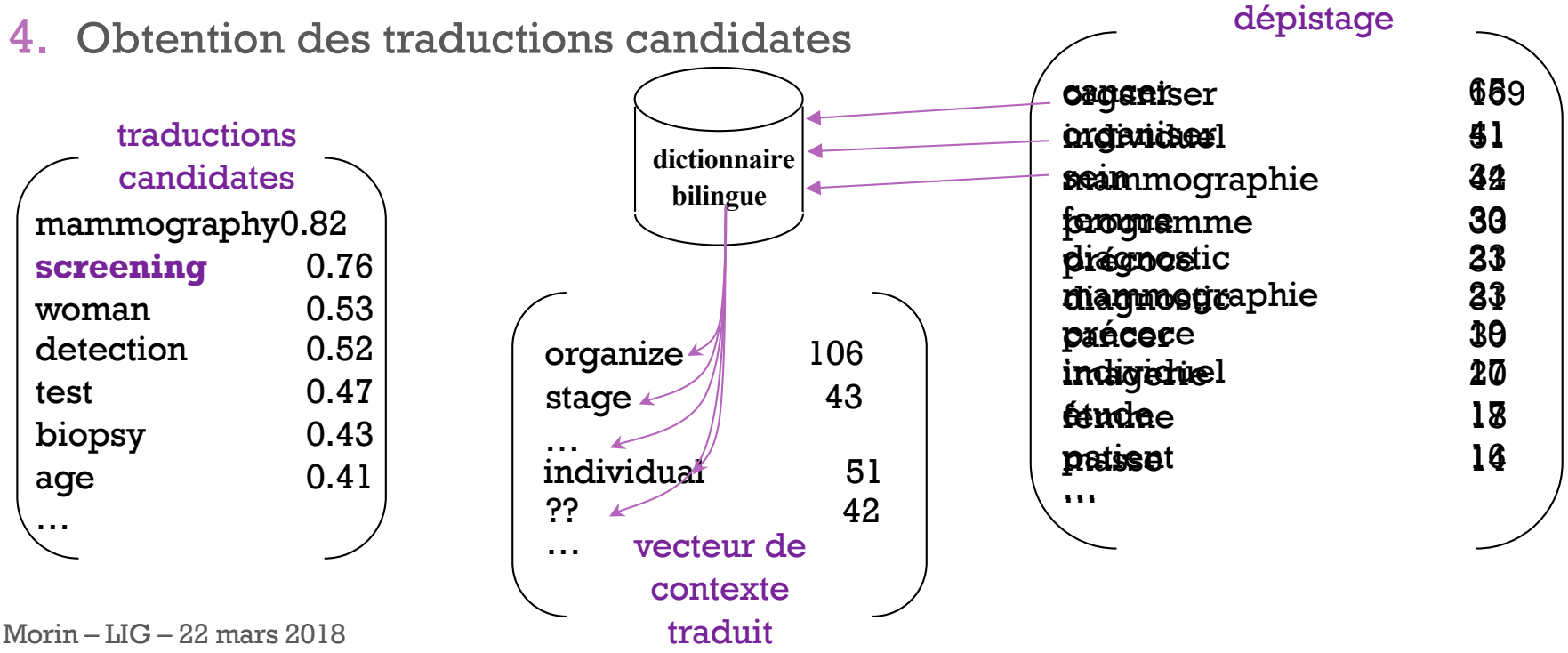
où $|Q|$ représente le nombre de termes à traduire et $Rank_i$ le rang de la bonne traduction parmi les traduction candidates

Méthodes d'alignement

- Méthodes d'extraction de lexiques bilingues à partir de corpus comparables :
  - basées sur une analyse du contexte lexical des mots et reposent sur la simple observation qu'un mot et sa traduction tendent à apparaître dans les mêmes contextes lexicaux
  - trouvent un ancrage dans la proposition de Firth (1957) : « *You shall know a word by the company it keeps* »

Standard Approach – SA (1/2)

- Approche standard (Fung & McKeown, 1997; Rapp, 1999) :
  1. Identification des contextes lexicaux
  2. Transfert d'une unité à traduire
  3. Identification des vecteurs proches de l'unité à traduire
  4. Obtention des traductions candidates



Standard Approach – SA (2/2)

	SA
MI/Cos	22.6
Odds/Cos	24.8
Log/Jac	27.9

- Approche qui est très sensible au choix de paramètres : taille de la fenêtre contextuelle, mesures d'association et de similarité… (cf. Laroche et Langlais (2010) pour une étude sur l'influence des paramètres)
- Les vecteurs sont creux, de grande dimension et explicites
- Les termes les plus fréquents sont les mieux traduits

Smoothing and Prediction Techniques (2/2) – BUCC et IJCNLP 2013

- Est-il possible de ré-estimer les cooccurrences des mots du corpus de spécialité en les observant dans un corpus de plus grande taille ?
- Les techniques de lissage et prédiction (Good, 1953), utilisées pour mieux estimer les probabilités lorsque les données sont insuffisantes, sont bien adaptés à cette ambition (en outre ils tendent à rendre les distributions plus uniformes)
- Nous utilisons ses techniques comme une étape de prétraitement de l'approche standard (sans prendre en compte les cooccurrences non observées) pour lisser les cooccurrences

Smoothing and Prediction Techniques (2/2) – BUCC et IJCNLP 2013

- Techniques de lissage

	SA	Add1	GT	JM	Katz	Kney
MI/Cos	22.6	24.8	25.6	**29.5**	25.9	9.1
Odds/Cos	24.8	24.4	25.2	23.3	**25.3**	14.1
Log/Jac	27.9	**30.6**	21.4	21.2	21.2	22.9

- Techniques de prédiction

	SA	Max	Mean	LReg	MAI
MI/Cos	22.6	**27.2**	20.3	26.7	26.4
Odds/Cos	24.8	22.9	19.8	**27.6**	20.9
Log/Jac	27.9	11.6	24.6	22.6	15.6

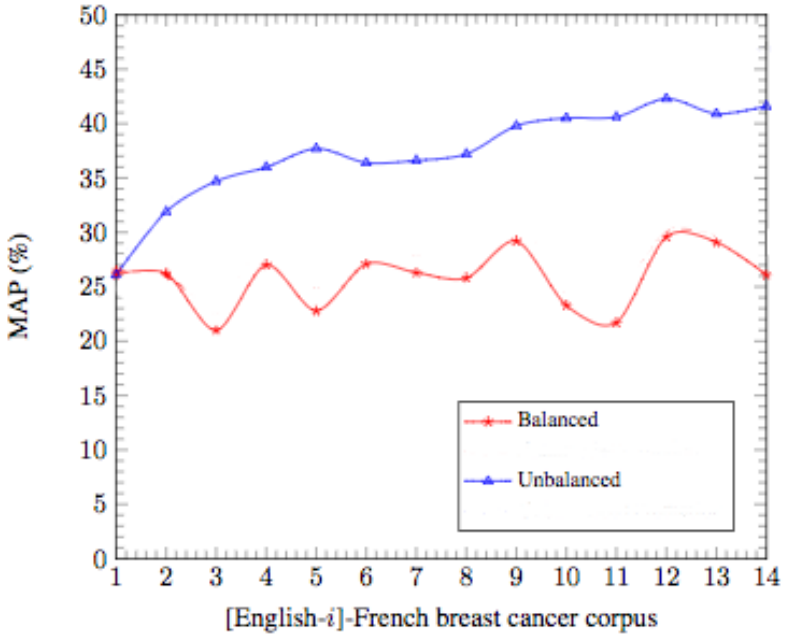
Unbalanced Specialized Comparable Corpora (1/2) – ACL 2014

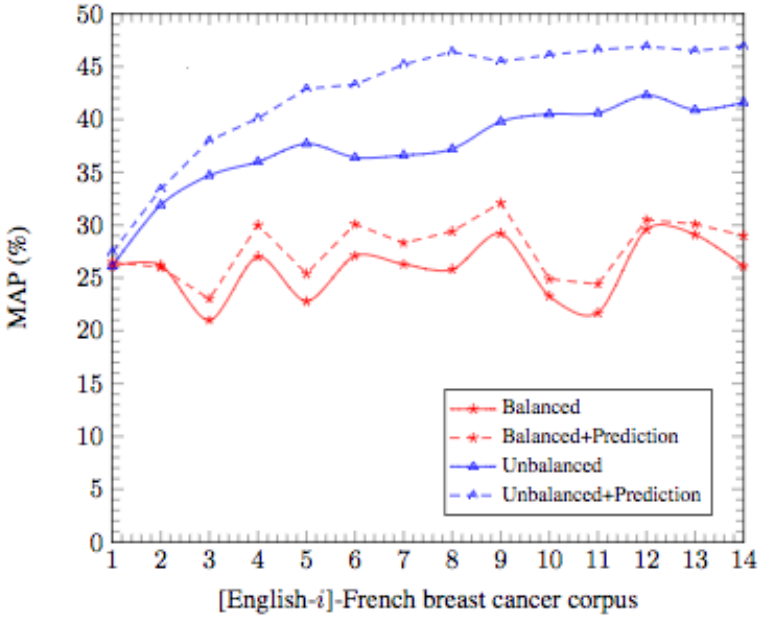
- Est-ce qu'un corpus comparable doit être équilibré (avoir la même taille en langue source comme cible) ?

References	Domain	Languages	Source/Target Sizes
Tanaka and Iwasaki (1996)	Newspaper	EN/JP	30/33 million words
Fung and McKeown (1997)	Newspaper	EN/JP	49/60 million bytes of data
Rapp (1999)	Newspaper	GE/EN	135/163 million words
Chiao and Zweigenbaum (2002)	Medical	FR/EN	602,484/608,320 words
Déjean et al. (2002)	Medical	GE/EN	100,000/100,000 words
Morin et al. (2007)	Medical	FR/JP	693,666/807,287 words
Otero (2007)	European Parliament	SP/EN	14/17 million words
Ismail and Manandhar (2010)	European Parliament	EN/SP	500,000/500,000 sentences
Bouamor et al. (2013)	Financial	FR/EN	402,486/756,840 words
-	Medical	FR/EN	396,524/524,805 words

Unbalanced Specialized Comparable Corpora (2/2) – ACL 2014

- Expérience simple : augmenter la taille de la partie anglaise du corpus comparable





Autres méthodes

- Quelques améliorations de l'approche standard :
  - Chiao (2004) introduit une hypothèse de symétrie distributionnelle
  - Prochasson et Morin (2009) introduisent la notion de point d'ancrage
  - Bouamor et al. (2013) utilisent Wikipédia pour la désambiguïsation des éléments à traduire
- Autres méthodes :
  - méthode par similarité interlangue et ses variantes (Déjean et Gaussier 2002, Daille et Morin, 2005; Hazem et al. 2009)
  - Méthode numérique CCA, ICA, PLSA… (Gaussier et al, 2004; Hazem et Morin 2012)

⇒ Ces méthodes semblent avoir atteint leur limite en termes de performance

Word Embedding Approach (1/2)
– IJCNLP 2017

- Comment rendre les vecteurs denses et moins grands ?
- Mikolov et al. (2013) : méthode pour apprendre une transformation linéaire de la langue source vers cible :
  - Apprentissage : pour toutes les paires de mots dictionnaire bilingue, les embedding de mots dans la langue source et les embedding de mots traduction dans la langue cible sont calculés. Une matrice de transformation W telle que Wx_i se rapproche de z_i est alors apprise par une fonction objectif :

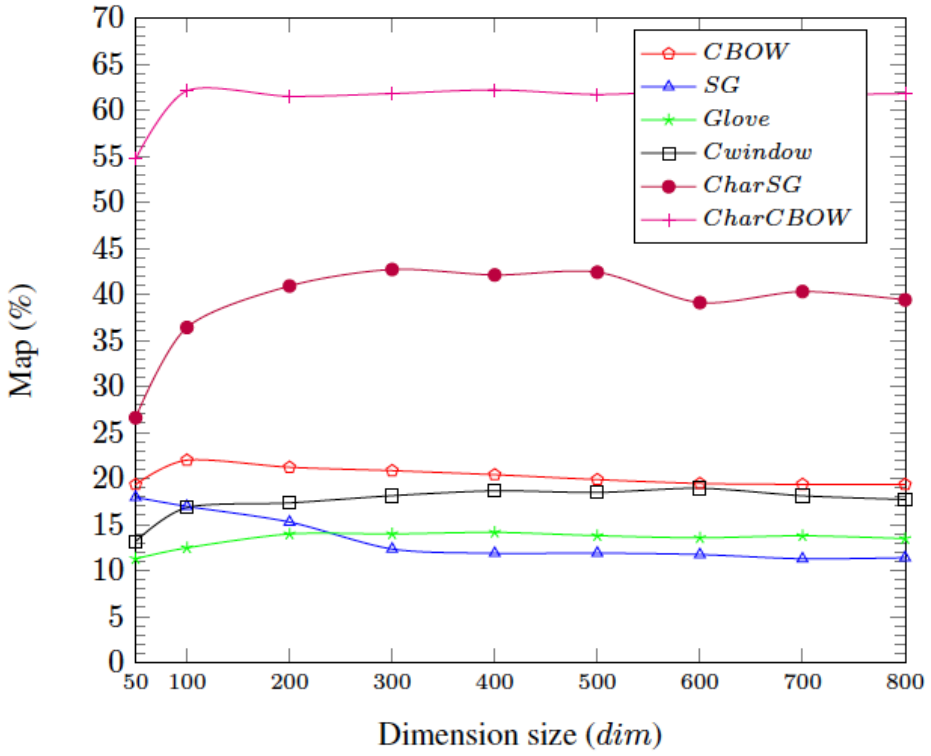
$\{x_i, z_i\}_{i=1}^{n}$
$x_i \in \mathbb{R}^{d_1}$ not i
$z_i \in \mathbb{R}^{d_2}$

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

  - Prédiction : le word embedding x d'un mot à traduire est transféré en utilisant la matrice de traduction telle *que* $z = Wx$. Les traductions candidates sont obtenues en classant les mots cibles les plus proches à z selon une mesure de similarité.

Word Embedding Approach (2/2) – IJCNLP 2017

- Modèles étudiés :
  - CBOW and Skip-gram (Mikolov et al., 2013)
  - Glove (Pennington et al., 2014)
  - Structured Embeddings (Ling et al., 2015)
  - Character n-gram Embeddings (Bojanowski et al., 2016)
- Autres modèles étudiés :
  - dependency-based model (Levy and Goldberg, 2014)
  - generalized-based model (Li et al., 2017)



Global and Selective Standard Approach (1/2) – COLING 2016

- Comment combiner autrement un corpus comparable spécialisé avec un corpus de langue générale ?
- Global Standard Approach (GSA)
  - Les vecteurs de contexte sont construits à partir d'un corpus réunissant les deux corpus (spécialisé et général)
- Selective Standard Approach (SSA)
  - Les vecteurs de contexte sont construits indépendamment pour les deux corpus (spécialisé et général)

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

S (resp. *G*) est le vocabulaire du corpus spécialisé (resp. général), *w* un mot à caractériser pour représenter et *c* un mot qui apparaît dans le contexte de *w*

Global and Selective Standard Approach (2/2) – COLING 2016

	BC	JRC	CC
SA	27.0	52.0	75.5
GSA		61.7	80.2
SSA		**66.6**	**82.3**

- Pour mémoire
  - en domaine de spécialité il est exclu d'avoir recours à un corpus de langue générale, c'est même une idée assez à contre courant…

Combination Using Neural Network Models (1/2) – IJCNLP 2017

- Global Data Combination (GDC)
  - Les modèles Skip-gram et CBOW sont appris à partir d'un corpus réunissant les deux corpus (spécialisé et général)
- Specific Data Combination (SDC)
  - Les modèles Skip-gram et CBOW sont appris indépendamment pour les deux corpus (spécialisé et général)
  - Les vecteurs des modèles sont concaténés en langue source et cible (à la manière de Garten et al. 2015)

Combination Using Neural Network Models (1/2) – IJCNLP 2017

	BC	JRC	CC
SA	27.0	52.0	75.5
CBow	17.1	40.3	60.9
CharCBow	60.8	35.3	57.4
GCD (Cbow)		49.9	67.7
GCD (Char Cbow)		52.9	73.9
SDC (Cbow)		53.7	70.7
SDC (CharCBow)		**64.9**	**74.9**

Conclusion

- Panorama (non exhaustif) de nos recherches en extraction de lexiques bilingues à partir de corpus comparables spécialisés
- Les résultats avec l'approche standard restent compréhensifs à la différence avec l'approche par embeddings qui restent surprenants
- Les méthodes par combinaison sélective améliorent grandement les résultats
- La direction la plus prometteuse n'est pas tant la combinaison de modèle avec des corpus de plus en plus volumineux que de choisir les données les plus appropriées
- Quid des termes complexes…