REVUE DE LITTÉRATURE - CORPUS COMPARABLES

Julie Roy

April 23, 2015

Université de Montréal

Introduction

- Matière première MT : corpus parallèles.
- Intérêt des corpus comparables : corpus parallèles et dictionnaires bilingues

Plan

1. Quelques définitions
2. Défis
3. Travaux et approches tentées
4. Résumé et quelques réflexions...
5. Conclusion

QUELQUES DÉFINITIONS

Corpus parallèle vs comparable

Selon P.Fung et P.Cheung (2004)[1]

,

Parallèle :
-Phrases sont alignées
-Contient traduction bilingue d'un même doc.

Comparable :
-Phrases pas nécessairement alignées et/ou traduites
-Alignement fait sur domaine, sujet, période, etc.

Corpus "noisy" parallèle vs quasi-comparable

Selon P.Fung et P.Cheung (2004)[1]

Parallèle "bruité" :
-Non-alignés, mais traductions moins fidèles.
-Même sujet, mais doc contiennent insertion/supression

Quasi-comparable :
-Non-alignés, pas de doc. source traduit.
-Composé de sections "in-topic" et "off-topic"

PRINCIPAUX PROBLÈMES

PROBLÈMES

BESOIN : -Méthode permet retrouver information isolée

MAIS

- environnement bruité.
- contenu similaire, mots très différents
- corpus intéressants, grande taille => besoin algo efficaces
- Données test ”in-domain”, données d’entraînement ”out-of-domain” => mauvaise performance de traduction
  Alignement des phrases est donc tâche difficile et coûteuse

TRAVAUX ET APPROCHES TENTÉES

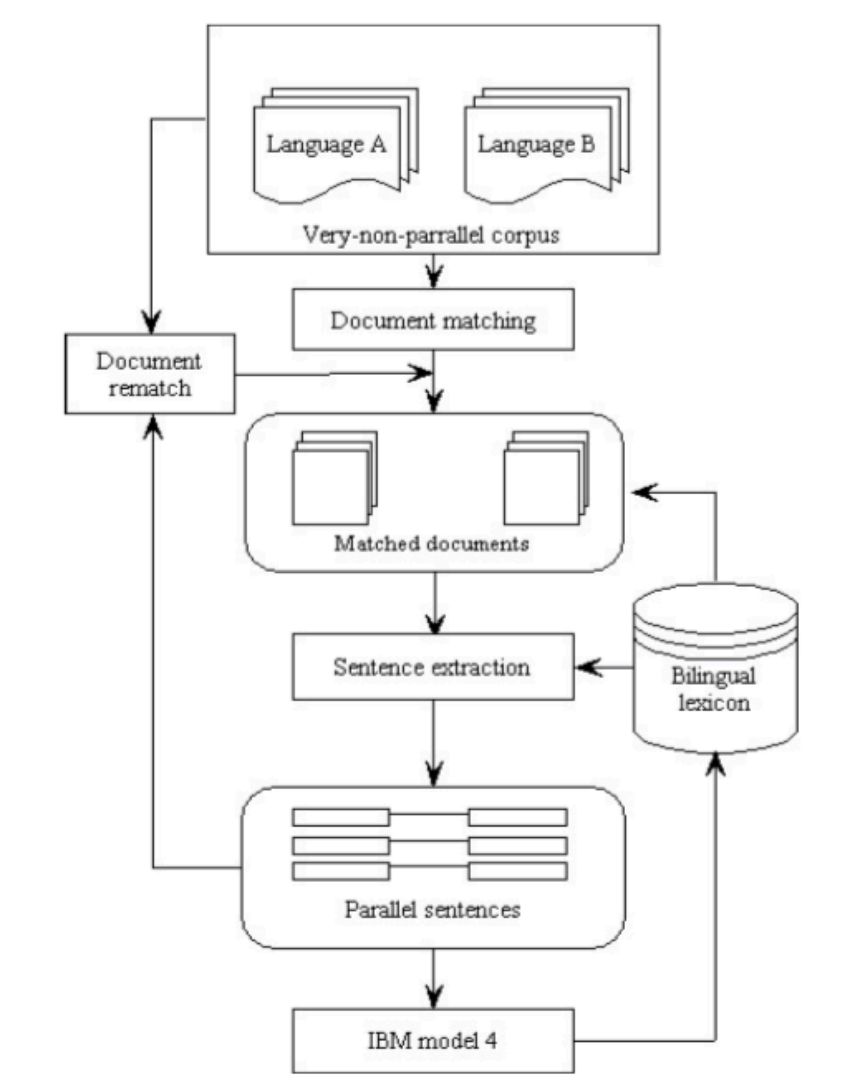
Extractions phrases parallèles

- P.Fung et Cheung (2004)[1]***
- Smith et al. (2010)[2]
- Expérience de Munteanu et Marcu 2005[3]

P.Fung et P.Cheung (2004)

- Utilisation corpus quasi-comparables (ensemble de nouvelles en-zh)
- Désambiguïsation de la terminologie
- Quantification du taux de parallélisme à l'aide de lexique
- Utilisation du "bootstrapping" + Modèle IBM : voir fig. 1
- Évaluation : compare à méthode "find-topic-extract-sentence" + manuelle

Approche Fung



Smith et al. (2010)

- Exploitation de structures des documents (wiki)
- Hypothèse : phrases parallèles observées dans certaine fenêtre
- Données : 20 paires d'articles dans 3 paires langues
- Évaluation intrinsèque et évaluation traduction.
  1. Comparaison résultats classifieur, rankers, modèle CRF (pour phrases parallèle)
  2. Precision moyenne
  3. Utilisation dans système de traduction (BLEU)
  4. Amélioration de l'État de l'art

Travaux de Munteanu et Marcu

- Utilisation d'arbre de suffixes bilingues (bilingual suffix trees (2002)[4].

Travaux de Munteanu et Marcu

- Utilisation d'arbre de suffixes bilingues (bilingual suffix trees (2002)[4].
- Entraînement d'un classifieur à maximum d'entropie(2004)[5]

Travaux de Munteanu et Marcu

- Utilisation d'arbre de suffixes bilingues (bilingual suffix trees (2002)[4].
- Entraînement d'un classifieur à maximum d'entropie(2004)[5]
- Entraînement d'un classifieur à maximum d'entropie (2005)[3]

Expérience de Munteanu et Marcu 2005[3]

- Plus cité - État de l'art
- Utilisation ME classifieur, non-basé sur contexte.
- Petit corpus parallèle et grand corpus non-parallèle (Bon pour paires de langues rares).
- Problème "in-domain" données test vs "out-of-domain" données d'entraînement
- Évaluation : phrases produites alimentent système de traduction. Amélioration de 50% sur le baseline
- Système robuste

lexiques parallèles de Munteanu

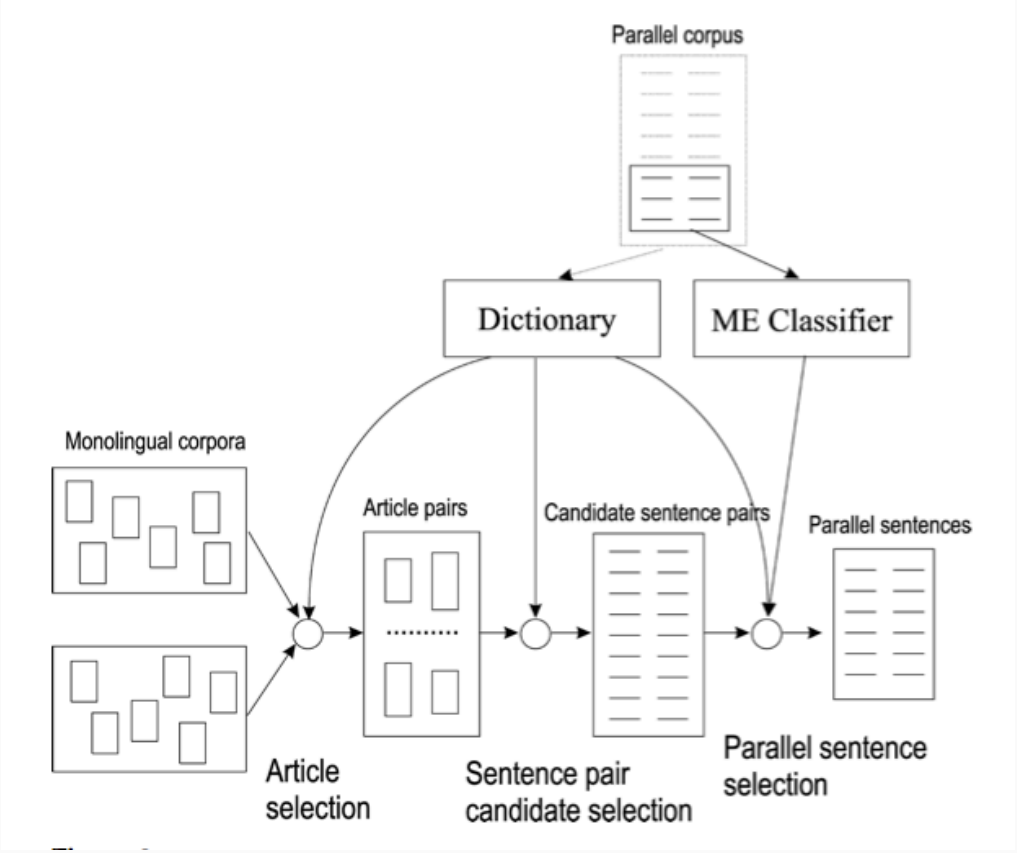


Figure:

Extraction de segments parallèles I

- Tanaka(2002) - Utilisation de mots composés + contexte afin de localiser traductions[6]
- Smith(2002) - Déterminer si 2 textes sont·translationally equivalentöu pas.[7]
- Munteanu et Marcu (2006)- Utilisation de technique du traitement de signal pour l'extraction de fragments parallèles [8]
- Quirk et al.(2007) - 2 algo basés sur modèle génératif - très basé sur Munteanu 2006 [9]
- Kumano et Tokunaga (2007) - Modèle "phrase-based" probabilités jointes sans dictionnaire [10]

Munteanu(2006) - Quirk(2007)

- Munteanu : 1ere tentative d'extraction fragments
  1. Paires d'articles à l'aide de méthode RI
  2. Sélection de paire dont recoupement de mots élévé
  3. Annotation de mots avec valeur entre 1 et -1 (score LLR - indication vraisemblance)
  4. Utilise approche inspirée du traitement de signal
  5. Un "span" de valeur positive plus grand que 3 => traduction
  6. Fragments ajoutés dans système de traduction : augmentation du score BLEU

Munteanu(2006) - Quirk(2007)

- Quirk :
  1. Désire améliorer méthode Munteanu avec modèle + efficace, bases théoriques fondées.
  2. Présente deux modèles d'alignement : génération conditionnelle et jointe.
  3. Évaluation en alimentant système de traduction.
  4. Idée intéressante, mais beaucoup de place l'amélioration.

Technique de RI - Rauf et Schwenk(2009)

- Technique semblable à Munteanu(2005)
  - sauf pas de dictionnaire bilingue
  - ME remplacé par mesure du ”Word error rate”(WER) et ”translation error rate”(TER)
- Système de traduction construit à partir d'un petit corpus parallèle traduit doc. source
- Utilise technique de RI et filtres simple pour créer données parallèles
- Évaluation : utilisation données dans système de traduction.

RÉSUMÉ ET QUELQUES RÉFLEXIONS

Quelques réflexions...

1. Utilisations de dictionnaires bilingues et corpus parallèles
2. Généralisation méthode difficile
3. Variabilités des données, des évaluations.
4. Utilisation de thesaurus?

CONCLUSION

Conclusion

Points importants :

1. Utilisations des corpus comparables est pertinente
2. Méthodes montrent avancées
3. Travaux de Munteanu, Fung et Smith plus cités.
4. Problèmes d'uniformité dans expérimentation

Questions?

BIBLIOGRAPHIE

Référence I

- P. Fung and P. Cheung, “2004b. multilevel bootstrapping for extracting parallel sentences from a quasi-comparable corpus,” in *In COLING 2004*, pp. 1051–1057, 2004.
- J. R. Smith, C. Quirk, and K. Toutanova, “Extracting parallel sentences from comparable corpora using document level alignment,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403–411, Association for Computational Linguistics, 2010.
- D. S. Munteanu and D. Marcu, “Improving machine translation performance by exploiting non-parallel corpora,” *Computational Linguistics*, vol. 31, no. 4, 2005.

Référence II

D. S. Munteanu and D. Marcu, "Processing comparable corpora with bilingual suffix trees," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Philadelphia), pp. 289–295, Association for Computational Linguistics, July 2002.

D. S. Munteanu, A. Fraser, and D. Marcu, "Improved machine translation performance via parallel sentence extraction from comparable corpora," in *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2004.

Référence III

- T. Tanaka, “Measuring the similarity between compound nouns in different languages using non-parallel corpora,” in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2002.
- N. A. Smith, “From words to corpora: Recognizing translation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Philadelphia), pp. 95–102, Association for Computational Linguistics, July 2002.

Référence IV

- D. S. Munteanu and D. Marcu, “Extracting parallel sub-sentential fragments from non-parallel corpora,” in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 81–88, Association for Computational Linguistics, 2006.
- C. Quirk, R. Udupa, and A. Menezes, “Generative models of noisy translations with applications to parallel fragment extraction,”
- T. Kumano, H. Tanaka, and T. Tokunaga, “Extracting phrasal alignments from comparable corpora by using joint probability smt model,” *Proceedings of TMI*, 2007.