Fiche de lecture : Unsupervised Morphology-Based Vocabulary Expansion

Informations :

Titre : Unsupervised Morphology-Based Vocabulary Expansion

Auteurs : Mohammad Sadegh Rasooli, Thomas Lippincott, Nizar Habash & Owen Rambow

Conférence ou revue : The 52nd Annual Meeting of the Association for Computational Linguistics (2014)

Année : 2014

Objectif de l'article :

Créer de nouveaux mots pour étendre le vocabulaire de langues dont les ressources sont insuffisantes.

Le but de beaucoup de technologies du langage humain est de générer un texte dans une langue cible, en utilisant son orthographe standard. Aujourd'hui les systèmes les plus performants pour ces applications se basent sur une grande quantité de données. Plus il y a de données, meilleurs sont les résultats. Pour des langues avec beaucoup de ressources, plus de données est souvent la meilleure solution puisque ces données sont disponibles. Ainsi dans ce domaine, les améliorations de résultats sont souvent liées à l'utilisation de plus larges ensembles de données. Concernant les langues avec peu de ressources disponibles, cette solution n'est pas envisageable. Si l'on souhaite développer des technologie de langage humain pour plus de langues (incluant celle disposant de moins de ressources), une alternative à la solution « plus de données » devient importante. Pour les langues disposant de peu de ressources, on manque également d'outils, et de descriptions de la morphologie. Le défi est alors d'utiliser au mieux les données. C'est le but de cet article : proposer une nouvelle approche pour générer des mots de la langue cible qui n'ont pas été vus parmi les données d'entraînement. Cette approche est entièrement non-supervisée.

Définitions :

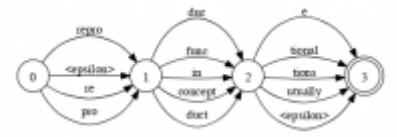
Ségmentation morphologique : principe de segmenter chaque mot d'un corpus en séquences de préfixes, stem et suffixes, où les affixes sont indépendamment optionnels.

Modèle d'extension affixe-fixé : Technique d'extension morphologique où l'on construit un ensemble unique de préfixes (resp. suffixes) à partir de chaque préfixe (resp. suffixe) unique des données d'entraînement. Selon ce modèle, chaque mot n'a qu'un seul préfixe et un seul sufixe (chacun des deux pouvant être indépendamment vide).

Segmentation morphologiques de quelques mots

re+ pro+ duc +e
func +tion +al
re+ duc +e
re+ duc +tion +s
in
pro+ duct
concept +u +al + ly

Modèle d'extension affixe-fixé correspondant



Modèle d'extension affixe-bigram : Technique d'extension morphologique où l'on fait la même chose que dans le modèle affixe-fixé, mais pour les préfixes et suffixes, on crée un modèle de langue bigram. L'avantage de cette technique est que des combinaisons inconnues d'affixes peuvent être générées. On attend de ce modèle d'avoir un meilleur rappel pour générer de nouveaux mots dans la langue grâce la flexibilité de son affixation.

Segmentation morphologiques de quelques mots

re+ pro+ duc +e
func +tion +al
re+ duc +e
re+ duc +tion +s
in
pro+ duct
concept +u +al + ly

Modèle d'extension affixe-bigram correspondant

Difficulté de la tâche :

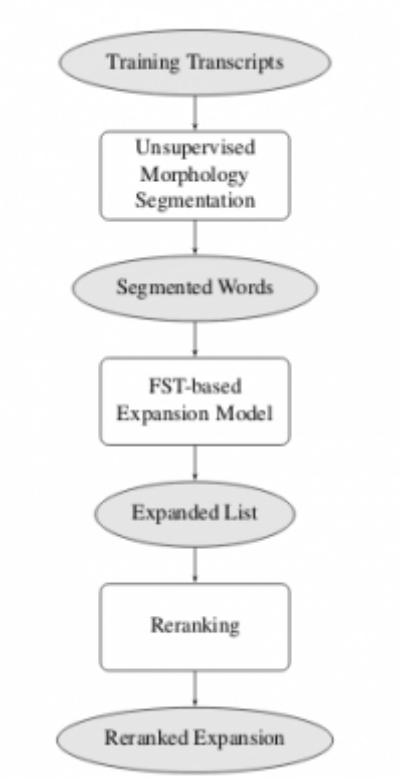
Je n'ai pas pu cerner de difficulté notoire mentionnée dans l'article.

Apport de l'article :

Étendre le vocabulaire de la langue cible peut être utile de différentes manières. Pour les applications de reconnaissance automatique de la parole ou de reconnaissance optique de caractères, un vocabulaire étendu de la langue cible peut être exploité sans avoir besoin de changer de technologie : les nouveaux mots ont seulement besoin d'être ajoutés dans les ressources concernées avec des probabilités appropriées estimées. Dans le cas de la traduction automatique vers une langue morphologiquement riche mais avec peu de ressources, la segmentation morphologique est utilisée pour réduire la dispersion dans les modèles, mais ne garantie pas qu'on ne génère pas de mauvaises combinaisons de mots. Les combinaisons de mot étendues peuvent être utilisé pour étendre les modèles de la langue utilisés par la traduction automatique pour biaiser contre des nouvelles séquences de mots segmentés hypothétiques incohérentes. Cet article propose d'utiliser les résultats de la segmentation morphologique dans le but de générer des mots inconnus de la langue.

Méthode :

1. Segmentation morphologique non supervisée. 2. Extension morphologique. 3. Reclassement des nouveaux mots.



Résultats :

De 65K à 115K tokens de différentes langues ont été morphologiquement segmentés. Un petit ensemble de donné (de 50K à 100K tokens) a été évalué en mesurant la réduction hors-vocabulaire. Les meilleurs résultats ont été obtenus avec le modèle affixe-fixé avec un reclassement trigram. La précision des mots est toujours un gros problème (moins de 30% des top 50K types générés peuvent être analysés par un analyseur morphologique Turc).

Conclusion sur l'article

Cet article a présenté un approche pour générer de nouveaux mots. Cette approche est utile pour des langues morphologiquement riches mais disposant de peu de ressources. Elle fournit des mots qui peuvent être utilisés dans les technologies du langage humain qui nécessitent de la génération dans cette langue, comme la reconnaissance automatique de la parole, le traitement optique de caractères ou la traduction automatique.