Fiche de lecture : Étiquetage multilingue en parties du discours avec MElt

Informations :

Titre : Étiquetage multilingue en parties du discours avec MElt

Auteur : Benoît Sagot

Conférence ou revue : 23ème Conférence sur le Traitement Automatique des Langues Naturelles

Année : 2016

Objectif de l'article

L'objectif de l'article est de montrer l'efficacité de l'étiqueteur morphosyntaxique MElt lorsqu'on lui fournit des ressources lexicales externes.

Difficulté de la tâche

La difficulté a été de définir une évaluation pertinente de l'étiqueteur morphosyntaxique MElt comparativement aux outils états-de-l'art.

Apport de l'article

Sans ressources lexicales externes, le modèle statistique utilisé dans les étiqueteurs morphosyntaxiques état-de-l'art est plus performant que celui sur lequel s’appuie MElt. Cependant, MElt permet d'intégrer des informations lexicales externes de manière performante, ce qui lui permet alors (souvent) de passer devant les étiqueteurs état-de-l'art.

Méthode

MElt est un système d'étiquetage de séquences qui repose sur les modèles markoviens à maximisation d'entropie. Un tel modèle peut tirer parti non seulement des informations extraites du corpus d'entraînement mais également d'informations issues d'un lexique extérieur varié. Les performances obtenues sont meilleures si l'intégration de ces dernières se fait sous forme de traits supplémentaires, par opposition à l'utilisation du lexique externe comme source de contraintes au moment de l'étiquetage (par exemple, en n'autorisant pas l'étiquetage d'un mot connu du lexique par une étiquette que le lexique ne lui associe pas).

MElt a été développé selon trois axes principaux. Tout d'abord, de nouveaux traits (préfixes et suffixes du mot à droite du mot courant) ont été ajoutés, et certaines propriétés de l'espace des traits ont été altérées. Notamment la longueur maximale des préfixes et des suffixes pris en compte par le modèle (pour le mot courant et celui à sa droite).

Par ailleurs, le moteur d'étiquetage a été encapsulé dans des outils permettant l'identification de certains types d'entités nommées (dates, adresses…) qui sont alors considérés comme des mots uniques par le modèle d'étiquetage et dont l'étiquetage interne est effectué grâce à des règles dédiées.

Enfin, le moteur d'étiquetage a été encapsulé dans des outils permettant le traitement de corpus bruités tels que trouvés sur le web. Mais cette option n'a pas été activée pour les travaux décrits dans cet article.

Résultats

Entraîné uniquement sur les corpus d'entraînement, sans lexique externe, MElt est moins bon que l'étiqueteur morphosyntaxique état-de-l'art utilisé en comparaison, mais avec un lexique externe, il devient meilleur.

Conclusion sur l'article

MElt est un étiqueteur morphosyntaxique à privilégier lorsqu'on travaille avec un lexique externe.