Lucía Ormaechea, Pierrette Bouillon, Benjamin Lecouteux et Didier Schwab

Vers une simplification automatique de la parole en français

Les enjeux de l'extraction des données d'apprentissage pour la simplification linguistique

Mots-clés : traitement automatique des langues naturelles ; tâches sous-dotées ; simplification automatique de textes ; exploitation de corpus comparables ; adaptation à une modalité orale

Introduction

La Simplification Automatique de Textes (SAT) est un domaine du TAL qui vise à réduire automatiquement la complexité linguistique des textes, sans pour autant perdre leur signification originale. Bien qu'il s'agisse d'une tâche importante d'un point de vue sociétal et computationnel, automatiser la simplification linguistique est souvent contrainte par la rareté de corpus parallèles associant des phrases *complexes et simples*. Ceci est encore plus prégnant dans le cas du français, où les ressources existantes sont insuffisantes pour l'entraînement de modèles basés sur l'apprentissage automatique (Brouwers et al., 2012 ; Cardon & Grabar, 2019). De plus, la majorité des travaux précédents se sont penchés sur la simplification linguistique de sources écrites et peu d'études ont examiné des méthodes servant à simplifier la parole (Buet & Yvon, 2021).

Notre travail cherche à pallier ces deux lacunes de manière séquentielle. Tout d'abord, nous proposons une méthode d'exploitation de corpus permettant d'extraire automatiquement des paires de phrases pertinentes pour la SAT. Cela facilite ensuite l'entraînement de modèles de simplification phrastique et permet d'étendre par la suite la tâche de la simplification automatique à une modalité orale.

Méthodologie

Dans cette présentation, nous examinons la première de ces deux lacunes, et détaillons une méthode de filtrage spécifiquement conçue pour l'identification des paires de phrases *complexes-simples* (affichée sur la Fig. 1). Pour cela, nous nous appuyons sur des corpus comparables différenciés en registre de langue, qui associent des textes standards à leurs versions simplifiées, comme Wikipédia et Vikidia, où la dernière vise à rendre les articles de Wikipédia plus facilement compréhensibles pour un public jeune. Sur cette base, nous examinons séquentiellement les deux conditions que doit remplir une phrase simplifiée pour être considérée comme étant valable au regard de sa référence :

1. *Préservation du sens original*. Nous vérifions si la phrase Vikidia est sémantiquement équivalente à celle de Wikipédia, avec l'usage des similarités cosinus basées sur SBERT et d'une annotation manuelle.
2. *Gain en simplicité linguistique*. Si la condition préalable est observée, nous vérifions que la phrase Vikidia est effectivement une version simplifiée de la phrase d'origine. Pour cela, nous avons conçu des modèles pour qualifier et quantifier la simplicité des phrases en entrée, de manière absolue et relative (*càd*, en contexte), notamment en effectuant un affinage des modèles de langage BERT pour les tâches de classification et de régression.

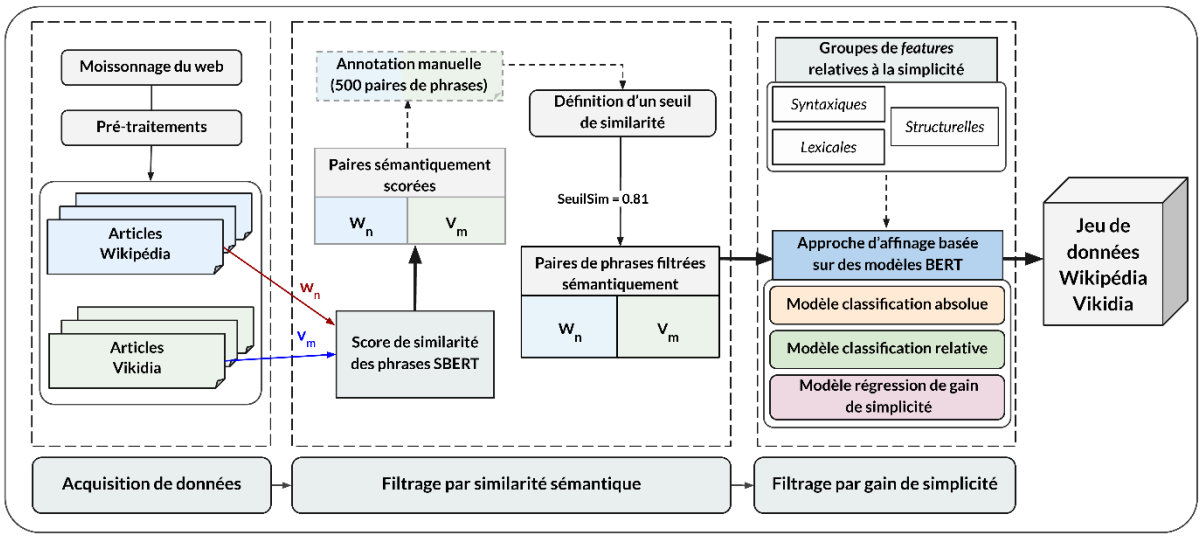

Fig. 1 : Vue d'ensemble de la chaîne d'obtention de données et le filtrage en deux étapes pour obtenir des paires de phrases *complexes-simples* à partir des éditions françaises de Wikipédia et Vikidia

Résultats

En utilisant cette approche de filtrage en deux étapes, nous avons créé un jeu de données monolingues parallèles, basée sur des ressources textuelles existantes. Cet ensemble de phrases peut servir à entraîner des modèles de simplification phrastique, ou pour affiner un grand modèle de langage pré-entraîné pour notre tâche en aval.

Discussion

Dans cette étude, nous avons proposé une méthode permettant d'exploiter des corpus comparables de manière à faciliter la constitution des données alignées pour l'entraînement ultérieur de modèles SAT. Cette approche s'avère particulièrement intéressante dans le cadre d'une langue naturelle raisonnablement bien dotée en ressources comme le français, mais dans laquelle la tâche spécifique à résoudre est très sous-dotée.

Dans nos perspectives, nous souhaitons aborder le deuxième verrou scientifique de notre étude, à savoir l'adaptation de la SAT à une modalité orale, qui présente la particularité d'être souvent caractérisée d'un point de vue grammatical par la présence de disfluences, telles que les hésitations, les ellipses ou les faux départs.

Remerciements

Ce travail a bénéficié d'un financement du Fond National Suisse (No. 197864) et de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005).

Références

Brouwers, L., Bernhard, D., Ligozat, A.-L., & François, T. (2012). Simplification syntaxique de phrases pour le français. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, *2*, 211-224.

Buet, F., & Yvon, F. (2021). Toward Genre Adapted Closed Captioning. *Interspeech 2021*, 4403-4407.

Cardon, R., & Grabar, N. (2019). Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification. *Proceedings - Natural Language Processing in a Deep Learning World*, 168-177.

Coordonnées

Lucía Ormaechea
Université de Genève, Université Grenoble-Alpes
lucia.ormaecheagrijalba@unige.ch

Pierrette Bouillon
Université de Genève
pierrette.bouillon@unige.ch

Benjamin Lecouteux
Université Grenoble-Alpes
benjamin.lecouteux@univ-grenoble-alpes.fr

Didier Schwab
Université Grenoble-Alpes
didier.schwab@univ-grenoble-alpes.fr