Université Paris Nanterre
École doctorale 139 – Connaissance, Langage, Modélisation

La résolution de la polysémie à l'aide de modèles de vecteur de mots et la visualisation de données : le cas des postpositions adverbiales *-ey*, *-eyse*, et *-(u)lo* en coréen

par Seongmin Mun

Thèse présentée et soutenue publiquement le 18 juin 2021
en vue de l'obtention du grade de
docteur en Traitement Automatique des Langues
sous la direction de Guillaume Desagulier

Membres du jury :

Directeur : Dr. Guillaume Desagulier	Université Paris VIII & UMR 7114, MoDyCo
Rapporteur : Prof. Iksoo Kwon	Hankuk University of Foreign Studies
Rapporteur : Prof. Laurent Prévot	Aix-Marseille Université
Examinatrice : Dr. Caroline Brun	Naver Labs Europe
Examinatrice : Prof. Iris Taravella	Université Paris Nanterre & UMR 7114, MoDyCo
Examinatrice : Prof. Delphine Battistelli	Université Paris Nanterre & UMR 7114, MoDyCo

Résumé

Ce projet de thèse présente des comptes rendus informatiques de la résolution de la polysémie au niveau des mots dans une langue peu étudiée—le Coréen. Les postpositions, qui se caractérisent par une correspondance forme-fonction multiple et qui sont donc polysémiques par nature, posent un défi à l'analyse automatique et à la performance des modèles pour identifier leurs fonctions. Dans ce projet, je consolide les modèles existants de classification de vecteur au niveau du mot (*Positive Pointwise Mutual Information* et *Singular Value Decomposition* ; *Skip-Gram and Negative Sampling*) en tenant compte du Window du contexte, et j'introduis un modèle de classification de vecteur au niveau de la phrase (*Bidirectional Encoder Representations from Transformers* (BERT)) dans le cadre de la modélisation sémantique distributionnelle. Par ailleurs, je développe deux systèmes de visualisation qui montrent (i) les relations entre les postpositions et leurs mots co-occurrents pour les modèles de vecteur au niveau du mot, et (ii) les clusters entre les phrases pour le modèle de vecteur au niveau de la phrase. Ces systèmes de visualisation ont l'avantage de mieux comprendre comment ces modèles de classification classent les fonctions prévues de ces postpositions. Les résultats montrent que, alors que la performance des modèles de vecteur au niveau du mot est modulée par la taille des corpus d'entraînement contenant les fonctions spécifiques des postpositions, le modèle de vecteur au niveau des phrases est stable (i.e., moins affecté par la taille du corpus) et simule la façon dont les humains reconnaissent la polysémie des postpositions adverbiales coréennes de façon plus appropriée que les modèles de vecteur au niveau du mot.

Mots-clés : polysémie, traitement automatique des langues, classification, modèles de vecteur de mots, visualisation de données, Coréen

1 Introduction

La polysémie, qui est un type d'ambiguïté, se produit lorsqu'une forme exprime des significations/fonctions multiples.

Tableau 1 – Liste de fréquence des sous-fonctions de *-ey*, *-eyse*, et *-(u)lo* dans le corpus validé par croisement

-ey		*-eyse*		*-(u)lo*	
Fonction	Fréquence	Fonction	Fréquence	Fonction	Fréquence
LOC	1,780	LOC	4,206	FNS	1,681
CRT	1,516	SRC	647	DIR	1,449
THM	448			INS	739
GOL	441			CRT	593
FNS	216			LOC	158
EFF	198			EFF	88
INS	69				
AGT	47				
Total	4,715	Total	4,853	Total	4,708

Note. Abréviation : AGT= agent ; CRT= critère ; DIR= direction ; EFF= effecteur ; FNS= état final ; GOL= but ; INS= instrument ; LOC= localisation ; SRC= source ; THM= thème

Bibliographie