COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques Mathématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
Spécialité : *Informatique*

Par

Mérième BOUHANDI

Amélioration endogène des modèles de langue

Application aux domaines de spécialité

Thèse présentée et soutenue à Nantes, le 14 février 2023
Unité de recherche : Laboratoire des Sciences du Numériques de Nantes (LS2N)

Rapporteurs avant soutenance :

Iris ESHKOL-TARAVELLA	Professeure des Universités, Université Paris Nanterre (MODYCO)
Olivier FERRET	Directeur de recherche, CEA (LIST)

Composition du Jury :

Président :	Prénom Nom	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Anne VILNAT	Professeure des Universités, Université Paris-Saclay (LISN)
	Iris ESHKOL-TARAVELLA	Professeure des Universités, Université Paris Nanterre (MODYCO)
	Olivier FERRET	Directeur de recherche, CEA (LIST)
	Thierry CHARNOIS	Professeur des Universités, Université Paris Sorbonne Nord (LIPN)
Directeur de thèse :	Emmanuel MORIN	Professeur des Universités, Nantes Université (LS2N)
Encadrant de thèse :	Thierry HAMON	Maître de conférences, Université Paris-Saclay (LISN)

Titre : Amélioration endogène des modèles de langue : Application aux domaines de spécialité

Mot clés : Traitement automatique du langage naturel, modèles de langue, modèles neuronaux à base de graphes, plongements de mots, domaine spécialisé

Résumé : Aujourd'hui, le champ de recherche sur la modélisation de la langue a atteint une certaine maturité : plusieurs modèles de langue profonds sont disponibles sur plusieurs langues et dans plusieurs domaines. Les performances de ces modèles ont d'ailleurs nettement progressé ces dernières années. Cependant, un enjeu principal demeure : les méthodes et les techniques actuellement utilisées pour construire ou adapter ces modèles, telle l'adaptation fine ou *fine-tuning*, donnent la priorité au volume des données à partir desquelles ils sont construits. Dans le cas des domaines spécialisés, les corpus pour l'entraînement ou l'adaptation de ces modèles sont généralement de taille plus modeste, et ces méthodes se révèlent moins efficaces. Nous avançons ainsi que, dans le cas des petits corpus ou des domaines de spécialité peu dotés, une partie de la structure syntaxique et sémantique du texte n'est pas exploitée lors de l'adaptation fine. Ainsi, un travail d'adaptation au domaine s'avère donc nécessaire. Dans ce travail de thèse, nous proposons une méthode d'adaptation des modèles de langue profonds pour obtenir de meilleurs résultats sur des tâches de spécialité, en prenant compte des informations globales issues de graphes de vocabulaire. Nous réinjectons cette connaissance dans les modèles de langue profonds, améliorant les résultats sur un ensemble de tâches spécialisées. Afin d'évaluer notre méthode, nous menons des expériences sur plusieurs tâches de spécialité. Nous réalisons un premier ensemble d'expériences afin d'adapter nos modèles au domaine. Puis, nous réalisons un autre ensemble d'expériences pour effectuer des analyses quantitatives, montrant que les modèles de langue profonds peuvent bel et bien être adaptés au domaine en utilisant l'approche à base de graphes que nous proposons.

Title: Endogenous Adaptation of Language Models: Application to Specialised Domains

Keywords: Natural language processing, language models, graph neural networks, word embeddings, specialised domain

Abstract: The field of research on language modelling has reached a certain maturity: several deep language models are available in several languages and domains. The performance of these models has improved significantly in recent years. However, one main issue remains. Most of the methods and techniques currently used to build and then adapt these models, such as fine-tuning, rely on significant volumes of data. In the case of specialised domains, the corpora for training or adapting these models are generally smaller, so these methods are often less efficient and produce weaker results on downstream tasks. We argue that in the case of small corpora and specialised domains, some of the syntactic and semantic structure of the text is not exploited during fine-tuning, and more specific domain adaptation is therefore necessary. In this work, we propose a method for adapting deep language models to obtain better results on specialised tasks, considering global information learned using vocabulary graphs. We feed this knowledge to the deep language model, improving the results on most of our specialised tasks. In order to evaluate our method, we conduct experiments on several specialised tasks. We perform the first experiments to adapt our models to the domain. Then, we conduct another set of experiments to perform quantitative analyses, showing that deep language models can be adapted to the domain using our proposed graph-based approach.