Plongements de mots translingues et alignement de textes parallèles pour les dialectes alsaciens

Projet encadré par : Delphine Bernhard (LiLPa, dbernhard@unistra.fr)

Programmation :	▮▮
Analyse / Développement de méthodes originales :	▮▮▮
Synthèse bibliographique :	▮▮
Collecte et annotation de données :	▮▮▮

Les dialectes alsaciens sont actuellement parlés par environ 500 000 locuteurs en Alsace. Il s'agit de dialectes non écrits pour lesquels il existe de nombreuses scripturalisations possibles. L'exemple ci-contre donne un aperçu de la variété de formes qu'il est possible de trouver pour le nom de la ville de Mulhouse.

Cette variabilité à l'écrit et le manque de ressources numériques (corpus de textes notamment) constituent des défis importants pour le développement d'applications de traitement automatique des langues (TAL) pour les dialectes alsaciens.

Il s'agira d'explorer l'utilisation de **plongements de mots translingues** pour les dialectes alsaciens. Les plongements de mots (*word embeddings*) permettent de représenter les mots par des vecteurs de nombres réels et sont nécessaires dans les modèles neuronaux appliqués à diverses tâches (étiquetage morphosyntaxique, analyse de sentiments, recherche d'information, etc.). Les représentations obtenues sont telles que les mots qui partagent des contextes d'occurrence similaires ont également des vecteurs proches et sont donc peu distants dans l'espace vectoriel (voir par exemple « cat » et « kitten » dans la Figure 1 infra)

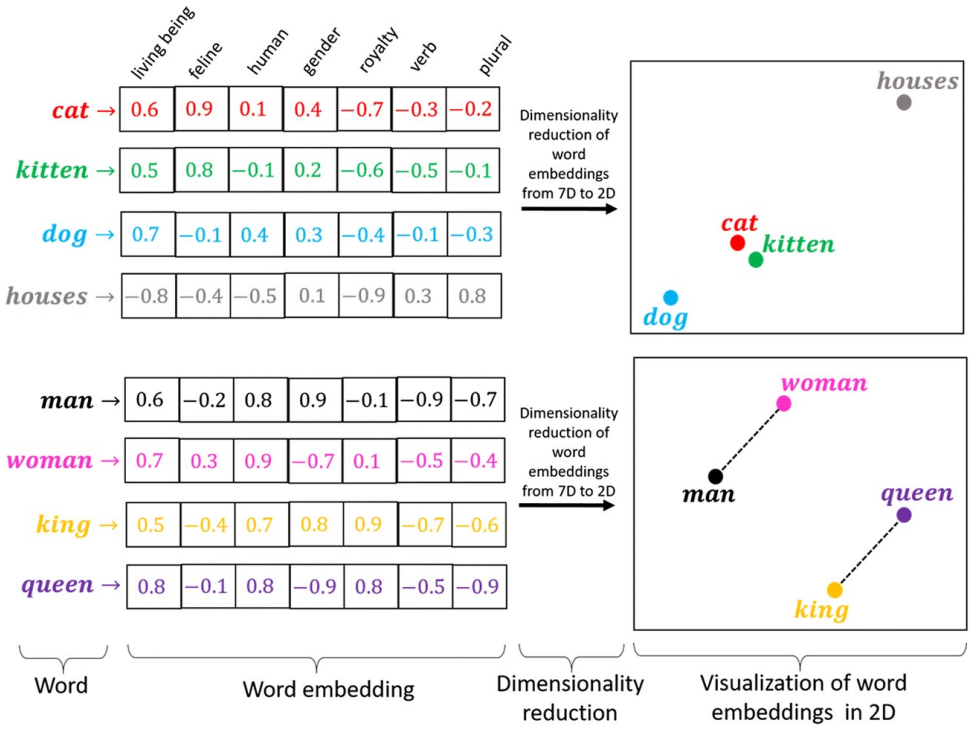


*Figure 1: Illustration des plongements de mots. Source : https://miro.medium.com/max/2598/1*sAJdxEsDjsPMioHyzlN3_A.png*

Les plongements de mots translingues permettent de comparer des mots issus de langues différentes (Ruder et al ., 2017 ;Conneau et al., 2018 ; Doval et al., 2018 ; Wada et al., 2019), grâce à des représentations dans le même espace vectoriel. Dans ce cas, les mots qui sont des traductions sont, idéalement, proches dans l'espace (par exemple « chien » et « hund » dans la Figure 2 infra).

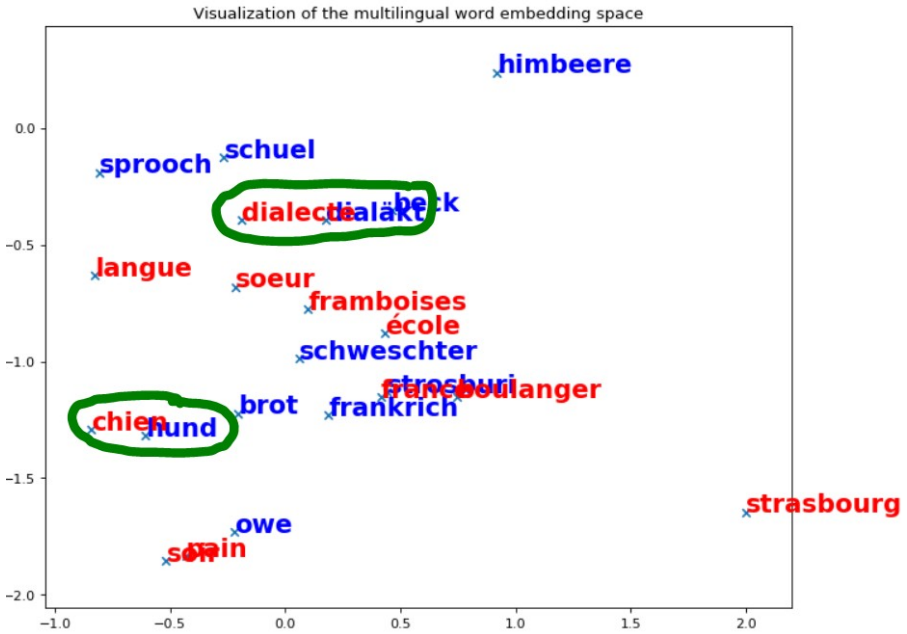
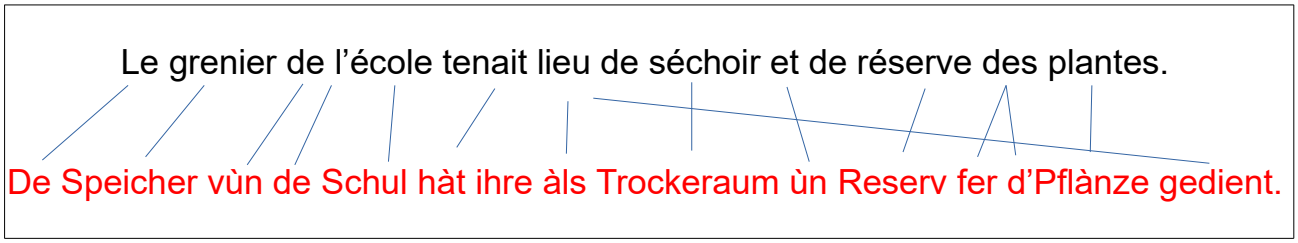


Figure 2: Exemple d'espace bilingue français-alsacien.

L'objectif du projet sera de tester, évaluer et comparer les méthodes proposées pour les langues dites « peu dotées » (c'est-à-dire pour lesquelles il existe peu de données) en les appliquant aux dialectes alsaciens. Il s'agira également de tester des modèles qui sont robustes à la variation graphique (par exemple Doval et al. 2019).
Les plongements de mots translingues seront comparés aux méthodes d'**alignement automatique dans des textes parallèles**. Après alignement, les paires de mots français-alsacien pourront être ajoutées à des lexiques bilingues (Poerner et al., 2018). Le défi sera de parvenir à obtenir des alignements sous-phrastiques pertinents en raison des ressources parallèles en quantité limitée et de la forte variation graphique en alsacien.



Les méthodes sélectionnées devront permettre d'enrichir des lexiques bilingues français-alsacien (voire également allemand-alsacien).

La maîtrise des dialectes alsaciens n'est pas requise. La connaissance de l'allemand sera toutefois un atout non négligeable pour l'analyse des résultats.

Références

- Conneau A., Lample G., Ranzato M., Denoyer L., et Jégou H. (2018). Word Translation Without Parallel Data. *In Proceedings of ICLR*, Consultable à http://arxiv.org/abs/1710.04087 [Accédé le 9 août 2019].
- Doval Y., Camacho-Collados J., Anke L.E., et Schockaert S. (2018). Improving Cross-Lingual Word Embeddings by Meeting in the Middle. *In Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 294–304.

- Doval Y., Vilares J., et Gómez-Rodríguez C. (2019). Towards robust word embeddings for noisy texts, *arXiv:1911.10876 [cs]*. Consultable à http://arxiv.org/abs/1911.10876 [Accédé le 15 janvier 2020].
- Poerner N., Sabet M.J., Roth B., et Schütze H. (2018). Aligning Very Small Parallel Corpora Using Cross-Lingual Word Embeddings and a Monogamy Objective, *arXiv preprint arXiv:1811.00066*.
- Ruder S., Vulić I., et Søgaard A. (2017). A Survey Of Cross-lingual Word Embedding Models, *arXiv:1706.04902 [cs]*. Consultable à http://arxiv.org/abs/1706.04902 [Accédé le 24 juillet 2019].
- Wada T., Iwata T., et Matsumoto Y. (2019). Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pages 3113-3124. Consultable à https://www.aclweb.org/anthology/P19-1300 [Accédé le 14 octobre 2019].