Wordsurf : un outil pour naviguer dans un espace de « Word Embeddings »

Philippe Suignard
EDF R&D, 7 boulevard Gaspard Monge, 91120 Palaiseau, France
philippe.suignard@edf.fr

RESUME

Dans cet article, nous présentons un outil appelé « Wordsurf » pour faciliter la phase d'exploration et de navigation dans un espace de « Word Embeddings » préalablement entrainé sur des corpus de textes avec Word2Vec.

ABSTRACT

Wordsurf : a tool to surf in a "word embeddings" space
In this article we present a tool called "Wordsurf" to facilitate the exploration and navigation phase in a "Word Embeddings" space previously trained on textual corpus with Word2Vec.

MOTS-CLÉS : Word2Vec, GloVe, word embeddings, plongement de mots
KEYWORDS: Word2Vec, GloVe, word embeddings

1 Contexte

Depuis quelques années, sont apparues des méthodes appelées « Word Embeddings » (notées WE par la suite) ou méthodes de plongement de mots en français, comme Word2Vec (Mikolov et al. 2013) ou Glove (Pennington et al. 2014), permettant de transformer des mots en vecteurs, c'est-à-dire de passer d'un espace de représentation discontinu (les mots) à un espace continu (espace vectoriel de grande dimension). Pour ce faire, Word2Vec s'appuie sur des corpus volumineux de données textuelles et utilise un réseau de neurones peu profond cherchant à prédire le mieux possible le contexte des mots. La représentation numérique de ces mots est ensuite très utile pour différentes tâches comme la classification, le clustering, la traduction, l'analyse d'opinions, etc.

Par ailleurs, EDF doit maintenant gérer des corpus textuels de plus en plus nombreux et variés : réclamations de clients, question posées à Laura l'avatar ou chatbot situé sur le site web d'« EDF Particuliers », comptes rendus d'interventions techniques, etc. En compléments des techniques d'exploration de corpus dites « classiques », comme le clustering proposé par Iramuteq (Ratinaud 2009), les technologies de type WE commencent à être utilisées à EDF pour « découvrir » ou explorer le contenu de ces différents corpus.

Pour faciliter cette phase d'exploration et de découverte, et comme les WE sont assez difficiles à interpréter, nous avons développé un outil appelé « Wordsurf » permettant de naviguer ou de « surfer » dans de telles bases ou espaces de mots.

2 L'outil Wordsurf

Wordsurf est un logiciel interactif développé en Java. Il permet de lire un modèle de mots préalablement entrainé par Word2Vec sur un corpus de textes. La démonstration sera illustrée par des résultats obtenus sur un corpus de questions posées à l'agent conversationnel ou chatbot Laura. Les fonctionnalités de Wordsurf sont les suivantes :

- Pour un mot donné, pouvoir rechercher les mots les plus « similaires » et les afficher sous la forme de nuage ou de liste de mots ;
- Visualiser l'évolution de ce nuage de mots en fonction des itérations du modèle ;
- Calculer des similarités entre phrases. Pour cela, trois mesures de similarité ont été implémentées issues de (SONG et al. 2016) ainsi que deux autres reposant sur la comparaison de la distribution des voisins des mots constituant les deux phrases ;
- Un peu plus anecdotique, voire parfois poétique, une fonction permet de calculer les interpolations entre deux mots donnés ;
- Pour un mot particulier, une liste de mots ou tous les mots du corpus, calculer un graphe où les nœuds sont les mots et les arcs traduisent la proximité entre ces mots. Le graphe peut être exporté vers le logiciel Gephi (Bastian et al. 2009) ;
- Pour un mot particulier, une liste de mots ou tous les mots du corpus, générer une visualisation des mots via l'algorithme t-SNE (Maaten et al. 2008) ;
- Pouvoir comparer qualitativement deux modèles « Word2Vec » entrainés avec des paramètres différents : pour cela, le logiciel liste les mots dont la représentation est la plus éloignés entre les deux modèles.

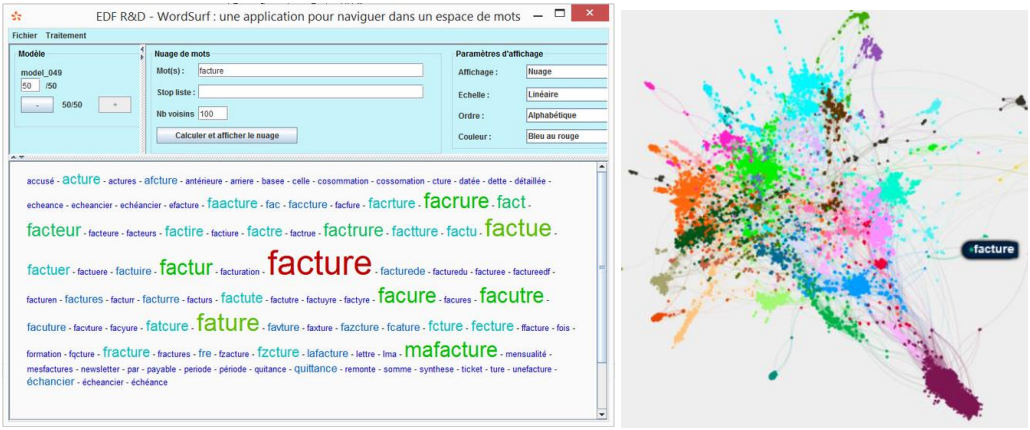

Figure 1 : Interface de l'outil Wordsurf et un exemple d'export de mots via Gephi.

La partie gauche de la Figure 1 présente l'interface de l'outil Wordsurf. On distingue un nuage de mots constitués des 100 mots les plus similaires au mot « facture ». La partie droite de la Figure 1 représente les mots du modèle ainsi que leur proximité exportés au format HTML grâce à un plug-in du logiciel Gephi. Les différentes couleurs représentent les univers lexicaux : les appareils électriques (four, radiateur, etc.), le bâti (toit, coffret, câble, poteau, etc.), les offres, les factures, les justificatifs, etc.

Références

BASTIAN M., HEYMANN S., & JACOMY M. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, 361-362.

MAATEN L. V. D., & HINTON G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S., & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

PENNINGTON J., SOCHER R., & MANNING C. D. (2014, October). Glove: Global Vectors for Word Representation. In EMNLP (Vol. 14, pp. 1532-1543).

RATINAUD P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. *Téléchargeable à l'adresse : http://www.iramuteq.org*

SONG Y., MOU L., YAN R., YI, L., ZHU Z., HU X., & ZHANG M. (2016). Dialogue session segmentation by embedding-enhanced texttiling. *arXiv preprint arXiv:1610.03955*.