Apprentissage de représentations de mots à partir de graphes

Équipe d'encadrement : Thomas Bonald (Telecom Paris), Matthieu Labeau (Telecom Paris), Antoine Saillenfest (OnePoint)

Les représentations de mots (word embeddings) ont suivi depuis leur popularisation en 2013 avec Word2vec une évolution très rapide, et dans de nombreuses directions. C'est un champ de recherche très actif, en particulier dans le cadre de travaux en liens avec l'apprentissage profond (deep learning). Ce champ de recherche suscite également un vif intérêt pour ses nombreuses applications dans le champ industriel : systèmes d'analyse de commentaires clients, systèmes de recommandation de produits, analyse de larges corpus de documents, etc.

La plupart des représentations de mots apprises de manière non-supervisée à partir de grandes quantités de texte peuvent être réparties en deux catégories : les représentations classiques (comme Word2vec, GloVe, Fasttext) et les représentations contextuelles (ELMO, mais surtout BERT et ses dérivés). Les premières sont apprises à partir des comptes de co-occurences des mots dans un corpus, avec des méthodes allant de la factorisation de matrices (SVD -singular value decomposition) à l'utilisation d'architectures neuronales pour prédire les mots à partir de leur contexte immédiat. Les représentations contextuelles, elles, sont obtenues via des architectures neuronales basées sur le mécanisme d'attention qui permettent d'inclure dans la représentation d'un mot des informations sur l'ensemble de la phrase (voir [Vaswani et al, 2017] pour l'utilisation de l'auto-attention dans l'architecture Transformer). Voir par exemple [Liu et al, 2020] pour une vue d'ensemble.

Il existe de nombreux liens entre l'apprentissage de représentations de graphes et de mots. Outre la factorisation de la matrice de co-occurence, des modèles de marche aléatoire (random walk), classiques dans le domaine de l'analyse de graphes, ont été associés à un modèle génératif pour apprendre des représentations de mots [Arora et al, 2015]. Inversement, l'idée de "prédiction du voisin" associée à un échantillonnage d'exemples négatifs (negative sampling), centrale dans l'apprentissage des représentations de mots dans Word2vec, a été transposée aux graphes [Grover et al, 2016]. Il existe par ailleurs plusieurs travaux qui ont cherché à utiliser les méthodes prévues pour des graphes pour obtenir des représentations de mots en se servant de bases de connaissances présentant des informations lexicales et sémantiques sous une structure de graphe telles que WordNet. [Salawa et al, 2019] en offre une vue d'ensemble, en cherchant à les comparer aux représentations classiques, tandis que [Daix et al, 2019] et [Sen et al, 2019] cherchent à les combiner. D'autres ont cherché à exploiter les structures de graphes naturellement présentes dans les données ; on peut citer par exemple, [Vashishth et al, 2019] qui propose de construire des représentations de mots à l'aide d'un graphe basé sur l'analyse en dépendance du contexte, et de les améliorer à l'aide d'un graphe de

vocabulaire basé sur le corpus. Dans [Ryabinin et al, 2020], les auteurs imposent aux représentations qu'elles suivent aussi une structure de graphe.

Il existe différentes pistes possibles de recherche en ce qui concerne l'utilisation des graphes pour l'apprentissage de représentations de mots :

Partant de l'idée que la self-attention utilisée pour obtenir les représentations contextuelles peut être vu comme l'apprentissage de poids sur un graphe connexe dont les nodes sont les mots de la phrase [Joshi et al, 2020], on peut réfléchir à la construction d'un modèle ressemblant à BERT, dont le but est de prédire les mots manquants d'une phrase, mais basé sur des graphes. Parallèlement, on pourra s'intéresser à l'apprentissage contrastif (voir par exemple [van der Oord et al, 2018]) pour chercher à améliorer les méthodes d'apprentissage basées sur les graphes, et introduire de la supervision venant de bases de connaissances.

Une seconde direction possible est de s'intéresser à la polysémie et à la détermination du ou des sens d'un mot, en travaillant directement sur le graphe biparti mot-contexte (comme l'avait fait [Schutze, 1998] initialement, bien avant Word2Vec), le contexte pouvant prendre différentes formes (mots proches comme dans Word2Vec, phrase, paragraphe, document, etc.) et les arêtes pondérées en conséquence. Un clustering "soft" de ce graphe, par exemple par la méthode de Louvain, permet d'identifier très rapidement les mots apparaissant dans différents contextes, sans passer par le clustering dans l'espace vectoriel des mots [Chang et al, 2018], lent et coûteux. En distinguant ces contextes, on pourrait ainsi calculer une représentation pour chaque sens d'un mot, permettant d'avoir une bien meilleure représentation vectorielle de l'ensemble du vocabulaire.

Les approches neuronales pour la génération de représentations de mots, d'abord classiques, permettent d'effectuer des tâches de calcul de similarité de mots - et aussi, d'analogie (ainsi, la relation de presque-linéarité entre les couples de vecteurs formant une analogie semble avoir été récemment expliquée par la théorie [Allen et Hospedales, 2019]). Dans le cas des représentations contextuelles, on pourra interroger les modèles en interprétant les structures d'attention des "transformers" de ces modèles [Clark et al, 2019]. Bien qu'il existe des façon d'obtenir des informations humainement interprétables sur les relations entre les représentations via des tâches spécifiquement crées (de "probing", pour des phénomènes linguistiques [Tenney et al, 2019] ou des faits [Petroni et al, 2019]) qui sont indépendantes du modèles, des efforts devraient être fait pour les inclure dans la procédure d'apprentissage, dans cadre d'une demande de plus en forte d'explicabilité dans ces systèmes [Googman et Flaxman, 2016].

Ainsi, une troisième piste à explorer serait d'utiliser la structure présente dans les données sous forme de graphe pour enrichir les représentations apprises avec des informations humainement compréhensibles sur les relations qui les relient, tout en préservant les performances du modèle d'apprentissage. [Ryabinin et al, 2020] a effectué un premier pas dans une direction parallèle, en

forçant les représentations apprises à s'organiser sous forme de graphe, par souci d'interprétabilité.

1 Références

[Liu et al, 2020] Qi Liu, Matt J. Kusner†, Phil Blunsom, A Survey on Contextual Embeddings

[Arora et al, 2015] Sanjeev Arora, Y. Li, Yingyu Liang, Tengyu Ma, Andrej Risteski, RANDWALK : A Latent Variable Model Approach to Word Embeddings

[Grover et al, 2016] Aditya Grover, Jure Leskovec, node2vec : Scalable Feature Learning for Networks

[Salawa et al, 2019] Małgorzata Salawa, António Branco, Ruben Branco, João António Rodrigues, Chakaveh Saedi, Whom to Learn From ? Graph- vs. Text-based Word Embeddings

[Daix et al, 2019] Pierre Daix-Moreux, Matthias Gallé, Joint Semantic and Distributional Word Representations with Multi-Graph Embeddings

[Sen et al, 2019] Procheta Sen, Debasis Ganguly, Gareth Jones, Word-Node2Vec : Improving Word Embedding with Document-Level Non-Local Word Co-occurrences

[Vashishth et al, 2019] Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, Partha Talukdar, Incorporating Syntactic and Semantic Information in Word Embeddings using Graph Convolutional Networks

[Ryabinin et al, 2020] Max Ryabinin, Sergei Popov, Liudmila Prokhorenkova, Elena Voita, Embedding Words in Non-Vector Space with Unsupervised Graph Learning

[Joshi et al, 2020] Chaitanya Joshi, Transformers are Graph Neural Networks

[van der Oord et al, 2018] Aaron van den Oord, Yazhe Li, Oriol Vinyals, Representation Learning with Contrastive Predictive Coding

[Vaswani et al, 2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762 (2017).

[Schutze, 1998], Hinrich Schütze, Automatic word sense discrimination

[Chang et al, 2018], Haw-Shiuan Chang, Amol Agrawal, Ananya Ganesh, Anirudha Desai, Vinayak Mathur, Alfred Hough, Andrew McCallum, Efficient Graph-based Word Sense Induction by Distributional Inclusion Vector Embeddings

[Allen and Hospedales, 2019] Allen, C Hospedales, T, Analogies Explained : Towards Understanding Word Embeddings

[Clark et al, 2019] Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning, What Does BERT Look At ? An Analysis of BERT's Attention

[Tenney et al, 2019] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas

McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick, What do you learn from context ? Probing for sentence structure in contextualized word representations

[Petroni et al, 2019] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller, Language Models as Knowledge Bases ?

[Googman Flaxman, 2016] Goodman, B Flaxman, S. (2016) European Union regulations on algorithmic decision-making and a ”right to explanation”