Apprentissage et évaluation de plongements lexicaux sur un corpus SNCF en langue spécialisée

Nicolas Dugué*, Nathalie Camelin*, Luce Lefeuvre**, Xining Li*, Coralie Reutenauer**, Cyndel Vaudapiviz**

*Le Mans Université, LIUM, EA 4023,
Laboratoire d'Informatique de l'Université du Mans
**SNCF INNOVATION & RECHERCHE
1-3, avenue Francois Mitterrand,93210 SAINT-DENIS

Résumé. Au sein du groupe SNCF, le programme PRISME d'excellence sécurité intègre une démarche de simplification de l'accès à l'information et de la production de contenus dans la documentation métier. Dans ce contexte, nous avons mis en œuvre des traitements sur un corpus de référentiels métiers SNCF afin de guider l'utilisateur dans sa recherche documentaire. Les travaux présentés visent à évaluer l'usage des plongements lexicaux pour générer des représentations sémantiques denses sur lesquelles se baseront des méthodes de deep learning pour structurer le corpus SNCF. Le protocole mis en place consiste en l'évaluation empirique des voisinages de mots par des experts. Dans cette étude, nous montrons les difficultés d'apprentissage et d'évaluation inhérentes à ce type de corpus avec de nombreux mots soit très spécifiques, soit polysémiques, rendant la construction d'un espace de représentations robuste difficile.

1 Introduction

Au sein du groupe SNCF, la documentation métier est aujourd'hui en pleine mutation, avec des métiers qui se digitalisent, plus mobiles et marqués par de nouveaux modes de consommation de l'information. Dans le cadre du programme PRISME de transformation en matière de sécurité ferroviaire, SNCF cherche à simplifier l'accès à l'information et la production de contenus dans la documentation métier. L'évaluation de nouveaux systèmes intelligents d'accès aux contenus, d'aide à l'interprétation et à la saisie participe à cette démarche.

Dans ce contexte, nous avons mis en œuvre des traitements sur un corpus de référentiels métiers SNCF pour répondre à deux objectifs. Le premier est celui de guider l'utilisateur dans sa recherche documentaire à travers la structuration des résultats de recherche, en regroupant en thématiques (classification *non supervisée*) les documents retournés en réponse à une requête. Le second objectif est d'aider les rédacteurs à qualifier automatiquement de nouveaux documents selon des thèmes définis dans une arborescence construite par les experts métiers (classification *supervisée*). Les enjeux dans notre contexte d'application industrielle sont de disposer de systèmes performants (réponse rapide et pertinente), maîtrisés (résultats interprétables et explicables) et adaptés aux besoins des utilisateurs de ces systèmes (pertinence des

descripteurs par rapport aux connaissances de l'utilisateur). Dans le cas des deux objectifs présentés, et dans ce contexte applicatif, le choix de la représentation des contenus (descripteurs lexicaux) constitue une étape préalable à l'application d'algorithmes d'apprentissage.

Les travaux présentés ici évaluent l'usage des *plongements lexicaux* (word embeddings) comme moyen de représenter les contenus. Ils abordent en particulier la question de l'apprentissage et de l'évaluation de ces plongements sur un corpus en langue française et spécialisée. Si nous nous intéressons à ces plongements, c'est tout d'abord parce qu'ils aboutissent à des représentations des contenus en faible dimension, ce qui permet d'accélérer les traitements et donc de proposer des réponses rapides à l'utilisateur. Par ailleurs, cette efficacité en calcul ne se fait pas au détriment de la qualité des résultats puisque ces vecteurs encapsulent une information sémantique riche contrairement aux représentations creuses dites *one-hot*.

Nous décrirons Section 2 les spécificités du corpus sur lequel nous travaillons et verrons que ces particularités ont donné lieu à des pré-traitements adaptés. Ensuite, nous détaillerons Section 3 la méthode d'apprentissage utilisée pour apprendre les plongements lexicaux. Nous discuterons Section 4 le protocole d'évaluation ainsi que les premiers résultats. Enfin, nous ouvrirons la discussion en considérant la polysémie des mots du vocabulaire spécialisé.

2 Corpus et pré-traitements

Les données sont constituées de 7029 textes, avec un contenu technique relatif à la sécurité, à l'exploitation et à l'utilisation du réseau ferroviaire. À cette base documentaire s'ajoutent un lexique ferroviaire et une base d'acronymes fournis par la SNCF.

Initialement au format pdf, le corpus a été converti au format txt et nettoyé (élimination d'erreurs liées à des problèmes de conversion, suppression de métadescripteurs des documents). Plusieurs pré-traitements ont ensuite été appliqués au corpus pour obtenir une représentation vectorielle robuste du contenu des documents. Le contenu a été découpé en unités lexicales selon les standards d'Unitex. Le corpus a ensuite été lemmatisé par une version du Lefff (Lexique des Formes Fléchies du Français, (Sagot (2010)) adaptée pour prendre en compte certaines spécificités du corpus SNCF : les problématiques SNCF étant étroitement liées à un ancrage territorial, les noms de communes françaises ont été ajoutés ; des variantes de graphie récurrentes dans le corpus (même mot avec ou sans accent, ou graphie tantôt avec oe ou avec œ) ont également été ajoutées. Le lexique a ensuite été filtré, pour conserver uniquement les lemmes de type noms propres, noms communs, verbes, adverbes et adjectifs, et les termes répertoriés dans les ressources lexicales fournies par SNCF (lexique et acronymes).

À l'issue des pré-traitements, la taille du vocabulaire est de 18k mots et la taille du corpus de l'ordre de 10^7. Ces pré-traitements ont été appliqués itérativement de façon à améliorer la qualité des embeddings, notamment en ce qui concerne le vocabulaire très spécifique du corpus (voir Section 4.2). Empiriquement, nous avons constaté une réduction du bruit dans les résultats de l'apprentissage que nous décrivons à la section suivante.

3 Plongements lexicaux

L'objectif est d'apprendre des vecteurs denses pour représenter notre vocabulaire : un espace de représentation des contenus de dimension réduite permet des temps de réponse plus

rapide qu'avec une matrice creuse de grande dimension, et les résultats de classification et de clustering bénéficient de l'utilisation des plongements (Kim (2014)). Dans notre cas, le corpus est de petite taille en comparaison des corpus utilisés pour l'apprentissage de plongements lexicaux à l'état de l'art (Mikolov et al. (2013)). Nous privilégions donc une approche basée sur la décomposition en valeur singulière (SVD) et décrite par Levy et al. (2015). Celle-ci est efficace sur de petits corpus et ses performances sont comparables à l'état de l'art.

Dans un premier temps, nous créons la matrice de cooccurrence termes-termes en utilisant une taille de fenêtre contextuelle de 5 mots. La matrice ainsi générée est une matrice symétrique creuse, de dimension égale à la taille du vocabulaire. Dans un second temps, les fréquences de cooccurrence de chaque paire de termes sont pondérées de façon à refléter leur significativité. Pour ce faire, nous calculons une variante de l'information mutuelle, la *PPMI*. Enfin, nous procédons à une réduction de dimension par application de la SVD. La taille du nouvel espace de représentation est un paramètre du modèle, elle est dans notre cas fixée arbitrairement à 200, une valeur prise dans l'intervalle des valeurs communes de l'état de l'art.

Nous souhaitons attirer l'attention sur le fait que, sur un tel corpus, il est important d'apprendre des plongements spécifiques. À titre d'exemple, prenons le mot courant *manette*. Parmi ses 10 plus proches voisins avec notre approche, 7 sont des acronymes SNCF, ce qui aurait été impossible d'obtenir avec des plongements appris sur un autre corpus.

4 Évaluation des plongements lexicaux

4.1 Protocole d'évaluation

Afin d'évaluer la qualité de l'espace de représentation construit sur le corpus SNCF, nous proposons dans cette section un protocole d'évaluation sollicitant plusieurs experts SNCF. L'évaluation consiste à valider la pertinence de l'association de deux mots donnés. Par exemple, l'association de *train* et *wagon* est pertinente, tandis que l'association de *billet* et *passage à niveau* ne l'est pas. Nous proposons de solliciter l'expert sur la pertinence de l'association entre un mot donné et ses 6 plus proches voisins dans l'espace de représentation généré via l'approche décrite Section 3. Le résultat d'évaluation de l'association est binaire : pertinente ou non pertinente. Ce protocole permet de simplifier le travail des experts dont le temps est précieux. Ceci nous permet de couvrir suffisamment le corpus et ainsi de garantir la significativité statistique des résultats. Enfin, une tâche claire laisse moins de place au subjectif (c'est le cas avec plusieurs degrés de similarité).

Les mots évalués ont été regroupés et proposés aux experts selon 4 catégories : Mots issus du lexique SNCF (*Lexique*) ; Acronymes SNCF polysémiques (*Acr. poly*) ; Acronymes SNCF non polysémiques (*Acr.*) ; N-grammes fréquents dans le corpus (*ngram fréq.*, $n \in \{1, 2, 3\}$). Pour chaque catégorie, on considère la fréquence d'apparition des mots dans le corpus pour échantillonner notre vocabulaire. Deux fois 20 mots de chacune des catégories sont présentés aux experts : 20 parmi les mots les plus fréquents, et 20 parmi ceux dont la fréquence se situe au niveau de la médiane de la distribution. Ainsi chaque expert se voit proposer 160 mots auxquels sont associés les 6 mots les plus proches, soit 960 paires de mots à juger comme étant pertinentes ou non. Le nombre d'associations différentes proposées à l'évaluation est détaillée par catégorie et fréquence Table 1.

	Lexique		Acr.		Acr. poly		ngram fréq.	
Fréquence	++	=	++	=	++	=	++	=
#Asso unique	585	635	1068	978	968	980	952	982

TAB. 1: Nombre d'associations différentes proposées à l'évaluation des experts selon les 4 catégories et selon la fréquence des termes : haute (++) ou médiane (=).

Les experts. Neuf experts SNCF ont participé. Il s'agit de responsables ou chefs en poste depuis en moyenne 10 ans (de 9 mois à 25 ans à la SNCF) exerçant à des postes variés : documentation métier, sécurité système, qualité et performance, organisation de travaux, *etc.*

L'interface. Une plateforme regroupant plusieurs formulaires web a été développée par le LIUM. Pour chacune des 4 catégories, un tableau est proposé contenant le mot et ses 6 voisins. Les 6 voisins sont associés à une case à cocher. Si l'expert estime que le mot voisin n'est pas en relation avec le mot courant alors il coche la case. S'il estime que l'association des deux mots est correcte, il n'a aucune action à faire. De plus, une fonctionnalité permettant d'indiquer que le mot n'est pas connu est proposée afin de bien faire la différence entre une association qui ne serait pas pertinente et une association qui ne peut être évaluée car au moins l'un des mots n'est pas connu. Nous estimions le temps d'annotation d'un formulaire à moins d'une heure (environ 20 secondes pour les 6 associations à un mot). Les experts ont en effet mis de une demi-heure à une cinquantaine de minutes pour annoter un formulaire.

4.2 Résultats

Analyse quantitative. La Table 2 présente les résultats numériques issus de l'évaluation par les experts. La somme des pourcentages des associations pertinentes et non pertinentes donne 100, les associations inconnues ont été écartées. En premier lieu, on constate de fortes différences entre les deux catégories d'acronymes d'une part, et les catégories *lexique* et *ngram fréq* d'autre part. Dans les deux premières catégories, le pourcentage d'associations inconnues est très élevée (entre 40 et 50%) tandis que dans les autres, il est faible (moins de 10% dans le cas des expressions fréquentes). On constate également cette dichotomie sur l'accord inter-évaluateur, plus faible sur les catégories d'acronymes que sur les autres.

	Lexique		Acr.		Acr. poly		ngram fréq.	
Fréquence	++	=	++	=	++	=	++	=
% asso ok	69	59	58	73	65	59	70	66
% asso ko	31	41	42	27	35	41	30	34
% asso ? ?	**4**	7	37	49	45	56	**8**	21
% p. accord	**71**	68	55	56	58	67	**73**	57

TAB. 2: Pourcentages d'associations validées (*ok*), jugées non pertinentes (*ko*) ou contenant un mot inconnu (*? ?*). Pourcentage d'accord inter-évaluateurs (*p. accord*).

En second lieu, on constate que les termes fréquents (++) sont mieux connus que les autres et donnent ainsi lieu à moins d'associations inconnues. L'accord inter-évaluateur est également

plus élevé dans cette sous-catégorie, quelque soit la catégorie des termes. Ces deux constats montrent la difficulté de conduire une évaluation concernant le vocabulaire spécifique : les acronymes qui sont très spécifiques sont souvent inconnus (en particulier les polysémiques), le vocabulaire moins fréquent est moins connu.

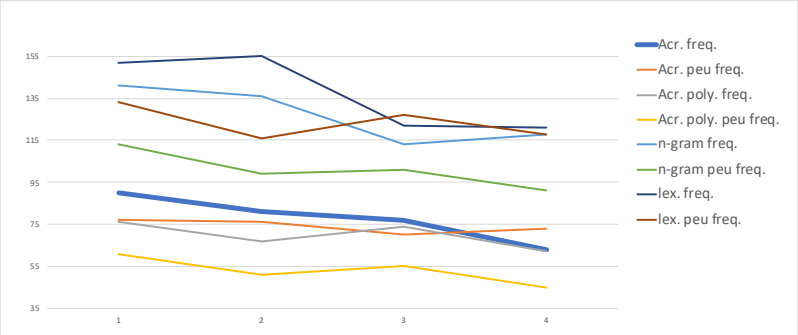

FIG. 1: Nombres d'associations jugées pertinentes en fonction de la catégorie (couleur) et de la proximité du voisinage (abscisse) comprise entre 1 (le mot le plus proche), et 4 (le quatrième mot le plus proche dans l'espace appris). Les mots les plus proches sont jugés plus pertinents.

Les résultats de l'évaluation montrent que 30 à 40% des paires proposées ne sont pas jugées pertinentes. Ces chiffres pris dans l'absolu ne sont pas très informatifs, ils mettent simplement en avant la difficulté de la tâche. En revanche, ils vont servir de référence pour les prochaines évaluations prévues avec les experts. Cela montre également que le nombre de 6 associations par terme est peut-être trop élevé et conduit à plus d'erreurs. En effet, la Figure 1 nous montre que les premières associations proposées sont systématiquement plus pertinentes.

Analyse qualitative. La catégorie des acronymes a été jugée comme étant la plus compliquée à évaluer. Ce résultat n'est en outre pas surprenant étant donné l'importante polysémie des acronymes au sein du groupe. L'acronyme n'est bien souvent pas connu de l'évaluateur. Deux stratégies sont alors mises en œuvre : i) l'expert recherche la signification de l'acronyme et évalue les associations proposées ; ii) il indique via l'interface que l'acronyme lui est inconnu. Dans de rares cas, les mots proposés ont permis d'élucider le sens d'un acronyme. Dans le cas des associations pertinentes, les experts remarquent que les mots proposés apparaissent potentiellement au sein de la même phrase. Cela signifie que les experts réfléchissent au contexte de l'acronyme pour évaluer son sens. D'autre part, les stratégies mises en oeuvre ici pour inférer le sens des acronymes lorsqu'il est inconnu montrent qu'il est difficile d'évaluer une paire de mots de manière indépendante des autres paires de mots.

La spécificité locale (au sens géographique) de certains acronymes a été évaluée comme surprenante et non pertinente par rapport à la réalité du terrain. Par exemple, pour le sigle *GL* (Grande Ligne) est proposé *Saint Denis*. En réalité, les trains grandes lignes passent bien par Saint-Denis, mais une proposition comme *Gare du Nord* aurait été jugée plus pertinente et plus dimensionnante. Cet exemple illustre la difficulté à intégrer des connaissances métiers dans la modélisation du vocabulaire de spécialité.

Afin de mettre en évidence la difficulté de l'évaluation sur les acronymes et leur polysémie, nous discutons quelques exemples. Dans le cas de *CTRL*, l'apprentissage a capté le sens de la touche du clavier *control*. Parmi les voisins proposés et validés par un évaluateur, on observe *cliquer*, *menu*, *sélectionner*, *clavier numérique*. Or, l'acronyme au sens SNCF était défini comme *CTRL - Channel tunnel rail link*. On constate ainsi l'importance de gérer la polysémie sous peine de créer de faux positifs. Au contraire, dans le cas de *TIS*, le sens qui ressort de notre apprentissage est l'un de ceux défini dans le lexique SNCF comme *TIS - Technique d'intervention de la surveillance générale*. Les voisins proposés sont *enseignement général*, *épreuve*, *épreuve orale*, *évaluation finale*, *jury final* et *épreuve finale* qui semblent tous pertinents. Pourtant, l'un des évaluateurs a considéré toutes ces associations comme non pertinentes. Il ne connaissait pas ce sens de l'acronyme TIS, ce qui a conduit à ces faux négatifs.

5 Conclusion

Ce travail met en évidence la difficulté de travailler sur un corpus en langue spécialisée. L'évaluation avec des experts s'est révélée difficile : les niveaux d'accord entre évaluateurs sont bas, leurs connaissances ne leur permettent d'annoter qu'une partie des associations - les autres mots de vocabulaires leur sont inconnus, et la polysémie des acronymes crée des faux positifs (*CTRL*) ainsi que des faux négatifs (*TIS*). Dans la suite de notre travail, nous souhaitons donc travailler sur l'apprentissage de plongements multi-prototypes (Tian et al. (2014)). Dans l'état de l'art, le nombre de sens (donc de prototypes) pour chaque mot est fixé arbitrairement. Les ressources lexicales de la SNCF nous donnent pour chaque acronyme, la liste de ses *sens*, chaque sens étant une liste des mots auxquels correspondent les initiales de l'acronyme. Nous prévoyons ainsi d'utiliser ces ressources pour guider l'apprentissage.

Références

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

Levy, O., Y. Goldberg, et I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL 3*, 211–225.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *LREC*.

Tian, F., H. Dai, J. Bian, B. Gao, R. Zhang, E. Chen, et T.-Y. Liu (2014). A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pp. 151–160.

Summary

This paper investigates the evaluation of word embeddings trained on a SNCF corpora with a very specific vocabulary. In particular, we consider this task regarding SNCF acronyms and lexicon, and show its hardness. We also highlight the relevance of considering polysemy.