Construction de plongements de concepts médicaux sans textes

Vincent Claveau

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Construction de plongements de concepts médicaux sans textes

Vincent Claveau
Univ. de Rennes, CNRS, IRISA
Campus de Beaulieu
35042 Rennes, France
vincent.claveau@irisa.fr

RÉSUMÉ

Dans le domaine médical, beaucoup d'outils du TAL reposent désormais sur des plongements de concepts issus de l'UMLS. Les approches existantes pour générer ces plongements nécessitent de grandes quantités de documents médicaux. Au contraire des ces approches, nous proposons dans cet article de nous appuyer sur les traductions en japonais, plus précisément en kanjis, disponibles dans l'UMLS pour générer ces plongements. Testée sur différents jeux d'évaluation proposés dans la littérature, notre approche, qui ne requiert donc aucun texte, donne de bons résultats comparativement à l'état-de-l'art. De plus, nous montrons qu'il est intéressant de les combiner avec les plongements – contextuels – existants.

ABSTRACT

Embedding medical concepts without texts.

In the medical field, many TAL tools are now based on embeddings of concepts from the UMLS. Existing approaches to generate these embeddings require large amounts of medical data. Contrary to these approaches, we propose in this article to rely on Japanese translations of the concepts, more precisely in Kanjis, available in the UMLS to generate these embeddings. Tested on different evaluation tasks proposed in the literature, our approach, which therefore requires no text, yields good results compared to the state of the art. Moreover, we show that it is interesting to combine them with existing — contextual-based — embeddings.

MOTS-CLÉS : TAL médical, plongements de concepts, CUI, UMLS, kanjis.

KEYWORDS: Biomedical NLP, concept embedding, CUI, UMLS, kanjis.

1 Introduction

De nombreuses techniques de TAL reposent désormais sur les réseaux de neurones ; ceux-ci nécessitent en entrée des représentations numériques des mots – les plongements de mots. Le domaine médical ne fait pas exception et la plupart des travaux exploitent donc désormais des approches neuronales. La question de la construction de plongements adaptés au vocabulaire médical a donc été posée depuis quelques années.

Plus particulièrement, plusieurs travaux se sont intéressés à produire des plongements des concepts médicaux identifiés dans l'UMLS (CUI - *Concept Unique Identifiers*). Le MetaTheusaurus de l'UMLS (Tuttle *et al.*, 1990) recense en effet des entités et des termes médicaux (termes simples ou complexes, voire portion de phrases) dans de nombreuses langues, et rattachés à ces identifiants uniques de

Volume 2 : Traitement Automatique des Langues Naturelles, pages 181–188.
HAL : hal-02784766.

concepts.

Dans cet article, nous revenons sur la question de ces représentations de CUI en proposant une nouvelle approche pour leur construction. Au contraire des travaux existants qui exploitaient de très grosses quantités de textes du domaine pour produire les plongements de CUI, notre approche ne nécessite aucun texte d'entraînement et repose uniquement sur l'UMLS et en particulier les traductions en kanjis. Deux versions de cette approche, l'une produisant une représentation creuse et l'autre dense, sont ici décrites et comparées aux plongements existants. Enfin, nous proposons une combinaison de nos plongements et de ceux de l'état-de-l'art, reposant sur une approche distributionnelle standard.

Nous revenons dans la section suivante sur les travaux existants cherchant à produire des plongements de CUI et sur ceux exploitant les propriété des kanjis. Nous présentons ensuite successivement nos deux représentations, dont les performances sont comparées à l'état-de-l'art en section 4. Nous terminons en donnant quels pistes de travaux qui nous semblent prometteuses.

2 Travaux connexes

L'UMLS est une ressource largement utilisée pour des tâches de TAL biomédical. Son MetaThesaurus rassemble de nombreuses terminologies dans plusieurs langues qui sont agrégées grâce à des identifiants de concepts : chaque terme est rattaché à un identifiant de concept ou CUI. Les CUI permettent au sein d'une langue de trouver les variantes d'un terme (plusieurs termes de la même langue ayant un même CUI), de même qu'ils permettent de trouver des traductions (un CUI partagé par des termes de langues différentes). Par ailleurs, les CUI sont rattachés à des types sémantiques organisés hiérarchiquement.

La construction de plongements de CUI a fait l'objet de plusieurs travaux (De Vine *et al.*, 2014; Choi *et al.*, 2016; Beam *et al.*, 2020). Tous adoptent une approche similaire et les outils usuels de cette tâche, tels que Glove (Pennington *et al.*, 2014) ou Word2vec (Mikolov *et al.*, 2013). Ainsi, ils sont tous construits à partir de textes médicaux anglais. La seule particularité, par rapport à un plongement lexical standard, est la phase de pré-traitement permettant de repérer les termes de l'UMLS dans les phrases et de les associer à leur CUI. Cette étape est faite avec l'outil Metamap (Aronson, 2001).

Il est important de noter plusieurs points communs à ces travaux. Ils reposent tous sur l'anglais, du fait de l'outil MetaMap qui n'existe que pour l'anglais, et de la disponibilité de quantités de textes cliniques.

Ainsi, De Vine *et al.* (2014) ont utilisé 350 000 résumés de journaux du domaine médical pour générer avec word2vec (skip-gram) des plongements pour 60 000 concepts (différentes dimensionnalités ont été testées, 200 étant celle obtenant les meilleures performances).

Choi *et al.* (2016) utilisent également word2vec et une SVD sur 4 millions de donnés d'assurances santé et 20 millions de notes cliniques pour générer 28 000 plongements de dimension 200.

Enfin, les plongements les plus performants et les plus utilisés sont CUI2vec (Beam *et al.*, 2020). Il s'agit de plongements dans $\mathbb{R}^{500}$ de 110 000 CUI qui ont nécessité une quantité gigantesque de textes (plus de 80 millions de documents, incluant ceux des travaux précemmet cités) pour être générés.

Utiliser des représentations par kanjis de termes médicaux a déjà été fait dans un contexte d'analyse morphologique (Claveau & Kijak, 2013). Dans ce travail, les kanjis servaient d'indices pour aider à

décomposer les termes morphologiquement complexes (eg. photo|chimio|thérapie). Bien que pour des buts différents, notre approche repose sur la même idée d'utiliser les kanjis comme des éléments atomiques de sens.

3 Plongements par kanjis

Comme nous l'avons expliqué, l'idée directrice de notre travail est d'exploiter les termes en kanjis pour s'en servir comme des représentations sémantiques des concepts (CUI). De telles traductions sont disponibles directement dans le MetaThesaurus de l'UMLS dans lequel des termes en plusieurs langues sont rattachés aux identifiants de concept CUI. L'intérêt des kanjis est qu'ils offrent une représentation ayant une portée sémantique (chaque kanji pris isolément a un – ou plusieurs – sens), ne sont pas soumis à de la variation morphologique, et sont en nombre limités (moins de 2 000 dans l'UMLS) comparativement aux mots en anglais par exemple. Dans l'UMLS version 2019AB que nous utilisons pour les expériences, il y a 72 000 CUI ayant au moins un terme en japonais. Par exemple, le concept identifié par le CUI C0031740, qui concerne la photochimiothérapie, a plusieurs réalisations en kanjis dans l'UMLS : 光化学法, 光力学的法, 光力学治, 光力学的治, 光力学的治, 光力学的治法.

3.1 Plongements creux

Le premier plongement que nous proposons est simplement la projection des CUI dans l'espace des kanjis. C'est-à-dire qu'un CUI est décrit par le vecteur pondéré construit à partir du sac de kanjis des termes relevant de ce CUI. Le concept C0031740 est donc représenté par un vecteur nul sauf aux dimensions correspondant aux kanjis 光, 化, 学, , 法, 力, 治, ... L'espace de représentation, après suppression des kanjis n'apparaissant qu'une fois, est $\mathbb{R}^{1462}$.

Plusieurs pondérations ont été testées : binaire (ou *one-hot*), Hellinger, TF, TF-IDF, Okapi-BM25 (Robertson *et al.*, 1998). Pour des raisons de place, seule la plus performante, Okapi-BM25, est rapportée ici. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par kanjis.

3.2 Plongements denses

Les vecteurs obtenus par l'approche précédente sont de dimension importante, mais creux (beaucoup de dimensions à 0). Il peut être intéressant d'en proposer une version dense et de dimension inférieure, notamment si la perspective est d'utiliser ces plongements dans des réseaux de neurones pour une application donnée.

Pour ce faire, on recourt à des techniques de réductions de dimensions (Sarveniazi, 2014). Plusieurs techniques standard on été testées. Pour des raisons de place nous ne rapportons les résultats que pour l'Analyse en Composantes Principales (ACP) qui a obtenu parmi les meilleurs résultats avec des temps de calcul courts. Sauf indication contraire, dans les expériences rapportées ci-dessous, nous avons choisi de réduire en 500 dimensions, à des fins de comparaison avec CUI2vec. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par ACP.

3.3 Fusion de plongements

Les plongements que nous proposons reposent sur des indices très différents des indices contextuelles habituellement exploités par Glove, word2vec ou autre. Notre approche ne repose donc pas sur l'hypothèse distributionnelle pour capturer le sens des CUI. On peut ainsi espérer que les notions encodées par les plongements par kanjis et les plongements contextuels de l'état-de-l'art soient complémentaires. Pour tester cette hypothèse, nous proposons de fusionner le plongement CUI2vec (Beam *et al.*, 2020) et nos plongements denses. Cela est fait simplement par concaténation des vecteurs ; le plongement obtenu est donc dans $\mathbb{R}^{1000}$.

Toujours dans un soucis de dimensionnalité moindre, nous proposons aussi de tester une deuxième version de plus petite dimensionnalité. Le plongement précédent, obtenu par concaténation, est donc ramené par ACP d'une dimension 1 000 à une dimension de 500. Dans les expériences rapportées dans la section suivante, cette approche est appelée approche par fusion.

4 Expérimentations

Dans cette section, nous évaluons les performances des plongements proposés dans la section précédente. À des fins de comparaison avec les plongements CUI2vec (Beam *et al.*, 2020), largement employés, nous choisissons de tester des plongements obtenus par ACP avec la même dimensionnalité (500).

4.1 Jeux d'évaluation

Nous reprenons plusieurs jeux d'évaluation proposés dans la littérature. Nous les présentons brièvement ci-dessous, le lecteur intéressé peut se reporter à Beam *et al.* (2020) pour plus de détails.

Corrélation avec des jugements humains. Pakhomov *et al.* (2010) ont développé un jeu de données dans lequel des médecins ont indiqué leur perception de proximité entre 566 paires de concepts UMLS. Chaque paire de concepts a ainsi une mesure moyenne de la façon dont ils sont jugés similaires (*similarity*) ou liés (*relatedness*). Nous rapportons la corrélation ρ de Spearman entre ces deux scores d'évaluation humaine (noté ρ *sim.* et ρ *rel.*) et la similarité (cosinus) des plongements.

Types sémantiques. Les types sémantiques sont des méta-informations sur la catégorie à laquelle appartient un concept, et ces catégories sont organisées hiérarchie. Comme proposé par Beam *et al.* (2020), nous avons extrait le type sémantique le plus spécifique disponible pour chaque concept (à partir du fichier MRSTY fourni par UMLS). Nous évaluons la capacité de nos plongements à retrouver, dans les plus proches voisins (au sens du cosinus) d'un concept donné, des concepts partageant le même type sémantique. Cela est évalué avec des mesures de précision sur la liste ordonnée des plus proches voisins des concepts. Par exemple, le concept C0031740 (photochiomiothérapie) vu précédemment est de type *Therapeutic or Preventive Procedure*, tout comme C010009 (lutéolyse), C0043308 (radiothérapie X), etc.

Relations UMLS. Le dernier jeu d'évaluation exploite les relations encodées entre les concepts dans l'UMLS. Ces relations sont générales (e.g. *is-a*) ou spécifiques au domaine médical (e.g. *diagnoses*), la liste complète est disponible à `https://www.nlm.nih.gov/research/umls/META3_`

Plongements	ρ *sim.*	ρ *rel.*	types sémantiques P@5	types sémantiques P@10	relations UMLS P@5	relations UMLS P@10
De Vine *et al.* (2014)	0,455	0,423	0,3940	0,3751	0,1631	0,1275
Choi *et al.* (2016) (claims)	**0,552**	0,384	0,5784	0,5559	0,2444	0,1906
Beam *et al.* (2020)	0,522	0,430	0,5095	0.4781	0,2645	0,2069
kanjis (sec. 3.1)	0,296	0,317	0,6378	0,6117	0,3991	0,3110
ACP (sec. 3.2)	0,228	0,163	0,6213	0,6051	0.3557	0,2814
fusion (sec. 3.3)	0,538	**0,481**	0,6507	0,6138	**0,4265**	**0,3299**
fusion ACP (sec. 3.3)	0,525	0,474	**0,6518**	**0,6158**	0,4180	0.3238

TABLE 1 – Résultats des plongements proposés et de ceux de l'état-de-l'art sur les jeux d'évaluation. Les meilleurs résultats sont indiqués en gras.

`current_relations.html`. Nous évaluons donc la capacités de nos plongements à retrouver, dans les plus proches voisins (au sens du cosinus) d'un concept donné, des concepts liés par n'importe quelle relation. Comme précédemment, cela est mesuré par des précisions a différents seuils sur la liste ordonnée des voisins des concepts. Par exemple, le concept C0031740 (Photochimiothérapie) est en relation *is-a* avec C0087111 (Traitement), et en ralation *has-clinical-form* avec C0034172 (Photothérapie UVA), etc.

4.2 Résultats

Le tableau 1 recense les résultats des différents plongements de la littérature et ceux proposés en section 3. Nous rapportons également les résultats des plongements disponibles de l'état-de-l'art, à savoir les plongements de Choi *et al.* (2016) appris sur les données d'assurance (*claims*), de De Vine *et al.* (2014) et CUI2vec (Beam *et al.*, 2020).

Approches par kanjis. De ces résultats, on peut noter que les approches fondées sur les kanjis se comportent différemment des plongements de l'état-de-l'art. Sur les tâches de comparaison avec l'impression de proximité donnée par des médecins (deux premières colonnes), les approches par kanjis sont largement moins performantes que les autres. Il semble que pour cette tâche, les indices contextuels soient très importants. En revanche, sur les deux autres tâches, fondées sur des proximités encodées dans l'UMLS (types sémantiques ou relations sémantiques), les approches par kanjis apportent des gains, et plus particulièrement sur les relations UMLS. Concernant la différence entre la représentation creuse et celle dense obtenue par ACP, on observe une baisse très légère sur l'ensemble des tâche, du fait de la perte d'information dans le processus de réduction de dimension.

Intérêt de la fusion. Les plongements construits classiquement sur l'hypothèse distributionnelle et ceux construits sur la représentation par kanjis semblent avoir des performances complémentaires. Leur fusion, par concaténation suivie ou non d'une ACP, tire le meilleur parti de chacun et permettent donc d'obtenir de bons résultats sur l'ensemble des jeux d'évaluation. Comme précédemment, la réduction de dimension dégrade très peu les résultats.

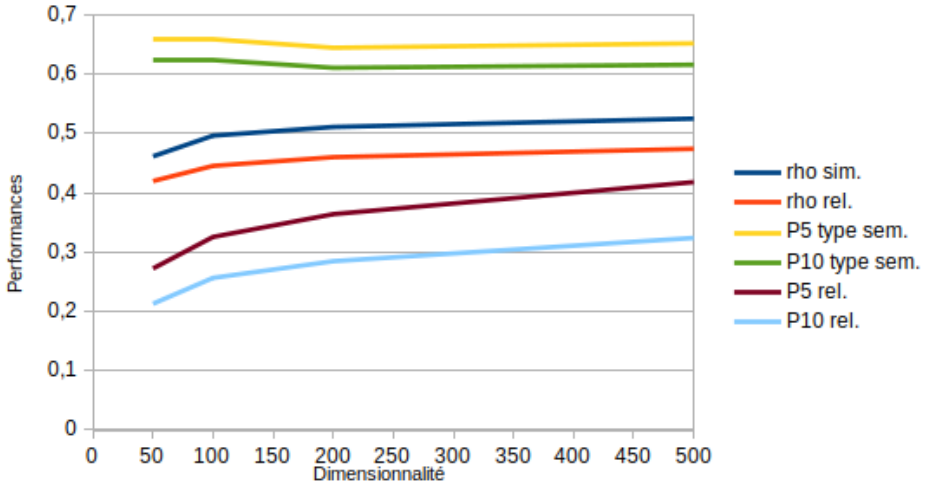

FIGURE 1 – Performances de la fusion de plongements selon la dimensionnalité obtenue par ACP.

Influence de la dimensionnalité. La figure 1 présente l'évolution des performances du plongement obtenu par ACP de la concaténation de cui2vec avec notre représentation ACP/kanjis. Comme attendu, on observe une diminution des performances lorsque le nombre de dimensions retenues à l'ACP finale diminue. Cette baisse de performances reste relativement faible, sauf pour l'évaluation via l'ensemble des relations UMLS. La variété des relations prises en compte dans cette évaluation semble difficile à concilier avec de plus faibles nombres de dimensions.

5 Conclusion et travaux futurs

L'approche que nous proposons est conceptuellement très simple, n'exploite que des données facilement accessibles (issues de l'UMLS) et ne nécessite que peu de temps de calcul. Elle offre pourtant de bons résultats par rapport aux approches existantes qui reposaient sur d'énormes quantités de textes médicaux. Les indices exploités (kanjis vs. contextes distributionnels) étant de nature différente, les performances sont variables selon les jeux d'évaluation. Sur la base de ce constat, nous avons aussi montré l'intérêt de les combiner au sein d'un plongement.

Les plongements existants sont dépendants de l'apparition des termes UMLS dans les corpus sur lesquels ils sont appris. Bien que jouissant d'une bonne couverture, comparable aux plongements de l'état-de-l'art, notre approche est quant à elle complètement dépendante de la disponibilité des termes japonais dans l'UMLS. L'anglais étant plus largement présent dans l'UMLS, des systèmes de traductions peuvent permettre de générer une ou plusieurs traductions candidates en kanjis et ainsi générer les plongements correspondants. Pour résoudre ce problème, nous sommes en train de développer une approche neuronale de génération de représentation kanjis. Les performances des plongements obtenus tendant à être sensiblement inférieurs à ceux obtenus directement des termes en kanjis, mais cela permet d'accroître considérablement la couverture à l'ensemble des CUI de l'UMLS.

D'autres indices, comme la structure même de l'UMLS (graphe de liens typés entre les concepts) peuvent également être utilisés en plus des liens de traductions. Ils peuvent notamment efficacement exploités par des approches de plongements de graphes ou de bases de connaissances (Wang *et al.*, 2018, *inter alia*). Mais cela pose alors la question de l'évaluation, puisque certains jeux d'évaluation

reposent actuellement sur ce graphe.

Récemment, le développement de plongements dynamiques tels que Bert (Devlin *et al.*, 2019) et leurs pendants dans le domaine biomédical (BioBert (Lee *et al.*, 2019) ou ClinicalBert (Alsentzer *et al.*, 2019)...) ont ouvert de nouvelles voies de recherche sur les plongements. Ils reposent cependant sur la même hypothèse distributionnelle et le même besoin de grandes quantités de textes. Par ailleurs, les jeux d'évaluation actuels ne sont pas très adaptés à leur fonctionnement (puisque l'évaluation des plongements est fait hors contexte dans les jeux de données utilisés dans cet article). Cependant, même pour ces approches, l'adjonction d'indices autres tels que nos représentations par kanjis pourraient potentiellement améliorer les représentations - uniquement contextuelles - apprises par ces approches.

Enfin, les différents plongements testés (incluant ceux de l'état de l'art), le code pour les générer (requiert un accès à l'UMLS), et le code pour les évaluer (requiert un accès à l'UMLS) sont disponibles pour la recherche sur demande auprès de l'auteur.

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JIN D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

ARONSON A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. In *Actes de AMIA 2001*, p. 17–21.

BEAM A. L., KOMPA B., SCHMALTZ A., FRIED I., WEBER G., PALMER N. P., SHI X., CAI T. & KOHANE I. S. (2020). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Proceedings of the Pacific Symposium on BioComputing*, p. 295–306, Hawaï, USA.

CHOI Y., CHIU C. Y.-I. & SONTAG D. A. (2016). Learning low-dimensional representations of medical concepts. In *Clinical Research Informatics* : AMIA.

CLAVEAU V. & KIJAK E. (2013). Analyse morphologique non supervisée en domaine biomédical. Application à la recherche d'information. *Traitement Automatique des Langues*, **54**(1), 13–45. HAL : hal-00912301.

DE VINE L., ZUCCON G., KOOPMAN B., SITBON L. & BRUZA P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, p. 1819–1822, New York, NY, USA : ACM. DOI : 10.1145/2661829.2661974.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**. DOI : 10.1093/bioinformatics/btz682.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU,

Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, p. 3111–3119.

PAKHOMOV S., MCINNES B., ADAM T., LIU Y., PEDERSEN T. & MELTON G. (2010). Semantic similarity and relatedness between clinical terms : An experimental study. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, **2010**, 572–576.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

ROBERTSON S. E., WALKER S. & HANCOCK-BEAULIEU M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199–210.

SARVENIAZI A. (2014). An actual survey of dimensionality reduction. *American Journal of Computational Mathematics*, **04**, 55–72. DOI : 10.4236/ajcm.2014.42006.

TUTTLE M., SHERERTZ D., OLSON N., ERLBAUM M., SPERZEL D., FULLER L. & NESLON S. (1990). Using meta-1 – the 1st version of the UMLS metathesaurus. In *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, p. 131–135, Washington, USA.

WANG Z., LV Q., LAN X. & ZHANG Y. (2018). Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 349–357, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1032.