Réduction de dimensionnalité des embeddings au moyen des ensembles approximatifs

Juan-Manuel Torres[1,†], Javier Ramírez-Rodríguez[1,2,†] and Martha-Lorena Avendaño-Garrido[1,3,*,†]

[1]*Laboratoire Informatique d'Avignon, Universitẽ d'Avignon, 84000 Avignon, France*
[2]*Departamento de Sistemas, Universidad Autónoma Metropolitana-Azcapotzalco, 02128, CDMX, México*
[3]*Facultad de Matemáticas, Universidad Veracruzana, 91097, Xalapa, México*

Abstract

Notre proposition cherche à réduire la dimensionnalité de la représentation des plongements de mots (embeddings) au moyen des ensembles approximatifs. Nous voulons garder les performances des algorithmes d'apprentissage profond sur une tâche classique de similarité sémantique entre les mots.

Keywords

Embeddings, Similitude sémantique, Réduction de dimensions, Ensembles approximatifs, GRASP

1. Introduction

Les plongements de mots (embeddings) sont des représentations denses des mots. Il s'agit d'un modèle utilisé dans le traitement automatique de langues (TAL) basé sur les réseaux neuronaux, les modèles probabilistes et la cooccurrence des mots. Les embeddings sont des vecteurs de dimension n avec des composantes dans $\mathbf{R}$; c'est-à-dire, les mots sont plongés dans un espace vectoriel continu à grande dimension. En outre, le vecteur représentant un mot i encode sa signification et sa distribution sémantique en fonction du contexte du mot. Les embeddings sont entraînés sur de grands corpus textuels, par exemple GloVe [1], Word2Vec [2], BERT [3] ou CAMEMBERT [4]. Par la suite, on a utilisé des techniques de post-traitement pour les embeddings pré-entraînés. Certaines de ces méthodes sont appelées techniques de réduction de dimensionnalité. Parmi elles on compte Analyse des Composants Principaux (PCA, pour son acronyme en anglais) [5] et t-Voisin Stochastique Distribué (t-SNE, anglais) [6]. En suivant cette idée, nous proposons de réduire la dimensionnalité des embeddings à l'aide de la théorie des ensembles approximatifs (Rough Sets) [7]. Un tel ensemble est une approximation d'un ensemble conventionnel en termes d'une paire d'ensembles appelée approximation inférieure et supérieure de l'ensemble en question. L'ensemble approximatif sera calculé au moyen d'une adaptation de la méthode Procédure de Recherche Adaptative Aléatoire Vorace (GRASP, anglais) [8]. Nous

Conférence en Recherche d'Information et Applications (CORIA), Avril 3-4, 2024, La Rochelle, France
*Corresponding author.
†These authors contributed equally.
juan-manuel.torres@univ-avignon.fr (J. Torres); jararo@azc.uam.mx (J. Ramírez-Rodríguez); maravendano@uv.mx (M. Avendaño-Garrido)
0000-0002-4392-1825 (J. Torres); 0000-0003-3986-1417 (J. Ramírez-Rodríguez); 0000-0001-7956-8958 (M. Avendaño-Garrido)

lançons donc l'hypothèse qu'une réduction pertinente des $n' < n$ dimensions ne degradera pas les performances des embeddings dans des tâches classiques de similarité sémantique de mots [5, 9].

2. Notre proposition

Les données sont codées dans une matrice où chaque ligne représente un objet et chaque colonne un attribut (une variable). Dans notre cas, l'ensemble U sont des embeddings de mots (lignes) avec $|U| = m$ et A est l'ensemble d'attributs avec $|A| = n$, soit la dimension des embeddings.

Avec cette information on construit la matrice de discernabilité D de dimension $\binom{m}{2} \times n$ ou ses éléments sont désignés par $d_{i,j}^k$, avec k l'indice de la colonne et le couple i, j l'indice de la ligne, correspondant à la discernabilité entre des embeddings i y j, c'est à dire

$$d_{i,j}^k = \begin{cases} 1 & si\ |u_i^k - u_j^k| > \epsilon, \forall i = 1, ...m-1, j = i+1, ..., m, k = 1, ..., n \\ 0 & \text{autrement} \end{cases} \tag{1}$$

avec u_i^k la valeur correspondant du embedding i dans sa *k-ième* dimension, et ϵ un réel positif proche de zéro. L'ensemble des vecteurs binaires de dimension n, noté V, sera l'espace de recherche. On definit la fonction d'aptitude (fitness) f, comme dans [10], pour les éléments de V qu'on veut maximiser

$$f(v) = \frac{n - L_v}{n} + \frac{C_v}{(m^2 - m)/2} \tag{2}$$

où L_v est le nombre de 1 dans v et C_v le nombre de combinaisons d'objets que v peut discerner. Le premier terme de (2) favorise la compacité de la solution candidate. Le second détermine dans quelle mesure cette solution peut discerner ces objets. f est calculée pour chaque voisine, ce qui permettra de trouver les meilleures candidates pour una solution réduite. Une telle solution doit pouvoir discerner parmi tous les objets de D.

Adaptation de GRASP

Nous avons adapté GRASP aux besoins de notre problème. À partir d'une solution candidate v initialement aléatoire telle que $L_v \leq \frac{n}{2} - 1$. Ceci réduit au maximum à la moitié le nombre de dimensions, ce qui permettra de se comparer avec des méthodes état de l'art [5]. On génère ainsi un échantillon aléatoire de solutions voisines, obtenues en remplaçant les 1 par des 0 et vice versa. Dans notre cas, la taille de l'échantillon sera $l < n$, un nombre approprié qui permettra réduire le coût de calculs. Avec cet ensemble on produit une liste restreinte de candidates (LRC) présentant les meilleures valeurs de f. Pour construire LCR, une borne d'inclusion b est définie:

$$b = c^* - \alpha(c^\star - c_*) \tag{3}$$

où c^* et c_*, sont les valeurs maximale et minimale de f, des solutions candidates, et $0 \leq \alpha \leq 1$ est une valeur réelle. Il est à noter que si $\alpha = 0$, on aurait un algorithme glouton, et si $\alpha = 1$, un algorithme completement aléatoire. Nous allons garder les candidates dont $f > b$. Une solution est choisie au hasard parmi les solutions de LRC. Avec cette nouvelle candidate, un nouveau

echantillonage du voisinage est généré. Ceci est un processus itératif de recherche locale, où le voisinage d'une solution sera exploré à plusieurs reprises à la recherche d'un maximum local. Si la solution n'est pas améliorée dans des voisinages successifs, on change de solution initiale.

Application

Nous savons qu'en réduisant la dimension des embeddings, les algorithmes utilisés pour calculer la similarité sémantique seront plus efficaces pour effectuer la tâche [9]. En particulier, nous voulons explorer jusqu'où peut on aller dans cette réduction dimensionnelle, sans perte notable de performances. [5] rapporte des performances égales voire supérieures de similarité en utilisant des méthodes de réduction (300 vers 150 dimensions) basées sur PCA, ce qui est assez impresionnant. Or, à notre connaissance, les ensembles approximatifs n'ont pas été utilisés dans cette tâche de similarité. Nous allons mesurer empiriquement les complexités en temps et en espace, qui risquent d'être différentes d'une réduction PCA. La finalité de cette recherche est de répondre à la tâche de similarité sémantique des mots dans un espace dimensionnel réduit ($n' < 150$) des embeddings. Cette réduction peut être appliquée aussi à d'autres tâches TAL, telles que le résumé automatique [11] ou la recherche d'informations sur des données massives [12].

References

[1] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Empirical Methods in Natural Language Processing, ACL, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013, p. 3111–3119.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1, ACL, Minneapolis, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[4] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 7203–7219. URL: https://www.aclweb.org/anthology/2020.acl-main.645.

[5] V. Raunak, V. Gupta, F. Metze, Effective dimensionality reduction for word embeddings, in: 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), ACL, Florence, Italy, 2019, pp. 235–243. doi:10.18653/v1/W19-4328.

[6] V. Warmerdam, T. Kober, R. Tatman, Going beyond T-SNE: Exposing whatlies in text embeddings, in: E. L. Park, M. Hagiwara, D. Milajevs, N. F. Liu, G. Chauhan, L. Tan (Eds.), Second Workshop for NLP Open Source Software (NLP-OSS), Association for Computational Linguistics, Online, 2020, pp. 52–60. doi:10.18653/v1/2020.nlposs-1.8.

[7] Z. Pawlak, Rough sets, International Journal of computer Information Sciences 11 (1982) 341–356. doi:10.1007/BF01001956.
[8] M. G. C. Resende, C. C. Ribeiro, Greedy Randomized Adaptive Search Procedures: Advances and Extensions, Springer International Publishing, Cham, 2019, pp. 169–220. doi:10.1007/978-3-319-91086-4_6.
[9] E. Linhares Pontes, S. Huet, A. C. Linhares, J.-M. Torres-Moreno, Predicting the semantic textual similarity with Siamese CNN and LSTM, in: TALN Vol 1, ATALA, Rennes, France, 2018, pp. 311–320. URL: https://aclanthology.org/2018.jeptalnrecital-court.13.
[10] A. T. Bjorvand, J. Komorowski, Practical applications of genetic algorithms for efficient reduct computation, Wissenschaft & Technik Verlag 4 (1997) 601–606.
[11] J.-M. Torres-Moreno, Automatic Text Summarization, Wiley, London, 2014.
[12] M.-R. Amini, G. Eric, Recherche d'Information - applications, modèles et algorithmes, Algorithmes, Eyrolles, 2013. URL: https://hal.science/hal-00881257, i-XIX, 1-233.