Connecting Words and Linked Data Concepts by Latent Features

Márius Šajgalík[1], Michal Barla[1], Mária Bieliková[1], and Julian Szymański[2]

[1] Faculty of Informatics and Information Technologies
Slovak University of Technology, Ilkovičova 2, 842 16 Bratislava 4
{marius.sajgalik,michal.barla,maria.bielikova}@stuba.sk
[2] Department of Computer Architecture
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
julian.szymanski@eti.pg.gda.pl

Few decades ago, we did not have efficient methods to automatically process raw unstructured data, but we were good at handling small graphs and creating rule-based methods to perform logical inference on them. Thus, linked data was born to create a structured representation of (ideally) all the information hidden in the raw texts. Despite the extensive manual labour needed to create it, we are now flooded with vast number of linked data sources on the Web. What started as small graphs [9] is now regarded as a problem of big data. Management of diverse linked data sources becomes a problem, since we often need to cope with low quality, duplicate, incomplete or even contradictory data [7, 4].

Recently, unsupervised learning of word features has come into attention [3, 8]. It is based on distributional hypothesis [6], which states that features of a word can be learned from its context. [5] argues that there is not always a single meaning of a word, though placed in a context. Distributed representation allows a single word to aggregate multiple senses or express multiple concepts, however, it is difficult to manually design all the features.

That brings us to our vision. Imagine one unified model that understands raw text on the fly and automatically infers the mentions and descriptions of the respective linked data concepts. Imagine a model that can also learn new unseen concepts, which are being created right now. Imagine a model, which can understand relations between concepts and automatically infers new ones. Imagine a model, which can generate descriptions of words, concepts and relations on request. Such unified model would enable much more efficient processing of new unannotated texts (an example of its potential application can be found in [10]) and management of big linked data sources. Moreover, it turns out we are not so far away from fulfilling this vision.

There are already several approaches that combine unsupervised learning of latent features with linked data to learn a joint model of words and concepts. In [1] we can see one of the first attempts to train such model, which learns latent features of words and relations. They also demonstrate the possibility of extracting additional knowledge facts from raw text. Approach in [11] presents a new neural tensor network model, which models feature vectors of an entity as an average over its word vectors and models each relation by a tensor. The authors

demonstrate power of their model on classification of unseen relationships. Approach in [13] learns joint model of words from unstructured text, concepts from linked data and relations between those concepts. In this case words, concepts and relations are all embedded into the same joint latent feature space.

Yet still, these approaches fail to fulfil our described vision entirely. All of them are just static models, which operate on word level. Thus, they are limited to static non-expandable vocabulary. However, we can go further and create a character-wise temporal model. Such model would be limited only by the feature space, not the vocabulary. It could also easily handle misspellings and learn the syntactical features (like those hidden in common prefixes or suffixes) much more easily. A temporal model could also generate the envisioned descriptions. There is a lot of recent approaches for different tasks that are successful at modelling such sequences of data (e.g. use of RNN models in [2, 12]). Having a joint model of words and concepts would greatly enhance human interaction in natural language with vast amount of information contained in linked data.

References

1. A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *AAAI Conference on Artificial Intelligence*, 2011.
2. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
3. R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of ICML*, pages 160–167. ACM, 2008.
4. K. Draszawka, J. Szymanski, and H. Krawczyk. Towards increasing density of relations in category graphs. In *Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation*, pages 51–60. 2014.
5. K. Erk. What is word meaning, really?: (and how can distributional models help us describe it?). In *Proc. of Workshop on GEMS*, pages 17–26. ACL, 2010.
6. Z. S. Harris. Distributional structure. *Word*, 1954.
7. M. Holub, O. Proksa, and M. Bieliková. Detecting identical entities in the semantic web data. In *SOFSEM 2015: Theory and Practice of Computer Science*, volume 8939 of *LNCS*, pages 519–530. Springer, 2015.
8. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proc. of Workshop at ICLR*, 2013.
9. G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.
10. M. Šajgalík, M. Barla, and M. Bieliková. Exploring multidimensional continuous feature space to extract relevant words. In *SLSP*, pages 159–170. Springer, 2014.
11. R. Socher, D. Chen, C. D. Manning, and A. Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, pages 926–934, 2013.
12. K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.
13. J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973*, 2013.