Entity Modelling through Ontologies and Large Language Models

Fabio Antonio Yáñez-Romero

Department of Software and Computing Systems, University of Alicante, Spain

Abstract

The aim of this paper is to present a line of research focused on improving the knowledge represented in natural language processing tasks through the use of ontologies, combining these with machine learning techniques. It is expected that with this kind of techniques it will be possible to fight against phenomena such as the hallucination present in current generative language models and to reach the state of the art in different tasks taking into account semantic knowledge. Initially, we will try to solve the problem of semantics in specific areas such as medicine, where the external knowledge that can be incorporated would help to provide knowledge that does not exist in unstructured data such as all ICD-10 codes. Therefore, we expect obtain enough conclusions to apply this methodology with other dominions.

Keywords

Embeddings, Generative Language Models, Graph Neural Networks, Knowledge Bases, Knowledge Graphs, Natural Language Processing, Ontologies, Semantics

1. Research Justification

Transformer-based models have reached the state of the art in many natural language processing (NLP) tasks. These models use an encoder-decoder architecture and add a self attention mechanism that allows a probabilistic model to be generated based on a large training corpus while retaining important information from large amounts of text [1]. The trend in recent years to improve the state of the art is to create models with this architecture using more layers, having more parameters to train and using larger corpora, as can be seen in cases such as Generative Pretrained Transformers (GPT). [2]. Although these models reach the state of the art, they are only probabilistic models that determine the response based on occurrences within the training text, are not capable of reasoning and do not have common or domain-specific knowledge, which leads to problems when certain knowledge is required to be taken into account in NLP tasks, representing syntax well but failing from a semantic point of view. This problem motivates research into methods for improving language models by incorporating knowledge from different sources and even employing alternative architectures that take semantic knowledge into account from their conception. The external knowledge used comes from two main sources:

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain

✉ fabio.yanez@ua.es (F. A. Yáñez-Romero)

🌐 https://www.linkedin.com/in/fabio-yañez-romero-0b00b9169/ (F. A. Yáñez-Romero)

iD 0009-0005-5494-0166 (F. A. Yáñez-Romero)

1. Corpus that implicitly include the knowledge to be transmitted. In this case the knowledge needed is not directly accessible, but it has the advantage of not requiring to prepare the knowledge in a suitable format. Sometimes the information coming from corpora is processed by creating a data structure typical of a knowledge base ([3]).

2. Knowledge bases that represent this knowledge explicitly, through databases. This knowledge is usually represented through graphs, allowing a wide variety of heterogeneous relationships between entities to be captured, and also great flexibility in representing this knowledge. The preparation of the knowledge structure can be time-consuming and computationally expensive. An incorrect representation of the knowledge means dragging an error throughout the learning stage of machine learning models.

In addition to adding external knowledge to improve the efficiency of language models, new architectures are also created taking this type of knowledge into account from their conception [4].

In specific areas such as medicine or economics the use of pre-trained models together with transfer learning allows high success rates, however, there are specific ontologies for these areas and common knowledge ontologies that can be used to improve the state of the art [5]. NLP tasks that do not use machine learning achieve much lower accuracy, which leads to the combined use of ontologies with deep language models and other models such as graph neural networks, designed to act directly on the structure of a graph.

The aim of this project is the modelling of digital entities, allowing them to be processed by machine learning algorithms, giving the models used semantic knowledge. The architecture used in ontologies will provide the necessary richness to correctly represent these entities. The integration with artificial intelligence models is the key point to be addressed.

Our intention is to use existing ontologies such as UMLS in the medical field or BabelNet for common knowledge and try to integrate it with neural networks used in natural language processing. If the interpretation of digital entities can be improved, it is hoped that this will improve the state of the art and bring semantic interpretation to the models used.

It is worth mentioning that this thesis is at an early stage, so many modifications are expected when it comes to tackling the problem of modelling digital entities, both in the way these entities are represented and in their use with artificial intelligence models.

2. Related Work

Representing the knowledge in ontologies (mainly in the form of graphs) as embedding vectors has been a line of research widely covered in the last decade. These techniques focus on creating embedding vectors that represent the nodes and edges present in this knowledge base and allow the inference of new relationships through these models. The trend of the latest models is to represent complex spaces that allow to cover the existing relationships between nodes with a lower computational cost, while improving the success rate when representing knowledge in different benchmarks such as FB15K-237, WN18RR, YAGO3-10 [6]. The techniques used to cover the representation of knowledge in the form of embedding are very varied, with translational models [7] , tensor factorisation models [8], other deep learning models [9] and rule-based models [10] standing out. The advantage of these models is that they allow knowledge to be added to machine learning models based on knowledge graphs, which are extremely expressive data structures.

Other studies try to represent knowledge from embedding vectors obtained directly from a corpus, this is the case of the first embedding models for text such as Word2Vec [11] and GloVe [12], being widely used by the scientific community today. In this case, the advantage of these models is that they provide vectors that represent the interactions between words within a specific document, without the need to start from a structured knowledge base.

Another approach is to represent text entities using the latent knowledge of pre-trained language models such as BERT to obtain vectors in the same space as the language model.

The existing knowledge in the text can be extracted by other procedures such as the identification of entities, determining the lexical or semantic dependence of each sentence, the causal relationships between different sentences, etc [13].

In other cases, the vectors that include the knowledge to be represented come from an ensemble of different models, being able to represent the knowledge from ontologies and extracting the information from the text [14].

The techniques that use knowledge in machine learning models do not only consist of using embedding vectors with this knowledge included, there is another branch of research focused on training deep learning models taking into account the existing knowledge by designing the loss function for this purpose [15].

Some of the architectures try to incorporate knowledge from the training phase of the language model, in the case of [16] a BERT-type architecture is used where whole entities and phrases are masked as well as randomly masked words in order for the model to learn the existing relationships between the different elements. [4] modifies the BERT architecture, using another encoder where an embedding generated by TransE [7] is processed for each entity of the ontology used, representing the entity of a knowledge graph, this encoder is used after the characteristic encoder of BERT, which processes the tokens of each sentence.

3. Hypothesis and Objectives

The hypothesis behind this line of research is that language models that only take into account the probability that a series of words appear in sequence, do not correctly capture semantic knowledge, and that semantic knowledge must be provided through well-structured knowledge that represents the semantic relationships between different entities, i.e. modelled entities.

The aim of this research is to take advantage of the knowledge of specific ontologies to increase the accuracy of language models, and provide more semantic knowledge to models that reach the state of the art, thus avoiding mistakes in different natural language processing tasks. Among the different tasks involved in this general objective we can consider the following:

1. Understanding the different ontologies to be used and the semantic relationships between the entities represented in these ontologies.

2. Extraction of the existing knowledge in these ontologies directly through embedding vectors or indirectly by representing this knowledge with logical rules during the training

of the model.

3. Experiment with deep neural network architectures in order to adapt semantic knowledge extracted from specific ontologies.

4. Document a general methodology for incorporating ontological knowledge into language models.

The use of external knowledge on language models will not only improve the results obtained from the point of view of knowledge and accuracy. It allows new lines of research where is not necessary to retrain the models from scratch, this being one of the current problems in virtual assistants based on GPT-3.

Having up-to-date knowledge at all times in this kind of models is expected to be useful to perform more complex tasks within NLP such as Fact Checking through different sources.

4. Methodology

Among the different techniques being considered to achieve the proposed objectives are the following:

- Integration of ontologies with language models, using embedding vectors obtained with graph neural networks or other machine learning techniques that manage to correctly represent the knowledge of the ontologies in the vector space of the language models.

- Modification and creation of language model architectures that allow the incorporation of semantic knowledge, being able to represent relations of hierarchy, antonymy, synonymy, etc. that enrich the model semantically.

- Use and integration of latent logic rules in the sources used to train language models to capture semantic knowledge (lexical and semantic trees, causality relations, etc).

For the first step, inserting domain-specific knowledge into generative language models, we are working with The Unified Medical Language System (UMLS) [5] as an ontology for the different biomedical terms and we intend to use a language model specific to the biomedical field such as BioBERT [17]. We will start with simple NLP tasks such as named entity recognition, word sense disambiguation or semantic role labelling. The aim is to build a knowledge graph suitable for the task based on information from UMLS.

The next step will be to test the transmission of semantic knowledge to generative language models with more complex tasks that require encoder-decoder architecture such as text generation.

Finally, it is proposed to carry out this task with common knowledge and larger models such as RoBERTa, making the appropriate modifications to try to represent the knowledge well. In this case, the use of common knowledge ontologies such as BabelNet is proposed [18].

The use of graph neural networks to process large graphs is discarded due to the high computational cost involved. This problem could be solved by limiting the knowledge graphs used to small instances for each particular step and relying on the latent knowledge of language models. [19].

5. Conclusions and Future Work

This publication indicates the research framework in which my thesis will be developed after studying the state of the art in the use of knowledge for the improvement of different tasks within the field of natural language processing.

The first objective will be to improve knowledge representation and state of the art in specific fields of language processing such as Biomedicine, where providing this knowledge requires less computational resources, and subsequently extrapolate these results to common knowledge, using ontologies and larger models.

The semantic improvement in language models is expected to allow better automation of tasks involving the use of natural language processing, as well as better use for knowledge inference and classification. A language model that correctly interprets semantics will allow the creation of virtual assistants that provide truthful, logical and less biased information derived from solely probabilistic models.

As future work, we hope that the results of this research will serve to contrast the veracity of different texts based on semantic knowledge in different information sources, being able to carry out a Fact Checking task to combat misinformation in different media.

Acknowledgments

The research developed in this paper has been funded by the PROMETEO project promoted by the *Generalitat Valenciana*, specifically the NL4DISMIS project, which aims to combat misinformation and disinformation in different media. Also, thanks to the Department of Language Processing and Information Systems, specifically the Natural Language Processing Team (GPLSI) for their support and advice during the development of the PhD.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[3] F. N. Al-Aswadi, H. Y. Chan, K. H. Gan, Automatic ontology construction from text: A review from shallow to deep learning trend, Artif. Intell. Rev. 53 (2020) 3901–3928. URL: https://doi.org/10.1007/s10462-019-09782-9. doi:10.1007/s10462-019-09782-9.

[4] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities, 2019. arXiv:1905.07129.

[5] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology., Nucleic Acids Res. 32 (2004) 267–270. URL: http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04.

[6] Z. Cao, Q. Xu, Z. Yang, X. Cao, Q. Huang, Geometry interaction knowledge graph embeddings, 2022. arXiv:2206.12418.

[7] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[8] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, Madison, WI, USA, 2011, p. 809–816.

[9] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, D. Phung, A novel embedding model for knowledge base completion based on convolutional neural network, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 327–333. URL: https://aclanthology.org/N18-2053. doi:10.18653/v1/N18-2053.

[10] M. Qu, J. Tang, Probabilistic logic neural networks for reasoning, 2019. arXiv:1906.08495.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.

[12] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[13] X. Liu, X. You, X. Zhang, J. Wu, P. Lv, Tensor graph convolutional networks for text classification, 2020. arXiv:2001.05313.

[14] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, M. Zhou, Compare to the knowledge: Graph neural fake news detection with external knowledge, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 754–763.

[15] T. Goodwin, D. Demner-Fushman, Enhancing question answering by injecting ontological knowledge through regularization, in: Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, Online, 2020, pp. 56–63. URL: https://aclanthology.org/2020.deelio-1.7. doi:10.18653/v1/2020.deelio-1.7.

[16] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, H. Wu, Ernie: Enhanced representation through knowledge integration, 2019. arXiv:1904.09223.

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtz682. doi:10.1093/bioinformatics/btz682.

[18] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, F. Cecconi, Ten years of babelnet: A survey, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence

Organization, 2021, pp. 4559–4567. URL: https://doi.org/10.24963/ijcai.2021/620. doi:10.24963/ijcai.2021/620, survey Track.

[19] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, 2021. arXiv:1812.08434.