

Complementing Language Embeddings with Knowledge Bases for Specific Domains

Paolo Tenti, Gabriella Pasi and Rafael Peñaloza

IKR3 Lab, University of Milano-Bicocca, Milan, Italy

Abstract

Language embeddings are a promising approach for handling natural language expressions. Current embeddings encompass a large language corpus, and need to be retrained to deal with specific sub-domains. On the other hand, these embeddings often disregard even basic domain knowledge, making them specially fragile when handling technical, specific, knowledge domains, and requiring costly re-training. To alleviate this issue, we propose a combined approach where the embedding is seen as a model of a logical knowledge base. Through a continuous learning approach, the embedding improves its satisfaction of the knowledge base, and in turn produces better training examples by labelling previously unseen text. In this position paper we describe the general framework for this continuous learning, along with its main features.

Keywords

Language embedding, Knowledge Bases, Natural Language Understanding, Neuro-Symbolic Learning

1. Introduction

Natural Language Understanding (NLU) is the mechanical act of understanding language expressions, which is pivotal to many text related applications (e.g., text classification, information retrieval, question-answering). These applications require features that faithfully represent text meaning, to use them in relevant algorithms.

Language embeddings (LE) [1, 2, 3] are dense representations of textual expressions, that capture their distributional semantics by *pre-training* a language model over large corpora of general language (e.g., Wikipedia). Their pre-trained nature allows to conveniently use LE in many down-stream tasks as representations of language expressions (known as *transfer learning*). However, pre-trained LE are challenged by domain-specific language in several applications. There are three main reasons for this.

First, LE do not capture domain-specific language, and require *re-training* over domain-specific corpora of unstructured text. However, re-training is computationally expensive, and the available datasets are not always large enough to effectively re-train LE. Second, LE capture the sense of words from their context, as distributional semantics. However, many domain-specific tasks require a more precise understanding of text. To cope with these problems one

DAO-XAI 2021

✉ p.tenti1@campus.unimib.it (P. Tenti); gabriella.pasi@unimib.it (G. Pasi); rafael.penalozan@unimib.it (R. Peñaloza)

🌐 <https://ikr3.disco.unimib.it/people/paolo-tenti/> (P. Tenti); <https://ikr3.disco.unimib.it/people/gabriella-pasi/>

(G. Pasi); <https://rpenalozan.github.io/> (R. Peñaloza)

🆔 0000-0002-9421-8566 (P. Tenti); 0000-0002-6080-8170 (G. Pasi); 0000-0002-2693-5790 (R. Peñaloza)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

can take advantage of *structured information*, such as key phrases or ontological categories. Third, even general domain applications often require to represent larger fragments of text (e.g., sentences, paragraphs). In such cases NLU techniques need to capture deep, semantic rooted understanding of text structures that are more complex than simple bags of tokens. Although several methods have been proposed in the literature, they are often task specific and increase design complexity of down-stream models in practice.

To mitigate these challenges, we propose to **complement LE with Knowledge Bases (KB)**—that is, formal representations of knowledge. Specifically, we aim to improve language understanding by (i) acting on the LE’s ability to represent domain-specific language expressions, and (ii) linking KB symbols (i.e., entities and relations) to language expressions. Intuitively, we consider language expressions as instances of KB assertions (*interpretations*), hence defining KB embeddings (KBE) (i.e., dense representations of KB symbols) as a function of LE. We jointly train both from supervised data to maximise logical satisfaction of the KB. We argue that this approach helps to address the LE challenges highlighted above.

First, we propose to use the KB to extract a domain-specific supervised dataset to *pre-train* our model. By doing so we decouple the problem of re-training LE over domain-specific data from task-specific data-sets that, as said, could not be large enough. Second, we propose to learn representations of text fragments longer than a token by means of a knowledge-aware task mediated by the KB, that is learning a representation of KB symbols (i.e., KBE). We argue that such representations are more useful to down-stream tasks than the ones obtained with the next sentence prediction task, which mainly capture syntactic properties rather than semantics. Third, we propose to use the combined knowledge and language embeddings (i.e., KBE and LE) to associate KB symbols to language expressions, and use them as features in down-stream models. We argue that representing language expressions with symbols, in addition to LE, is a step forward to precisely characterising their meaning compared to distributional semantics alone.

In addition, note that by using KB symbols as features we improve interpretability of down-stream models. Moreover, as we will explain in more details, interpretability enables a **continuous learning** framework to mutually improve LE and KBE.

2. Related work

Complementing KB and LE is not new. Some works [4, 5, 6] focus on the LE’s challenge to represent domain-specific language, and use a knowledge-aware tasks (mediated by a KB) to re-train LE; while [7] specifically focuses on continuous learning. The problem of complementing Knowledge Graph (KG) embeddings with textual information, such as names and descriptions of KG entity and relations, to improve the KG Completion task is studied in [8, 9, 10]. However, none of those works address the problem of complementing domain-specific LE with the extraction of symbolic features from language expressions for NLU.

Petroni et al. study the **ability of word embeddings to capture relational knowledge**, similarly to what a KB does, by focusing on general language. They highlight that word embeddings can capture lightweight KB capabilities. From this perspective, [12] proposes encoding relational knowledge in a separate word embedding learned from co-occurrence

statistics, complementary to a given standard word embedding. The analysis presented by the authors shows that relational word vectors do indeed capture information that is complementary to what is encoded in standard word embeddings. We argue that formal representations of knowledge are not matched by distributional semantics out of the box.

Information Extraction (IE) aims to extract structured information from unstructured text. Most work focuses on unsupervised methods, to face the challenges of compiling supervised datasets and obtaining a KB upfront [13, 14]. Open Information Extraction (OIE) [15, 16, 17] extracts relational facts from unstructured text as surface patterns (i.e., spans of pure unstructured text), without linking them to an existing KB. We argue that using KBs to formally describe domain knowledge is beneficial to enforce control over the IE process. In fact, KBs can assist in the compilation of supervision, simplify the evaluation process of IE results by letting to focus on fewer, well-known symbols rather than the more widespread surface patterns, and improve explainability of down-stream tasks.

Several tasks focus on extracting KB resources from unstructured text; e.g., Named Entity Recognition (NER), Named Entity Linking (NEL), and Relation extraction (RE). Traditional approaches use extraction pipelines that treat NER, RE and NEL as separate tasks, suffering from error propagation and ignoring synergies between sub-tasks. In addition, these methods depend heavily on complex features. Thus, recent works focus on building joint, neural models [18]. These models are either task-purposed (i.e., they only focus on entities [19] or relations [20, 18]) or domain-specific [21, 22]. We are interested, instead, on the more general problem of modelling synergies between language expressions and KBs, to extract complete relational facts from language expressions for any domain, similarly to KG completion [23].

Providing dense representations of KG resources (subjects, objects and relations) has been widely considered [24, 25]. In such models, **KG embeddings** (KGE) are learnt from relational facts, to optimise a predetermined embedding function. However, domain-specific background knowledge is usually formalised through hierarchies, taxonomies and logical rules, which are typical of KBs rather than KGs—the latter store large collections of relational facts instead. Gutiérrez-Basulto and Schockaert [26] showed that KGE models hardly capture even the most basic logical properties of KBs, and propose to represent KB resources through convex regions in a semantic space, which is seen as an interpretation of the KB. In addition, they describe how to keep the embedding model open to external resources; e.g., language expressions represented by embeddings. Kulmanov et al. [27] apply a similar KB embedding model to a domain-specific task for KB completion. Although [26, 27] are related to our study, we focus on NLU.

Similar approaches combining logic and real world objects represented by embeddings were studied in [28, 29]. These differ by the methods used to enforce logical consistency (i.e., fuzzy logic or probability) in contrast to the geometric properties by [26, 27]. In addition, [28] have not been fully studied to interpret language expressions for NLU [30, 31].

3. Model Description

Our main goal is to infer KB symbols (i.e., entities and relations) from *surface patterns*, that is, text spans of arbitrary length from unstructured text. As a simple example, consider the sentence *The city of lights has been the capital of France for many centuries*, where *city of lights*

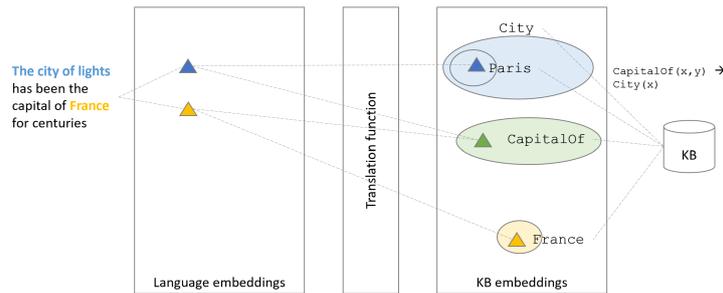


Figure 1: An intuitive representation of the model for a positive case: parameters of a function to translate the language embeddings semantic space and regions representing KB symbols are optimized to satisfy the KB from language expressions labeled with KB symbols.

and *France* are surface patterns that should be meant as *Paris* and *France* respectively, and the sentence as $\text{CapitalOf}(\text{Paris}, \text{France})$.

We consider surface patterns as possible *interpretations* of KB symbols. Recall that a KB is a partial representation of the world, which usually introduces restrictions on the possible meanings of the symbols it uses. Hence, KB semantics is typically defined by means of interpretations. In essence, an interpretation describes all the instances of interest and their relationship within all the properties expressed in the KB. Slightly more formally, an interpretation consists of an *interpretation domain*, which describes the objects in the world, and an *interpretation function*, which describes the meaning of each symbol within this world. This interpretation is a *model of the KB* if it satisfies all the constraints imposed by the KB [26, 32].

We consider (the set of representations of) surface patterns as an interpretation domain, and aim to learn from data a suitable interpretation function guaranteeing that the resulting interpretation is in fact a model. To achieve this, we propose to (see Figure 1):

- encode surface patterns in a language semantic space, using a pre-trained LE model;
- encode KB symbols as regular regions in a KB semantic space as in [26]. Specifically, relations are interpreted as convex regions;
- build an interpretation function bridging both semantic spaces.

We use a supervised dataset to *jointly* train the parameters of the model, using a loss based on the violation of the KB constraints. The supervised dataset labels the unstructured text fragments with markers of surface patterns and their relative symbols in the KB. The pre-trained model can be used to obtain domain-specific language embeddings, and to infer KB symbols over natural language expressions by using regular regions. In addition, regular regions can be used as KBE in down-stream tasks (e.g., KB Completion).

Importantly, to deal with the problem of polyonymity, notice that entities are also considered (unary) relations, and they are represented as regular regions. In fact entities in the KB are singleton symbols (e.g., *Paris*) but they are representative of potentially many different surface patterns (e.g., the city of lights, *Paris*, the capital of France). We also emphasise the restriction to regular regions for interpreting relations. There are three main reasons for this choice. First, regular regions allow for better interpretability of the representations; second, the results allow



Figure 2: A framework for continuous learning

generalizability by avoiding over-fitting through complex regions; and third, regular regions can be succinctly described through a few parameters.

Indeed, learning regular regions over the original language embedding space would be desirable, because the interpretation function would be reduced to the identity function. This would be possible for entities, as LE guarantee that similar entities lay close in the semantic space. However, it might be not possible for relations of arity greater than 1. For one, relations with the same domain and range would have overlapping regions. In addition, relations with a wide domain or range would have very large regions. In both cases we would lose representativity. Thus, we need either to increase the dimensions of the embeddings ($n - 1$ dimensions on the number of KB symbols are needed [26], leading to a sparse representation space) or to use a non-linear, relation-specific transformation to encode inputs.

4. Continuous Learning

To train the proposed model from supervised datasets and to make use of pre-existing KBs are certainly two strong assumptions, as they might be expensive to obtain. Still, domain-specific applications exist where highly-qualified, labour-intensive, error-prone human interventions are usually employed. Two examples are offered by the manual screening of scientific publications to be included in literature reviews, and by manual labelling of unstructured text. Such human activities could be shifted to higher level interventions, such as maintaining KBs and supervised datasets; keeping a degree of control over the inference process through interpretable models is desirable in such scenarios, when compared to completely unsupervised approaches.

We propose a continuous learning framework, which iteratively refines the supervised dataset and the knowledge base. This framework, depicted in Figure 2, is organised into the following steps:

- background knowledge is formalised through a KB containing relational data (assertions) and Datalog rules (axioms);
- assertions from the KB are used to extract a distantly supervised dataset from a domain-specific corpus of unstructured text fragments;
- the supervised dataset is used to re-train the model, and the model is used to comprehend new text fragments by inferring KB symbols;
- inferred KB symbols can be analysed by humans, and used to maintain the supervised dataset and the KB.

We use **distance supervision** to select sentences from unstructured text corpora that match named entities from KB assertions. A known challenge of this approach is to discriminate if

matching sentences have a meaning which is coherent to the assertion under scrutiny. Observe that compiling a good dataset for supervised learning is more related to precision than recall: capturing all possible good sentences is less desirable than capturing a few high quality sentences representing the assertions.

In our view, this distance supervision problem can be successfully addressed by considering it as a search problem: we view any given assertion as a query made over a corpus of (natural language) sentences. [33] suggest that re-ranking models (i.e., BM25+CE [33], ColBERT [34]) works well in combination with pre-trained LE, showing good generalization capabilities over unseen datasets and domains.

5. Conclusions and Future Work

We propose a framework to learn from supervised data a model aimed to align language embeddings and KB representations; such a framework can be useful in two ways. First, we obtain domain-relevant, knowledge-aware language embeddings by continuously re-training them with a KB mediated task; second, we obtain KB embeddings that provide a model of the KB over language expressions, which can be used to infer KB symbols (i.e., relations and entities) over language expressions. LE and KB symbols can be used in domain-specific downstream applications as features. This model can improve the effectiveness of natural language understanding methods in domain-specific applications and, by using KB symbols as features, improve interpretability. We also propose distant supervision to compile a dataset for training, and to use text ranking techniques to improve precision.

One potential application field is in the area of literature reviews, where all publications related to a specific topic need to be analysed. In this case, our methods automatically find and recommend scientific publications that match the topic of interest, among the huge amount of existing publications. Importantly, current literature reviews require extensive interventions from highly-qualified human experts to discern whether a publication is indeed related to the topic studied, and also to evaluate its importance and relevance.

As future work, we plan to implement such a model and test its performance in potential down-stream tasks.

References

- [1] J. Devlin, M.-W. Chang, K. Lee, K. N. Toutanova, 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), 2018, pp. 4171–4186.
- [2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018. URL: <https://openai.com/blog/language-unsupervised/>.
- [3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, 2018, pp. 2227–2237.

- [4] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities., in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.
- [5] B. He, D. Zhou, J. Xiao, X. jiang, Q. Liu, N. J. Yuan, T. Xu, Integrating graph contextualized knowledge into pre-trained language models, arXiv preprint arXiv:1912.00147 (2019).
- [6] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-bert: Enabling language representation with knowledge graph, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2901–2908.
- [7] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, H. Wang, Ernie 2.0: A continual pre-training framework for language understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8968–8975.
- [8] H. Xiao, M. Huang, X. Zhu, Ssp: Semantic space projection for knowledge graph embedding with text descriptions., in: AAAI, 2016, pp. 3104–3110.
- [9] D. Nozza, E. Fersini, E. Messina, Cage: Constrained deep attributed graph embedding, *Information Sciences* 518 (2020) 56–70.
- [10] H. Zhong, J. Zhang, Z. Wang, H. Wan, Z. Chen, Aligning knowledge and text embeddings by entity descriptions, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 267–272.
- [11] F. Petroni, T. Rocktäschel, P. S. H. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases, in: In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. (pp. pp. 2463–2473). Association for Computational Linguistics: Hong Kong, China. (2019), 2019, pp. 2463–2473.
- [12] J. Camacho-Collados, L. E. Anke, S. Schockaert, Relational word embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3286–3296.
- [13] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: Commonsense transformers for automatic knowledge graph construction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4762–4779.
- [14] S. Yu, T. He, J. R. Glass, Constructing a knowledge graph from unstructured documents without external alignment., arXiv: Computation and Language (2020).
- [15] C. Niklaus, M. Cetto, A. Freitas, S. Handschuh, A survey on open information extraction, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3866–3878.
- [16] M. Mausam, Open information extraction systems and downstream applications, in: IJCAI’16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 4074–4077.
- [17] P. Hohenecker, F. Mtumbuka, V. Kocijan, T. Lukasiewicz, Systematic comparison of neural architectures and training approaches for open information extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8554–8565.
- [18] Y. Yuan, X. Zhou, S. Pan, Q. Zhu, Z. Song, L. Guo, A relation-specific attention network for joint entity and relation extraction, in: International Joint Conference on Artificial

- Intelligence-Pacific Rim International Conference on Artificial Intelligence 2020, volume 4, 2020, pp. 4054–4060.
- [19] I. O. Mulang, K. Singh, C. Prabhu, A. Nadgeri, J. Hoffart, J. Lehmann, Evaluating the impact of knowledge graph context on entity disambiguation models, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2157–2160.
- [20] T. Nayak, H. T. Ng, Effective modeling of encoder-decoder architecture for joint entity and relation extraction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8528–8535.
- [21] F. Li, M. Zhang, G. Fu, D. Ji, A neural joint model for entity and relation extraction from biomedical text, *BMC Bioinformatics* 18 (2017) 198–198.
- [22] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, J. A. Kors, Knowledge-based extraction of adverse drug events from biomedical text, *BMC Bioinformatics* 15 (2014) 64–64.
- [23] B. D. Trisedya, G. Weikum, J. Qi, R. Zhang, Neural relation extraction for knowledge base enrichment, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 229–240.
- [24] Y. Dai, S. Wang, N. N. Xiong, W. Guo, A survey on knowledge graph embedding: Approaches, applications and benchmarks, *Electronics* 9 (2020) 750.
- [25] S. M. Kazemi, D. Poole, Simple embedding for link prediction in knowledge graphs, in: NIPS’18 Proceedings of the 32nd International Conference on Neural Information Processing Systems, volume 31, 2018, pp. 4289–4300.
- [26] V. Gutiérrez-Basulto, S. Schockaert, From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules, in: KR, 2018, pp. 379–388.
- [27] M. Kulmanov, W. Liu-Wei, Y. Yan, R. Hoehndorf, El embeddings: Geometric construction of models for the description logic el++, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019, pp. 6103–6109.
- [28] L. Serafini, A. S. d’Avila Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge., *NeSy@HLAI* (2016).
- [29] M. Richardson, P. Domingos, Markov logic networks, *Machine Learning* 62 (2006) 107–136.
- [30] F. Bianchi, M. Palmonari, P. Hitzler, L. Serafini, Complementing logical reasoning with sub-symbolic commonsense, in: RuleML+RR - 3rd International Joint Conference on Rules and Reasoning, volume 11784, 2019, pp. 161–170.
- [31] I. Donadello, L. Serafini, A. S. d’Avila Garcez, Logic tensor networks for semantic image interpretation, in: Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 1596–1602.
- [32] D. Calvanese, G. Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The dl-lite family, *Journal of Automated Reasoning* 39 (2007) 385–429.
- [33] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models, *arXiv preprint arXiv:2104.08663* (2021).
- [34] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextual-

ized late interaction over BERT, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48.