24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Enriching Language Models with Semantics

Tobias Mayer[1]

Abstract.

Recent advances in language model (LM) pre-training from large-scale corpora have shown to improve various natural language processing tasks. They achieve performances comparable to non-expert humans on the GLUE benchmark for natural language understanding (NLU). While the improvement of the different contextualized representations comes from (i) the usage of more and more data, (ii) changing the types of lexical pre-training tasks or (iii) increasing the model size, NLU is more than memorizing word co-occurrences. But how much world knowledge and common sense can those language model capture? How much can those models infer from just the lexical information? To overcome this problem, some approaches include semantic information in the training process. In this paper, we highlight existing approaches to combine different types of semantics with language models during the pre-training or fine-tuning phase.

1 Introduction

Pre-trained contextualized language models such as ELMo, OpenAI GPT, BERT and XLNet have been shown to effectively capture language representation and helped advance the state-of-the-art in many natural language processing (NLP) tasks. A central discipline in NLP is natural language understanding (NLU), which is a prerequisite for other downstream tasks, like natural language generation or question answering. NLU requires a sophisticated comprehension of natural language. For NLU tasks various branches of transformer models are spearheading the leaderboard of the General Language Understanding Evaluation (GLUE) benchmark. This benchmark consists of 9 language understanding tasks representing problems like sentiment analysis, semantic similarity, paraphrasing and inference. Recent transformer models surpass the performance of non-expert humans making this benchmark no longer a suitable metric [7]. The newer SuperGLUE [7] benchmark comprises harder tasks like reading comprehension, common sense reasoning or textual entailment to better quantify the performance of the understanding. While for most tasks the leaders of the GLUE benchmark also performed reasonably well [2], they are significantly worse than humans on the causal reasoning task and co-reference dependent reading comprehension, where the human baseline is at 100% accuracy. Besides a deep understanding of the discourse, these problems require common sense and world knowledge. It has been shown that BERT-based models in the higher layers do capture some kind of semantic abstraction [6] and the performance of the models on the aforementioned tasks is also high. But how well do these models understand the interactions in a discourse and how much common sense and world knowledge can be learnt from just word co-occurrences? More importantly, how can the major limitation of being trained only with character-based features be enhanced to capture more of this information?

[1] Univ. Côte d'Azur, Inria, CNRS, I3S, France, email: tmayer@i3s.unice.fr

One option is to include semantic information in the training process. We show here recent approaches in this direction. Section 2 describes one approach to incorporate world knowledge from knowledge bases during pre-training. The next section highlights discourse and semantic aware approaches, which inject the semantic information during either pre-training or fine-tuning.

2 Semantics from Knowledge Bases

Additionally to the aforementioned problems of understanding the discourse, knowledge dependent tasks, like fine-grained relation classification or entity typing, pose challenges for models trained solely on contextual character-based features. For example in *Bob Dylan wrote Blowin' in the Wind in 1962*, it is hard to determine if Bob Dylan is a writer or songwriter without knowing that *Blowin' in the Wind* is a song. This knowledge is available in knowledge bases (KB). The semantic web is full of structured world knowledge, which can be exploited. One approach to incorporate such external knowledge into language models is Enhanced language RepresentatioN with Informative Entities (ERNIE) [8]. The idea is to stack a knowledge encoder consisting of multiple aggregators on top of the encoder layers of a transformer model, where the knowledge encoder fuses knowledge graph embeddings with the contextualized embeddings into one united feature space. As a first step, named entity mentions in the text are aligned with their KB entries. The aligned named entities are represented with knowledge graph embeddings using TransE. Each aggregator takes the contextualized token embeddings from the transformer encoder and the entity embeddings and feeds them into a multi-head self-attention layer, respectively. An information fusion layer integrates the different representations coming from the two self-attention layers into one feature space. The output embeddings for each token and entity are the input for the next aggregator. The output of the last aggregator is used as the final embedding representation. For more details we refer the reader to the original paper [8]. Like BERT, the pre-training for ERNIE is done with cloze test like tasks[2]. Similar to the masked language model (MLM), they employ a knowledge masking task, where either one entity of the entity alignment is replaced with a random entity, a token-entity alignment is masked, or the alignment stays unchanged. For ERNIE 1.0 the pre-training comprises MLM, next sentence prediction (same as for BERT) plus the knowledge masking task, while ERNIE 2.0 consists of more tasks[3]. Adding only the knowledge masking to the pre-training, ERNIE 1.0 significantly outperforms BERT on entity typing and relation classification datasets while still delivering comparable results on GLUE.

[2] Cloze tests are fill-in-the-blank tests, which require an understanding of the context and are commonly used in language learning.

[3] The discourse semantic aware pre-training task will be quickly highlighted in the following section.

With ERNIE being a first step towards integrating heterogeneous information coming from world knowledge databases, the next step is to inject common sense knowledge in a similar fashion. There are available resources providing this knowledge to a certain extend, e.g. ConceptNet [4] or ATOMIC [3] in form of cause and effect relations.

3 Contextual Semantics

One approach to include contextual semantics to language modelling is SemBERT [9], motivated by the semantically incomplete answer spans of BERT on the Stanford Question Answering Dataset (SQuAD), where single semantic discourse units were broken down and only parts were classified as the answer to the question. For example, answering *How many people does the Greater Los Angeles Area have?* with *17.5 million* instead of *over 17.5 million*. To overcome this problem, the authors integrated information from semantic role labeling (SRL) in the sequence encoding. As a first preprocessing step, the input sentences are annotated with a semantic role labeler. Each token is assigned a list of labels, where the length of the list is the number of semantic structures output by the semantic role labler. The embeddings of each semantic role label are learnt via a BiGRU and subsequently fed into a linear layer to obtain one joint representation for each word in the sequence. In parallel, the subword-level representations from the BERT encoder are converted to word-level using a CNN with max pooling to match the token length of the SRL output. The contextualized and semantic embeddings are concatenated to form the final embedding. While the BERT encoder is initialized with pre-trained weights, the weights for the BiGRU are learnt during the fine-tuning on a specific task. SemBERT outperformed the existing models on GLUE and SQuAD[4].

Another way to inject discourse knowledge is discourse-aware semantic self-attention [1], which replaces the basic multi-head self-attention block in the transformer encoder. Here, the motivation comes from integrating discourse information into reading comprehension to better understand interactions, causation and temporal sequences in the text. For example, given the context: *Jacob frequently visits Jeff and Kenny, who are serving time in a juvenile hall. Jacob initially threatens them, until eventually Jeff commits suicide.* To answer *Why did Jeff commit suicide?* one needs to understand that the suicide is caused (*until eventually*) by Jacob threatening Jeff (*them*). For this, structured knowledge about entity co-reference and their semantic roles are required as much as information about the discourse relations between text sequences. To learn all this information, the proposed self-attention gets three additional inputs[5], which are represented by one embedding vector, respectively: 1) semantic role label; similar to the aforementioned approach, embeddings for the semantic roles are learnt. 2) discourse relation label; following 15 fine-grained discourse relation sense types from the Penn Discourse Tree Bank annotation scheme, such as *causation* or *contrast*. 3) label of the co-reference cluster; where tokens referring to the same entity are assigned to the same cluster. Using these linguistic annotations, the model outperforms the same model with the basic self-attention by +3.43 Rouge-L on NarrativeQA reading comprehension. Concerning the impact of the individual linguistic information, the authors found that information about the SRL improves *who* and *when* questions, while information about the discourse relations is beneficial to answer *why* and *where* questions.

Similar to the discourse-aware semantic self-attention, ERNIE 2.0 [5] takes advantage of information about the discourse relations. One of the added tasks for pre-training with respect to the previous version, is the discourse relation classification task. Here, the model has to predict the marker, e.g. *but*, for an explicit discourse relation between two sentences. Together with the continual learning strategy and the other added pre-training tasks related to lexical, structural and semantic information, ERNIE 2.0 shows significant improvement compared with the previous version.

4 Conclusion

In this paper, we have presented various ways to combine information about semantics and discourse with current state-of-the-art transformer-based language models. Furthermore, we have shown one example of how to inject world knowledge coming from KBs into LM pre-training. We consider the addition of semantics to LMs trained on only contextualized character-based features an important and inevitable step towards natural language understanding. Especially with respect to common sense, world knowledge and co-referential discourse, current contextualized representations cannot solve the challenges of general language understanding alone.

ACKNOWLEDGEMENTS

This work is partly funded by the French government labelled PIA program under its IDEX UCA JEDI project (ANR-15-IDEX-0001).

REFERENCES

[1] Todor Mihaylov and Anette Frank, 'Discourse-aware semantic self-attention for narrative reading comprehension', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2541–2552, Hong Kong, (November 2019). Association for Computational Linguistics.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[3] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi, 'ATOMIC: an atlas of machine commonsense for if-then reasoning', *CoRR*, **abs/1811.00146**, (2018).

[4] Robyn Speer and Catherine Havasi, 'Representing general relational knowledge in ConceptNet 5', in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 3679–3686, Istanbul, Turkey, (May 2012). European Language Resources Association (ELRA).

[5] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang, 'ERNIE 2.0: A continual pre-training framework for language understanding', *CoRR*, **abs/1907.12412**, (2019).

[6] Ian Tenney, Dipanjan Das, and Ellie Pavlick, 'BERT rediscovers the classical NLP pipeline', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, (July 2019). Association for Computational Linguistics.

[7] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, 'Superglue: A stickier benchmark for general-purpose language understanding systems', *CoRR*, **abs/1905.00537**, (2019).

[8] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu, 'ERNIE: Enhanced language representation with informative entities', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1441–1451, Florence, Italy, (July 2019). Association for Computational Linguistics.

[9] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou, 'Semantics-aware BERT for language understanding', in *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*, (2020).

[4] While later models like XLNet and RoBERTa outperform SemBERT, they still do not consider semantic information. The proposed approach to inject semantics can be implemented in these LMs as well.

[5] Linguistic annotation is a pre-processing and relational annotations spanning multiple sentences are projected from paragraph-level to token-level.