ETH zürich

Department of Computer Science

Towards a Deeper Understanding of Semantic Comprehension in Language Models Paired with Semantic Graphs

Matthias Kleiner[1], Ahmet Özüdogru[1], David Zollikofer[1]

[1]authors contributed equally, names sorted alphabetically

1 Introduction

Language models that are pretrained on large amounts of non-annotated data have proven themselves useful in numerous downstream tasks. Although they successfully learn good representations for syntax they struggle with capturing semantic meaning (Tenney et al., 2019; Wu et al., 2021).

To circumvent this, (Wu et al., 2021) proposes infusing additional semantic information into language models. Concretely, they use RoBERTa as a starting point, with a relational graph convolutional network (RGCN) (Schlichtkrull et al., 2018) which is stacked on-top of the transformer. Using DELPH-IN minimal recursion semantics (DM) (Ivanova et al., 2012; Oepen et al., 2014) they build a semantic graph whose nodes are populated with the corresponding contextualized embeddings from the RoBERTa transformer.

2 Our Contribution

Our contribution lies in using abstract meaning representation (AMR), a high level semantic abstraction (Banarescu et al., 2013), instead of DM as used in (Wu et al., 2021) for infusing semantic information into language models. This is based on the hypothesis that AMR graphs contain additional relevant semantic information compared to DM graphs.

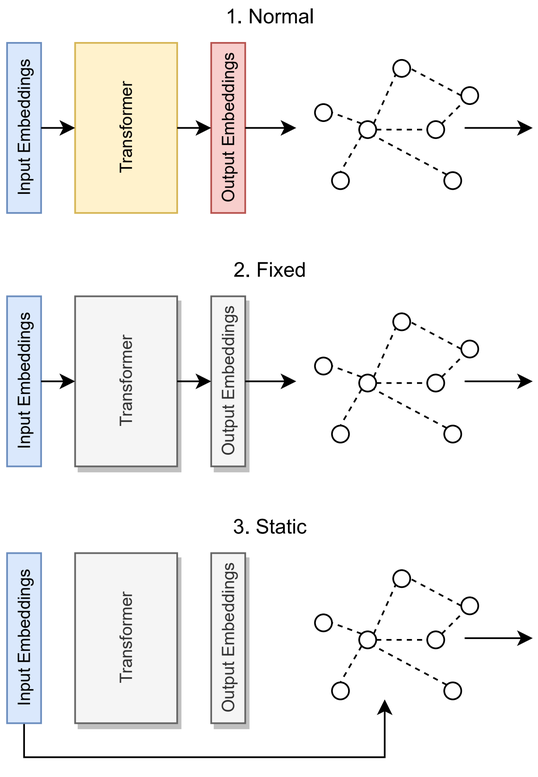

Figure 1: Illustration of the three modes we are evaluating (the graph is either AMR or DM). The weights of the non coloured components are frozen and hence they are not being trained.

We perform detailed ablations on semantic understanding in language models paired with a graph neural network on multiple GLUE (Wang et al., 2018) subtasks RTE, QNLI and MNLI (reporting accuracies on ID as well as OOD). We investigate the impact of AMR versus DM in a number of different settings: (1) fine-tuning the underlying transformer, (2) freezing transformer weights and (3) using non-contextualized embeddings.

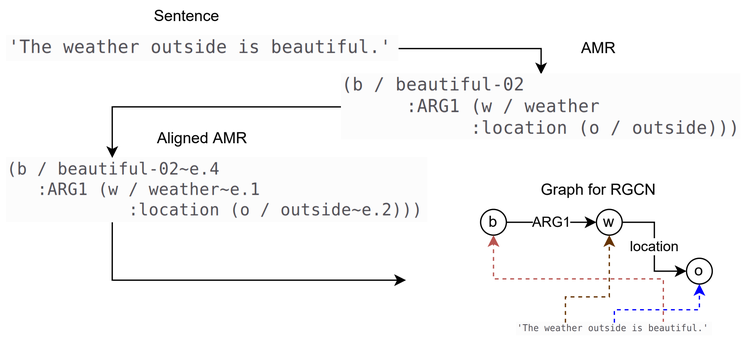

Figure 2: Illustration of our AMR preprocessing pipeline.

To generate AMR representations, we build a data processing pipeline using (Jascob, 2022; Goodman, 2020). As AMR abstracts the semantics of a sentence, there are no longer direct alignments between AMR graph nodes and parts of a sentence, as we had with DM. Hence, we use a *Rule Based Word Aligner* to assign tokens of the sentence to the individual nodes of the AMR graph.

3 Results and Analysis

			MNLI	
Models	**RTE**	**QNLI**	**ID.**	**OOD.**
RoBERTa by Wu et al. (2021)	$79.0_{\pm 1.6}$	$93.0_{\pm 0.3}$	$87.7_{\pm 0.2}$	$87.3_{\pm 0.3}$
DM	$79.66_{\pm 0.55}$	$92.61_{\pm 0.01}$	$87.15_{\pm 0.10}$	$87.09_{\pm 0.04}$
AMR	$79.18_{\pm 2.17}$	$\mathbf{92.89}_{\pm 0.13}$	$87.15_{\pm 0.01}$	$87.20_{\pm 0.08}$
DM static embeddings	$50.78_{\pm 0.21}$	$70.53_{\pm 0.23}$	$\mathbf{70.20}_{\pm 0.18}$	$\mathbf{70.19}_{\pm 0.02}$
AMR static embeddings	$\mathbf{52.95}_{\pm 0.91}$	$69.70_{\pm 0.01}$	$67.46_{\pm 0.13}$	$67.52_{\pm 0.18}$
DM fixed encoder	$61.61_{\pm 2.11}$	$84.66_{\pm 0.23}$	$\mathbf{79.04}_{\pm 0.23}$	$\mathbf{79.27}_{\pm 0.08}$
AMR fixed encoder	$59.93_{\pm 2.01}$	$84.47_{\pm 0.14}$	$77.59_{\pm 0.01}$	$78.34_{\pm 0.20}$

Figure 3: Analogous to the SIFT paper, we report mean ± standard deviation; for each bold entry of the DM or AMR model, the corresponding mean minus the standard deviation is no worse than the corresponding mean, of the opposite AMR or DM, plus standard deviation.

We evaluate all our models on the GLUE (Wang et al., 2018) subtasks RTE, QNLI and MNLI (reporting accuracies on ID as well as OOD).

Results of the Ablation Study Looking at the results, we see that RoBERTa performs at least 15 percent better than all of our models in the static embedding training mode (where no language models are used). Hence, our first result is that **the contribution in capturing semantic meaning of the RGCN is very small compared to the contribution of the language model**.

Next, we investigate which part of a language model contributes to the success in capturing semantics. We observe that all of our models in the fixed encoder training mode perform at least 6 percent better than the corresponding ones in the static embedding training mode. Likewise, our models in normal training mode perform at least 7 percent better than the corresponding ones in fixed encoder training mode. Thus, we conclude **that both using the contextualized embeddings, and finetuning the weights of RoBERTa play a significant role in capturing semantic information** as their respective performance improvements are consistent throughout all tests conducted.

Comparing DM and AMR The most significant case where AMR models perform better than the DM is in the static embedding training mode on the RTE dataset. They achieve an accuracy 2 percent better than the DM model.

However, on the MNLI dataset DM dominates AMR models by 2-3 percent in static and fixed training modes, but while almost having the same performance in the normal training mode.

Hence, **we are unable to conclude that one representation is better than the other**.

4 Conclusions

We conclude that the impact of infusing additional semantic information is minor in comparison to the effect of a language model such as RoBERTa. Moreover, between DM and AMR it is unclear which one provides more utility to our model as their performance ranges are very similar and often overlap.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Michael Wayne Goodman. 2020. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. 2012. Who did what to whom? a contrastive study of syntacto-semantic dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 2–11, Jeju, Republic of Korea. Association for Computational Linguistics.

Brad Jascob. 2022. AMRLib. [Online; accessed 22. Apr. 2022].

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.