Project Proposal

Computational Semantics for Natural Language Processing
ETH Zürich - Spring 2022 - Prof. Mrinmaya Sachan

Matthias Kleiner, Ahmet Özüdogru, David Zollikofer
{makleine, oahmet, zdavid}@student.ethz.ch

April 22, 2022

1 Overview

Background Previous research [1] has shown that pre-trained language models are limited in the semantic understanding they capture by comparing ceiling and probe models, and suggests augmenting language models with semantic graph structures for a wide range of downstream language tasks.

Tasks such as natural language inference require a deep understanding of the semantics of given segments of text. The SIFT method introduced in [1] combines a language model which provides embeddings for a semantic graph on which a graph neural network is trained. This construction outperforms classical fine-tuning of language models on a wide variety of GLUE tasks.

Motivation Natural language inference, specifically entailment, requires a deep semantic understanding of the given texts. As a result, we deem it interesting to specify to what degree adding more semantic information to these graph structures can potentially help in tasks which likely depend on a good semantic understanding, in our case inference and entailment.

Our Contribution Concretely, we build on top of the SIFT method and plan to add more semantic information to the graph structures in the form of abstract meaning representation (AMR) graphs. AMR graphs are independent of the sentence structure and capture semantic meaning.

Using a pretrained AMR graph generator such as [2] or the recently published [3] we aim to extract AMR information from the entailment tasks specified in the data section below.

How exactly we will add the information gathered from the AMR graph will be an active area of research of our project. We note that there has been previous research on information extraction using AMR graphs [4] and GNNs, but we have not found any prior research focusing on natural language inference (NLI).

Project Outcomes The project has two possible outcomes: Either enhancing the graph structure with AMR shows possibly significant improvements in performance in

Heuristic	Premise	Hypothesis	Label	RoBERTa	SIFT
Lexical Overlap	The banker near the judge saw the actor.	The banker saw the actor.	E	98.3	**98.9**
	The judge by the actor stopped the banker.	The banker stopped the actor.	N	68.1	**71.0**
Sub-sequence	The artist and the student called the judge.	The student called the judge.	E	99.7	**99.8**
	The judges heard the actors resigned.	The judges heard the actors.	N	25.8	**29.5**
Constituent	Before the actor slept, the senator ran.	The actor slept.	E	**99.3**	98.8
	If the actor slept, the judge saw the artist.	The actor slept.	N	**37.9**	37.6

Table 1: HANS heuristics, RoBERTa-base and SIFT's accuracy. "E": entailment. "N": non-entailment. Table from [1] (Wu et al. 2021)

the entailment tasks, which likely confirms the hypothesis that adding additional semantic information to the graph structure aids performance. On the other hand, if no improvement in performance can be found, it is possible that additional semantic information does not benefit entailment tasks or that current models, such as our baseline, already implicitly incorporated the semantic information we added. Meaning, the model structure cannot extract additional useful information from the AMR graphs compared to what was previously already provided by the DM graph structure. We however deem the latter to be unlikely, as the comparison between ceiling and probe models in [1] has shown. We however deem the latter to be unlikely given the comparison between ceiling and probe models in [1].

2 Data

The datasets we intend to use are HANS[5] and GLUE[6] subtasks, in particular QNLI, WNLI and MNLI, which are all entailment tasks. Our approach is to first focus on MNLI and HANS. Once we have a satisfactory model for those, we intend to expand our efforts to the remaining entailment tasks. Furthermore, if time allows for it, we deem it desirable to look into the implications of our modeling design decisions for non-entailment based tasks, by measuring change of the performance in the respective downstream tasks. We might also use the SNLI[7] dataset for additional pretraining, as commonly done, before fine-tuning on a specific downstream task.

As previously mentioned, our interest lies in semantic structure relevance for NLI, in particular entailment tasks. The HANS heuristic results in table 1, from the SIFT[1] paper, illustrate the struggle of existing models with non-entailment samples.

A sample input is constructed as follows: We use a premise and hypothesis pair as presented in table 1, concatenate the two separated with the special token, and finally use the RoBERTa tokenizer to arrive at our input embeddings. From the two sentences we can then also create the DM and AMR graphs which are used to infuse semantic information.

3 Model

We directly build upon the SIFT model a presented in [1] which is composed of a language model (RoBERTa), a parser to generate the semantic DM[8] graphs as well as a graph neural network, namely the RGCN[9] for the existing SIFT structure, which runs on the aforementioned graph.

Concretely, we add a third component to the model, namely the AMR graph generator. We will then augment the existing DM[8] graph with information obtained from the AMR graph. How this augmentation is performed, is the main focus of our research and is purposely left open-ended. Possible design questions include whether to do AMR graph node to sequence alignment, whether to combine the two sequences or purposefully keep them separated in the GNN and how GNN readout as well as final classification is performed.

By design of SIFT, the tasks give us two sequences to create AMR graphs out of, we therefore, we have to do alignment of BERT-like outputs and AMR graph in a smart way. Next, we have to decide whether to combine the two sequences or purposefully keep them separated in the GNN and add a head on top of the GNN output, which will eventually perform the entailment classification.

As we are augmenting and possibly changing the graph structure, we will likely change the GNN model to be able to take advantage of the newly added information.

4 Methodology

4.1 Evaluation & Baseline

We are evaluating accuracy as well as F1 scores for the individual NLI classification tasks. Accuracy as it is often used in literature and allows comparing our results across paper; on the other hand we also report F1 scores as we deem the harmonic mean between precision and recall to be more indicative of performance especially as the datasets are not necessarily evenly distributed between entailment and non-entailment samples.

We will be using the SIFT models provided by [1] as a baseline, as it is the basis on which we add more semantic information using AMR graphs.

4.2 Tools

We will be using a Python-centric workflow. To build our models we will extend code already provided by [1] on GitHub which builds on-top of the Huggingface transformers library, DGL, as well as PyTorch.

For the AMR graph construction we can use the established library AMRlib [2] which provides AMR graphs as well as different alignment methods. If possible we would additionally like to work with [3] which has been published recently, though pretrained models might not be available.

In order to make the training feasible, we will be working with GPUs to speed up our models.

References

[1] Z. Wu, H. Peng, and N. A. Smith, “Infusing finetuning with semantic dependencies,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 226–242, 2021.

[2] “AMRLib,” Mar. 2022, [Online; accessed 22. Apr. 2022]. [Online]. Available: https://amrlib.readthedocs.io/en/latest

[3] J. Zhou, T. Naseem, R. F. Astudillo, Y.-S. Lee, R. Florian, and S. Roukos, "Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based amr parsing," *arXiv preprint arXiv:2110.15534*, 2021.

[4] Z. Zhang and H. Ji, "Abstract meaning representation guided graph encoding and decoding for joint information extraction," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 39–49.

[5] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," *arXiv preprint arXiv:1902.01007*, 2019.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[7] Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D., "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, 2015.

[8] A. Ivanova, S. Oepen, L. Øvrelid, and D. Flickinger, "Who did what to whom? a contrastive study of syntacto-semantic dependencies," in *Proceedings of the sixth linguistic annotation workshop*, 2012, pp. 2–11.

[9] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference.* Springer, 2018, pp. 593–607.