Exposé
Predicting Veridicality in a Joint Textual-Symbolic Neural Inference and Generation Architecture

Philipp Leon Alfred Meier
Universität Heidelberg
{meier}@cl.uni-heidelberg.de

1. Introduction

This master thesis aims to explore the veridicality phenomenon as a Natural Language Inference (NLI) task in an architecture for joint textual-symbolic neural inference and generation. A BART model (AMRBART) specialized on processing Abstract Meaning Representations (AMR) will be fine-tuned on MNLI data and used for classification and generation tasks. AMRBART will receive different inputs for the downstream tasks. The research question is does the abstraction level of AMR graphs support semantic tasks like classification and generation?

2. Background

Veridicality is a linguistic phenomenon. In Natural Language Inference (NLI) a context is called *veridical* if the reader can make an assumption about the truthfullness of an assertion. Figure 1 shows an example of veridical inference: Sentence A says that Jo knows that Ann and Bob left. Its counterpart A' says 'Jo hopes that Ann and Bob left.' We can only derive the truthfulness for complement B 'Ann and Bob left' from sentence A, since it is 'known' that both left. This allows the inference C, that 'Ann left.'.

This phenomenon is interesting since it challenges the generalization capability of modern systems as shown in recent work. Mastering this task is helpful for relation and event extraction systems.

2.1. AMR graphs

Abstract Meaning Representation is a semantic representation language, where each AMR graph represents one english sentence. It captures 'who is doing what to whom' in a sentence. Figure 2 shows an AMR in Penman notation. The characters 'w', 'b' and 'g' are variables and correspond to the nodes in figure 3. The slash '/' is the abbreviation of ':instance'. An AMR graph is rooted, labeled and acyclic as demonstrated in figure 3. AMR graph realise PropBank

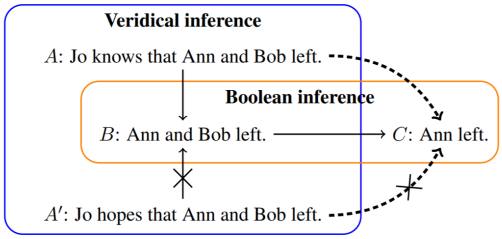

Figure 1. Inference example (Yanaka et al., 2021)

frames (Palmer et al., 2005).

```
(w / want-01
  :arg0 (b / boy)
  :arg1 (g / go-01
           :arg0 b))
```

Figure 2. Example of an AMR in Penman Notation (Banarescu et al., 2013)

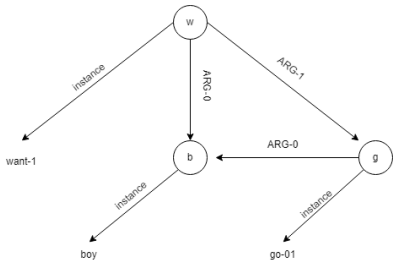

Figure 3. Example of an AMR graph

AMRs were created by (Banarescu et al., 2013) with the intention to unify semantic annotation. Before the introduction of AMR, there was a lack of an ~~approachable~~ annotated corpus (also called sembank). (Banarescu et al., 2013) introduced ~~through AMR~~ a ~~simple readable~~ sembank to encourage natural language understanding (NLU).

3. Encoding veridicality

Little work has been done to incorporate the veridicality phenomenon in AMR graphs. Williamson et al. (2021) propose to encode non-veridicality through a new role named :content in AMR-graphs. (Crouch and Kalouli, 2018) propose named graph structures with more expressive power.

This work proposes a more simple encoding of veridicality through a relation called 'veridical'as illustrated in figure 4. In this case, the value of 'veridical' is negative, because it cannot be infered that the complement is true. The variable can take three values: Positive (+), Negative (-) and Neutral (o).

```
(b/believe-01
    :ARG0 (b2/boy)
    :ARG1 (s/sick-05
        :ARG1 b2)
        :veridicality -)
```

Figure 4. Example of veridicality encoding, Sentence is 'The boy believes he's sick.'

This has the advantage that the implementation is relatively easy while being only a small change to the current AMR representation scheme.

4. System

The AMRBART model (Bai et al., 2022) serves as the base system for the experiments in this thesis. BART stands for 'Bidirectional and Auto-Regressive Transformer' which combines a Bidirectional Encoder (e.g BERT) and an Autoregressive Decoder (e.g GPT). Pre-training objective is the reconstruction of corrupted documents.
AMRBART allows AMR-to-text generation or AMR parsing. Figure 5 provides an overview over the two mentioned tasks. Bai et al. introduce four pre-training tasks for text and graph integration for AMRBART. The tasks are:

- Graph augmented text denoising: AMR graph supports text reconstruction
- Text augmented graph denoising: Text supports masked graph reconstruction
- Noisy graph augmented text denoising: Target text is generated based on masked text and graph
- Noisy text augmented graph denoising: Target graph is generated based on masked text and masked graph

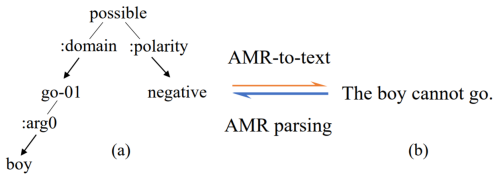

Figure 5. Overview of AMR tasks (Bai et al., 2022)

5. Data

(Ross and Pavlick, 2019) introduced a new Natural Language Inference (NLI) test set dervied from the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2017) for veridicality evaluation. In NLI, a model should be able to determine whether a hypothesis is true (entailment), false (contradiction) or neutral, given a premise. AMRBART will be fine-tuned for classification and generation on MNLI data. An example of NLI data is shown in table 1.

This dataset contains 1500 sentence pairs, including 137 verbs and their corresponding signatures. These signatures a verb allows positive (+), negative (-) or neutral (0) assertions. For example, a signature like +/+ displays a complement is positively in positive and in negative environments. Table 2 shows each signature with an example. Consider the sentences 'John knows that Ann left the party.' and 'John did not know that Ann left the party.'. 'Knows' has a +/+ signature, which means the complement, 'Ann left the party' holds true in positive and negative environment. For each sentence, the set contains a negated counterpart and the complement of the sentence.
In their work, a BERT model was fine-tuned on MNLI data in order to predict over three classes: entailment (+), contradiction(-) and neutral(o). Moreover, the dataset contains the verb, sentence, the negated sentence and the complement. Additionally, it contains only specific verb-complement constructions with 'to' or 'that'.

The Linguistic Data Consortium (LDC) provides AMR 2.0 and AMR 3.0 corpora which was used in (Bai et al., 2022) for pre-training for the AMRBART model.

6. Tasks

In this thesis, we aim to solve two tasks. In the classification task the signatures mentioned in section 5 should be correctly classified. The generation task demonstrates how a model performs in NLI entailment and in veridicality inference generation.

6.1. Classification

As pre-step in the classification task, the AMRBART model and BART-large are fine-tuned on MNLI data and then tested on NLI data. This setting demonstrates the AMRBART's capability for a NLI setting. Additionally, this experiment gives first insights about using AMR graphs as input.

Veridicality is related to Natural Language Inference since veridicality influences which inferences can be drawn. Figure 1 demonstrates this dependency. Sentence A allows to draw the inference 'Ann left'. On the contrary, A' does not allow this inference due to 'hope'. It is not confirmed, that Ann and Bob left the party. The capability to draw such inferences is fundamental for NLI systems.

In the next step, AMRBART is again fine-tuned on MNLI data and evaluated on the test set of (Ross and Pavlick, 2019). Similar to the pre-step, the model is trained to make a softmax classfication over the three classes Entailment, Contradiction and Neutral.
In (Ross and Pavlick, 2019) a BERT takes the positive or negative sentence and the complement as input. Then a prediction over the above mentioned classes is made. Two examples for the input for BERT can be seen below, the true label is on the right side, where [CLS] stands for Classification:

- [CLS] Nike declined to be a sponsor [SEP] Nike was a sponsor → Contradiction (-) (positive environment)
- [CLS] Nike did not decline to be a sponsor [SEP] Nike was a sponsor → Neutral (o) (negative environment)

Decline has the signature -/o, which means in a positive environment the complement is not true, therefore the correct prediction is contradiction (-). In a negative environment, one cannot make an assumption about the truthfulness of the complement, therefore the model should predict neutral (o). Entailment corresponds to the sign (+).

AMRBART is a sequence-to-sequence model, unlike BERT which is a encoder-only model. For this reason, the input differs for BART. In fine-tuning BART for a sequence classification task, the same input is fed to encoder and decoder. The final decoder token is then used to in linear classifier. The fine-tuned model has then the following input:

- Nike declined to be a sponsor [SEP] Nike was a sponsor [EOS] → Contradiction (-) (positive environment)
- Nike did not decline to be a sponsor [SEP] Nike was a sponsor [EOS] → Neutral (o) (negative environment)

For AMRBART, the input can be purely text or AMR graphs or a combination of both. In the combination, the sentence and the complement are given as text and graph: (Sentence-Text, Sentence-AMR) [SEP] (Complement-Text, Complement-AMR). These experiments aim to answer the question whether AMR input supports the classification for veridicality setting.

6.2. Generation

In the generation task we aim to explore the AMRBART's entailment generation and veridicality inference generation capabilities. There are two tasks for generation:

- NLI conclusion generation: Given a premise, generate a hypothesis, e.g A person on a horse jumps over an obstacle → A person is outdoors, on a horse.
- Veridicality (inference) generation: Given a premise, generate an entailed proposition: Nike declined to be a sponsor → Nike is not a sponsor.

To test how well the system can predict veridicality entailments by generating them (as entailed conclusions), the model is fine-tuned by giving premise as input and expect it to generate an entailed proposition. Again, MNLI data is used for the generation tasks.

When using BART for generation tasks, the model generates a manipulated input string through copying information from the input string. For both tasks the input is the premise encoded as text, AMR or both and the model should then generate a hypothesis or the entailed proposition as text.

Table 3 gives an overview over the planned tasks with its corresponding input and output.

7. Evaluation

To evaluate the performance in the classification task Accuracy, Precision, Recall and F1-Score are used. Additionally, Smatch, S2Match and Spearman's ρ are used to evaluate AMR graphs.

The evaluation in the NLI conclusion generation task has to show if the generated hypothesis is correct given the premise. Golden references are provided, therefore metrics like BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) can used to calculate the overlap between the generated hypothesis and the gold reference.

8. Ablation

The ablation study aims to answer the research question: Does the [illegible]action level of AMRs support the models in tasks like classification or generation? Since AMRBART received different input combinations during the experiments, the ablation study will show possible advantages and disadvantages for the tasks.

Premise	Label	Hypothesis
Because it plays on with my childhood imagination	Neutral	The art plays on my young imagination.
The man should have died instantly	Contradiction	The man was perfectly fine.
Well you can see that on television also.	Entailment	You can see that on television, as well.

Table 1. Example from the MNLI corpus (Williams et al., 2017)

Verb	Signature	Sentence	Complement
know that	+/+	John knows that Ann left the party. John did not know that Ann left the party.	Ann left the party. (entailment) Ann left the party. (entailment)
happen to	+/-	At that moment, I happened to look up. At that moment, I did not happen to look up.	At that moment, I looked up. (entailment) ¬ At that moment, I looked up. (contradiction)
forget to	-/+	She forgot to turn on the lights. She did not forget to turn on the lights.	¬ She turned on the lights. (contradiction) She turned on the lights. (entailment)
suspect to	o/+	I suspect that that is precisely the song's origin. I do not suspect that that is precisely the song's origin.	Cannot be infered (neutral) That is precisely the song's origin. (entailment)
attempt to	o/-	Four members attempted to solve the riddle. Four members did not attempt to solve the riddle.	Cannot be infered. (neutral) ¬ Four members solved the riddle. (contradiction)
refuse to	-/o	Four members refuse to solve the riddle. Four members did not refuse to solve the riddle.	¬ Four members solved the riddle. (contradiction) Cannot be infered. (neutral)
show that	+/o	The studies show that the results can be positive. The studies do not show that the results can be positive.	The results can be positive. (entailment) Cannot be infered. (neutral)
show that	o/o	They try to see her views differently. They do not try to see her views differently.	Cannot be infered. (neutral) Cannot be infered. (neutral)

Table 2. Examples of signatures (Ross and Pavlick, 2019).

Task	Input example	Target
Classification	Nike declined to be a sponsor [SEP] Nike was a sponsor [EOS]	Contradiction
Classification	(d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor)) [SEP] (b/be-01 :arg0 (n/Nike) :arg1 (s/sponsor)) [EOS]	Contradiction
Classification	Nike declined to be a sponsor (d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor)) [SEP] Nike was a sponsor (b/be-01 :arg0 (n/Nike) :arg1 (s/sponsor))[EOS]	Contradiction
NLI conclusion generation	Nike declined to be a sponsor [EOS]	Nike is a sponsor.
NLI conclusion generation	(d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor)) [EOS]	Nike is a sponsor.
NLI conclusion generation	Nike declined to be a sponsor. [SEP] (d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor)) [EOS]	Nike is a sponsor. [SEP] (b/be-01 :arg0 (n/Nike) :arg1 (s/sponsor)
Veridicality inference gen.	Nike declined to be a sponsor [EOS]	Nike is not a sponsor.
Veridicality inference gen.	(d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor) :veridicality -) [EOS]	Nike is not a sponsor.
Veridicality inference gen.	Nike declined to be a sponsor. [SEP] (d/decline-01 :arg0 (n/Nike) :arg1 (s/sponsor) :veridicality -) [EOS]	Nike is not a sponsor. [SEP] (b/be-01 :arg0 (n/Nike) :arg1 (s/sponsor) :polarity -)

Table 3. AMRBART input examples.

References

X. Bai, Y. Chen, and Y. Zhang. Graph pre-training for amr parsing and generation. *arXiv preprint arXiv:2203.07836*, 2022.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909`.

R. Crouch and A.-L. Kalouli. Named graphs for semantic representation. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 113–118, 2018.

M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

A. Ross and E. Pavlick. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, 2019.

A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

G. Williamson, P. Elliott, Y. Ji, and J. D. Choi. Intensionalizing abstract meaning representations: Non-veridicality and scope. *arXiv preprint arXiv:2109.09858*, 2021.

H. Yanaka, K. Mineshima, and K. Inui. Exploring transitivity in neural nli models through veridicality. *arXiv preprint arXiv:2101.10713*, 2021.