AMREx: AMR for Explainable Fact Verification

Chathuri Jayaweera, Sangpil Youm, Bonnie Dorr
University of Florida, Gainesville, FL,USA
{chathuri.jayawee, youms, bonniejdorr}@ufl.edu

Abstract

With the advent of social media networks and the vast amount of information circulating through them, automatic fact verification is an essential component to prevent the spread of misinformation. It is even more useful to have fact verification systems that provide explanations along with their classifications to ensure accurate predictions. To address both of these requirements, we implement AMREx, an Abstract Meaning Representation (AMR)-based veracity prediction and explanation system for fact verification using a combination of Smatch, an AMR evaluation metric to measure meaning containment and textual similarity, and demonstrate its effectiveness in producing partially explainable justifications using two community standard fact verification datasets, FEVER and AVeriTeC. AMREx surpasses the AVeriTec baseline accuracy showing the effectiveness of our approach for real-world claim verification. It follows an interpretable pipeline and returns an explainable AMR node mapping to clarify the system's veracity predictions when applicable. We further demonstrate that AMREx output can be used to prompt LLMs to generate natural-language explanations using the AMR mappings as a guide to lessen the probability of hallucinations.

1 Introduction

With the vast amount of information circulating on social media and the constantly changing Claims about various topics, automatic fact verification has become crucial for preventing the spread of misinformation. To address this need, automatic fact-checking task (Vlachos and Riedel, 2014) and several shared tasks have been introduced to encourage NLP researchers to develop systems that gather Evidence (Fact extraction) for a given Claim and classify it (Fact verification) as to its predicted veracity. Examples include FEVER (Thorne et al., 2018b, 2019) and the current AVeriTec task (Schlichtkrull et al., 2023, 2024), which employ the labels `Supports`, `Refutes`, `NotEnoughInfo` (NEI) or `ConflictingEvidence/CherryPicking`.

Natural Language Inference (NLI) systems, which assess whether a premise semantically entails a given hypothesis (Bowman et al., 2015), have been used for fact verification, yielding demonstrably strong results in the FEVER shared task. However, there has been limited focus on the explainability of these implementations. Recent studies (Gururangan et al., 2018; McCoy et al., 2019) have highlighted NLI models' tendency to rely on spurious cues for entailment classification making it important to provide clear explanations alongside fact verification predictions.

We design and implement a new, deterministic NLI system based on Abstract Meaning Representation (AMR), dubbed AMREx, and test it on the FEVER and AveriTeC fact-checking datasets. AMR is a rooted, directed, acyclic graph with nodes representing concepts and edges denoting the relations (Banarescu et al., 2013). This representation captures semantic relationships among entities that can be difficult to identify in a syntactic representation (Ma et al., 2023). We apply an existing AMR evaluation metric (Cai and Knight, 2013), to map Claims (e.g., *X was produced Y*) to relevant Evidence (e.g., *X is a film produced by Y*). We incorporate this mapping into our AMREx system to yield partially explainable fact verification.

We assume Evidence collection has already been completed, as our focus is on the potential for *explainability* of our fact-checking results, independent of the degree of *correctness* with respect to a ground truth. This, in fact, is the key contribution of this paper: We demonstrate that explainability is valuable regardless of performance levels. If performance is high, explainability supports an exploration of the factors contributing to the algorithm's success. If performance is low, it serves as a diagnostic tool to understand what went wrong.

Fig. 1 illustrates our explainable output using

Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 234–244
November 15, 2024

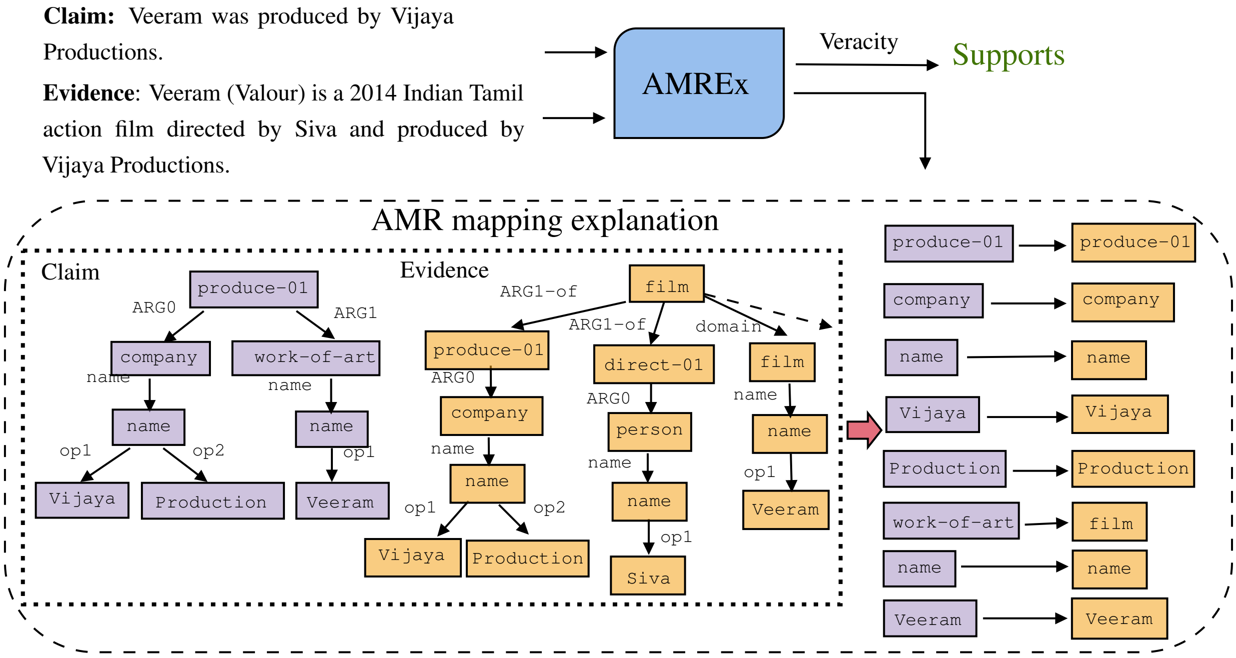

Figure 1: Explainable fact verification pipeline. Lower Left: AMR graph for the Claim, Lower Middle: AMR graph for the Evidence, Lower Right: The AMR graph mapping to explain the model's prediction as "Supports"

AMR graph mapping. We quantify the degree to which the Claim AMR is contained in the Evidence AMR and present the mappings identified in this process to demonstrate whether the Claim is embedded within the Evidence. For example, the Claim *Veeram was produced by Vijay Productions* and Evidence *Veeram (Valour) is a 2014 Indian Tamil action film directed by Siva and produced by Vijaya Productions* are represented as AMRs and processed through the Smatch algorithm. This identifies similar substructures between them, showing that both texts mention a production (rooted by `produce-01` predicate) with similar attributes and refer to the same film (through substructures rooted by `work-of-art` and `film` in the Claim and Evidence AMRs). AMREx uses this high-level notion of meaning containment, along with a textual similarity score, to produce the veracity prediction "Supports".

Section 2 reviews existing NLI implementations and explainable representations used in fact verification. Section 3 provides a detailed description of AMREx system and the experiments conducted. Section 4 presents an analysis and discussion of the results, with conclusions in Section 5.

2 Related Work

Below we explore existing studies related to NLI for fact verification, Explainable representation of fact verification, and AMR.

2.1 NLI for Fact Verification

NLI models have been employed for fact verification by assessing whether a given premise p logically infers hypothesis h (Bowman et al., 2015; Zeng and Zubiaga, 2024). These models usually classify Claim veracity using labels: `Supports`, `Refutes` and `NEI`. Thorne et al. (2018b) has developed a large-scale fact verification dataset with balanced label distribution across various domains. In this study, we adopt a 3-way (FEVER) and 4-way (AVeriTec) classification for fact verification.

With the development of fact verification datasets, fine-tuned language models (e.g., BERT, XLNet) have been applied to verify facts, improving generalizability without the need for manually crafted rules (Chernyavskiy and Ilvovsky, 2019; Nie et al., 2019; Portelli et al., 2020; Zhong et al., 2020). These BERT-based models use the Claim and potential Evidence as inputs and determine the final labels. Recently, Pan et al. (2023) fine-tuned a small dataset to enhance the performance of BERT-based models, aiming to develop domain-specific models and improving generalizability. We transcend this work by employing semantic similarity in the embedding space between Claim and Evidence, along with structural similarity.

Using pre-trained models, graph neural networks

(GNNs) have been employed to enhance reasoning for fact verification (Zhong et al., 2020; Zhou et al., 2019). These models represent Evidence as nodes within a graph, enabling information exchange between nodes, thereby improving reasoning capabilities to determine the final label. Zhong et al. (2020) use Semantic Role Labeling (SRL), assigning semantic roles to both Claim and Evidence sentences for graph construction. Building on the concept of deeper reasoning for fact verification, we apply AMR to assess sentence similarities through the lens of sentence structure.

Large Language Models (LLMs) have been utilized for fact verification by augmenting verification sources. LLMs enable more realistic fact verification by considering the date of Claims and using only the information available prior to the Claim (Chen et al., 2024). LLMs generate Claim-focused summaries, which are then used as inputs for classifiers to determine the veracity of Claims (Zhao et al., 2024). Although LLMs have demonstrated improved performance in fact verification, they still rely on classifiers that operate based on the outputs of a *black box* model.

2.2 Explainable Representations on Fact Verification

Creating explainable justifications for fact verification predictions is an essential aspect of the task as it highlights the reasons behind a veracity prediction and presents it comprehensibly and faithfully. Several attempts have been made to create such explanations using varying techniques such as interpretable knowledge graph-based rules, attention weights, and natural-language explanations using extractive and abstractive summarization, etc.

Ahmadi et al. (2019) implement an interpretable veracity prediction pipeline using Knowledge Graphs (KG) and probabilistic answer set programming that handles the uncertainties in rules created based on KGs and facts mined from the web. The resulting explanations are not in natural language but still possess a degree of interpretability. Lu and Li (2020) implement a graph-based fact verification model with attention-based explanations that highlight evidential words and users when detecting fake news in tweets. Natural logic theorem proving (Krishna et al., 2022) produces structured explanations using an alignment-based method similar to AMREx, but it operates at the sentential level, whereas AMREx uses semantic representations to create alignments. AMREx focuses on relationships among textual entities through node mapping. Similarly, Vedula and Parthasarathy (2021) combine structural knowledge with text embeddings to generate natural language explanations, akin to AMREx. However, their approach introduces a black-box relationship between the prediction process and explanation generation.

Recent developments in language models have paved the way for natural-language explanation generations where both extractive and abstractive summarization are utilized for creating explanations. Atanasova et al. (2020a) train a joint model for explanation generation and veracity classification where the extractive explanations are created by selecting the most relevant ruling comments out of a collection of them for a given Claim while Kotonya and Toni (2020) further extends this technique to create abstractive summaries for health-related Claims. Even though Large Language Models (LLMs) possess impressive generation capabilities Kim et al. (2024) show that zero-shot prompting of LLMs returns erroneous explanations due to hallucinations and focuses on generating faithful explanations using a multi-agent refinement feedback system. To address these shortcomings of LLMs, AMREx uses a linguistic approach to create a mapping of AMR graphs that explains our model's veracity predictions. We also show the potential of the mapping to be used as a prompt to generate natural-language explanations.

2.3 Abstract Meaning Representation (AMR)

AMR is a rooted, directed, and acyclic semantic representation that captures the meaning of a text through concepts and the relations that connect them (Banarescu et al., 2013). It has been used for various NLP applications such as text summarization, argument similarity detection, aspect-based sentiment classification, and natural language inference (Dohare et al., 2017; Opitz et al., 2021; Ma et al., 2023; Opitz et al., 2023), due to its ability to capture key relationships among entities and generalize meaning regardless of syntax. In AMREx, we focus on measuring the similarity between two AMRs using the Smatch score (Cai and Knight, 2013), which is designed to identify structural similarities of AMRs, effectively comparing concept relations between pairs of texts.

3 Experiment

This section presents the details behind datasets used in our experiments, along with the experimental steps carried out to build AMREx model.

3.1 Datasets

We use two fact-checking datasets to test the effectiveness of our model in verifying the veracity of Claims, as described below. For both datasets, we assume the gold Evidence for each Claim has been collected and thus focus only on verifying the Claim's veracity.

3.1.1 FEVER dataset

The FEVER dataset (Thorne et al., 2018a) consists of more than 1.8k Claims generated by altering sentences from Wikipedia. These Claims are classified into three classes: `Supports` ("S"), `Refutes` ("R") and `NotEnoughInfo` ("N"). The dataset includes relevant Evidence from Wikipedia articles for Claims in the first two classes. Some Claims require multi-hop inference/reasoning to verify their veracity.

3.1.2 AVeriTeC dataset

AVeriTeC (Schlichtkrull et al., 2024) is a newly released dataset containing 4568 real-world Claims. This dataset addresses several issues associated with previous datasets, such as inclusion of Evidence published after the Claim and artificially generated Claims. The Claims fall into four categories: `Supported` ("S"), `Refuted` ("R"), `NotEnoughEvidence` ("N") and `ConflictingEvidence/Cherrypicking` ("C"), where `ConflictingEvidence/Cherrypicking` represents Claims that have both supporting and refuting Evidence. Unlike previous datasets, AVeriTeC employs a question-answering approach to build the reasoning process for fact verification, encouraging researchers to formulate questions that support Evidence extraction and to find their answers on the web.

3.2 AMREx Model

We present the design of the AMREx for verification of Claim veracity. The underlying model is an NLI model based on a combination of an AMR evaluation metric and cosine similarity on SBERT (Reimers and Gurevych, 2019) embeddings that predicts entailment for a single (Claim, Evidence) pair. These predictions are then aggregated per claim to predict the veracity. The last stage of the

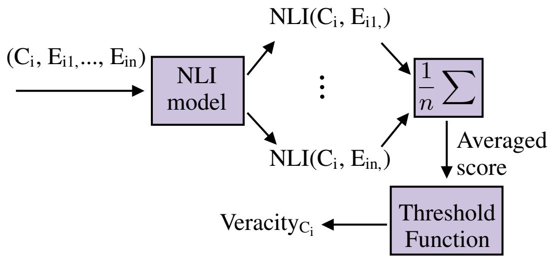

Figure 2: AMREx model: The model aggregates all the entailment predictions from the NLI model for a claim and returns the final veracity prediction

model is customized to suit the different dataset formats. (See Fig. 2 for overall AMREx pipeline).

3.2.1 NLI model

Although semantic entailment does not always correspond to a strict subsumption relationship between sentences, we adopt a simplifying assumption that entailment aligns with subsumption. Specifically, our NLI model is based on the hypothesis that if SentenceA (s_A) semantically entails SentenceB (s_B), then the meaning of s_B is contained inside that of s_A. This simplification allows our implementation to be built upon structured semantic concepts. Mapping this to AMR graph representations where g_A and g_B are the respective representations for s_A and s_B, we hypothesize that g_B is a subset of g_A. To assess how much of g_B's meaning is contained in g_A, we use the Smatch (Cai and Knight, 2013) precision score between g_A and g_B, combined with the cosine similarity of SBERT embeddings of s_A and s_B (as shown in Eq. 1) to calculate the entailment score ($f(s_A, s_B)$) between s_A and s_B. Note that the Smatch precision score is asymmetrical. So, s_A is considered the premise and s_B, the hypothesis. We then apply a threshold function (See Eq. 2) to the resulting score to classify s_A as either entailing (+1) or not entailing (-1) s_B, as shown in Eq. 3 (See Fig. 3).

$$f(s_A, s_B) = \lambda * Smatch_P(g_A, g_B) + (1 - \lambda) * Cosine_{SBERT}(s_A, s_B) \quad (1)$$

$$th_1(f(s_A, s_B)) = \begin{cases} +1, & f(s_A, s_B) \geq 0.6 \\ -1, & f(s_A, s_B) < 0.6 \end{cases} \quad (2)$$

$$NLI(s_A, s_B) = th_1\Big(f(s_A, s_B)\Big) \quad (3)$$

However, as the two datasets use slightly different labeling schemes (FEVER uses a 3-way classifi-

Dataset	S	R	N	C	Total # sentences
FEVER	3281	3270	3284	-	9835
AVeriTec	649	1166	115	226	2156

Table 1: Label distribution of FEVER and AVeriTec datasets: Supports (S), Refutes (R), Not Enough Evidence (N), Conflicting Evidence (C).

cation format, while AVeriTeC uses a 4-way classification format) and a Claim may involve multiple pieces of Evidence in the entailment process, the fact verification approach needs to be customized for each dataset. This customization will be described in Sections 3.2.2 and 3.2.3. As observed in this implementation, minor variations in verdict labels may exist across different datasets, we believe these differences are not substantial, as all labels pertain to assessing the truth value of a claim. Therefore, the threshold function can be readily adjusted to accommodate new verdict labels.

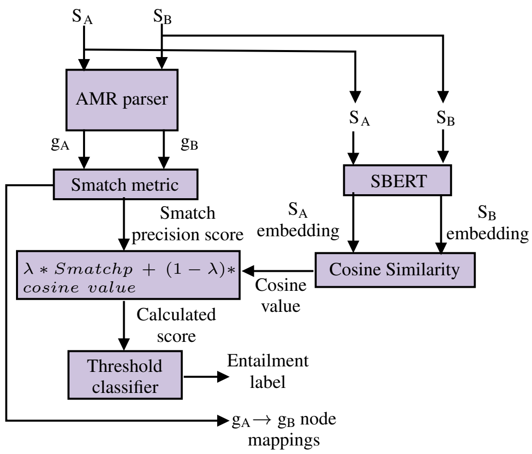

Figure 3: NLI model pipeline. S_A refers to SentenceA and S_B refers to SentneceB. g_A refers to AMR graphA from S_A and g_B refers to AMR graphB from S_B

3.2.2 Fact verification for FEVER

The FEVER dataset categorizes Claims and Evidence into three classes (`Supports`, `Refutes`, `NotEnoughInfo`. Each Claim may have one or more pieces of Evidence, while those labeled NEI lack any Evidence. To address the lack of Evidence for the NEI class, we use the modified FEVER dataset provided by Atanasova et al. (2020b), which includes Evidence for NEI class.

Given a pair (C_i, E_{ij}) where C_i is a Claim and E_{ij} is its jth Evidence, we use the NLI pipeline shown in Fig. 3 to compute the entailment between them. Here, C_i is treated as the hypothesis and E_{ij} as the premise. If E_{ij} entails C_i, it returns +1. If not, it returns -1, as outlined in Eq. 3. AMREx then averages the results across all Evidence for C_i from the NLI model, to determine the overall entailment (e), and classify that into one of the three classes using a threshold classifier to return the veracity of C_i, as shown in Eq. 4 and 5. When deciding the thresholds for the labels, "Supports" and "Refutes" are given the positive and negative extremes, respectively, whereas "Not enough Info" is assigned the middle range. This is based on the assumption that evidence with insufficient information will exhibit lower structural and textual similarity scores without extreme contradictions. The exact threshold values were determined experimentally.

$$th_{2_{FV}}(e) = \begin{cases} \text{"S"}, & e \geq 0.1 \\ \text{"N"}, & -0.1 < e < 0.1 \\ \text{"R"}, & e \leq -0.1 \end{cases} \tag{4}$$

$$Veracity_{C_i} = th_{2_{FV}}\Big(\frac{1}{n}\sum_{j=1}^{n} NLI(C_i, E_{ij})\Big) \tag{5}$$

3.2.3 Fact verification for AVeriTeC

The AVeriTeC dataset requires a Claim extraction system to first create questions to aid in finding Evidence related to a Claim, and then locate relevant documents and sentences to answer those questions, which are considered Evidence for the Claim. Since we assume the correct questions and answers are already provided for each Claim, we calculate the overall entailment between a Claim and Evidence using Eq. 5. However, we apply a customized threshold function for the AVeriTec dataset as it includes four veracity labels (Eq. 6). Additionally, the dataset features three types of Evidence: Boolean, Abstractive, and Extractive. Since Boolean Evidence (Yes/No answers) is incompatible with both AMRs and our entailment pipeline, we focus on abstractive and extractive Evidence in the experiment to fully measure our pipeline's ability to represent sentential Evidence. Table 1 shows the label distribution of both datasets.

$$th_{2_{AV}}(e) = \begin{cases} \text{"S"}, & e \geq 0.5 \\ \text{"C"}, & 0.1 < e < 0.5 \\ \text{"N"}, & -0.1 \leq e \leq 0.1 \\ \text{"C"}, & -0.5 < e < -0.1 \\ \text{"R"}, & e \leq -0.5 \end{cases} \tag{6}$$

Model	lambda	S	R	N	C	Macro F1	Acc.
FEVER baseline	_	_	_	_	_	_	0.88
$AMREx_{FEVER_{acc,f1}}$	0	0.52	0.39	0.41	_	0.44	0.44
AVeriTec baseline	_	0.48	0.74	0.59	0.15	0.49	0.49
$AMREx_{AVeriTec_{acc}}$	0.9	0.10	0.67	0.04	0.02	0.21	**0.50**
$AMREx_{AVeriTec_{f1}}$	0	0.25	0.61	0.06	0.11	0.26	0.43

Table 2: Accuracy and Macro F1 scores of veracity prediction for each veracity label. Only accuracy is reported in FEVER baseline.

4 Results and Analyses

We experiment with λ values in the [0,1] range for Eq. 1 on both FEVER and AVeriTec datasets to find the best combination of AMR graph intersection and textual similarity measurement. The results for both datasets are in Table 2. We selected the best-performing models based on both the highest accuracy and macro F1 score, leading to two AMREx implementations for each dataset.

For the FEVER dataset, the best accuracy and macro F1 score are achieved when $\lambda = 0$, suggesting that the Smatch precision score has minimal impact on predicting the veracity of (Claim, Evidence) pairs. The label-wise performance shows that $AMREx_{FEVERacc,f1}$ is more effective at identifying supporting (Claim, Evidence) pairs but struggles with refuting instances.

However, the AVeriTec dataset exhibits different behavior, with $\lambda = 0.9$ yielding the best accuracy and $\lambda = 0$ producing the best macro F1 score. $AMREx_{AVeriTec_{acc}}$ also manages to surpass the AVeriTec accuracy baseline. $AMREx_{AVeriTec_{f1}}$ performs comparably to the AVeriTec baseline in recognizing refutable (Claim, Evidence) pairs and those with conflicting evidence. However, with greater emphasis on the Smatch precision score when $\lambda = 0.9$, $AMREx_{AVeriTec_{f1}}$ improves in identifying refutable (Claim, Evidence) pairs, albeit at the cost of performance on other label instances.

Through an error analysis, we identify several cases where AMREx fails to accurately predict the veracity and we explore their potential causes. Consider the following supporting (Claim, Evidence) pair from the FEVER dataset, Claim: *"Wish Upon was released in the 21st century."*, Evidence: *"It is set to be released in theaters on July 14, 2017, by Broad Green Pictures and Orion Pictures"* (See Fig. 4 for corresponding AMRs in Penman notation (Goodman, 2020)). AMREx returns the following mapping for this instance with a Smatch precision score of 0.53 and a textual similarity score of 0.38.

```
a0(release-01) -> b2(release-01)
a1(music) -> b1(it)
a2(name) -> b10(name)
a3(Wish) -> b11(Orion)
a4(Upon) -> b12(Pictures)
a5(date-entity) -> b14(date-entity)
```

```
AMR Corresponding to the Claim:
(a0/release-01
   :ARG1 (a1/music
      :name (a2/name
         :op1 (a3/Wish)
         :op2 (a4/Upon)))
   :time (a5/date-entity
      :century 21))

AMR Corresponding to the Evidence:
(b0/set-08
   :ARG1 (b1/it)
   :ARG2 (b2/release-01
      :ARG0 (b3/and
         :op1 (b4/company
            :name (b5/name
               :op1 (b6/Broad)
               :op2 (b7/Green)
               :op3 (b8/Pictures)))
         :op2 (b9/company
            :name (b10/name
               :op1 (b11/Orion)
               :op2 (b12/Pictures))))
      :ARG1 i
      :location (b13/theater)
      :time (b14/date-entity
         :day 14
         :month 7
         :year 2017)))
```

Figure 4: Abstract Meaning Representations (AMRs) for Claim: "Wish Upon was released in the 21st century." and Evidence: "It is set to be released in theaters on July 14, 2017, by Broad Green Pictures and Orion Pictures"

The AMR node mapping correctly identifies that both texts are related to a release event (with the `a0` node mapping to the `b2` node), connects "music" in Claim AMR to "it" in Evidence AMR, and recognizes that both texts mention a `date-entity`.

However, it fails to map "the 21st-century" in the Claim with the date in the Evidence AMR. The Smatch precision score indicates a higher level of meaning entailment compared to the textual similarity score, but it is not high enough to meet the entailment threshold with any λ value, leading AMREx to incorrectly predict "Refutes". This reveals a limitation of the Smatch algorithm in inferring that the year 2017 falls within the 21st century, as it is a concept mapping algorithm. We note that SBERT contextual embeddings also fail to capture this detail and give an even lower similarity assessment.

Another example reveals that high structural similarity between AMRs, despite a few factual differences, can result in incorrect meaning containment assessments. Consider the Claim: *"Marnie is a romantic film."* and the Evidence: *"Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock."* with the gold veracity label "Refutes" (See Fig. 5 for AMRs). The resulting AMR node mappings are as follows:

```
a0(film) -> b0(film)
a1(romantic-03) -> b1(direct-01)
a2(name) -> b11(name)
a3(Marnie) -> b12(Marnie)
```

```
AMR Corresponding to the Claim:
(a0/film
   :ARG0-of (a1/romantic-03)
   :name (a2/name
      :op1 (a3/Marnie)))

AMR Corresponding to the Evidence:
(b0/film
   :ARG1-of (b1/direct-01
      :ARG0 (b2/person
         :name (b3/name
            :op1 (b4/Alfred)
            :op2 (b5/Hitchcock))))
   :mod (b6/thriller
      :mod (b7/psychological))
   :mod (b8/country
      :name (b9/name
         :op1 (b10/America)))
   :name (b11/name
      :op1 (b12/Marnie))
   :time (b13/date-entity
      :year 1964))
```

Figure 5: Abstract Meaning Representations (AMRs) for Claim: "Marnie is a romantic film." and Evidence: "Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock."

In the Claim AMR, "Marnie" being a "romantic film" is represented by the `romantic-03` node, while in the Evidence, it being a "Psychological thriller" is represented by a modifier to the root `film`. Due to this structural discrepancy, the Smatch algorithm fails to distinguish between the two genres and instead maps `romantic-03` to `direct-01` with a similar structure that still correctly creates a mismatch, but for the wrong reason. However, most concepts in the Claim AMR match those in the Evidence AMR, leading to a high Smatch precision score of 0.75. The textual similarity score also returns a 0.70. Hence, any λ combination of the two scores surpasses the entailment threshold, yielding a "Supports" prediction.

These examples reveal that the AMR and textual similarity-based approach of AMREx struggles with instances involving implied meaning or those with high structural similarity but factual differences, indicating areas that need improvement.

4.1 Explainability of the Model

The model's explainability stems from two key aspects. First, the deterministic nature of the model's calculations allows us to trace how a particular prediction was calculated. This provides a comprehensive explanation of the entire system pipeline and tracks the process at each step. Second, the visual mapping between the AMRs of Claims and Evidence, as shown in Fig. 1, helps clarify why the model returns a particular prediction for a (Claim, Evidence) pair in terms of structural similarity. This explanation is partial and post hoc, relying only on AMR node mappings for generation. However, it is integrated into the system, as AMR representations influence both the veracity prediction and explanation generation. An example illustrating AMREx's explanations is discussed below.

Consider Claim *"Rabies is a ride at Six Parks."* and the Evidence, *"Rabies is a viral disease that causes inflammation of the brain in humans and other mammals."* The corresponding AMRs for Claim and Evidence are shown in Fig. 6. When these two AMRs are processed through the Smatch algorithm, the resulting AMR node mapping is as follows:

```
a0(ride-01) -> b0(disease)
a1(disease) -> b8(disease)
a2(name) -> b9(name)
a3(Rabies) -> b10(Rabies)
a4(amusement-park) -> b2(inflame-01)
a5(name) -> b4(and)
a6(Parks) -> b6(mammal)
```

As the mapping reveals, the only shared meaning

```
AMR Corresponding to the Claim:
(a0/ride-01
   :ARG1 (a1/disease
      :name (a2/name
         :op1 (a3/Rabies)))
   :location (a4/amusement-park
      :name (a5/name
         :op1 6
         :op2 (a6/Parks))))

AMR Corresponding to the Evidence:
(b0/disease
   :ARG0-of (b1/cause-01
      :ARG1 (b2/inflame-01
         :ARG1 (b3/brain)
         :part-of (b4/and
            :op1 (b5/human)
            :op2 (b6/mammal
               :mod (b7/other)))))
   :domain (b8/disease
      :name (b9/name
         :op1 (b10/Rabies)))
   :mod (b11/virus))
```

Figure 6: Abstract Meaning Representations (AMRs) for Claim: “Rabies is a ride at Six Parks.” and Evidence: “Rabies is a viral disease that causes inflammation of the brain in humans and other mammals.”

between these two AMRs is that both sentences are related to a disease called Rabies (with the `a1`-`a3` nodes mapping to the `b8`-`b10` nodes). This leads to a low Smatch precision score of 0.46 and a textual similarity score of 0.59. When combined with any λ value, this results in a low entailment value, causing the threshold function to predict non-entailment (-1). Hence, the second threshold function will return “Refutes” as the veracity of this (Claim, Evidence) pair.

This process, together with the AMR node mapping, demonstrates the model’s overall logic and final prediction. We further demonstrate the effectiveness of the AMR node mapping in generating natural-language explanations by using it to prompt ChatGPT to generate justifications. An excerpt of the generated explanation for the (Claim, Evidence) pair above is shown in Fig. 7. Hence, we argue that AMREx’s explainable output can serve as a guide for generating natural-language explanations for veracity detection, helping to reduce hallucinations in LLM models.

5 Conclusion

We implement AMREx, an Abstract Meaning Representation-based veracity prediction and explanation system for fact verification, and show its effectiveness in producing explainable justifications using two fact verification datasets. Although its performance is lower compared to the baselines, its partially explainable output could still be used as a diagnostic tool to perform error analyses on the veracity prediction system to understand the areas to improve. We further demonstrate that AMREx output can guide LLMs to generate natural-language explanations using the AMR mappings.

AMR Graph Mapping Analysis:

...

Key Mappings:

`a0(ride-01)-b0(disease)`: The concept "ride" in the claim is mapped to "disease" in the evidence, which indicates a mismatch. The claim is about a ride, whereas the evidence is about a disease.

`a1(disease)-b8(disease)`: Both the claim and the evidence refer to "Rabies" as a disease. However, in the claim, it’s tied to a "ride" rather than a health context.

`a2(name)-b9(name)`: The name "Rabies" is consistently mapped in both the claim and the evidence.

`a3(Rabies)-b10(Rabies)`: The specific instance "Rabies" is correctly identified in both the claim and the evidence.

`a4(amusement-park)-b2(inflame-01)`:The location "amusement park" in the claim is mapped to "inflammation" in the evidence, which again indicates a significant mismatch in context.

`a5(name)-b4(and),a6(Parks)-b6(mammal)`: The reference to "Six Parks" in the claim is incorrectly mapped to a reference to "mammals" in the evidence.

Explanation:

Mismatch in Core Concepts: The claim describes "Rabies" as a "ride" at an "amusement park," whereas the evidence discusses "Rabies" as a "viral disease." The AMR mappings reveal that the central concepts of "ride" and "amusement park" in the claim do not align with the "disease" and "medical context" in the evidence.

...

Classification: Given that the evidence contradicts the central concept of the claim by presenting a different context (medical vs. amusement), this pair should be classified as REFUTES. The evidence does not support the claim that "Rabies is a ride at Six Parks" and instead presents a fact that contradicts this claim.

Figure 7: Natural Language explanation generated by ChatGPT based on the AMREx’s AMR node mapping output.

Limitations

In its current form, AMREx performs best across all classes when its score is less influenced by the structural similarity assessment. This makes the AMR node mappings less useful as explanations, despite achieving the highest performance. Therefore, further exploration is needed to adjust the structural similarity assessment to better suit the veracity detection task. Since AMREx relies heavily on AMRs, it is crucial to use a high-performing AMR parser when converting the sentences to AMRs. Therefore, the overall performance of the system depends on the accuracy of the AMR parser. Furthermore, the AMR mapping algorithm is more effective when applied to text instances with a high degree of structural similarity, which may not always be the case with real-world data. The AMR node mappings provide a partial, post hoc explanation of the system, while the interpretability of the entire system fully encompasses the prediction process. An evaluation of the explainable aspect of AMREx model in comparison to current structural explainable fact verification systems is also necessary. We expect to address these limitations in future modifications to the system.

Ethical Statement

We utilize ChatGPT responses as a demonstration of the effectiveness of AMREx in creating natural-language explanations for veracity predictions. We acknowledge that there is a possibility for ChatGPT to generate hallucinated, or toxic content. However, one of the key objectives of our study is to develop an explainable system whose output can guide the reduction of hallucinations in LLM-generated outputs, including ChatGPT. We believe this approach contributes to the generation of content that is both faithful and safe. Additionally, we manually check the ChatGPT-generated content in this study for hallucinated or toxic content and can confirm that the presented examples are free of such issues.

Acknowledgement

This material is based upon work supported, in part, by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. In *Conference on Truth and Trust Online*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364. Association for Computational Linguistics.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177. Association for Computational Linguistics.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752. Association for Computational Linguistics.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587. Association for Computational Linguistics.

Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and aggregate: A novel domain-independent approach to factual data verification. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 69–78. Association for Computational Linguistics.

Shibhansh Dohare, Harish Karnick, and Vivek Gupta. 2017. Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

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. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754. Association for Computational Linguistics.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514. Association for Computational Linguistics.

Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu, and Lijie Wen. 2023. AMR-based network for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 322–337. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35. Association for Computational Linguistics.

Juri Opitz, Shira Wein, Julius Steen, Anette Frank, and Nathan Schneider. 2023. AMR4NLI: Interpretable and robust NLI measures from semantic graphs. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 275–283. Association for Computational Linguistics.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023. Investigating zero- and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–524. Association for Computational Linguistics.

Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the evidence to augment fact verification models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 47–51. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6. Association for Computational Linguistics.

Nikhita Vedula and Srinivasan Parthasarathy. 2021. Face-keg: Fact checking explained using knowledge graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 526–534, New York, NY, USA. Association for Computing Machinery.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22. Association for Computational Linguistics.

Xia Zeng and Arkaitz Zubiaga. 2024. MAPLE: Micro analysis of pairwise language evolution for few-shot claim verification. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1177–1196. Association for Computational Linguistics.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573. ELRA and ICCL.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901. Association for Computational Linguistics.