Explainable Unsupervised Argument Similarity Rating with Abstract Meaning Representation and Conclusion Generation

Juri Opitz\textsuperscript{1} Philipp Heinisch\textsuperscript{2} Philipp Wiesenbach\textsuperscript{1} Philipp Cimiano\textsuperscript{2} Anette Frank\textsuperscript{1}

\textsuperscript{1}Dept. of Computational Linguistics, Heidelberg University
\textsuperscript{2}CITEC, Bielefeld University

\textsuperscript{1}lastname@cl.uni-heidelberg.de, \textsuperscript{2}\{pheinisch,cimiano\}@techfak.uni-bielefeld.de

Abstract

When assessing the similarity of arguments, researchers typically use approaches that do not provide interpretable evidence or justifications for their ratings. Hence, the features that determine argument similarity remain elusive. We address this issue by introducing novel argument similarity metrics that aim at high performance and explainability. We show that Abstract Meaning Representation (AMR) graphs can be useful for representing arguments, and that novel AMR graph metrics can offer explanations for argument similarity ratings. We start from the hypothesis that similar premises often lead to similar conclusions—and extend an approach for AMR-based argument similarity rating by estimating, in addition, the similarity of conclusions that we automatically infer from the arguments used as premises. We show that AMR similarity metrics make argument similarity judgements more interpretable and may even support argument quality judgements. Our approach provides significant performance improvements over strong baselines in a fully unsupervised setting. Finally, we make first steps to address the problem of reference-less evaluation of argumentative conclusion generations.

1 Introduction

Rating the similarity of arguments (Reimers et al., 2019) is a core task in argument mining and argument search (Maturana, 1988; Wachsmuth et al., 2017; Ajjour et al., 2019). Argument similarity ratings are also needed for (case-based) argument retrieval (Rissland et al., 1993; Chesnevar and Maguitman, 2004), data exploration via argument clustering, and even automated debaters (Slonim et al., 2021): to counter an opponent’s argument, one may retrieve an argument similar to theirs, but of opposite stance to the topic (Wachsmuth et al., 2018).

Typically, argument similarity ratings are computed over ‘bag-of-word’ argument representations, or else over argument representations inferred with language models such as BERT (Devlin et al., 2019) or InferSent (Conneau et al., 2017). Two key advantages of such approaches are due to their unsupervised setup: First, unsupervised methods do not rely on human annotations, which are expensive and can be subject to noise and biases. Second, it has been shown for previous supervised methods that they have learned less about argumentation tasks than had been assumed, by exploiting spurious clues and artifacts from manually created data (Opitz and Frank, 2019; Niven and Kao, 2019). This has led to a recent interest in solving argumentation tasks in an unsupervised manner, e.g., by logical reasoning (Jo et al., 2021).

In this paper we will highlight that previous methods for rating argument similarity suffer from a common flaw: beyond shallow statistics (word matches in bag-of-word models, or word similarities in distributional space), they do not provide any rationale for their predictions, and the prediction process is in general not transparent. Therefore, we know only little about the following question:

• Which argument features correlate with human argument similarity decisions?

In this work, we undertake a first attempt at answering this question, by testing two hypotheses:

i) Representing arguments with Abstract Meaning Representations (AMRs) and using AMR graph metrics improves argument similarity rating and provides explanatory information.

ii) Extending arguments with inferred conclusions can improve argument similarity rating.

In the following §2 we discuss related work. §3 introduces our two key hypotheses, and §4 presents our argument similarity rating model and its implementation. In §5 we compare our model against strong baselines from prior work. In §6 we conduct several analyses to show how our approach can contribute to a better understanding of arguments, their
conclusions and argument similarity ratings: we i) assess predictors of human argument similarity ratings to investigate the criteria that correlate with human ratings of argument similarity; ii) discuss potential advantages of using AMR for graph-based argumentation tasks in a concrete example, and iii) investigate how interpretable argument similarity computation can help assess the quality and usefulness of conclusions drawn from arguments in a reference-less conclusion evaluation setup.\footnote{https://github.com/Heidelberg-nlp/amr-argument-sim}

2 Related work

Argument similarity and search Assessing argument similarity is a key task in argument mining \citep{reimers2019}, Lenz et al. \citeyear{lenz2019} and can enhance argument search \citep{maturana1988, rissland1993, wachsmuth2017, ajjour2019, chesnevar2004}. Yet, while delivering solid performance on benchmarks, current methods fail to provide any deeper rationale for their predictions. It is thus not clear whether and to what extent spurious clues or other artifacts may influence the similarity decision \citep{opitz2019, niven2019}. In this paper, we aim at alleviating these issues by i) representing arguments with Abstract Meaning Representation \citep{banarescu2013} and conducting similarity assessment using well-defined graph metrics that provide explanatory AMR structure alignments; and ii) by investigating to what extent argument similarity can be projected to inferred conclusions.

Explanations in argumentation Until recently, the quest for explanations in argumentation was mainly focused on theory development. The Toulmin model, e.g., offers a theory of what is needed to make an argument complete \citep{toulmin2003}. Argumentation schemes, which develop taxonomies of argument types and argumentation fallacies \citep{walton2005, walton2008} can be viewed as mechanisms for explaining functions, strengths and weaknesses of arguments. Other research aims at studying the computational and formal aspects of argumentation, e.g. abstract argumentation \citep{dung1995} and Bayesian argumentation \citep{zenker2013}. Research in empirical argument mining led researchers to investigate practical methods for explanations \citep{lawrence2021, becker2021, gunning2019, rago2021, vassilades2021}. While most approaches focus on the analysis of linguistic aspects \citep{lauscher2021}, e.g., by extracting selected features \citep{aker2017, lugini2018} or leveraging discourse knowledge in language models \citep{opitz2019}, others exploit large background knowledge graphs \citep{kobbe2019, paul2020, yuan2021} such as ConceptNet \citep{liu2004, speer2017} or DBpedia \citep{mendes2012}. An advantage of our approach is the explicit graph alignment between two arguments’ meaning graphs that better marks related structures, and that can help explain argument similarity judgements.

Argument mining with graphs There is growing interest in extracting graph structures from natural language arguments. Lenz et al. \citeyear{lenz2020}, e.g., propose a pipeline for detecting and linking argumentative discourse units (ADUs). Al-Khatib et al. \citeyear{al-khatib2020} detect textual phrases and link them with POS/NEG relations, where POS indicates a positive influence and NEG a negative influence (inhibition), e.g., sports NEG health issues. However, such approaches lack finer semantic assessment: they do not distinguish word senses, and the linked entities (phrases or ADUs) are taken as atoms, which hampers explainability: when linking sports and health issues with a NEG relation, we cannot differentiate sports NEG issues and sports NEG health (only the former is correct). We target a finer analysis of argumentative texts, by representing them with dense AMR graphs. Additionally, by aligning graph representations of several arguments, our work paves the way for improved argument knowledge graph construction, aided by, or based on, AMR.

Generation of argumentative conclusions The task of conclusion generation has been recently investigated by Alshomary et al. \citeyear{alshomary2020, alshomary2021}, and allows us to infer conclusions from given premises. Conclusion generation can be seen as the inverse of argument generation \citep{sato2015, schiller2020}. In this work, we show that by considering conclusions inferred from pairs of arguments, we can improve our argument similarity ratings.

3 Hypotheses

We base our models for explanatory argument similarity assessment on two hypotheses.

Hypothesis 1: Abstract Meaning Representation \citep{banarescu2013} of arguments supports explainable argument similarity assessment AMRs are directed, rooted and acyclic
graphs that aim at capturing a sentence’s meaning in a machine-readable format. Edges are labeled with semantic relation types (e.g., negation, cause, etc.) and vertices denote either variables or concepts (variables are instances of concepts and allow us to capture coreferences). Hence, the AMR formalism captures various semantic phenomena that can play a role when assessing argument similarity.

E.g., besides the obviously useful aspect of negation, AMR captures semantic roles and predicate senses (Kingsbury and Palmer, 2002). While it is clear that similar arguments tend to involve similar predicates and predicate senses, semantic structure and role assignment may also play a role. For instance, the claims: *consumption of alcohol leads to depression vs. depression leads to consumption of alcohol* are clearly distinct, while sharing the same concepts. Other AMR facets may also be useful. E.g., AMR captures coreferences and resolving them in different ways can induce significant meaning differences. Finally, AMR includes key semantic relations (location, cause, possession, etc.) that are often implicit or underspecified in language, hence their explicit representation in AMR provides a rich basis for assessing arguments.

Arguments represented with AMR can be compared with AMR graph metrics (Cai and Knight, 2013; Damonte et al., 2017; Opitz et al., 2020) that also induce an explicit alignment between two argument graphs.

Hypothesis II: similar arguments lead to similar conclusions. We hypothesize that a key feature of similar arguments is that they invite for similar conclusions. Analogously, dissimilar arguments tend to lead to differing conclusions. Consider the following two arguments:

i) *Cannabis can have negative effects on brain development of teens.*

ii) *Smoking cannabis is harmful for the lungs.*

The arguments are dissimilar, even though they share the same (negative) stance and argue from a similar perspective (health). This dissimilarity is also reflected in the conclusions that can be inferred from them: from i) we can infer that, i.e., *Cannabis consumption should be strictly controlled for age* or *Cannabis can have a negative impact on the brain*—while from ii) we could infer that *Cannabis, if consumed, should not be smoked* or *Cannabis smokers should get their lungs checked.*

As a complementary example, the similarity of two arguments may be reinforced by the similarity of their inferred conclusions, as shown below:

i) *Fracking can contaminate water and water wells and suck towns dry.*

ii) *As a water-poor state, fracking and its toxic wastewater presents a serious danger to our communities and ecosystems.*

Arguments i) and ii) are rated as similar, presumably because they point at detrimental ramifications of fracking related to water issues. This similarity is likely to be reflected in conclusions drawn from them, such as: i) *Fracking can lead to water issues* or ii) *Fracking poses dangers for water-poor states.*

4 Argument Similarity via AMR Metrics

According Hyp I, we represent arguments with AMR graphs and rate their similarity with AMR metrics. To test Hyp II we infer conclusions from arguments with language models and compute similarity on arguments extended with their conclusion.

4.1 Models

We propose three model variants that aim at explaining argument similarity. Given two arguments $a, a'$ and their extrapolated conclusions $c = conclusion(a)$, $c' = conclusion(a')$, we compute similarity in the space of abstract meaning representation using a similarity function $f$ in three alternative ways: i) $f(a, a')$, between the two arguments, ii) $f(c, c')$ between their conclusions, iii) $f(a \oplus c, a' \oplus c')$, i.e., between the combinations of argument $a$ and its derived conclusion $c$, where we use a simple decomposable weighting:

$$f(a \oplus c, a' \oplus c') = \lambda f(a, a') + (1 - \lambda) f(c, c') \quad (1)$$

If not specified otherwise, $\lambda$ is set to 0.95. The AMR metric $f$ will be described in the following.

4.2 Implementation

AMR parser We parse all arguments from the data with the parser from amrlib$^3$, a fine-tuned T5 sequence-to-sequence model that achieves high scores on AMR benchmarks.

$^3$https://github.com/bjascob/amrlib

$^2$We choose a high value of $\lambda$ since, clearly, the premises are bound to host the primary evidence for similarity, while a conclusion may serve as auxiliary information. In our experiments, we also consider extreme decompositions ($\lambda \in \{0, 1\}$).
AMR metric We use $S^2$MATCH (Opitz et al., 2020), which is based on the AMR graph matching metric SMATCH (Cai and Knight, 2013), but admits graded concept similarity by matching concept nodes with GloVe embeddings (Pennington et al., 2014) and cosine similarity. To find an optimal graph mapping, exactly like SMATCH, it leverages a hill-climber to approximate the NP-hard problem of aligning AMR graphs. Following the alignment step, the (soft) matching of propositions (triples) are scored with an F1 score. Since, so-far, little is known about the trade-off and interface between concrete and abstract semantics in human mental representations (Mkrtchian et al., 2019), we introduce two more variants that assess similarity from complementary perspectives: $S^2$MATCH$^{Concept}$ and $S^2$MATCH$^{Struct}$. The first metric variant focuses on conceptual overlap (Fig. 1, middle), i.e. the more concrete semantic aspects, by putting a triple weight on concept matches. The second variant focuses on structural matches (Fig. 1, bottom), i.e., the more abstract semantic aspects, by putting triple weight on relation matches.

Conclusion generator We generate conclusions from arguments using the T5 model (Raffel et al., 2020) pre-trained on summarization tasks. To encourage the model to generate informative conclusions (as opposed to summaries), we further fine-tune it on premise-conclusion samples from Stab and Gurevych (2017), which contain intelligible and rational conclusions of high linguistic quality.

5 Argument Similarity Prediction with AMR Metrics: Experiments

5.1 Setup

Data set and evaluation metric We use the UKP aspect corpus (Reimers et al., 2019), which contains 3,596 argument pairs on 28 topics that have been assigned a four-way similarity rating: highly similar (HS), somewhat similar (SS), not similar (NS), different topic/‘can’t decide’ (DTORCD). Following Reimers et al. (2019), we frame the task as a binary prediction problem: highly similar (HS, SS) and non-similar (NS, DTORCD), and we conduct evaluation via cross validation with 4 folds. In every iteration, 7 topics serve as testing data, while the other 21 topics serve to tune a decision threshold of the metric score. As in Reimers et al. (2019), we evaluate the F1 score for each of the two labels and the arithmetic F1 mean (macro F1).

Baselines We compare to previously established unsupervised baselines (Reimers et al., 2019): i) Tfidf calculates cosine similarity between Tfidf-weighted bag-of-word vectors; ii) InferSent (FastText|Glove) leverages sentence embeddings produced by the InferSent model (Conneau et al., 2017) based on either FastText (Bojanowski et al., 2016) or GloVe (Pennington et al., 2014) vectors, which are compared with cosine similarity; iii) (GloVe|ELMo|BERT) Embedding uses averaged GloVe embeddings or averaged contextualized embeddings from ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) language models.

5.2 Results

Best system Table 1 shows our main results. The AMR-based approach that is based on concept-focused $S^2$MATCH scores, taking both the argument and its inferred conclusion into account, obtains rank 1 (68.70 macro F1) and outperforms all baselines, including the BERT baseline. The difference is significant with $p < 0.005$ (Student t-test). This system is closely followed by other AMR-based systems, e.g., using concept-focused $S^2$MATCH that sees only the argument (68.17 macro

5For further detail on this fine-tuning step see Appendix.

6Strictly speaking, this is not a fully unsupervised setup, however, we stick to this framing of the task to facilitate comparison to the previous work (Reimers et al., 2019).
Table 1: Main results.

| predictor          | $f(a,a)$ | $f(c,c)$ | $f(ac,a')c$ |
|--------------------|----------|----------|-------------|
| Concepts           | 0.492$^*$ | 0.290$^*$ | 0.492$^*$   |
| Sem. Role Labels   | 0.400$^*$ | 0.185$^*$ | 0.402$^*$   |
| Predicate Frames   | 0.355$^*$ | 0.232$^*$ | 0.357$^*$   |
| Reentrancies (Coref.) | 0.235$^*$ | 0.085$^*$ | 0.235$^*$   |
| Named Entity (NER) | 0.076$^*$ | 0.052$^*$ | 0.077$^*$   |
| Negations          | 0.042$^*$ | -0.011    | 0.042$^*$   |

Table 2: Semantic predictors of human argument similarity. $^*$: significant with $p < 0.05/p < 0.005$.

6 Analyses & Explainability

While these model ablations provide a global view of what matters in argument similarity rating, we now analyze the impact of finer semantic features.

6.1 Fine predictors of argument similarity

The previous experiment suggests that human argument similarity ratings can be modeled through a combination of different meaning facets, with a focus on concepts. We will now investigate how human argument similarity ratings correlate with specific meaning aspects represented in AMR graphs.

Setup For this we leverage fine-grained AMR metrics (Damonte et al., 2017) and compute semantic similarity with respect to 6 meaning aspects i) named entities (NER); ii) negation; iii) lexical concepts; iv) predicate frames; v) coreference and vi) semantic roles (SRL). Instead of merging the labels somewhat similar and similar, we keep them distinct and use a three-point Likert scale: 0 means not similar or unrelated, 0.5 means somewhat similar, and 1 means highly similar. To assess the correlation, we use Pearson’s correlation coefficient.

Results of this univariate feature analysis are displayed in Table 2. As expected from the earlier experiment, shared concepts are strong predictors for argument similarity (Concepts, $r=0.49$). Also more abstract semantic features, such as similar semantic roles, have a solid signalling effect (SRL, $r=0.40$). Similarly, coreferences have predictive capacity, though at a lower range ($r=0.23$). On the other hand, negation or shared named entities do exhibit only small (yet still significant) predictive capacity (Negation, $r=0.04$ and NER, $r=0.08$).

Motivated by this result, we conduct two extreme ablations: concept-only and structure-only metrics. While the structure-only variant shows worse results than AMR S-focus (macro F1 $\Delta f(a,a')$: -2.7), concept-only variant and concept-focused are more or less on par (macro F1 $\Delta f(a,a')$: -0.2).
The low correlation of NE overlap with human similarity ratings can in part be explained by the fact that we do not find many arguments where this could potentially matter (in our data, only 1 to 2 out of 1,000 nodes represent person NEs). However, if humans were to rate argument similarity in a dataset that features many arguments from expert opinion (Godden and Walton, 2006; Wagemans, 2011), named entity overlap may have a significant predictive capacity. Also, negation might be more important than what we see in this analysis, since it can be expressed in alternative ways (e.g., through antonyms) that are not encoded as such in AMR.

### 6.2 Example case with alignment

To illustrate the potential of using AMR for connecting and assessing arguments, we study an example case in Fig. 2. It shows the graphs and graph alignments that were found, for the actual arguments and their automatically induced conclusions, for our running example on fracking.

**Observations about argument alignment** The top figure shows the alignment of the two argument graphs, where important substructures have been linked. *Contamination of water and water wells* is linked to *endangering of our communities and ecosystems* (orange nodes and alignment). It is also appropriate that *towns that are sucked dry* is linked to *water poor state* (blue). This link is very valuable since these statements stand in a semantic EXACERBATE-relationship that may be important for the arguments’ similarity (the water-poverty of states is exacerbated if towns are sucked dry). Ideally, we would like such alignments to be labeled with a corresponding semantic relation. In future work, we plan to achieve this by leveraging commonsense knowledge graphs like ConceptNet.

**Observations about conclusion alignment** The bottom figure shows the alignment of the automatically deduced conclusions. For the left argument, the conclusion fails to produce an abstraction and more or less repeats the argument. For the argument on the right-hand side, however, the conclusion generator produced a more informative conclusion. From the input argument it concludes that *Fracking and its toxic wastewater are a threat to the environment*—focusing on the negative environmental impact of fracking. This triggers a graph alignment which adds valuable new information (see clouds with dotted margins). The alignment makes explicit that *water wells* and *toxic wastewater* stand in a correspondence in the context of *fracking*. Specifically, we see how the *contamination of wells* (left graphs) happens: *wells are polluted with toxic wastewater* (right graphs). Additionally, the left graph helps explain parts of the meaning of the right graph: Fracking and toxic wastewater are a threat because fracking contaminates water and water wells.

### 6.3 Investigations of conclusion quality

An inferred conclusion can be more or less abstract or dissimilar from the input argument. This raises the question of the *quality* of an inferred conclusion. In fact, we can apply our AMR similarity metrics to quantify the similarity of an argument and its inferred conclusion—formally: \( f(a, c) \)—which may be indicative of the *novelty* of a conclusion in relation to its premise. Hence, we investigate how AMR similarity metrics can be used to measure the *novelty* of a conclusion relative to its premise. Another aspect of conclusion quality is its *validity* or *justification*, i.e., to what extent it can be trusted. Clearly, a conclusion that is very similar to the premise has a high chance of being valid (as long as the premise is), whereas this is uncertain for parts of its meaning that do not match the premise.

In current research, not much is known about how to rate the quality of a conclusion drawn from an argument. We explore this question by performing a manual assessment of different quality aspects of conclusions, and investigate to what extent these can be assessed with our AMR similarity metrics.

We randomly sample 100 argument-conclusion pairs per topic. The pairs are given to two annotators whom we ask to assign binary ratings regarding two questions: i) Is the conclusion *justified* based on the premise? With this we aim to assess whether the argument legitimizes the conclusion; and ii) Does the conclusion introduce some *novelty* relative to the argument? This should be denied if, e.g., the conclusion repeats the premise.

As shown in Fig. 3, we measure moderate IAA, with slightly higher agreement for *novelty*. The results show that T5 often manages to produce either valid (*justification*, \( \approx 65-75\% \) of cases) or novel content (*novelty*, \( \approx 50-60\% \)), but struggles to produce conclusions that fulfill both criteria (*justification & novelty*: \( \approx 25-35\% \) of cases).
and threatens the environment with its toxic wastewater. Fracking contaminates water and sucks towns dry.

As a water-poor state, fracking and its toxic wastewater presents a serious danger to our communities and ecosystems.

Fracking contaminates water and water wells and sucks towns dry.

Fracking and its toxic wastewater are a threat to the environment.

Figure 2: Full example (edge-labels omitted for simplified display) of explicit alignments between argument graphs (top) and automatically induced conclusions (bottom). Here, the conclusions help explaining argument similarity, since the alignment connects fracking in both graphs, as well as water wells and toxic wastewater, showing how contaminating of the wells (left graphs) actually happens: wells are polluted with toxic wastewater (right graphs).

Figure 3: Annotation results of two quality aspects with IAA: $K=0.49$ (justification) and $K=0.57$ (novelty).

### 6.4 Can we predict conclusion quality?

We now extend the use of our metrics to assess conclusion quality by computing the similarity of argument and conclusion: $f(a, c)$\(^9\). We calculate six graph similarity statistics of their AMRs to finally produce an aggregate score assessment: i) $|a \cap c|/|a|$ measures the relative amount of premise content that is contained in the conclusion ('precision'); ii) $|a \cap c|/|c|$ measures the relative amount of conclusion content contained in the premise ('recall'); iii) the harmonic mean of i) and ii) corresponds to main metric $f(a, c)$; and features iv-vi apply a non-linear function to i-iii), measuring the proximity to the feature means\(^{10}\), which expresses the idea that a conclusion that is both novel and justified may be situated at mean similarity of premise and conclusion, measured by $f(a, c)$.

We use a Linear SVM for predicting, in three binary classification tasks, either justification; novelty or both, using the feature set i)-vi).\(^{11}\) Results are seen in Table 3. Despite the small training data, performance is good for predicting justified (max. 68.6 F1) or novel (max. 70.0 F1). But predicting a conclusion to be novel & justified yields substantially lower performance (max. 58.3 F1), while still above baseline. Feature correlations show that novel is negatively (-) associated with $f(a, c)$ (i-iii), while justified is positively (+) correlated with $f(a, c)$ (i-iii). We find much weaker correlation for novel&justified, tending to mean similarity (iv-vi).

\(^{10}\)I.e., given the mean $\mu$ of a feature $x$, the new value $x'_i$ of datum $i$ is $x'_i = 1 - (\mu - x_i)^2$.

\(^{11}\)We average all results over 25 runs of leave-one-out cross validations. When predicting either justification, or novelty, we average over the two annotators; when predicting justification and novelty, to increase the positive class labels slightly, the gold target are cases where one or two annotators annotated both novel and justified.

\(^9\)AMR metrics have been previously used in NLG evaluation by Opitz and Frank (2021).
Table 3: Macro F1 scores for predicted conclusion quality using AMR-based models $f(a, c)$, assessing various aspects. For single features, + show positive correlation; - negative correlation (levels 0.05, 0.005, 0.0005).

| feat. id | random | justified | novel | both |
|----------|--------|-----------|-------|------|
| i) $|a \cap c|/|a|$ | 59.0 ++ | 58.7 -- | 53.4 |
| ii) $|a \cap c|/|c|$ | 68.6 +++ | 64.3 --- | 52.4 |
| iii) harm. mean i, ii) | 61.3 +++ | 61.9 --- | 52.2 |
| iv) proximity to mean i | 49.8 | 55.1 -- | 56.5 + |
| v) proximity to mean ii | 35.9 | 52.0 | 58.3 + |
| vi) proximity to mean iii | 30.5 | 54.4 - | 58.1 + |

Table 4: Predictors of conclusion usefulness.

| feat. id | random | justified | novel | both |
|----------|--------|-----------|-------|------|
| f(a) | | | | |
| f(a', c) | | | | |
| f(a, a') - f(c, c')/2 | | | | |
| human | | | | |

Table 5: AMR metrics detecting dissimilar arguments.

| similarity (SIM) features | feat. id | random | justified | novel | both |
|---------------------------|----------|--------|-----------|-------|------|
| $f(a)$ | | | | |
| $f(a', c)$ | | | | |
| $f(a, a') - f(c, c')/2$ | | | | |
| human | | | | |

Our analyses support Hyp1 in that AMR metrics are able to rate similarity of arguments, of conclusions and of argument-conclusion pairs, and this also allows us to determine if a conclusion is novel or justified. While many justified conclusions are highly similar to the premise, deciding their justification is difficult if they involve novelty. We argue this is because justification cannot be determined from premises alone, but requires external knowledge. We leave this issue for future work.

6.5 Conclusion usefulness

Finally, we revisit our Hyp2, that by extending arguments with inferred conclusions, we can support assessment of argument similarity. This raises the issue of the usefulness of a conclusion, in terms of achieving good performance and interpretability of an argument similarity method. The aspect of the usefulness of a conclusion clearly differs from the question of its quality. For one, it is possible that a good conclusion is not useful for argument similarity rating, simply because the assessment of the paired argument premises already provides a confident and precise similarity judgement. On the other hand, a mediocre conclusion could provide complementary indications that can support the similarity judgement. In this final section we aim to assess factors that can determine this usefulness.

Operationalizing conclusion usefulness We define a score $U$ for the usefulness of a conclusion, based on a human rating $y$, the conclusion similarity $f(c, c')$ and argument similarity $f(a, a')$, as

$$U = \frac{1}{1 + \left( y - f(a, a') \right)^2} \cdot \left( y - f(a, a') \right)^2, \quad (2)$$

where $U$ is maximized iff the automatic similarity rating of the conclusions does not differ from the human rating, while the automatic similarity rating of the premises differs maximally from the human rating. It is in exactly these situations that a conclusion assessment will prove most useful.

Features for assessing conclusion usefulness $U$

We assume the following features for modeling the usefulness of a conclusion, which we compute with our similarity function $f$: i) the similarity of the arguments $f(a, a')$; ii) the similarity of the conclusions $f(c, c')$; iii) the (signed) difference between the argument and the conclusion similarities $f(a, a') - f(c, c')$; iv) we compute the (signed) difference between the similarity of $(a, c)$ and $(a', c')$: $f(a, c) - f(a', c')/2$, finally, v) $y$ is the human rating.

Results Table 4 shows that the highest predictive power for conclusion usefulness is feature iii): the similarity of the two arguments minus the similarity of the two conclusions. It exhibits a highly significant positive correlation with conclusion usefulness, and relates to the following scenario: If two arguments are considered to be similar, but the conclusions as dissimilar, this may signal that the arguments are rated dissimilar by the human, and the high initial rating may be reconsidered.

Table 5 shows a data sample where the conclusions help to correct an initial, over-optimistic similarity rating of the premises. The premises are rated dissimilar by the human, but since they contain similar concepts, such as saving lives, the AMR metric assigns a high similarity rating (0.7) to the pair $(a, a')$. However, the automatically generated conclusions $(c, c')$ are assigned low(er) similarity (0.2). The low rating can be explained by the fact that the conclusion generator has distilled different conclusions from the premises that reflect the dif-
ferent foci of the arguments: the first proposes that organ donations are good for saving lives, while the second argument proposes that generally more medical research should be conducted.

7 Discussion

Explanation dimensions Our argument similarity rating approach may provide explanations in various dimensions. i) First and foremost, the explicit alignment and similarity computation based on AMR and AMR graph metrics, by relating similar concepts between arguments and their conclusions, provides insight into which components of two argument AMRs relate to each other, with individual alignment scores, and to what extent they contribute to the overall score. Especially in light of recent observations showing supervised models to be prone to superficial cues in data sets (Opitz and Frank, 2019; Niven and Kao, 2019; Heinzerling, 2020; Jo et al., 2021), this property is desirable. ii) We apply the fine-grained AMR decomposition of Damonte et al. (2017) in terms of semantic phenomena, such as negation or semantic roles. This can further illuminate in which ways an argument pair is similar/dissimilar. iii) By taking into account the similarity of automatically inferred conclusions, the similarity computed for premises may be re-adjusted in case the similarity of the inferred conclusions strongly differs.

On a related note, the AMR similarity statistics enabled us to gain some first indications of what could be considered a good conclusion (without even matching against a reference): e.g., our qualitative evaluations indicate that good conclusions tend to be neither very similar, nor very dissimilar to the premise. This seems plausible, since (too) high similarity may indicate a mere summary (reducing novelty), while (too) low similarity may indicate a lack of coherence (reducing validity).

Perspectives of improvement and future work A key component of our approach that influences all aspects of explainability described above, are the similarity metrics computed over AMRs. While we proposed one variant of $S^2$MATCH that focuses specifically on the similarity of concepts, further variations could be explored. We may also consider more recent AMR metrics that measure meaning similarity via graph kernels (Opitz et al., 2021).

Our approach also hinges on the quality of the inferred conclusions. The conclusions we obtained are often either justified or novel, but less often satisfy both conditions. In addition, we find that the degree of novelty is often rather small, perhaps reflecting that the T5 generator was pre-trained on summarization data and hence may tend to produce inferences that are not novel, since novelty is not a common characteristics of a summary. On the positive side, our approach can be fueled by an increasing amount of research on argument conclusion generation (Alshomary et al., 2020, 2021). In general, and particularly for our approach, it will be interesting to work with systems that produce not only a single, but multiple valid conclusions. Considering relations across and within two conclusion sets inferred from two premises may provide key information on argument similarity.

Finally, by measuring the similarity of premises and their conclusions, our approach could shed light on another important question: how to assess novelty and justification of a conclusion without a reference? This is an important question for research on argument conclusion generation since it lacks methods that can judge the quality of conclusions in the absence of (costly) references.

8 Conclusion

In this paper, we investigated two hypotheses: i) AMR meaning representation and graph metrics help in assessing argument similarity, ii) automatically inferred conclusions can aid or reinforce the similarity assessment of arguments. We find solid evidence for the first hypothesis, especially when slightly adapting AMR metrics to focus more on concrete concepts found in arguments. We find weak evidence that supports the second hypothesis, i.e., metrics improve consistently, but by small margins, when they are allowed to additionally consider the AMRs of automatically inferred conclusions. We believe, however, that more substantial gains may be obtained in future work, by improving conclusion generation models such that they produce content that is both valid and novel. Finally, we have made first steps towards a reference-less metric for assessing novelty and justification of generated conclusions.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable comments. This work has been partially funded by the DFG through the project ACCEPT as part of the Priority Program “Robust Argumentation Machines” (SPP1999).
References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args. me corpus. In Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), pages 48–59. Springer.

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyyedeh Ziyaei, and Mina Gholabadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In Proceedings of the 4th Workshop on Argument Mining, pages 91–96, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7367–7374.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 224–233, Online. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4334–4345, Online. 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 to propbank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’05), pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Carlos Iván Chesnevar and Ana G Maguitman. 2004. Arguenet: An argument-based recommender system for solving web search queries. In 2004 2nd International IEEE Conference on Intelligent Systems. Proceedings (IEEE Cat. No. 04EX791), volume 1, pages 282–287. IEEE.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 536–546, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence, 77(2):321–357.

David M Godden and Douglas Walton. 2006. Argument from expert opinion as legal evidence: Critical questions and admissibility criteria of expert testimony in the american legal system. Ratio Juris, 19(3):261–286.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. Science Robotics, 4(37).

Benjamin Heinzerling. 2020. Nlp’s clever hans moment has arrived. Journal of Cognitive Science, 21(1):159–168.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard H. Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. CoRR, abs/2105.07571.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02).
Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. 2019. Exploiting Background Knowledge for Argumentative Relation Classification. In 2nd Conference on Language, Data and Knowledge (LDK 2019), volume 70 of OpenAccess Series in Informatics (OASIcs), pages 8:1–8:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2021. Scientia potentia est – on the role of knowledge in computational argumentation.

John Lawrence. 2021. Explainable argument mining. Ph.D. thesis, University of Dundee.

Mirko Lenz, Stefan Ollinger, Premtim Sahitaj, and Ralph Bergmann. 2019. Semantic textual similarity measures for case-based retrieval of argument graphs. In International Conference on Case-Based Reasoning, pages 219–234. Springer.

Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an argument mining pipeline transforming texts to argument graphs. Computational Models of Argument: Proceedings of COMMA 2020, 326:263.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In Proceedings of the 5th Workshop on Argument Mining, pages 57–67, Brussels, Belgium. Association for Computational Linguistics.

Humberto R Maturana. 1988. Reality: The search for objectivity or the quest for a compelling argument. The Irish journal of psychology, 9(1):25–82.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 1813–1817, Istanbul, Turkey. European Language Resources Association (ELRA).

Nadezhda Mkrtuchyan, Evgeny Blagovechtchenski, Dinana Kurmakueva, Daria Gnedyk, Svetlana Kostromina, and Yury Shytrov. 2019. Concrete vs. abstract semantics: from mental representations to functional brain mapping. Frontiers in human neuroscience, 13:267.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Juri Opitz. 2019. Argumentative relation classification as plausibility ranking. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, pages 193–202, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-leman in the bamboo: Novel amr graph metrics and a benchmark for amr graph similarity. arXiv preprint arXiv:2108.11949.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In Proceedings of the 6th Workshop on Argument Mining, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1504–1518, Online. Association for Computational Linguistics.

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. Transactions of the Association for Computational Linguistics, 8:522–538.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative relation classification with background knowledge. In Computational Models of Argument, pages 319–330. IOS Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

A Rago, O Cocarascu, C Bechliulianidis, D Lagnado, and F Toni. 2021. Argumentative explanations for interactive recommendations. Artificial Intelligence, 296:1–22.
Whenever we encounter multiple premises or supportive claims of a single claim, we concatenate them in document order.