Target Inference in Argument Conclusion Generation

Milad Alshomary
Paderborn University
Department of Computer Science
milad.alshomary@mail.upb.de

Shahbaz Syed
Leipzig University
Department of Computer Science
shahbaz.syed@uni-leipzig.de

Martin Potthast
Leipzig University
Department of Computer Science
martin.potthast@uni-leipzig.de

Henning Wachsmuth
Paderborn University
Department of Computer Science
henningw@upb.de

Abstract

In argumentation, people state premises to reason towards a conclusion. The conclusion conveys a stance towards some *target*, such as a concept or statement. Often, the conclusion remains implicit, though, since it is self-evident in a discussion or left out for rhetorical reasons. However, the conclusion is key to understanding an argument, and hence, to any application that processes argumentation. We thus study the question to what extent an argument's conclusion can be reconstructed from its premises. In particular, we argue here that a decisive step is to infer a conclusion's target, and we hypothesize that this target is related to the premises' targets. We develop two complementary target inference approaches: one ranks premise targets and selects the top-ranked target as the conclusion target, the other finds a new conclusion target in a learned embedding space using a triplet neural network. Our evaluation on corpora from two domains indicates that a hybrid of both approaches is best, outperforming several strong baselines. According to human annotators, we infer a reasonably adequate conclusion target in 89% of the cases.

1 Introduction

The conclusion (or claim) of a natural language argument conveys a pro or con stance towards some *target*, such as a controversial concept or statement (Bar-Haim et al., 2017). It is inferred from a set of premises. Conclusions are key to understanding arguments, and hence, critical for any downstream application that processes argumentation. The task of *identifying* conclusions has been studied intensively in the context of argument mining (Stab and Gurevych, 2014) and automatic essay assessment (Falakmasir et al., 2014). In genres other than essays, however, conclusions often remain implicit, since they are clear from the context of a discussion (Habernal and Gurevych, 2015) or hidden on purpose for rhetorical reasons, as is often the case in news editorials (Al Khatib et al., 2016). This alters the task entirely to become a *synthesis* task: Given an argument's premises, generate its conclusion.

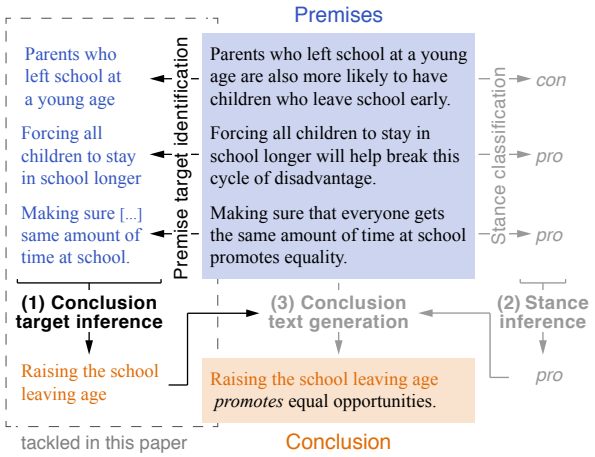

Figure 1: Illustration of our full model of generating an argument's conclusion from its premises. This paper focuses on the identification and inference of targets.

As detailed in Section 2, research on argumentation synthesis is still limited. Existing approaches focus on generating single claims (Bilu and Slonim, 2016), new arguments (Reisert et al., 2015), counterarguments (Hua et al., 2019), or argumentative texts (Wachsmuth et al., 2018). Closer to conclusion generation, Egan et al. (2016) summarized the main points of online debates, and Wang and Ling (2016) worked on identifying the main claim of an argument through abstractive summarization. To our knowledge, however, no approach so far reconstructs an argument's conclusion from its premises.

In general, we consider the synthesis task outlined above. Conceptually, we decompose this task into three steps, as depicted in Figure 1: (1) inferring the conclusion's target from the premises, (2) inferring the conclusion's stance, and (3) generating the conclusion's text with the inferred stance

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4334–4345

and the inferred target. In this paper, we focus on the first step by proposing two computational approaches for conclusion target inference.

As sketched in Figure 1, we hypothesize that the conclusion target is related to the targets of the argument's premises. To obtain premise targets, we train a state-of-the-art sequence labeling model (Akbik et al., 2018) on target-annotated claims (Bar-Haim et al., 2017). Since the exact relation of premise and conclusion targets is unknown, we develop two complementary inference approaches: One approach ranks premise targets based on their likelihood of being a conclusion target. The other one employs a triplet neural network (Hoffer and Ailon, 2015) that generates a conclusion target embedding from the premise targets in a learned target embedding space. A unique facet of the latter is the integration of the network with a knowledge base of targets (built from any training set), namely, the approach returns the known target whose embedding is closest to the generated embedding.

We compare the approaches against several baselines, including an existing sequence-to-sequence model for argument summarization (with and without encoded premise targets). For evaluation purposes, we study argument corpora from two genres where the correct conclusions are given: student essays (Stab and Gurevych, 2014) and debate portals (Wang and Ling, 2016). On these corpora, we empirically test how often an inferred target matches the target found in the ground-truth conclusion. Moreover, we let human annotators manually check the adequacy of the inferred targets.

In our experiments, both approaches consistently outperform sequence-to-sequence generation, justifying the explicit modeling of the relation between premise and conclusion targets. According to manual evaluation, a combined version of the two approaches infers an at least somewhat adequate target in 89%, and a *fully* adequate target in 55% of the cases, indicating the practical applicability of our target inference in conclusion generation.

In summary, the contributions of this paper are:[1]

1. A conceptual model of the task of generating an argument's conclusion from its premises.
2. Two complementary approaches that infer a conclusion's target from premises effectively.
3. Empirical evidence for the importance of modeling targets in conclusion generation.

[1]Resources: https://webis.de/publications.html?q=ACL+2020
Code base: https://github.com/webis-de/ACL-20

2 Related Work

Arguments have been modeled in different ways, focusing on the roles of their components (Toulmin, 1958), their inference scheme (Walton et al., 2008), or the interplay between their pro and con components (Freeman, 2011). On an abstract level, the models all share that they consider an argument as a *conclusion* (in terms of a claim) and a set of *premises* (reasons to support or object the claim). We restrict our view to this abstract model here.

Even though this paper is about inferring conclusion targets, our ultimate goal is to reconstruct the whole *conclusion* of an argument. Computational approaches to identify conclusions in a text have been pioneered research on student essay assessment (Burstein and Marcu, 2003). Falakmasir et al. (2014) show the importance of essay conclusions in applications, whereas Jabbari et al. (2016) specifically target an essay's overall conclusion, i.e., its *thesis* (also known as major, main, or central claim). Given the importance of theses, we dedicate one experiment particularly targeting them below.

The classification of argument components (as theses, conclusions, premises, etc.) is a core task in argument mining (Stede and Schneider, 2018) and has been approached for different genres (Stab and Gurevych, 2014; Peldszus and Stede, 2015). As Habernal and Gurevych (2015) observe, though, real-world arguments often leave the conclusion implicit, particularly where it is clear in the context of a discussion. In genres such as news editorials, conclusions may even be left out on purpose, in order to persuade readers in a "hidden" manner (Al Khatib et al., 2016). If an implicit conclusion is needed, it hence needs to be synthesized.

Argumentation synthesis research is on the rise. Early argument generation approaches relied on rule-based discourse planning techniques (Zukerman et al., 2000). Later, Reisert et al. (2015) generalized target-stance relations from claims and used them to automatically create new arguments. The relations were curated manually, though. An approach that finds the best conclusion for generation among a set of candidate claims was presented by Yanase et al. (2015). Sato et al. (2015) built upon this approach to phrase texts with multiple arguments. Others recycled targets and predicates of claims in new claims (Bilu and Slonim, 2016), generated arguments with specific inference schemes for user-defined content (Green, 2017), modeled rhetorical aspects in synthesis (Wachsmuth et al.,

2018), and composed arguments that follow a strategy (El Baff et al., 2019). All these methods synthesize *new* argumentative content. In contrast, we aim for the missing components of given arguments.

As such, our task resembles enthymeme reconstruction. An enthymeme is an implicit premise, usually the *warrant* (or *major premise*) that clarifies how a conclusion is inferred from the given premises (Walton et al., 2008). Motivated by the importance of finding the thesis, Boltuzic and Šnajder (2016) study how to identify such enthymemes given the other components. Similarly, Habernal et al. (2018) present the task of identifying the correct warrant from two options, and Rajendran et al. (2016) aim to generate the premise connecting an aspect-related opinion to an overall opinion. Instead of missing premises, we aim to synthesize (parts of) an argument's *conclusion*.

For any text generation task, a candidate technique is sequence-to-sequence models (Sutskever et al., 2014). Relevant in the given context, Hua and Wang (2018) used such models to generate counter-arguments, and Hua et al. (2019) extended this approach by planning and retrieval mechanisms. With a comparable intention, Chen et al. (2018) modified the bias of news headlines from right-to-left or vice versa. Closest to our work is the approach of Wang and Ling (2016) whose sequence-to-sequence model generates summaries for opinionated and argumentative text. Like us, the authors face the problem of varying numbers of input components, and tackle this using an importance-based sampling method. For their evaluation, they crawled arguments from *idebate.org*. We use this dataset in our experiments. Unfortunately, their manual evaluation considers opinionated text only, leaving the semantic adequacy of the generated argument summaries unclear.

The exact connection to summarization is unclear, which is why we include an approximation of the model of Wang and Ling (2016) as a baseline in our experiments. General research on summarization is manifold and beyond the scope of this work. For a survey, we refer the reader to Gambhir and Gupta (2017). In recent work, we summarize the core of an argument to be used as a snippet in the context of argument search by a two-sentence extract (Alshomary et al., 2020) and Egan et al. (2016) create abstractive summaries of the main points in a debate. We hypothesize a dependency between the target and stance of a conclusion and those of the premises. At a high level, this resembles the work of Angelidis and Lapata (2018) where aspects and sentiments are modeled for the extractive summarization of opinions.

We focus on the inference of conclusion targets in this work. Our approach builds upon ideas of Bar-Haim et al. (2017), who classify the stance of premises to a conclusion. To do so, they identify and relate targets in these components, and model stance with sentiment. We do not explicitly tackle stance inference here, because our focus is a conclusion's *target*. To identify premise targets, we first train a state-of-the art sequence tagger using contextualized word embeddings (Akbik et al., 2018) on the corpus of Bar-Haim et al. (2017). From these premise targets, we then infer the conclusion target, as explained below.

3 Data

Before discussing our target inference approach in Section 4, this section briefly introduces the datasets that we use in our analyses and experiments. To allow for evaluating the given task, the conclusion is always given in these datasets.

3.1 Wikipedia Claims with Targets

The *Claim Stance Dataset* (Bar-Haim et al., 2017) contains 2,394 claims referring to 55 topics from Wikipedia articles. Not only the stance of premises towards their topics is manually annotated, also a phrase is marked in each claim as being a target. We use this dataset to train and evaluate a target phrase tagging model for the purpose of *identifying* targets in the given premises of an argument. As Bar-Haim et al., we take all premises associated to 25 conclusions for training and the rest for testing.

3.2 Debate Portal Conclusions

The *iDebate Dataset* (Wang and Ling, 2016) consists of 2,259 pro and con points for 676 controversial issues from the online debate portal *idebate.org*. Each point comes with a one-sentence conclusion (called *central claim* by the authors) and an argumentative text supporting the conclusion. Each sentence is seen as one premise of the conclusion (called *argument*), resulting in a total of 17,359 premises. We use this dataset for training, optimizing, and evaluating all approaches to conclusion target inference. Following its authors, we split the dataset based on debates: 450 debates for training, 67 for validation, and 150 for testing.

3.3 Essay Theses and Conclusions

The *Argument Annotated Essays* corpus (Version 2; Stab and Gurevych (2014)) includes 402 persuasive student essays. Each essay was segmented manually into subsentence-level argument components: theses (called *major claims*), conclusions (*claims*), and premises. We use this corpus to study target inference in a second domain. To analyze different types of argument relations, we derive two datasets from the corpus: *Essay Conclusions* for conclusions and their premises with 1,530 training, 256 validation, and 234 test cases, and *Essay Theses* for theses and the underlying conclusions with 300 training, 50 validation, and 52 test cases.

4 Approach

We now present our approach to infer the target of an argument's conclusion from its premises. Based on a premise target identifier, it employs two complementary sub-approaches: One ranks premise targets by their potential representativeness for the (later unknown) conclusion, and then picks the top-ranked premise target. The other predicts candidate embeddings for the conclusion target from the top-ranked premise targets, and then picks the conclusion target from a knowledge base of targets whose embedding is most similar to those embeddings.

4.1 Premise Target Identification

To model the relation between premises and conclusion target, we first identify the premises' targets. The task of identifying target phrases in argumentative text has been introduced by Bar-Haim et al. (2017). We here tackle it as BIO sequence labeling, classifying each token as being the beginning, inside, or outside of a target. Since premise target identification is not our main focus, we simply train a state-of-the-art neural sequence tagger (Akbik et al., 2018) on the claim stance dataset and then use it to automatically annotate targets in all input premises.[2]

4.2 Inference by Premise Target Ranking

A reasonable hypothesis is that one of the premise targets of an argument represents an adequate conclusion target. Our first sub-approach thus simplifies the given task into selecting the premise target that most likely represents the conclusion target.

[2]Despite domain differences to the other datasets, we see in Section 5 that the tagger works rather reliably across datasets.

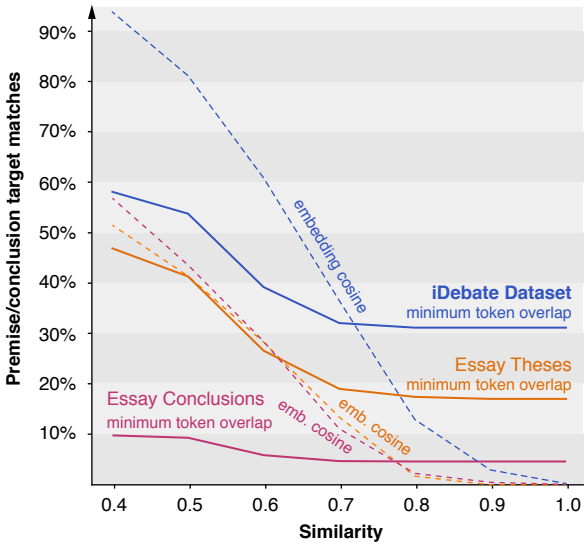

Figure 2: Percentage of training arguments in the given datasets where the conclusion target matches any of the premise targets, assuming a match when either a certain *minimum token overlap* (solid lines) or some *embedding cosine* similarity (dashed lines) is given.

Since there is no training data that reflects this likelihood, we follow the idea of importance sampling of Wang and Ling (2016): Given the output of our target identifier on a training instance, we use the percentage of content tokens overlapping between premise targets and the conclusion target as a representativeness label (quantified as Jaccard distance). Then, we learn a ranking model to predict the representativeness of a candidate premise target based on four features:

1. The average *embedding cosine similarity* of the candidate to the other candidates,
2. the *number of words* in the candidate,
3. the relative start and end character *position of the candidate* in the covering premise, and
4. the *number of sentiment words* (positive, negative, and neutral) in that premise.

The input of the ranking model are premise targets grouped by argument. During training, a probability is learned to reflect the ordering between each pair of premise targets in an argument with respect to conclusion target representativeness. Then, the model utilizes a cross-entropy loss function to minimize the difference between learned and the desired probability.

The effectiveness of this approach is naturally limited by the percentage of cases where the conclusion target actually matches any premise target.

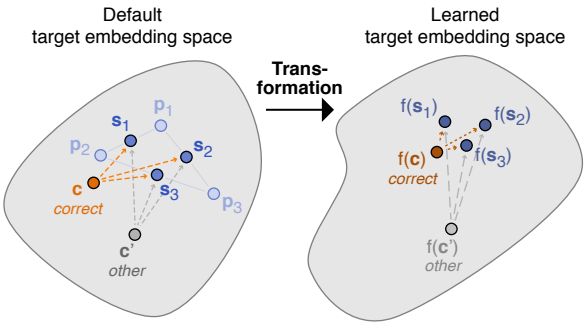

Figure 3: Sketch of the target embedding space transformation. The distance from the averages $\mathbf{s}_1, \mathbf{s}_2, \ldots$ of the premise targets to the correct conclusion target $\mathbf{c}$ is minimized, the distance to other targets $\mathbf{c}'$ maximized.

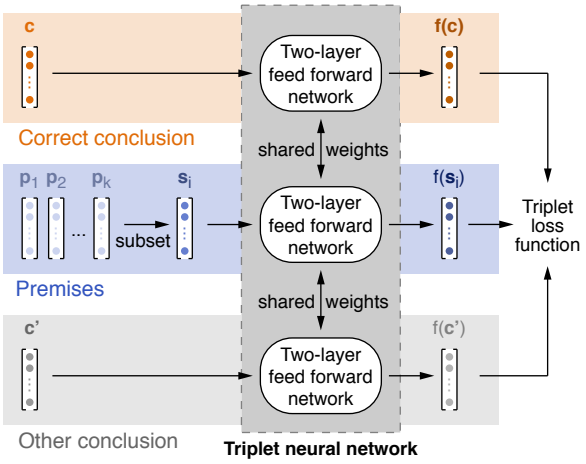

Figure 4: Our approach to learn conclusion target embeddings. The triplet neural network makes the average embedding $\mathbf{s}$ of a subset of the premise targets similar to the correct embedding $\mathbf{c}$, and dissimilar to others.

For a rough estimation, Figure 2 shows, based on two different similarity measures, how often at least one premise target matches the conclusion target in the three given training sets. Naturally, it is unclear in general how high the similarity needs to be for actual semantic equivalence.

4.3 Inference by Target Embedding Learning

To overcome the outlined shortcoming of being restricted to premise targets, we investigate a second hypothesis: An adequate conclusion target can be found in other arguments. To this end, we integrate a neural model with a knowledge base of targets in a novel way.

In particular, our second sub-approach tackles the given task by producing candidate conclusion target embeddings from the (top-ranked) premise targets, and then picking the target from a knowledge base whose embedding is most similar to the candidates. In principle, the knowledge base can be built from any corpus of argumentative texts based on our target identifier. In our experiments, we simply use all conclusion targets extracted from the training split of the datasets.

Now, to predict a conclusion target embedding, we first get the top $k > 1$ premise targets using our ranking approach and create average embeddings $\mathbf{s}_1, \mathbf{s}_2, \ldots$ of all $\binom{k}{m}$ possible subsets of these targets with $m > 1$. Then, we learn a function f on training arguments that maps each $\mathbf{s}_i$ to a transformed embedding space where it resembles the correct conclusion target $\mathbf{c}$ and differs more from other targets $\mathbf{c}'$. Figure 3 sketches this idea. The best k and m are found by tuning in validation.

As depicted in Figure 4, we model f as a *triplet neural network* (Hoffer and Ailon, 2015) with three vectors as an input: an anchor $\mathbf{s}_i$, a positive $\mathbf{c}$, and a negative $\mathbf{c}'$, where $\mathbf{c}'$ is a randomly sampled target from the target knowledge base. During training, we create $\binom{k}{m}$ triplets from each argument. Based on these, we utilize the following triplet loss function to minimize the cosine distance d between $\mathbf{s}_i$ and $\mathbf{c}$, and to maximize d between $\mathbf{s}_i$ and $\mathbf{c}'$:

$$\max\{d(f(\mathbf{s}_i), f(\mathbf{c})) - d(f(\mathbf{s}_i), f(\mathbf{c}')) + d_{max}, 0\}$$

Here, d_{max} represents the maximum distance to be considered, also determined during validation.

During prediction, we employ the trained network to map the average embeddings $\mathbf{s}_1, \mathbf{s}_2, \ldots$ of all premise target subsets to the transformed embedding space, and compute the average $avg(f(\mathbf{s}_i))$ of all mapped embeddings $f(\mathbf{s}_i)$. Then, we pick the conclusion target $\mathbf{c}$ from the knowledge base whose mapped embedding $f(\mathbf{c})$ has the minimum cosine distance to $avg(f(\mathbf{s}_i))$. This way, we ensure that we always end up with a meaningful target. Figure 5 sketches the conclusion target inference on the left and exemplifies it on the right.

4.4 A Hybrid of Both Sub-Approaches

The reasonableness of the conclusion target inferred by the second sub-approach depends on the quality of the knowledge base. To avoid inferring fully unrelated targets, we also consider a simple hybrid of our two approaches below: If the target inferred by the embedding learning approach overlaps with the (full) text of any premise in at least one content token, it is taken. Otherwise, the target inferred by the premise ranking is taken. More elaborated heuristics are left to future work.

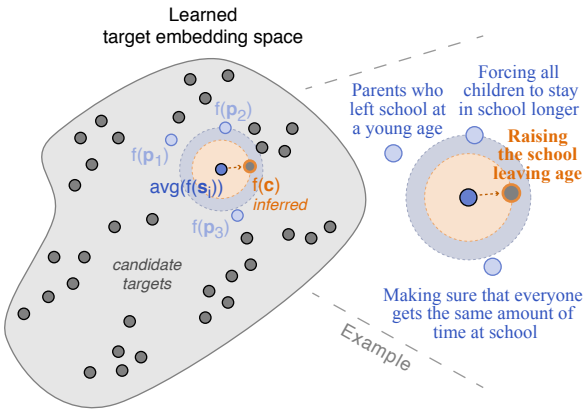

Figure 5: Sketch of inferring a conclusion target from an argument's premises. Given a knowledge base of candidates, the target is chosen whose learned embedding $f(\mathbf{c})$ is closest to the learned average $avg(f(\mathbf{s}_i))$ of premise targets. An example is shown on the right.

5 Automatic Evaluation

In this section, we report on empirical experiments, along with their results, performed to evaluate our approaches to target inference.

5.1 Premise Target Identification

We implemented the target identifier as a BiLSTM-CRF with hidden layer size 256, using the pretrained contextual string embedding model of Akbik et al. (2018). We trained the model on the training set of the Claim Stance Dataset with batch size 16 and a learning rate of 0.1 for five epochs.

Results On the Claim Stance test set, the identifier achieved an F1-score of 0.77. To assess its effectiveness in other domains, we let human annotators evaluate the identified targets of a random sample of 100 conclusions from the iDebate dataset. Each instance was evaluated by three annotators. Based on the majority agreement, the tagger identified 72% of the cases correctly.[3]

5.2 Conclusion Target Inference

To evaluate target inference, we use the iDebate Dataset and the two essay datasets. As no ground-truth conclusion targets are provided, we used our target identifier to extract targets from the conclusions and compared them to the output of our approaches. In some cases, particularly where targets were not explicitly phrased, our target identifier did not annotate any token. Hence, we eliminated those cases from the test set.[4]

Approaches For the premise target ranking approach, we trained LambdaMART (Burges, 2010) on each training set with 1000 estimators and a learning rate of 0.02. We refer to this approach below as *Premise Targets (ranking)*.

For target embedding learning, we used the pretrained FastText embeddings with 300 dimensions (Bojanowski et al., 2017) to initially represent each target. To obtain a knowledge base of candidate targets, we applied the target identifier to all conclusions of all training sets.[5] The resulting lexicon contains 1,780 targets, each is represented by its FastText embedding. We implemented the triplet neural network as three feed-forward neural networks, each with two layers and shared weights. We call this approach *Target Embedding (learning)*.

The simple hybrid of both approaches introduced above is denoted *Hybrid (ranking & embedding)*.

Baselines On one hand, we compare to the state-of-the-art sequence-to-sequence argument summarizer of Wang and Ling (2016). Since its code is not available, we approximately reimplemented it.[6] Specifically, we replicated the importance sampling with the same features (also on five premises) but no regularization. For generation, we used three LSTM layers with hidden size 150 and a pretrained embedding of size 300. Extra features of the original approach were left out, as they did not help much in our case. We trained the model with batch size 48 and learning rate 0.1 using the Adagrad optimizer (Duchi et al., 2011). For translation, we followed Wang and Ling. To identify targets in the generated summaries, we employed our target identifier. We refer to this baseline as *Seq2Seq*.

To test our hypothesis on the relation of premise and conclusion targets, we extended *Seq2Seq* by a pointer generator (See et al., 2017) and an extra binary feature that encodes whether a token belongs to a target or not, allowing the model to learn this relation. We call this *Seq2Seq (w/ premise targets)*.

On the other hand, we complemented our approaches with simpler variants, in order to check whether learning is needed. Instead of premise tar-

[3] In terms of Fleiss' κ, the agreement was 0.39, which is not high but still seems reasonable, given that we did not train annotators. Notice that this agreement value has *no* effect at all on the evaluation of our target inference approaches below.

[4] Example conclusion where no target was identified: *"It makes it more difficult for extremists to organize and spread their message when blocked"*.

[5] More elaborated knowledge bases are left to future work.

[6] The authors did not respond to our requests.

#	Approach	Scenario	iDebate dataset			Essay Conclusions			Essay Theses		
			bleu	meteor	accur.	bleu	meteor	accur.	bleu	meteor	accur.
b1	Seq2Seq	–	0.7	0.01	0%	–	–	–	–	–	–
b2	Seq2Seq (w/ premise targets)	–	4.4	0.07	5%	–	–	–	–	–	–
b3	Premise Targets (random)	–	3.9	0.11	8%	2.2	0.09	3%	8.8	0.19	17%
b4	Target Embedding (average)	Optimistic	7.2	0.16	18%	**8.3**	0.12	**8%**	15.3	0.24	21%
		Pessimistic	6.4	0.15	17%	4.1	0.12	6%	15.3	0.24	21%
a1	Premise Targets (ranking)	–	9.7	**0.16**	17%	4.1	0.11	5%	17.3	0.25	24%
a2	Target Embedding (learning)	Optimistic	9.2	0.15	18%	**8.3**	0.12	**8%**	**27.9**	**0.29**	**27%**
a2		Pessimistic	7.2	0.13	16%	3.4	0.09	5%	13.6	0.23	21%
a1&a2	Hybrid (ranking & embedding)	Optimistic	**10.0***	**0.16**	**20%***	8.2	**0.13**	**8%**	**27.9**	**0.29**	**27%**
		Pessimistic	8.1	0.15	18%	3.4	0.10	5%	13.6	0.23	21%
	Oracle (theoretic upper bound)	Optimistic	94.3	0.85	100%	98.9	0.95	100%	98	0.90	100%
		Pessimistic	35.8	0.58	65%	34.2	0.59	49%	26	0.52	48%

Table 1: Effectiveness of the evaluated target inference approaches in terms of BLEU, METEOR, and accuracy on the test sets of the iDebate dataset and the two essay datasets. The best value in each column is marked bold. Values of *a1&a2* marked with * are significantly better than the best baseline *b4* at $p < 0.05$ (student t-test). The bottom rows show the effectiveness of an oracle that selects those conclusion targets, which maximize each score.

get ranking, our baseline *Premise Targets (random)* simply chooses a premise target randomly. Instead of target embedding learning, we simply pick the target from the target space whose embedding is most similar to the average premise target embedding, called *Target Embedding (average)*.

Measures We use two common complementary evaluation measures, BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007). BLEU counts n-gram matches (we include 1- and 2-grams) focusing on precision, while METEOR is recall-oriented. Following the idea of Figure 2, we also report accuracy, where a given target is correct if it has 50%+ content overlap with the ground truth.

Experiments We tuned all approaches on the respective validation sets, and then evaluated them on the test set. Since *Seq2Seq* requires much training data, we evaluated both variants on iDebate only.

Before the inference of *Target Embedding (learning)*, the corresponding premise targets were added to the knowledge base as candidates for a conclusion target. Below, we consider two scenarios, an *optimistic* and a *pessimistic* one: In the former, the ground-truth target is added to the knowledge base, in the latter not. The optimistic scenario thus reflects the effectiveness of the approach regardless of the limitations of the knowledge base.

Results Table 1 lists the results. Clearly, encoding premise targets into *Seq2Seq* boosts its effectiveness, indicating the importance of modeling premise targets. However, both *Seq2Seq* variants perform poorly compared to our approaches. While

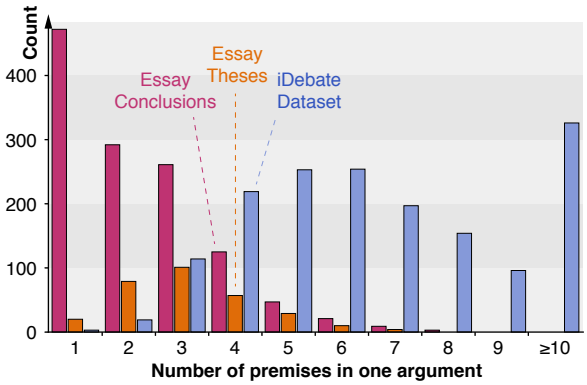

Figure 6: Histogram of the number of arguments with a specific number of premises in the three given datasets.

the limited training data size is one reason, this also indicates that pure sequence-to-sequence generation may not be enough.

On iDebate, both approaches are better than all baselines in terms of BLEU score. The best results are achieved by *Hybrid (ranking & embedding)* in terms of all measures (significantly for BLEU and accuracy). Even in the pessimistic scenario, its BLEU score of 8.1 outperforms all baselines.

In the optimistic scenario on the essay datasets, *Target Embedding (learning)* is strongest for most scores. The hybrid approach hardly achieves any improvement. Due to the small dataset size, no significance was found, though. In the pessimistic scenario, *Premise Target (ranking)* seems more suitable. The lower scores on Essay Conclusions can be attributed to the low number of premises (see Figure 6), which makes finding an adequate conclusion target among the premise targets less likely.

#	iDebate dataset		Essay Conclusions		Essay Theses	
	New	Exact	New	Exact	New	Exact
b4	5%	6%	3%	2%	0%	6%
a1	0%	**9%**	0%	1%	0%	9%
a2	**24%**	7%	**25%**	**6%**	**12%**	**12%**
a1&a2	9%	8%	15%	**6%**	**12%**	**12%**

Table 2: Percentage of test cases where each approach picked a *new* target (not a premise target) and where the picked target is an *exact* match of the ground-truth target. The highest value in each column is marked bold.

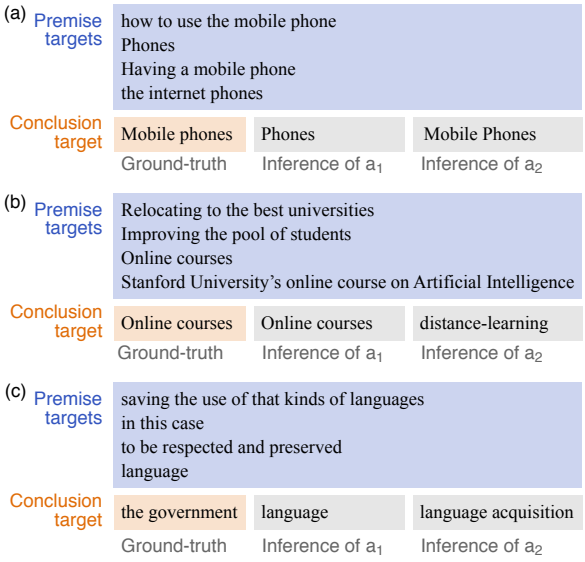

Figure 7: Three examples of premise targets from the datasets, the associated ground-truth conclusion target, and the conclusion targets inferred by our approaches.

As Table 1 shows, all approaches are much worse than theoretically possible (*oracle*) in terms of automatic metrics. However, the manual evaluation below reveals that the inferred conclusion targets actually compete with the ground truth.

Analysis To illustrate the behavior of selected approaches, Table 2 compares the percentages of cases where they pick a new target as well as where they pick the exact ground-truth conclusion target (in the optimistic scenario). Befittingly, target embedding learning (*a2*) is most "exploratory" regarding new targets. On the essay datasets, where the conclusion target only sometimes occurs in the premises, *a2* is also best in inferring the exact target. Still, premise target ranking (*a1*) may pick the ground truth, if it matches any premise target. The hybrid seems a suitable balance between both.

Figure 7(a) exemplifies the ability of *a2* to infer the correct conclusion target even if it does not match a premise target exactly. Example (b) stresses the limitation of automatic evaluation: "distance-learning" (inferred by *a2*) does not overlap with the ground truth, but it semantically matches well. In (c), the ground-truth target was barely inferable from the premise targets.[7]

#	Scenario	Fully	Somewhat	Not	Majority
b2	–	5%	18%	76%	93 / 100
a1	–	56%	33%	11%	91 / 100
a2	Optimistic	50%	28%	22%	92 / 100
	Pessimistic	49%	27%	24%	93 / 100
a1&a2	Optimistic	55%	34%	11%	89 / 100
	Pessimistic	56%	32%	12%	90 / 100
Ground-truth		62%	29%	10%	84 / 100

Table 3: Majority agreement for how adequate (*fully*, *somewhat*, *not*) are the conclusion targets of baseline *b2*, our approaches, and the ground truth. The right column lists the number of cases where majority is given.

6 Manual Evaluation

To assess the actual quality of the inferred conclusion targets, we manually evaluated our approaches (optimistic and pessimistic scenario) and the baseline *b2* (*Seq2Seq (w/ premise targets)*) in comparison to the ground-truth targets using Amazon Mechanical Turk. For this, we sampled 100 random instances from the iDebate test set. In a single task, an argument's premises were given along with the conclusion target of either approach. Annotators had to judge the adequacy of the target for the given premises as *fully*, *somewhat*, or *not* adequate. Each instance was judged by five annotators. No one judged multiple targets for the same argument.[8]

Table 3 shows the distribution of majority judgments for each approach. Only 23% of the *b2* targets were considered fully or somewhat adequate, i.e., pure text generation seems insufficient. In contrast, our sub-approaches' targets are competitive to the ground truth, which was not always adequate either (likely due to errors in target identification). The high performance of *a1* (*Premise Targets (ranked)*) might be explained by the inferred targets being part of the premises, affecting annotators' preferences. Still, the targets of *a2* (*Target Embedding (learning)*) are seen as adequate in 78% of the cases (50% fully), with the ability of infer-

[7] Full example arguments found in supplementary material.

[8] We paid $0.40 per task, restricting access to annotators with an approval rate of at least 95% and 5000 approved tasks. To ensure correct annotations, a reason had to be given.

ring conclusion targets that are not explicitly stated in the premises. Even in the pessimistic scenario, the inferences of *a1* and *a1&a2* remain stable.

7 Conclusion

An argument's conclusion comprises its stance towards the target it discusses. Still, the conclusion is often left implicit in real life, because it is clear for humans or hidden for rhetorical reasons. We have conceptualized the task of reconstructing the conclusion from the argument's premises as (1) inferring the conclusion's target, (2) inferring its stance, and (3) phrasing its actual text. Then we have focused on the first step in which we infer the conclusion target given a set of premises.

Hypothesizing that the conclusion target depends on the premise targets, we have developed two new and complementary target inference approaches: *Premise Targets (ranking)* returns the premise target that is most likely adequate for the conclusion, while *Target Embedding (learning)* generates a conclusion target embedding from the premises and matches it against a target knowledge base.

On three datasets from two domains (debate portals and student essays), our approaches outperform several baselines, including a state-of-the-art neural sequence-to-sequence summarizer. The latter also benefits from modeling premise targets, additionally supporting our hypothesis. In terms of BLEU, METEOR, and accuracy, *Target Embedding (learning)* and a hybrid of both approaches turned out particularly strong, whereas *Premise Targets (ranking)* was best in a manual evaluation. Overall, we manage to infer an at least somewhat adequate conclusion target in 89% of all cases, indicating the practical applicability of our approaches.

Combining target inference with stance classification in future work, we can already generate basic conclusions, say, "Raising the school leaving age is *good*". A more elaborate phrasing approach may take over context information from the premises.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) within the collaborative research center "On-The-Fly Computing" (SFB 901/3, project no. 160364472). We thank students from Paderborn University for evaluating our target identifier: Denis Kuchelev, Christin Löer, Natalie Lüke, Avishek Mishra, Enri Ozuni, René Scherf, Harsh Shah, Nikit Srivastava, Martin Wegmann.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. 2020. Extractive snippet generation for arguments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, to appear.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Filip Boltuzic and Jan Šnajder. 2016. Fill the gap! Analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133. Association for Computational Linguistics.

Christopher J.C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview.

Jill Burstein and Daniel Marcu. 2003. A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Mohammad Hassan Falakmasir, Kevin D Ashley, Christian D Schunn, and Diane J Litman. 2014. Identifying thesis and conclusion statements in student essays to scaffold peer review. In *International Conference on Intelligent Tutoring Systems*, pages 254–259. Springer.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Nancy L. Green. 2017. Argumentation scheme-based argument generation to support feedback in educational argument modeling systems. *International Journal of Artificial Intelligence in Education*, 27(3):515–533.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230. Association for Computational Linguistics.

Fattaneh Jabbari, Mohammad Hassan Falakmasir, and Kevin D. Ashley. 2016. Identifying thesis statements in student essays: The class imbalance challenge and resolution. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pages 220–225.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating Toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55. Association for Computational Linguistics.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and

Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. 2015. Learning sentence ordering for opinion generation of debate. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 94–103, Denver, CO. Association for Computational Linguistics.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Supplementary material: Example arguments with inferred conclusion targets

Example 1

Argument: Relocating to the best universities is a budgetary concern , but also family and social relations concern for many people , which prevents all the best people from even applying to universities that would suit them the best . Online courses can recruit students from anywhere in the world much easier than traditional universities can because students do n't need to travel far away for the best education . This then ensures that universities have better access to the brightest people . For instance , Stanford University 's online course on Artificial Intelligence enabled people from 190 countries to join , and none of students receiving a score of 100 percent where from Stanford -LSB- 14 -RSB- . Improving the pool of students would automatically result in better academics , professionals and science , which would benefit the society better .

Ground truth conclusion: Online courses are a way to higher academic excellence

Premise Targets (ranked): Online courses

Target Embedding (learning): distance-learning

Example 2

Argument: Having a mobile phone helps us to learn in a lot of different ways . First we learn about technology ; about how to use the mobile phone . Second most phones today have apps -LRB- programs -RRB- to enable learning using the phone , or else through the internet . Phones can access online courses and lessons which can be provided in fun ways and can in some cases instantly tell you if you have the right answer . It may even sometimes be possible to do homework on a phone and send it to your teacher . Even without the internet phones can be used to provide short assignments , or to provide reminders to study .

Ground truth conclusion: Mobile phones help us to learn

Premise Targets (ranked): Phones

Target Embedding (learning): Mobile phones

Example 3

Argument: students who used to prepare Microsoft PowerPoint presentation for their school projects , get an edge over others at an early stage of their career When children are allowed to play around with computer from a very early age , they get acquainted with the previously mentioned skills and become expert before facing professional world computers enable people to prepare presentations , draw complex graphs and pictures , document thesis in a simple though efficient way

Ground truth conclusion: it 's clear that computer has a positive effect on the children

Premise Targets (ranked): students who used to prepare Microsoft PowerPoint presentation for their school projects

Target Embedding (learning): future prospects of computers

Table 4: Example arguments chosen from the test dataset, where premise targets and the conclusion target are highlighted in each argument. Along with that, we show the conclusion targets inferred by our approaches.