Neural Argumentation Mining on Essays and Microtexts with Contextualized Word Embeddings

Anonymous ACL submission

Abstract

Detecting the argument components Claim and Premise is a central task in argumentation mining. Working with two annotated corpora from the genre of short argumentative texts, we extend a BiLSTM-CRF neural tagger to identify argumentative units and to classify their type (claim vs. premise). For the corpora we use, Persuasive Essays and Argumentative Microtexts, current methods relied on pre-computed non-contextual word embeddings such as Glove. In this paper, we adopt contextual word embeddings (Bert, RoBerta) and cast the problem as a sequence labeling task. We show that this step improves the state of the art for the Persuasive Essays, and we present strong initial results on applying the same approach to the Argumentative Microtexts.

1 Introduction

The task of finding argumentation structures in text has received increasing attention over the last years. In contrast to most other NLP problems, it is not a single, well-demarcated task but a constellation of subtasks, combinations of which can be employed for specific applications (Lippi and Torroni, 2016; Stede and Schneider, 2018). These subtasks are:

- Find argument components (ACs): Given a text, which spans correspond to argumentative material?
- Classify ACs: Does an AC constitute a claim being made, or a premise being given to support or undermine a claim?
- Detect relations among ACs: Various relations can hold between ACs; mostly, just *support* and *attack* are being distinguished.
- Build argumentation graph: Combine the results of the aforementioned subtasks into a well-formed graph structure representing the argumentation that is performed in the text. (Notice that argumentation can be recursive: Claim C is supported by premise E1, which is in turn supported by premise E2, so that E1 has two functions.)
- Classify argumentation schemes: Provide labels for the reasoning patterns underlying claim-evidence pairs.
- Argument quality: Work out various attributes for the arguments and/or relations, such as the strength of an argument, etc.

One view of thinking about argumentation mining is that of an extension of sentiment analysis. In a broad sense, sentiment analysis cares about “what people think about some entity X”, whereas argumentation mining extends this to the question “why people think Y about X”; thus it can unveil more complex reasoning processes rather than just detect opinions and sentiment.

In this paper, we concentrate on the ‘core’ subtasks that any application will need: Finding ACs in text, and labelling them as either *claim* or *premise*. This in line with the common definition of an argument (e.g., (van Eemeren and Grootendorst, 2004)) as consisting minimally of one claim and one statement of evidence, which we here call a premise.[1]

We will be using two datasets that have been among the earliest that were made available, and at the same time are among the most “deeply” annotated, in the sense that full argumentation graphs are provided. These are the persuasive essay (PE) corpus by (Stab and Gurevych, 2017) and the argumentative microtext (AMT) corpus by (Peldszus

[1] More generally, ‘premise’ covers statements that can either support or attack a claim. This distinction is subject to the relation classification, which we do not address in the present paper.

and Stede, 2016). As indicated, for the present purpose we use only the labeling of argument components as claim vs. evidence, though.

Our contributions are (i) we present new state-of-the-art results on argument component detection and type classification on the PE corpus; and (ii) we show the first results for mapping that analysis procedure to the AMT corpus, i.e., in a combined detection and classification task. (Previous research on AMT has so far started from gold-annotated components and focused on building complete tree structures.)

In the following, we first summarize the relevant related work (Section 2), and then describe the two corpora in more detail (Section 3). This is followed by a presentation of our experiments and results (Section 4) and conclusions (Section 5).

2 Related Work

Both the PE and the AMT corpora have been used in a variety of approaches to argument mining tasks. Some have concentrated on subtasks that proceed from already-given argument components, which are then classified as claim or evidence (and afterwards, relations are built). This holds for (Peldszus and Stede, 2015), (Potash et al., 2017), and (Afantenos et al., 2018). The first end-to-end systems, comprising argument component identification as well as role and relation classification, were presented by (Persing and Ng, 2016) and (Stab and Gurevych, 2017) for the PE corpus, both using linguistic feature engineering, and ILP as optimization tool. Focusing on component and role identification (i.e., the task that we address here), the current state of the art results on the PE corpus were achieved by the neural systems of (Eger et al., 2017), who compared several DL approaches and found LSTM-ER most successful, and by (Chernodub et al., 2019), who used a BiLSTM-CNN-CRF. We will compare our own results to these in Section 4. Recently, (Wambsganss et al., 2020) used a similar technical setup as we do, but they focus solely on the identification of argument components (i.e., they do not distinguish claim and evidence), and thus their results are not directly comparable.

For the AMT corpus, all previous work that we are aware of has started from the argumentative discourse units (ADUs) given by the corpus annotation and then distinguished the types of argument components (Peldszus and Stede, 2015; Stab and Gurevych, 2017; Potash et al., 2017). By transferring our approach from PE to AMT, our experiments reported below are thus the first that include the argument component detection step, and hence we cannot compare our results to a previous state of the art.

Recent interesting work, which is not directly comparable to ours, was done by (Persing and Ng, 2020), who suggest an unsupervised approach for claim/evidence and relation labeling on the PE corpus, and (Alhindi and Ghosh, 2021), who employ BERT-based transfer learning on a new corpus of student essays.

3 Text Corpora

Persuasive Essays. The PE corpus consists of 402 argumentative essays (2235 Paragraphs) that were written by learners of English in response to a given prompt. (Stab and Gurevych, 2017) collected the essays from a website and provided annotations of argumentation graphs. Essays started with a question, and contain a claim and a constellation of evidence, possibly with substructure. Some sentences can be non-argumentative, as they merely provide background or elaborations of minor significance. In addition, for the whole text there is a main claim, usually located at the end of the text, and which is supported by the paragraph-level claims. In the interest of compatibility with other work, we here treat the types 'main claim' and 'claim' as equivalent and perform classification on paragraph level, i.e., the task is to label the ACs in each paragraph.

Argumentative Microtexts. The AMT corpus by (Peldszus and Stede, 2016) consists of 112 short texts (each of about 3–5 sentences) that have been labelled with full argumentation tree structures. Similar to PE, the AMT texts were written by students in response to a prompt. However, students wrote in their native language German, and the texts were later professionally translated to English. The annotations are very similar to those in PE, except that (i) there is no 'main claim' (instead, each text has one single claim), and (ii) AMT texts do not contain any non-argumentative material; in other words, the argumentation is "dense". We treat an AMT text as technically corresponding to a paragraph from a PE text.

Corpus Statistics. Table 1 provides information on the sizes of the Persuasive Essays and Microtext corpus. The train, development, and test splits

	Train	Dev	Test
PE	1587	199	449
MT	80	9	23

Table 1: Corpus statistics (number of paragraphs)

represent comparable proportions of the total, but overall the PE corpus is substantially larger.

4 Experiments and Results

We first describe the task of mapping the corpora to a common format, then explain our technical approach to claim/premise identification, and afterwards describe the experiment and its results.

PE Preprocessing. The corpus uses a token-oriented, tab-separated (CoNLL-like) format, whose two columns are the word (token) and its label. The label consists of a component type (Major-Claim, Claim and Premise). As stated above, we mapped 'Major Claim' to 'Claim', so for our task we have two labels for classification: Claim (C) and Premise (P). Overall, there are 2257 claims, and 3832 premises. In order to train using Flair[2], we used the spaCy toolkit [3] to add part-of-speech information, distribute the claim/premise classes to token-level BIO annotations, and then encode the PE data as a sequence of triples, $(Token, PoS, BIO)$.

AMT Preprocessing. The Argumentative Microtext corpus comes in an XML format, which we converted to the same format as that described above for PE. Overall, AMT has 112 claims (one for each paragraph), and 464 premises.

Approach. Following the approach of (Chernodub et al., 2019), we implement a BiLSTM-CRF neural tagger for identifying argumentative units and for classifying them as claims or premises. The BiLSTM-CRF method is a popular sequence tagging approach and achieves almost state-of-the-art performance for tasks like named entity recognition (NER). Further, we tested two versions of pre-computed contextual word embeddings; Bert (Devlin et al., 2018) and RoBERTa (Liu et al., 2019).

Experiment. We train on-the-fly in each training mini-batch. it means that embeddings would not get stored in memory. The advantage is that this keeps your memory requirements low. We apply the same experimental settings of the earlier research quoted above: a fixed 70/20/10 train/dev/test split on the PE, and we used the same distribution for AMT. The hyper-parameters were: Optimizer: SGD; learning rate: 0.1; dropout: 0.1; number of hidden units: 256.

Results. Table 2 shows a comparison of our best performing models on the Persuasive Essays dataset to the best results provided by the (Eger et al., 2017) and (Chernodub et al., 2019), as well as our results on AMT. On the PE corpus, Bert embeddings performed best and on AMT corpus RoBerta yields the best results. As the table shows, our approach on PE improves F1-score performance considerably from 0.645 reported by (Eger et al., 2017) to 0.715. Applying our approach using RoBerta on AMT gives 0.718 F1-score, which we consider promising. This result is, to best of our knowledge, the first that has been reported for this particular task on the AMT corpus.

Method	F1(PE)	F1(AMT)
STag (BiLSTM-CRF-CNN)	0.647	-
TARGER (using Glove)	0.645	-
Our Model (using Bert)	**0.715**	0.619
Our Model (using RoBerta)	0.675	**0.718**

Table 2: Comparison of our model performance (micro F1-Score) on PE, AMT to the best approaches from (Eger et al., 2017) and (Chernodub et al., 2019) on span level

5 Conclusion and outlook

Contextual word embeddings have been shown to yield state-of-the-art results for many NLP tasks, and in this paper we found that they also outperform previous work (using non-contextual embeddings) on identifying claims and premises in argumentative essays. For the Persuasive Essay corpus we were thus able to achieve a new state of the art for the combination of the two subtasks "detect argument components" and "classify argument components", which we implemented as one joint sequence-labeling task.

We argue that this joint task is in fact highly relevant for practical applications of argument mining on other genres as well: Given the customary definition of argument as a claim and at least one premise, these need to be identified and distinguished in running text, whether it is some social media contri-

[2]https://github.com/flairNLP/flair
[3]https://spacy.io

bution, a legal document, or a newspaper editorial. We thus think it is appropriate to apply this task also on the argumentative microtext corpus (Peldszus and Stede, 2016), which in previous work has been studied only by exploiting two simplifications: there is no non-argumentative material, and pre-annotated ADU boundaries are used – in other words, the detection of argument components has not been performed. For a realistic setting, these simplifications should be dropped, however. We therefore applied our approach also to the microtexts, even though we are solving a somewhat "inflated" problem: We classify claim/premise/other on texts that – somewhat artificially – do not contain any "other". Our results are, to our knowledge, the first that have been provided for this new perspective on the corpus.

Our next steps are: (i) We plan to add the step of relation identification, which is necessary for a more fine-grained representation of argumentation structure in texts that may contain multiple claims and/or recursive structures. (ii) We will further explore the issue of domain adaptation by experimenting with cross-domain train/test settings for the PE and AMT corpora, and possibly for an additional corpus.

References

Stergos Afantenos, Andreas Peldszus, and Manfred Stede. 2018. Comparing decoding mechanisms for parsing argumentative structures. *Argument & Computation*, 9(3):177–192.

Tariq Alhindi and Debanjan Ghosh. 2021. "sharks are not the threat humans are": Argument component segmentation in school student essays. arXiv:2103.04518.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova Tetreault. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computation and Language*, arXiv:1810.04805. Version 2.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press, Cambridge.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

M. Lippi and P. Torroni. 2016. Argumentation mining: State of the art and emerging trends. *JACMTransactions on Internet Technology (TOIT)*, 16(2):1–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pre-training approach. *Computation and Language*, arXiv:1907.11692.

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, Lisbon, Portugal. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–816, London. College Publications.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2020. Unsupervised argumentation mining in student essays. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6795–6803, Marseille, France. European Language Resources Association.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool, San Rafael, CA.

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020. Unlocking transfer learning in argumentation mining: A domain-independent modelling approach. In *Proceedings of the 15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.