Fine and Coarse Granular Argument Identification and Classification in Persuasive Essays

Exposé for the Module Research Case Studies (WS 2021)

Arturo Figueroa, Ashwin Pednekar, and Bilal Mehyar

Trier University, 54286 Trier, Germany
{s4arfigu, s4aspedn, s4bimehy}@uni-trier.de

1 Introduction

Argumentation is a verbal activity that aims at increasing or decreasing the acceptability of a controversial standpoint (Van Eemeren et al., 2013). Argumentation is necessary for humans to communicate thoughts in order to establish a point, persuade others or form logical reasoning in order to justify decision making (Peldszus and Stede, 2013; Stab and Gurevych, 2014; Habernal and Gurevych, 2017). Well-reasoned arguments are not only important for making justified decisions but also play a crucial role in drawing widely-accepted conclusions and deriving novel knowledge in epistemic activities. Moreover, the ability to develop well-reasoned arguments is a fundamental requirement for learning itself Davies (2009) because it requires critical thinking and the ability to reason rationally.

Argumentative writing is complex, consisting of several statements and objections stretching out over a long discourse in a chain of reasoning, sometimes depending on the genre. Since arguments can be so complex, it can be difficult to keep track of their development over an extended discourse, such as an essay, thesis, debate or legal case (Botley and Hakim, 2014). The structure of an *argument* consist of some *argument components*. These argument components are referred to in literature as Argumentative Discourse Units (ADUs) (Lawrence and Reed, 2020). An argument consists a claim and one or more premises (Govier, 2009). The **claim** is a controversial statement and the central component of an argument, and **premises** are reasons for justifying or refuting the claim. Moreover, arguments have directed argumentative relations, describing the relationships one component has with another. Each such relation indicates that the source component is either a justification for or a refutation of the target component.

Argument Mining, or sometimes referred to as Argumentation Mining, is the automatic identification and extraction of ADUs from unstructured natural language text (Lawrence and Reed, 2020). The field has expanded considerably in the recent years, and has seen applications in various domains such as legal decision support, policy making, debating technologies, computer-assisted writing and automated essay scoring (Stab and Gurevych, 2017; Nguyen and Litman, 2018). The task of Argument Mining can typically be decomposed into the following sub-tasks (Eger et al., 2017; Persing and Ng, 2016):

1. Identifying the locations of ADUs
2. Classification of ADUs into premises or claims
3. Identifying relations between ADUs
4. Classifying the relationship between the ADUs whether it is a supporting or attacking relationship.

This research focuses on the first two sub-tasks of Argument Mining, namely, *1. Identifying the locations of ADUs* and *2. Classification of ADUs into premises or claims*. The overall goals and contributions of this research would be to:

1. Study various approaches of segmenting text in order to identify the locations of ADUs, and find the best approach.
2. Provide an effective set of features and classification approach for classifying ADUs into claims and premises.
3. Provide a suitable measure of error for training and evaluating the model.

2 Foundations and Related Work

This section discusses the knowledge foundations and previous works related to the research.

2.1 Argument Mining in Persuasive Essays

Persuasive essays are used to assess the students' language proficiency, understanding of the subject matter and argumentation skills (Ghosh et al., 2016). Persuasive essays specifically ask students to write down their opinions and support them using convincing statements (Nguyen and Litman, 2015) and therefore contains explicit argumentative structures unlike other types of text (Trautmann et al., 2020). Some applications of Argument Mining include an automated essay scoring system (Nguyen and Litman, 2018) and an argumentative writing support tool that provides students feedback on their argumentation structure during writing (Wambsganss et al., 2020).

2.2 Unit Segmentation

A major challenge in the task of identifying the locations of ADUs is how to segment the text into meaningful units to be used for argumentation tasks (Trautmann et al., 2020; Stab et al., 2018). Peldszus and Stede (2013) suggest that the task of segmentation may be similar to the task of identifying Elementary Discourse Units (EDUs) which are atomic, non-overlapping units of text that carry a piece of information. There is still debate on what should be considered as EDUs, among notions such as clauses, sentences, phrases or turns of talk (Lawrence and Reed, 2020). Furthermore, it is also unclear how many EDUs should an ADU consist of (Peldszus and Stede, 2013). Many of previous works have relied on unit segmentation of text at sentence level (Stab and Gurevych, 2014; Nguyen and Litman, 2015; Ghosh et al., 2016; Nguyen and Litman, 2018; Wambsganss et al., 2020; Lenz et al., 2020). This approach is simple to perform but has the following disadvantages: (1) it produces units that are often longer than the actual ADUs (Lenz et al., 2020), and (2) it is unable to differentiate multiple ADUs in a single sentence, resulting many ADUs being grouped into one ADU (Trautmann et al., 2020).

An ADU should be grammatically correct in isolation (Stab and Gurevych, 2015). Adhering to this rule, it is suggested that all spans should be clauses (Trautmann et al., 2020). In rare cases, there are ADUs that are not clauses but noun phrases or clauses missing the subject, which usually is the topic of the claim itself (Trautmann et al., 2020).

Peldszus and Stede (2013) suggest segmenting the text as clause-based EDUs and then use some kind of relation to decide when it would be necessary to combine EDUs into an ADU.

Stab and Gurevych (2017) consider the problem of argument identification as a sequence labeling problem. They segment the text into of token level and perform classification on each token.

2.3 ADU Classification

Different types of corpora are inherently different and would require different approaches to identify arguments Teufel and Moens (2002). News articles may have arguments positioned at the beginning of the text whereas that is not the case for scientific articles (Teufel and Moens, 2002). In biomedical texts, the demonstrative pronoun "this" serves as a good indicator for claims since they are found following in semantic sequences such as *this method...* or *this conclusion...* (Houngbo and Mercer, 2014). Persuasive essays often lack section titles and headings unlike scientific or news articles, but they follow a specific prompt that requires students to state their claims and support them with premises (Nguyen and Litman, 2015). Online user comments on the other hand have arbitrary structure and often lack explicit arguments or contain claims with missing justifications (Park and Cardie, 2014).

Stab and Gurevych (2014) considered a multi-class classification problem in the context of persuasive essays, and have categorized features into groups as shown in Table 1. These categorization of features have been studied, utilized and further expanded by several authors in the field Nguyen and Litman (2015); Stab and Gurevych (2017); Aker et al. (2017); Habernal and Gurevych (2017); Wambsganss et al. (2020)

Structural features	Features related to the structure of the text such as token statistics, position of the unit in the paragraph, number of punctuation.
Syntactic features	Features extracted from grammatical rules and dependency trees, such as the number of sub-clauses and the depth of the parse tree, and the tense of the sentence.
Lexical features	Features relate to the semantics such as n-grams, existence of certain n-grams, certain adverbs or modals.
Indicators	Existence of specific words in a pre-defined list that frequently signal ADUs, including words like "therefore", "consequently", "because".
Contextual features	Features extracted from the text surrounding the unit, such as number of tokens, number of punctuation or existence of modal verbs.

Table 1: Categorization of Features (Source: Stab and Gurevych (2014))

Nguyen and Litman (2015) state that structural and syntactic features generally perform well on most types of texts. Aker et al. (2017) investigate the effectiveness of the feature categories proposed by previous works which includes the ones proposed by Stab and Gurevych (2014). Their results showed that for persuasive essays, structural features performed the best followed by lexical features. They claim that syntactic features do not perform well and have little significance in identification of ADUs.

Aside from the features introduced by Stab and Gurevych (2014), several authors Aker et al. (2017); Habernal and Gurevych (2017); Stab and Gurevych (2017); Wambsganss et al. (2020) use embeddings as an additional feature, since embeddings help to achieve state-of-the-art performance in several Natural Language Processing tasks (Habernal and Gurevych, 2017). Word embeddings (Mikolov et al., 2013b) are vector representations of words that capture syntactic and semantic relationships of a word. Word embeddings however have a limitation that it does not capture information of the word order and has difficulties representing phrases Mikolov et al. (2013b). Embeddings generated from variants of the BERT architecture have been applied with great success in recent works in the field of Argument Mining (Ein-Dor et al., 2020; Trautmann et al., 2020). One of the variants of BERT, Sentence-BERT (SBERT) is able to generate sentence embeddings, vectors representations that capture semantic relationships of whole sentences (Reimers and Gurevych, 2019).

Opitz and Frank (2019) categorize the features into three types. Content-ignorant features are derived from the context of the ADU span, i.e. from the leading and trailing tokens of the argumentative unit. Content-based features are derived from the ADU span itself. Lastly, full access features are a combination of both, and include the features that capture discourse structures between the boundaries of an ADU and it's surroundings.

Nguyen and Litman (2015) perform post-processing, using LDA to extract domain words specific to the corpus and use them to derive features for argument mining.

Song et al. (2014) performed argument mining on a corpus of argumentative essays using sentence segmentation. They employed an L2 penalized logistic regression binary classifier.

Levy et al. (2014) studied context-dependent claim detection; to detect claims in relation to a given topic of discourse. They present a funnel method, using less demanding classifiers (such as logistic regression) at the first stages of the model, passing on only sentences with high scores to more computationally demanding classifiers.

Lippi and Torroni (2015) argue that most methods rely on highly engineered features and are mostly domain-dependent. They argue that humans can distinguish that a sentence "sounds like" a claim even without knowing the context because argumentative sentences contain a common rhetorical structure and hence propose a method to exploit these structures in order to identify claims without knowing any context in advance. Results showed that some positive results actually contained a claim but they were labeled as negative examples in the corpus, due to the context dependent nature of the annotations. Mikolov et al. (2013a)

3 Research Questions

The challenge of unit segmentation is faced in the task of identifying the locations of ADUs. In the task of classifying ADUs, feature vectors need to be constructed to represent natural language text in a structured form. The set of features needs to be devised, along with the classification approach as well as a method to train and evaluate the model.

In this research, The Argument Annotated Essays (version 2.0) [1] is used, which is a corpus of 402 annotated persuasive essays consisting of 2,257 claims and 3,832 premises, to address the following research questions:

1. Which level of granularity of unit segmentation is effective for identifying locations of ADUs?
2. Regarding the tasks of identifying the locations of ADUs and classification of ADUs into claims and premises, what features are suitable?
3. What classification approach and which classifiers perform well in classifying ADUs into claims, premises or non-ADUs?

[1] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422

4. What should be a good measure of error to use for training and evaluation of the model?

4 Methodology and Approach

This section describes the procedure that will be used to find solutions to the research questions.

1. **Literature Search**

   A literature search would be carried out on articles, journals and books related to the field of Argument Mining and Natural Language Processing, in order to gain background knowledge, a deeper understanding of the problem statement, current methods that are employed as well as knowledge gaps present in the field.

2. **Setup Environment**

   A programming environment with necessary libraries, a collaborative platform and version control would be set up for the team to coordinate their work efforts.

3. **Data Exploration**

   This step will be carried out to understand the representation of the training data and the format in which the gold standard is labeled as well as perform preliminary pre-processing of the data.

4. **Unit Segmentation**

   Various levels of unit segmentation shall be performed and investigated for their effectiveness in classification. An ideal yet unrealistic case would be to generate all possible n-grams as segments, one of which has to be the correct segmentation but it is not possible since it will be extremely computationally intensive. Therefore, heuristics will be needed to get close to the correct segmentation of argument components.

   Unit segmentation would be experimented with at varying levels of granularity. Starting from the most coarse-grained level of paragraphs, then sentences, clauses and finally the level of tokens. Sentence segmentation can be performed by using sentence delimiters which can be done easily by existing Natural Language Processing libraries such as spaCy. The clause level could be performed through a rule-based approach relying on the syntax of English grammar, and using a pattern matcher to find the position of the clauses. The most fine-grained segmentation would be at the token level. In this case the plan is to treat the problem as a sequence labeling problem, using the BIO format to classify the tokens, where the tokens at the beginning of the argument component are to be labeled as "B", subsequent tokens inside the ADU as "I" and non-ADU tokens as "O" (Stab and Gurevych, 2017).

5. **Determining Features**

   After creating functions to segment the text into units, functions would need to be coded to extract features from text. The features would be obtained through experimentation and literature search. As a start, the well-known categorization of features (Structural, Lexical, Indicator, Syntactic, Contextual) provided by Stab and Gurevych (2014) would be used as a guideline to generate our own features. Another type of features would be utilized which is embeddings such as word2vec (Mikolov et al., 2013b), SBERT Reimers and Gurevych (2019) and RoBERTa (Liu et al., 2019) where an implementation is available in one of spaCy's language models.

6. **Classifiers**

   Non-neural network classifiers are considered to be used since the obvious benefit is the runtime. The current choice of classifiers is Logistic Regression, Random Forests, Naive Bayes, XGBoost and SVM, which have been used several time in past research (Lenz et al., 2020; Stab and Gurevych, 2017, 2014; Levy et al., 2014). a combination of multiple classifier is also considered to be experimented.

7. **Classification Approach**

   The plan is to use the following classification approaches:

   (a) A single binary classifier, classifying text as ADUs or non-ADUs.

   (b) A single multi-class classifier, classifying text as claims, premises or non-ADUs.

   (c) Two Binary classifiers, the first classifies text into ADUs or non-ADUs, the second classifier classifies the ADUs into claims or premises.

8. **Evaluation**

   By using different methods of unit segmentation, our units would be different from the labeled data and a binary measure of accuracy would not suffice. An appropriate measure is needed to devise to use as a loss function for training the model as well as evaluating the accuracy of the classification. The start will be with the basic idea of counting the number of tokens from the unit to the ADU contained within the unit (token distance) or out of all the tokens in the unit, what percentage of them have the correct label (percentage correctness). More sophisticated measures might be proposed, for example, the extended accuracy **??**, which would punish misclassifications in shorter units more severely than longer units.

Let d be the number of tokens to the correct position of the ADU Let U be the total number of tokens in the unit

$$\frac{1}{(d+1)^{\frac{log_2(d+1)}{log_2(U+1)}}}$$

5 Time Schedule

What	**Who**	**When**
Literature Search and writing Exposé	All	September-December
Set up programming environment	All	December
Literature Search for segmentation approaches and evaluation measures	All	December-January
Coding Functions to extract features	Ashwin	December-January
Experiment with methods of unit segmentation	Bilal	January
Training using different classification approaches and evaluation measures	Arturo	January
Finalize implementations based on results and experience	All	January-February
Write paper	All	February
Revise paper based on feedback	All	March

Table 2: Time Schedule Proposal Research Case Studies

Bibliography

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 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, 2017.

Simon Philip Botley and Faizal Hakim. Argument structure in learner writing: A corpus-based analysis using argument mapping. *Kajian Malaysia*, 32(1): 45–77, 2014.

Peter Davies. Improving the quality of students' arguments through 'assessment for learning'. *Journal of Social Science Education (JSSE)*, 8(2):94–104, 2009.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*, 2017.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. Corpus wide argument mining—a working solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7683–7691, 2020.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, 2016.

T. Govier. *A Practical Study of Argument*. Cengage Learning, 2009. ISBN 9780495603405. URL https://books.google.is/books?id=IUxXbUvDiswC.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.

Hospice Houngbo and Robert E Mercer. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the first workshop on argumentation mining*, pages 19–23, 2014.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

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

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, 2014.

Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a. URL http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Huy Nguyen and Diane Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, 2015.

Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Juri Opitz and Anette Frank. Dissecting content and context in argumentative relation analysis. *arXiv preprint arXiv:1906.03338*, 2019.

Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, pages 29–38, 2014.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

Isaac Persing and Vincent Ng. 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, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, 2014.

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

Christian Stab and Iryna Gurevych. Guidelines for annotating argumentation structures in persuasive essays. *Ubiquitous Knowledge Processing Lab (UKP Lab) Computer Science Department, Technische Universität Darmstadt*, 2015.

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

Christian Stab, Tristan Miller, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*, 2018.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguistics*, 28(4):409–445,

2002. https://doi.org/10.1162/089120102762671936. URL https://doi.org/10.1162/089120102762671936.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056, 2020.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge, 2013.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.