Argument component identification and its application in feedback on Dutch essays

Liqin Zhang, Howard Spoelstra
Open University of the Netherlands
liqin.zhang,howard.spoelstra@ou.nl

Marco Kalz
Open University of the Netherlands, Heidelberg University of Education
kalz@ph-heidelberg.de

ABSTRACT: Assessment for feedback on argumentative essays is challenging. However, current research on argument mining opens possibilities to developing automated argumentation assessment tools. Currently these tools are mainly available for the English language. Additional effort is required to adapt the techniques to e.g. Dutch. This study focuses on argument component identification on Dutch essays and how to present the results so they can be used as formative feedback.

Keywords: argument mining; Dutch essay; formative feedback

1 INTRODUCTION

Writing argumentative essays is challenging to both students and teachers. A survey among teachers (Authors, submitted) shows that both difficulty and time consumption for providing feedback on arguments are high. While automated assessment of argumentation and formative feedback generation would be able to alleviate the pressure on teachers, it is still not commonly available. On the one hand, argument mining, whose goal is to extract argumentation features, is still emerging (Lippi & Torroni, 2016), on the other hand, the state-of-the-art technology is mostly developed in the context of English, meaning that the application of the methods to other languages, such as Dutch, is restricted. We aim to develop a model to support argumentation analysis of essays in Dutch that can provide formative feedback as both quality indication and guidelines for improvement. Specifically, an argumentation component identification model for Dutch argumentative essays is developed by adopting previous work based on the English language (Stab & Gurevych, 2017a). The model is expected to function as an argumentation analysis component in a Dutch writing analysis tool. Based on the affordances of the model we explore the possible approaches to generate formative feedback.

2 RELATED STUDIES

2.1 Argument mining and formative feedback

Previous studies have explored the application of argument mining for essay assessment in terms of holistic essay quality and argumentation quality (Ghosh, Khanam, Han, & Muresan, 2016; Wachsmuth et al., 2017; Toledo et al., 2019). Additionally, giving effective formative feedback is already one of the main challenges in the current development of automated essay assessment (Strobl et al., 2019), which is particularly true for formative feedback on the argument components in texts. The study of applying argument mining for relevant feedback generation is still rare. One of the few studies into this matter is by Stab & Gurevych (2017b), who attempt to provide feedback on arguments according to the output of an argumentation analysis model (Stab & Gurevych, 2017a). Only recent studies into argument mining began to focus on developing such models aiming to translate the results of analysis into useful formative feedback for further improment of argumentative writing. For instance, Carlile et al. (2018) and Gao et al. (2019) developed rubric-based corpora and taxonomies which are used to classify various argumentation features, such as persuasiveness, coverage, coherence etc. into quality levels and it is beneficial for further studies of providing feedback based on the classification.

2.2 Argument mining in multilingual settings

Because of the lack of the argument mining research in non-English context, there are few human-annotated argumentation corpora available for multilingual argument mining research. Since the workload to create such human-annotated corpora is heavy, Eger et al. (2018) explore an alternative approach which involves machine translation and tag projection in order to extend existing models to afford multilingual application. The approach is as follows: A human-annotated L1 corpus (source language) with tags for major claims, claims, and premises is translated to L2 (target language) using Google Translate. By using fast-align (Dyer, Chahuneau, & Smith, 2013), each token in L1 is then aligned to the corresponding parallel text in L2. As each token in L1 is annotated by the aforementioned tags, these tags from L1 are projected[1] to the corresponding aligned tokens in L2. Eger et al. (2018) have shown that using such a machine-generated corpus to train a model for argument component identification performs comparably to the models trained by using a human-annotated corpus. This approach is also successfully applied to identify argument component relations for other languages, such as Portuguese (Rocha, Stab, Cardoso, & Gurevych, 2019).

3 ARGUMENT COMPONENT IDENTIFICATION FOR DUTCH

The identification of argument component identification is one of the basic tasks in argument mining. This study starts with developing an argumentation component identification model for Dutch, based on the studies of Stab & Gurevych (2017a) and Eger et al. (2018).

3.1 Data

Human-annotated argumentation corpora for Dutch argument mining are not available. To create a corpus of Dutch essays with argumentation structure annotations, a human-annotated corpus of 402 essays in English (Stab & Gurevych, 2014) is translated into Dutch with Google Translate. Afterwards, following the approach by Eger et al. (2018) the tags from the essays in English are projected into the Dutch texts. As a result, we obtain a corpus in Dutch containing 402 persuasive essays, annotated with argumentation components (major claim, claim, and premise) on the token-level. The annotations are in Inside-outside-beginning (IOB) format (Ramshaw & Marcus, 1999).

[1] For the detail of the projection method please see Eger et al. (2018).

3.2 Implementation of the model

When a training dataset is available in which the labels are on the token level, the implementation of a model is a task of token-level sequence tagging. The model is a bidirectional Long-Short-Term-Memory (LSTM) neural network with a CRF layer. The parameters set to train the model are derived from Eger et al. (2018). The model is trained in 5 runs, with 50 epochs in each run. The Dutch words are represented with 300-dimension vectors which is trained on the Wikipedia Dutch articles using a skip-gram model (Bojanowski et al., 2017). The framework of Kahse (2018) is applied for the implementation of the model. The adapted Dutch data is split into train/develop/test sets, and the performance of the model is evaluated by precision/recall/F1 scores.

4 PRESENTATION OF FORMATIVE FEEDBACK

Based on the affordances of the model, we propose to provide formative feedback which we expect to be useful for teachers and students for them to improve their performance in writing argumentatively. As the current model identifies the argument components and types, it is feasible to highlight them in the text. Descriptive statistics (such as the number of premises and claims) can also be provided as feedback, as these relate to argument quality. Visualizing the argumentation structure (see Figure 1) is also effective to provide insight into the argumentation structure of the essay (Chiang, Fan, Liu, & Chen, 2016).

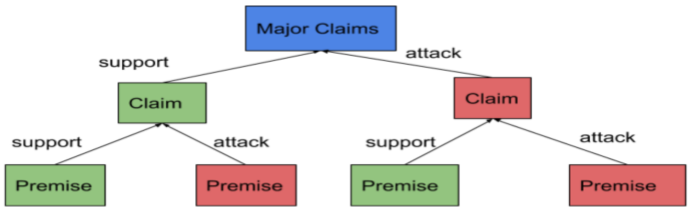

Figure 1: An example of visualization of argumentation structure.

5. LIMITATIONS AND FUTURE WORK

The possibility of generating effective formative feedback depends on the performance of the argumentation component identification model. The performance of the model needs to have a sufficiently accurate to use it for formative feedback purposes. Meanwhile, the variety in the formative feedback is possibly limited by the affordances of the model. In the future word we will therefore focus on the assessment of the reliability of the approach and the different ways to present feedback based on results of the analysis. We suggest that the model can be improved by analyzing the relations between argument components with assessment rubrics, allowing the student to understand their performance. Last but not the least, the usefulness of the formative feedback generated by the argument analysis approaches should be evaluated by conducting an empirical experiment involving students and teachers.

REFERENCES

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.

Carlile, W., Gurrapadi, N., Ke, Z., & Ng, V. (2018, July). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 621-631).

Chiang, K. H., Fan, C. Y., Liu, H. H., & Chen, G. D. (2016). Effects of a computer-assisted argument map learning strategy on sixth-grade students' argumentative essay reading comprehension. *Multimedia Tools and Applications*, 75(16), 9973-9990.

Dyer, C., Chahuneau, V., & A Smith, N. (2013). A simple, fast, and effective reparameterization of ibm model 2.

Eger, S., Daxenberger, J., Stab, C., & Gurevych, I. (2018). Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!. *arXiv preprint arXiv:1807.08998*.

Gao, Y., Driban, A., McManus, B. X., Musi, E., Davies, P., Muresan, S., & Passonneau, R. J. (2019, August). Rubric Reliability and Annotation of Content and Argument in Source-Based Argument Essays. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 507-518).

Ghosh, D., Khanam, A., Han, Y., & Muresan, S. (2016, August). 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)* (pp. 549-554).

Kahse, T. (2019). Multi-task learning for argumentation mining. *arXiv preprint arXiv:1904.10162.*

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2), 10.

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157-176). Springer, Dordrecht.

Rocha, G., Stab, C., Cardoso, H. L., & Gurevych, I. (2018, November). Cross-Lingual Argumentative Relation Identification: from English to Portuguese. In *Proceedings of the 5th Workshop on Argument Mining* (pp. 144-154).

Stab, C., & Gurevych, I. (2014, August). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1501-1510).

Stab, C., & Gurevych, I. (2017a). Parsing argumentation structures in persuasive essays. *Computational Linguistics, 43*(3), 619-659.

Stab, C., & Gurevych, I. (2017b). Training Argumentation Skills with Argumentative Writing Support. *Computational Linguistics*, *43*(3), 619–659.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131,* 33-48.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., ... & Slonim, N. (2019). Automatic Argument Quality Assessment--New Datasets and Methods. *arXiv preprint arXiv:1909.01007.*

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., ... & Stein, B. (2017, April). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: *Volume 1, Long Papers* (pp. 176-187).