Recognizing Insufficiently Supported Arguments in Argumentative Essays

Christian Stab[†] and **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

In this paper, we propose a new task for assessing the quality of natural language arguments. The premises of a well-reasoned argument should provide enough evidence for accepting or rejecting its claim. Although this criterion, known as sufficiency, is widely adopted in argumentation theory, there are no empirical studies on its applicability to real arguments. In this work, we show that human annotators substantially agree on the sufficiency criterion and introduce a novel annotated corpus. Furthermore, we experiment with feature-rich SVMs and convolutional neural networks and achieve 84% accuracy for automatically identifying insufficiently supported arguments. The final corpus as well as the annotation guideline are freely available for encouraging future research on argument quality.[1]

1 Introduction

Argumentation is an omnipresent routine and an integral part of our daily verbal communication. It is a verbal activity that aims at increasing or decreasing the plausibility of a controversial standpoint (van Eemeren et al., 1996, p. 5). Well-reasoned arguments of high quality are not only important for making thoughtful decisions and persuading a particular audience but also play a major role for drawing widely accepted conclusions. Computational argumentation is a recent research field in natural language processing that focuses on the analysis of arguments in natural language texts. Novel advances have a broad application potential in various areas like debating technologies (Levy et al., 2014; Rinott et al., 2015), policy making (Sardianos et al., 2015), information retrieval (Carstens and Toni, 2015), and legal decision support (Mochales-Palau and Moens, 2009). Recently, computational argumentation is receiving increasing interest in *intelligent writing assistance* (Song et al., 2014; Stab et al., 2014) since it enables *argumentative writing support* systems that provide tailored feedback about arguments in student essays.

Most of the existing approaches in computational argumentation consider argumentation as discourse structures and focus on the identification of arguments in natural language texts. For instance, existing approaches classify text units as argumentative or non-argumentative (Moens et al., 2007), recognize argument components such as *claims* or *premises* at the sentence-level (Mochales-Palau and Moens, 2009; Kwon et al., 2007; Eckle-Kohler et al., 2015) or clause-level (Levy et al., 2014; Sardianos et al., 2015), or identify argument structures by classifying pairs of argument components (Stab and Gurevych, 2014). However, these approaches are of limited use for argumentative writing support systems since they do not recognize the weak points of arguments.

Despite the comprehensive theoretical framework on argument quality in logic and argumentation theory (van Eemeren et al., 1996; Damer, 2009), there are only few computational approaches that focus on the assessment of arguments in natural language texts. These existing approaches either identify undisputed arguments in online communities (Cabrio and Villata, 2012), assess the persuasiveness of arguments (Wei et al., 2016), compare and rank arguments regarding their convincingness (Habernal and Gurevych, 2016b), or summarize the argumentation strength

[1] https://www.ukp.tu-darmstadt.de/data/argumentation-mining

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 980–990,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

of an entire essay in a single holistic score (Persing and Ng, 2015). Our approach is based on the theoretical framework proposed by Johnson and Blair (2006). In particular, we focus on the *sufficiency criterion* that an argument fulfills if its premises provide enough evidence for accepting or rejecting the claim. The following example argument illustrates a violation of the sufficiency criterion:

> **Example 1**: "*It is an undeniable fact that tourism harms the natural habitats of the destination countries. As Australia's Great Barrier Reef has shown, the visitors cause immense destruction by breaking corals as souvenirs, throwing boat anchors or dropping fuel and other sorts of pollution.*"

The premise of this argument represents a particular example (second sentence) that supports a general claim in the first sentence. The argument is a generalization from one sample to the general case. However, a single sample is not enough to support the general case. Therefore, the argument does not comply with the sufficiency criterion.

> **Example 2**: "*Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells of the patient. In addition, it shortens the healing process.*"

Example 2 illustrates a sufficiently supported argument. It is reasonable to accept that transplantation patients will benefit from cloning if it enables a better match and an accelerated healing process.

Our primary motivation is to create an argument analysis method for argumentative writing support systems that classifies an argument as *sufficient* if its premises provide enough evidence for accepting its claim (example 2) or as *insufficient* if its premises do not provide enough evidence (example 1). Therefore, our first research question is whether human annotators can reliably apply the sufficiency criterion to real arguments and if it is possible to create annotated data of high quality. The second research question addresses the automatic recognition of insufficiently supported arguments. We investigate if, and how accurately, insufficiently supported arguments can be identified by computational techniques.

The contribution of this paper is threefold: first, we investigate to what extent human annotators agree on the sufficiency criterion. We present the results of an annotation study with three annotators and show that our annotation guideline successfully guides annotators to substantial agreement. Second, we show that insufficiently supported arguments can be identified with high accuracy using convolutional neural networks (CNN). The experimental results show that a CNN significantly outperforms several challenging baselines and manually created features. Third, we introduce a novel corpus for studying the quality of arguments.

2 Related Work

Previous works in computational argumentation focused primarily on approaches for *argument mining*. These include, for example, methods for the identification of arguments in legal texts (Moens et al., 2007), news articles (Eckle-Kohler et al., 2015; Sardianos et al., 2015), or user-generated web discourse (Habernal and Gurevych, 2016a). Other approaches address the classification of argument components into claims and premises (Mochales-Palau and Moens, 2009), supporting and opposing claims (Kwon et al., 2007), or backings, rebuttals and refutations (Habernal and Gurevych, 2016a). Levy et al. (2014) recognize context-dependent claims and Rinott et al. (2015) retrieve several types of evidence from Wikipedia. Approaches for identifying the structure of arguments recognize argumentative relations between argument components using context-free grammars (Mochales-Palau and Moens, 2009), pair classification (Stab and Gurevych, 2014), or maximum spanning trees (Peldszus and Stede, 2015). However, none of these approaches consider the quality of arguments.

Similarly, most existing corpora in computational argumentation are only annotated with argument components (Habernal and Gurevych, 2016a; Aharoni et al., 2014; Mochales-Palau and Moens, 2009) or argument structures (Reed et al., 2008; Stab and Gurevych, 2014; Peldszus and Stede, 2015) and do not include annotations of argumentative quality issues. Other resources in the field contain arguments annotated with different properties such as emotions and sarcasm (Walker et al., 2012), the type of reasoning (Reed et al.,

2008) or the stance on a topic (Somasundaran and Wiebe, 2009). However, there is no corpus of arguments annotated with the sufficiency criterion.

Currently there are only few approaches that focus on the automatic assessment of argument quality. Cabrio and Villata (2012) employed textual entailment for identifying undisputed arguments in online discussions. They built a graph that represents attack and support relations between arguments and applied the abstract argumentation framework (Dung, 1995) for identifying accepted arguments. Although their approach is capable of finding undisputed arguments among a given set of arguments, it does not answer why a specific argument is of inferior quality than another argument. Thus, their approach is of limited use for guiding students since it does not pinpoint particular weaknesses of arguments.

Park and Cardie (2014) proposed an approach for classifying propositions as verifiable (experiential and non-experiential) or unverifiable. Their best approach based on a support vector machine achieves a macro F1 score of .690. Although the verifiability of propositions enables to determine appropriate types of support, it does not answer if an argument is sufficiently supported or not.

Persing and Ng (2015) introduced an approach for recognizing the argumentation strength of an essay. They found that pos n-grams, prompt adherence features, and predicted argument components perform best. However, their model determines a single holistic score that summarizes the argumentation quality of the entire essay. Consequently, it does not provide formative feedback that guides students to improve their arguments.

Recently, researchers proposed approaches for automatically assessing the persuasiveness of arguments. For instance, Wei et al. (2016) proposed an approach for ranking user comments taken from online fora and found that argumentation related features are effective for this task. Cano-Basave and He (2016) ranked speakers in political debates by using semantic frames which indicate persuasive argumentation features, and Habernal and Gurevych (2016b) compared the convincingness of argument pairs using feature-rich SVMs and bidirectional LSTMs. However, the persuasiveness score of an argument is only of limited use for argumentative writing support, since it summarizes various quality criteria and does not explain why an argument is weak.

3 Argument Quality: Theoretical Background

An argument consists of several *argument components*. It includes a claim and one or more premises. The *claim* (also called *conclusion*) is a controversial statement and the central component of an argument. The *premises* constitute the reasons for believing the claim to be true or false (Damer, 2009, p. 14). Assessing the quality of arguments is a complex task since arguments in natural language are hardly ever in a standardized form (Damer, 2009; Govier, 2010). Moreover, argument quality is a product of many different criteria (Johnson and Blair, 2006). The quality of an argument depends, for instance, on its lexical clarity and phrasing (representation), the level of trust that the audience has in the arguer (ethos), and the emotions and values appealed by the argument (pathos). The *logical quality* of arguments (logos) is, however, independent of all other merits, defects and external influence factors (Johnson and Blair, 2006, p. 50). Certainly, external factors or the presentation style can have a strong influence on the persuasive power of arguments. However, these factors can at most masquerade an illogical argument but not improve its logical quality. Therefore, the logical quality is most suitable for assessing the (intrinsic) quality of arguments and for providing feedback about written arguments respectively.

Traditionally, there are two different perspectives on the logical quality of arguments: (*i*) the formal logic perspective and (*ii*) the informal logic perspective. The objective of *formal logic approaches* is to distinguish deductively valid arguments from invalid arguments (van Eemeren et al., 1996, chapter 1.2), i.e. to recognize if the claim of an argument follows necessarily from its premises. However, formal logic approaches cannot be applied to everyday arguments since the vast majority of arguments do not follow deductive inference rules (Damer, 2009; van Eemeren et al., 1996).

Informal logic aims at developing theoretical frameworks for analyzing arguments in ordinary natural language (Groarke, 2015). These include, for example, *fallacy theories* which focus on determining particular argumentative mistakes that can be observed with a marked degree of frequency. Current theories list various forms of fallacious arguments. For instance, the framework proposed by Damer (2009) describes 61 different fallacy

types. However, fallacy theories are not appropriate for recognizing logically good arguments (van Eemeren et al., 1996, p. 178) since it is unknown if all fallacies are already known. To overcome this limitation, Johnson and Blair (2006) proposed three binary criteria, known as RAS-criteria, that a logically good argument needs to fulfill:

- *Relevance*: An argument fulfills the relevance criterion, if all of its premises count in favor of the truth (or falsity) of the claim.
- *Acceptability*: An argument fulfills the acceptability criterion if its premises represent undisputed common knowledge or facts.
- *Sufficiency*: An argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim.

The relevance criterion addresses the relation between each premise and the claim whereas the acceptability criterion focuses on the truthfulness of each individual premise. Both need to be evaluated independently for each premise of the argument. The sufficiency criterion addresses the premises of an argument together. It is fulfilled if the relevant premises of an argument are enough for justifying (or rejecting) the claim. The sufficiency criterion presupposes a non-empty set of relevant premises. However, an argument can violate the relevance criterion and comply with the sufficiency criterion at the same time. For instance, an argument can have several relevant premises that are sufficient for accepting the claim and additional premises that are not relevant to the claim. This also implies that a sufficient argument has a non-empty set of relevant premises but it is unknown if all premises of a sufficient argument are relevant to the claim.

In contrast to fallacy theories, the RAS-criteria enable to distinguish good from bad arguments with respect to logical quality since each argument that complies with all three criteria is a logically good one (Govier, 2010; Johnson and Blair, 2006). Moreover, the RAS-criteria attribute a particular defect to the relation between individual premises and the claim (relevance), the truthfulness of individual premises (acceptability), or the premises considered together (sufficiency). Therefore, they enable purposeful feedback for resolving particular defects of weak arguments and are well suited for argumentative writing support systems.

4 Corpus Creation

We conducted our annotation on a corpus of 402 argumentative essays that has been previously annotated with argumentation structures (Stab and Gurevych, 2016). By analyzing the annotated argumentation structures, we found that each body paragraph contains at least one argument and only 4.3% of all body paragraphs include several arguments, i.e. claims supported by premises. Therefore, we considered each body paragraph as an individual argument. This approximation has additional practical advantages for the identification of insufficiently supported arguments since it does not require the identification of argumentation structures in advance and prevents potential error propagation. Following this procedure, we extracted 1,029 arguments with an average length of 94.6 tokens and 4.5 sentences per argument.

4.1 Annotation Study

Three non-native annotators with excellent English proficiency independently annotated the arguments as sufficient or insufficient. We used 64 arguments from the corpus for elaborating the annotation guideline and 20 arguments for collaborative training sessions with the annotators. In these sessions, all three annotators collaboratively analyzed arguments for resolving disagreements and obtaining a common understanding of the annotation guideline. For the actual annotation task, we used the freely available brat rapid annotation tool (Stenetorp et al., 2012).

4.1.1 Inter-Annotator Agreement

All three annotators independently annotated an evaluation set of 433 arguments. We evaluated the agreement between the annotators using several inter-annotator agreement measures implemented in DKPro Agreement (Meyer et al., 2014). We used observed agreement and the two chance-corrected measures Fleiss' κ (Fleiss, 1971) and Krippendorff's α with nominal distance function (Krippendorff, 1980). The three annotators agreed on 91.07% of all 433 arguments (observed agreement). The chance-corrected agreement scores of $\kappa = .7672$ and $\alpha = .7673$ show substantial agreement between the annotators which allows "*tentative conclusions*" (Krippendorff, 1980). Therefore, we conclude that human annotators can reliably identify insufficiently supported arguments in argumentative essays.

4.1.2 Analysis of Disagreements

In order to identify the reasons for the disagreements, we manually investigated all arguments on which the annotators disagreed. We found that a high proportion of these arguments include modal verbs in their claims. The following example illustrates such an argument:

> "*Watching television too often can have a negative effect on communication abilities. For instance, children often prefer watching cartoons or movies instead of meeting their classmates and thus they will not learn how to communicate properly.*"

Due to the modal verb "*can*" in the claim of this argument (first sentence), it is sufficient to provide one specific example as premise. However, annotators tend to overlook modal verbs and overhastily annotate these arguments as insufficient.

The second most frequent cause of disagreements is due to the length of the arguments. In particular, one annotator annotated remarkably fewer arguments as insufficient. These arguments exhibit a comparatively large number of premises. This indicates that longer arguments are more likely to be perceived as sufficient than shorter arguments.

We also observed that several disagreements are due to hard cases. For instance, assessing the sufficiency of the following argument depends on the subjective interpretation of the undetermined quantification "*many*" in the claim:

> "*Living in big cities provides many opportunities. First of all, it will be easier to find a job in a city. Also there are various bars and clubs where you can meet new people. Above all there are shopping malls and cinemas for spending your free time.*"

We also found that annotators do not agree on arguments including terms such as "*some*", "*various*", or "*large number*". Thus, extending the annotation guideline with an explanation of how to handle modal verbs, the number of premises and undetermined qualifiers could further improve the agreement between the annotators in future annotation studies.

4.2 Creation of the Final Corpus

We merged the annotations of the three annotators on the evaluation set using majority voting. The remaining arguments have been annotated by the two annotators with the highest pairwise agreement on the evaluation set ($\alpha = .815$). Disagreements on the remaining arguments have been manually resolved in discussions among the two annotators. Table 1 shows an overview of the corpus.

size	
tokens	97,370
sentences	4,593
arguments	1,029
class distribution	
sufficient	681 (66.2%)
insufficient	348 (33.8%)

Table 1: Size of the final corpus and class distribution of sufficiency annotations.

The class distribution is skewed towards sufficiently supported arguments. However, the proportion of 33.8% insufficiently supported arguments indicates that students frequently do not support their claims with sufficient evidence.

5 Experiments

We consider the identification of insufficiently supported arguments as a binary classification task and label each body paragraph as *sufficient* or *insufficient*. For preventing errors in model assessment due to a particular data splitting (Krstajic et al., 2014), we used a repeated 5-fold cross-validation setup and ensured that arguments from the same essay are not distributed over the train, test and development sets. We repeated the cross-validation 20 times which yields a total of 100 folds. As evaluation scores, we used accuracy and macro F1 score as well as the F1 score, precision and recall of the class "insufficient". Whereas the precision indicates the performance of the model to identify arguments that are really in need of revision, recall shows how well the model recognizes all insufficiently supported arguments in an essay. All evaluation scores are reported as average including the standard deviation over the 100 folds. In order to determine the macro F1 score, we employ macro-averaging as proposed by Sokolova and Lapalme (2009, p. 430). For model selection and hyperparameter tuning, we randomly sampled 10% of the training set of each

fold as a development set. For significance testing, we employ Wilcoxon signed-rank test on macro F1 scores with a significance level of $\alpha = .005$.

We employ several models from the DKPro Framework (Eckart de Castilho and Gurevych, 2014) for preprocessing. We use the language tool segmenter[2] for tokenization and sentence splitting. We employ the Stanford parser (Klein and Manning, 2003) and named entity recognizer (Finkel et al., 2005) for constituency parsing and recognizing organizations, persons and locations. Note that only the model described in Section 5.2 requires all preprocessing steps. All other models use only the tokenization of the language tool segmenter.

5.1 Baselines

For our experiments, we use the following two baselines: First, we employ a majority baseline that classifies each argument as sufficient. Second, we use a support vector machine with polynomial kernel implemented in the Weka framework (Hall et al., 2009). We employ the $4{,}000$ most frequent lowercased words as binary features and refer to this model as SVM-bow.

5.2 Manually Created Features (SVM)

Our first system is based on manually created features. As a learner, we use the same support vector machine as for SVM-bow. For feature extraction and experimentation, we use the DKPro TC text classification framework (Daxenberger et al., 2014). We tried various features which have been used previously for assessing the quality or the persuasiveness of arguments (cf. Section 2). For instance, we experimented with argument structures (Stab and Gurevych, 2014), transitional phrases (Persing and Ng, 2015), semantic roles (Das et al., 2014) and discourse relations (Lin et al., 2014). However, we found that only the following features are effective for recognizing insufficiently supported arguments:

Lexical: To capture lexical properties, we employ the $4{,}000$ most frequent lowercased words as binary features analogous to SVM-bow.

Length: We use the number of tokens and the number of sentences as features since sufficiently supported arguments might exhibit more premises than insufficiently supported arguments (cf. Section 4.1.2).

Syntax: For capturing syntactic properties, we extract binary production rules from the constituent parse trees of each sentence of the argument as described by Stab and Gurevych (2014).

Named Entities (ner): We assume that arguments with insufficient support refer to particular entities in order to justify more general claims (cf. example 1 in Section 1). Thus, we add the number of named entities appearing in the argument and the average occurrence of named entities per sentence to our feature set. We consider organizations, persons and locations separately. Thus the named entity features comprise six features in total, i.e. three binary and three numeric features.

5.3 Convolutional Neural Network (CNN)

Our second model is a convolutional neural network with max-over time pooling (Collobert et al., 2011). We use the implementation provided by Kim (2014). The selection of this model is motivated by the excellent performance that the model achieves in many different classification tasks like sentiment classification of question classification. We found in our experiments that instead of using several convolutional layers with different window sizes, a single convolutional layer with a window size of 2 and 250 feature maps performs best. For representing each word of an argument, we use word embeddings trained on the google news data set by Mikolov et al. (2013). In order to adapt these vectors to the identification of insufficient arguments, we use non-static word vectors as proposed by Kim (2014). We train the network with stochastic gradient descent over shuffled mini-batches with the Adadelta update rule (Zeiler, 2012), a dropout rate of $.5$ and a mini-batch size of 50. For finding the best model, we apply early stopping on the development sets.

5.4 Results

Table 2 shows the results of the model assessment on the test sets. The SVM-bow model with unigram features achieves $.755$ macro F1 score and $.785$ accuracy. It significantly outperforms the majority baseline by $.357$ macro F1 score which indicates that lexical features are informative for identifying insufficiently supported arguments. The support vector machine with manually created features significantly outperforms both the majority baseline and SVM-bow. It achieves $.798$ accuracy and $.770$ macro F1 score and thus outperforms the SVM-bow model by $.015$ macro F1

[2] https://www.languagetool.org/

	Accuracy	*Macro F1*	*F1 Insufficient*	*Precision*	*Recall*
Human Upper Bound*	.911±.022	.887±.026	.940±.015	.863±.058	.808±.109
Baseline Majority	.662±.033	.398±.012	0	0	0
Baseline SVM-bow †	.785±.029	.755±.034	.661±.051	.709±.067	.624±.067
SVM †‡	.798±.028	.770±.032	.681±.047	.731±.060	.641±.061
CNN †‡	**.843**±.025	**.827**±.027	**.770**±.039	**.762**±.054	**.784**±.068

Table 2: Results of model assessment on the test sets and comparison to human upper bound († significant improvement over baseline majority; ‡ significant improvement over Baseline SVM-bow; *determined on a subset of 433 arguments).

score. We obtain the best performance by using the CNN model. It significantly outperforms all other models with respect to all evaluation scores and achieves .827 macro F1 score and an accuracy of .843. The results also show that the SVM model with manually created features achieves a considerably lower recall compared to precision. Thus, the model is less suitable for exhaustively finding all insufficiently supported arguments. In contrast, the CNN model is more balanced with respect to precision and recall and considerably outperforms the recall of the SVM model. Therefore, the CNN model outperforms the SVM model in finding insufficiently supported arguments in argumentative essay and performs better for recognizing arguments that are really in need of revision.

We determine the human upper bound by averaging the evaluation scores of all three annotator pairs on the 433 independently annotated arguments (cf. Section 4). Human annotators achieve an accuracy of .911. The CNN model yields only .068 less accuracy compared to the human upper bound and thus achieves 92.5% of human performance.

5.5 Feature Analysis

Although the CNN model outperforms the support vector machine with manual features, we analyzed the features for gaining a better understanding of insufficiently supported arguments and to investigate which linguistic properties are informative for recognizing arguments with insufficient support. Table 3 shows the macro F1 scores of the support vector machine using individual features and the results of feature ablation tests on the development sets.

The results show that lexical features are most effective for identifying insufficiently supported arguments. They achieve the best macro F1 score of .749 when used individually. Removing lexical features from the feature set also yields the highest decrease in macro F1 score compared to the other features. The second best features are named entities. Using only named entity features yields a macro F1 score of .681. Thus, we can confirm our assumption that named entities are informative features for assessing the sufficiency of arguments. Syntactic features are also effective for recognizing insufficiently supported arguments. They yield .640 macro F1 score when used individually. The results also show that the length of an argument is only marginally informative for assessing the sufficiency of arguments. Using the length features individually yields only a slight improvement of the macro F1 score over the majority baseline. However, removing the length from the entire feature set causes a slight decrease of .002 in the macro F1 score compared to the system which uses all features. We achieve the best results by combining all features.

	Macro F1	*F1 Insuf.*	*F1 Suf.*
BS Majority	.396±.020	0	.793±.041
only lexical	.749±.048	.649±.070	.835±.040
only length	.397±.023	.002±.015	.792±.040
only syntax	.640±.063	.502±.101	.767±.047
only ner	.681±.059	.410±.114	.823±.039
all w/o lexical	.658±.059	.529±.093	.776±.045
all w/o length	.766±.049	.674±.068	.847±.040
all w/o syntax	.755±.049	.659±.070	.839±.040
all w/o ner	.760±.050	.666±.069	.843±.041
all features	.768±.049	.677±.068	.848±.040

Table 3: Results of the SVM using individual features and feature ablation tests on the dev sets.

For gaining further insights into the characteristics of insufficiently supported arguments, we ranked all unigrams using information gain. The top ten words are "*example*", "*my*", "*was*", "*instance*", "*i*", "*for*", "*me*", "*friend*", "*he*", and "*did*". This might be an indication that *examples* (signaled by the terms "*example*" and "*instance*")

or *personal experiences* (signaled by terms such as "*me*", "*my*", "*friend*" or "*he*") are not sufficient for developing strong arguments.

5.6 Error Analysis

In order to analyze the most frequent errors of the convolutional neural network, we manually investigated all arguments which are wrongly classified in each run of the repeated cross-validation experiment. In total, we found 41 sufficient arguments which are consistently misclassified as insufficient (false positives) and 28 insufficient arguments that are always misclassified as sufficient (false negatives).

Among the false positives, we observed that 35 arguments include examples as evidence which are signaled by terms like "*example*" or "*instance*". Thus, the model tends to overemphasize the presence of particular lexical indicators. Most of these arguments either refer to an example in addition to other premises which are already sufficient to support the claim or include an example for specifying another premise. However, we also found several false negatives which include examples as evidence. Thus, the model does not solely rely on these lexical clues.

Among the 28 false negatives, we found 8 arguments that refer to multi-word named entities which are not captured by word embeddings. Another 5 false negatives support the claim by means of personal experience and 3 ones cite numbers, i.e. previous studies or empirical evidence.

6 Discussion

Although the convolutional neural network achieves promising results, the sufficiency criterion is only one of three criteria that a logically good argument needs to fulfill. Thus, our approach is not yet able to separate logically good from illogical arguments. In our experiments, we also analyzed arguments with respect to the relevance and acceptability criterion. In particular, we conducted several annotation studies with varying guidelines and two annotators on a set of 100 arguments. For annotating the relevance criterion, we presented the annotated structure of each argument to the annotators and asked them to assess the relevance of each premise for the claim individually. In order to evaluate the acceptability criterion, we asked the annotators to mark each premise as acceptable if it represents undisputed common knowledge or a fact. However, we found that human annotators hardly agree on these criteria. We obtained low agreement scores of $\kappa = .435$ for the relevance criterion and $\kappa = .259$ for the acceptability criterion, which is not sufficient for creating a reliable corpus. In addition, we found that the violations of the relevance and acceptability criteria are less frequent than violations of the sufficiency criterion in argumentative essays. We observed that only 15% of the arguments include a premise that violates the relevance criterion and 14% of all premises violate the acceptability criterion. Although this imbalance explains the low agreement scores (Artstein and Poesio, 2008, p. 573), it also poses additional requirements for the size of the corpus and for computational models.

Although we didn't obtain adequate agreement scores for the acceptability and relevance criteria, we implemented a system that identifies insufficiently supported arguments in argumentative essays with a reasonable accuracy. Given that sufficiency flaws are the most frequent quality defects in argumentative essays, our system represents an important milestone for realizing argumentative writing support systems.

7 Conclusion

We presented a novel approach for assessing the quality of natural language arguments. In particular, we focused on the sufficiency criterion that each logically good argument needs to fulfill. Previous approaches on argument quality are of limited use for argumentative writing support systems since they are not capable of recognizing particular weaknesses in argumentative texts. To overcome this limitation, we conducted an empirical study on the applicability of the sufficiency criterion to real arguments in argumentative essays. The inter-annotator agreement of $\alpha = .7673$ shows that human annotators substantially agree in this annotation task and confirms that humans can reliably separate sufficiently supported arguments from insufficiently supported arguments. We introduced a novel corpus annotated with the sufficiency criterion for studying logical mistakes in argumentation. This corpus is freely available for ensuring the reproducibility of our results and to encourage future research on argument quality. Furthermore, we presented the results of our experiments for automatically rec-

ognizing if an argument is sufficiently supported or not. We found that convolutional neural networks significantly outperform challenging baselines and manually created features with a macro F1 score of .827 and an accuracy of .843. Moreover, we showed that insufficiently supported arguments frequently exhibit particular lexical indicators. In addition, the feature analysis revealed that named entities and syntactic features are good indicators for separating sufficiently supported arguments from insufficiently supported arguments.

For future work, we plan to continue with our experiments with the relevance and acceptability criteria. In addition, we plan to integrate our method in writing environments for evaluating its effectiveness for supporting authors.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus project AWS under grant No. 01|S12054. We thank our annotators Can Diehl and Radhika Gaonkar for their valuable contributions and the anonymous reviewers for their helpful comments.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, MD, USA.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '12, pages 208–212, Jeju Island, Korea.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, USA.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Reasoning*. Wadsworth Cengage Learning, 6th edition.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40:1:9–56.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based framework for supervised learning experiments on textual data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, ACL '14, pages 61–66, Baltimore, MD, USA.

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

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 2236–2242, Lisbon, Portugal.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 363–370, Ann Arbor, Michigan.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, 7th edition.

Leo Groarke. 2015. Informal logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Ivan Habernal and Iryna Gurevych. 2016a. Argumentation mining in user-generated web discourse. *Computational Linguistics, arXiv preprint arXiv:1601.02403v4*, page (in press).

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Ralph H. Johnson and Anthony J. Blair. 2006. *Logical Self-Defense*. International Debate Education Association.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1746–1751, Doha, Qatar.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Sapporo, Japan.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage.

Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(10).

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 76–81, Philadelphia, PA, USA.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 1489–1500, Dublin, Ireland.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. Dkpro agreement: An open-source java library for measuring inter-rater agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 105–109, Dublin, Ireland.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, Barcelona, Spain.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, Stanford, CA, USA.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, MA, USA.

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*, EMNLP '15, pages 938–948, Lisbon, Portugal.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL '15, pages 543–552, Beijing, China.

Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pages 2613–2618, Marrakech, Morocco.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 440–450, Lisbon, Portugal.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO, USA.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, ACL '09, pages 226–234, Suntec, Singapore.

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

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 '14, pages 46–56, Doha, Qatar.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *arXiv preprint arXiv:1604.07370*.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 40–49, Bertinoro, Italy.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Avignon, France.

Frans H. van Eemeren, Rob Grootendorst, and Francisca Snoeck Henkemans. 1996. *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*. Routledge, Taylor & Francis Group.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany.

Matthew D. Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.