# Towards segment-based recognition of argumentation structure in short texts

**Andreas Peldszus**
Applied Computational Linguistics
University of Potsdam
peldszus@uni-potsdam.de

## Abstract

Despite recent advances in discourse parsing and causality detection, the automatic recognition of argumentation structure of authentic texts is still a very challenging task. To approach this problem, we collected a small corpus of German microtexts in a text generation experiment, resulting in texts that are authentic but of controlled linguistic and rhetoric complexity. We show that trained annotators can determine the argumentation structure on these microtexts reliably. We experiment with different machine learning approaches for automatic argumentation structure recognition on various levels of granularity of the scheme. Given the complex nature of such a discourse understanding tasks, the first results presented here are promising, but invite for further investigation.

## 1 Introduction

Automatic argumentation recognition has many possible applications, including improving document summarization (Teufel and Moens, 2002), retrieval capabilities of legal databases (Palau and Moens, 2011), opinion mining for commercial purposes, or also as a tool for assessing public opinion on political questions.

However, identifying and classifying arguments in naturally-occurring text is a very challenging task for various reasons: argumentative strategies and styles vary across texts genres; classifying arguments might require domain knowledge; furthermore, argumentation is often not particularly explicit – the argument proper is being infiltrated with the full range of problems of linguistic expression that humans have at their disposal.

Although the amount of available texts featuring argumentative behaviour is growing rapidly in the web, we suggest there is yet one resource missing that could facilitate the development of automatic argumentation recognition systems: Short texts with explicit argumentation, little argumentatively irrelevant material, less rhetorical gimmicks (or even deception), in clean written language.

For this reason, we conducted a text generation experiment, designed to control the linguistic and rhetoric complexity of written ‘microtexts’. These texts have then been annotated with argumentation structures. We present first results of automatic classification of these arguments on various levels of granularity of the scheme.

The paper is structured as follows: In the next section we describe related work. Section 3 presents the annotation scheme and an agreement study to prove the reliability. Section 4 describes the text generation experiment and the resulting corpus. Section 5 and 6 present the results of our first attempts in automatically recognizing the argumentative structure of those texts. Finally, Section 7 concludes with a summary and an outlook on future work.

## 2 Related Work

There exist a few ressources for the study of argumentation, most importantly perhaps the AIF database, the successor of the Araucaria corpus (Reed et al., 2008), that has been used in different studies. It contains several annotated English datasets, most interestingly for us one covering online newspaper articles. Unfortunately, the full source text is not part of the downloadable database, which is why the linguistic material surrounding the extracted segments is not easy to retrieve for analysis. Instead of manually annotating, Cabrio and Villata (2012) created an argumentation resource by extracting argumentations from collaborative debate portals, such as debatepedia.org, where arguments are already classified into pro and con classes by the

*Proceedings of the First Workshop on Argumentation Mining*, pages 88–97,
Baltimore, Maryland USA, June 26, 2014. ©2014 Association for Computational Linguistics

users. Unfortunately, those arguments are themselves small texts and their internal argumentative structure is not marked up. Finally, to the best of our knowledge, the only existing corpus of German newspaper articles, essays or editorials annotated with argumentation structure is that used by Stede and Sauermann (2008), featuring ten commentaries from the Potsdam Commentary Corpus (Stede, 2004). Although short, these texts are rhetorically already quite complex and often have segments not relevant to the argument.[1]

In terms of automatic recognition, scientific documents of different fields have been studied intensively in the Argumentative Zoning approach or in similar text zoning approaches (Teufel and Moens, 2002; Teufel et al., 2009; Teufel, 2010; Liakata et al., 2012; Guo et al., 2013). Here, sentences are classified into different functional or conceptual roles, grouped together with adjacent sentences of the same class to document zones, which induces a flat partitioning of the text. A variety of machine learning schemes have been applied here.

Another line of research approaches argumentation from the perspective of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and works with argumentation-enriched RST trees (Azar, 1999; Green, 2010). However, we do not consider RST to be the best level for representing argumentation, due to its linearization constraints (Peldszus and Stede, 2013a, sec. 3). Nevertheless, noteworthy advances have been made recently in rhetorical parsing (Hernault et al., 2010; Feng and Hirst, 2012). Whether hybrid RST argumentation structures will profit similarly remains to be shown. A more linguistically oriented approach is given with the TextCoop platform (Saint-Dizier, 2012) for analyzing text on the discourse level with emphasis on argumentation.

One step further, Feng and Hirst (2011) concentrate on types of arguments and use a statistical approach to classify already identified premises and conclusions into five common argumentation schemes (Walton et al., 2008).

## 3 Annotation Scheme

Our representation of the argumentation structure of a text is based on Freeman's theory of argumentation structure (Freeman, 1991; Freeman, 2011).[2] Its central idea is to model argumentation as a hypothetical dialectical exchange between the proponent, who presents and defends his claims, and the opponent, who critically questions them in a regimented fashion. Every move in such a dialectical exchange corresponds to a structural element in the argument graph. The nodes of this graph represent the propositions expressed in text segments (round nodes are proponent's nodes, square ones are opponent's nodes), the arcs between those nodes represent different supporting (arrow-head links) and attacking moves (circle-head links). The theory distinguishes only a few general supporting and attacking moves. Those could be specified further with a more fine grained set, as provided for example by the theory of argumentation schemes (Walton et al., 2008). Still, we focus on the coarse grained set, since this reduces the complexity of the already sufficiently challenging task of automatic argument identification and classifcation. Our adaption of Freeman's theory and the resulting annotation scheme is described in detail and with examples in (Peldszus and Stede, 2013a).

### 3.1 Reliability of annotation

The reliability of the annotation scheme has been evaluated in two experiments. We will first recapitulate the results of a previous study with naive annotators and then present the new results with expert annotators.

**Naive annotators**: In (Peldszus and Stede, 2013b), we presented an agreement study with 26 naive and untrained annotators: undergraduate students in a "class-room annotation" szenario, where task introduction, guideline reading and the actual annotation is all done in one obligatory 90 min. session and the subjects are likely to have different experience with annotation in general, background knowledge and motivation. We constructed a set of 23 microtexts (each 5 segments long) covering different linearisations of several combinations of basic argumentation constructs. An example text and the corresponding argumentation structure graph is shown in Figure 1. On these texts, the annotators achieved moderate agreement[3] for certain aspects of the ar-

[1] We intend to use this resource, when we move on to experiment with more complex texts.

[2] The theory aims to integrate the ideas of Toulmin (1958) into the argument diagraming techniques of the informal logic tradition (Beardsley, 1950; Thomas, 1974) in a systematic and compositional way.

[3] Agreement is measured in Fleiss $\kappa$ (Fleiss, 1971).

gument graph (e.g. $\kappa$=.52 in distinguishing proponent and opponent segments, or $\kappa$=.58 in destinguishing supporting and attacking segments), yet only a marginal agreement of $\kappa$=.38 on the full labelset describing all aspects of the argument graph. However, we could systematically identify subgroups performing much better than average using clustering techniques: e.g. a subgroup of 6 annotators reached a relatively high IAA agreement of $\kappa$=.69 for the full labelset and also high agreement with gold data.

**Expert annotators**: Here, we present the results of an agreement study with three expert annotators: two of them are the guideline authors, one is a postdoc in computational linguistics. All three are familiar with discourse annotation tasks in general and specifically with this annotation scheme. They annotated the same set of 23 microtexts and achieved a high agreement of $\kappa$=.83 on the full labelset describing all aspects of the argument graph. The distinction between supporting and attacking was drawn with very high agreement of $\kappa$=.95, the one between proponent and opponent segments even with perfect agreement.

Since argumentation structures can be reliably annotated using this scheme, we decided to create a small corpus of annotated microtexts.

## 4 Dataset

The corpus used in this study consists of two parts: on the one hand, the 23 microtexts used in the annotation experiments just described; on the other hand, 92 microtexts that have been collected in a controlled text generation experiment. We will describe this experiment in the following subsection.

### 4.1 Microtext generation experiment

We asked 23 probands to discuss a controversial issue in a short text of 5 segments. A list of 17 of these issues was given, concerning recent political, moral, or everyday's life questions. Each proband was allowed to discuss at maximum five of the given questions. Probands were instructed to first think about the pros & cons of the controversial question, about possible refutation and counter-refutations of one side to the other. On this basis, probands should decide for one side and write a short persuasive text (corresponding to the standards of the written language), arguing in favour of their chosen position.

The written texts were required to have a length of five segments. We decided not to bother our probands with an exact definition of a segment, as this would require the writers to reliably identify different complex syntactic constructions. Instead, we simply characterized it as a clause or a sentence, expressing an argumentative point on its own. We also required all segments to be argumentatively relevant, in the sense that they either formulate the main claim of the text, support the main claim or another segment, or attack the main claim or another segment. This requirement was put forward in order to prevent digression and argumentatively irrelevant but common segment types, such as theme or mood setters, as well as background information. Furthermore, we demanded that at least one possible objection to the main claim be considered in the text, leaving open the choice of whether to counter that objection or not. Finally, the text should be written in such a way that it would be understandable without having the question as a headline.

In total, 100 microtexts have been collected. The five most frequently chosen issues are:

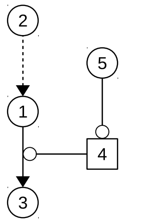
- Should the fine for leaving dog excrements on sideways be increased?
- Should shopping malls generally be allowed to open on Sundays?
- Should Germany introduce the death penalty?
- Should public health insurance cover treatments in complementary and alternative medicine?
- Should only those viewers pay a TV licence fee who actually want to watch programs offered by public broadcasters?

### 4.2 Cleanup and annotation

Since we aim for a corpus of clean, yet authentic argumentation, all texts have been checked for spelling and grammar errors. As a next step, the texts were segmented into elementary units of argumentation. Due to the (re-)segmentation, not all texts conform to the length restriction of five segments, they can be one segment longer or shorter. Unfortunately, some probands wrote more than five main clauses, yielding texts with up to ten segments. We decided to shorten these texts down to six segments by removing segments that appear redundant or negligible. This removal also required modifications in the remaining segments to maintain text coherence, which we made as

[*Energy-saving light bulbs contain a considerable amount of toxic substances.*]$_1$ [*A customary lamp can for instance contain up to five milligrams of quicksilver.*]$_2$ [*For this reason, they should be taken off the market,*]$_3$ [*unless they are virtually unbreakable.*]$_4$ [*This, however, is simply not case.*]$_5$

(a)



(b)

| node id | rel. id | full label | target |
|---|---|---|---|
| 1 | 1 | PSNS | (n+2) |
| 2 | 2 | PSES | (n-1) |
| 3 | 3 | PT | (0) |
| 4 | 4 | OAUS | (r-3) |
| 5 | 5 | PARS | (n-1) |

(c)

Figure 1: An example microtext: the (translated) segmented text in (a), the argumentation structure graph in (b), the segment-based labeling representation in (c).

minimal as possible. Another source of problems were segments that do not meet our requirement of argumentative relevance. Some writers did not concentrate on discussing the thesis, but moved on to a different issue. Others started the text with an introductory presentation of background information, without using it in their argument. We removed those segments, again with minimal changes in the remaining segments. Some texts containing several of such segments remained too short after the removal and have been discarded from the dataset.

After cleanup, 92 of the 100 written texts remained for annotation of argumentation structure. We found that a few texts did not meet the requirement of considering at least one objection to the own position. In a few other texts, the objection is not present as a full segment, but rather implicitly mentioned (e.g. in a nominal phrase or participle) and immediatly rejected in the very same segment. Those segments are to be annotated as a supporting segment according to the guidelines, since the attacking moves cannot be expressed as a relation between segments in this case.

We will present some statistics of the resulting dataset at the end of the following subsection.

## 5 Modelling

In this section we first present, how the argumentation structure graphs can be interpreted as a segment-wise labelling that is suitable for automatic classification. We then describe the set of extracted features and the classifiers set up for recognition.

### 5.1 Preparations

In the annotation process, every segment is assigned one and only one function, i.e. every node in the argumentative graph has maximally one outgoing arc. The graph can thus be reinterpreted as a list of segment labels.

Every segment is labeled on different levels: The ‘role’-level specifies the dialectical role (proponent or opponent). The ‘typegen’-level specifies the general type, i.e. whether the segment presents the central claim (thesis) of the text, supports or attacks another segment. The ‘type’-level additionally specifies the kind of support (normal or example) and the kind of attack (rebutter or undercutter). Whether a segment’s function holds only in combination with that of another segment (combined) or not (simple) is represented on the ‘combined’-level. The target is finally specified by a position relative identifier: The offset *-x. . . 0. . . +x* identifies the targeted segment, relative from the position of the current segment. The prefix ‘n’ states that the proposition of the node itself is the target, while the prefix ‘r’ states that the relation coming from the node is the target.[4]

The labels of each separate level can be merged to form a complex tagset. We interpret the result as a hierarchical tagset as it is presented in Figure 2. The label ‘PSNS(n+2)’ for example stands for a proponent’s segment, giving normal, non-combined support to the next but one segment, while ‘OAUS(r-1)’ represents an opponent’s segment, undercutting the relation established by the immediately previous segment, not combined. Figure 1c illustrates the segment-wise labelling for the example microtext.

The dataset with its 115 microtexts has 8183 word tokens, 2603 word types and 579 segments in total. The distribution of the basic labels and the complex ‘role+type’ level is presented in Table 1. The label distribution on the ‘role+type’ level shows that most of the opponent’s attacks are rebutting attacks, directed against the central claim

[4] Segments with combined function (as e.g. linked supporting arguments) are represented by equal relation ids, which is why segments can have differing node and relation ids. However, for the sake of simplicity, we will only consider example of non-combined nature in this paper.

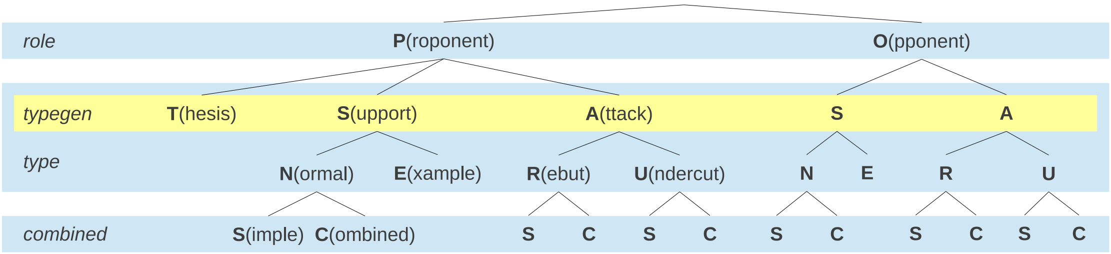


Figure 2: The hierarchy of segment labels.

or its premises directly (OAR>OAU). In contrast, the proponent's counters of these attack are typically untercutting attacks, directed against the attack relation (PAU>PAR). This is due to the author's typical strategy of first conceding some aspect in conflict with the main claim and then rendering it irrelevant or not applicable without directly challenging it. Note however, that about 40% of the opponents objections have not been countered by the proponent (OA*>PA*).

### 5.2 Features

All (unsegmented) texts have been automatically split into sentences and been tokenized by the OpenNLP-tools. The mate-pipeline then processed the tokenized input, yielding lemmatization, POS-tags, word-morphology and dependency parses (Bohnet, 2010). The annotated gold-standard segmentation in the dataset was then automatically mapped to the automatic sentence-splitting/tokenization, in order to be able to extract exactly those linguistic features present in the gold-segments. Using this linguistic output and several other resources, we extracted the following features:

**Lemma Unigrams**: We add a set of binary features for every lemma found in the present segment, in the preceding and the subsequent segment in order to represent the segment's context in a small window.

**Lemma Bigrams**: We extracted lemma bigramms of the present segment.

**POS Tags**: We add a set of binary features for every POS tag found in the present, preceding and subsequent segment.

**Main verb morphology**: We added binary features for tempus and mood of the segment's main verb, as subjunctive mood might indicate anticipated objections and tempus might help to identify the main claim.

**Dependency triples**: The dependency parses were used to extract features representing dependency triples (relation, head, dependent) for each token of the present segment. Two features sets were built, one with lemma representations, the other with POS tag representations of head and dependent.

**Sentiment**: We calculate the sentiment value of the current segment by summing the values of all lemmata marked as positive or negative in SentiWS (Remus et al., 2010).[5]

**Discourse markers**: For every lemma in the segment that is listed as potentially signalling a discourse relation (cause, concession, contrast, asymmetriccontrast) in a lexicon of German discourse markers (Stede, 2002) we add a binary feature representing the occurance of the marker, and one representing the occurance of the relation. Again, discourse marker / relations in the preceding and subsequent segment are registered in separate features.

**First three lemmata**: In order to capture sentence-initial expressions that might indicate argumentative moves, but are not strictly defined as discourse markers, we add binary features representing the occurance of the first three lemmata.

**Negation marker presence**: We use a list of 76 German negation markers derived in (Warzecha, 2013) containing both closed class negation operators (negation particles, quantifiers and adverbials etc.) and open class negation operators (nouns like "denial" or verbs like "refuse") to detect negation in the segment.

**Segment position**: The (relative) position of the segment in the text might be helpful to identify typical linearisation strategies of argumentation.

In total a number of ca. 19.000 features has been extracted. The largest chunks are bigrams and lemma-based dependencies with ca. 6.000 features each. Each set of lemma unigrams (for

[5] We are aware that this summation is a rather trivial and potentially error-prone way of deriving an overall sentiment value from the individual values of the tokens, but postpone the use of more sophisticated methods to future work.

| level | role | typegen | type | comb | target | role+type |
|---|---|---|---|---|---|---|
| labels | P (454) | T (115) | T (115) | / (115) | n-4 (26) | PT (115) |
| | O (125) | S (286) | SN (277) | S (426) | n-3 (52) | PSN (265) |
| | | A (178) | SE (9) | C (38) | n-2 (58) | PSE (9) |
| | | | AR (112) | | n-1 (137) | PAR (12) |
| | | | AU (66) | | 0 (115) | PAU (53) |
| | | | | | n+1 (53) | OSN (12) |
| | | | | | n+2 (35) | OSE (0) |
| | | | | | r-1 (54) | OAR (100) |
| | | | | | r-2 (7) | OAU (13) |
| | | | | | . . . | |
| # of lbls | 2 | 3 | 5 | 3 | 16 | 9 |

Table 1: Label distribution on the basic levels and for illustration on the complex ‘role+type’ level. Labels on remaining complex level combine accoringly: ‘role+type+comb’ with in total 12 different labels and ‘role+type+comb+target’ with 48 different labels found in the dataset.

the present, preceding, and subsequent segment) has around 2.000 features.

### 5.3 Classifiers

For automatic recognition we compare classifiers that have frequently been used in related work: Naïve Bayes (NB) approaches as in (Teufel and Moens, 2002), Support Vector Machines (SVM) and Conditional Random Fields (CRF) as in (Liakata et al., 2012) and maximum entropy (MaxEnt) approaches as in (Guo et al., 2013) or (Teufel and Kan, 2011). We used the Weka data mining software, v.3.7.10, (Hall et al., 2009) for all approaches, except MaxEnt and CRF.

**Majority**: This classifier assignes the most frequent class to each item. We use it as a lower bound of performance. The used implementation is Weka’s ZeroR.

**One Rule**: A simple but effective baseline is the one rule classification approach. It selects and uses the one feature whose values can describe the class majority with the smallest error rate. The used implementation is Weka’s OneR with standard parameters.

**Naïve Bayes**: We chose to apply a feature selected Naïve Bayes classifier to better cope with the large and partially redundant feature set.[6] Before training, all features are ranked accoring to their information gain observed on the training set. Features with information gain $\not> 0$ are excluded.

**SVM**: For SVMs, we used Weka’s wrapper to LibLinear (Fan et al., 2008) with the Crammer and Singer SVM type and standard wrapper parameters.

**MaxEnt**: The maximum entropy classifiers are trained and tested with the MaxEnt toolkit (Zhang, 2004). We used at maximum 50 iterations of L-BFGS parameter estimation without a Gaussian prior.

**CRF**: For the implementation of CRFs we chose Mallet (McCallum, 2002). We used the SimpleTagger interface with standard parameters.

Nonbinary features have been binarized for the MaxEnt and CRF classifiers.

## 6 Results

All results presented in this section have been produced in 10 repetitions (with different random seeds) of 10-fold cross validation, i.e. for each score we have 100 fold-specific values of which we can calculate the average and the standard deviation. We report A(ccuracy), micro-averaged F(1-score) as a class-frequency weighted measure and Cohen’s $\kappa$ (Cohen, 1960) as a measure focussing on less frequent classes. All scores are given in percentages.

### 6.1 Comparing classifiers

A comparison of the different classifiers is shown in Table 2. Due to the skewed label distribution, the majority classifier places the lower bounds already at a quite high level for the ‘role’ and ‘comb’-level. Also note that the agreement between predicted and gold for the majority classifier is equivalent to chance agreement and thus $\kappa$ is 0 on every level, even though there are F-scores near the .70.

Bold values in Table 2 indicate highest average. However note, that differences of one or two percent points between the non-baseline classifiers are not significant, due to the variance over the

[6] With feature selection, we experienced better scores with the Naïve Bayes classifier, the only exception being the most complex level ‘role+type+comb+target’, where only very few features reached the information gain threshold.

| level | Majority | | | OneR | | | CRF | | |
|---|---|---|---|---|---|---|---|---|---|
| | A | F | $\kappa$ | A | F | $\kappa$ | A | F | $\kappa$ |
| role | 78±1 | 69±1 | 0±0 | 83±3 | 79±4 | 33±13 | 86±5 | 84±6 | 49±16 |
| typegen | 49±1 | 33±1 | 0±0 | 58±3 | 47±3 | 23±7 | 68±7 | 67±8 | 46±12 |
| type | 48±1 | 31±1 | 0±0 | 56±3 | 45±3 | 22±6 | 62±7 | 58±8 | 38±10 |
| comb | 74±1 | 62±1 | 0±0 | 81±4 | 77±4 | 44±12 | **84**±5 | **81**±7 | 55±13 |
| target | 24±1 | 9±1 | 0±0 | 37±5 | 29±4 | 24±6 | 47±11 | **45**±11 | **38**±12 |
| role+typegen | 47±1 | 30±1 | 0±0 | 56±3 | 45±3 | 22±6 | 67±7 | 65±8 | 49±11 |
| role+type | 46±1 | 29±1 | 0±0 | 54±3 | 43±3 | 21±6 | 61±7 | 56±8 | 38±11 |
| role+type+comb | 41±1 | 24±1 | 0±0 | 50±4 | 38±3 | 19±6 | 56±7 | 51±8 | 36±9 |
| role+type+comb+target | 20±1 | 7±1 | 0±0 | 28±4 | 19±3 | 18±5 | 36±10 | 30±9 | 28±10 |
| level | Naïve Bayes | | | MaxEnt | | | LibLinear | | |
| | A | F | $\kappa$ | A | F | $\kappa$ | A | F | $\kappa$ |
| role | 84±5 | 84±5 | **52**±14 | **86**±4 | **85**±5 | **52**±15 | **86**±4 | 84±4 | 50±14 |
| typegen | **74**±5 | **74**±5 | **57**±8 | 70±6 | 70±6 | 51±10 | 71±5 | 71±5 | 53±9 |
| type | **68**±5 | **67**±5 | **52**±8 | 63±6 | 62±6 | 43±9 | 65±6 | 62±6 | 44±9 |
| comb | 74±6 | 75±5 | 42±11 | **84**±5 | **81**±7 | **56**±12 | **84**±3 | **81**±4 | 54±10 |
| target | 38±6 | 38±6 | 29±6 | 47±8 | 44±8 | 37±9 | **48**±5 | 44±5 | **38**±6 |
| role+typegen | **69**±6 | **69**±6 | **55**±9 | 68±7 | 67±7 | 51±10 | **69**±5 | 67±6 | 52±9 |
| role+type | 61±5 | **61**±5 | **45**±7 | 63±6 | **61**±6 | **45**±9 | **64**±5 | 60±5 | **45**±8 |
| role+type+comb | 53±6 | 51±6 | 36±8 | 58±6 | 54±7 | 41±8 | 61±5 | 56±5 | 44±8 |
| role+type+comb+target | 22±4 | 19±4 | 16±4 | 36±6 | **33**±6 | 29±6 | **39**±5 | 32±4 | **31**±5 |

Table 2: Classifier performance comparison: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of A(ccuracy), micro averages of F1-scores, and Cohen's $\kappa$.

folds on this rather small dataset.

The Naïve Bayes classifier profits from the feature selection on levels with a small number of labels and gives best results for the 'type(gen)' and 'role+typegen' levels. On the most complex level with 48 possible labels, however, performance drops even below the OneR baseline, because features do not reach the information gain threshold. The MaxEnt classifier performs well on the 'role' and 'comb', as well as on the 'role+type' levels. It reaches the highest F-score on the most complex level, although the highest accuracy and agreement on this levels is achieved by the SVM, indicating that the SVM accounted better for the less frequent labels. The SVM generally performs well in terms of accuracy and specifically on the most interesting levels for future applications, namely in target identification and the complex 'role+type' and 'role+type+comb+target' levels. For the CRF classifier, we had hoped that approaching the dataset as a sequence labelling problem would be of advantage. However, applied out of the box as done here, it did not perform as well as the segment-based MaxEnt or SVM classifier.

### 6.2 Feature ablation on 'role+type' level

We performed feature ablation tests with multiple classifiers on multiple levels. For the sake of brevity, we only present the results of the SVM and MaxEnt classifiers here on the 'role+type' level. The results are shown in Table 3. Bold values indicate greatest impact, i.e. strongest loss in the upper leave-one-feature-out half of the table and highest gain in the lower only-one-feature half of the table.

The greatest loss is produced by leaving out the discourse marker features. We assume that this impact can be attributed to the useful abstraction of introducing the signalled discourse relation as a features, since the markers are also present in other features (as lemma unigrams, perhaps first three lemma or even lemma dependencies) that produce minor losses.

For the single feature runs, lemma unigrams produce the best results, followed by discourse markers and other lemma features as bigrams, first three lemma and lemma dependencies. Note that negation markers, segment position and sentiment perform below or equal to the majority baseline. Whether at least the sentiment feature can prove more useful when we apply a more sophisticated calculation of a segment's sentiment value is something we want to investigate in future work. POS-tag based features are around the OneR baseline in terms of F-score and $\kappa$, but less accurate.

Interestingly, when using the LibLinear SVM, lemma bigrams have a larger impact on the overall performance than lemma based dependency triples in both tests, even for a language with a relatively free word order as German. This indicates that the costly parsing of the sentences might not be required after all. However, this difference is not

| Features | LibLinear | | | MaxEnt | | |
|---|---|---|---|---|---|---|
| | A | F | $\kappa$ | A | F | $\kappa$ |
| all | 64±5 | 60±5 | 45±8 | 63±6 | 61±6 | 45±9 |
| all w/o dependencies lemma | 64±5 | 60±5 | 46±8 | 62±6 | 60±6 | 44±9 |
| all w/o dependencies pos | 65±5 | 61±5 | 46±8 | 63±6 | 61±7 | 45±9 |
| all w/o discourse markers | **62**±5 | **59**±5 | **43**±8 | **61**±7 | **58**±7 | **42**±9 |
| all w/o first three lemma | 64±5 | 60±5 | 44±8 | 63±6 | 60±7 | 44±9 |
| all w/o lemma unigrams | 63±5 | 60±5 | 45±8 | 62±6 | 60±7 | 44±9 |
| all w/o lemma bigrams | 63±5 | 60±5 | 44±8 | 62±6 | 60±6 | 44±9 |
| all w/o main verb morph | 64±5 | 60±5 | 45±8 | 62±6 | 60±6 | 43±9 |
| all w/o negation marker | 64±5 | 60±6 | 45±8 | 63±6 | 61±7 | 45±9 |
| all w/o pos | 64±5 | 61±5 | 45±8 | 63±6 | 60±7 | 44±8 |
| all w/o segment position | 64±5 | 60±5 | 45±8 | 63±6 | 61±6 | 45±9 |
| all w/o sentiment | 64±5 | 60±5 | 45±8 | 62±6 | 60±6 | 44±9 |
| only dependencies lemma | 56±4 | 47±4 | 27±6 | 56±6 | 49±7 | 30±8 |
| only dependencies pos | 42±6 | 41±6 | 18±8 | 41±7 | 40±7 | 16±9 |
| only discourse markers | 56±6 | 53±6 | 34±9 | 53±6 | 52±7 | 30±10 |
| only first three lemma | 54±6 | 52±6 | 33±9 | 50±6 | 48±6 | 26±8 |
| only lemma unigrams | **59**±5 | **55**±5 | **37**±8 | **59**±6 | **56**±7 | **38**±8 |
| only lemma bigrams | **59**±4 | 53±5 | 34±8 | 55±7 | 51±7 | 30±9 |
| only main verb morph | 49±6 | 39±4 | 16±7 | 52±5 | 41±6 | 20±6 |
| only negation marker | 25±14 | 19±8 | 0±4 | 46±5 | 29±5 | 0±0 |
| only pos | 45±6 | 45±6 | 24±9 | 46±8 | 45±7 | 23±10 |
| only segment position | 31±12 | 25±10 | 4±7 | 46±5 | 29±6 | 0±0 |
| only sentiment | 22±14 | 15±11 | -1±3 | 46±5 | 29±6 | 0±0 |

Table 3: Feature ablation tests on the 'role+type' level: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of A(ccuracy), micro averages of F1-scores, and Cohen's $\kappa$.

as clear for the MaxEnt classifier.

### 6.3 Class specific results

Finally, we present class-specific results of the MaxEnt classifier for the 'role+type' level in Table 4. Frequent categories give good results, but for low-frequency classes there are just not enough instances in the dataset. We hope improve this by extending the corpus by corresponding examples.

| label | Precision | Recall | F1-score |
|---|---|---|---|
| PT | 75±12 | 74±13 | 74±11 |
| PSN | 65±8 | 79±7 | 71±6 |
| PSE | 1±6 | 1±6 | 1±6 |
| PAR | 12±29 | 12±27 | 11±24 |
| PAU | 57±26 | 49±24 | 50±22 |
| OSN | 1±12 | 1±12 | 1±12 |
| OAR | 54±18 | 42±16 | 46±13 |
| OAU | 8±27 | 7±23 | 7±23 |

Table 4: MaxEnt class-wise results on the 'role+type' level: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of Precision, Recall and F1-score.

## 7 Summary and outlook

We have presented a small corpus of German microtexts that features authentic argumentations, yet has been collected in a controlled fashion to reduce the amount of distracting or complicated rhetorical phenomena, focussing instead on the argumentative moves. The corpus has been annotated with a scheme that –as we have shown– can be reliably used by trained and experienced annotators. To get a first impression of the performance of frequently used modelling approaches on our dataset, we experimented with different classifiers with rather out-of-the-box parameter settings on various levels of granularity of the scheme. Given the complex nature of such a discourse understanding tasks, the first results presented here are promising, but invite for further investigation.

We aim to generate a significantly larger corpus of argumentative microtexts by a crowd-sourced experiment. For the improvement of models, we consider various strategies: Integrating top down constraints on the argumentation structure, as done in (Guo et al., 2013) for the zoning of scientific documents, is one option. Hierarchical models that apply classifiers along the levels of our label hierarchy are another option. Furthermore, we want to explore sequence labelling models in more detail. Ultimately, the goal will be to apply these methods to authentic news-paper commentaries.

## Acknowledgments

Thanks to Manfred Stede and to the anonymous reviewers for their helpful comments. The author was supported by a grant from Cusanuswerk.

## References

Moshe Azar. 1999. Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, 13:97–114.

Monroe C. Beardsley. 1950. *Practical Logic*. Prentice-Hall, New York.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In Luc De Raedt, Christian Bessiere, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 205–210. IOS Press.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

James B. Freeman. 1991. *Dialectics and the Macrostructure of Argument*. Foris, Berlin.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Argumentation Library (18). Springer.

Nancy L. Green. 2010. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24:181–196.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937, Atlanta, Georgia, June. Association for Computational Linguistics.

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, November.

Hugo Hernault, Hemut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):15–22.

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

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta, May. European Language Resources Association (ELRA).

Patrick Saint-Dizier. 2012. Processing natural language arguments with the TextCoop platform. *Journal of Argumentation and Computation*, 3(1):49–82.

Manfred Stede and Antje Sauermann. 2008. Linearization of arguments in commentary text. In *Proceedings of the Workshop on Multidisciplinary Approaches to Discourse*. Oslo.

Manfred Stede. 2002. DiMLex: A Lexical Approach to Discourse Markers. In Vittorio Di Tomaso Alessandro Lenci, editor, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria, Italy.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Simone Teufel and Min-Yen Kan. 2011. Robust argumentative zoning for sensemaking in scholarly documents. In Raffaella Bernadi, Sally Chambers, Björn Gottfried, Frédérique Segond, and Ilya Zaihrayeu, editors, *Advanced Language Technologies for Digital Libraries*, volume 6699 of *Lecture Notes in Computer Science*, pages 154–170. Springer Berlin Heidelberg.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445, December.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1493–1502, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. CSLI Publications.

Stephen N. Thomas. 1974. *Practical Reasoning in Natural Language*. Prentice-Hall, New York.

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

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

Saskia Warzecha. 2013. Klassifizierung und Skopusbestimmung deutscher Negationsoperatoren. Bachelor thesis, Potsdam University.

Le Zhang, 2004. *Maximum Entropy Modeling Toolkit for Python and C++*, December.