Joint prediction in MST-style discourse parsing for argumentation mining

Andreas Peldszus
Applied Computational Linguistics
UFS Cognitive Science
University of Potsdam
peldszus@uni-potsdam.de

Manfred Stede
Applied Computational Linguistics
UFS Cognitive Science
University of Potsdam
stede@uni-potsdam.de

Abstract

We introduce a new approach to argumentation mining that we applied to a parallel German/English corpus of short texts annotated with argumentation structure. We focus on structure prediction, which we break into a number of subtasks: relation identification, central claim identification, role classification, and function classification. Our new model jointly predicts different aspects of the structure by combining the different subtask predictions in the edge weights of an evidence graph; we then apply a standard MST decoding algorithm. This model not only outperforms two reasonable baselines and two data-driven models of global argument structure for the difficult subtask of relation identification, but also improves the results for central claim identification and function classification and it compares favorably to a complex mstparser pipeline.

1 Introduction

Argumentation mining is a task that has drawn increased interest in the last years. In its full-fledged version, it seeks to automatically recognize the structure of argumentation in a text by identifying and connecting the central claim of the text, supporting premises, possible objections, and counter-objections to these objections.[1]

A variety of applications can profit from access to the argumentative structure of text, including the retrieval of relevant court decisions from legal databases (Palau and Moens, 2011), automatic document summarization systems (Teufel and Moens, 2002), the analysis of scientific papers in biomedical text mining (Teufel, 2010; Liakata et al., 2012), or essay scoring. Importantly, argument analysis can also be an extension of opinion mining applications.

To make argumentation structures available for these applications, their robust automatic recognition is required, a task that is very challenging: argumentative strategies and styles vary across text genres and languages; classifying arguments might require domain knowledge; furthermore, argumentation can often rely on implicitly conveyed messages.

The full-fledged task can be decomposed into several subtasks:

- Segmentation: splitting the text into elementary discourse units (EDUs as used in general kinds of discourse parsing, typically sentences or clauses)
- Identification of argumentative discourse units (ADUs): discarding argumentatively irrelevant EDUs, joining adjacent EDUs to form larger ADUs
- ADU type classification: determining the type of argumentative unit; different schemes have been proposed, involving stance, evidence types, rhetorical status, argumentative function
- Relation identification: building a connected tree- or graph-structure to represent argumentative relations between the ADUs
- Relation type classification: determining the type of argumentative relation (e.g. supporting versus attacking relations or more fine-grained types)

In this paper, we address the last three subtasks: Given a text segmented into relevant ADUs, identify the argumentation structure. We will work with a bilingual corpus of short texts that have been generated in a text production experiment.

[1] A comprehensive overview of the research field is given in (Peldszus and Stede, 2013).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 938–948,
Lisbon, Portugal, 17-21 September 2015.

The next section describes related work. In section 3, we present the dataset used in our experiments. Section 4 gives a more detailed description of the task. The baselines and the models are presented in section 5. We then report the result of our experiments in section 6 and close with some concluding remarks.

2 Related Work

In our discussion of related work, we focus on the three subtasks addressed in this paper:

ADU type classification: One typical classification task concerns the properties of a segment in the argumentation structure: Burstein and Marcu (2003) trained classifiers for identifying thesis and conclusion statements in student essays, using additional automatic discourse parse features and cue words, resulting in an average F-score of 53% for thesis and 80% for conclusion segments. For legal texts, Palau and Moens (2011) demonstrated in their influential work how to classify the segment of a text into premises and conclusions, obtaining an F-score of 68% and 74% for the two classes. More recently, Stab and Gurevych (2014) classified segments in student essays into the classes major claim (of the text), claim (of the paragraph), premise and irrelevant. The macro average F-score for all classes is 73%, the F-score for the claim of the paragraph 54% and for the major claim 63%.

Besides structural segment-wise classification tasks, there is also work on more semantic tasks: The rhetorical status of a segment is classified in the argumentative zoning approaches (Teufel and Moens, 2002; Teufel and Kan, 2011; Liakata et al., 2012), where certain coarse-grained patterns of argumentation in scholarly papers can be captured. Park and Cardie (2014) focus on supporting segments and classify which type of evidence is presented in it. Finally, stance classification (Hasan and Ng, 2013) might be of interest to identify possible objections, although it is typically applied on full comments and not on single segments.

Relation identification: Much less prior work can be found for the process of building argumentation structures. Palau and Moens (2011) used a hand-written context-free grammar to predict argumentation trees on legal documents, achieving an accuracy of 60%. Only recently, data-driven approaches have been applied. Lawrence et al. (2014) construct tree structures on philosophical texts using unsupervised methods based on topical distance between the segments. The relations in the tree are neither labeled not directed. Unfortunately, the method was evaluated on only a few annotated items, which is why we cannot comment on the results. Finally, Stab and Gurevych (2014) present a supervised data-driven approach for relation identification. They predict attachment for support-graphs spanning over paragraphs of English essays and obtain a macro F1 score of 72%, and an F1 score of 52% for positive attachment. No decoding is used to optimize global predictions per text.

Relation type classification: The only study on explicitly classifying argumentative relations we are aware of is (Feng and Hirst, 2011). They classify pairs of premise and conclusion from newswire text into a set of five frequently used argumentation schemes in the sense of Walton et al. (2008). In one-against-others classification, the system yields best average accuracies of over 90% for two schemes, while for the other three schemes the results are between 63% and 70%.

To the best of our knowledge, no data-driven model of argumentation structure has been proposed yet that would optimize argumentation structure globally for the complete input text, as it is done in other discourse parsing tasks, e.g. in (Muller et al., 2012).

3 Dataset

Texts: We use the *arg-microtext* corpus (Peldszus and Stede, to appear), a freely available[2] parallel corpus of 112 short texts with 576 ADUs. The texts are authentic discussions of controversial issues. They were originally written in German and have been professionally translated to English, preserving the segmentation and if possible the usage of discourse markers. The texts have been collected in a controlled text generation experiment, with the result that all of them fulfill the following criteria: (i) The length of each text is about 5 ADUs (henceforth: segments). (ii) One segment explicitly states the central claim. (iii) Each segment is argumentatively relevant. (iv) At least one objection to the central claim is considered.

Scheme: The argumentation structure of every text has been annotated according to a scheme (Peldszus and Stede, 2013) based on Freeman's theory of argumentation structures (Free-

[2]https://github.com/peldszus/arg-microtexts

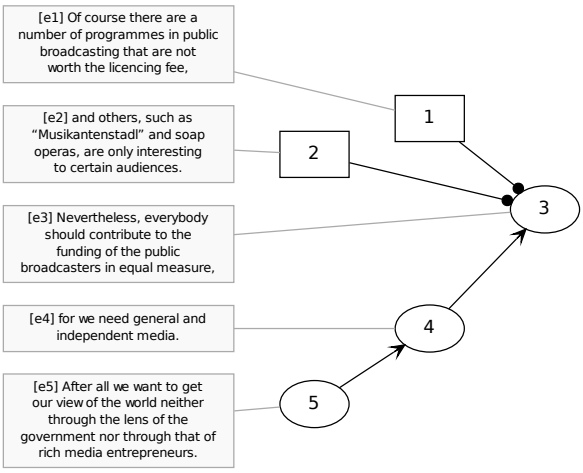

Figure 1: An example text and its reduced argumentation structure: Text segments, proponent (round) and opponent (box) nodes, supporting (arrow-head) and attacking (circle-head) relations.

man, 1991; Freeman, 2011), that has been proven to yield reliable structures in annotation experiments (Peldszus, 2014). The argumentation structure of a text is defined as a graph with the text segments as nodes. Each node is associated with one argumentative *role*: the proponent who presents and defends the central claim, or the opponent who critically questions the proponent's claims. Edges between the nodes represent argumentative relations, and each edge is of one specific argumentative *function*: support or attack. The scheme allows to discriminate between "rebutting" attacks, targeting another node and thereby challenging its acceptability, and "undercutting" attacks, targeting an edge and thereby challenging the acceptability of the inference from the source to the target node. It can also represent linked support, where multiple premises jointly support a claim.

Transformation: The annotated graph structures can be quite complex, especially when they involve undercutting relations and linked support. For the purpose of this study, we thus reduce the graphs to a simpler tree-like representation. All relations pointing to edges are rewritten to point to the source node of the original target edge, which enables the use of standard graph algorithms (like MST). Also, this is a loss-less mapping, given that every segment has only one outgoing arc (as generally done in argumentation models). Furthermore, the set of relation types is reduced to the simple binary distinction between support and attack. We think this is a reasonable simplification that facilitates comparisons with slightly different approaches/datasets (we are not aware of any dataset that makes use of the full granularity proposed in our scheme).

An example text from the corpus in its reduced form is shown in Figure 1. Text boxes are EDUs, each of which constitutes also an ADU. Proponent ADUs are round nodes, opponent ADUs are box nodes. Supporting relations have a normal arrowhead, while attacking relations have a circle arrowhead.

All statistics on the annotated argumentation structures apply equally for the German and the English version of the parallel corpus.

4 Task

Identifying the structure of argumentation according to our scheme involves choosing one segment as the central claim of the text, deciding how the other segments are related to the central claim and to each other, identifying the argumentative role of each segment, and finally the argumentative function of each relation.

Our prior experiments on automating the recognition of argumentation structure approached the problem as a segment-wise classification task (Peldszus, 2014). Formulating the task this way was successful for the recognition of argumentative role and function of a segment. For the automation of the structure building however, the segment-wise classification of attachment with only a small context window around the target segment proved to be a very hard task. This is due to the long-distance dependencies frequently found in argumentation graphs. For example, 46% of the relations marked in the corpus used for this study involve non-adjacent segments. For longer texts this number might increase further: Stab and Gurevych (2014) report a rate of 63% of non-adjacent relations in their corpus.

In this study we therefore frame the task of attachment classification as a binary decision, where the classifier, when given a pair of a source and a target segment, chooses whether or not to establish a relation from the source to the target. Since these relations can hold not only between adjacent but between arbitrary segments of the text, all possible combinations of segments are required to be tested. Consequently, the class distribution is very skewed.

- **attachment (at):** Is there an argumentative

connection between the source and the target segment? In the corpus, a relation has been annotated for 464 segment pairs, no relation has been annotated for the combinatorially remaining 2000 pairs of segments.

In this paper we first address only the task of attachment classification, and then the prediction of the full graph, involving all other levels:

- **central claim (cc):** Is the current segment the central claim of the text? In our data 112 of the 576 segments are central claims.
- **role (ro):** Does the current segment present a claim of the proponent or the opponent? In our data 451 of the 576 segments are proponent segments and 125 are opponent segments.
- **function (fu):** Has the current segment a supporting or an attacking function? In our data, 290 segments are supports, 174 are attacks and 112 are the central claim and thus have no own function.

5 Models

We compare two heuristic baseline models and different data-driven models that we developed, each of them trained and evaluated separately on both language versions of the corpus. All models are evaluated on the basis of 10 iterations of 5x3-fold nested cross validation (CV). The outer 5-fold CV is for evaluation only, i.e. to ensure that the model is trained only on training data and tested only on test data. If a model requires hyperparameters to be tuned or multiple passes, then this is achieved via one (or multiple) inner 3-fold CV over the training data only. The folding is stratified, randomly distributing the texts of the corpus while aiming to reproduce the overall label distribution in both training and test set.

5.1 Baseline: attach to first

In the English-speaking school of essay writing and debating, there is the tendency to state the central claim of a text or a paragraph in the very first sentence, followed by supporting arguments. It is therefore a reasonable baseline to assume that all segments attach to the first segment. In our corpus, the first segment is the central claim in 50 of the 112 texts (44.6%).

This baseline (**BL-first**) will not be able to capture serial argumentation, where one more general argument is supported or attacked by a more specific one. However, it will cover convergent argumentation, where separate arguments are put forward in favor of the central claim (given that it is expressed in the first segment). It will always produce flat trees. In our corpus, 176 of the 464 relations (37.9%) attach to the first segment.

5.2 Baseline: attach to preceding

A typically very strong baseline in discourse parsing is attaching to the immediately preceding segment (Muller et al., 2012). Possibly, this holds more for corpora with relations often or always being adjacent, as in rhetorical structure trees. Since argumentation structures often exhibit non-adjacent relations (see above), this heuristic might be easier to beat in our scenario.

This baseline (**BL-preced.**) will always produce chain trees and thus cover serial argumentation, but not convergent argumentation. In our corpus, 210 of all 464 relations (45.3%) attach to the preceding segment.

5.3 Learned attachment without decoding

We train a linear log-loss model (**simple**) using stochastic gradient descent (SGD) learning, with elastic net regularization, the learning rate set to optimal decrease and class weight adjusted according to class distribution (Pedregosa et al., 2011). The following hyper parameters are tuned in the inner CV: the regularization parameter alpha, the elastic net mixing parameter and the number of iterations. We optimize macro averaged F1-score.

For each text segment, we extract binary features for lemma, pos-tags, lemma- and pos-tag-based dependency-parse triples and the main verb morphology (Bohnet, 2010), and discourse connectives (Stede, 2002), furthermore simple statistics like relative segment position, segment length and punctuation count. For each pair of text segments, we extract relative distance between the segments and their linear order (is the source before or after the target). The feature vector for each pair then contains both the pair features and the segment features for source and target segment and their adjacent segments.[3]

[3] We experimented with several features, some of which were dismissed from the final evaluation runs due to lacking impact: sentiment values and the presence of negation for

5.4 Learned attachment with MST decoding

The simple model just described might be able to learn which segment pairs actually attach, i.e., correspond to some argumentative relation in the corpus. However it is not guaranteed to yield predictions that can be combined to a tree structure again. A more appropriate model would enforce global constraints on its predictions. In the **simple+MST** model, this is achieved by a *minimum spanning tree* (MST) decoding, which has first been applied for syntactic dependency parsing (McDonald et al., 2005a; McDonald et al., 2005b) and later for discourse parsing (Baldridge et al., 2007; Muller et al., 2012). First, we build a fully-connected directed graph, with one node for each text segment. The weight of each edge is the attachment probability predicted by the learned classifier for the corresponding pair of source and target segment. We then apply the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to determine the minimum spanning tree, i.e., the subgraph connecting all nodes with minimal total edge cost (in our case highest total edge probability). This resulting tree then represents the best global attachment structure for a text given the predicted probabilities.

5.5 Joint prediction with MST decoding

All models presented in the previous subsections have in common that they do not rely on other features of the argumentation graph. However, it is fair to assume that knowledge about the argumentative role and function of a segment or its likeliness to be the central claim might improve the attachment classification. Consequently, our next model considers not only the predicted probability of attachment for a segment pair, but also the predicted probabilities of argumentative role, function and of being the central claim for each segment. The predictions of all levels are combined in one *evidence graph*.

Additional segment-wise base classifiers: We train base classifiers for the role, function and central claim level using the same learning regime as described in Section 5.3. Contrary to the attachment classification, the items are not segment pairs but single segments. We thus extract all segment-based features as described above for the target segment and its adjacent segments.

segments, and distance measures between pairs of segments in terms of word-overlap, td-idf and LDA distributions.

Combining segment and segment-pair predictions: Our goal in this model is to combine the predicted probabilities of all levels in one edge score, so that the MST decoding can be applied as before. Figure 2 depicts the situation before and after the combination, first with separate prediction for segments and segment pairs and then with the combined edge scores.

The evidence graph is constructed as follows: First, we build a fully connected multigraph over all segments with as many edges per segment-pair as there are edge types. In our scenario there are two edge types, supporting and attacking edges. Then we translate the segment-wise predictions into level-specific edge scores.

The edge-score for the central claim level $\overline{\text{cc}}_{i,j}$ is equal to the probability of the edge's source not being the central claim, which is capturing the intuition that central claims are unlikely to have outgoing edges:

$$\overline{\text{cc}}_{i,j} = \quad p(\text{cc}_i = \text{no}) \tag{1}$$

The edge-score for the argumentative function level $\overline{\text{fu}}_{i,j}$ is equal to the probability of the source being the corresponding segment for the edge type:

$$\overline{\text{fu}}_{i,j} = \begin{cases} p(\text{fu}_i = \text{sup}) & \text{for sup. edges} \\ p(\text{fu}_i = \text{att}) & \text{for att. edges} \end{cases} \tag{2}$$

The edge-score for the argumentative role level $\overline{\text{ro}}_{i,j}$ is also determined by the edge type. Attacking edges involve a role switch (proponent or opponent would not attack their own claims), while supporting edges preserve the role (proponent or opponent will only support their own claims):

$$\overline{\text{ro}}_{i,j} = \begin{cases} p(\text{ro}_i = \text{pro}) \times p(\text{ro}_j = \text{pro}) + \\ p(\text{ro}_i = \text{opp}) \times p(\text{ro}_j = \text{opp}) & \text{for sup. edges} \\ \\ p(\text{ro}_i = \text{pro}) \times p(\text{ro}_j = \text{opp}) + \\ p(\text{ro}_i = \text{opp}) \times p(\text{ro}_j = \text{pro}) & \text{for att. edges} \end{cases} \tag{3}$$

Finally, of course the edge-score for the attachment level $\overline{\text{at}}_{i,j}$ is equal to the probability of attachment between the segment pair:

$$\overline{\text{at}}_{i,j} = \quad p(\text{at}_{i,j} = \text{yes}) \tag{4}$$

The combined score of an edge $w_{i,j}$ is then defined as the weighted sum of the level-specific edge score:

$$w_{i,j} = \frac{\phi_1 \overline{\text{ro}}_{i,j} + \phi_2 \overline{\text{fu}}_{i,j} + \phi_3 \overline{\text{cc}}_{i,j} + \phi_4 \overline{\text{at}}_{i,j}}{\sum \phi_n} \tag{5}$$

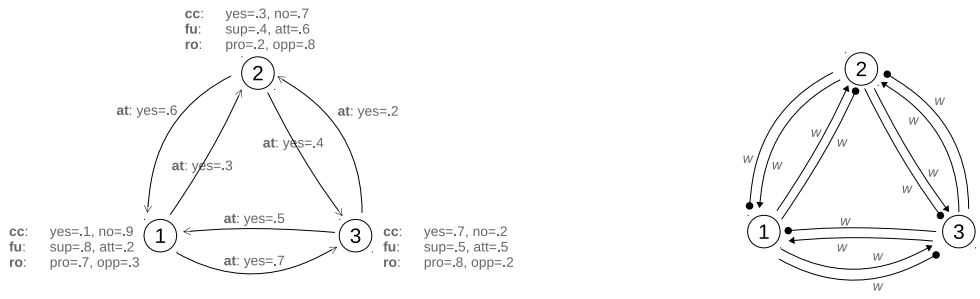

Figure 2: An example evidence graph before (left) and after (right) the predicted probabilities of the different levels have been combined in a single edge score.

In our implementation, the combined evidence graphs can be constructed without a weighting, and then be instantiated with a specific weighting to yield the combined edge scores $w_{i,j}$.

Procedure: As before, we first tune the hyperparameters in the inner CV, train the model on the whole training data and predict probabilities on all items of the test set. Also, we predict all items in the training data "as unseen" in a second inner CV using the best hyperparameters. This procedure is executed for every level. Using the predictions of all four levels, we then build the evidence graphs for training and test set.

Finding the right weighting: We evaluate two versions of the evidence graph model. The first version (**EG equal**) gives equal weight to each level-specific edge score. The second version (**EG best**) optimizes the weighting of the base classifiers with a simple evolutionary search on all evidence graphs of the training set, i.e. it searches for a weighting that maximizes the average level evaluation score of the decoded argumentation structures in the training set. Finally, all evidence graphs of the test set are instantiated with the selected weighting (the equal one or the optimized one) and evaluated.

5.6 Comparison: MST parser

Finally, we compare our models to the well-known mstparser[4], which was also used in the discourse parsing experiments of Baldridge et al. (2007). The mstparser applies 1-best MIRA structured learning, a learning regime that we expect to be superior over the simple training in the previous models. In all experiments in this paper, we use 10 iterations for training, the non-projective 1-best MST decoding, and no second order features. The base mstparser model (**MP**) evaluated here uses the same features as above, as well as its own features extracted from the dependency structure. Second, we evaluate a pre-classification scenario (**MP+p**), where the predictions of the base classifiers trained in the above models for central claim, role and function are added as additional features. We expect this to improve the central claim identification as well as the edge labeling.

For the full task involving all levels, we combine the mstparser with an external edge labeler, as the internal edge labeler is reported to be weak. In this setting (**MP+r**), we replace the edge labels predicted by the mstparser with the predictions of the base classifier for argumentative function. Furthermore, the combination of pre-classification, mstparser and external relation labeler (**MP+p+r**) is evaluated. Finally, we evaluate a scenario (**MPϵ+p+r**) where the mstparser has access only to its own features and to those of the pre-classification, but not to the features described in Section 5.3, and the external relation labeller is used. In this scenario, the mstparser exclusively serves as a meta-model on the base classifier's predictions.

6 Results

All results are reported as average and standard deviation over the 50 folds resulting from 10 iterations of (the outer) 5-fold cross validation. We use the following metrics: macro averaged F1, F1 for positive attachment, and Cohen's Kappa κ. For significance testing, we apply the Wilcoxon signed-rank test on the macro averaged F1 scores and assume a significance level of α=0.01.

6.1 Attachment task

Table 1 shows the results in the attachment task. The rule-based baseline scores are equal for both

[4] http://sourceforge.net/projects/mstparser/

	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best	MP	MP+p
F1 macro	.618±.041	.662±.025	.679±.025	.688±.032	.712±.026	.710±.028	.724±.030	**.728**±.033
attach F1	.380±.067	.452±.039	.504±.038	.494±.053	.533±.042	.530±.044	.553±.048	**.559**±.053
κ	.236±.081	.325±.050	.365±.048	.377±.064	.424±.052	.421±.055	.449±.060	**.456**±.066
trees	100%	100%	15.4%	100%	100%	100%	100%	100%
	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best	MP	MP+p
F1 macro	.618±.041	.662±.025	.663±.030	.674±.036	.692±.034	.693±.031	.707±.035	**.720**±.034
attach F1	.380±.067	.452±.039	.478±.049	.470±.058	.501±.056	.502±.052	.524±.056	**.546**±.056
κ	.236±.081	.325±.050	.333±.059	.347±.071	.384±.068	.386±.063	.414±.070	**.440**±.069
trees	100%	100%	11.6%	100%	100%	100%	100%	100%

Table 1: Results for the attachment task: for German (above) and English (below), best values highlighted.

	German		English	
total graphs	1120	100.0%	1120	100.0%
rooted	1091	97.4%	1088	97.1%
cycle free	1059	94.6%	995	88.8%
full span	908	81.1%	864	77.1%
out degree	298	26.6%	283	25.3%
trees	173	15.4%	120	10.7%

Table 2: Number and percentage of valid trees for the "simple" attachment model

languages, since they rely only on the annotated structure of the parallel corpus. Here, attach-to-first is the lower bound, attach-to-preceding is a more competitive baseline, as we had hypothesized in section 5.2.

The learned classifier (simple) beats both baselines in both languages, although the improvement is much smaller for English than for German. In general, the classifier lacks precision compared to recall: It predicts too many edges. As a result, the graph constructed from the predicted edges for one text very often does not form a tree. In Table 2, we give a summary of how often tree constraints are fulfilled, showing that without decoding, valid trees can only be predicted for 15.4% of the texts in German and for 10.7% of the texts in English. The most frequently violated constraint is "out degree", stating that every node in the graph should have at most one outgoing edge. Note that all other models, the baselines as well as the MST decoding models, are guaranteed to predict tree structures.

The simple+MST model yields slightly lower F1-scores for positive attachment than without decoding, trading off a loss of 10 points in recall of the over-optimistic base classifier against a gain of 5 in precision. However, the output graphs are constrained to be trees now, which is rewarded by a slight increase in the summarizing metrics macro F1 and κ.

The evidence graph models (EG equal & EG best) clearly outperform the simple and simple+MST model, indicating that the attachment classification can benefit from jointly predicting the four different levels. Note, that the EG model with equal weighting scores slightly better than the one with optimized weighting for German but not for English. However, this difference is not significant ($p>0.5$) for both languages, which indicates that the search for an optimal weighting is not necessary for the attachment task.

The overall best result is achieved by the mstparser model. We attribute this to the superior structured learning regime. The improvement of MP over EP equal and best is significant in both languages ($p<0.008$). Using pre-classification further improves the results, although difference is neither significant for German ($p=0.4$) nor for English ($p=0.016$).

6.2 Full task

Until now, we only focused on the attachment task. In this subsection we will present results on the impact of joint prediction for all levels.

The results in Table 3 show significant improvements of the EG models over the base-classifiers on the central claim, the function and the attachment levels ($p<0.0001$). This demonstrates the positive impact of jointly predicting all levels. The EG models achieve the best scores in central claim identification and function classification, and the second best result in role identification. The differences between EG equal and EG best are not significant on any level, which again indicates that

		simple	EG equal	EG best	MP	MP+p	MP+r	MP+p+r	MPϵ+p+r
cc	maF1	.849±.035	.879±.042	**.890**±.037	.825±.055	.855±.055	.825±.055	.855±.055	.854±.053
	κ	.698±.071	.759±.085	**.780**±.073	.650±.111	.710±.110	.650±.111	.710±.110	.707±.105
ro	maF1	**.755**±.049	.737±.052	.734±.046	.464±.042	.477±.047	.656±.054	.669±.062	.664±.053
	κ	**.511**±.097	.477±.103	.472±.092	.014±.049	.022±.063	.315±.106	.340±.122	.330±.105
fu	maF1	.703±.046	.735±.045	**.736**±.043	.499±.054	.527±.047	.698±.054	.723±.052	.723±.050
	κ	.528±.068	.573±.066	**.570**±.063	.293±.056	.326±.056	.522±.076	.557±.075	.560±.073
at	maF1	.679±.025	.712±.026	.710±.028	.724±.030	**.728**±.033	.724±.030	**.728**±.033	.724±.029
	κ	.365±.048	.424±.052	.421±.055	.449±.060	**.456**±.066	.449±.060	**.456**±.066	.448±.059
		simple	EG equal	EG best	MP	MP+p	MP+r	MP+p+r	MPϵ+p+r
cc	maF1	.817±.045	.860±.051	**.869**±.053	.780±.063	.831±.059	.780±.063	.831±.059	.823±.063
	κ	.634±.090	.720±.103	**.737**±.107	.559±.126	.661±.118	.559±.126	.661±.118	.647±.122
ro	maF1	**.750**±.045	.721±.051	.720±.047	.482±.053	.475±.047	.620±.064	.638±.057	.641±.062
	κ	**.502**±.090	.445±.098	.442±.092	.024±.068	.015±.060	.243±.126	.280±.114	.285±.122
fu	maF1	.671±.049	.707±.048	**.710**±.050	.489±.062	.514±.059	.642±.057	.681±.057	.677±.059
	κ	.475±.074	.529±.070	**.530**±.072	.254±.058	.296±.063	.440±.081	.491±.083	.486±.083
at	maF1	.663±.030	.692±.034	.693±.031	.707±.035	**.720**±.034	.707±.035	**.720**±.034	.713±.033
	κ	.333±.095	.384±.068	.386±.063	.414±.070	**.440**±.069	.414±.070	**.440**±.069	.427±.066

Table 3: Results for the full task: for German (above) and English (below), best values highlighted.

we can dispense with the extra step of optimizing the weighting and use the simple equal weighting. These result are consistent across both languages.

The pure labeled mstparser model (MP) performs worse than the base classifiers on all levels except for the attachment task. Adding pre-classification yields significant improvements on all levels but role identification. Using the external relation labeler drastically improves function classification and indirectly also role identification. The combined model (MP+p+r) yields best results for all mstparser models, but is still significantly outperformed by EG equal in all tasks except attachment classification. There, the mstparser models achieve best results, the improvement of MP+p+r over EG equal is significant for English ($p<0.0001$) and for German ($p=0.001$). Interestingly, the meta-model (MPϵ+p+r) which has access to its own features and to those of the pre-classification, but not to the features described in Section 5.3, performs nearly as good as or equal to the combined model (MP+p+r).

The only level not benefiting from any MST model in comparison with the base classifier is the role classification: In the final MST, the role of each segment is only implicitly represented, and can be determined by following the series of the role-switches of each argumentative function from the segment to the root. The loss of accuracy for predicting the argumentative role is much smaller for German than for English, probably due to the better attachment classification in the first place.

Finally, note that the EG best model gives the highest total score when summed over all levels, followed by EG equal and then MP+p+r.

Projecting further improvements: We have shown that joint prediction of all levels in the evidence graph models helps to improve the classification on single levels. To measure exactly how much a level contributes to the predictions of other levels, we simulate better base classifiers and study their impact. To achieve this, we artificially improved the classification of one target level by overwriting a percentage of its predictions with ground truth. The overwritten predictions where drawn randomly, corresponding to the label distribution of the target level. E.g. for a 20% improvement on the argumentative function level, the predictions of 20% of the true "attack"-items were set to attack and the predictions of 20% of the true "support"-items were set to support, irrespective of whether the classifier already chose the correct label.

The results of the simulations are presented in Figure 3 for English only, due to space constraints. The results for German exhibit the same trends. The figure plots the κ-score on the y-axis against the percentage of improvement on the x-axis. Artificially improved levels are drawn as a dashed line. As the first plot shows, function classifica-

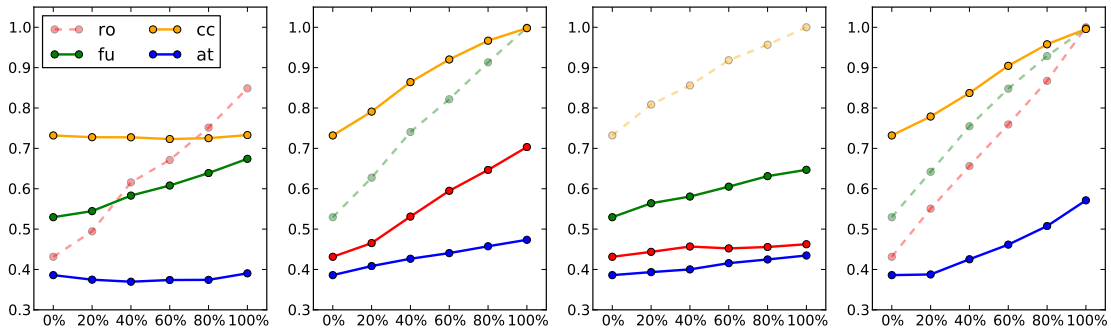

Figure 3: Simulations of the effect of better base classifiers in the EG equal model for English: dashed levels artificially improved, x = number of predictions overwritten with ground truth; y = average κ score in 10 iterations of 5fold CV.

tion is greatly improved by a better role classification (due to the logical connection between them), whereas the other levels are unaffected. In contrast, all levels would benefit from a better function classification, most importantly even the attachment classification. Potential improvements in the central claim identification mostly affect function classification (as these classification tasks partly overlap: central claims will not be assigned a function they cannot have). Finally, a combined improvement on the logically coupled task of role and function identification, would even more help the attachment classification. It might thus be useful to work on a better joint role and function classifier in near future.

Evidence combination: As pointed out by one reviewer, combining the evidence in an edge score as a weighted sum, see (5), instead of a product of probabilities might be inadequate and could result in a model that optimizes the highest scored but not the most probable structure. We compared the EG equal against an EG model with a product of probability. The model scores are nearly identical and do not show a significant difference.

7 Summary and Outlook

We introduced a new approach to argumentation mining that we applied to a parallel German/English corpus of 112 short texts. For the purposes of automatic mining, the original more fine-grained annotation in the corpus was reduced to a slightly simplified scheme consisting of support and attack relations among argumentative discourse units. We did not address the segmentation step here but focused on structure prediction, which we broke into a number of subtasks. Our new *evidence graph* model jointly predicts different aspects of the structure by combining the different subtask predictions in the edge weights of an evidence graph; we then apply a standard MST decoding algorithm. This model not only outperforms two reasonable baselines and two simple models for the difficult subtask of attachment/relation identification, but also improves the results for central claim identification and relation classification, and it compares favorably to a 3-pass mstparser pipeline.

To the best of our knowledge, this is the first data-driven model of argumentation structure that optimizes argumentation structure globally for the complete sequence of input segments. Furthermore, it is the first model jointly tackling segment type classification, relation identification and relation type classification.

Although a direct comparison with results from related work on other corpora is not possible, we can draw indirect comparisons. The first learned model without decoding (simple) is similar to the one presented by Stab and Gurevych (2014). Since it is outperformed by our joint MST decoding model on our data, we assume similar gains could be accomplished on their student essay dataset.

Our next step is to apply the method to other corpora and to more complex text, where the identification of non-participating segments (which are irrelevant for the argumentation) needs to be accounted for. Furthermore, we plan to investigate structured models that not only jointly predict but jointly learn the different aspects of the argumentation graph.

Acknowledgments

We are grateful to the anonymous reviewers for their thoughtful comments. We want to thank Željko Agić, Stergos Afantenos and Christoph Teichmann for fruitful discussions and Wladimir Sidorenko for feedback on an earlier version of the paper. The first author was supported by a grant from Cusanuswerk.

References

Jason Baldridge, Nicholas Asher, and Julie Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft*, 26:213–239.

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.

Jill Burstein and Daniel Marcu. 2003. A machine learning approach for identification thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467.

Y. J. Chu and T. H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.

Jack Edmonds. 1967. Optimum Branchings. *Journal of Research of the National Bureau of Standards*, 71B:233–240.

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.

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.

Saidul Kazi Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356. Asian Federation of Natural Language Processing.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June. Association for Computational Linguistics.

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.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005a. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005b. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December. The COLING 2012 Organizing Committee.

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

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, Maryland, June. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Andreas Peldszus and Manfred Stede. 2013. 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. to appear. An annotated corpus of argumentative microtexts. In *Proceedings of the First European Conference on Argumentation: Argumentation and Reasoned Action*, Lisbon, Portugal, June 2015.

Andreas Peldszus. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, U.S., June. Association for Computational Linguistics.

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)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

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.

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

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