

Automatic Detection of Arguments in Legal Texts

Marie-Francine Moens
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven, Belgium
marie-france.moens@law.kuleuven.be

Raquel Mochales Palau
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven, Belgium
raquel.mochalespalau@law.kuleuven.be

Erik Boiy
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven, Belgium
erik.boiy@law.kuleuven.be

Chris Reed
School of Computing
University of Dundee
Dundee DD1 4HN, Scotland, UK
chris@computing.dundee.ac.uk

ABSTRACT

This paper provides the results of experiments on the detection of arguments in texts among which are legal texts. The detection is seen as a classification problem. A classifier is trained on a set of annotated arguments. Different feature sets are evaluated involving lexical, syntactic, semantic and discourse properties of the texts. The experiments are a first step in the context of automatically classifying arguments in legal texts according to their rhetorical type and their visualization for convenient access and search.

Keywords

Argument recognition, discourse analysis, information extraction, machine learning.

1. INTRODUCTION

Information searches comprise a substantial amount of time of a legal professional. The general goal of information seeking is to build an argument to answer the problem at hand. He wants to find a viable argument that will support his or others' claims. Hence the current interest in information systems, such as the Araucaria project [23], that visualize the argument structure of a text. The manual structuring of an argumentative text into a graph visualization as is done in the Araucaria research is a very costly job. The Araucaria tool assists the drafting of the argumentation structure of a text by allowing to manually drag text into a graph structure that represents the argumentation.

We want to support this process with automated means, i.e., the automated recognition of an argumentation structure and its arguments in a legal text and the classification of an argument or set of arguments according to its argument type (e.g., counter argument, rebuttal). If we succeed, we can improve the visualization of the content, the indexing of

the texts for retrieval purposes, the automatic comparison of the content, and the assessment of the influence of arguments in decisions. Recognizing the argument structure in cases is also a valuable step in automatically qualifying arguments and describing them with conceptual factors from a legal ontology that play a role in the decision, and reasoning with these factors in a case-based reasoning system.

Given the practical need for automatic detection and classification of arguments in legal texts, we study a number of fundamental research questions of this task in the ACILA project (2006-2010). They regard the study of legal argumentation structures, the construction of a taxonomy of rhetorical discourse structures for the legal field and their linguistic cues, the natural language processing of legal texts, the automatic classification of the arguments according to their rhetorical type, and the convenient and user-friendly visualization and presentation to the user.

In this paper, the detection of arguments in texts among which are legal texts is seen as a classification problem. A classifier is trained on a set of annotated arguments. Different feature sets are evaluated involving lexical, syntactic, semantic and discourse properties of the texts. We give first some general background on the problem of argument detection and classification, followed by our methods for detecting arguments in texts, which include sections on feature selection and machine learning techniques. The next sections describe our experiments, results and their discussion. Before presenting the conclusions and our plans for future research, we give an overview of related research.

2. ARGUMENT DETECTION AND CLASSIFICATION

Argumentative texts are found in our daily discourse. We use arguments each time we want to persuade a party in the communication process. Argumentative texts are also common in the legal domain. For instance, one can find arguments in legislative texts (e.g., arguments sustaining a certain norm), in case law (e.g., arguments provided by the different parties or by the judge) and doctrinal texts (e.g., arguments with regard to a certain interpretation of a legal principle). Studies on legal reasoning by [2, 3, 21, 24, 25, 28] and others have built theoretical models of legal reasoning and represented argumentation structures in a logical formalism (e.g., propositional, first order predicate, deontic

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICAIL '07 June 4-8, Palo Alto, CA USA

Copyright 2007 ACM 978-1-59593-680-6/07/0006 ...\$5.00.

and defeasible logic, or claim lattices). Legal reasoning is usually performed in a context of debate and disagreement. Accordingly, such notions as arguments, moves, attacks, dialogue and burden of proof are studied. On a practical scale this research has resulted in dialogue and argumentation systems (e.g., [10]) that offer a useful interface by which users are guided when founding a hypothesis or conclusion, and as such have evaluated the theoretical models.

Based on recent work in legal argumentation theory [9, 26] combined with analysis of a corpus of texts, it may be possible to build a taxonomy (or ontology) of types of legal arguments and their relations, in much the same way as research that has been attempting to classify general argumentation schemes [22, 14]. In this article, we do not yet focus on this part of the research, and concentrate instead on how legal argumentation is reflected in written language. Although we lack studies on this subject, literature on general discourse theory and rhetorical structure analysis is inspiring. The rhetorical structure of a text is a main indicator of how information in that text is ordered into a coherent informational structure [17, 27, 13].

Linguistic phenomena that signal rhetorical relations are lexical cues, pronouns and other forms of phoric reference, and tense and aspect [13]. The most prominent indicators of rhetorical structure are lexical cues [1], most typically expressed by conjunctions and by certain kinds of adverbial groups. Research has shown that it is usually possible for humans to read a text and correctly assign one or more applicable rhetorical relations between the discourse segments, but having the machine to perform this task is much more difficult, because many textual cues are often missing or are ambiguous. Research also stresses the possibility of learning corpus-specific rhetorical cues from a corpus of training data [18] where it is possible to take into account many more features than the ones commonly considered.

Nevertheless, the idea to initiate research on argument detection and classification in text from the angle of rhetorical structure analysis seems to us a valuable one and is motivated as follows. First of all, a theory of the discourse structure of legal argumentation is currently lacking [5], while rhetorical discourse structures in general text are well studied and are backed by well-founded theories. Besides, these rhetorical relations, that claim to be based on taxonomies of general inference processes that go back to Hume and Aristotle [29, 15], might be mappable to argumentation types defined in legal theory. Secondly, linguistic methods such as a rhetorical structure analysis provides us a generic methodology for legal text analysis. This is especially interesting from a commercial point of view. In order to make systems for argument recognition in legal texts commercially viable, they need to be grounded in a generic framework. It is unthinkable that separate detection systems will be developed for separate domains of law. Language technology can be applied on all kinds of texts that are written in a certain language, hereby not a priori excluding that domain-specific argument types in legal texts might demand for some tailored approaches.

Even if we succeed in correctly typing the found arguments, there will still be the problem of implicit arguments that for their correct interpretation depend on external knowledge such as custom, historical development, models of social justice, and the purpose underlying legal warrants. This poses interesting questions on the world and common sense

knowledge that we need, and whether or not we can acquire this knowledge automatically from large text corpora.

Once the rhetorical relations are detected, an important issue is to define a representation that is suitable to capture the discourse structure and that allows easily accessing the information or making additional computations. We think here first of all of visualizing the argument structure of a text, which should be valuable for a legal professional when searching and assessing arguments for his or her legal case at hand. The Araucaria system mentioned above is a good example of presenting argument information to a user. An interesting research topic here is computing the salience of an argument in a discourse (e.g., arguments on issues that are most in dispute) and discovering how saliency can change over time, when the impact of an argument shifts by the influence of subsequent cases.

3. METHODS

In the experiments described in this paper, our objective is to detect argumentative sentences, where sentences are considered in isolation. We represent a sentence as a vector of features and train a classifier on examples that were manually annotated. We define generic features that can easily be extracted from the texts and study their contribution in the classification of sentences as argumentative. The feature vectors of these training examples will serve as input for state of the art classification algorithms. We obtained the best results with a multinomial naive Bayes classifier and a maximum entropy model.

3.1 Feature extraction

We use and motivate the following features:

- **Unigrams:** Each word in the sentence. This simple feature (along with the following two) can be considered as a baseline for more specific features. Punctuation marks (,;:?!) are not included in the basic version; if they are this is signaled with “(+p)”.
- **Bigrams:** Each pair of successive words.
- **Trigrams:** Each three successive words.
- **Adverbs:** Adverbs are detected with a part-of-speech (POS) tagger (in the experiments below we use QTag). They can signal argumentative information (e.g., in a “how” or “where” statement).
- **Verbs:** Verbs are also detected with a POS tagger. Only the main verbs (excluding “to be”, “to do” and “to have”) are considered. We would like to see how verbs compare to other parts-of-speech, like adverbs, for the detection of arguments.
- **Modal auxiliary:** A binary feature that indicates if a modal auxiliary is present. Modal auxiliaries indicate the level of necessity:
 - must/need to/have to = obligation, requirement, no choice
 - should/ought to/had better = recommendation
 - can/could = it is possible
 - may/might = option, choice
 - will/shall = intention, order
 - would = (counterfactual) condition

These verbs may give some indication for argumentation presence and are detected by the POS tagger.

Table 1: Examples of argumentative sentences per text type.

Text type	Argument	Non-argument
Discussion fora	“On this occasion, however, I shall not vote for any individual or party but will spoil my paper.”	“I have been voting since 1964 and at one time worked for my chosen party.”
Legal judgments	“He is aware of the risks involved, and he should bear the risks.”	“Let there be any misunderstanding one point should be clarified at the outset.”
Newspapers	“Labor no longer needs the Liberals in the Upper House.”	“The independents were a valuable sounding board for Labor’s reform plans.”
Parliamentary records	“I have accordingly disallowed the notice of question of privilege.”	“Copies of the comments of the Ministers have already been made available to Dr. Raghuvansh Prasad Singh.”
Weekly magazines	“But for anyone who visits Rajasthan’s Baran district, the apathy of the district administration and the failure of the Public Distribution System (pds) is clear to see”	“This time in Rajasthan.”

- **Word couples:** All possible combinations of two words in the sentence are considered. This approach captures more context than bigrams, at the expense of increasing the feature vector size substantially. For this reason we also did tests with cleaned couples. For the cleaning we removed the verbs “to be”, “to do” and “to have”, general determiners (a, the, this, that), proper nouns, pronouns and symbols.
- **Text statistics:** The following are considered:
 - *Sentence length:* well-built arguments may require more words than the average sentence.
 - *Average word length:* “difficult” words might make the argument look more impressive.
 - *Number of punctuation marks:* argumentation presence may increase the amount of punctuation needed in the sentence.
- **Punctuation:** The sequence of punctuation marks pre-sent in the sentence is used as a feature (e.g. “:.”). When a punctuation mark occurs more than once in a row, it is considered the same pattern (e.g. two or more successive commas both result in “,+”). Some of the resulting patterns may reflect a punctuation structure inside arguments, while others may rarely occur in argumentation.
- **Key words:** Keywords refer to 286 words or word sequences obtained from a list of terms indicative for argumentation [16]. Examples from the list are “but”, “consequently”, and “because of”.
- **Parse features:** In the parse tree of each sentence (Charniak [7]) we used the depth of the tree and the number of subclauses as features.

3.2 Classification algorithms

3.2.1 Multinomial naive Bayes classifier

A *naive Bayes* classifier is an example of a generative classifier which learns a model of the joint probability $p(x, y)$ and makes its predictions by using Bayes’ rule to calculate $p(y|x)$ and then selects the most likely label y . It makes the simplifying (naive) assumption that the individual features are conditionally independent given the class. The features are typically represented as binary values, i.e., the presence or absence of a feature in the object (here sentence). In a variation of this model, which is called *multinomial naive Bayes* classifier [20] (MNB) and which is often used in text categorization tasks, the number of occurrences of each feature is captured in the feature vector.

3.2.2 Maximum entropy model

This classifier adheres to the *maximum entropy* principle [4]. This principle states that, when we make inferences based on incomplete information, we should draw them from the probability distribution that has the maximum entropy permitted by the information we have. In natural language we often deal with incomplete patterns in our training set given the variety of natural language patterns that signal similar content. Hence, this type of classifier is often used in information extraction from natural language texts, which motivates our choice of this classifier. The maximum entropy classifier (called “Maxent” in the tables below) is an example of a discriminative classifier, which models the posterior probability $p(y|x)$ directly while learning a direct map from inputs x to the class label y . The features are described by binary variables called feature functions.

4. EXPERIMENTS AND DISCUSSION

4.1 The corpus

The Araucaria corpus comprises two distinct sets of data: a structured set in English collected and analysed according to a specific methodology as a part of a project at the University of Dundee, and an unstructured multi-lingual set of user-contributed analyses. Only the structured data was used for our analysis. The data was collected over a six week period in 2003, during which time a weekly regime of data collection scheduled regular harvest of one argument from each of 19 newspapers (from the UK, US, India, Australia, South Africa, Germany, China, Russia and Israel, in their English editions where appropriate), from 4 parliamentary records (in the UK, US and India), from 5 court reports (from the UK, US and Canada), from 6 magazines (UK, US and India), and from 14 further online discussion boards and “cause” sources such as Human Rights Watch and GlobalWarming.org. These sources were selected because they offered (a) long-term online archive of material; (b) free access to archive material; (c) reasonable likelihood of argumentation. Each week, the first argument enountered in each source was identified and analysed by hand. The skill of distinguishing argument from non-argument is sophisticated and requires training: it is a typical learning outcome of an undergraduate critical thinking course. The analysis of argument, including the categorisation of text by argumentation scheme, is more challenging yet, and faces the additional problem that multiple analyses may be possible (thereby reducing intercoder reliability). The Araucaria corpus analysis employed the rigorous scheme-based analysis approach of [14] to mitigate these problems. In the corpus there are 1899 sentences that contain an argument and 827

sentences without arguments, which we used for our experiments. 1072 new sentences containing no argument were extracted from the same sources as the ones used for the Araucaria project and added to the corpus. The sentences were classified by their text type: newspapers, parliamentary records, legal judgments, weekly magazines, discussion fora, cause information sources and speeches. We built a subcorpus for each text type by picking sentences as to have a maximum balanced set of positive and negative examples.

4.2 Evaluation

In an intrinsic evaluation correspondence is sought whether the argumentative sentences detected by the system correspond with the ones that were manually annotated. We compute here the accuracy of the detection as the number of correctly classified sentences divided by the number of sentences that were classified, and average the values obtained in a ten-fold cross-validation. We performed tests on the complete sentence set (Table 2) and on each subcorpus representing a specific text type (Table 3).

4.3 Results

As seen in Table 2, the simple, shallow features already yield acceptable results. Compared to a baseline where an off-the-shelf list of 286 keywords is used [16] (accuracy of 57.98% for the Maxent classifier), using the words of a text (unigram feature) results in a classification accuracy slightly above 70%. Unigrams include more word cues for the rhetorics than the keyword list and the keyword list does not perfectly fit argumentation discovery. Still, a better score is obtained by considering bi-grams or word couples in the category of lexical features. When considering syntactic features, the table shows that verbs and adverbs contribute to the argument classification, but on their own, these features are insufficiently discriminative. Parse features, i.e., exploiting the depth of the parse tree and the number of sub-clauses, are weaker patterns for argument detection. We did not test all feature combinations. Among the ones tested, we see that if we combine lexical features such as unigrams or word couples with knowledge of text statistics, punctuation, verbs or modal auxiliaries, we improve the accuracy. The best results are obtained by combined word couples, verbs and statistics on sentence length, average word length and number of punctuation marks (accuracy of 73.75%). For instance, the sentence “But, I’m still convinced there will be no solution without a 24/7 approach to conflict resolution by the United States.” could be classified as an argument based on the combination of these features, but was wrongly classified by solely using the individual features.

Table 3 gives us the results for the individual text types, where we have an indication that arguments in newspapers and arguments in legal texts are respectively the most easy and most difficult to detect. Explanations for the lower accuracy obtained for legal texts are: the small amount of training examples and - at least in this test set - more ambiguous argumentation patterns. Although many training examples are available, discussion fora score also lower than news. They contain more ambiguous and less well-formed texts compared to news. We did not train a classifier for finding arguments in the text types “cause information sources” and “speeches” because their number of training examples are smaller than 50.

A manual screening of 10 sentences from legal texts did

Table 2: Results in terms of average accuracy of training a classifier with different types of features. “# Fs” stands for the number of features, “# Ss” for the number of sentences.

Features	# Fs	MNB	Maxent
Unigrams	9238	73.06%	71.14%
Unigrams (+p)	9244	72.59%	71.35%
Bigrams	38078	71.09%	70.62%
Trigrams	50929	64.24%	64.67%
Adverbs	435	55.74%	58.87%
Verbs	2043	60.19%	61.16%
Modal auxiliary	1	49.76%	57.35%
Word couples	381429	71.17%	72.93%
Word couples (cleaned)	259857	70.38%	69.88%
Word couples (+p)	402193	71.67%	72.90%
Punctuation	147	57.29%	54.21%
Key words	108	53.32%	57.98%
Text statistics	3	58.48%	50.95%
Parse features	2	50.54%	50.26%
Unigrams & Text statistics	9247	73.12%	70.98%
Unigrams & Punctuation	9466	73.70%	71.03%
Word couples (+p) & Modal auxiliary	402194	71.64%	72.91%
Word couples (+p) & Text statistics	402196	73.70%	73.22
Word couples (+p) & Punctuation	402419	71.91%	73.28%
Word couples (+p) & Text statistics & Punctuation	402422	73.70%	73.28%
Word couples (+p) & Modal auxiliary & Text statistics	402197	73.70%	73.38%
Word couples (+p) & Verbs & Text statistics	404236	73.75%	72.59%
Word couples (+p) + Keywords & Text statistics & Verbs	404344	73.46%	72.72%

not give a derivation from accuracy figures found in the literature both for POS tagging (above 95% accuracy) and sentence parsing (around 90% recall and precision) [8].

4.4 Potential and limitations of the current approach

Our results confirm that simple features already are discriminative for argument detection and rhetorical structure recognition in general, which is in line with findings in the literature. Wolf and Gibson [30] give a list of conjunctive terms that illustrate coherence relations in text, many of them are word couples or single words with POS tag adverb or verb. Also, Marcu and Echihiabi [19] demonstrate that simple features can successfully identify discourse relations, if sufficient training data are available.

A set of 98 sentences that were erroneously classified (false positives and false negatives) were manually examined to identify the reason of the misclassification errors (Table 4). Some 21.4% of the sentences that are erroneously classified could be resolved, if one would consider the previous

Table 3: Results in terms of average accuracy of training a classifier with Word couples (+p), Verbs & Text statistics. “# Fs” stands for the number of features, “# Ss” for the number of sentences.

Text type	# Ss	# Fs	MNB	Maxent
Discussion fora	750	89613	71.73%	68.40%
Legal judgments	138	39681	65.94%	68.12%
Newspapers	702	119942	76.35%	73.22%
Parliamentary records	184	31207	72.83%	67.93%
Weekly magazines	176	33525	69.89%	69.32%

Table 4: Classification errors of a sample of 98 wrongly classified sentences when training a classifier with Word couples (+p), Verbs & Text statistics. “# Ss” stands for the number of sentences.

Cause of error	# Ss	%S
Modal verbs	19	19.3
The word “but”	9	9.2
Lack of context information	21	21.4
Ambiguous examples	18	18.3
No apparent indication found	31	31.6
Total	98	100

content in the discourse. In our experiments we have not yet included features that refer to content in previous sentences. As reported in the literature textual cues with regard to discourse structure detection, and to rhetorical structure detection in particular, are often ambiguous. This finding is confirmed with our results. Features such as “modal verbs” and the adverb “but” often behave ambiguously with regard to argument formulation (responsible for 28.5% of the errors in this set). For instance, in the sentence “I just wanted to make it clear, should it not have been crystal in my original post, I am not attacking T.J.” the modal verb “should” wrongly detects an argument. Combined with the 18.3% that represents other ambiguous sentences, ambiguity of the linguistic markers seem to be a major source of errors. For instance, the system incorrectly assigns an argument role to the sentence “You are treating me, father, more harshly than I deserve” based on the presence of the adverb “more”. The most difficult category of errors are those where the text does not give any cue for identifying an argument. Some reasoning steps are left implicit, and the precise logical connection between individual reasoning steps is not spelled out. This was also the case in some of the legal sentences in our error set. The most difficult case to solve is when there are no linguistic markers in the text and the argument detection depends on world and common sense knowledge which is not present in the text. Inevitably, the lack of linguistic markers or ambiguity of these markers also lead to different interpretations of the texts.

5. RELATED RESEARCH

Research in argument detection and classification in the legal domain and beyond is very limited. Perhaps the closest to our work is work done by Ben Hachey and Claire Grover

[11] at the University of Edinburgh. Their system trains a classifier on 141 House of Lords judgments and recognizes the rhetorical status of sentences of 47 judgments based on a number of textual features, where each judgment contains 105 sentences on average. A limited set of rhetorical labels is compiled composing of e.g., fact (the sentence recounts the events or circumstances which give rise to the legal proceedings), proceedings (the sentence describes legal proceedings taken in lower courts), background (the sentence is a direct quotation or citation of source of law material), framing (the sentence is part of the Lords’ argumentation), disposal (the sentence either credits or discredits a claim or previous ruling), textual (the sentence signals the structure of the document or contains formal content unrelated to a case), and other (default class). The authors rely on very simple features such as: location of a sentence within a document and within subsections and paragraphs; sentence length; whether the sentence contains a word from the title; whether the sentence contains significant terms spotted by the $tf \times idf$ (term frequency \times inverse document frequency) metric; whether the sentence contains a citation; linguistic features of the first finite verb; cue phrases; and the presence of certain named entity types. The authors trained different classifiers: decision tree learning algorithms, naive Bayes classification, support vector machines and maximum entropy modeling. Among the best results, the maximum entropy classifier shows a precision of 51% at a recall of 17% when precision and recall are averaged over the sentence categories mentioned above. Disposal sentences are most accurately recognized, while fact and background sentences yield a precision and recall of 0%. Argumentative sentences were recognized with 25% precision at a recall of 6%. A low recall indicates a lack of sufficient patterns for training. Precision errors might be due to the quite simplistic approach of feature extraction that was inspired by the classification of text segments in scientific articles. The authors do not analyze in depth the causes of the errors.

Beyond the legal field argument detection is also considered as a very innovative task. First attempts to detect arguments in mathematical discourse are described by [12]. We are not aware of argument recognition in other types of texts (e.g., speeches, political texts, ...).

The work of Stefanie Brüninghaus and Kevin Ashley [6] on classifying factual patterns (which often constitute arguments) with descriptive legal concepts (which the authors refer to as *factors*) is also worth mentioning. These authors address this difficult task of concept recognition in text with machine learning techniques.

6. CONCLUSIONS

The research in the ACILA project introduces the innovative area of research of argument detection and classification in text. The experiments on which we report here give us an initial assessment on the types of features that play a role in identifying legal arguments in single sentences, compare the results with identifying arguments in other types of texts, and pinpoint obvious problems when automatically extracting arguments from texts.

The simple features that we have tested yield already promising results while attaining an accuracy of the classification of almost 74% averaged over a variety of text types. For legal texts, this figure drops to 68%, but this might be due to the lack of sufficient training examples.

In our future work, argument detection and classification will be applied to the processing of multiple sentences (and clauses) in a discourse. Along comes the problem of segmentation. We will also focus on the classification of different types of arguments. We will start with the recognition of generic rhetorical relations in the texts and map these on the legal argument types in the literature. We will perform more fine-grained experiments with feature selection and extraction. This could also include the development of specific classifiers inspired by context dependent classification (e.g., relational Markov networks). We also expect that the meaning of the content of sentences will play a role when aiming at a more accurate classification.

The many practical consequences of our research have been discussed above. In addition, the research has the potential to bridge past theoretical research on argumentation and reasoning models in law and the current demand for automation in the legal practice. Finally, the content analysis of texts will continue to pinpoint problems of the use of natural language that hamper its automated processing and give guidelines on how to improve legal language when it is used in electronic communication.

7. ACKNOWLEDGMENTS

We would like to thank Xu Zhang and Pieter Hens who assisted with the implementation. The research was financed by the K.U.Leuven grant OT 03/06.

8. REFERENCES

- [1] J. Allen. *Natural Language Understanding (2nd ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1995.
- [2] K. Ashley. *Reasoning with Cases and Hypotheticals*. MIT, Cambridge, MA, 1990.
- [3] T. J. M. Bench-Capon. Argument in artificial intelligence and law. *Artif. Intell. Law*, 5(4):249–261, 1997.
- [4] A. L. Berger, S. D. Pietra, and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
- [5] K. Branting. *Reasoning with Rules and Precedents: a Computational Model of Legal Analysis*. Kluwer Academic Publishers, Dordrecht ; Boston, 2000.
- [6] S. Brüninghaus and K. D. Ashley. Generating legal arguments and predictions from case texts. In *ICAIL '05*, pages 65–74, New York, NY, USA, 2005. ACM.
- [7] E. Charniak. A maximum-entropy-inspired parser. Technical Report CS-99-12, 1999.
- [8] K. Deschacht and M.-F. Moens. *Text Processing Tools*. Technical Report, K.U.Leuven, 2006.
- [9] E. T. Feteris. A dialogical theory of legal discussions: Pragma-dialectical analysis and evaluation of legal argumentation. *Artif. Intell. Law*, 8(2/3):115–135, 2000.
- [10] T. Gordon. *The Pleadings Game*. Kluwer Academic Publishers, Boston, MA, 1995.
- [11] B. Hachey and C. Grover. Automatic legal text summarisation: Experiments with summary structuring. In *ICAIL '05*, pages 75–84, New York, NY, USA, 2005. ACM.
- [12] H. Horacek and M. Wolska. Interpreting semi-formal utterances in dialogs about mathematical proofs. *Data Knowl. Eng.*, 58(1):90–106, 2006.
- [13] E. H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.
- [14] J. Katzav and C. Reed. On argumentation schemes and the natural classification of arguments. *Argumentation*, 18, June 2004.
- [15] W. Kintsch. *Comprehension: A Paradigm for Cognition*. Cambridge University, 1998.
- [16] A. Knott and R. Dale. Using linguistic phenomena to motivate a set of rhetorical relations. Technical Report HCRC/RP-39, Edinburgh, Scotland, 1993.
- [17] W. Mann and S. Thompson. Rhetorical structure theory: Toward a functional theory of organization. Technical Report, 1988.
- [18] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT, Cambridge, MA, 2000.
- [19] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations, 2002.
- [20] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.
- [21] H. Prakken. Analysing reasoning about evidence with formal models of argumentation. *Law, Probability and Risk*, 3:33–50, 2004.
- [22] H. Prakken, C. Reed, and D. Walton. Argumentation schemes and generalisations in reasoning about evidence. In *ICAIL '03*, pages 32–41, New York, NY, USA, 2003. ACM.
- [23] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, (14 (3 - 4)):961 – 980, 2004.
- [24] G. Sartor. Legal reasoning, a cognitive approach to the law. In *A Treatise of Legal Philosophy and General Jurisprudence, Volume 5*. Springer, 2005.
- [25] S. E. Toulmin. *The Uses of Argument*. University Printing House, Cambridge, Bentley House, 200 Euston Road, London N.W.I., 1958.
- [26] B. Verheij. *Virtual Arguments: On the Design of Argument Assistants for Lawyers and Other Arguers*. T M C Asser, The Hague, 1998.
- [27] C. M. W. Mann and S. Thompson. Rhetorical structure theory and text analysis. In *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pages 39–78. Pragmatics and Beyond New Series, 1992.
- [28] D. Walton. *Argumentation Methods for Artificial Intelligence in Law*. Springer, 2005.
- [29] T. Winograd. *Understanding Natural Language*. Academic, Inc., Orlando, FL, USA, 1972.
- [30] F. Wolf and E. Gibson. *Coherence in Natural Language*. MIT, Cambridge, Massachusetts, USA, 2006.