Identifying High-Level Organizational Elements in Argumentative Discourse

Nitin Madnani Michael Heilman Joel Tetreault
Educational Testing Service
Princeton, NJ, USA
{nmadnani,mheilman,jtetreault}@ets.org

Martin Chodorow
Hunter College of CUNY
New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

Argumentative discourse contains not only language expressing claims and evidence, but also language used to organize these claims and pieces of evidence. Differentiating between the two may be useful for many applications, such as those that focus on the content (e.g., relation extraction) of arguments and those that focus on the structure of arguments (e.g., automated essay scoring). We propose an automated approach to detecting high-level organizational elements in argumentative discourse that combines a rule-based system and a probabilistic sequence model in a principled manner. We present quantitative results on a dataset of human-annotated persuasive essays, and qualitative analyses of performance on essays and on political debates.

1 Introduction

When presenting an argument, a writer or speaker usually cannot simply state a list of claims and pieces of evidence. Instead, the arguer must explicitly structure those claims and pieces of evidence, as well as explain how they relate to an opponent's argument. Consider example 1 below, adapted from an essay rebutting an opponent's argument that grizzly bears lived in a specific region of Canada.

> **The argument states that** based on the result of the recent research, there probably were grizzly bears in Labrador. **It may seem reasonable at first glance, but actually, there are some logical mistakes in it.** . . . **There is a possibility that** they were a third kind of bear apart from black and grizzly bears. Also, the explorer accounts were recorded in the nineteenth century, which was more than 100 years ago. . . . **In sum, the conclusion of this argument is not reasonable since the account and the research are not convincing enough.** . . .

The argument begins by explicitly restating the opponent's claim, prefacing the claim with the phrase "The argument states that." Then, the second sentence explicitly marks the opponent's argument as flawed. Later on, the phrase "There is a possibility that" indicates the subsequent clause introduces evidence contrary to the opponent's claim. Finally, the sentence "In sum, . . ." sums up the arguer's stance in relation to the opponent's claim.[1]

As illustrated in the above example, argumentative discourse can be viewed as consisting of language used to express claims and evidence, and language used to organize them. We believe that differentiating organizational elements from content would be useful for analyzing persuasive discourse.

[1]The word *Also* signals that additional evidence is about to be presented and should also be marked as shell. However, it was not marked in this specific case by our human annotator (§3.2).

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 20–28,
Montréal, Canada, June 3-8, 2012.

We refer to such organizational elements as **shell**, indicating that they differ from the specific claims and evidence, or "meat," of an argument. In this work, we develop techniques for detecting shell in texts. We envision potential applications in political science (e.g., to better understand political debates), information extraction or retrieval (e.g., to help a system focus on content rather than organization), and automated essay scoring (e.g., to analyze the quality of a test-taker's argument), though additional work is needed to determine exactly how to integrate our approach into such applications.

Detecting organizational elements could also be a first step in parsing an argument to infer its structure. We focus on this initial step, leaving the other steps of categorization of spans (as to whether they evaluate the opponent's claims, connect one's own claims, etc.), and the inference of argumentation structure to future work.

Before describing our approach to identifying shell, we begin by defining it. Shell refers to sequences of words used to refer to claims and evidence in persuasive writing or speaking, providing an organizational framework for an argument. It may be used by the writer or the speaker in the following ways:

- to declare one's own claims (e.g., "There is the possibility that")
- to restate an opponent's claims (e.g., "The argument states that")
- to evaluate an opponent's claims (e.g., "It may seem reasonable at first glance, but actually, there are some logical mistakes in it")
- to present evidence and relate it to specific claims (e.g., "To illustrate my point, I will now give the example of")

There are many ways of analyzing discourse. The most relevant is perhaps rhetorical structure theory (RST) (Mann and Thompson, 1988). To our knowledge, the RST parser from Marcu (2000) is the only RST parser readily available for experimentation. The parser is trained to model the RST corpus (Carlson et al., 2001), which treats complete clauses (i.e., clauses with their obligatory complements) as the elementary units of analysis. Thus, the parser treats the first sentence in example 1 as a single unit and does not differentiate between the main and subordinate clauses. In contrast, our approach distinguishes the sequence "The argument states that . . . " as shell (which is used here to restate the external claim). Furthermore, we identify the entire second sentence as shell (here, used to evaluate the external claim), whereas the RST parser splits the sentence into two clauses, "It may seem . . ." and "but actually . . .", linked by a "contrast" relationship.[2] Finally, our approach focuses on explicit markers of organizational structure in arguments, whereas RST covers a broader range of discourse connections (e.g., elaboration, background information, etc.), including implicit ones. (Note that additional related work is described in §6.)

This work makes the following contributions:

- We describe a principled approach to the task of detecting high-level organizational elements in argumentative discourse, combining rules and a probabilistic sequence model (§2).
- We conduct experiments to validate the approach on an annotated sample of essays (§3, §4).
- We qualitatively explore how the approach performs in a new domain: political debate (§5).

2 Detection Methods

In this section, we describe three approaches to the problem of shell detection: a rule-based system (§2.1), a supervised probabilistic sequence model (§2.2), and a simple lexical baseline (§2.3).

2.1 Rule-based system

We begin by describing a knowledge-based approach to detecting organizational elements in argumentative discourse. This approach uses a set of 25 hand-written regular expression patterns.[3]

In order to develop these patterns, we created a sample of 170 annotated essays across 57 distinct prompts.[4] The essays were written by test-takers of a standardized test for graduate admissions. This sample of essays was similar in nature to but did not overlap with those discussed in other sections

[2] We used the RST parser of Marcu (2000) to analyze the original essay from which the example was adapted.

[3] We use the PyParsing toolkit to parse sentences with the grammar for the rule system.

[4] Prompts are short texts that present an argument or issue and ask test takers to respond to it, either by analyzing the given argument or taking a stance on the given issue.

MODAL → do | don't | can | cannot | will | would | ...

ADVERB → strongly | totally | fundamentally | vehemently | ...

AGREEVERB → disagree | agree | concur | ...

AUTHORNOUN → writer | author | speaker | ...

SHELL → I [MODAL] [ADVERB] AGREEVERB with the AUTHORNOUN

Figure 1: An example pattern that recognizes shell language describing the author's position with respect to an opponent's, e.g., *I totally agree with the author* or *I will strongly disagree with the speaker*.

of the paper (§2.2, §3.2). The annotations were carried out by individuals experienced in scoring persuasive writing. No formal annotation guidelines were provided. Besides shell language, there were other annotations relevant to essay scoring. However, we ignored them for this study because they are not directly relevant to the task of shell language detection.

From this sample, we computed lists of n-grams ($n = 1, 2, \ldots, 9$) that occurred more than once in essays from at least half of the 57 distinct essay prompts. We then wrote rules to recognize the shell language present in the n-gram lists. Additional rules were added to cover instances of shell that we observed in the annotated essays but that were not frequent enough to appear in the n-gram analysis.

We use "Rules" to refer to this method.

2.2 Supervised Sequence Model

The next approach we describe is a supervised, probabilistic sequence model based on conditional random fields (CRFs) (Lafferty et al., 2001), using a small number of general features based on lexical frequencies. We assume access to a labeled dataset of N examples $(\mathbf{w}, \mathbf{y})$ indexed by i, containing sequences of words $w^{(i)}$ and sequences of labels $y^{(i)}$, with individual words and labels indexed by j (§3 describes our development and testing sets). $y^{(i)}$ is a sequence of binary values, indicating whether each word $w_j^{(i)}$ in the sequence is shell ($y_j^{(i)} = 1$) or not ($y_j^{(i)} = 0$). Following Lafferty et al. (2001), we find a parameter vector θ that maximizes the following log-likelihood objective function:

$$L(\theta|\mathbf{w}, \mathbf{y}) = \sum_{i=1}^{N} \log p\left(y^{(i)} \mid w^{(i)}, \theta\right) \quad (1)$$
$$= \sum_{i=1}^{N} \left(\theta^\top \mathbf{f}(w^{(i)}, y^{(i)}) - \log Z^{(i)}\right)$$

The normalization constant Z_i is a sum over all possible label sequences for the ith example, and $\mathbf{f}$ is a feature function that takes pairs of word and label sequences and returns a vector of feature values, equal in dimensions to the number of parameters in θ.[5]

The feature values for the jth word and label pair are as follows (these are summed over all elements to compute the values of $\mathbf{f}$ for the entire sequence):

- The relative frequency of $w_j^{(i)}$ in the British National Corpus.
- The relative frequency of $w_j^{(i)}$ in a set of 100,000 essays (see below).
- Eight binary features for whether the above frequencies meet or exceed the following thresholds: $10^{\{-6,-5,-4,-3\}}$.
- The proportion of prompts for which $w_j^{(i)}$ appeared in at least one essay about that prompt in the set of 100,000.
- Three binary features for whether the above proportion of prompts meets or exceeds the following thresholds: $\{0.25, 0.50, 0.75\}$.
- A binary feature with value 1 if $w_j^{(i)}$ consists only of letters a-z, and 0 otherwise. This feature distinguishes punctuation and numbers from other tokens.

[5] We used CRFsuite 0.12 (Okazaki, 2007) to implement the CRF model.

- A binary feature with value 1 if the rule-based system predicts that $w_j^{(i)}$ is shell, and 0 otherwise.
- A binary feature with value 1 if the rule-based system predicts that $w_{j-1}^{(i)}$ is shell, and 0 otherwise.
- Two binary features for whether or not the current token was the first or last in the sentence, respectively.
- Four binary features for the possible transitions between previous and current labels ($y_j^{(i)}$ and $y_{j-1}^{(i)}$, respectively).

To define the features related to essay prompts and lexical frequencies in essays, we created a set of 100,000 essays from a larger set of essays written by test-takers of a standardized test for graduate admissions (the same domain as in §2.1). The essays were written in response to 228 different prompts that asked students to analyze various issues or arguments. We use additional essays sampled from this source later to acquire annotated training and test data (§3.2).

We developed the above feature set using cross-validation on our development set (§3). The intuition behind developing the word frequency features is that shell language generally consists of chunks of words that occur frequently in persuasive language (e.g., "claims," "conclude") but not necessarily as frequently in general text (e.g., the BNC). The sequence model can also learn to disprefer changes of state, such that multi-word subsequences are labeled as shell even though some of the individual words in the subsequence are stop words, punctuation, etc.

Note there are a relatively small number of parameters in the model,[6] which allows us to estimate parameters on a relatively small set of labeled data. We briefly experimented with adding an ℓ_2 penalty on the magnitude of θ in Equation 2, but this did not seem to improve performance.

When making predictions $\hat{y}^{(i)}$ about the label sequence for a new sentence, the most common approach is to find the most likely sequence of labels y given the words $w^{(i)}$, found with Viterbi decoding:

$$\hat{y}^{(i)} = \operatorname*{argmax}_y p_\theta(y \mid w^{(i)}) \quad (2)$$

We use "CRF$_v$" to refer to this approach. We use the suffix "+R" to denote models that include the two rule-based system prediction features, and we use "-R" to denote models that exclude these two features.

In development, we observed that this decoding approach seemed to very strongly prefer labeling an entire sentence as shell or not, which is often not desirable since shell often appears at just the beginnings of sentences (e.g., "The argument states that").

We therefore test an alternative prediction rule that works at the word-level, rather than sequence-level. This approach labels each word as shell if the sum of the probabilities of all paths in which the word was labeled as shell—that is, the marginal probability—exceeds some threshold λ. Words are labeled as non-shell otherwise. Specifically, an individual word $w_j^{(i)}$ is labeled as shell (i.e., $\hat{y}_j^{(i)} = 1$) according to the following equation, where $1(q)$ is an indicator function that returns 1 if its argument q is true, and 0 otherwise.

$$\hat{y}_j^{(i)} = \quad 1\left(\left(\sum_y p_\theta(y \mid w^{(i)})\, y_j\right) \geq \lambda\right) \quad (3)$$

We tune λ using the development set, as discussed in §3.

We use "CRF$_m$" to refer to this approach.

2.3 Lexical Baseline

As a simple baseline, we also evaluated a method that labels words as shell if they appear frequently in persuasive writing—specifically, in the set of 100,000 unannotated essays described in §2.2. In this approach, word tokens are marked as shell if they belonged to the set of k most frequent words from the essays. Using the development set discussed in §3.2, we tested values of k in $\{100, 200, \ldots, 1000\}$. Setting $k = 700$ led to the highest F_1.

We use "TopWords" to refer to this method.

[6] There were 42 parameters in our implementation of the full CRF model. Excluding the four transition features, each of the 19 features had two parameters, one for the positive class and one for the negative class. Having two parameters for each is unnecessary, but we are not aware of how to have the crfsuite toolkit avoid these extra features.

3 Experiments

In this section, we discuss the design of our experimental evaluation and present results on our development set, which we used to select the final methods to evaluate on the held-out test set.

3.1 Metrics

In our experiments, we evaluated the performance of the shell detection methods by comparing token-level system predictions to human labels. Shell language typically occurs as fairly long sequences of words, but identifying the exact span of a sequence of shell seems less important than in related tagging tasks, such as named entity recognition. Therefore, rather than evaluating based on spans (either with exact or a partial credit system), we measured performance at the word token-level using standard metrics: precision, recall, and the F_1 measure. For example, for precision, we computed the proportion of tokens predicted as shell by a system that were also labeled as shell in our human-annotated datasets.

3.2 Annotated Data

To evaluate the methods described in §2, we gathered annotations for 200 essays that were not in the larger, unannotated set discussed in §2.2. We split this set of essays into a development set of 150 essays (68,601 word tokens) and a held-out test set of 50 essays (21,277 word tokens). An individual with extensive experience at scoring persuasive writing and familiarity with shell language annotated all tokens in the essays with judgments of whether they were shell or not (in contrast to §2.1, this annotation only involved labeling shell language).

From the first annotator's judgments on the development set, we created a set of annotation guidelines and trained a second annotator. The second annotator marked the held-out test set so that we could measure human agreement. Comparing the two annotators' test set annotations, we observed agreement of $F_1 = 0.736$ and Cohen's $\kappa = 0.699$ (we do not use κ in our experiments but report it here since it is a common measure of human agreement). Except for measuring agreement, we did not use the second annotator's judgments in our experiments.[7]

[7]In the version of this paper submitted for review, we measured test set agreement with judgments from a third individual, who was informally trained by the first, without the formal guidelines. Agreement was somewhat lower: $F_1 = 0.668$ and $\kappa = 0.613$.

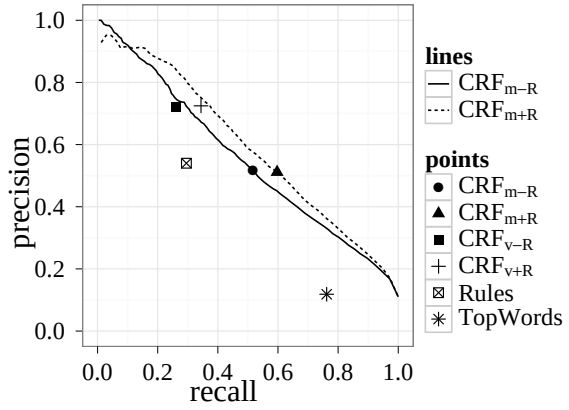

Figure 2: Precision and recall of the detection methods at various thresholds, computed through cross-validation on the development set. Points indicate performance for the rule-based and baseline system as well as points where F_1 is highest.

3.3 Cross-validation Results

To develop the CRF's feature set, to tune hyperparameters, and to select the most promising systems to evaluate on the test set, we randomly split the sentences from the development set into two halves and conducted tests with two-fold cross-validation.

We tested thresholds for the CRF at $\lambda = \{0.01, 0.02, \ldots, 1.00\}$.

Figure 2 shows the results on the development set. For the rule-based system, which did not require labeled data, performance is computed on the entire development set. For the CRF approaches, the precision and recall were computed after concatenating predictions on each of the cross-validation folds.

The TopWords baseline performed quite poorly, with $F_1 = 0.205$. The rule-based system performed much better, with $F_1 = 0.382$, but still not as well as the CRF systems. The CRF systems that predict maximum sequences had $F_1 = 0.382$ without the rule-based system features (CRF$_{v-R}$), and $F_1 = 0.467$ with the rule-based features (CRF$_{v+R}$). The CRF systems that made predictions from marginal scores performed best, with $F_1 = 0.516$ without the rule-based features, and $F_1 = 0.551$ with the rule-based features. Thus, both the rule-based sys-

Method	P	R	F_1	Len
TopWords	0.125	0.759	0.214 $*$	2.80
Rules	0.561	0.360	0.439 $*$	4.99
CRF$_{v-R}$	0.729	0.268	0.392 $*$	15.67
CRF$_{v+R}$	0.763	0.369	0.498 $*$	13.30
CRF$_{m-R}$	0.586	0.574	0.580	9.00
CRF$_{m+R}$	0.556	0.670	0.607	9.96
Human	0.685	0.796	0.736 $*$	7.91

Table 1: Performance on the held-out test set, in terms of precision (P), recall (R), F_1 measure, and average length in tokens of sequences of one or more words labeled as shell (Len). $*$ indicates F_1 scores that are statistically reliably different from CRF$_{m+R}$ at the $p < 0.01$ level.

tem features and the marginal prediction approach led to gains in performance.

From an examination of the predictions from the CRF$_{m+R}$ and CRF$_{m-R}$ systems, it appears that a major contribution of the features derived from the rule-based system is to help the hybrid CRF$_{m+R}$ system avoid tagging entire sentences as shell when only parts of them are actually shell. For example, consider the sentence "According to this statement, the speaker asserts that technology can not only influence but also determine social customs and ethics" (typographical errors included). CRF$_{m-R}$ tags everything up to "determine" as shell, whereas the rule-based system and CRF$_{m+R}$ correctly stop after "asserts that."

4 Test Set Results

Next, we present results on the held-out test set. For the CRF$_m$ systems, we used the thresholds that led to the highest F_1 scores on the development set ($\lambda = 0.26$ for CRF$_{m+R}$ and $\lambda = 0.32$ for CRF$_{m-R}$). Table 1 presents the results for all systems, along with results comparing the second annotator's labels ("Human") to the gold standard labels from the first annotator.

The same pattern emerged as on the development set, with CRF$_{m+R}$ performing the best. The F_1 score of 0.607 for the CRF$_{m+R}$ system was relatively close to the F_1 score of 0.736 for agreement between human annotators. To test whether CRF$_{m+R}$'s relatively high performance was due to chance, we computed 99% confidence intervals for the differences in F_1 score between CRF$_{m+R}$ and each of the other methods. We used the bias-corrected and accelerated (BC$_a$) Bootstrap (Efron and Tibshirani, 1993) with 10,000 rounds of resampling at the sentence level for each comparison. A difference is statistically reliable at the α level (i.e., $p < \alpha$) if the $(1 - \alpha)\%$ confidence interval for the difference does not contain zero, which corresponds to the null hypothesis. Statistically reliable differences are indicated in Table 1. The only system that did not have a reliably lower F_1 score than CRF$_{m+R}$ was CRF$_{m-R}$, though due to the relatively small size of our test set, we do not take this as strong evidence against using the rule-based system features in the CRF.

We note that while the CRF$_{m+R}$ system had lower precision (0.556) than the CRF$_{v+R}$ system (0.763), its threshold λ could be tuned to prefer high precision rather than the best development set F_1. Such tuning could be very important depending on the relative costs of false positives and false negatives for a particular application.

We also computed the mean length of sequences of one or more contiguous words labeled as shell. Here also, we observed that the CRF$_{m+R}$ approach provided a close match to human performance. The mean lengths of shell for the first and second annotators were 8.49 and 7.91 tokens, respectively. For the CRF$_{m+R}$ approach, the mean length was slightly higher at 9.96 tokens, but this was much closer to the means of the human annotators than the mean for the CRF$_{v+R}$ system, which was 13.30 tokens. For the rule-based system, the mean length was 4.99 tokens, indicating that it captures short sequences such as "In addition," more often than the other systems.

5 Observations about a New Domain

In this section, we apply our system to a corpus of transcripts of political debates[8] in order to understand whether the system can generalize to a new domain with a somewhat different style of argumentation. Our analyses are primarily qualitative in nature due to the lack of gold-standard annotations. We chose two historically well-known debates

[8]The Lincoln–Douglas debates were downloaded from `http://www.bartleby.com/251/`. The other debates were downloaded from `http://debates.org/`.

(Lincoln–Douglas from 1858 and Kennedy–Nixon from 1960) and two debates that occurred more recently (Gore–Bush from 2000 and Obama–McCain from 2008). These debates range in length from 38,000 word tokens to 65,000 word tokens.

Political debates are similar to the persuasive essays we used above in that debate participants state their own claims and evidence as well as evaluate their opponents' claims. They are different from essays in that they are spoken rather than written—meaning that they contain more disfluencies, colloquial language, etc.—and that they cover different social and economic issues. Also, the debates are in some sense a dialogue between two people.

We tagged all the debates using the CRF_{m+R} system, using the same parameters as for the test set experiments (§4).

First, we observed that a smaller percentage of tokens were tagged as shell in the debates than in the essays. For the annotated essay test set (§3.2), the percentage of tokens tagged as shell was 14.0% (11.6% were labeled as shell by the first annotator). In contrast, the percentage of tokens tagged as shell was 4.2% for Lincoln–Douglas, 5.4% for Kennedy–Nixon, 4.6% for Gore–Bush, and 4.8% for Obama–McCain. It is not completely clear whether the smaller percentages tagged as shell are due to a lack of coverage by the shell detector or more substantial differences in the domain.

However, it seems that these debates genuinely include less shell. One potential reason is that many of the essay prompts asked test-takers to respond to a particular argument, leading to responses containing many phrases such as "The speaker claims that" and "However, the argument lacks specificity . . . ".

We analyzed the system's predictions and extracted a set of examples, some of which appear in Table 2, showing true positives, where most of the tokens appear to be labeled correctly as shell; false positives, where tokens were incorrectly labeled as shell; and false negatives, where the system missed tokens that should have been marked.

Table 2 also provides some examples from our development set, for comparison.

We observed many instances of correctly marked shell, including many that appeared very different in style than the language used in essays. For example, Lincoln demonstrates an aggressive style in the following: "Now, I say that there is no charitable way to look at that statement, except to conclude that he is actually crazy." Also, Bush employs a somewhat atypical sentence structure here: "It's not what I think and its not my intentions and not my plan."

However, the system also incorrectly tagged sequences as shell, particularly in short sentences (e.g., "Are we as strong as we should be?"). It also missed shell, partially or entirely, such as in the following example: "But let's get back to the core issue here."

These results suggest that although there is potential for improvement in adapting to new domains, our approach to shell detection at least partially generalizes beyond our initial domain of persuasive essay writing.

6 Related Work

There has been much previous work on analyzing discourse. In this section, we describe similarities and differences between that work and ours.

Rhetorical structure theory (Mann and Thompson, 1988) is perhaps the most relevant area of work. See §1 for a discussion.

In research on intentional structure, Grosz and Sidner (1986) propose that any discourse is composed of three interacting components: the linguistic structure defined by the actual utterances, the intentional structure defined by the purposes underlying the discourse, and an attentional structure defined by the discourse participants' focus of attention. Detecting shell may also be seen as trying to identify explicit cues of intentional structure in a discourse. Additionally, the categorization of shell spans as to whether they evaluate the opponents claims, connect ones own claims, etc., may be seen as determining what Grosz and Sidener call "discourse segment purposes" (i.e., the intentions underlying the segments containing the shell spans).

We can also view shell detection as the task of identifying phrases that indicate certain types of speech acts (Searle, 1975). In particular, we aim to identify markers of assertive speech acts, which declare that the speaker believes a certain proposition, and expressive speech acts, which express attitudes toward propositions.

Shell also overlaps with the concept of discourse markers (Hutchinson, 2004), such as "however" or

LINCOLN (L) — DOUGLAS (D) DEBATES	
TP	**L: Now, I say that there is no charitable way to look at that statement, except to conclude that he is actually crazy.** **L: The first thing I see fit to notice is the fact that** ...
FP	**D:** He became noted as the author of the scheme to ... **D:** ... such amendments were to be made to it as would render it useless and inefficient ...
FN	**D: I wish to impress it upon you,** that every man who voted for those resolutions ... **L: That statement he makes, too, in the teeth of the knowledge that I had made the stipulation to come down here** ...

KENNEDY (K) — NIXON (N) DEBATES	
TP	**N: I favor that because I believe that's the best way to** aid our schools ... **N: And in our case, I do believe that** our programs will stimulate the creative energies of ...
FP	**N:** We are for programs, in addition, which will see that our medical care for the aged ... **K:** Are we as strong as we should be?
FN	**K: I should make it clear that I do not think** we're doing enough ... **N: Why did Senator Kennedy take that position then? Why do I take it now?**

BUSH (B) — GORE (G) DEBATES	
TP	**B: It's not what I think and its not my intentions and not my plan.** **G:** And FEMA has been a major flagship project of our reinventing government efforts. **And I agree, it works extremely well now.**
FP	**B: First of all,** most of this is at the state level. **G:** And it focuses not only on increasing the supply, **which I agree we have to do,** but also on ...
FN	**B: My opponent thinks** the government—the surplus is the government's money. **That's not what I think** **G: I strongly support** local control, **so does Governor Bush.**

OBAMA (O) — MCCAIN (M) DEBATES	
TP	**M: But the point is—the point is,** we have finally seen Republicans and Democrats sitting down and negotiating together ... **O: And one of the things I think we have to do is make** sure that college is affordable ...
FP	**O:** ... but in the short term there's an outlay and we may not see that money for a while. **O:** We have to do that now, because it will actually make our businesses and our families better off.
FN	**O: So I think the lesson to be drawn is that** we should never hesitate to use military force ... to keep the American people safe. **O: But let's get back to the core issue here.**

PERSUASIVE ESSAYS (DEVELOPMENT SET, SPELLING ERRORS INCLUDED)	
TP	**However, the argument lacks specificity and relies on too many questionable assumptions to make a strong case for** adopting an expensive and logistically complicated program. **I believe that both of these claims have been made in hase and other factors need to be considered.**
FP	Since they are all far from now, **the prove is not strong enough to support the conclusion.** As we know that one mind can not think as the other does.
FN	**History has proven that** ... **The given issue which states that** in any field of inquiry ... **is a controversional one.**

Table 2: Examples of CRF_{m+R} performance. Underlining marks tokens predicted to be shell, and bold font indicates shell according to human judgments (our judgments for the debate transcripts, and the annotator's judgments for the development set). Examples include true positives (**TP**), false positives (**FP**), and false negatives (**FN**). Note that some **FP** and **FN** examples include partially accurate predictions.

"therefore." Discourse markers, however, are typically only single words or short phrases that express a limited number of relationships. On the other hand, shell can capture longer sequences that express more complex relationships between the components of an argumentative discourse (e.g., "But let's get back to the core issue here" signals that the following point is more important than the previous one).

There are also various other approaches to analyzing arguments. Notably, much recent theoretical research on argumentation has focused on argumentation schemes (Walton et al., 2008), which are high-level strategies for constructing arguments (e.g., argument from consequences). Recently, Feng and Hirst (2011) developed automated methods for classifying texts by argumentation scheme. In similar work, Anand et al. (2011) use argumentation schemes to identify tactics in blog posts (e.g., moral appeal, social generalization, appeals to external authorities etc.). Although shell language can certainly be found in persuasive writing, it is used to organize the persuader's tactics and claims rather than to express them. For example, consider the following sentence: "**It must be the case that** this diet works **since** it was recommended by someone who lost 20 pounds on it." In shell detection, we focus on the lexico-syntactic level, aiming to identify the bold words as shell. In contrast, work on argumentation schemes focuses at a higher level of abstraction, aiming to classify the sentence as an attempt to persuade by appealing to an external authority.

7 Conclusions

In this paper, we described our approach to detecting language used to explicitly structure an arguer's claims and pieces of evidence as well as explain how they relate to an opponent's argument. We implemented a rule-based system, a supervised probabilistic sequence model, and a principled hybrid version of the two. We presented evaluations of these systems using human-annotated essays, and we observed that the hybrid sequence model system performed the best. We also applied our system to political debates and found evidence of the potential to generalize to new domains.

Acknowledgments

We would like to thank the annotators for helping us create the essay data sets. We would also like to thank James Carlson, Paul Deane, Yoko Futagi, Beata Beigman Klebanov, Melissa Lopez, and the anonymous reviewers for their useful comments on the paper and annotation scheme.

References

P. Anand, J. King, J. Boyd-Graber, E. Wagner, C. Martell, D. Oard, and P. Resnik. 2011. Believe me–we can do this! annotating persuasive acts in blog text. In *Proc. of AAAI Workshop on Computational Models of Natural Argument*.

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proc. of the Second SIGdial Workshop on Discourse and Dialogue*.

B. Efron and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

V. W. Feng and G. Hirst. 2011. Classifying arguments by scheme. In *Proc. of ACL*.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Comput. Linguist.*, 12(3):175–204.

B. Hutchinson. 2004. Acquiring the meaning of discourse markers. In *Proc. of ACL*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

W. C. Mann and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3).

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

N. Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

J. R. Searle. 1975. A classification of illocutionary acts. *Language in Society*, 5(1).

D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.