LAMP-TR-147 January 2008
HCIL-2008-02

Elements of a Computational Model for Multi-Party Discourse: The Turn-Taking Behavior of Supreme Court Justices

Timothy Hawes, Jimmy Lin, and Philip Resnik

University of Maryland
College Park, Maryland, USA
E-mail: {twhawes,jimmylin}@umd.edu, resnik@glue.umd.edu

Abstract

This paper explores computational models of multi-party discourse, using transcripts from U.S. Supreme Court oral arguments. The turn-taking behavior of participants is treated as a supervised sequence labeling problem and modeled using first- and second-order Conditional Random Fields. We specifically explore the hypothesis that discourse markers and personal references provide important features in such models. Results from a sequence prediction experiment demonstrate that incorporating these two types of features yields significant improvements in performance. This work is couched in the broader context of developing tools to support legal scholarship, although we see other NLP applications as well.

Publication Date: January 14, 2008
Keywords: Discourse Markers, Personal References, U.S. Supreme Court, Conditional Random Fields

Please cite as: Timothy Hawes, Jimmy Lin, and Philip Resnik. Elements of a Computational Model for Multi-Party Discourse: The Turn-Taking Behavior of Supreme Court Justices. Technical Report LAMP-TR-147/HCIL-2008-02, University of Maryland, College Park, January 2008.

1 Introduction

Legal scholarship, like many other areas of the social sciences, is a text-centered endeavor. Typical activities include combing through court opinions, litigant briefs, secondary review articles, and other documents in search of support for a particular hypothesis. Despite the growing availability of these texts in electronic formats, analysis is still largely performed by hand.

In recent years, however, an increasing number of social scientists have turned to automatic methods to analyze large text collections, often teaming up with researchers in computational linguistics, information retrieval, and related areas. In general, these collaborations are mutually beneficial: social scientists gain the power of new tools that enable inquiries not possible before. Researchers in text processing benefit from a new domain that challenges the robustness and generality of their algorithms.

Our work, which falls within this collaborative tradition, focuses on oral arguments of the U.S. Supreme Court. Transcripts from these proceedings provide a good opportunity to develop computational models of multi-party discourse. In particular, we view the turn-taking behavior of Supreme Court justices as a supervised sequence modeling problem. This work explores the hypothesis that discourse markers and personal references provide useful features in such a modeling task. We have consciously adopted an empirical, data-driven approach that requires few annotations and makes minimal theoretical commitments. Ultimately, this approach to computational discourse modeling can provide a basis for answering interesting questions in legal scholarship, and we see broader applicability in other NLP tasks as well.

2 Related Work

The application of automated content analysis techniques to the social sciences is, of course, not new. Perhaps the best known example is the analysis of function words to determine the authorship of a group of essays in *The Federalist* (Mosteller and Wallace, 1964), and more recently social scientists have developed large categorized lexicons in order to connect language use with multiple kinds of individual differences (Slatcher et al., 2007). Over the past few years, there has been work in domains closely related to our own. Laver et al. (2003) characterized the ideological positions ("liberal" vs. "conservative") of various European political parties from the content of their manifestos. Thomas et al. (2006) investigated the use of machine learning techniques to determine support or opposition for proposed legislation based on transcripts of U.S. Congressional floor debates. Most closely related to our work is that of Evans et al. (2007), who examined a series of affirmative action cases argued in the U.S. Supreme Court. They applied a naive Bayes classifier to determine the ideological position of third-party briefs submitted to the court.

In many ways, political science (broadly defined) provides an ideal domain for automated text analysis. The texts involved are often government records, which simplifies corpus acquisition. In many cases, records of actions taken by important actors (e.g., votes) are also available, and provide a source of annotations for training and evaluation. Finally, politically-oriented text is rich in phenomena that computational linguists are interested in, e.g., opposing sentiments and different characterizations of the same factual event (Greene, 2007).

Beyond the social sciences, there is also a tradition of leveraging computational techniques in the humanities, commonly known as *digital humanities*. One example is the analysis of phrase occurrences in the works of Shakespeare (Lancashire, 1997); cf. (Craig, 2004). However, the goals of humanities scholars are often different from those of social scientists.

Regardless of discipline, there appears to be relatively little prior work on the dynamics of multi-party discourse involving three or more participants. There is considerable literature on discourse analysis for single parties; within the computational literature, for example, Rhetorical Structure Theory (RST) provides a foundation for implemented tools that identify relationships such as ELABORATION, RESULT, PURPOSE, and the like between clauses for the same speaker or author (Marcu, 1997; Corston-Oliver, 1998). For two parties, the literature on dialogue and conversation analysis is quite extensive. Just to cite a few papers: speaker recognition in telephone conversations (Przybocki and Martin, 1998), tagging dialogue acts in conversational speech (Stolcke et al., 2000), and managing discourse in spoken dialogue systems (Williams and Young, 2007). Conversations with three or more participants are less well studied, although some corpora have begun to emerge (MacWhinney et al., 2004) and initial progress has been made on applications such as automatic speaker recognition and transcription in multi-party settings such as meeting rooms (Stanford et al., 2003). To our knowledge, the dynamics of larger multi-party conversations is relatively unexplored.

3 Background

According to conventional wisdom in judicial politics, oral arguments play little if any role in how the Supreme Court makes decisions (Rohde and Spaeth, 1976). Since substantive legal arguments have already been presented in litigant and amici briefs, which are submitted to the court well in advance of the oral sessions, it is thought that they have little impact on the final decision (and some go as far as claiming that justices have already made up their minds). This view, however, has recently been challenged by a number of scholars using manual corpus analysis methods (Johnson, 2001; Shullman, 2004; Johnson et al., 2006). They present evidence that oral arguments are used by justices to gather additional information, and that the quality of the arguments presented does have an impact on outcome. Scholars have hypothesized that oral arguments are the scene of both tactical and strategic maneuvering as justices jostle for support—the process centers around conversations *among* the justices themselves as much as it does on arguments made by lawyers representing both sides. As one might expect, these hypotheses are exceedingly difficult to test using conventional tools of judicial scholarship.

This work takes a step in developing new computational tools that will help judicial scholars better understand the complex interactions that take place during an oral argument session. Our approach centers around building computational models of multi-party discourse, treating analysis as a supervised sequence labeling problem. We have begun by focusing on the turn-taking behavior of Supreme Court justices, so our labels denote "who's currently holding the floor". According to political scientists we are collaborating with, identifying patterns of discussion will likely add significant value for scholars studying the legal and intellectual dynamics of the court.[1] The composition of the court in recent years has yielded a long series of narrowly split (5–4) decisions, indicating the importance of "swing votes". Computational models of discourse would allow our colleagues to tackle the following hypotheses for the first time:

- In divisive cases (as compared to those with broader consensus), justices will devote more effort to convincing colleagues—which would manifest in patterns of interactions, length of turns, complexity of utterances, etc. Probing "questions" might in fact be intended to provoke responses from other justices.
- A justice who ultimately writes a solo opinion (either concurrence or dissent), will engage in more complex (or more frequent) questioning than usual, since his or her position ultimately differs substantially from that of colleagues.
- For a justice who ultimately votes against an established ideological inclination, one would expect cognitive dissonance—which might manifest in questions posed to one side or another. For example, a typically "conservative" justice might probe the lawyer for the "liberal" side more intensely than usual.

[1] Personal communication, Wayne McIntosh and Michael Evans, our political science collaborators.

JUSTICE ALITO: Suppose you have a librarian in a courthouse and the librarian is charging lawyers 25 cents a page for photocopies, but there's some library rule that says the fee is supposed to be 10 cents a page. Now is that, is that a RICO?
MR. TRIBE: If the librarian thinks that, the legislature is not giving us enough money, so I'm going to deliberately use my authority to get an extra five cents from everybody, I suppose if you could prove willfulness, which is an important element of Hobbs, and if there were several librarians and there was a pattern and you could establish the other prerequisites of RICO, it could be a RICO violation.
JUSTICE BREYER: Well, the two cases you cite, the first one is the person who was charged with extortion is a judge –
MR. TRIBE: That's right.
JUSTICE BREYER: And he was extorted on the ground that he told the plaintiff to pay the defendant, so the money wasn't given to the government. So I don't see that that's a difference. And then the second case –
MR. TRIBE: The second case –
JUSTICE BREYER: – it may have been, but you say "See Also," which is a sign to me there's something wrong with that case. (Laughter.)

Figure 1: Transcript sample taken from *Wilkie v. Robbins* (06-219).

To the extent that *any* patterns of interactions or regularities can be captured by our models, computational methods will contribute to the foundations of a new approach to issues in judicial scholarship.

4 From Transcripts to Sequences

Transcripts of Supreme Court oral arguments provide the starting point for our work. In this section, we describe how raw transcripts become a set of labeled sequences for use in model development and testing, and briefly describe some relevant properties of the sequence set.

Argument transcripts are provided online the same day a case is argued, by the current courtroom reporter, Alderson Reporting Company. Since transcripts are distributed as PDF's, we first used an off-the-shelf utility to convert the documents into XML format. The transcripts are segmented by speaker, with the speaker for each turn labeled individually. We treat this segmentation as roughly equivalent to the actual turn sequence during the case (i.e., each speaker segment is treated as one *turn*). In addition to the speaker turns, the body of the transcripts contains time stamps at the beginning and end of each case, as well as section headers before the oral and rebuttal arguments of the case parties. This paper focuses on 188 cases from the 2004–2006 terms; prior to 2004, justices were not individually identified in the transcripts, and at the time of data collection, the 2007 term had not concluded yet.

An extract from a transcript is shown in Figure 1, taken from *Wilkie v. Robbins* (06-219).[2] This example illustrates that in many cases justices aren't asking direct questions of the lawyer (Mr. Tribe in this case) as much as they are commenting on the merits of the case. Note that the switch in justices during this sequence might indicate interest in a particular thread of argument.

In order to investigate judicial turn-taking via sequence modeling, we define unlabeled turns as the observable units in the sequence, each to receive a speaker label, much as words receive grammatical category labels in part-of-speech tagging. Formally, we extract from each unit x a set $\bar{x}$ of features (Section 5), and our models predict the labels y_i for a sequence, yielding $\{(\bar{x}_1, y_1), \ldots, (\bar{x}_n, y_n)\}$. The labels y_i comprise a set of 15 symbols: 11 for the justices (1 for each), 1 to represent the lawyers (either

[2] RICO is the Racketeer Influenced and Corrupt Organizations Act of 1970, a law intended to eradicate organized crime. Clearly Mr. Tribe is much better at law than at arithmetic.

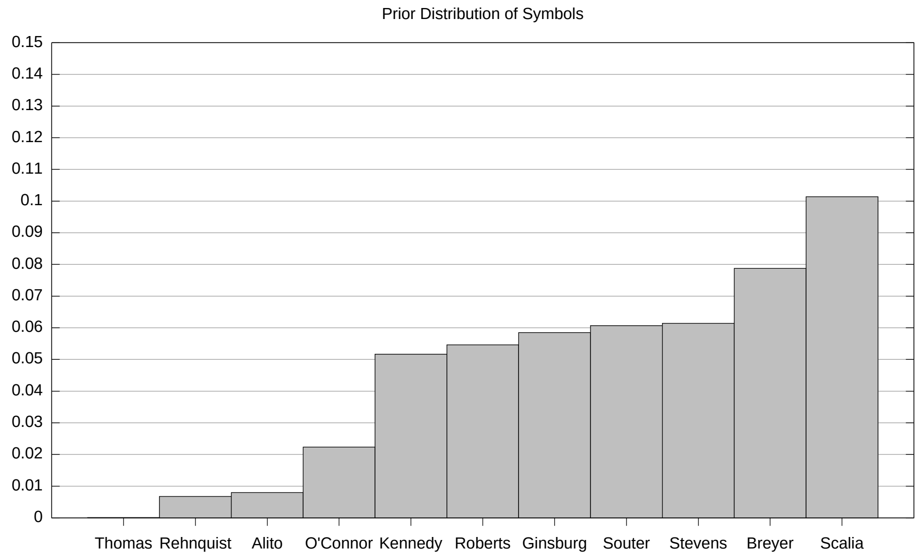

Figure 2: Prior probability for each justice holding a turn in our data set.

on behalf of the petitioner or respondent), plus 1 special symbol for time stamps and 2 additional special symbols to encode the section headings.

The prior probability of symbols representing each justice (across all cases) is shown in Figure 2, which corresponds to how frequently each justice engaged in the oral arguments. Lawyers from both sides combine to represent 47.4% percent of all turns; the other symbols comprise 2.2% of the data. Although there are only nine justices in the U.S. Supreme Court at any given time, we observe 11 distinct symbols due to changes in court membership. Alito replaced O'Connor and Roberts replaced Rehnquist, so neither of those justices spanned the entire data set (this has significant impact on our cross-validation experiments later). Finally, we observe that Thomas rarely spoke (only 5 turns observed in the entire data set).

In the sections that follow, we describe the features $\bar{x}$ used to represent each turn, the models we train on labeled sequences $\{(\bar{x}_1, y_1), \ldots, (\bar{x}_n, y_n)\}$, and our experiments in predicting the label sequence $\{y_1, \ldots, y_m\}$ given unlabeled turns $\{\bar{x}_1, \ldots, \bar{x}_m\}$.

5 Feature Set

Unigram tokens in the turn, filtered for stopwords, represent an obvious baseline representation of turn content. Going beyond this baseline feature set, we are interested primarily in two properties of turns: first, discourse markers that provide cues to turn-taking behavior, and, second, reference to people both within and external to the conversation. We describe these in turn.

Utterances in a coherent discourse—whether prose written by a single author, two-way conversations, or multi-party dialogues—are connected via sets of relations, which must be understood in order to successfully interpret a conversation. Although there are a variety of theoretical characterizations of discourse relations, few of these characterizations are amenable to experimentation using automatic methods. In this work, therefore, we minimize theoretical commitments and reliance on analysis tools, relying only on overt cues to the underlying relations that may hold between utterances.[3]

[3] To the extent that we are successful, accurate identification of the underlying relations, e.g., (Marcu, 1997), should only add value.

Overt cues for discourse relations are generally known in the literature as discourse markers, although other names have been used, e.g., discourse particles, discourse connectives, pragmatic markers, etc. They are generally syntactically optional and semantically vague (or meaningless). However, discourse markers are pragmatically meaningful and serve to link clauses in the discourse structure (Schiffrin, 1987). The transcript fragment in Figure 1 contains a couple of examples of discourse markers. For example, Breyer begins his turn with *well*, a commonly-cited marker, in this case signaling disagreement. His next utterance begins with *and*, signaling an attempt to "hold the floor" after an interruption. For use in this study, we defined a set of discourse markers using a collection of approximately 700 markers compiled from prior literature (Marcu, 1997; Oates, 2001) together with manual examination of arguments. Approximately 50% of turns in our corpus begin with a discourse marker from this collection.

A second set of features is derived from personal references made in the utterances, consisting of four types: names of the justices, honorifics ("Your Honor"), second person pronouns, and a feature for any reference to a non-justice individual by name. Including these features represents a first-pass attempt at developing a model of speaker address, i.e., to whom one is speaking. These features of reference are closely related to speech-based sources of information for addressee detection discussed in Jovanovic and op den Akker (2004); note, however, that the visual cues they suggest, including gaze and gesture, are absent from our transcripts.

Because our approach does not specifically distinguish instances of direct address, the relation between personal references made in the utterances and turn-taking behavior is not obvious. A reference may indeed signal direct address, when a question or comment is intended for a specific recipient in the multi-party conversation. In this case, the referenced individual is more likely to respond (although not always). However, references can also be made with respect to a third party's comment, position, etc., e.g., "concerning Justice Breyer's earlier comments..." Both kinds of personal reference are potentially important clues to the conversational dynamics; we do not distinguish them here. Because referencing conventions in this context are highly regularized, we were able to use regular expression pattern matching to extract features in this set. We find that approximately 40% of turns contain at least one instance of personal reference.

In summary, our experiments involved the following sets of Boolean features:

- **Unigrams.** The unigram feature set ("unigrams") consists of one feature for each unique unigram observed in the collection; its value corresponds to presence or absence of that unigram in the utterance(s) comprising that turn. Including such features essentially gives rise to simple, speaker-specific language models.
- **Discourse markers.** The set of discourse marker features ("DM") was extracted using a constrained string search for markers in the list described above. We only identified markers in turn-initial substrings, allowing for matching of multiple adjacent items. Consider an example from *Kansas v. Marsh (Reargued)* (04-1170): "JUSTICE BREYER: *Okay, well*, what do you say to –", from which we extract two discourse markers (italicized). We expect turn-initial markers to be overt indicators of discourse relations between turns.
- **Personal references.** Because personal reference features ("Ref") are less informative as potential address information for the turn they occur with (justices do not often reference themselves by name or in the second person), each turn includes features corresponding to the absence or presence of references in the previous, current, and subsequent turn.

In addition, for practical reasons, an additional contentless feature, TURN, was included for every turn. This ensured that all turns were represented in the feature sequence during testing.

6 Experiments

The treatment of discourse modeling as a sequence labeling problem makes Conditional Random Fields (Lafferty et al., 2001) an attractive tool for analysis. CRF's are a type of undirected graphical model, preferable to Hidden Markov Models in many sequence labeling tasks since they relax stringent conditional independence assumptions made by generative models. Indeed, CRF's have been empirically shown to work well for a variety of NLP tasks, including part-of-speech tagging (Lafferty et al., 2001), shallow parsing (Sha and Pereira, 2003), and named-entity recognition in the biomedical domain (Settles, 2004).

Using the MALLET implementation of Conditional Random Fields we constructed a number of models from the symbol/feature sequences.[4] Each Supreme Court case in our collection defines a single sequence. To determine the benefit provided by discourse markers and personal references we tested four combinations of feature sets: unigrams, unigrams plus discourse marker features, unigrams plus personal reference features, and unigrams with both discourse marker and personal reference features. Both first- and second-order CRF's were tested. Evaluation for the first-order CRF's was conducted using 10-fold cross-validation. However, due to the significantly longer training time, second-order CRFs were evaluated using 2-fold cross-validation. While folds were split randomly, we used the same split for all trials to ensure results would be comparable for all justices, and most importantly for those justices whose terms do not cover the entire corpus.

As an evaluation measure, we use the quality of prediction for each justice, as quantified by the F-score (harmonic mean of precision and recall, per justice, weighted equally). Predicting the next speaker, *per se*, is not of particular interest to judicial scholars.[5] However, improving the quality of next-speaker predictions does provide evidence that our models are better capturing the dynamics of the multi-party discourse and therefore potentially useful in advancing judicial scholarship.

Figure 3 shows the results for 10-fold cross-validation using first-order CRFs. Of our alphabet of 15 symbols, we report 10. Not reported are the 3 symbols for time and section headings, the 1 symbol for lawyers, and Justice Clarence Thomas for whom we lacked sufficient data to predict. Bars in the graph have been annotated with the relative improvement of using all features compared to unigram features only. As predicted, in all instances the inclusion of additional features provides a boost in overall F-score over the unigrams baseline. While the improvement provided by discourse markers or personal reference varies by justice, the use of both provides improvement over all other conditions for all justices.

Figure 4 shows results for 2-fold cross-validation using second-order CRFs. The graph has been annotated with relative improvements in the same way. We observe a similar pattern, as well as an overall boost in performance (over first-order models) for all justices but Alito and Rehnquist. The differences for these two justices are expected, given the prior probabilities of them speaking and the reduction in the number of cross-validation folds. The fact that we get an overall increase in performance, in addition to a similar pattern, demonstrates that increasing history in a model of justice turn-taking is beneficial. The second-order models capture both the tendency of justices to continue holding the floor and more complex interactions between different justices—both effects are central to hypotheses that legal scholars are interested in.

Summarizing results from both sets of experiments, we compare overall prediction accuracy in Figure 5. Prediction accuracy is simply computed as the fraction of all predictions that were correct (for each condition). With these results, we are able to determine the statistical significance of performance differences. The error bars denote the 95% confidence intervals, as computed by the Clopper-Pearson method (1934) for inferring exact binomial confidence intervals. The confidence intervals indicate that

[4] http://mallet.cs.umass.edu

[5] Although we suggest it might have direct value in other applications, e.g. speaker identification in meeting-room settings.

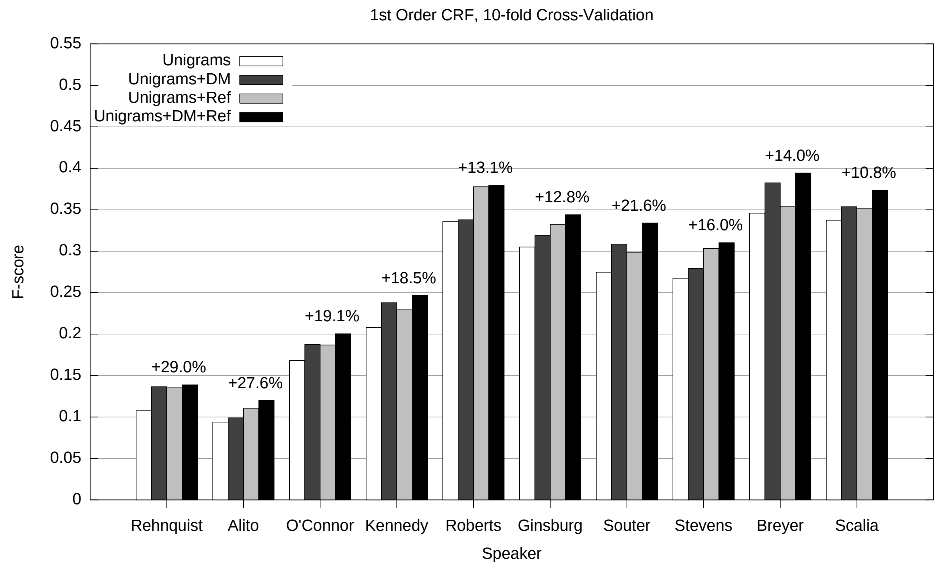

Figure 3: F-scores for first-order CRFs with 10-fold cross-validation

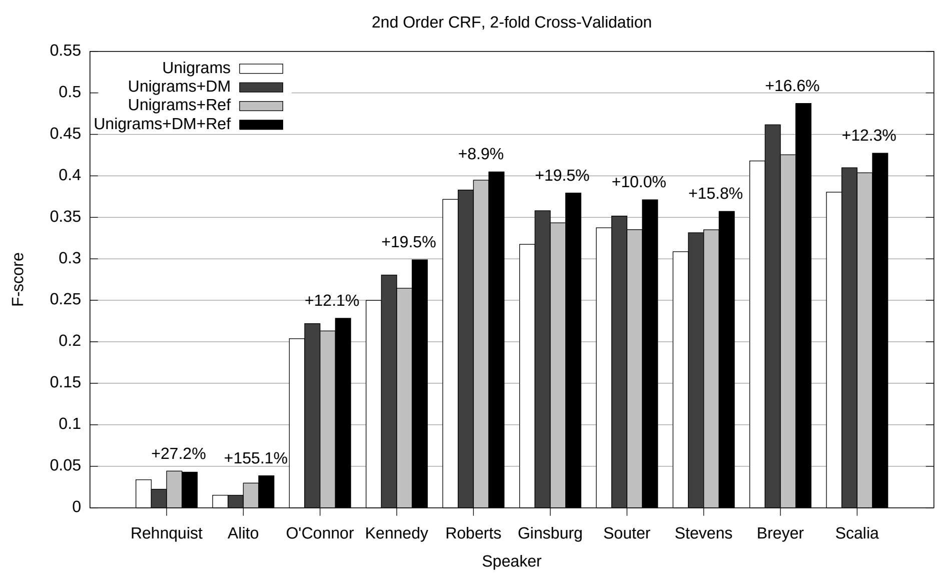

Figure 4: F-scores for second-order CRFs with 2-fold cross-validation

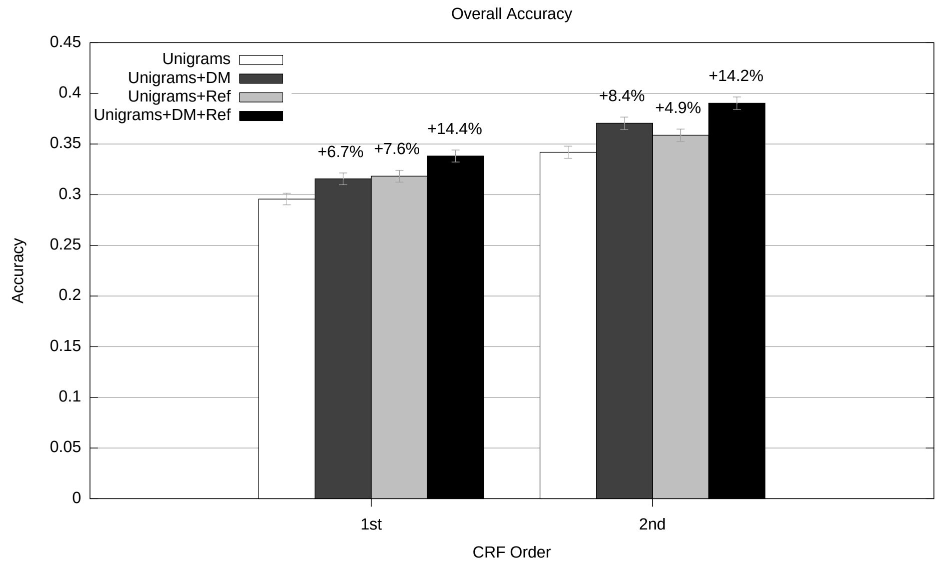

Figure 5: Overall prediction accuracy

for both first- and second-order models, discourse marker features and personal reference features yield a statistically significant improvement in performance (but differences between the two feature sets are not significant). Combining both feature sets results in significant improvements over using either set individually.

7 Conclusions

What is the significance of these results? It must be acknowledged from the start that absolute performance on the evaluation task is quite low. However, we would argue that "success" in this instance is not defined by prediction accuracy itself, but rather by progress toward computational tools that will enable legal scholars to achieve new insights.

To that end, our results support two conclusions:

- There are indeed predictable patterns of interactions between justices in the oral arguments, as evidenced by the performance of the second-order models.
- Discourse markers and personal references provide important clues for underlying discourse relations, at least those relevant in governing turn-taking behavior.

It is worth noting that there is no obvious way to quantify the upper bound in our sequence prediction task. In fact, it might surprise some legal scholars that we have obtained any generalizations about speaker predictability at all. Moreover, while these results are only preliminary, they help make the point that computational analysis can help to detect patterns of activity in the conversational dynamics of Supreme Court justices—even in the absence of detailed domain-specific knowledge or theoretically motivated discourse annotations. Further progress on computational modeling tools in this domain can provide the catalyst for hypothesis generation, deeper analysis, and better understanding of the Court's decision-making process.

Future work aims to augment experimental results with interpretive analysis by our political science colleagues, who bring to bear significant amounts of domain knowledge. Such a combination has led to insightful conclusions in previous work (Evans et al., 2007). We present a few possibilities:

- Frequently-observed sequences of exchanges between justices would be of significant interest, as they may highlight inter-personal dynamics (e.g., ideological conflicts).
- Analysis of individual feature contributions could allow scholars to characterize the idiolect of individual justices and how they respond to their colleagues under different circumstances.
- Correlation of features within our model to votes could determine if aspects of justices' behavior predict their final decisions.

Additionally, we envision that our computational models will serve to guide the acquisition of annotations along relevant dimensions, e.g., issue area, specific discourse relations, etc. This data-driven, bottom-up approach to discourse modeling ensures that annotation efforts remain grounded in tasks and hypotheses that are relevant for the domain experts.

Going beyond this domain, we would argue that our general approach has potential in other areas of language technology. As the accuracy of automatic speech recognition (ASR) systems improves, a number of speech-based NLP applications have become increasingly viable. One such application would be improving the capabilities of intelligent conference rooms and other "intelligent room" environments (Coen, 1998; Stanford et al., 2003); for example, improving conversation tracking or weighting speaker-specific models based on the likelihood of who will speak next. Modeling of this kind also contributes to offline ASR systems as the field continues to move toward spontaneous conversational speech settings, such as talk shows. More generally, the approach we are taking has the potential to improve any multi-agent system in which each software agent must maintain the current state of conversation.

Finally, it is worth emphasizing the cross-disciplinary value of our approach. Although our collaborators currently consist of political scientists, computational models of multi-party discourse may provide new and useful scientific tools for sociologists, anthropologists, and other researchers who wish to better understand human behavior in group dynamics.

Acknowledgments

This research was conducted with support from the National Science Foundation, under award BSC-062467. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the sponsoring institution. We would like to thank our collaborators, Wayne McIntosh and Michael Evans, for valuable discussions and feedback. The second author would like to thank Esther and Kiri for their kind support.

References

C. J. Clopper and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413.

Michael Coen. 1998. Design principles for intelligent environments. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence (AAAI/IAAI 1998)*, 547–554, Madison, Wisconsin.

Simon H. Corston-Oliver. 1998. Identifying the linguistic correlates of rhetorical relations. In *Proceedings of the Workshop on Discourse Relations and Discourse Markers at COLING/ACL 1998*, 8–14, Montreal, Quebec, Canada.

Hugh Craig. 2004. Stylistic analysis and authorship studies. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, 273–288. Blackwell Publishing, Malden, Massachusetts.

Michael Evans, Wayne McIntosh, Jimmy Lin, and Cynthia Cates. 2007. Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4):1007–1039.

Stephan Greene. 2007. *Spin: Lexical Semantics, Transitivity, and the Identification of Implicit Sentiment*. Ph.D. thesis, University of Maryland, College Park.

Timothy R. Johnson, Paul J. Wahlbeck, and James F. Spriggs II. 2006. The influence of oral arguments on the U.S. Supreme Court. *American Political Science Review*, 100(1):99–113.

Timothy R. Johnson. 2001. Information, oral arguments, and Supreme Court decision making. *American Politics Research*, 29(4):331–351.

Natasa Jovanovic and Rieks op den Akker. 2004. Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT/NAACL 2004*, 89–92, Cambridge, Massachusetts.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, 282–289, Williamstown, Massachusetts.

Ian Lancashire. 1997. Empirically determining Shakespeare's idiolect. *Shakespeare Studies*, 25:171–185.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Brian MacWhinney, Steven Bird, Christopher Cieri, and Craig Martell. 2004. TalkBank: Building an open unified multimodal database of communicative interaction. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 525–528.

Daniel Marcu. 1997. The rhetorical parsing of unrestricted natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, 96–103, Madrid, Spain.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Massachusetts.

Sarah Oates. 2001. A listing of discourse markers. Technical Report ITRI-01-26, Information Technology Research Institute, University of Brighton.

Mark A. Przybocki and Alvin F. Martin. 1998. The NIST speaker recognition evaluations. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998)*, 331–335, Grenada, Spain.

David Rohde and Harold Spaeth. 1976. *Supreme Court Decision Making*. Freeman, San Francisco.

Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press, Cambridge, U.K.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP 2004)*, 107–110, Geneva, Switzerland.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2003)*, 134–141, Edmonton, Alberta, Canada.

Sarah Levien Shullman. 2004. The illusion of devil's advocacy: How the justices of the Supreme Court foreshadow their decisions during oral argument. *The Journal of Appellate Practice and Process*, 6:271–293.

Richard B. Slatcher, Cindy K. Chung, James W. Pennebaker, and Lori D. Stone. 2007. Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality*, 41:63–75.

Vincent M. Stanford, John S. Garofolo, Olivier Galibert, Martial Michel, and Christophe D. Laprun. 2003. The NIST Smart Space and Meeting Room projects: signals, acquisition annotation, and metrics. In *Proceedings of the 2003 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'03)*, volume 4, 736–739.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 327–335, Sydney, Australia.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422.