Recognizing Authority in Dialogue with an Integer Linear Programming Constrained Model

Elijah Mayfield
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
elijah@cmu.edu

Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
cprose@cs.cmu.edu

Abstract

We present a novel computational formulation of speaker *authority* in discourse. This notion, which focuses on how speakers position themselves relative to each other in discourse, is first developed into a reliable coding scheme (0.71 agreement between human annotators). We also provide a computational model for automatically annotating text using this coding scheme, using supervised learning enhanced by constraints implemented with Integer Linear Programming. We show that this constrained model's analyses of speaker authority correlates very strongly with expert human judgments (r^2 coefficient of 0.947).

1 Introduction

In this work, we seek to formalize the ways speakers position themselves in discourse. We do this in a way that maintains a notion of discourse structure, and which can be aggregated to evaluate a speaker's overall stance in a dialogue. We define the body of work in positioning to include any attempt to formalize the processes by which speakers attempt to influence or give evidence of their relations to each other. Constructs such as Initiative and Control (Whittaker and Stenton, 1988), which attempt to operationalize the authority over a discourse's structure, fall under the umbrella of positioning. As we construe positioning, it also includes work on detecting certainty and confusion in speech (Liscombe et al., 2005), which models a speaker's understanding of the information in their statements. Work in dialogue act tagging is also relevant, as it seeks to describe the actions and moves with which speakers display these types of positioning (Stolcke et al., 2000).

To complement these bodies of work, we choose to focus on the question of how speakers position themselves as *authoritative* in a discourse. This means that we must describe the way speakers introduce new topics or discussions into the discourse; the way they position themselves relative to that topic; and how these functions interact with each other. While all of the tasks mentioned above focus on specific problems in the larger rhetorical question of speaker positioning, none explicitly address this framing of authority. Each does have valuable ties to the work that we would like to do, and in section 2, we describe prior work in each of those areas, and elaborate on how each relates to our questions.

We measure this as an *authoritativeness ratio*. Of the contentful dialogue moves made by a speaker, in what fraction of those moves is the speaker positioned as the primary authority on that topic? To measure this quantitatively, we introduce the Negotiation framework, a construct from the field of systemic functional linguistics (SFL), which addresses specifically the concepts that we are interested in. We present a reproducible formulation of this sociolinguistics research in section 3, along with our preliminary findings on reliability between human coders, where we observe inter-rater agreement of 0.71. Applying this coding scheme to data, we see strong correlations with important motivational constructs such as Self-Efficacy (Bandura, 1997) as well as learning gains.

Next, we address automatic coding of the Negotiation framework, which we treat as a two-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1018–1026,
Portland, Oregon, June 19-24, 2011.

dimensional classification task. One dimension is a set of codes describing the authoritative status of a contribution[1]. The other dimension is a segmentation task. We impose constraints on both of these models based on the structure observed in the work of SFL. These constraints are formulated as boolean statements describing what a correct label sequence looks like, and are imposed on our model using an Integer Linear Programming formulation (Roth and Yih, 2004). In section 5, this model is evaluated on a subset of the MapTask corpus (Anderson et al., 1991) and shows a high correlation with human judgements of authoritativeness (r^2 = 0.947). After a detailed error analysis, we will conclude the paper in section 6 with a discussion of our future work.

2 Background

The Negotiation framework, as formulated by the SFL community, places a special emphasis on how speakers function in a discourse as sources or recipients of information or action. We break down this concept into a set of codes, one code per contribution. Before we break down the coding scheme more concretely in section 3, it is important to understand why we have chosen to introduce a new framework, rather than reusing existing computational work.

Much work has examined the emergence of discourse structure from the choices speakers make at the linguistic and intentional level (Grosz and Sidner, 1986). For instance, when a speaker asks a question, it is expected to be followed with an answer. In discourse analysis, this notion is described through dialogue games (Carlson, 1983), while conversation analysis frames the structure in terms of adjacency pairs (Schegloff, 2007). These expectations can be viewed under the umbrella of conditional relevance (Levinson, 2000), and the exchanges can be labelled *discourse segments*.

In prior work, the way that people influence discourse structure is described through the two tightly-related concepts of *initiative* and *control*. A speaker who begins a discourse segment is said to have initiative, while control accounts for which speaker is being addressed in a dialogue (Whittaker and Stenton, 1988). As initiative passes back and forth between discourse participants, control over the conversation similarly transfers from one speaker to another (Walker and Whittaker, 1990). This relation is often considered synchronous, though evidence suggests that the reality is not straightforward (Jordan and Di Eugenio, 1997).

Research in initiative and control has been applied in the form of mixed-initiative dialogue systems (Smith, 1992). This is a large and active field, with applications in tutorial dialogues (Core, 2003), human-robot interactions (Peltason and Wrede, 2010), and more general approaches to effective turn-taking (Selfridge and Heeman, 2010). However, that body of work focuses on influencing discourse structure through positioning. The question that we are asking instead focuses on how speakers view their authority as a source of information about the topic of the discourse.

In particular, consider questioning in discourse. In mixed-initiative analysis of discourse, asking a question always gives you control of a discourse. There is an expectation that your question will be followed by an answer. A speaker might already know the answer to a question they asked - for instance, when a teacher is verifying a student's knowledge. However, in most cases asking a question represents a lack of authority, treating the other speakers as a source for that knowledge. While there have been preliminary attempts to separate out these specific types of positioning in initiative, such as Chu-Carroll and Brown (1998), it has not been studied extensively in a computational setting.

Another similar thread of research is to identify a speaker's *certainty*, that is, the confidence of a speaker and how that self-evaluation affects their language (Pon-Barry and Shieber, 2010). Substantial work has gone into automatically identifying levels of speaker certainty, for example in Liscombe et al. (2005) and Litman et al. (2009). The major difference between our work and this body of literature is that work on certainty has rarely focused on how state translates into interaction between speakers (with some exceptions, such as the application of certainty to tutoring dialogues (Forbes-Riley and Litman, 2009)). Instead, the focus is on the person's self-evaluation, independent of the influence on the speaker's positioning within a discourse.

Dialogue act tagging seeks to describe the moves people make to express themselves in a discourse.

[1] We treat each line in our corpus as a single contribution.

This task involves defining the role of each contribution based on its function (Stolcke et al., 2000). We know that there are interesting correlations between these acts and other factors, such as learning gains (Litman and Forbes-Riley, 2006) and the relevance of a contribution for summarization (Wrede and Shriberg, 2003). However, adapting dialogue act tags to the question of how speakers position themselves is not straightforward. In particular, the granularity of these tagsets, which is already a highly debated topic (Popescu-Belis, 2008), is not ideal for the task we have set for ourselves. Many dialogue acts can be used in authoritative or non-authoritative ways, based on context, and can position a speaker as either giver or receiver of information. Thus these more general tagsets are not specific enough to the role of authority in discourse.

Each of these fields of prior work is highly valuable. However, none were designed to specifically describe how people present themselves as a source or recipient of knowledge in a discourse. Thus, we have chosen to draw on a different field of sociolinguistics. Our formalization of that theory is described in the next section.

3 The Negotiation Framework

We now present the Negotiation framework[2], which we use to answer the questions left unanswered in the previous section. Within the field of SFL, this framework has been continually refined over the last three decades (Berry, 1981; Martin, 1992; Martin, 2003). It attempts to describe how speakers use their role as a source of knowledge or action to position themselves relative to others in a discourse.

Applications of the framework include distinguishing between focus on teacher knowledge and student reasoning (Veel, 1999) and distribution of authority in juvenile trials (Martin et al., 2008). The framework can also be applied to problems similar to those studied through the lens of initiative, such as the distinction between authority over discourse structure and authority over content (Martin, 2000).

A challenge of applying this work to language technologies is that it has historically been highly qualitative, with little emphasis placed on reproducibility. We have formulated a pared-down, reproducible version of the framework, presented in Section 3.1. Evidence of the usefulness of that formulation for identifying authority, and of correlations that we can study based on these codes, is presented briefly in Section 3.2.

3.1 Our Formulation of Negotiation

The codes that we can apply to a contribution using the Negotiation framework are comprised of four main codes, K1, K2, A1, and A2, and two additional codes, ch and o. This is a reduction over the many task-specific or highly contextual codes used in the original work. This was done to ensure that a machine learning classification task would not be overwhelmed with many infrequent classes.

The main codes are divided by two questions. First, is the contribution related to exchanging information, or to exchanging services and actions? If the former, then it is a K move (knowledge); if the latter, then an A move (action). Second, is the contribution acting as a primary actor, or secondary? In the case of knowledge, this often correlates to the difference between assertions (K1) and queries (K2). For instance, a statement of fact or opinion is a K1:

g	K1	well i've got a great viewpoint here just below the east lake

By contrast, asking for someone else's knowledge or opinion is a K2:

g	K2	what have you got underneath the east lake
f	K1	rocket launch

In the case of action, the codes usually correspond to narrating action (A1) and giving instructions (A2), as below:

g	A2	go almost to the edge of the lake
f	A1	yeah

A challenge move (ch) is one which directly contradicts the content or assertion of the previous line, or makes that previous contribution irrelevant. For instance, consider the exchange below, where an instruction is rejected because its presuppositions are broken by the challenging statement.

g	A2	then head diagonally down towards the bottom of the dead tree
f	ch	i have don't have a dead tree i have a dutch elm

[2] All examples are drawn from the MapTask corpus and involve an instruction giver (g) and follower (f). Within examples, discourse segment boundaries are shown by horizontal lines.

All moves that do not fit into one of these categories are classified as other (o). This includes back-channel moves, floor-grabbing moves, false starts, and any other non-contentful contributions.

This theory makes use of discourse segmentation. Research in the SFL community has focused on intra-segment structure, and empirical evidence from this research has shown that exchanges between speakers follow a very specific pattern:

o* X2? o* X1+ o*

That is to say, each segment contains a primary move (a K1 or an A1) and an optional preceding secondary move, with other non-contentful moves interspersed throughout. A single statement of fact would be a K1 move comprising an entire segment, while a single question/answer pair would be a K2 move followed by a K1. Longer exchanges of many lines obviously also occur.

We iteratively developed a coding manual which describes, in a reproducible way, how to apply the codes listed above. The six codes we use, along with their frequency in our corpus, are given in Table 1. In the next section, we evaluate the reliability and utility of hand-coded data, before moving on to automation in section 4.

3.2 Preliminary Evaluation

This coding scheme was evaluated for reliability on two corpora using Cohen's kappa (Cohen, 1960). Within the social sciences community, a kappa above 0.7 is considered acceptable. Two conversations were each coded by hand by two trained annotators. The first conversation was between three students in a collaborative learning task; inter-rater reliability kappa for Negotiation labels was 0.78. The second conversation was from the MapTask corpus, and kappa was 0.71. Further data was labelled by hand by one trained annotator.

In our work, we label conversations using the coding scheme above. To determine how well these codes correlate with other interesting factors, we choose to assign a quantitative measure of authoritativeness to each speaker. This measure can then be compared to other features of a speaker. To do this, we use the coded labels to assign an *Authoritativeness Ratio* to each speaker. First, we define a function $A(S, c, L)$ for a speaker, a contribution, and a set of labels $L \subseteq \{K1, K2, A1, A2, o, ch\}$ as:

Code	Meaning	Count	Percent
K1	Primary Knower	984	22.5
K2	Secondary Knower	613	14.0
A1	Primary Actor	471	10.8
A2	Secondary Actor	708	16.2
ch	Challenge	129	2.9
o	Other	1469	33.6
	Total	4374	100.0

Table 1: The six codes in our coding scheme, along with their frequency in our corpus of twenty conversations.

$$A(S, c, L) = \begin{cases} 1 & c \text{ spoken by } S \text{ with label } l \in L \\ 0 & \text{otherwise.} \end{cases}$$

We then define the Authoritativeness ratio $Auth(S)$ for a speaker S in a dialogue consisting of contributions $c_1 ... c_n$ as:

$$Auth(S) = \frac{\sum_{i=1}^{n} A(S, c_i, \{K1, A2\})}{\sum_{i=1}^{n} A(S, c_i, \{K1, K2, A1, A2\})}$$

The intuition behind this ratio is that we are only interested in the four main label types in our analysis - at least for an initial description of authority, we do not consider the non-contentful o moves. Within these four main labels, there are clearly two that appear "dominant" - statements of fact or opinion, and commands or instructions - and two that appear less dominant - questions or requests for information, and narration of an action. We sum these together to reach a single numeric value for each speaker's projection of authority in the dialogue.

The full details of our external validations of this approach are available in Howley et al. (2011). To summarize, we considered two data sets involving student collaborative learning. The first data set consisted of pairs of students interacting over two days, and was annotated for aggressive behavior, to assess warning factors in social interactions. Our analysis

showed that aggressive behavior correlated with authoritativeness ratio ($p < .05$), and that less aggressive students became less authoritative in the second day ($p < .05$, effect size $.15\sigma$). The second data set was analyzed for Self-Efficacy - the confidence of each student in their own ability (Bandura, 1997) - as well as actual learning gains based on pre- and post-test scores. We found that the Authoritativeness ratio was a significant predictor of learning gains ($r^2 = .41$, $p < .04$). Furthermore, in a multiple regression, we determined that the Authoritativeness ratio of both students in a group predict the average Self-Efficacy of the pair ($r^2 = .12$, $p < .01$).

4 Computational Model

We know that our coding scheme is useful for making predictions about speakers. We now judge whether it can be reproduced fully automatically. Our model must select, for each contribution c_i in a dialogue, the most likely classification label l_i from $\{K1, K2, A1, A2, o, ch\}$. We also build in parallel a segmentation model to select s_i from the set $\{new, same\}$. Our baseline approach to both problems is to use a bag-of-words model of the contribution, and use machine learning for classification.

Certain types of interactions, explored in section 4.1, are difficult or impossible to classify without context. We build a contextual feature space, described in section 4.2, to enhance our baseline bag-of-words model. We can also describe patterns that appear in discourse segments, as detailed in section 3.1. In our coding manual, these instructions are given as rules for how segments should be coded by humans. Our hypothesis is that by enforcing these rules in the output of our automatic classifier, performance will increase. In section 4.3 we formalize these constraints using Integer Linear Programming.

4.1 Challenging cases

We want to distinguish between phenomena such as in the following two examples.

f	K2	so I'm like on the bank on the bank of the east lake
g	K1	yeah

In this case, a one-token contribution is indisputably a K1 move, answering a yes/no question. However, in the dialogue below, it is equally inarguable that the same move is an A1:

g	A2	go almost to the edge of the lake
f	A1	yeah

Without this context, these moves would be indistinguishable to a model. With it, they are both easily classified correctly.

We also observed that markers for segmentation of a segment vary between contentful initiations and non-contentful ones. For instance, filler noises can often initiate segments:

g	o	hmm...
g	K2	do you have a farmer's gate?
f	K1	no

Situations such as this are common. This is also a challenge for segmentation, as demonstrated below:

g	K1	oh oh it's on the right-hand side of my great viewpoint
f	o	okay yeah
g	o	right eh
g	A2	go almost to the edge of the lake
f	A1	yeah

A long statement or instruction from one speaker is followed up with a terse response (in the same segment) from the listener. However, after that backchannel move, a short floor-grabbing move is often made to start the next segment. This is a distinction that a bag-of-words model would have difficulty with. This is markedly different from contentful segment initiations:

g	A2	come directly down below the stone circle and we come up
f	ch	I don't have a stone circle
g	o	you don't have a stone circle

All three of these lines look like statements, which often initiate new segments. However, only the first should be marked as starting a new segment. The other two are topically related, in the second line by contradicting the instruction, and in the third by repeating the previous person's statement.

4.2 Contextual Feature Space Additions

To incorporate the insights above into our model, we append features to our bag-of-words model. First, in our classification model we include both lexical bigrams and part-of-speech bigrams to encode further lexical knowledge and some notion of syntactic structure. To account for restatements and topic shifts, we add a feature based on cosine similarity (using term vectors weighted by TF-IDF calculated

over training data). We then add a feature for the predicted label of the previous contribution - after each contribution is classified, the next contribution adds a feature for the automatic label. This requires our model to function as an on-line classifier.

We build two segmentation models, one trained on contributions of less than four tokens, and another trained on contributions of four or more tokens, to distinguish between characteristics of contentful and non-contentful contributions. To the short-contribution model, we add two additional features. The first represents the ratio between the length of the current contribution and the length of the previous contribution. The second represents whether a change in speaker has occurred between the current and previous contribution.

4.3 Constraints using Integer Linear Programming

We formulate our constraints using Integer Linear Programming (ILP). This formulation has an advantage over other sequence labelling formulations, such as Viterbi decoding, in its ability to enforce structure through constraints. We then enhance this classifier by adding constraints, which allow expert knowledge of discourse structure to be enforced in classification. We can use these constraints to eliminate label options which would violate the rules for a segment outlined in our coding manual.

Each classification decision is made at the contribution level, jointly optimizing the Negotiation label and segmentation label for a single contribution, then treating those labels as given for the next contribution classification.

To define our objective function for optimization, for each possible label, we train a one vs. all SVM, and use the resulting regression for each label as a score, giving us six values $\vec{l_i}$ for our Negotiation label and two values $\vec{s_i}$ for our segmentation label. Then, subject to the constraints below, we optimize:

$$\arg\max_{l \in \vec{l_i}, s \in \vec{s_i}} l + s$$

Thus, at each contribution, if the highest-scoring Negotiation label breaks a constraint, the model can optimize whether to drop to the next-most-likely label, or start a new segment.

Recall from section 3.1 that our discourse segments follow strict rules related to ordering and repetition of contributions. Below, we list the constraints that we used in our model to enforce that pattern, along with a brief explanation of the intuition behind each.

$$\begin{aligned} &\forall c_i \in s, (l_i = K2) \Rightarrow \\ &\forall j < i, c_j \in t \Rightarrow (l_j \neq K1) \end{aligned} \tag{1}$$

$$\begin{aligned} &\forall c_i \in s, (l_i = A2) \Rightarrow \\ &\forall j < i, c_j \in t \Rightarrow (l_j \neq A1) \end{aligned} \tag{2}$$

The first constraints enforce the rule that a primary move cannot occur before a secondary move in the same segment. For instance, a question must initiate a new segment if it follows a statement.

$$\begin{aligned} &\forall c_i \in s, (l_i \in \{A1, A2\}) \Rightarrow \\ &\forall j < i, c_j \in s \Rightarrow (l_j \notin \{K1, K2\}) \end{aligned} \tag{3}$$

$$\begin{aligned} &\forall c_i \in s, (l_i \in \{K1, K2\}) \Rightarrow \\ &\forall j < i, c_j \in s \Rightarrow (l_j \notin \{A1, A2\}) \end{aligned} \tag{4}$$

These constraints specify that A moves and K moves cannot cooccur in a segment. An instruction for action and a question requesting information must be considered separate segments.

$$\begin{aligned} &\forall c_i \in s, (l_i = A1) \Rightarrow ((l_{i-1} = A1) \vee \\ &\forall j < i, c_j \in s \Rightarrow (l_j \neq A1)) \end{aligned} \tag{5}$$

$$\begin{aligned} &\forall c_i \in s, (l_i = K1) \Rightarrow ((l_{i-1} = K1) \vee \\ &\forall j < i, c_j \in s \Rightarrow (l_j \neq K1)) \end{aligned} \tag{6}$$

This pair states that two primary moves cannot occur in the same segment unless they are contiguous, in rapid succession.

$$\begin{aligned} &\forall c_i \in s, (l_i = A1) \Rightarrow \\ &\forall j < i, c_j \in s, (l_j = A2) \Rightarrow (S_i \neq S_j) \end{aligned} \tag{7}$$

$$\begin{aligned} &\forall c_i \in s, (l_i = K1) \Rightarrow \\ &\forall j < i, c_j \in s, (l_j = K2) \Rightarrow (S_i \neq S_j) \end{aligned} \tag{8}$$

The last set of constraints enforce the intuitive notion that a speaker cannot follow their own secondary move with a primary move in that segment (such as answering their own question).

Computationally, an advantage of these constraints is that they do not extend past the current segment in history. This means that they usually are only enforced over the past few moves, and do not enforce any global constraint over the structure of the whole dialogue. This allows the constraints to be flexible to various conversational styles, and tractable for fast computation independent of the length of the dialogue.

5 Evaluation

We test our models on a twenty conversation subset of the MapTask corpus detailed in Table 1. We compare the use of four models in our results.

- **Baseline:** This model uses a bag-of-words feature space as input to an SVM classifier. No segmentation model is used and no ILP constraints are enforced.
- **Baseline+ILP:** This model uses the baseline feature space as input to both classification and segmentation models. ILP constraints are enforced between these models.
- **Contextual:** This model uses our enhanced feature space from section 4.2, with no segmentation model and no ILP constraints enforced.
- **Contextual+ILP:** This model uses the enhanced feature spaces for both Negotiation labels and segment boundaries from section 4.2 to enforce ILP constraints.

For segmentation, we evaluate our models using exact-match accuracy. We use multiple evaluation metrics to judge classification. The first and most basic is accuracy - the percentage of accurately chosen Negotiation labels. Secondly, we use Cohen's Kappa (Cohen, 1960) to judge improvement in accuracy over chance. The final evaluation is the r^2 coefficient computed between predicted and actual Authoritativeness ratios per speaker. This represents how much variance in authoritativeness is accounted for in the predicted ratios. This final metric is the most important for measuring reproducibility of human analyses of speaker authority in conversation.

We use SIDE for feature extraction (Mayfield and Rosé, 2010), SVM-Light for machine learning (Joachims, 1999), and Learning-Based Java for ILP inference (Rizzolo and Roth, 2010). Performance is evaluated by 20-fold cross-validation, where each fold is trained on 19 conversations and tested on the remaining one. Statistical significance was calculated using a student's paired t-test. For accuracy and kappa, $n = 20$ (one data point per conversation) and for r^2, $n = 40$ (one data point per speaker).

Model	**Accuracy**	**Kappa**	r^2
Baseline	59.7%	0.465	0.354
Baseline+ILP	61.6%	0.488	0.663
Segmentation	*72.3%*		
Contextual	66.7%	0.565	0.908
Contextual+ILP	68.4%	0.584	0.947
Segmentation	*74.9%*		

Table 2: Performance evaluation for our models. Each line is significantly improved in both accuracy and r^2 error from the previous line ($p < .01$).

5.1 Results

All classification results are given in Table 2 and charts showing correlation between predicted and actual speaker Authoritativeness ratios are shown in Figure 1. We observe that the baseline bag-of-words model performs well above random chance (kappa of 0.465); however, its accuracy is still very low and its ability to predict Authoritativeness ratio of a speaker is not particularly high (r^2 of 0.354 with ratios from manually labelled data). We observe a significant improvement when ILP constraints are applied to this model.

The contextual model described in section 4.2 performs better than our baseline constrained model. However, the gains found in the contextual model are somewhat orthogonal to the gains from using ILP constraints, as applying those constraints to the contextual model results in further performance gains (and a high r^2 coefficient of 0.947).

Our segmentation model was evaluated based on exact matches in boundaries. Switching from baseline to contextual features, we observe an improvement in accuracy of 2.6%.

5.2 Error Analysis

An error analysis of model predictions explains the large effect on correlation despite relatively smaller

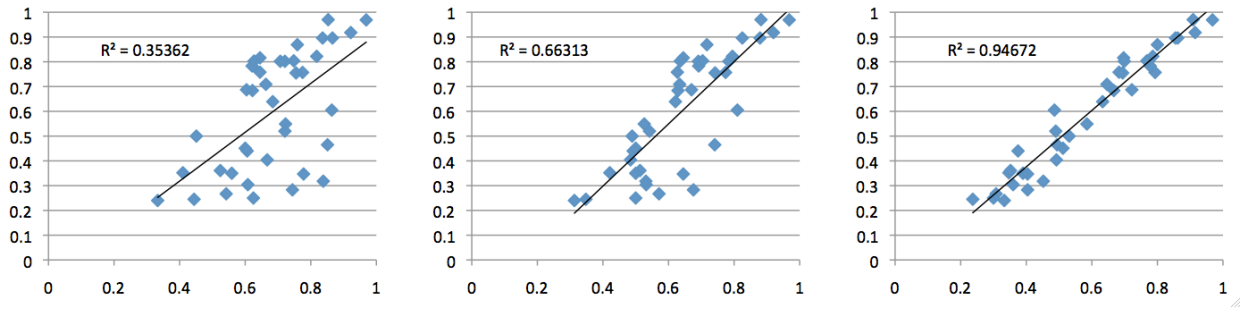

Figure 1: Plots of predicted (x axis) and actual (y axis) Authoritativeness ratios for speakers across 20 conversations, for the Baseline (left), Baseline+Constraints (center), and Contextual+Constraints (right) models.

changes in accuracy. Our Authoritativeness ratio does not take into account moves labelled o or ch. What we find is that the most advanced model still makes many mistakes at determining whether a move should be labelled as o or a core move. This error rate is, however, fairly consistent across the four core move codes. When a move is determined (correctly) to not be an o move, the system is highly accurate in distinguishing between the four core labels.

The one systematic confusion that continues to appear most frequently in our results is the inability to distinguish between a segment containing an $A2$ move followed by an $A1$ move, and a segment containing a $K1$ move followed by an o move. The surface structure of these types of exchanges is very similar. Consider the following two exchanges:

g	A2	if you come down almost to the bottom of the map that I've got
f	A1	uh-huh

f	K1	but the meadow's below my broken gate
g	o	right yes

These two exchanges on a surface level are highly similar. Out of context, making this distinction is very hard even for human coders, so it is not surprising then that this pattern is the most difficult one to recognize in this corpus. It contributes most of the remaining confusion between the four core codes.

6 Conclusions

In this work we have presented one formulation of authority in dialogue. This formulation allows us to describe positioning in discourse in a way that is complementary to prior work in mixed-initiative dialogue systems and analysis of speaker certainty. Our model includes a simple understanding of discourse structure while also encoding information about the types of moves used, and the certainty of a speaker as a source of information. This formulation is reproducible by human coders, with an inter-rater reliability of 0.71.

We have then presented a computational model for automatically applying these codes per contribution. In our best model, we see a good 68.4% accuracy on a six-way individual contribution labelling task. More importantly, this model replicates human analyses of authoritativeness very well, with an r^2 coefficient of 0.947.

There is room for improvement in our model in future work. Further use of contextual features will more thoroughly represent the information we want our model to take into account. Our segmentation accuracy is also fairly low, and further examination of segmentation accuracy using a more sophisticated evaluation metric, such as WindowDiff (Pevzner and Hearst, 2002), would be helpful.

In general, however, we now have an automated model that is reliable in reproducing human judgments of authoritativeness. We are now interested in how we can apply this to the larger questions of positioning we began this paper by asking, especially in describing speaker positioning at various instants throughout a single discourse. This will be the main thrust of our future work.

Acknowledgements

This research was supported by NSF grants SBE-0836012 and HCC-0803482.

References

Anne Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, et al. 1991. The HCRC Map Task Corpus. In *Language and Speech*.

Albert Bandura. 1997. *Self-efficacy: The Exercise of Control*

Margaret Berry. 1981. Towards Layers of Exchange Structure for Directive Exchanges. In *Network* 2.

Lauri Carlson. 1983. *Dialogue Games: An Approach to Discourse Analysis*.

Jennifer Chu-Carroll and Michael Brown. 1998. An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions. In *User Modeling and User-Adapted Interaction*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. In *Educational and Psychological Measurement*.

Mark Core and Johanna Moore and Claus Zinn. 2003. The Role of Initiative in Tutorial Dialogue. In *Proceedings of EACL*.

Kate Forbes-Riley and Diane Litman. 2009. Adapting to Student Uncertainty Improves Tutoring Dialogues. In *Proceedings of Artificial Intelligence in Education*.

Barbara Grosz and Candace Sidner. 1986. Attention, Intentions, and the Structure of Discourse. In *Computational Linguistics*.

Iris Howley and Elijah Mayfield and Carolyn Penstein Rosé. 2011. Missing Something? Authority in Collaborative Learning. In *Proceedings of Computer-Supported Collaborative Learning*.

Thorsten Joachims. 1999. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning*.

Pamela Jordan and Barbara Di Eugenio. 1997. Control and Initiative in Collaborative Problem Solving Dialogues. In *Proceedings of AAAI Spring Symposium on Computational Models for Mixed Initiative Interactions*.

Stephen Levinson. 2000. *Pragmatics*.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. 2005. Detecting Certainness in Spoken Tutorial Dialogues. In *Proceedings of Interspeech*.

Diane Litman and Kate Forbes-Riley. 2006. Correlations betweeen Dialogue Acts and Learning in Spoken Tutoring Dialogue. In *Natural Language Engineering*.

Diane Litman, Mihai Rotaru, and Greg Nicholas. 2009. Classifying Turn-Level Uncertainty Using Word-Level Prosody. In *Proceedings of Interspeech*.

James Martin. 1992. *English Text: System and Structure.*

James Martin. 2000. Factoring out Exchange: Types of Structure. In *Working with Dialogue*.

James Martin and David Rose. 2003. *Working with Discourse: Meaning Beyond the Clause*.

James Martin, Michele Zappavigna, and Paul Dwyer. 2008. Negotiating Shame: Exchange and Genre Structure in Youth Justice Conferencing. In *Proceedings of European Systemic Functional Linguistics*.

Elijah Mayfield and Carolyn Penstein Rosé. 2010. An Interactive Tool for Supporting Error Analysis for Text Mining. In *Proceedings of Demo Session at NAACL*.

Julia Peltason and Britta Wrede. 2010. Modeling Human-Robot Interaction Based on Generic Interaction Patterns. In *AAAI Report on Dialog with Robots*.

Lev Pevzner and Marti Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. In *Computational Linguistics*.

Heather Pon-Barry and Stuart Shieber. 2010. Assessing Self-awareness and Transparency when Classifying a Speakers Level of Certainty. In *Speech Prosody*.

Andrei Popescu-Belis. 2008. Dimensionality of Dialogue Act Tagsets: An Empirical Analysis of Large Corpora. In *Language Resources and Evaluation*.

Nick Rizzolo and Dan Roth. 2010. Learning Based Java for Rapid Development of NLP Systems. In *Language Resources and Evaluation*.

Dan Roth and Wen-Tau Yih. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proceedings of CoNLL*.

Emanuel Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*.

Ethan Selfridge and Peter Heeman. 2010. Importance-Driven Turn-Bidding for Spoken Dialogue Systems. In *Proceedings of ACL*.

Ronnie Smith. 1992. A computational model of expectation-driven mixed-initiative dialog processing. Ph.D. Dissertation.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, et al. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In *Computational Linguistics*.

Robert Veel. 1999. Language, Knowledge, and Authority in School Mathematics. In *Pedagogy and the Shaping of Consciousness: Linguistics and Social Processes*

Marilyn Walker and Steve Whittaker. 1990. Mixed Initiative in Dialogue: An Investigation into Discourse Structure. In *Proceedings of ACL*.

Steve Whittaker and Phil Stenton. 1988. Cues and Control in Expert-Client Dialogues. In *Proceedings of ACL*.

Britta Wrede and Elizabeth Shriberg. 2003. The Relationship between Dialogue Acts and Hot Spots in Meetings. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.