LEARNING THE STRUCTURE OF MIXED INITIATIVE DIALOGUES USING A CORPUS OF ANNOTATED CONVERSATIONS[1]

Giovanni Flammia and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{flammia,zue}@sls.lcs.mit.edu

ABSTRACT

This paper reports an ongoing effort to derive linear discourse structures from a corpus of telephone conversations. First, we would like to determine how reliably human annotators can tag discourse segments in dialogues. Second, we begin to investigate how to build machine models for performing this annotation task. To carry out our research, we use a corpus of transcribed and annotated human-human dialogues in a specific information retrieval domain (Movie theater schedules). We conducted an experiment in which 25 different dialogues have each been annotated by at least seven different people. We found that the average precision and recall among annotators in placing segment boundaries is 84.3%, and in assigning segment purpose labels is 80.1%. A simple discourse segment parser based on finite state machines is able to cover 56% of the same dialogues. When the finite state grammar is able to analyse a dialogue, it agrees with human annotators in placing segment boundaries with 59.4% precision and 66.4% recall, and it agrees in segment label accuracy at the 59% level.

1. INTRODUCTION

Since 1989, our group has been involved in the development of conversational systems that can help users access information and solve problems using verbal input. A key research challenge facing us is the development of a dialogue management component that can guide the user towards a successful conclusion of an interaction by offering helpful suggestions and issuing clarification requests. In the research reported here, we have taken the approach of developing a human-machine dialogue model based on analyses of human-human dialogues when solving the same tasks. Admittedly, human-human dialogues can be quite variable, containing frequent interruptions, speech overlaps, incomplete or unclear sentences, incoherent segments, and topic switches. Some of these variabilities satisfy communication needs and may not contribute directly to goal-directed problem solving. However, we believe, as do others, that studying human-human dialogue and comparing it to human-machine dialogue can provide valuable insights [1].

To carry out our research, we need to first obtain properly annotated language resources. Since dialogue annotation can be extremely time consuming, it is essential that we develop the necessary tools to maximize efficiency and consistency. We have previously reported the development of a dialogue annotation tool called *Nb* which has been used extensively for dialogue annotation in our group and several other institutions. *Nb* is freely available for Unix and Windows [2].

The research reported in this paper is a continuation of our previous work. Specifically, we seek answers to two questions. First, is it possible to obtain consistent annotations from many subjects? While our previous results in this regard were not very encouraging we attributed the equivocal results to the small sample size and the lack of training for the annotators. In the experiments reported here, we used a much larger number of annotators and dialogues, together with a revised set of instructions.

Second, is it possible to build discoure segment models (semi)-automatically from an annotated corpus? To be sure, we do not expect that all parts of human-human conversations can be accurately tagged with specific goal-oriented purposes along the few dimensions that we choose to annotate. Our goal is to determine empirically the extent to which a low-dimensional discourse segment structure can be extracted from annotated transcriptions of spontaneous, natural dialogues.

The rest of the paper is organized as follows. In the next section, we revisit the consistency question and report a series of dialogue annotation experiments, utilizing segment as well as semantic tags. Then, we describe briefly our initial attempt at automatically developing a finite-state dialogue model using the maually annotated reference material. Finally, we evaluate the adequacy of the automatic segmentation with respect to the manual segmentation of the reference material.

2. MANUAL ANNOTATION

2.1. Corpus

To carry out our research, we are making use of a corpus of 122 orthographically transcribed and annotated telephone conversations. The text data are faithful transcriptions of actual telephone conversations between customers and operators of the BellSouth *Movies Now* service, collected by BellSouth Intelliventures in 1994. The *Movies Now* service is a telephone number that people can call to get information about current movie schedules in Atlanta. Discourse annotation is based on the text transcriptions alone, without listening to the corresponding audio recordings. The text transcriptions include punctuation

[1] This research is supported by DARPA, under contract N 66001-94-C-6040, monitored by Naval Command Control and Ocean Surveillance Center.

	Phone Number For Theater
CPhnLocSet	I am looking for the number */Number* to the Gwinnet cinema */Location*
A	OK.
	List Movies Playing At Theater
CMov	Also, do you know what movies */Movie* are playing?
AMovLoc	Yes, I can tell you what's */Movie* playing there */Location*.
C	OK.
AMovSet	Pulp Fiction */Movie* and Time Cop */Movie*
	Phone Number For Theater
CPhn	And the number */Phone* please.
APhnSet	OK, the number is 356-8753 */Phone*.

Figure 1: Sample annotated dialogue. Each sentence is annotated with topics (in italics) and assigned a symbol (first column). Discourse segments are assigned to different purposes (in bold face above the transcription).

marks, sentence segmentation, speaker change markers and other markers for long pauses, overlapped and interrupted speech, non-speech events, and unclear speech. A sample from this corpus is displayed in Figure 1. The average number of dialogue turns (speaker changes) per dialogue was 40.

2.2. Segment Tags

Tagging discourse segment boundaries and purposes is a difficult and subjective cognitive task. The reliability of this task depends on the linguistic variability of the corpus and on the level of detail of the annotation. Swerts and Ostendorf studied highly structured question-and-answer human-machine dialogues from the ATIS airline reservation task [7]. They found that human annotators are reliable when tagging simple discourse segments that refer to the same flight, although they do not report precision/recall results on boundary placements. Hearst reports a study in multi-paragraph discourse segmentation of professionally written narrative text [4]. She found that human annotators agree with a reference segmentation with 81% precision and 71% recall. She also reports a machine performance of 66% precision and 61% recall. Litman and Passoneau studied extensively discourse segmentation of spontaneous spoken narrative by non-professional speakers [6]. They report human agreement with a reference segmentation of 72% precision and 63% recall, and machine performance of 63% precision and 46% recall. Hirschberg, Nakatani, and Grosz [5] analyzed hierarchical discourse segmentation of spontaneous task-oriented speech monologues by non-professional speakers. While they use different measures for computing agreement, they found that different annotators may segment discourse at different level of granularity, and that discourse segmentation is more reliable when performed by listening to the acoustic signal as well as reading the text transcription.

Our initial annotation experiment, cited in [2], reported a best case 60% pairwise agreement among annotators in placing segment boundaries. In that experiment we used a minimal set of instructions, we allowed hierarchical segmentations with any level of nesting, we used conversations from multiple domains as our data, and a small number of annotators with different levels of linguistic knowledge. Since then, we have conducted more focused experiments by developing a set of instructions tailored towards the movie domain, by enrolling as paid volunteers 16 graduate students with some knowledge of computer science and linguistics, and by constraining the annotation task. In the instructions, we assumed that a conversation can be decomposed sequentially by purpose [3], that is, a conversation can be modeled by a sequence of one or more segments, each segment having the role to fulfill one specific purpose. The annotation task was further constrained by allowing only linear (non-hierarchical) segmentations, and by limiting the choices to a small set of segment purposes that were determined from knowledge about the agent's task.[2] By constraining the segmentation to be linear, we limited the degrees of freedom and the associated cognitive load for the annotators, at the expense of the expressive power of the annotation.

We selected 25 dialogues from the corpus. For each dialogue we asked between seven and nine different people to break the text transcription into a sequence of one or more discourse segments, according to a on-line instruction manual. In each annotation session, the on-line annotation tool included hypertext instructions and a set of drill annotations. Before completing the actual annotation task, each annotator was prompted by the tool to discover the "correct" segmentation for four training dialogues that were previously tagged by one of the authors. All annotators found the task challenging. On average, each annotation session lasted 1 and 1/2 hours, which included reading the on-line instruction manual, completing the four drill exercises, and annotating nine different dialogues.

For each dialogue and each pair of annotators, we have evaluated the agreement in placing segment boundaries and labeling segments with segment purposes. Agreement in placing segment boundaries has been computed by averaging precision and recall values. For each pair of annotators, we computed the number of boundaries that were proposed by both of them as well as the boundaries proposed by only one of them. We then computed precision and recall values using each annotator in turn as reference. Over all pairs of coders, the average precision and recall value was 84.3% (standard deviation 5.8%). Agreement in labeling segment purposes has been computed by extracting the sequence of segment purpose symbols, and by running the NIST alignment program on each pair of symbol sequences. Agreement between two annotators has been evaluated as the symbol accuracy, defined as the difference between the number of matched segments and the number of inserted segments. The average pairwise symbol accuracy for segment purposes was 80.1% (6.6% substitutions, 7.3% insertions, 6.6% deletions). Figure 2 is a scatter plot of all pairwise agreements. Each point in the plot is the agreement among two different annotators

[2] The purposes are: (1) List movies playing at theater, (2) Where is this movie playing, (3) Is this movie playing at this theater, (4) Show times at this theater, and (5) Phone number for theater.

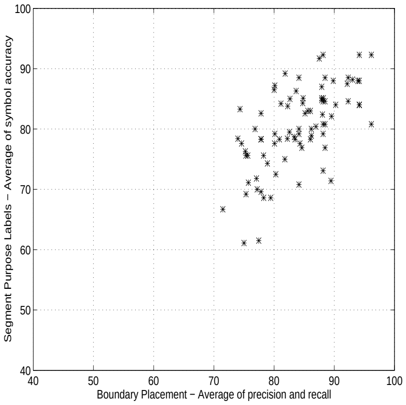

Figure 2: Pairwise agreement among annotators in tagging segment purposes and segment boundaries. Each point in the plot displays the agreement among two annotators.

on the dialogues they both annotated. Disagreement in placing segment boundaries mostly occurred around spontaneous speech communication events such as incomplete sentences, restarts, repeated questions, and speech repairs. Disagreements in segment purposes mostly occured when a surface linear segmentation represented an underlying hierarchical segmentation. Other disagreements in assuming a linear segmentation occured when the conversation was switching back and forth between two purposes before completing either of them. This made the linear annotation task somewhat more difficult, with annotators deleting or inserting either segments.

2.3. Semantic Tags

To aid the selection of input features for an automatic segmentation algorithm, we annotate semantic labels (topics) from individual phrases in the text transcriptions. Tagging semantic labels is accomplished by selecting specific noun phrases and prepositional phrases in each sentence and tagging them with one out of four topics.[3] Tagged phrases can either be specific named entities (e.g., *the movie Stargate*), indefinite entities (*a movie*, *one*), wh-words (*what*'s playing, *which one*) or anaphora (*it*'s playing *there*, the *first one*). We conducted a pilot experiment enrolling four paid graduate students. We followed the same annotation protocol as for the segmentation task, and asked each of them to tag the text transcriptions of the same three dialogues from our corpus. We then extracted a sequence of symbols, one for each sentence annotated with semantic topics, computed by combining the manually annotated tags with the automatically annotated part of speech tags and speaker information (see Figure 1). The pairwise agreement among annotators was computed as the average symbol accuracy by running the NIST alignment program on each pair of symbol sequences. Symbol accuracy was computed by subtracting the number of inserted or deleted symbols from the the number of symbols that were proposed by both annotators. The experiment showed that this annotation can be done fast, at an average of 8 dialogue turns per minute of annotation. The average pairwise accuracy scores among annotators was 86.3% (standard deviation 6%, average correct match score 90%). The disagreements consisted in omissions and insertions when tagging some of the anaphoric expressions, pronouns, and wh-words, while named entities were always tagged correctly by all of the annotators.

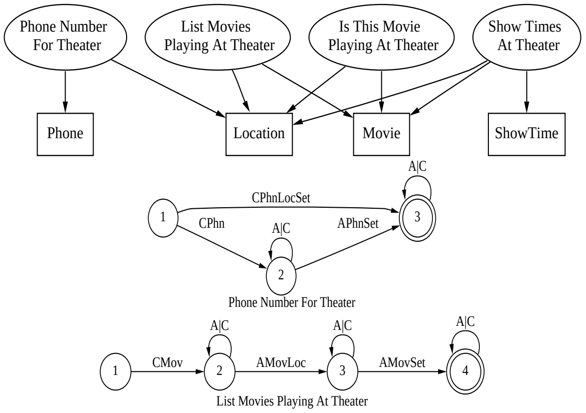

Figure 3: A task model for the movies domain. Segment purposes (elliptical nodes) require setting values for one or more topics (boxes). At the discourse level, each segment is realized by a sequence of dialogue turns, modeled here by finite state machines.

3. AUTOMATIC SEGMENTATION

3.1. Discourse Segment Models

As a preliminary experiment, we have set our goal on determining whether a simple model such as a finite state machine is sufficient to model the variability within and across discourse segments. We chose a finite state model because of its simplicity and because it is used in some human-machine dialogue models. The finite state machines model the typical turn-by-turn communication process between the agent and the customer.

The computational model we use is divided into two related levels. The first level is the task or planning level, and the second level is the discourse level. At the task level, we model simple transactions in which the agent can accomplish one or more purposes, or goals, during a conversation. The top diagram in Figure 3 enumerate four possible purposes in the Movies domain (elliptical nodes). Each purpose requires to retrieve from a relational database the specific values for one or more topics, indicated by boxes. The detailed computational structure of this level is beyond the scope of this paper. At the discourse level, each purpose is realized by a discourse segment, and each discourse segment is realized by a sequence of one or more dialogue turns. The detailed structure of the finite state machines is learned by examples. The combination of semantic and part-of-speech tags and speaker information are used to assign each sentence to a particular symbol from a finite alphabet (first column in Figure 1). Each symbol provides a low-dimensional discrete approximation (vector quantization) of the surface content of each individual dialogue turn. The symbols are

[3] The topics are: (1) location, (2) movie title, (3) phone number and (4) show time or date.

Parse	Boundaries		Segment Label
Completed	Precision	Recall	Accuracy
56% (14/25)	59.4%	66.4%	59%

Table 1: Average agreement between the finite state model and human annotators in placing segment boundaries and labeling segment purposes. Agreement data has been computed for the 14 dialogues that were fully analyzed by the parser.

used as labels for the state transitions of the finite state machines that model each individual discourse segment. For example, the two machines displayed in Figure 3 are derived from the dialogue in Figure 1.

3.2. Training And Evaluation

A training set of 97 dialogue transcriptions annotated with segment boundaries, segment purposes, semantic tags and part-of-speech tags is used to provide training samples for each one of the discourse segments. The finite state machines are currently built by simply enumerating all possible training samples, and then minimizing the computed structure. All the observed sequences of segments in the training data are used to build the top-level finite state grammar of discourse segments. The model did not attach probabilities to state transitions.

We tested the models by parsing each one of the 25 dialogues that were annotated by the human annotators, using the learned finite state grammar in combination with an Earley parser. When the parser was able to complete the segmentation of a dialogue, all possible proposed segmentations were matched against the segmentation proposed by each annotator, and the segmentation that most closely matched the human segmentation was used to compute the agreement between human and machine, using the human annotation as reference. We found that the simple finite state grammar was able to cover 56% of the dialogues, i.e., 14 out of the 25 dialogues produced an average of two segmentations, and the parser failed to complete the analysis for the other 11 dialogues. For the 14 dialogues for which a segmentation was proposed, we computed precision and recall to compare segment boundary placement, and symbol accuracy to compare the assignment of segment purpose labels. Average results are displayed in Table 1. When we examined the individual results for each dialogue, we found out that the parser had different behavior depending on the data. The finite state grammar was able to cover 20% of the dialogues with segment accuracy better than 70% level, while it was unable to cover accurately a majority of previously unseen data.

The three major limitations of this finite state model is that the grammar does not generalize, that it is ambiguous, and that it does not model new versus given information content at the global level of a discourse segment. Since the grammar encodes the training data exactly, it cannot generalize to previously unseen sequences of symbols unless they appear in the training data. Ambiguity is caused by the quantization error introduced by the symbol encoding scheme, and is affected by the sparse data problem as well. For example, if a sequence of symbols *A B* can be segmented as both *(A) (B)* and *(A B)* and the training data contains only examples of one type, the parser will produce either a segment deletion or a segment insertion. Ambiguity can be resolved by adding additional knowledge sources to the model and by adding features to the symbol encoding scheme. For example, to correctly segment a dialogue it is not sufficient to model the typical turn-by-turn sequence of topic labels. Within each topic, such as *Location* or *Movie*, it is also necessary to track whether any given hypothesized discourse segment represents new vs. given information.

4. DISCUSSION

Discourse segmentation of human-human dialogues is a difficult cognitive task. Given the limitations in the expressive power of linear discourse segmentations and the high degree of variability in spontaneous human-human dialogues, we are encouraged by the reliability of the annotation performed by humans. Human annotators use a holistic approach in segmenting a dialogue, evaluating the coherence of each hypothesized discourse segment with respect to the entire dialogue. We found that a simple finite state model that track state transitions at the local level of the dialogue turn is unable to match the performance of human annotators. We plan to refine our automatic modeling technique along four lines. First, we will double the size of the training data. Secondly, we plan to integrate some discourse markers into the symbolic encoding of each sentence. Thirdly, we will abandon a simple finite state model in favor of a context free model that includes probabilities. Finally, the model must include a holistic measure of new versus given information. This feature is currently missing from our model, and can be measured using topic identification and information retrieval techniques applied to entire discourse segments rather than to individual dialogue turns [4, 7].

REFERENCES

[1] Bernsen, N. O., Dybkjaer, L. and Dybkjaer, H. “Cooperativity in human-machine and human-human spoken dialogue.” *Discourse Processes* Vol. 21. No. 2. 1996. pp. 213-236.

[2] Flammia G. and Zue V. “Empirical Evaluation of Human Performance and Agreement in Parsing Discourse Constituents in Spoken Dialogue.” *Proc. Eurospeech-95*. 1995. pp. 1965-1968. *http://www.sls.lcs.mit.edu/flammia/Nb.html*

[3] Grosz B., Sidner C., “Attentions, Intentions and the Structure Of Discourse.” *Computational Linguistics*, Vol. 12. No. 3, 1986. pp. 175-204.

[4] Hearst M. *Context and Structure in Automated Full-Text Information Access* TR. UCB/CSD-94/836. University Of California, Berkeley, CA. 1994.

[5] Hirschberg J., Nakatani C., and Grosz B. “Conveying Discourse Structure Through Intonation Variation.” *ESCA Workshop on Spoken Dialogue*. 1995. pp. 189-192.

[6] Litman D.J. and Passonneau R.J. “Combining Multiple Knowledge Sources For Discourse Segmentation.” *ACL Proceedings*. 1995. pp. 108-115.

[7] Swerts M. and Ostendorf M. “Discourse prosody in human-machine interactions.” *ESCA Workshop on Spoken Dialogue*. 1995. pp. 205-208.