What do We Know about Conversation Participants: Experiments on Conversation Entailment

Chen Zhang **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{zhangch6, jchai}@cse.msu.edu

Abstract

Given the increasing amount of conversation data, techniques to automatically acquire information about conversation participants have become more important. Towards this goal, we investigate the problem of conversation entailment, a task that determines whether a given conversation discourse entails a hypothesis about the participants. This paper describes the challenges related to conversation entailment based on our collected data and presents a probabilistic framework that incorporates conversation context in entailment prediction. Our preliminary experimental results have shown that conversation context, in particular dialogue act, plays an important role in conversation entailment.

1 Introduction

Conversation is a joint activity between its participants (Clark, 1996). Their goals and their understanding of mutual beliefs of each other shape the linguistic discourse of conversation. In turn, this linguistic discourse provides tremendous information about conversation participants. Given the increasing amount of available conversation data (e.g., conversation scripts such as meeting scripts, court records, and online chatting), an important question is *what do we know about conversation participants?* The capability to automatically acquire such information can benefit many applications, for example, development of social networks and discovery of social dynamics.

Related to this question, previous work has developed techniques to extract profiling information about participants from conversation interviews (Jing et al., 2007) and to automatically identify dynamics between conversation participants such as agreement/disagreement from multiparty meeting scripts (Galley et al., 2004). We approach this question from a different angle as a *conversation entailment* problem: given a conversation discourse D and a hypothesis H concerning its participant, the goal is to identify whether D entails H. For instance, in the following example, the first hypothesis can be entailed from the dialogue segment while the second hypothesis cannot.

Example 1:

Dialogue Segment:

A: And where about were you born?
B: Up in Person Country.

Hypothesis:

(1) B was born in Person Country.
(2) B lives in Person Country.

Inspired by textual entailment (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007), conversation entailment provides an intermediate step towards acquiring information about conversation participants. What we should know or would like to know about a participant can be rather open. The type of information needed about participants is also application-dependent and difficult to generalize. In conversation entailment, we will not face this problem since hypotheses can be used to express any type of information about a participant one might be interested in. Although hypotheses are currently given in our investigation, they can potentially be automatically generated based on information needs and/or theories on cognitive status/mental models of conversation participants. The capability to make correct entailment judgements based on these hypotheses will benefit many applications such as information extraction, question answering, and summarization.

As a first step in our investigation, we collected a corpus of conversation entailment data from nineteen human annotators. Our data showed that conversation entailment is more challenging than

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 206–215,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

the textual entailment task due to unique characteristics about conversation and conversational implicature. To predict entailment, we developed a probabilisitic framework that incorporates semantic representation of conversation context. Our preliminary experimental results have shown that conversation context, in particular dialogue acts, play an important role in conversation entailment.

2 Related Work

Recent work has applied different approaches to acquire information about conversation participants based on human-human conversation scripts, for example, to extract profiling information from conversation interviews (Jing et al., 2007) and to identify agreement/disagreement between participants from multiparty meeting scripts (Galley et al., 2004). In human-machine conversation, inference about conversation participants has been studied as a part of user modeling. For example, earlier work has investigated inference of user intention from utterances to control clarification dialogue (Horvitz and Paek, 2001) and recognition of user emotion and attitude from utterances for intelligent tutoring systems (Litman and Forbes-Riley, 2006). In contrast to previous work, we propose a new angle to address information acquisition about conversation participants, namely, through conversation entailment.

This work is inspired by a large body of recent work on textual entailment initiated by the PASCAL RTE Challenge (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007). Nevertheless, conversation discourse is very different from written monologue discourse. The conversation discourse is shaped by the goals of its participants and their mutual beliefs. The key distinctive features include turn-taking between participants, grounding between participants, and different linguistic phenomena of utterances (e.g., utterances in conversation tend to be shorter, with disfluency, and sometimes incomplete or ungrammatical). It is the goal of this paper to explore how techniques developed for textual entailment can be extended to address these unique behaviors in conversation entailment.

3 Experimental Data

The first step in our investigation is to collect entailment data to help us better understand the problem and facilitate algorithm development and evaluation.

3.1 Data Collection Procedure

We selected 50 dialogues from the Switchboard corpus (Godfrey and Holliman, 1997). In each of these dialogues, two participants discuss a topic of interest (e.g., sports activities, corporate culture, etc.). To focus our work on the entailment problem, we use the transcribed scripts of the dialogues in our experiments. We also make use of available annotations such as syntactic structures, disfluency markers, and dialogue acts.

We had 15 volunteer annotators read the selected dialogues and create hypotheses about participants. As a result, a total of 1096 entailment examples were created. Each example consists of a snippet from the dialogue (referred to as *dialogue segment* in the rest of this paper), a hypothesis statement, and a truth value indicating whether the hypothesis can be inferred from the snippet given the whole history of that dialogue session. During annotation, we asked the annotators to provide balanced examples for each dialogue. That is, roughly half of the hypotheses are truly entailed and half are not. Special attention was given to negative entailment examples. Since any arbitrary hypotheses that are completely irrelevant can be negative examples, a special criteria is enforced that any negative examples should have a majority word overlap with the snippet. In addition, inspired by previous work (Jing et al., 2007; Galley et al., 2004), we particularly asked annotators to provide hypotheses that address the profiling information of the participants, their opinions and desires, as well as the dynamic communicative relations between participants.

A recent study shows that for many NLP annotation tasks, the reliability of a small number of non-expert annotations is on par with that of an expert annotator (Snow et al., 2008). It also found that for tasks such as affection recognition, an average of four non-expert labels per item are capable of emulating expert-level label quality. Based on this finding, in our study the entailment judgement for each example was further independently annotated by four annotators (who were not the original contributors of the hypotheses). As a result, on average each entailment example (i.e., a pair of snippet and hypothesis) received five judgements.

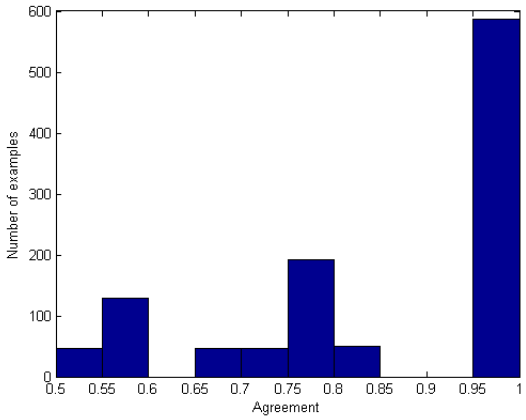


Figure 1: Agreement histogram of entailment judgements

3.2 Data and Examples

Figure 1 shows a histogram of the agreements of collected judgements. It indicates that conversation entailment is in fact a quite difficult task even for humans. Only 53% of all the examples (586 out of 1096) are agreed upon by all human annotators. The disagreement between users sometimes is caused by language ambiguity since conversation scripts are often short and without clear sentence boundaries. For example,

Example 2:

Dialogue Segment:

A: Margaret Thatcher was prime minister, uh, uh, in India, so many, uh, women are heads of state.

Hypothesis:

A believes that Margaret Thatcher was prime minister of India.

In the utterance of speaker A, the prepositional phrase *in India* is ambiguous because it can either be attached to the preceding sentence, which sufficiently entails the hypothesis; or it can be attached to the succeeding sentence, which leaves it unclear which country A believes Margaret Thatcher was prime minister of.

Difference in recognition and handling of conversational implicature is another issue that led to disagreement among annotators. For example:

Example 3:

Dialogue Segment:

A: Um, I had a friend who had fixed some, uh, chili, buffalo chili and, about a week before we went to see the movie.

Hypothesis:

A ate some buffalo chili.

Example 4:

Dialogue Segment:

B: Um, I've visited the Wyoming area. I'm not sure exactly where *Dances with Wolves* was filmed.

Hypothesis:

B thinks *Dances with Wolves* was filmed in Wyoming.

In the first example, a listener could assume that A follows the maxim of relevance. Therefore, a natural inference that makes "fixing of buffalo chili" relevant is that A ate the buffalo chili. Similarly, in the second example, the speaker A mentions a visit to Wyoming, which can be considered relevant to the filming place of DANCES WITH WOLVES. Some annotators recognized such relevance and some did not.

Given the discrepencies between annotators, we selected 875 examples which have at least 75% agreement among the judgements in our current investigation. We further selected one-third of this data (291 examples) as our development data. The experiments reported in Section 5 are based on this development set.

3.3 Types of Hypotheses

The hypotheses collected from our study can be categorzied into the following four types:

Fact. Facts about the participants. This includes: (1) profiling information about individual participants (e.g., occupation, birth place, etc.); (2) activities associated with individual participants (e.g., A bikes to work everyday); and (3) social relations between participants (e.g., A and B are coworkers, A and B went to college together).

Belief. Participants' beliefs and opinions about the physical world. Any statement about the physical world in fact is a belief of the speaker. Technically, the state of the physical world that involves the speaker him/herself is also a type of belief. However, here we assume a statement about oneself is true and is considered as a *fact*.

Desire. Participants' desire of certain actions or outcomes (e.g., A wants to find a university job). These desires represent the states of the world the participant finds pleasant (although they could be conflicting to each other).

Intent. Participants' deliberated intent, in particular communicative intention which captures the intent from one participant on the other participant such as whether A agrees/disagrees with B

on some issue, whether A intends to convince B on something, etc.

Most of these types are motivated by the Belief-Desire-Intention (BDI) model, which represents key mental states and reflects the *thoughts* of a conversation participant. *Desire* is different from *intention*. The former arises subconsciously and the latter arise from rational deliberation that takes into consideration desires and beliefs (Allen, 1995). The *fact* type represents the facts about a participant. Both thoughts and facts are critical to characterize a participant and thus important to serve many other downstream applications. The above four types account for 47.1%, 34.0%, 10.7%, and 8.2% of our development set respectively.

4 A Probabilistic Framework

Following previous work (Haghighi et al., 2005; de Salvo Braz et al., 2005; MacCartney et al., 2006), we approach conversation entailment using a probabilistic framework. To predict whether a hypothesis statement H can be inferred from a dialogue segment D, we estimate the probability

$$P(D \vDash H|D, H)$$

Suppose we have a representation of a dialogue segment D in m clauses $d_1, \ldots, d_m$ and a representation of the hypothesis H in n clauses $h_1, \ldots, h_n$. Since a hypothesis is the conjunction of the decomposed clauses, whether it can be inferred from a segment is equivalent to whether all of its clauses can be inferred from the segment. We further simplify the problem by assuming that whether a clause is entailed from a dialogue segment is conditionally independent from other clauses. Note that this conditional independence assumption is an over-simplification, but it gets things started. Therefore:

$$\begin{aligned}
&P(D \vDash H|D, H) \\
&\quad = P(d_1 \ldots d_m \vDash h_1 \ldots h_n|d_1, \ldots, d_m, h_1, \ldots, h_n) \\
&\quad = P(D \vDash h_1, \ldots, D \vDash h_n|D, h_1, \ldots, h_n) \\
&\quad = \prod_{j=1}^{n} P(D \vDash h_j|D = d_1 \ldots d_m, h_j) \\
&\quad = \prod_{j=1}^{n} P(d_1 \ldots d_m \vDash h_j|d_1, \ldots, d_m, h_j) \qquad (1)
\end{aligned}$$

If this likelihood is above a certain threshold (e.g., 0.5 in our experiments), then H is considered as a true entailment from D.

Given this framework, two important questions are: (1) how to represent and automatically create the clauses from each pair of dialogue segment and hypothesis; and (2) how to estimate probabilities as shown in Equation 1?

4.1 Clause Representation

Our clause representation is inspired by previous work on textual entailment (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007). Clause representation has several advantages. First, it can be acquired automatically from a parse tree (e.g., dependency parser). Second, it can be used to facilitate both logic-based reasoning as in (Tatu and Moldovan, 2005; Bos and Markert, 2005; Raina et al., 2005) or probabilistic reasoning as in (Haghighi et al., 2005; de Salvo Braz et al., 2005; MacCartney et al., 2006). The key difference between our work and previous work on textual entailment is the representation of conversation discourse, which has not been considered in previous work but is important for conversation entailment, as we will see later.

More specifically, a clause is made up by two components: **Term** and **Predicate**.

Term: A term can be an *entity* or an *event*. An *entity* refers to a person, a place, an organization, or other real world entities. This follows the concept of *mention* in the Automatic Content Extraction (ACE) evaluation (Doddington et al., 2004). An *event* refers to an action or an activity. For example, from the sentence "John married Eva in 1940" we can identify an event of marriage. Following the neo-Davidsonian representation (Parsons, 1990), all the events are reified as terms in our representation.

Predicate: A predicate represents either a *property* (i.e., unary) for a term or a *relation* (i.e., binary) between two terms. For example, an entity *company* has a property of *Russian* as in the phrase "a Russian company" (i.e., $Russian(company)$). An event *visit* has a property of *recently* (i.e., $recently(visit)$) as in the phrase "visit Brazil recently". From the phrase "Prime Minister recently visited Brazil", there are binary relations: *Prime Minister* is the subject of the event *visit* (i.e., $subj(visit, Prime\ Minister)$) and *Brazil* is the object of the *visit* (i.e., $obj(visit, Brazil)$).

This representation is a direct conversion from the dependency structure and can be used to represent the semantics of utterances in the dialogue

segments and the semantics of hypotheses. For example,

Example 5:

Dialogue Segment:

B: Have you seen *Sleeping with the Enemy*?
A: No. I've heard that's really great, though.
B: You have to go see that one.

Hypothesis:

B suggests A to watch *Sleeping with the Enemy*.

Appendix A shows the dependency structure of the dialogue utterances and the hypothesis from Example 5. Appendix B shows the corresponding clause representation of the dialogue segment and the hypothesis. Note that in this representation, *you* and *I* are replaced with the respective participants. Since the clauses are generated based on parse trees, most relational predicates are syntactic-driven.

To facilitate conversation entailment, we further augment the representation of a dialogue segment by incorporating conversation context. Appendix C shows the augmented representation for Example 5. It represents the following additional information:

- **Utterance:** A group of pseudo terms u_1, u_2, ... are used to represent individual utterances.
- **Participant:** A relational clause $speaker(\cdot,\cdot)$ is used to represent the speaker of this utterance, e.g., $speaker(u_1, B)$.
- **Content:** A relational clause $content(\cdot,\cdot)$ is used to represent the content of an utterance where the second term is the *head* of the utterance as identified in the parsing structure. e.g., $content(u_3, heard)$
- **Dialogue act:** A relational clause $act(\cdot,\cdot)$ is used to represent the dialogue act of the speaker for a particular utterance. e.g., $act(u_2, no\_answer)$. A set of 42 dialogue acts from the Switchboard annotation are used here (Godfrey and Holliman, 1997).
- **Utterance flow:** A relational clause $follow(\cdot,\cdot)$ is used to connect each pair of adjacent utterances. e.g., $follow(u_2, u_1)$. We currently do not consider overlap in utterances, but our representation can be modified to handle this situation by introducing additional predicates.

4.2 Entailment Prediction

Given the clause representation for a conversation segment and a hypothesis, the next step is to make an entailment prediction (as in Equation 1) based on two models: an *Alignment Model* and an *Inference Model*.

4.2.1 Alignment Model

The alignment model is to find alignments (or matches) between terms in the clause representation for a hypothesis and those in the clause representation for a conversation segment. We define an **alignment** as a mapping function g between a term x in the dialogue segment and a term y in the hypothesis. $g(x, y) = 1$ if x and y are aligned; otherwise $g(x, y) = 0$. Note that a verb can be aligned to a noun as in $g(sell, sale) = 1$. It is also possible that there are multiple terms from the segment mapped to one term in the hypothesis, or vice versa.

For any two terms x and y, the problem of predicting the alignment function $g(x, y)$ can be formulated as a binary classification problem. We used several features to train the classifier, which include whether x and y are the same (or have the same stem), whether one term is an acronym of the other, and their WordNet and distributional similarities (Lin, 1998).

Given an augmented representation with conversation context (as in Appendix C), we also align event terms in the hypothesis (e.g., *suggest* in Example 5) to (pseudo) utterance terms in the dialogue segment. We call it a *pseudo alignment*. This is currently done by a set of rules which associate event terms in the hypotheses with dialogue acts. For example, the event term *suggest* may be aligned to an utterance with dialogue act of *opinion*. Appendix D gives a correct alignment for Example 5, in which $g(u_4, x_1) = 1$ is a pseudo alignment.

4.2.2 Inference Model

As shown in Equation 1, to predict the inference of the entire hypothesis, we need to calculate the probability that the dialogue segment entails each clause from the hypothesis. More specifically, given a clause from the hypothesis h_j, a set of clauses from the dialogue segment $d_1, \ldots, d_m$, and an alignment function g between them derived by the method described in Section 4.2.1, we predict whether $d_1, \ldots, d_m$ entails h_j under the alignment g using two different classification models,

depending on whether h_j is a property or a relation (i.e. whether it takes one argument ($h_j(\cdot)$) or two arguments ($h_j(\cdot,\cdot)$)):

Given a property clause from the hypothesis, $h_j(x)$, we look for all the property clauses in the dialogue segment that describes the same term as x, i.e. a clause set $D' = \{d_i(x')|d_i(x') \in D, g(x', x) = 1\}$. Then we predict whether $h_j(x)$ can be inferred from the clauses in D' by binary classification, using a set of features similar to those used in the alignment model.

Given a relational clause from the hypothesis, $h_j(x, y)$, we look for the relation between the counterparts of x and y in the dialogue segment. That is, we find the set of terms $X' = \{x'|x' \in D, g(x', x) = 1\}$ and the set of terms $Y' = \{y'|y' \in D, g(y', y) = 1\}$ and look for the closest relation between these two sets of terms in the dependency structure. If there is a path between any $x' \in X'$ and any $y' \in Y'$ in the dependency structure with a length smaller than a threshold λ_L, we predict that $h_j(x, y)$ can be inferred. Note that our current handling of the relational clauses is rather simplified. It only captures whether two terms from an hypothesis are connected by any relation in the dialogue segment.

Appendix E shows the inference procedure of the four hypothesis clauses in Example 5. For each relational clause $h_j(x, y)$, the shortest path between the corresponding X' and Y' has a length of 3 or less, so each of these four clauses is entailed from the dialogue segment. Based on Equation 1 we can conclude that the overall hypothesis is entailed.

We trained the alignment model and the inference model (e.g.,the threshold λ_L) based on the development data provided by the PASCAL 3 challenges on textual entailment.

5 Experimental Results

To understand unique behaviors of conversation entailment, we focused our current experiments on the development dataset (see Section 3.2). We are particularly interested in how the techniques for textual entailment can be improved for conversation entailment. To do so, we applied our entailment framework on the test data of the PASCAL-3 RTE Challenge (Giampiccolo et al., 2007). Among 800 testing examples, our approach achieved an accuracy of 60.6%. This result is on par with the performance of the median system of accuracy 61.8% (z-test, p=0.63) in the PASCAL-3 RTE Challenge. Our current approach is very lean on the use of external knowledge. Its competitive performance sets up a reasonable baseline for our investigation on conversation entailment. This same system, modified to tailor linguistic characteristics of conversation (e.g., removal of disfluency), was used as the baseline in our experiments.

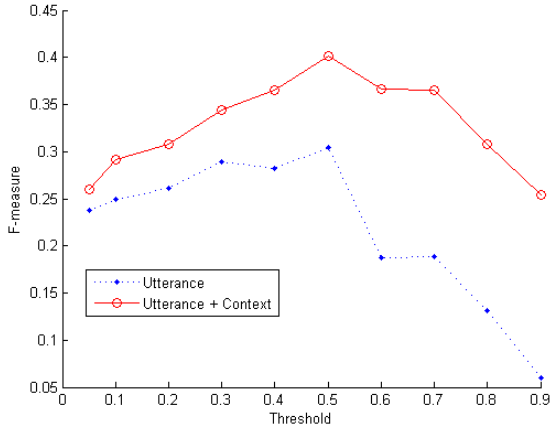
5.1 Event Alignment

To understand the effect of conversation context in the event alignment, we compared two configurations of alignment model for events. The first configuration is based on the clause representation of semantics of utterances (as shown in Appendix B). This is the same configuration as used in textual entailment. The second configuration is based on representation of both semantics from utterances and conversation context (as shown in Appendix C). We evaluate how well each configuration aligns the event terms based on the pairwise alignment decision: for any event term t_H in the hypothesis and any term t_D in the dialogue, whether the model can correctly predict that the two terms should be aligned.

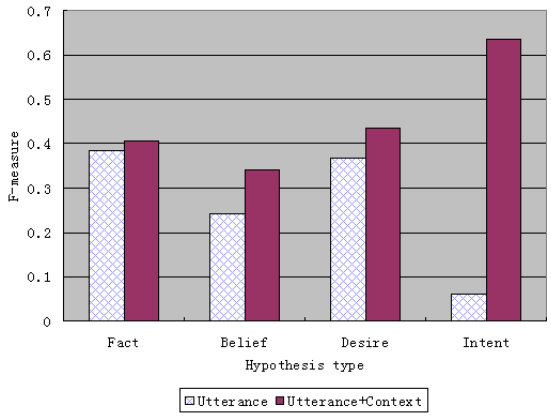
Figure 2(a) shows the comparison of F-measure between the two models. Depending on the threshold of alignment prediction, the precision and recall of the prediction vary. When the threshold is lower, the models tend to give more alignments, resulting in lower precision and higher recall. When the threshold is higher, the models tend to give fewer alignments, thus resulting in higher precision but lower recall. When the threshold is around 0.5, the alignment reaches its best F-measure. Regardless of what threshold is chosen, the model based on both utterance and context consistently works better. Figure 2(b) shows the breakdown based on the types of hypothesis (at threshold 0.5). The model that incorporates conversation context consistently performs better for all types. Its improvement is particularly significant for the *intent* type of hypothesis.

These results are not surprising. Many event terms in hypotheses (e.g., suggest, think, etc.) do not have their counterparts directly expressed in utterances in the dialogue discourse. Only through the modeling of dialog acts, these terms can be aligned to potential pseudo terms in the dialogue

segment. For the *fact* type hypotheses, the event terms in the hypotheses generally have their counterparts in the dialogue discourse. That explains why the improvement for the *fact* type using conversation context is minimal.



(a) Overall comparison on F-measure



(b) Comparison for different types of hypothesis

Figure 2: Experimental results on event alignment

5.2 Entailment Prediction

Given correct alignments, we further evaluated entailment prediction based on three configurations of the inference model: (1) the same inference model learned from the textual entailment data and tested on the PASCAL-3 RTE Challenge (Text); (2) an improved model incorporating a number of features relevant to dialogues (especially syntactic structure of utterances) based on representations without conversation context as in Appendix B (+Dialogue); (3) a further improved model based on augmented representations of conversation context and using dialogue acts during the prediction of entailment as in Appendix C (+Context).

System	Acc	Prec	Recall	F
Text	53.6%	71.6%	29.3%	41.6%
+Dialogue	58.4%	84.1%	32.3%	46.7%
+Context	67.7%	91.7%	47.0%	62.1%

Table 1: Experimental results on entailment prediction

For each configuration we present two evaluation metrics: an accuracy of the overall prediction and a precision-recall measurement for the positive entailment examples. All the evaluations are performed on our development data, which has 56.4% of positive examples and 43.6% of negative examples.

The evaluations results are shown in Table 1. The system learned from textual entailment performs lower than the prediction based on the majority class (56.4%). Incorporating syntactic features of dialogues did better but the difference is not statistically significant. Incorporating conversation context, especially dialogue acts, achieves significantly better performance (z-test, $p < 0.005$).

Table 2 shows the comparison of the three configurations based on different types of hypothesis. As expected, the basic system trained on textual entailment is not capable for any *intent* type of hypotheses. Modeling conversation context with dialogue acts improves inference for all types of hypothesis, with most significant improvement for the *belief*, *desire*, and *intent* types of hypothesis.

6 Conclusion

This paper describes our initial investigation on conversation entailment to address information acquisition about conversation participants. Since there are so many variables involved in the prediction, our experiments have been focused on a set of development data where most of the features are annotated. This allowed us to study the effect of conversation context in both alignment and entailment. Our future work will enhance the current approach by training the models based on our development data and evaluate them on the testing data. Conversation entailment is an important task. Although the current exercise is targeted to process conversation scripts from human-human conversation, it can potentially benefit human machine conversation by enabling automated agents to gain better understanding of their conversation

System	Fact		Belief		Desire		Intent	
	Acc	F	Acc	F	Acc	F	Acc	F
Text	58.4%	51.3%	52.5%	37.3%	51.6%	34.8%	33.3%	0
+Dialogue	68.6%	62.6%	53.5%	36.1%	48.4%	33.3%	33.3%	0
+Context	70.8%	64.9%	67.7%	62.8%	58.1%	47.8%	62.5%	60.9%

Table 2: Experimental results on entailment prediction for different types of hypotheses

partners.

Acknowledgments

This work was partially supported by IIS-0347548 and IIS-0840538 from the National Science Foundation. We thank the anonymous reviewers for their valuable comments and suggestions.

References

James Allen. 1995. *Natural language understanding*. The Benjamin/Cummings Publishing Company, Inc.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of HLT-EMNLP*, pages 628–635.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *PASCAL Challenges Workshop on Recognising Textual Entailment*.

Rodrigo de Salvo Braz, Roxana Girju, Vasin Punyakanok, Dan Roth, and Mark Sammons. 2005. An inference model for semantic entailment in natural language. In *Proceedings of AAAI*.

G. Doddington, A. Mitchell, M. Przybocki, and L. Ramshaw. 2004. The automatic content extraction (ace) programctasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of ACL*, pages 669–676.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9.

John J. Godfrey and Edward Holliman. 1997. *Switchboard-1 Release 2*. Linguistic Data Consortium, Philadelphia.

Aria Haghighi, Andrew Ng, and Christopher Manning. 2005. Robust textual inference via graph matching. In *Proceedings of HLT-EMNLP*, pages 387–394.

Eric Horvitz and Tim Paek. 2001. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *Proceedings of the 8th International Conference on User Modeling*, pages 3–13.

Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of ACL*, pages 1040–1047.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, pages 296–304.

Diane Litman and Katherine Forbes-Riley. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of HLT-NAACL*, pages 41–48.

Terence Parsons. 1990. *Events in the Semantics of English. A Study in Subatomic Semantics*. MIT Press.

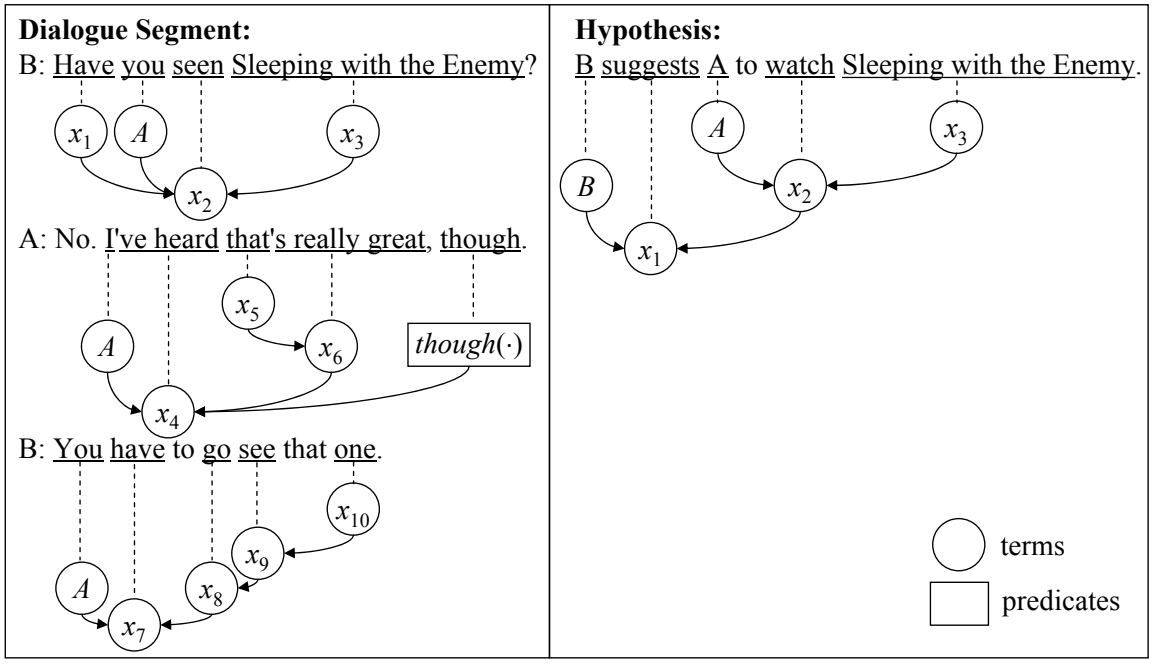
Rajat Raina, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via learning and abductive reasoning. In *Proceedings of AAAI*, pages 1099–1105.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of HLT-EMNLP*, pages 371–378.

APPENDIX

A Dependency Structure of Dialogue Utterances and Hypothesis in Example 5



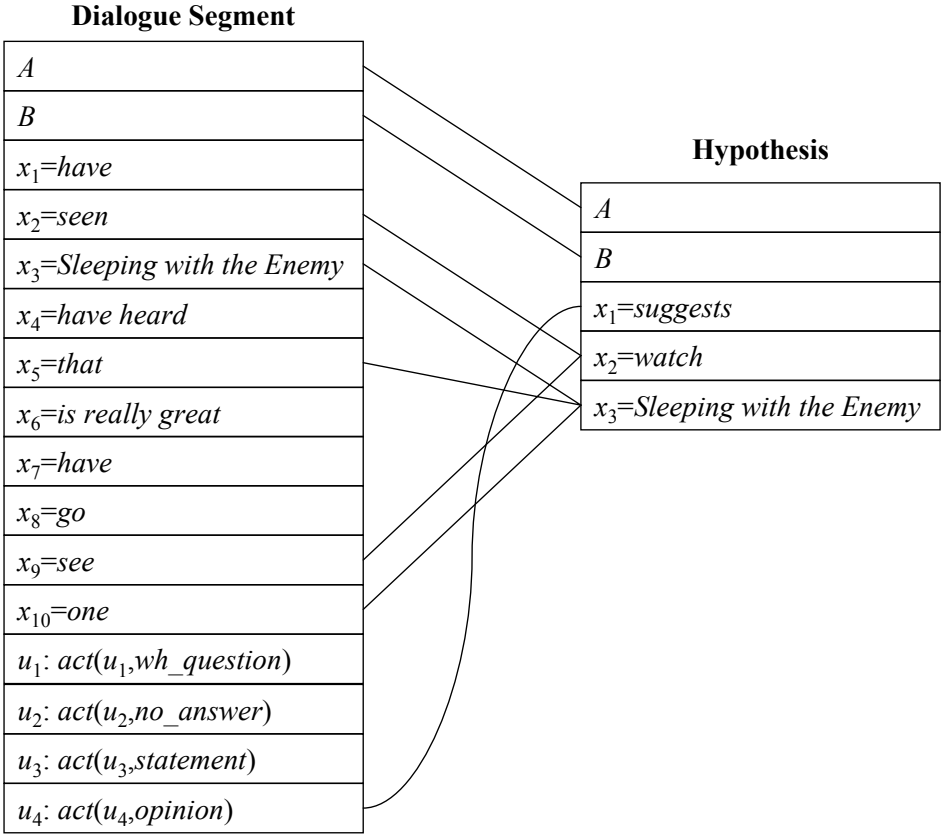
B Clause Representation of Dialogue Segment and Hypothesis for Example 5

Dialogue Segment:		
	Terms	Clauses
B:	x_1=*have*, x_2=*seen*, x_3=*Sleeping with the Enemy*, *A*	*subj*(x_2,*A*), *obj*(x_2,x_3), *aux*(x_2,x_1)
A:	x_4=*have heard*, x_5=*that*, x_6=*is really great*, *A*	*subj*(x_4,*A*), *obj*(x_4,x_6), *subj*(x_6,x_5), *though*(x_4)
B:	x_7=*have*, x_8=*go*, x_9=*see*, x_{10}=*one*, *A*	*subj*(x_7,*A*), *obj*(x_7,x_8), *obj*(x_8,x_9), *obj*(x_9,x_{10})
Hypothesis:		
	x_1=*suggests*, x_2=*watch*, x_3=*Sleeping with the Enemy*, *A*, *B*	*subj*(x_1,*B*), *obj*(x_1,*A*), *obj*(x_1,x_2), *obj*(x_2,x_3)

C Augmented Clause Representation of Dialogue Segment in Example 5

Dialogue Segment (with context representation):		
	Terms	Clauses
B:	u_1,x_1=*have*, x_2=*seen*, x_3=*Sleeping with the Enemy*, *A*, *B*	*speaker*(u_1,*B*), *content*(u_1,x_2), *act*(u_1,*wh_question*), *subj*(x_2,*A*), *obj*(x_2,x_3), *aux*(x_2,x_1)
A:	u_2, u_3, x_4=*have heard*, x_5=*that*, x_6=*is really great*, *A*	*speaker*(u_2,*A*), *content*(u_2,-), *act*(u_2,*no_answer*), *speaker*(u_3,*A*), *content*(u_3,x_4), *act*(u_3,*statement*), *subj*(x_4,*A*), *obj*(x_4,x_6), *subj*(x_6,x_5), *though*(x_4)
B:	u_4,x_7=*have*, x_8=*go*, x_9=*see*, x_{10}=*one*, *A*, *B*	*speaker*(u_4,*B*), *content*(u_4,x_7), *act*(u_4,*opinion*), *subj*(x_7,*A*), *obj*(x_7,x_8), *obj*(x_8,x_9), *obj*(x_9,x_{10})
		follow(u_2,u_1), *follow*(u_3,u_2), *follow*(u_4,u_3)

D The Alignment for Example 5



E The Prediction of Inference for the Hypothesis Clauses in Example 5

Hypothesis Clause	*subj*(x_1,*B*)		*obj*(x_1,*A*)		*obj*(x_1,x_2)		*obj*(x_2,x_3)	
Clause Type	relation		relation		relation		relation	
Terms in this Clause	x_1	*B*	x_1	*A*	x_1	x_2	x_2	x_3
Aligned Terms in the Dialogue Segment	u_4	*B*	u_4	*A*	u_4	x_2, x_9	x_2, x_9	x_3, x_5, x_{10}
Shortest Path between the Aligned Terms in the Dependency Structure of Dialogue Segment	*speaker*(u_4,*B*)		*content*(u_4,x_7), *subj*(x_7,*A*)		*content*(u_4,x_7), *obj*(x_7,x_8), *obj*(x_8,x_9)		*obj*(x_9,x_{10})	
Path Length	1		2		3		1	
Hypothesis Clause Entailed?	yes		yes		yes		yes	