Towards Conversation Entailment: An Empirical Investigation

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

While a significant amount of research has been devoted to textual entailment, automated entailment from conversational scripts has received less attention. To address this limitation, this paper investigates the problem of conversation entailment: automated inference of hypotheses from conversation scripts. We examine two levels of semantic representations: a basic representation based on syntactic parsing from conversation utterances and an augmented representation taking into consideration of conversation structures. For each of these levels, we further explore two ways of capturing long distance relations between language constituents: implicit modeling based on the length of distance and explicit modeling based on actual patterns of relations. Our empirical findings have shown that the augmented representation with conversation structures is important, which achieves the best performance when combined with explicit modeling of long distance relations.

1 Introduction

Textual entailment has received increasing attention in recent years (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2008; Bentivogli et al., 2009). Given a segment from a textual document, the task of textual entailment is to automatically determine whether a given hypothesis can be entailed from the segment. The capability of such kind of inference can benefit many text-based applications such as information extraction and question answering.

Textual entailment has mainly focused on inference from written text in monologue. Recent years also observed an increasing amount of conversational data such as conversation scripts of meetings, call center records, court proceedings, as well as online chatting. Although conversation is a form of language, it is different from monologue text with several unique characteristics. The key distinctive features include turn-taking between participants, grounding between participants, different linguistic phenomena of utterances, and conversation implicatures. Traditional approaches dealing with textual entailment were not designed to handle these unique conversation behaviors and thus to support automated entailment from conversation scripts.

Example 1:
Conversation Segment:
B: My mother also was very very independent. She had her own, still had her own little house and still driving her own car,
A: Yeah.
B: at age eighty-three.
Hypothesis:
(1) B's mother is eighty-three.
(2) B is eighty-three.

To address this limitation, our previous work (Zhang and Chai, 2009) has initiated an investigation on the problem of *conversation entailment*. The problem was formulated as follows: given a conversation discourse *D* and a hypothesis *H* concerning its participant, the goal was to identify whether *D* entails *H*. For instance, as in Example 1, the first hypothesis can be entailed from the

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 756–766,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

conversation segment while the second hypothesis cannot. While our previous work has provided some interesting preliminary observations, it mostly focused on data collection and initial experiments and analysis using a small set of development data. It is not clear whether the previous results are generally applicable, how different components in the entailment framework interact with each other, and how different representations may influence the entailment outcome.

To reach a better understanding of conversation entailment, we conducted a further investigation based on the larger set of test data collected in our previous work (Zhang and Chai, 2009). We specifically examined two levels of representations: a basic representation based on syntactic parsing from conversation utterances and an augmented representation taking into consideration of conversation structures. For each of these levels, we further explored two ways of capturing long distance relations: (1) implicit modeling based on the length of distance and (2) explicit modeling based on actual patterns of relations. Our empirical findings have shown that augmented representation with conversation structures is important in conversation entailment. Combining conversation structures with explicit modeling of long distance relations results in the best performance.

2 Related Work

Our work here is related to recent advances in textual entailment, automated processing of conversation scripts, and our initial investigation on conversation entailment.

There is a large body of work on textual entailment initiated by the Pascal Recognizing Textual Entailment (RTE) Challenges (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2008; Bentivogli et al., 2009). Different approaches have been developed, for example, based on logic proving (Tatu and Moldovan, 2005; Bos and Markert, 2005; Raina et al., 2005) and graph match (Haghighi et al., 2005; de Salvo Braz et al., 2005; MacCartney et al., 2006). Supervised learning approaches have also been applied to measure the similarities between training and testing pairs (Zanzotto and Moschitti, 2006). In the most recent RTE Challenge (Bentivogli et al., 2009), the best system achieves 73.5% of accuracy, while the median performance among all participants is 60.4%. These results indicate that, while progress has been made, textual entailment remains a challenging problem.

As more and more conversation data becomes available, researchers have investigated automated processing of conversation data to acquire useful information, for example, related to opinions (Somasundaran et al., 2007; Somasundaran et al., 2008; Somasundaran et al., 2009), biographic attributes (Garera and Yarowsky, 2009), social networks (Jing et al., 2007), and agreements and disagreements between participants (Galley et al., 2004). Recent studies have also developed approaches to summarize conversations (Murray and Carenini, 2008) and to model conversation structures (dialogue acts) from online Twitter conversations (Ritter et al., 2010). Here we address a different angle regarding conversation scripts, namely conversation entailment.

In our previous work (Zhang and Chai, 2009), we started an initial investigation on conversation entailment. We have collected a dataset of 875 instances. Each instance consists of a conversation segment and a hypothesis (as described in Section 1). The hypotheses are statements about conversation participants and are further categorized into four types: about their profile information, their beliefs and opinions, their desires, and their communicative intentions. We developed an approach that is motivated by previous work on textual entailment. We use clauses in the logic-based approaches as the underlying representation of our system. Based on this representation, we apply a two stage entailment process similar to MacCartney et al. (2006) developed for textual entailment: an alignment stage followed by an entailment stage.

Building upon our previous work, in this paper, we systematically examine different representations of the conversation segment and different modeling of long distance relations between language constituents. We compare the roles of these different representations on the performance of entailment prediction using a larger testing dataset that was not previously evaluated. This analysis allows better understanding of the problem and provides insight on potential solutions.

3 Overall Framework

In our previous work (Zhang and Chai, 2009), conversation entailment is formulated as the following: given a conversation segment D which is represented by a set of clauses $D = d_1 \wedge \ldots \wedge d_m$, and a hypothesis H represented by another set of clauses $H = h_1 \wedge \ldots \wedge h_n$, the prediction on whether D entails H is determined by the product of probabilities that each hypothesis clause h_j is entailed from all the conversation segment clauses $d_1 \ldots d_m$ as follows. This is based on a simple assumption that whether a clause is entailed from a conversation segment is conditionally independent from other clauses.

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

A clause here is similar to a sentence in first-order predicate calculus. It is made up by *terms* and *predicates*. A term is either: 1) an entity described by a noun phrase, e.g., *John Lennon*, *mother*, or *she*; or 2) an action or event described by a verb phrase, e.g., *marry* in "John married Eva in 1940". A predicate represents either: 1) a property (i.e., unary) for a term, e.g., $Russian(company)$, or $recently(visit)$; or 2) a relation (i.e., binary) between two terms, e.g., $subj(visit, Prime\ Minister)$ and $obj(visit, Brazil)$ in "Prime Minister recently visited Brazil".

Given the clause representation, we follow the idea similar to MacCartney et al. (2006), and predict the entailment decision in two stages of processing: (1) an alignment model aligns terms in the hypothesis to terms in the conversation segment; and (2) an inference model predicts the entailment based on the alignment between the hypothesis and the conversation segment.

3.1 Alignment Model

An **alignment** is defined as a mapping function g between a term x in the conversation segment and a term y in the hypothesis. $g(x, y) = 1$ if x and y are aligned; otherwise $g(x, y) = 0$. It is possible that multiple terms from the segment are mapped to one term in the hypothesis ($g(x_1, y) = g(x_2, y) = 1$), or vice versa ($g(x, y_1) = g(x, y_2) = 1$). To predict these alignments, the problem is formulated as binary classification: given any two terms x from the conversation and y from the hypothesis, decide the value of their alignment function $g(x, y)$.

3.2 Inference Model

Once an alignment between a hypothesis and a conversation segment is established, an inference model is applied to predict whether the conversation segment entails the hypothesis given such alignment. More specifically, as shown in Equation 1, given a clause from the hypothesis h_j, a set of clauses from the conversation segment $d_1, \ldots, d_m$, and an alignment g between them, the goal is to predict whether $d_1, \ldots, d_m$ entails h_j under the alignment g.

The prediction is treated differently according to different types of clauses. If h_j is a property clause (i.e., takes one argument $h_j(\cdot)$), a property inference model is applied; otherwise (i.e., relational clauses with two arguments $h_j(\cdot, \cdot)$), a relational inference model is applied.

In this paper we follow the same framework. However our focus here is on the new question that how different levels of semantic representation and different approaches of modeling long distance relationship affect the alignment and inference models as well as the overall entailment performance.

4 Semantic Representation

Given the clause representation described earlier, an important question is what information from the conversation segment should be captured and represented. To address this question, we examined two levels of shallow semantic representation. The first level is basic representation which only captures the information from all the utterances in the conversation segment. The second representation includes conversation structures (e.g., speakers and dialogue acts). Next we use Example 2 to illustrate these representations.

Example 2:
Conversation 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*.

4.1 Basic Representation

The first representation is based on the syntactic parsing from conversation utterances and we call it a *basic* representation. Figure 1(a) shows an example of dependency structures for several utterances that are derived from the Stanford parser (Klein and Manning, 2003), and Figure 1(b) shows the corresponding clause representation. In the dependency structure, the vertices represent entities (e.g., x_1) and actions (e.g., x_3) within an utterance. They correspond to terms in the clause representation. An edge between vertices captures a dependency relation and is represented as predicates in the clause representation. For example, the edge between x_1 and x_3 indicates x_1 is the subject of x_3, which is represented by the clause representation $subj(x_3, x_1)$. Similar representation also applies to the hypothesis as shown in Figure 1(c), 1(d).

4.2 Augmented Representation

The second representation is built upon the basic representation and incorporates conversation structure across turns and utterances. We call it an *augmented* representation. Figure 2(a) shows the augmented structures of the conversation segment and Figure 2(b) shows the corresponding clause representation. Compared to the basic representation, there are two additional types of vertices (i.e., terms) highlighted in the figures:

- Vertices representing utterances (e.g., $u_1 \ldots u_4$). Their corresponding terms capture the dialogue acts for the utterances (e.g. $u_1 = yes\_no\_question$). To focus our effort, currently we only apply annotated dialogue acts provided in the Switchboard corpus (Godfrey and Holliman, 1997). Two edges are added to connect different utterances. The first edge connects each utterance vertex to the head of the corresponding utterance to indicate the specific content of the utterance (e.g., $content(u_1, x_3)$). The second edge connects an utterance to its succeeding utterance to indicate the temporal progression of the conversation (e.g., $follow(u_2, u_1)$).
- Vertices representing speakers or participants (e.g., s_A, s_B). One edge is added to connect each utterance to its speaker (e.g., $speaker(u_1, s_B)$).

Note that since our clause representations are mainly based on the dependency relations, they are mostly syntactic-driven. However, it does capture some shallow semantics such as who is the agent (i.e., subject) or the patient (i.e., object) of an event. The incorporation of speakers and dialogue acts in our augmented representations provides additional semantics of conversation discourse.

5 Modeling LDR

A critical part in predicting entailment is to recognize the semantic relationship between two language constituents, especially when these two constituents are not directly related. In Figure 2(a), for example, we want to recognize that x_9 (*You*) is the (logical) subject of x_{11} (see). Here we experimented two ways of modeling such long distance relations (LDR).

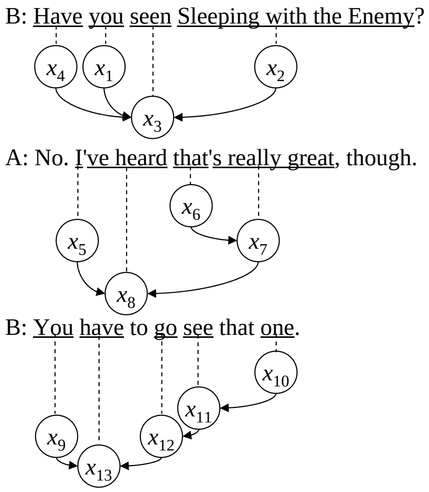
5.1 Implicit Modeling of LDR

The first method characterizes the relationship simply by the distance between two constituents in the dependency structure (or augmented structure). For example, in Figure 2(a) the distance between x_{11} and x_9 is 3. We call this method an *implicit* modeling of long distance relationship.

The advantage of implicit modeling is that it is easy to implement based on the dependency structure. However, its limitation is that the distance measure does not capture sufficient information of semantic relations between language constituents.

5.2 Explicit Modeling of LDR

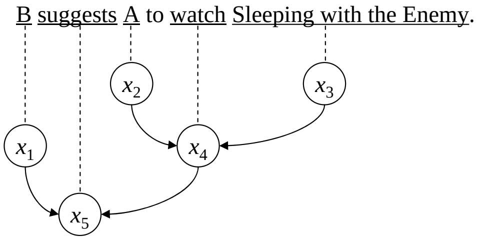
The second way of modeling long distance relationship is called *explicit* modeling. It uses a string to


B: Have you seen Sleeping with the Enemy?
x_4
x_1
x_2
x_3
A: No. I've heard that's really great, though.
x_6
x_5
x_7
x_8
B: You have to go see that one.
x_{10}
x_{11}
x_9
x_{12}
x_{13}

(a) dependency structure of the conversation utterances

Terms	Clauses
x_1=*A* x_2=*Sleeping with the Enemy* x_3=*seen*, x_4=*have*	*obj*(x_3,x_2) *subj*(x_3,x_1) *aux*(x_3,x_4)
x_5=*A*, x_6=*that* x_7=*is really great* x_8=*have heard*	*subj*(x_7,x_6) *obj*(x_8,x_7) *subj*(x_8,x_5)
x_9=*A*, x_{10}=*one*, x_{11}=*see*, x_{12}=*go*, x_{13}=*have*	*obj*(x_{11},x_{10}) *obj*(x_{12},x_{11}) *obj*(x_{13},x_{12}) *subj*(x_{13},x_9)

(b) basic representation of the conversation segment

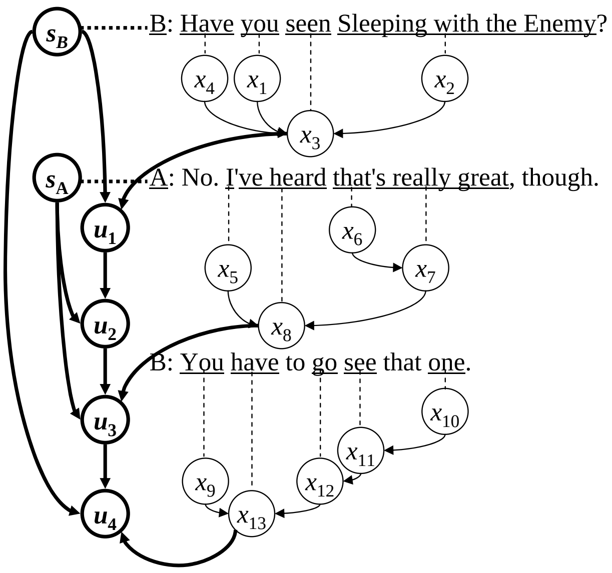


(c) dependency structure of the hypothesis

Terms	Clauses
x_1=*B*, x_2=*A* x_3=*Sleeping with the Enemy* x_4=*watch* x_5=*suggests*	*subj*(x_4,x_2) *obj*(x_4,x_3) *subj*(x_5,x_1) *obj*(x_5,x_4)

(d) representation of the hypothesis

Figure 1: The dependency structures and corresponding basic representation of Example 2


s_B
B: Have you seen Sleeping with the Enemy?
x_4
x_1
x_2
x_3
s_A
A: No. I've heard that's really great, though.
u_1
x_6
x_5
x_7
u_2
x_8
B: You have to go see that one.
u_3
x_{10}
x_{11}
x_9
x_{12}
x_{13}
u_4

(a) dependency and conversation structures of the conversation segment

Terms	Clauses

s_A*, *s_B
u_1=yes_no_question
	speaker(u_1,s_B)
	content(u_1, x_3)

x_1=*A*, x_3=*seen*, x_4=*have* x_2=*Sleeping with the Enemy*	*obj*(x_3,x_2), *subj*(x_3,x_1) *aux*(x_3,x_4)

u_2=no_answer	***speaker(u_2,s_A)***
	follow(u_2,u_1)

u_3=statement	***speaker(u_3,s_A)***
	content(u_3.x_8)
	follow(u_3,u_2)

x_5=*A*, x_7=*is really great* x_6=*that*, x_8=*have heard*	*subj*(x_7,x_6), *obj*(x_8,x_7) *subj*(x_8,x_5)

u_4=viewpoint	***speaker(u_4,s_B)***
	content(u_4,x_{13})
	follow(u_4,u_3)

x_9=*A*, x_{10}=*one*, x_{11}=*see* x_{12}=*go*, x_{13}=*have*	*obj*(x_{11},x_{10}), *obj*(x_{12},x_{11}) *obj*(x_{13},x_{12}), *subj*(x_{13},x_9)

(b) augmented representation of the conversation segment

Figure 2: The dependency and conversation structures and corresponding augmented representation of Example 2

describe the path from one constituent to the other: $v_1 e_1 \ldots v_{l-1} e_{l-1} v_l$, where $v_1, \ldots, v_l$ are the vertices on the path and $e_1, \ldots, e_{l-1}$ are the edges. Each v_i describes the type of the vertex in the dependency structure, which is either a noun (N), a verb (V), or an utterance (U). Each e_i describes whether the edge is forward (→) or backward (←). For example, in Figure 2(a), the path from x_{11} to x_9 is $V \rightarrow V \rightarrow V \leftarrow N$.

This kind of string representation of paths in syntactic parse is known as a way of modeling "shallow semantics" between any two constituents in a language structure. It is largely used in other NLP tasks such as semantic role labeling (Pradhan et al., 2008). The difference here is our paths are extracted from dependency parses as opposed to traditional constituent parses, and our paths also incorporate the representation of conversation structures (e.g., utterances and speakers).

6 Applications in Entailment Models

In this section we describe how different representations and modeling of LDR are used in the alignment and inference models.

6.1 Applications in Alignment Model

Although a noun and a verb can potentially be aligned, to simplify the problem, we restrict the problem to the alignment between two nouns or two verbs. We trained an alignment model for nouns and one for verbs separately.

Table 1 summarizes a set of features used in the alignment models. Most of these features are shared by the model for noun alignment and the model for verb alignment. These features include whether the two strings are the same, two terms have the same stem, the similarity between the two terms either based on WordNet or distributional statistics (Lin, 1998).

To learn the alignment model for nouns, we annotated the noun alignments for the development data used in PASCAL RTE-3 Challenge (Giampiccolo et al., 2007) and trained a logistic regression model based on the features in Table 1. Cross-validation on the same dataset shows relatively satisfying performance (96.4% precision and 94.9% recall). In this paper, we focus on the alignment between verbs since it appears more difficult.

	Noun Align.	Verb Align.
Verb *be* identification		✓
String equality	✓	✓
Stemmed equality	✓	✓
Acronym equality	✓	
Named entity equality	✓	
WordNet similarity	✓	✓
Distributional similarity	✓	✓
Subject consistency		✓
Object consistency		✓

Table 1: Features for alignment models

A major difference between noun alignment and verb alignment is that, for verb alignment the consistency of their arguments is also important. For two events (described by two verbs) to be aligned, at least their subjects (usually denoting the executers of actions) and objects (usually denoting the receivers of actions) should match to each other respectively. Note that, although actions/events also depend on other arguments or adjuncts, here we only consider the subjects and objects and leave the consistency check of other arguments/adjuncts to downstream processes. Based on two different ways of modeling long distance relationship (as described in Section 5), we explored two methods for modeling argument consistency (AC) in verb alignment models.

6.1.1 Implicit Modeling of AC

The first approach models argument consistency based on implicit modeling of the relationship between a verb and its aligned subject/object. Specifically, given a pair of verb terms (x, y) where x is from the conversation segment and y is from the hypothesis, let s_y be the subject of y and s_x be the aligned entity of s_y in the conversation (in case of multiple alignments, s_x is the one closest to x). The subject consistency of the verbs (x, y) is then measured by the distance between s_x and x in the dependency structure. Similarly, the distance between a verb and its aligned object is used as a measure of the object consistency.

In Example 2, to decide whether the conversation term *see* (x_{11} in Figure 1(a), 1(b), and 2) and the hypothesis term *watch* (x_4 in Figure 1(c), 1(d)) should be aligned, we first identify the subject of x_4 in the hypothesis, which is x_2 (A). We then look for

x_2's alignments in the conversation segment, among which x_9 (*You*) is the closest to x_{11} (*see*). In Figure 2(a), we find the distance between x_{11} and x_9 is 3.

Using the implicit modeling of argument consistency, we follow the same approach as in our previous work (Zhang and Chai, 2009) and trained a logistic regression model to predict verb alignment based on the features in Table 1.

6.1.2 Explicit Modeling of AC

The second approach captures argument consistency based on explicit modeling of the relationship between a verb and its aligned subject (or object). Given a pair of verb terms (x, y), let s_y be the subject of y and s_x be the aligned entity of s_y in the conversation closest to x, we use the string describing the path from x to s_x as the feature to capture subject consistency. For example, in Figure 2(a), the path from x_{11} to x_9 is $V \rightarrow V \rightarrow V \leftarrow N$.

This string representation of paths is used to capture both the subject consistency and the object consistency. Since they are non-numerical features, and the variability of their values can be extremely large, so we applied an instance-based classification model (e.g., k-nearest neighbor) to determine alignments between verb terms. We measure the distance between two path features by their minimal string edit distance, and then simply use the Euclidean distance to measure the closeness between any two verbs. Again this model is trained from our development data described in Zhang and Chai (2009).

Figure 3 shows an example of alignment between the conversation terms and hypothesis terms in Example 2. Note that in this figure the alignment between $x_5 = suggests$ from the hypothesis and $u_4 = opinion$ from the conversation segment is a *pseudo alignment*, which directly maps a verb term in the hypothesis to an utterance term represented by its dialogue act. This alignment is obtained by following the same set of rules learned from the development dataset as in (Zhang and Chai, 2009).

6.2 Applications in Inference Model

As mentioned earlier, once an alignment is established, the inference model is to predict whether each clause in the hypothesis is entailed from the conversation segment. Two separate models were used to handle the inference of property clauses ($h_j(x)$) and and the inference of relational clauses ($h_j(x, y)$). Property clauses involve less variables and are relatively simple, so we used the same property inference model as in (Zhang and Chai, 2009). Here we focus on relational inference model and examine how different modeling of long distance relationship may affect relation inference.

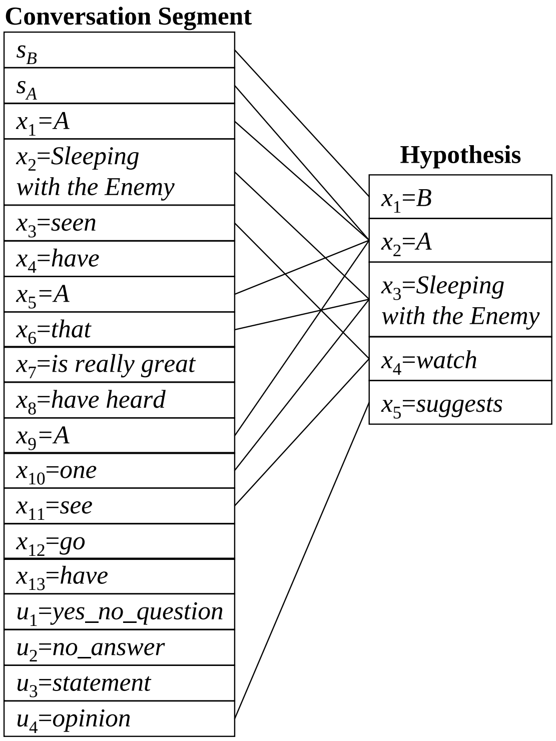


Figure 3: The alignment result for Example 2

For a relation h between x and y to be entailed from a conversation segment, we need to find a same or similar relation in the conversation segment between x's and y's counterparts (i.e., aligned entities of x and y in the conversation segment).

More specifically, given a relational clause from the hypothesis, $h_j(x, y)$, we find the sets of terms $X' = \{x'|x' \in D, g(x', x) = 1\}$ and $Y' = \{y'|y' \in D, g(y', y) = 1\}$, which are aligned with x and y, respectively. We then find the closest relation between these two sets of terms, (x^*, y^*), such that the distance between x^* and y^* is the smallest for any $x^* \in X'$ and $y^* \in Y'$. For instance, in the hypothesis of Example 2 there are terms x_5=*suggests* and x_4=*watch*, and a relational clause $obj(x_5, x_4)$ describing an action-object relation between them. Their counterparts in the conversation segment are $X' = \{u_4 = \mathit{viewpoint}\}$ and $Y' = \{x_3 = \mathit{seen}, x_{11} = \mathit{see}\}$. So the closest pair of terms between these two sets is u_4 and x_{11}. Consequently, whether the target relational clause $h_j(x, y)$ is entailed is determined by the relationship between x^* and y^*. Such relationship can be modeled either implicitly or explicitly.

6.3 Implicit modeling of relation inference

In this model we follow the simple idea that the shorter a path is between two terms, the more likely these two terms have a direct relationship. So we predefine a threshold, λ_L. We predict that $h_j(x, y)$ is entailed if the distance between x^* and y^* is smaller than λ_L. However, as can be seen, this distance does not reflect whether the type of relationship between x^* and y^* is similar to the relationship that holds between x and y.

6.4 Explicit modeling of relation inference

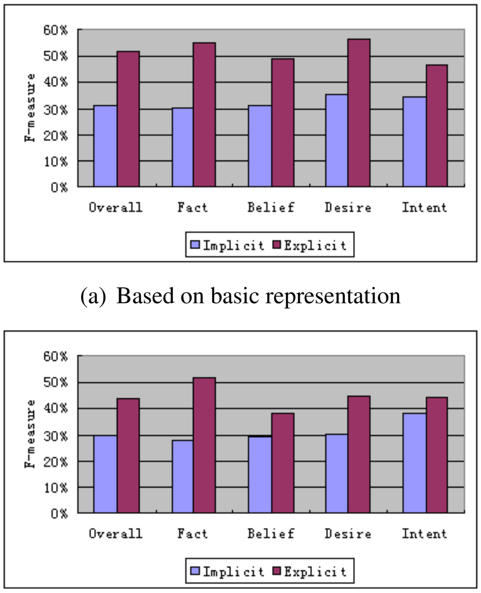
In order to capture more semantics from the relation between two terms, we use explicit modeling of the relationship between terms x^* and y^*. In the previous example, the relationship between u_4 and x_{11} is modeled by the path from u_4 to x_{11}, $U \leftarrow V \leftarrow V \leftarrow V$.

Given this characterization, the prediction of whether $h_j(x, y)$ is entailed from the conversation segment is formulated as a binary classification problem, using a k-nearest neighbor classification model with following features:

1. Explicit modeling of long distance relationship, i.e., the path from x^* to y^* in the dependency structure of the conversation segment;
2. The types (*N*, *V*, or *U*) of x, y, x^*, and y^*;
3. The type of relation between x and y, for example, *obj* in $obj(x, y)$;
4. The order (i.e., before or after) between x and y, and between x^* and y^*;
5. The specific type of the hypothesis.

7 Evaluation and Analysis

We evaluated different model configurations using our data[1]. This dataset consists of 291 development instances and 584 testing instances. The hypotheses were categorized into four types: (1) fact: profile and social relations of conversation participants (accounted for 47% of the development data and 49% of the testing data); (2) belief: participants' beliefs and opinions (34% and 35%); (3) desire: participants' desire of certain actions or outcomes (11% and 4%); (4) intent: communicative intent that captures some perlocutionary force from one participant to the other (e.g. A stops B from doing something; A disagreees with B on something, 8% and 12%)



(a) Based on basic representation

(b) Based on augmented representation

Figure 4: Evaluation of verb alignment

Note that in our original work (Zhang and Chai, 2009), only development data were used to show some initial observations. Here we trained our models on the development data and results shown are from the testing data.

7.1 Evaluation of Alignment Models

The evaluation of alignment models is based on pairwise decision. For each pair of terms (x, y), where x is from a conversation segment and y is from a hypothesis, we measure whether the model correctly predicts that the two terms should or should not be aligned. Because the alignment classification has extremely unbalanced classes, we use precision-recall of true alignments as evaluation metrics.

Figure 4(a) and 4(b) shows the comparison (F-measure) of two alignment models for verb align-

[1] The data is available for download at `http://links.cse.msu.edu:8000/lair/projects/conversationentailment_data.html`.

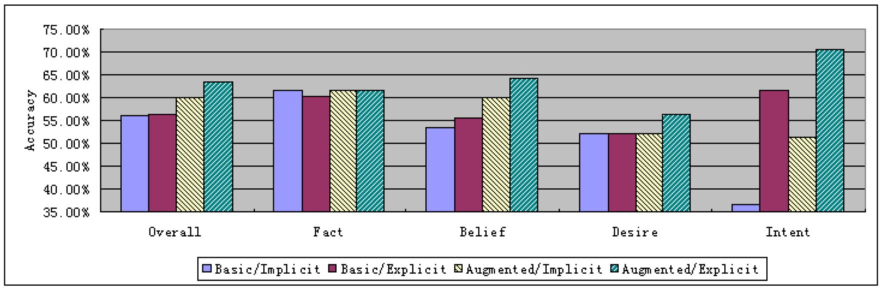


Figure 5: Evaluation of inference models based on different representations

ment, based on the basic representation and the augmented representation, respectively. Note that we cannot directly compare the results between these two figures since they involve different number of alignment instances[2]. Nevertheless, we can see the overall trend within each figure: the explicit model outperforms the implicit model. This suggests that the explicit modeling of semantic relationship between verbs and arguments works better than the implicit modeling used in previous work. Furthermore, the improvement is most noticeable when hypotheses are *facts* (24.8% with the basic representation and 24.1% with the augmented representation), and least when hypotheses are *intents* (12.2% with the basic representation and 6.2% with the augmented representation).

7.2 Evaluation of Inference Models

In order to compare different inference models, in this section (and this section only) we use gold-standard alignment results. They are obtained from manual annotation in our evaluation. We evaluated two inference models, one with implicit modeling of long distance relationship and one with explicit modeling. Evaluations were conducted based on both the basic representation and the augmented representation. Figure 5 shows the four groups of evaluation results.

Overall speaking, the augmented representation outperforms the basic representation for both implicit modeling and explicit modeling of long distance relationship (McNemar's tests, $p < 0.05$). The explicit model performs better than implicit model only based on augmented representation (McNemar's test, $p < 0.05$).

Clause Representation	Relation modeling		Improvement
	Implicit	Explicit	
Basic	53.9%	53.9%	0
Augmented	54.8%	58.7%	**3.9%**

Table 2: Entailment performance with different representations and LDR modeling

The results were further broken down by different hypothesis types. For the *fact* type of hypotheses, there is no difference between different representations and modeling of long distance relationship. This is not surprising since most hypotheses about partipants' profiling information can be inferred directly from the utterances. The augmented representation affects the *intent* type of hypothesis most significantly, so does the explicit modeling of long distance relationship.

7.3 Interaction between Clause Representations and LDR Modeling

It was shown in previous sections that the augmented representation helps entailment prediction compared to the basic representation. Here we want to study how they interact with other entailment components and what is their effect in the enhanced modeling of long distance relations. Specifically, we test the performance of implicit and explicit modeling of long distance relations under two different representation settings: the basic representation and the augmented representation.

Table 2 compares the performance (accuracy) of entailment models with different relationship modeling. We can see that the explicit model makes improvement over the implicit model for augmented representation (McNemar's test, $p < 0.05$), while no improvement is made for basic representation. These evaluation results appear to suggest that there

[2] The alignment based on the augmented representation in Figure 4(b) also includes pseudo alignments.

is an interaction between clause representations and semantic modeling of long distance relations: the modeling of long distance relations between language constituents appears only effective when conversation structure is incorporated in the representation.

It is interesting to see the difference in the prediction performances on *fact* hypotheses and *intent* hypotheses. For *fact*, the most benefit of incorporating explicit modeling of long distance relationship appears at the alignment stage, but not much at the inference stage. However, this situation is different for *intent*, where the benefit of explicitly modeling long distance relationship mostly happened at the inference stage. This observation suggests that the effects of different types of modeling may vary for different types of hypotheses, which indicates that hypothesis type dependent models may be beneficial.

8 Discussion and Conclusion

This paper presents an empirical investigation on conversation entailment. We specifically examine two levels of representation of conversation segments and two different ways of modeling long distance relations between language constituents. Our findings indicate that, although traditional architecture and approaches for textual entailment remain important, additional representation and processing that address conversation structures is critical. The augmented representation with conversation structures, together with explicit modeling of semantic relations between language constituents, results in the best performance (58.7% accuracy).

The work here only represents an initial step towards conversation entailment. Conversation phenomena are rich and complex. Conversation entailment is extremely difficult. Besides the same challenges faced by textual entailment, it is further complicated by conversation implicature. Although our current data enables us to start an initial investigation, its small size poses significant limitations on technology development and evaluation. For example, our studies have indicated hypothesis type-dependent approaches may be beneficial, however we do not have sufficient data to yield reasonable models. A more systematical approach to collect and create a larger set of data is crucial. Innovative community-based approaches (e.g., through web) for data collection and annotation can be pursued in the future. As more techniques in semantic processing (e.g., semantic role) become available, future work should also capture deeper semantics, address pragmatics, and incorporate richer world knowledge.

Finally, as the technology in conversation entailment is developed, its applications in NLP problems should be explored. Example applications include information extraction, question answering, summarization from conversation scripts, and modeling of conversation participants. These applications may provide new insights on the nature of the conversation entailment problem and its potential solutions.

Acknowledgments

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

References

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.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*.

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

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*.

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.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, Suntec, Singapore, August.

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.

Danilo Giampiccolo, Hoa Trang Dang, Bernardog Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

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.

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.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA.

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

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.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 773–782, Honolulu, Hawaii, October.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310.

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.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California, June.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, September.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK, August.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179, Singapore, August.

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

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 401–408, Morristown, NJ, USA.

Chen Zhang and Joyce Chai. 2009. What do we know about conversation participants: Experiments on conversation entailment. In *Proceedings of the SIGDIAL 2009 Conference*, pages 206–215.