An Evaluation Understudy for Dialogue Coherence Models

Sudeep Gandhe and **David Traum**
Institute for Creative Technologies
University of Southern California
13274 Fiji way, Marina del Rey, CA, 90292
{gandhe,traum}@ict.usc.edu

Abstract

Evaluating a dialogue system is seen as a major challenge within the dialogue research community. Due to the very nature of the task, most of the evaluation methods need a substantial amount of human involvement. Following the tradition in machine translation, summarization and discourse coherence modeling, we introduce the the idea of evaluation understudy for dialogue coherence models. Following (Lapata, 2006), we use the information ordering task as a testbed for evaluating dialogue coherence models. This paper reports findings about the reliability of the information ordering task as applied to dialogues. We find that simple n-gram co-occurrence statistics similar in spirit to BLEU (Papineni et al., 2001) correlate very well with human judgments for dialogue coherence.

1 Introduction

In computer science or any other research field, simply building a system that accomplishes a certain goal is not enough. It needs to be thoroughly evaluated. One might want to evaluate the system just to see to what degree the goal is being accomplished or to compare two or more systems with one another. Evaluation can also lead to understanding the shortcomings of the system and the reasons for these. Finally the evaluation results can be used as feedback in improving the system.

The best way to evaluate a novel algorithm or a model for a system that is designed to aid humans in processing natural language would be to employ it in a real system and allow users to interact with it. The data collected by this process can then be used for evaluation. Sometimes this data needs further analysis - which may include annotations, collecting subjective judgments from humans, etc. Since human judgments tend to vary, we may need to employ multiple judges. These are some of the reasons why evaluation is time consuming, costly and sometimes prohibitively expensive.

Furthermore, if the system being developed contains a machine learning component, the problem of costly evaluation becomes even more serious. Machine learning components often optimize certain free parameters by using evaluation results on held-out data or by using n-fold cross-validation. Evaluation results can also help with feature selection. This need for repeated evaluation can forbid the use of data-driven machine learning components.

For these reasons, using an automatic evaluation measure as an understudy is quickly becoming a common practice in natural language processing tasks. The general idea is to find an automatic evaluation metric that correlates very well with human judgments. This allows developers to use the automatic metric as a stand-in for human evaluation. Although it cannot replace the finesse of human evaluation, it can provide a crude idea of progress which can later be validated. e.g. BLEU (Papineni et al., 2001) for machine translation, ROUGE (Lin, 2004) for summarization.

Recently, the discourse coherence modeling community has started using the information ordering task as a testbed to test their discourse coherence models (Barzilay and Lapata, 2005; Soricut and Marcu, 2006). Lapata (2006) has proposed an au-

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 172–181,
Columbus, June 2008.

tomatic evaluation measure for the information ordering task. We propose to use the same task as a testbed for dialogue coherence modeling. We evaluate the reliability of the information ordering task as applied to dialogues and propose an evaluation understudy for dialogue coherence models.

In the next section, we look at related work in evaluation of dialogue systems. Section 3 summarizes the information ordering task and Lapata's (2006) findings. It is followed by the details of the experiments we carried out and our observations. We conclude with a summary future work directions.

2 Related Work

Most of the work on evaluating dialogue systems focuses on human-machine communication geared towards a specific task. A variety of evaluation metrics can be reported for such task-oriented dialogue systems. Dialogue systems can be judged based on the performance of their components like WER for ASR (Jurafsky and Martin, 2000), concept error rate or F-scores for NLU, understandability for speech synthesis etc. Usually the core component, the dialogue model - which is responsible for keeping track of the dialogue progression and coming up with an appropriate response, is evaluated indirectly. Different dialogue models can be compared with each other by keeping the rest of components fixed and then by comparing the dialogue systems as a whole. Dialogue systems can report subjective measures such as user satisfaction scores and perceived task completion. SASSI (Hone and Graham, 2000) prescribes a set of questions used for eliciting such subjective assessments. The objective evaluation metrics can include dialogue efficiency and quality measures.

PARADISE (Walker et al., 2000) was an attempt at reducing the human involvement in evaluation. It builds a predictive model for user satisfaction as a linear combination of some objective measures and perceived task completion. Even then the system needs to train on the data gathered from user surveys and objective features retrieved from logs of dialogue runs. It still needs to run the actual dialogue system and collect objective features and perceived task completeion to predict user satisfaction.

Other efforts in saving human involvement in evaluation include using simulated users for testing (Eckert et al., 1997). This has become a popular tool for systems employing reinforcement learning (Levin et al., 1997; Williams and Young, 2006). Some of the methods involved in user simulation are as complex as building dialogue systems themselves (Schatzmann et al., 2007). User simulations also need to be evaluated as how closely they model human behavior (Georgila et al., 2006) or as how good a predictor they are of dialogue system performance (Williams, 2007).

Some researchers have proposed metrics for evaluating a dialogue model in a task-oriented system. (Henderson et al., 2005) used the number of slots in a frame filled and/or confirmed. Roque et al. (2006) proposed hand-annotating information-states in a dialogue to evaluate the accuracy of information state updates. Such measures make assumptions about the underlying dialogue model being used (e.g., form-based or information-state based etc.).

We are more interested in evaluating types of dialogue systems that do not follow these task-based assumptions: systems designed to imitate human-human conversations. Such dialogue systems can range from chatbots like Alice (Wallace, 2003), Eliza (Weizenbaum, 1966) to virtual humans used in simulation training (Traum et al., 2005). For such systems, the notion of task completion or efficiency is not well defined and task specific objective measures are hardly suitable. Most evaluations report the subjective evaluations for appropriateness of responses. Traum et. al. (2004) propose a coding scheme for response appropriateness and scoring functions for those categories. Gandhe et. al. (2006) propose a scale for subjective assessment for appropriateness.

3 Information Ordering

The information ordering task consists of choosing a presentation sequence for a set of information bearing elements. This task is well suited for text-to-text generation like in single or multi-document summarization (Barzilay et al., 2002). Recently there has been a lot of work in discourse coherence modeling (Lapata, 2003; Barzilay and Lapata, 2005; Soricut and Marcu, 2006) that has used information ordering to test the coherence models. The information-bearing elements here are sentences rather than high-level concepts. This frees the models from having to depend on a hard to get training corpus which has been hand-authored for concepts.

Most of the dialogue models still work at the higher abstraction level of dialogue acts and intentions. But with an increasing number of dialogue systems finding use in non-traditional applications such as simulation training, games, etc.; there is a need for dialogue models which do not depend on hand-authored corpora or rules. Recently Gandhe and Traum (2007) proposed dialogue models that do not need annotations for dialogue-acts, semantics and hand-authored rules for information state updates or finite state machines.

Such dialogue models focus primarily on generating an appropriate coherent response given the dialogue history. In certain cases the generation of a response can be reduced to selection from a set of available responses. For such dialogue models, maintaining the information state can be considered as a secondary goal. The element that is common to the information ordering task and the task of selecting next most appropriate response is the ability to express a preference for one sequence of dialogue turns over the other. We propose to use the information ordering task to test dialogue coherence models. Here the information bearing units will be dialogue turns.[1]

There are certain advantages offered by using information ordering as a task to evaluate dialogue coherence models. First the task does not require a dialogue model to take part in conversations in an interactive manner. This obviates the need for having real users engaging in the dialogue with the system. Secondly, the task is agnostic about the underlying dialogue model. It can be a data-driven statistical model or information-state based, form based or even a reinforcement learning system based on MDP or POMDP. Third, there are simple objective measures available to evaluate the success of information ordering task.

Recently, Purandare and Litman (2008) have used this task for modeling dialogue coherence. But they only allow for a binary classification of sequences as either coherent or incoherent. For comparing different dialogue coherence models, we need the ability for finer distinction between sequences of information being put together. Lapata (2003) proposed Kendall's τ, a rank correlation measure, as one such candidate. In a recent study they show that Kendall's τ correlates well with human judgment (Lapata, 2006). They show that human judges can reliably provide coherence ratings for various permutations of text. (Pearson's correlation for inter-rater agreement is 0.56) and that Kendall's τ is a good indicator for human judgment (Pearson's correlation for Kendall's τ with human judgment is 0.45 ($p < 0.01$)).

Before adapting the information ordering task for dialogues, certain questions need to be answered. We need to validate that humans can reliably perform the task of information ordering and can judge the coherence for different sequences of dialogue turns. We also need to find which objective measures (like Kendall's τ) correlate well with human judgments.

4 Evaluating Information Ordering

One of the advantages of using information ordering as a testbed is that there are objective measures available to evaluate the performance of information ordering task. Kendall's τ (Kendall, 1938), a rank correlation coefficient, is one such measure. Given a reference sequence of length n, Kendall's τ for an observed sequence can be defined as,

$$\tau = \frac{\text{\# concordant pairs} - \text{\# discordant pairs}}{\text{\# total pairs}}$$

Each pair of elements in the observed sequence is marked either as concordant - appearing in the same order as in reference sequence or as discordant otherwise. The total number of pairs is $C_2^n = n(n-1)/2$. τ ranges from -1 to 1.

Another possible measure can be defined as the fraction of n-grams from reference sequence, that are preserved in the observed sequence.

$$b_n = \frac{\text{\# n-grams preserved}}{\text{\# total n-grams}}$$

In this study we have used, b_2, fraction of bigrams and b_3, fraction of trigrams preserved from the reference sequence. These values range from 0 to 1. Table 1 gives examples of observed sequences and

[1]These can also be at the utterance level, but for this paper we will use dialogue turns.

Observed Sequence	b_2	b_3	τ
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]	1.00	1.00	1.00
[8, 9, 0, 1, 2, 3, 4, 5, 6, 7]	0.89	0.75	0.29
[4, 1, 0, 3, 2, 5, 8, 7, 6, 9]	0.00	0.00	0.60
[6, 9, 8, 5, 4, 7, 0, 3, 2, 1]	0.00	0.00	-0.64
[2, 3, 0, 1, 4, 5, 8, 9, 6, 7]	0.56	0.00	0.64

Table 1: Examples of observed sequences and their respective b_2, b_3 & τ values. Here the reference sequence is [0,1,2,3,4,5,6,7,8,9].

respective b_2, b_3 and τ values. Notice how τ allows for long-distance relationships whereas b_2, b_3 are sensitive to local features only. [2]

5 Experimental Setup

For our experiments we used segments drawn from 9 dialogues. These dialogues were two-party human-human dialogues. To ensure applicability of our results over different types of dialogue, we chose these 9 dialogues from different sources. Three of these were excerpts from role-play dialogues involving negotiations which were originally collected for a simulation training scenario (Traum et al., 2005). Three are from SRI's Amex Travel Agent data which are task-oriented dialogues about air travel planning (Bratt et al., 1995). The rest of the dialogues are scripts from popular television shows. Fig 6 shows an example from the air-travel domain. Each excerpt drawn was 10 turns long with turns strictly alternating between the two speakers.

Following the experimental design of (Lapata, 2006) we created random permutations for these dialogue segments. We constrained our permutations so that the permutations always start with the same speaker as the original dialogue and turns strictly alternate between the speakers. With these constraints there are still $5! \times 5! = 14400$ possible permutations per dialogue. We selected 3 random permutations for each of the 9 dialogues. In all, we have a total of 27 dialogue permutations. They are arranged in 3 sets, each set containing a permutation for all 9 dialogues. We ensured that not all permutations in a given set are particularly very good or very bad. We used Kendall's τ to balance the permutations across the given set as well as across the given dialogue.

Unlike Lapata (2006) who chose to remove the pronouns and discourse connectives, we decided not do any pre-processing on the text like removing disfluencies or removing cohesive devices such as anaphora, ellipsis, discourse connectives, etc. One of the reason is such pre-processing if done manually defeats the purpose of removing humans from the evaluation procedure. Moreover it is very difficult to remove certain cohesive devices such as discourse deixis without affecting the coherence level of the original dialogues.

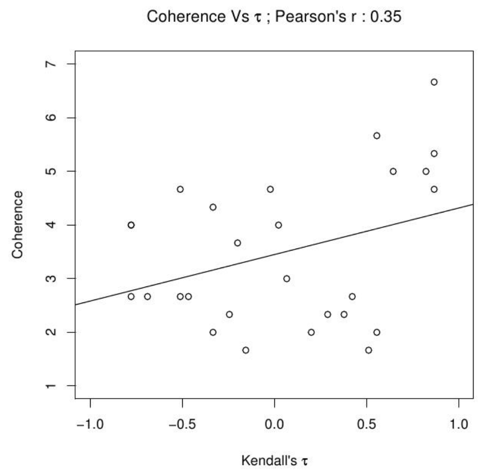
6 Experiment 1

In our first experiment, we divided a total of 9 human judges among the 3 sets (3 judges per set). Each judge was presented with 9 dialogue permutations. They were asked to assign a single coherence rating for each dialogue permutation. The ratings were on a scale of 1 to 7, with 1 being very incoherent and 7 being perfectly coherent. We did not provide any additional instructions or examples of scale as we wanted to capture the intuitive idea of coherence from our judges. Within each set the dialogue permutations were presented in random order.

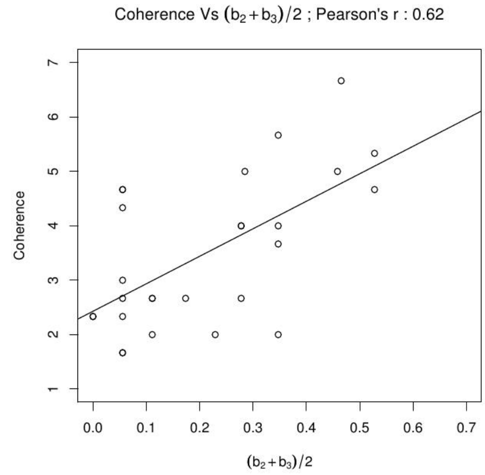
We compute the inter-rater agreement by using Pearson's correlation analysis. We correlate the ratings given by each judge with the average ratings given by the judges who were assigned the same set. For inter-rater agreement we report the average of 9 such correlations which is 0.73 (std dev = 0.07). Artstein and Poesio (2008) have argued that Krippendorff's α (Krippendorff, 2004) can be used for inter-rater agreement with interval scales like the one we have. In our case for the three sets α values were 0.49, 0.58, 0.64. These moderate values of alpha indicate that the task of judging coherence is indeed a difficult task, especially when detailed instructions or examples of scales are not given.

In order to assess whether Kendall's τ can be used as an automatic measure of dialogue coherence, we perform a correlation analysis of τ values against the average ratings by human judges. The Pearson's correlation coefficient is 0.35 and it is statistically not significant (P=0.07). Fig 1(a) shows the relationship between coherence judgments and τ values. This experiment fails to support the suitability

[2] For more on the relationship between b_2, b_3 and τ see row 3,4 of table 1 and figure 4.



(a) Kendall's τ does not correlate well with human judgments for dialogue coherence.



(b) Fraction of bigram & trigram counts correlate well with human judgments for dialogue coherence.

Figure 1: Experiment 1 - single coherence rating per permutation

of Kendall's τ as an evaluation understudy.

We also analyzed the correlation of human judgments against simple n-gram statistics, specifically $(b_2 + b_3)/2$. Fig 1(b) shows the relationship between human judgments and the average of fraction of bigrams and fraction of trigrams that were preserved in the permutation. The Pearson's correlation coefficient is 0.62 and it is statistically significant ($P<0.01$).

7 Experiment 2

Since human judges found it relatively hard to assign a single rating to a dialogue permutation, we decided to repeat experiment 1 with some modifications. In our second experiment we asked the judges to provide coherence ratings at every turn, based on the dialogue that preceded that turn. The dialogue permutations were presented to the judges through a web interface in an incremental fashion turn by turn as they rated each turn for coherence (see Fig 5 in the appendix for the screenshot of this interface). We used a scale from 1 to 5 with 1 being completely incoherent and 5 as perfectly coherent. [3] A total of 11 judges participated in this experiment with the first set being judged by 5 judges and the remaining two sets by 3 judges each.

For the rest of the analysis, we use the average coherence rating from all turns as a coherence rating for the dialogue permutation. We performed the inter-rater agreement analysis as in experiment 1. The average of 11 correlations is 0.83 (std dev = 0.09). Although the correlation has improved, Krippendorff's α values for the three sets are 0.49, 0.35, 0.63. This shows that coherence rating is still a hard task even when judged turn by turn.

We assessed the relationship between the average coherence rating for dialogue permutations with Kendall's τ (see Fig 2(a)). The Pearson's correlation coefficient is 0.33 and is statistically not significant ($P=0.09$).

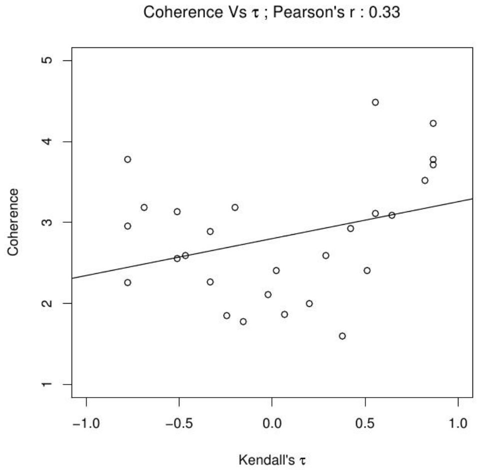
Fig 2(b) shows high correlation of average coherence ratings with the fraction of bigrams and trigrams that were preserved in permutation. The Pearson's correlation coefficient is 0.75 and is statistically significant ($P<0.01$).

Results of both experiments suggest that, $(b_2 + b_3)/2$ correlates very well with human judgments and can be used for evaluating information ordering when applied to dialogues.

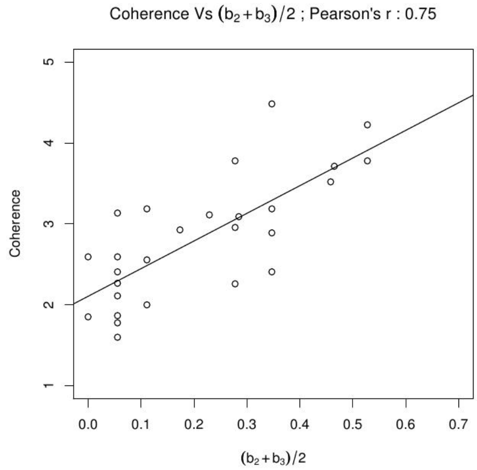
8 Experiment 3

We wanted to know whether information ordering as applied to dialogues is a valid task or not. In this experiment we seek to establish a higher baseline for

[3] We believe this is a less complex task than experiment 1 and hence a narrower scale is used.



(a) Kendall's τ does not correlate well with human judgments for dialogue coherence.



(b) Fraction of bigram & trigram counts correlate well with human judgments for dialogue coherence.

Figure 2: Experiment 2 - turn-by-turn coherence rating

the task of information ordering in dialogues. We presented the dialogue permutations to our human judges and asked them to reorder the turns so that the resulting order is as coherent as possible. All 11 judges who participated in experiment 2 also participated in this experiment. They were presented with a drag and drop interface over the web that allowed them to reorder the dialogue permutations. The reordering was constrained to keep the first speaker of the reordering same as that of the original dialogue and the re-orderings must have strictly alternating turns. We computed the Kendall's τ and fraction of bigrams and trigrams $(b_2 + b_3)/2$ for these re-orderings. There were a total of 11×9 = 99 reordered dialogue permutations. Fig 3(a) and 3(b) shows the frequency distribution of τ and $(b_2 + b_3)/2$ values respectively.

Humans achieve high values for the reordering task. For Kendall's τ, the mean of the reordered dialogues is 0.82 (std dev = 0.25) and for $(b_2 + b_3)/2$, the mean is 0.71 (std dev = 0.28). These values establish an upper baseline for the information ordering task. These can be compared against the random baseline. For τ random performance is 0.02 [4] and for $(b_2 + b_3)/2$ it is 0.11. [5]

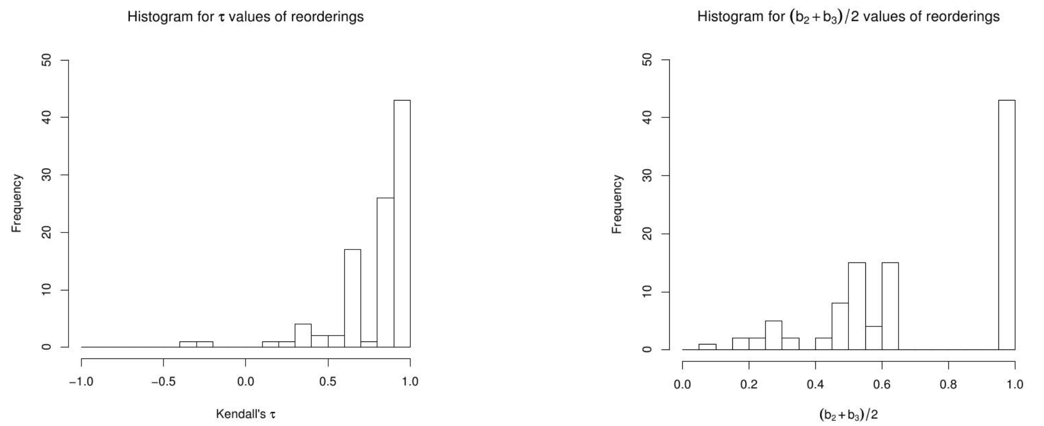
9 Discussion

Results show that $(b_2 + b_3)/2$ correlates well with human judgments for dialogue coherence better than Kendall's τ. τ encodes long distance relationships in orderings where as $(b_2 + b_3)/2$ only looks at local context. Fig 4 shows the relationship between these two measures. Notice that most of the orderings have τ values around zero (i.e. in the middle range for τ), whereas majority of orderings will have a low value for $(b_2 + b_3)/2$. τ seems to overestimate the coherence even in the absence of immediate local coherence (See third entry in table 1). It seems that local context is more important for dialogues than for discourse, which may follow from the fact that dialogues are produced by two speakers who must react to each other, while discourse can be planned by one speaker from the beginning. Traum and Allen (1994) point out that such social obligations to respond and address the contributions of the other should be an important factor in building dialogue systems.

The information ordering paradigm does not take into account the content of the information-bearing items, e.g. the fact that turns like "yes", "I agree",

[4] Theoretically this should be zero. The slight positive bias is the result of the constraints imposed on the re-orderings - like only allowing the permutations that have the correct starting speaker.

[5] This value is calculated by considering all 14400 permutations as equally likely.



(a) Histogram of Kendall's τ for reordered sequences

(b) Histogram of fraction of bigrams & trigrams values for reordered sequences

Figure 3: Experiment 3 - upper baseline for information ordering task (human performance)

”okay” perform the same function and should be treated as replaceable. This may suggest a need to modify some of the objective measures to evaluate the information ordering specially for dialogue systems that involve more of such utterances.

Human judges can find the optimal sequences with relatively high frequency, at least for short dialogues. It remains to be seen how this varies with longer dialogue lengths which may contain sub-dialogues that can be arranged independently of each other.

10 Conclusion & Future Work

Evaluating dialogue systems has always been a major challenge in dialogue systems research. The core component of dialogue systems, the dialogue model, has usually been only indirectly evaluated. Such evaluations involve too much human effort and are a bottleneck for the use of data-driven machine learning models for dialogue coherence. The information ordering task, widely used in discourse coherence modeling, can be adopted as a testbed for evaluating dialogue coherence models as well. Here we have shown that simple n-gram statistics that are sensitive to local features correlate well with human judgments for coherence and can be used as an evaluation understudy for dialogue coherence models. As with any evaluation understudy, one must be careful while using it as the correlation with human judgments is not perfect and may be inaccurate in some cases – it can not completely replace the need for full evaluation with human judges in all cases (see (Callison-Burch et al., 2006) for a critique of BLUE along these lines).

In the future, we would like to perform more experiments with larger data sets and different types of dialogues. It will also be interesting to see the role cohesive devices play in coherence ratings. We would like to see if there are any other measures or certain modifications to the current ones that correlate better with human judgments. We also plan to employ this evaluation metric as feedback in building dialogue coherence models as is done in machine translation (Och, 2003).

Acknowledgments

The effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We would like to thank Radu Soricut, Ron Artstein, and the anonymous SIGdial reviewers for helpful comments.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. In *To appear in Computational Linguistics*.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proc. ACL-05*.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument summarization. *JAIR*, 17:35–55.

Harry Bratt, John Dowding, and Kate Hunicke-Smith. 1995. The sri telephone-based atis system. In *Proceedings of the Spoken Language Systems Technology Workshop*, January.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. In *proceedings of EACL-2006*.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding*, pages 80–87, Dec.

Sudeep Gandhe and David Traum. 2007. Creating spoken dialogue characters from corpora without annotations. In *Proceedings of Interspeech-07*.

Sudeep Gandhe, Andrew Gordon, and David Traum. 2006. Improving question-answering with linking dialogues. In *International Conference on Intelligent User Interfaces (IUI)*, January.

Kalliroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *proceedings of Interspeech*.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2005. Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In *proceedings of IJCAI workshop*.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*.

Daniel Jurafsky and James H. Martin. 2000. *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Klaus Krippendorff. 2004. *Content Analysis, An Introduction to Its Methodology 2nd Edition*. Sage Publications.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Mirella Lapata. 2006. Automatic evaluation of information ordering. *Computational Linguistics*, 32(4):471–484.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1997. Learning dialogue strategies within the markov decision process framework. In *Automatic Speech Recognition and Understanding*, pages 72–79, Dec.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *In ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, July.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Technical Report RC22176 (W0109-022), IBM Research Division*, September.

Amruta Purandare and Diane Litman. 2008. Analyzing dialog coherence using transition patterns in lexical and semantic features. In *Proceedings 21st International FLAIRS Conference*, May.

Antonio Roque, Hua Ai, and David Traum. 2006. Evaluation of an information state-based dialogue manager. In *Brandial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue*.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *proceedings of HLT/NAACL*, Rochester, NY.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proc. ACL-06*.

David R. Traum and James F. Allen. 1994. Discourse obligations in dialogue processing. In *proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 1–8.

David R. Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *In Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1699–1702.

David Traum, William Swartout, Jonathan Gratch, and Stacy Marsella. 2005. Virtual humans for non-team interaction training. In *AAMAS-05 Workshop on Creating Bonds with Humanoids*, July.

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*.

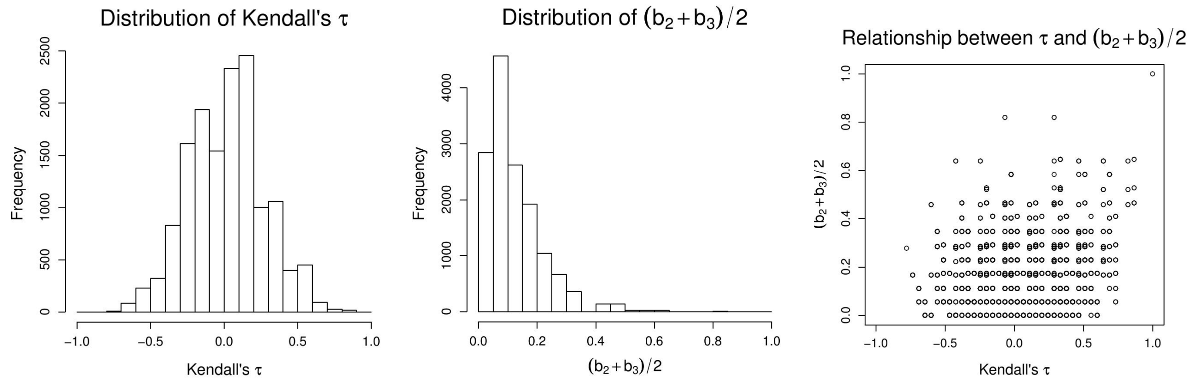
Richard Wallace. 2003. *Be Your Own Botmaster, 2nd Edition*. ALICE A. I. Foundation.

Joseph Weizenbaum. 1966. Eliza–a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, January.

Jason D. Williams and Steve Young. 2006. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21:393–422.

Jason D. Williams. 2007. A method for evaluating and comparing user simulations: The cramer-von mises divergence. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Appendix



(a) (b) (c)

Figure 4: Distributions for Kendall's τ, $(b_2 + b_3)/2$ and the relationship between them for all possible dialogue permutations with 10 turns and earlier mentioned constraints.

Speaker	Text	Coherence Rating
JACK	Yeah this is Bauer.	
TONY	Jack, this is Tony. You guys are around 10 minutes from the airport?	★★★★☆
JACK	Do everything you have to.	★☆☆☆☆
TONY	No, not yet. We're still working on it. Look- you'll link up with the assault team, they're set outside the emergency room. They're sweeping the ground now, they're starting to lock it down.	★★★☆☆
JACK	Okay, that sounds alright. Tony, contacted Kim yet?	★★★★☆
TONY	No, I haven't. I just called the sheriff's station, they sent out search teams. I just don't understand why she hasn't tried contacted us yet.	★★★★★
JACK	Yeah, that's seems about right. Look- you got a complete tail number on the aircraft yet?	★★☆☆☆
TONY	Believe me; I'm putting every resource we can spare on it.	★★★☆☆
JACK	She knows she can't go back to Los Angeles, and right now the LAPD's looking for her as a murder suspect. She's not gonna reach out to us- do you understand? We've got to find her. We've got to find her.	★★★☆☆
TONY	Okay.	★★★★☆

Figure 5: Screenshot of the interface used for collecting coherence rating for dialogue permutations.

Agent	AAA at American Express may I help you?
User	yeah this is BBB BBB I need to make some travel arrangements
Agent	ok and what do you need to do?
User	ok on June sixth from San Jose to Denver, United
Agent	leaving at what time?
User	I believe there's one leaving at eleven o'clock in the morning
Agent	leaves at eleven a.m. and arrives Denver at two twenty p.m. out of San Jose
User	ok
Agent	yeah that's United flight four seventy
User	that's the one

Doctor	hello i'm doctor perez how can i help you
Captain	uh well i'm with uh the local i'm i'm the commander of the local company and uh i'd like to talk to you about some options you have for relocating your clinic
Doctor	uh we're not uh planning to relocate the clinic captain what uh what is this about
Captain	well have you noticed that there's been an awful lot of fighting in the area recently
Doctor	yes yes i have we're very busy we've had many more casual+ casualties many more patients than than uh usual in the last month but uh what what is this about relocating our clinic have have uh you been instructed to move us
Captain	no but uh we just have some concerns about the increase in fighting xx
Doctor	are you suggesting that we relocate the clinic because we had no plans we uh we uh we're located here and we've been uh we are located where the patients need us
Captain	yeah but yeah actually it is a suggestion that you would be a lot safer if you moved away from this area we can put you in an area where there's n+ no insurgents and we have the area completely under control with our troops
Doctor	i see captain is this a is this a suggestion from your commander
Captain	i'm uh the company commander

Figure 6: Examples of the dialogues used to elicit human judgments for coherence