Towards Intelligent QA Interfaces: Discourse Processing for Context Questions

Mingyu Sun
Department of Linguistics
Michigan State University
East Lansing, MI 48824
sunmingy@msu.edu

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

ABSTRACT

Question answering (QA) systems take users' natural language questions and retrieve relevant answers from large repositories of free texts. Despite recent progress in QA research, most work on question answering is still focused on isolated questions. In a real-world information seeking scenario, questions are not asked in isolation, but rather in a coherent manner that involves a sequence of related questions to meet users' information needs. Therefore, to support coherent information seeking, intelligent QA interfaces will inevitably require techniques to support context question answering. To address this problem, this paper investigates approaches to discourse processing of a sequence of coherent questions and their implications on query expansion. In particular, we examine three models for query expansion that are motivated by Centering Theory. Our empirical results indicate that more sophisticated processing based on discourse transitions and centers can significantly improve the performance of document retrieval compared to models that only resolve references.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Theory and method, Natural language

H. 3.1 [**Content Analysis and Indexing**]: Linguistic processing

H. 3.3 [**Information Search and Retrieval**]: Query formulation

General Terms

Algorithms, Management, Experimentation.

Keywords

QA interfaces, question answering, context management, discourse processing

1. INTRODUCTION

Question answering (QA) systems take users' natural language questions and automatically locate answers from large collections of documents. The QA technologies have advanced tremendously in the past few years, partly motivated by a series of evaluations conducted at the Text REtrieval Conference (TREC) [25]. The state-of-the-art system can answer almost 85% of factoid questions such as "What Spanish explorer discovered the Mississippi River?" in an open domain [22].

Despite the tremendous progress in retrieving answers to isolated questions, the work that addresses a coherent sequence of questions is still limited. When seeking for information, a user very often has an overall information goal in mind, for example, information about the ecological system in Hawaii. To satisfy this information goal, the user may need to specify a sequence of related questions as follows:

(1) Q1: Where is Hawaii located?
Q2: What is the state fish?
Q3: Is it endangered?
Q4: Any other endangered species?

We consider this QA process coherent because the questions are not isolated, but rather evolving and related to serve a specific information goal. As the example shown above, each subsequent question relates to its preceding question(s). For example, question (1Q2) relates to (1Q1) since it asks about the state fish of Hawaii. Question (1Q3) relates to (1Q2) about the endangerment of the fish and (1Q4) relates to the whole discourse about other endangered species in Hawaii. This example indicates that each of these questions needs to be interpreted in a particular context as the question answering process proceeds. From a linguistic point of view, these questions are related to each other through different linguistic expressions and devices such as the definite noun phrase in (1Q2), pronoun in (1Q3), and ellipsis in (1Q4). A key question is how to use the discourse information to process these context questions and facilitate answer retrieval.

To address this question, we turn to Centering Theory that models local coherence of discourse [9]. Centering Theory describes how different linguistic devices (e.g., pronouns) are used to maintain the local coherence of a discourse and minimize hearer's inference load. In coherent question answering, users also tend to maintain the coherence of discourse. This is evident by the example (1) as well as the data provided in the TREC 2004 (described later). Therefore, this paper examines how Centering Theory can be used to process discourse and link key pieces of information together from a sequence of context questions. In particular, three models based on Centering Theory have been implemented to model the question discourse and guide query expansion: (1) a reference model that resolves pronoun references for each question, (2) a forward model that adds query terms from the previous question based on its forward looking centers, and (3) a transition model that selectively adds query terms according to the transitions identified between adjacent questions.

IUI'06, January 26–February 1, 2006, Sydney, Australia.

In our current investigation, rather than a complete end-to-end study, this paper focuses on discourse processing for query expansion. A good retrieval component based on the expanded queries can be integrated with other sophisticated answer extraction techniques to improve the end-to-end performance. In particular, we evaluated three models concerning their performance in document retrieval on two data sets: the data collected in our studies and the data provided by TREC 2004. The empirical results indicate that Centering Theory based approaches provide better performance for entity related context questions (e.g., about Hawaii) as opposed to event-based context questions (e.g., about presidential election in 2004). The transition model and the forward model consistently outperform the reference resolution model.

In the following sections, we first give a brief introduction to Centering Theory, then describe three models for discourse processing, and finally present our empirical studies and results.

2. RELATED WORK

The term *context* in *context question answering* can refer to different context such as user context [19] and discourse context [25]. In this paper, we focus on the discourse context, in particular the discourse of a sequence of questions. The question answering based on the discourse context was first investigated in TREC 10 Question Answering Track [25]. The context task was designed to investigate the system capability to track context through a series of questions. However, as described in [25], there were two unexpected results of this task. First, the evaluations of systems have shown that the ability to identify the correct answer to a question in the later series had no correlation with the capability to identify correct answers to the preceding questions. Second, since the first question in a series already restricted answers to a small set of documents, the performance was determined by whether the system could answer a particular type of question, rather than the ability to track context. The results from TREC 10 motivate more systematic studies of discourse processing for context question answering.

Discourse processing for context questions can range from shallow processing to deep processing. One example of shallow processing is an algorithm developed by Language Computer Corporation (LCC). In this algorithm, to process a given question, the system first identifies a prior question in the discourse that contains potential referent to a referring expression in the current question and then combines that prior question with the current question to retrieve documents. This algorithm has shown to be effective in TREC 10 context question evaluation [13]. In our previous work [4], we investigated the potential of deep processing for context management. In particular, a semantic rich discourse representation was motivated, which provides a basis for the work reported here.

There has been tremendous amount of work on discourse modeling in the area of natural language processing. The discourse research mainly addresses two important questions: 1) what information is to be captured from the discourse; and 2) how such information can be used for language interpretation and generation. Many theories have been developed for both texts (e.g., Hobbs theory [14] and Rhetorical Structure Theory [20]) and dialogues (e.g., Grosz and Sidner's conversation theory [10] and Discourse Representation Theory [18]). In this paper, our method is based on Centering Theory [9], a theory that relates the salience of entities in an utterance with the form of linguistic expression and the local coherence of discourse.

3. DISCOURSE PROCESSING

As shown in example (1), a sequence of context questions resembles a traditional coherent text discourse with similar linguistic devices such as reference and ellipsis. Therefore, approaches to model text discourse can be applied here. Specifically, we propose an entity-based discourse coherence structure to represent the discourse of questions. This structure consists of two levels of linguistic representations of the semantic relations between context questions. Pronoun reference is one of the linguistic devices that help form cohesion relations [12], a lower level mechanism to relate questions via lexical ties (ellipsis and repetition are other examples). One level above the cohesion relations, the topicality relations intend to identify the topic-driven semantic relations between adjacent questions. As shown in many examples of context questions (e.g., example (1)), both of these two levels are important given the fact that some questions have pronouns while others do not. Therefore, in this paper, we present a linguistically driven approach that aims to tie these two levels together based on Centering Theory [9]. Given a context question, our approach examines the discourse of questions that lead up to the current question and identifies key entities in the discourse to help form query terms for the current question. Next we first give a brief introduction to Centering Theory and then describe our models in detail.

3.1 Centering Theory

3.1.1 Background

Relying on situation semantics [2], Centering Theory and the centering framework discussed in [9] was developed from three sources: (1) the early centering theory [6][7][8], (2) the theory and focusing algorithm in capturing local discourse coherence [24], and (3) the relationship between the computational inference load and change of focusing state [16] [15]. As a computational model for discourse interpretation, Centering Theory aims to achieve the goal of identifying the mechanism of how a discourse maintains its local coherence (within a discourse segment) using various referring expressions. Centers or discourse entities are explicitly defined to capture the local coherence mechanism that explains how an utterance links to its preceding and succeeding discourse.

Basic framework of center management in [9] has two rules and a partial ordering on the forward-looking centers. The *forward looking centers* C_f (U_n) is defined as an ordered entity list corresponding to entities mentioned in utterance U_n. They are a set of entities that the succeeding discourse may be linked to. The *backward looking center* C_b (U_{n+1}) is defined as the most highly ranked entity of C_f (U_n) mentioned in the succeeding utterance U_{n+1}. The most highly ranked entity of C_f (U_n) is later defined by other researchers as C_p. The term *backward-looking center* and *forward-looking center* correspond to Sidner's discourse focus and potential foci [24].

Table 1. Extended transition states (adapted from [3])

	$C_b(U_i) = C_b(U_{i-1})$	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	Continue	Smooth-Shift
$C_b(U_i) \neq C_p(U_i)$	Retain	Rough-Shift

In Centering Theory, three types of transition relations are defined across two adjacent utterances: *continuation*, *retaining* and *shifting*. Later work [3] extended *shifting* to smooth shifting and rough shifting. Two criteria are used to determine the kind of transition: whether $C_b(U_{n+1})$ is the same as $C_b(U_n)$; whether $C_b(U_{n+1})$ is the most highly ranked entity of $C_f(U_{n+1})$ as shown in Table 1. Degree of coherence thus will be reflected assuming that two utterances are more coherent if they share the same C_b and least coherent if neither they share the same C_b nor the C_b coincides with C_p. This characteristic has been used in measuring coherence in some applications [21].

Based on the centers and transitions, there are two rules in Centering Theory: (1) If any element of $C_f(U_n)$ is realized by a pronoun in U_{n+1} then the $C_b(U_{n+1})$ must be realized by a pronoun also; (2) Sequences of *continuation* are preferred over sequences of *retaining*; and sequences of *retaining* are to be preferred over sequences of *shifting*. These rules can be applied to resolve references and determine the coherence of the discourse. Details on Centering Theory can be found in [9].

Following the entity-based assumption in Centering Theory, how to rank the entities has been discussed extensively in the literature. The ranking scheme based on grammatical relations is most widely implemented with different variations. For example, one ranking scheme indicates that an entity in a subject position is ranked higher than an entity in an object position, which is ranked higher than entities in other positions (i.e., *subject>object(s)>other*) [9]. In this paper, we adopt a more detailed ranking hierarchy proposed in [3] as follows:

subject > existential predicate nominal[1] *> object > indirect object > demarcated adverbial PP*[2]

3.1.2 Three Discourse Models based on Centering Theory for Query Expansion

In a sequence of context questions, each individual question may ask for some partial information related to an overall goal. One important feature of Centering Theory that coincides with that of a question sequence is that Centering Theory reflects dynamics between centers within a discourse segment. This is the motivation of using Centering Theory as our theoretical framework. In particular, we have developed three models for processing contextual questions. Given a question in a discourse, the first model forms query terms by resolving the pronouns and possessives (we name it the *reference model* in this paper) in the question. The second model incorporates the forward-looking centers from an adjacent preceding question with terms from the current question for query expansion (i.e., the *forward model*). The third model applies discourse transitions to selectively incorporate entities from the discourse for query expansion (i.e., the *transition model*).

Reference Model

In the reference model, we use the centering algorithm to resolve pronoun and possesive pronoun references. The algorithm we implemented was based on [3]. There are a few implementation details and modifications worth mentioning here: (1) Instead of only dealing with the adjacent utterance (the strict *local coherence* in [9]), our approach keeps looking back to all the previous questions till an antecedent is found; (2) The linguistic features used include gender, number and animacy; (3) The ranking scheme is based on the same grammatical role hierarchy of the discourse entities as proposed in [3] (mentioned above). At a higher level, this algorithm only assigns those highly ranked antecedents from the discourse to references that can form a more coherent discourse (as indicated by the transitions in Table 1). The detail of the algorithm is reviewed in [17]. Once a pronoun is resolved, its antecedent is used in the query formation for the current question.

Forward Model

In the forward model, query terms for a current question are formed by incorporating forward-looking centers $C_f(U_n)$ from its adjacent preceding question. Note that the forward-looking centers have already been resolved by the reference resolution algorithm , so this model is one step further from the reference model. The motivation for the forward model is based on our assumption that a question discourse is coherent. The forwarding centers from the previous adjacent question form the very *local entity context* for the current question.

Transition Model

Instead of incorporating forward looking centers from its adjacent preceding question as in the forward model, the transition model takes even one step further by selectively incorporating entities from the discourse based on discourse transitions. Centering Theory is used in this model to identify transitions.

As described earlier, the center movement from one utterance to the next implies the degree of discourse coherence, which is modeled by four different transitions: *continue*, *retain*, *smooth-shift* and *rough-shift*. The first two transitions mainly correspond to the situation where the user is continuing the topic and/or the focus from the preceding utterance; and the last two correspond to a certain type of shift of interest. For questions that involve pronouns, the transitions types are automatically identified by the reference resolution algorithm (see the algorithm in [17]). For questions that do not have pronouns, we use an entity-based algorithm that assumes the highest ranked entity is the centered

Table 2. Transition rules for questions without pronouns but with non-pronominal referring expressions

NP Modifier	NP head	Transition
Same	Same	Continue
Different	Same	Retain
Same	Different	Smooth-shift
Different	Different	Rough-shift

[1] A noun phrase that is used as a predicate in an existential sentence (*e.g. There is **a cat** in the house.*)

[2] A noun phrase that is used in an adverbial prepositional phrase separated from the main clause (*e.g. In **the parking lot**, there is an Acura.*)

Table 3. Query expansion strategies based on transition type

Transition	Strategy
Continue	Add the highest ranked proper name most recently introduced from the discourse
Retain	Inherit and then update (if necessary) the constraints from the discourse. Constraints are currently location and time.
Shift	Add the forwarding centers from the previous adjacent to the current question

entity or most accessible in terms of interpretation and understanding. We use the same ranking scheme as in the reference model to assign a rank to each entity. We then compare the highest ranked entities from the adjacent question pair and assign transition types according to Table 2.

More specifically, different transitions are determined based on the syntactic information of a noun phrase (NP) that realizes the C_p. A real world object or an entity can serve as a *center* depending on the NP that realizes it. NPs, especially referring expressions including non-pronominal definite referring expressions and pronominal expressions are the linguistic elements that are discussed initially within the centering framework [8]. Semantically the realization relation for the definite noun phrases may hold in three cases: (1) referentially as to denote an object; (2) attributively as to contribute to the semantic interpretation related to the descriptive content of the expressions; and (3) as the pragmatic reference that is essentially a "speaker's reference". The first two aspects motivate our approach to identify transitions based on NP expressions, in particular, the definite noun phrases.

Intuitively definite noun phrases that share the same NP head and modifier often refer to the same center, which results in a continuation according to centering. Similarly, attention will be retained if two similar entities referred to in two utterances have corresponding NP expressions that share the same NP head but different modifiers. NPs that have same modifier but different head often refer to different entities that share the same descriptive properties. In this case attention is more shifted from the retention case, less from the rough shift where attention on the properties of the entity as well as the entity itself has been shifted. Table 2 shows the four rules that are used to identify different types of transitions. A fifth transition *other* will be assigned to a question pair if none of the four rules can be applied, for example, questions don't have non-pronominal referring expressions. Once different types of transitions are identified, the next step is to apply different strategies to selectively incorporate entities from the discourse for query expansion. To this end, we have currently simplified the process by combining *smooth-shift*, *rough-shift*, and *other* together to a general type *shift*. The specific strategies for each transition type are shown in Table 3.

The strategy for the *continue* transition is motivated by the following two reasons. First, as pointed out in [9] there are cases where "the full definite noun phrases that realize the centers do more than just refer." Similarly we conjecture that the highest ranked proper name in a question sequence carries more information than just for referring. In other words, we believe that given questions that involve pronouns, a highest ranked proper name can provide adequate context if that proper name is not the antecedent of the pronoun and its status is not overwritten by the new information from the current question. Second, as described in [11] on topic status and in [1] on analysis of definite noun phrases, proper names should be given discourse prominence as an important definite noun phrase type. Since currently we do not resolve definite descriptions this strategy partially addresses the importance of definiteness status of other types of definite noun phrases besides pronouns and possessives.

(2) Q1: Where is Hawaii located?
Q2: What is the state fish?
Q3: Is it endangered?

In Example (2), we identify the transition between (2Q2) and (2Q3) as *continue* because *it* in (2Q3) and *the state fish* in (2Q2) refer to the same entity (i.e., the state fish) and this entity is also the C_p of (2Q3). According to the strategy for *continue*, when processing (2Q3), in addition to the query term *the state fish* (which is the antecedent for the pronoun *it* in (2Q3)), the proper name *Hawaii* from (2Q1) will also be inherited.

For the transition type *retain*, intuitively we believe if two questions are on similar but not the same entities (e.g., *the first debate* and *the second debate*), they should share a similar constraint environment (such as time, location[3], etc.). That particular constraint from a preceding question still applies to a current question unless its value is explicitly revised in the current question. The strategy for the *retain* transition was designed based on this intuition.

(3) Q1: Where was the 2nd presidential debate held in 2004?
Q2: Where was the 3rd debate held?

In Example (3), the transition between (3Q1) and (3Q2) is identified as *retain* because according to Table 2, expressions realizing C_p(3Q1) and C_p(3Q2), that is, *the 2nd president debate* and *the 3rd debate* share the same NP head but different modifiers. The strategy for *retain* will allow (3Q2) to inherit its time constraint *2004* from (3Q1).

For the transition type *shift*, currently we adopt the strategy in the forward model by incorporating forward-looking centers from the preceding question. Although the shift transition reflects the least local coherence between utterances, the preceding forward-looking centers are still important in terms of offering the local context information.

(4) Q1: When did Vesuvius destroy Pompeii the first time?
Q2: What civilization ruled at that time?

In Example (4), the transition between (4Q1) and (4Q2) is identified as rough shift according to Table 3 because NPs realizing C_p(4Q1) (i.e., *Vesuvius*) and C_p(4Q2) (i.e., *civilization*) neither share the same head nor the same modifiers. Following the strategy for the shift transition the resulting query terms inherit the forward-looking centers from a preceding question. In this case, query terms *Vesuvius* and *Pompeii* will be added to (4Q2) for document retrieval. Note that all the strategies described here are based on some linguistic observations. Other strategies can be experimented with, in the future.

4. DATA COLLECTION AND ANALYSIS

To support our investigation, we initiated a data collection effort through user studies. We designed the following four topics and

[3] We use simple regular expressions to identify constraints such as location and time.

Table 4. Characteristics comparison between our data and TREC 2004 data (including only factoid questions)

	Debate	Hawaii	Pompeii	Tom Cruise	Our data (overall)	TREC data
Number of topics	1	1	1	1	4	65
Number of question sets	22	22	22	21	87	65
Total number of questions	132	131	134	125	522	230
Type of topics	Event	Entity	Event /entity	Entity	Event /entity	Entity
Average question length	7.4	7.5	7.3	7.0	7.3	7.2
Percentage of context questions with pronouns	14.5%	26.6%	25.0%	81.7%	36.3%	73.9%
Percentage of questions where pronouns refer to topics	56.3%	60.7%	25.0%	73.3%	61.1%	96.0%
Number of Antecedent-in-previous/current question	12	19	20	79	130	126
Total Number of transitions	110	109	112	104	435	165
Number of *continue* transitions	21	19	26	69	135 (30%)	105 (64%)
Number of *retain* transitions	42	31	27	18	118 (27%)	30 (18%)
Number of *shift* transitions	47	59	59	17	182 (43%)	30 (18%)

prepared a set of documents which contain relevant facts about each of these four topics: (1) the presidential debate in 2004; (2) Hawaii; (3) the city of Pompeii; and (4) Tom Cruise. In total, 22 subjects participated in our study. Each subject was asked to put him/herself in a position to acquire information about these topics. And they were asked to specify their information need and provide a sequence of questions (no less than 6 questions) to address that need. As a result of this effort, we collected 87 sets (i.e., sequences of questions) with a total of 522 questions.

Specifically, we emphasized on the following issues during the data collection:

- The answer to each question should come from a different document to enforce the use of the context for the subsequent questions. We feel this design is closer to a natural scenario. This is because if some information has already been shown in the surroundings of the answer to a previous question, users may not even need to ask questions about that information. Users tend to ask questions about facts that he/she has not seen during the information seeking session.
- Each sequence of questions should be coherent in the sense that they should serve a certain information goal. We asked users to explicitly specify their information goals while they provided a sequence of questions.
- Since our goal is to investigate discourse processing for coherent question answering, we are specifically interested in concise questions that depend on the discourse. Therefore we asked users to provide questions that are as natural and concise as possible.

This methodology of collecting context questions is motivated by TREC evaluation where sequences of context questions were pre-defined by NIST staff. To extend context questions to interactive question answering, we are currently collecting data from an online interactive environment.

In addition to our data, we also tested our models on TREC 2004 data. The following is an example taken from TREC 2004.

(5) Q1: What film introduced Jar Jar Binks?
Q2: What actor is used as his voice?
Q3: To what alien race does he belong?

In TREC 2004, each set of questions comes with a predefined target (e.g., *Jar Jar Binks* for Example (5)). Since TREC data was also designed to test system capability of answering list and definition questions, which are not the focus of our work, we omitted those questions in our evaluation. In this paper, we only focus on the 230 factoid context questions in our analysis and evaluation.

Table 4 shows a comparison of two datasets: our data and TREC data. First of all, TREC data consists of 65 topics (i.e., targets) and each topic has one set of questions. In contrast, our data consists of only four topics where each topic comes with more than 20 sets of questions from different users. Question sets from multiple users on a same topic will allow us to test the generality of our discourse processing strategies across different users.

Unlike TREC data where each topic is about a single entity such as *the Black Panthers organization*, our data covers both event and entity. For example, the topic on the "presidential debate" is about an event, which can potentially relate to the facts (e.g., when, what, etc), the cause, and the consequence of the event. This variation will allow us to study the potential distinctions in processing different types of topics (in terms of event or entity) systematically.

From Table 4 we can see that, the surface characteristics across our data and TREC factoid questions are very similar in terms of the question length. However, TREC data has a higher percentage of pronoun usage in the context questions. In our data, only questions with the topic of "Tom Cruise" have the high percentage of pronouns, while the other topics have significantly lower percentage of pronouns. This indicates the potential different impact of pronoun resolution in different datasets.

Furthermore, the majority of the pronouns in the TREC data (96%) refer to the topic/target which has been provided to each set. Therefore, incorporating target terms for query expansion will have a same effect as a model that resolves pronouns. Each context question will then become an isolated factoid question and additional discourse processing may not be necessary. In our

Table 5. Overall performance of different models on document retrieval for our data and TREC data

Topic	Baseline	Ref. Auto	Ref. Key	Ref. % diff	For. Auto	For. Key	For. % diff	Trans. Auto	Trans. Key	Trans. % diff
Debate	0.044	0.043	0.048	11.6%	0.048	0.048	0%	0.042	0.042	0%
Hawaii	0.051	0.067	0.085	26.9%	0.080	0.085	6.2%	0.100	0.110	10.0%
Pompeii	0.132	0.118	0.149	27.3%	0.156	0.163	4.5%	0.185	0.186	0.5%
Tom Cruise	0.100	0.185	0.227	22.7%	0.220	0.227	3.2%	0.228	0.228	0%
Overall	0.080	0.102	0.115	12.7%	0.125	0.126	0.8%	0.138	0.140	1.4%
TREC	0.158	0.221	0.265	20.0%	0.283	0.288	1.7%	0.289	0.296	2.4%

data, the percentage of pronouns that refer to the topic is significant lower, which indicates a higher demand on discourse processing.

In term of transitions, the majority of TREC data has the *continuation* transition (64%), while our data exhibits more diverse behavior. By studying these different characteristics of the two datasets, we hope to learn their implications of specific strategies from our empirical evaluation.

5. EVALUATION

To evaluate our linguistically driven discourse processing, we conducted a series of experiments to compare the performance of the three models on both our data and TREC data. For our data, we incorporated documents with answers to each of the collected questions to the AQUAINT CD2 collection and the evaluation was done based on the updated CD2 collection (with a size about 1.8G). For the TREC questions, we used the entire AQUAINT collection (about 3G). In all the experiments, we used Lemur retrieval engine[4] for document retrieval. Since the first occurrence of a correct answer is important, we used Mean Reciprocal Ranking (MRR) as our first measurement. MRR is defined as:

$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$, where $rank_i$ is the rank of a retrieved document which provides the first correct answer for the ith question and N is the total number of questions evaluated.

Many observations can be made based on our experimental results. Due to space limitation we report the results that particularly address the following issues:

- How are different models based on Centering Theory compared to each other in terms of document retrieval performance? Will different models affect different types of questions? Are there any correlations between the characteristics of questions and the effectiveness of potential strategies?
- How sensitive is each model's response to performance limitation of automated discourse processing? In other words, what is the capability of each model in compensating the potential mistakes caused by machine processing (e.g., incorrectly resolving some pronouns)?

Note that our focus is not on document retrieval, but rather on the impact of the discourse processing on document retrieval. Therefore, the evaluation reported in this paper is based on the subsequent questions (435 in our data and 165 from TREC data) which exclude every first question in each set since processing the first question does not use any discourse information.

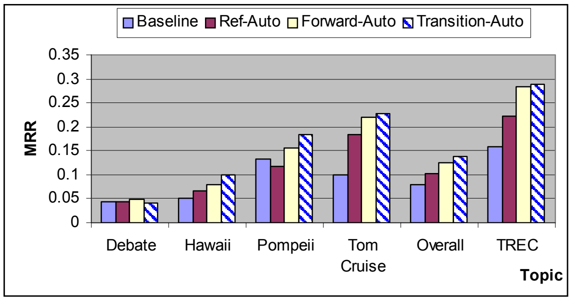


Figure 1. Overall comparison of four models based on automated processing

Table 5 shows the overall performance of all three models on the two datasets compared to a baseline model in terms of MRR. The baseline model simply incorporates the preceding question to the current question to form a query without any pronoun resolution. The motivation for this baseline strategy is that since most antecedents of pronoun references have occurred in the preceding questions (see Table 4, especially the TREC data) the preceding question can simply provide a context for the current question.

Since all three models based on Centering Theory rely on pronoun resolution, the performance of automated pronoun resolution algorithm directly impacts the final performance of document retrieval. Therefore in Table 5, along with the performance resulting from automated processing (i.e., marked with "auto" in the column title), we also provide retrieval results for each model based on manually annotated antecedents (with "key" in the column title), as well as the performance difference between the two (i.e., the % difference column).

To better present the results, Figure 1 shows a detailed comparison between four models as a result of automated processing. As shown in Figure 1, except for the *Debate* data the incremental increase in the complexity of discourse processing (e.g., from reference model, to forward model, to transition model) improves the overall performance. For the *Debate* data, different models performed comparably the same. In other words, any type of discourse processing has not shown a significant effect compared to the baseline model. One of the reasons is that, the sets of questions collected for *Debate* are somewhat different from the rest of the topics in terms of the content of the questions. The *Debate* data relates to an event while the rest of the data sets

[4] http://www-2.cs.cmu.edu/~lemur/. The Lemur toolkit supports indexing of large-scale text databases and the implementation of retrieval systems based on as a variety of retrieval models.

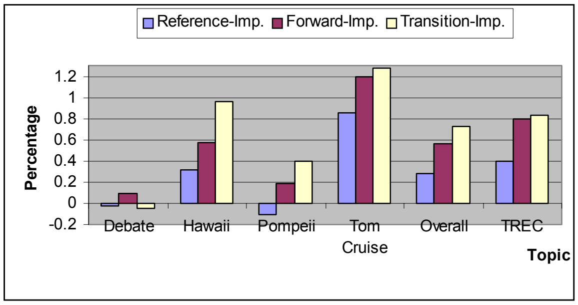


Figure 2. Our models based on the Centering Theory (automated processing) compared to the baseline

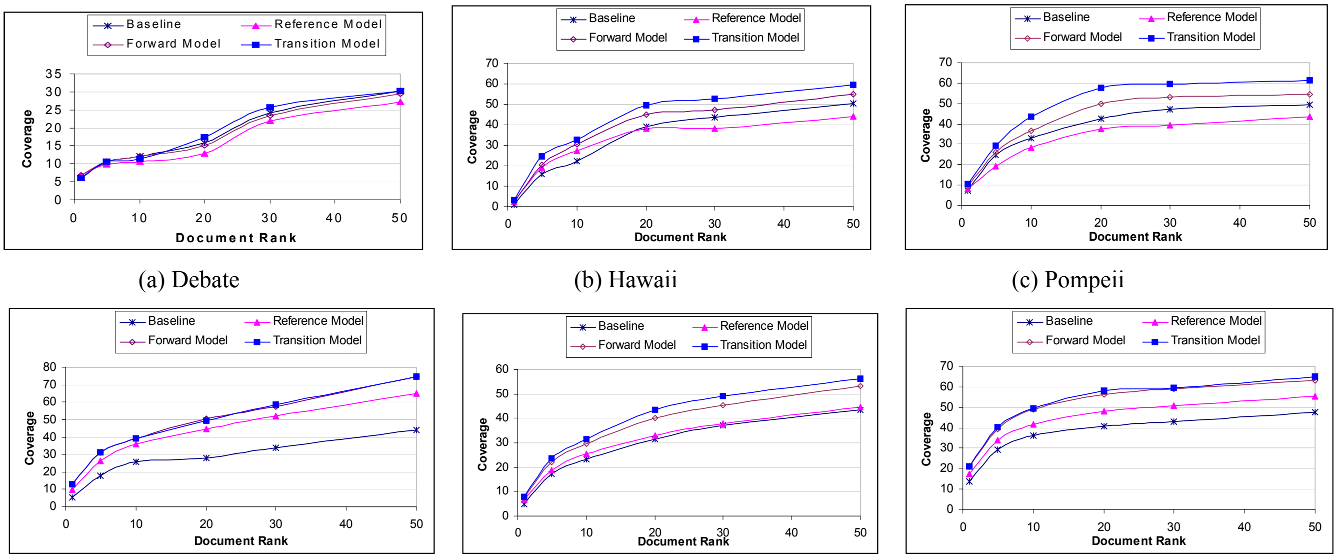
Table 6. Document retrieval performance based on the transition model and passage retrieval performance from the University of Sheffield on TREC data

Document Rank	Transition Model	Sheffield's Lucene[5]
1	20.87	12.17
5	40.43	32.17
10	49.57	39.56
20	58.26	47.39
30	59.57	51.30
50	64.78	55.65

relate to entities such as *place* or *person*. Since Centering Theory is mainly based on the transitions between discourse entities, it could be the case that our models would work better for entity related questions than event related questions. An event may involve more complicated transitions such as consequence, cause, and reason; other models utilizing relation-based coherence theories such as Rhetorical Structure Theory could be a potential approach. However, more in-depth analysis is necessary in order to reach a better understanding of event related questions and their implications on the automated discourse processing targeted to these questions.

To illustrate the contribution of each incremental processing, Figure 2 shows the percentage of improvement compared to the baseline model. First of all, it is possible that automated processing of pronoun resolution could result in wrong antecedents, therefore the reference model based on automated processing might hurt the retrieval performance compared to the baseline model. This is evident for the *Debate* and *Pompeii* data. The *Pompeii* data is a mixture of event and entity topic (e.g., it involves the event of volcano eruption) so the effect from our forward and transition models is also limited compared to the baseline. Furthermore, the additional contribution of the transition model is relatively less for the *Tom Cruise* data and the TREC data than that for the *Hawaii* and *Pompeii* data. A possible explanation is that both *Tom Cruise* and the TREC data have higher percentage of pronouns (see Table 4). The specific transitions identified between two adjacent questions largely depend on the resolution of those pronouns. Therefore, the reference model has already handled the functions provided by the transition model. However, in the *Hawaii* and *Pompeii* data, the occurrences of pronouns are relatively lower. The transition model can particularly accommodate entities that are not realized as pronouns such as definite descriptions (e.g., through the *continue* transition as discussed earlier).

From our experimental results, it is interesting to point out that the sensitivity of each model varies in response to the accuracy of automated discourse processing. From Table 5, in the reference model, a perfect pronoun resolution makes a big difference compared to an imperfect automated pronoun resolution (the performance difference is between 12-27% as shown in the "Ref % diff" column). However, the performance difference as a result of the capability of resolving pronouns becomes diminished in the forward and the transition models. This result indicates that by inheriting more context from the preceding questions as in the



(a) Debate

(b) Hawaii

(c) Pompeii

(d) Tom Cruise

(e) Our overall data

(f) TREC data

Figure 3. Coverage comparison between four models based on automated processing

forward and transition model, it can potentially compensate the inaccuracy in automated pronoun resolution.

To further examine the three models on document retrieval, we also evaluated document retrieval performance in terms of *coverage*. While *MRR* rewards the method that improves the ranking of the correct answers, *coverage* rewards methods that introduce the correct answer in the retrieved results. More specifically, coverage is defined as the percentage of questions for which a text returned at or below the given rank contains an answer [5]. Figure 3 shows the coverage measurement for each model on different topics. Overall speaking, we see that transition model is consistently better than the other models. The entity topic resemblance between *Tom Cruise* data and TREC data again results in similar performance (i.e., they both have large percentage of pronouns referring to the topic itself).

Given our experimental results described above, a natural question is how the retrieval performance from our models is compared to other retrieval performance. It is hard to achieve this kind of comparison because TREC 2004 did not provide document retrieval performance based on the context questions. The closest we can find is the "coverage" based on passage retrieval for TREC 2004 factoid questions provided by the University of Sheffield [5]. Table 6 shows our retrieval performance (from the transition model) and the Sheffield's retrieval performance (using the Lucene retrieval engine) in terms of coverage based on all 230 factoid questions. Note that since our system was evaluated on document retrieval and Sheffield's system was on passage retrieval, this is not a direct comparison. We list them together simply to have some sense about whether our performance is on the right track. Resources and initiatives to facilitate a direct comparison are in great need in order to enhance understanding on discourse processing for document retrieval.

6. CONCLUSION

To support coherent information seeking, intelligent QA interfaces will inevitably require techniques for context question answering. To address this need, this paper presents three models that process a sequence of questions. Since these models are based on Centering Theory that focuses on discourse entities, the state-of-the-art NLP techniques are sufficient for discourse processing. Our empirical results have shown the advantage of the forward model and the transition model compared to the model that only resolves pronoun references.

This paper presents our initial investigation on the role of discourse processing for context questions. Our future work will focus on coherent question answering in an online interactive setting and examine the need and implication of discourse processing for context questions.

ACKNOWLEDGEMENT

This work was supported in part by IIS-0347548 from the National Science Foundation. The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] Abbott, B. Definiteness and indefiniteness. In Laurence R. Horn and Gregory Ward (eds.), *Handbook of Pragmatics*. Oxford, Blackwell, 2004.

[2] Barwise, J., and Perry, J. *Situations and Attitudes.* MIT Press. 1983.

[3] Brennan, S. E., Friedman, M. W., and Pollard, C. A Centering approach to pronouns. In *ACL'87*, Stanford, CA, ACL, 1987, 155-162.

[4] Chai, J., and Jin, R. Discourse status for context questions. In *Proceedings of HLT-NAACL 2004 Workshop on Pragmatics in Question Answering* (Boston, MA. May 3-7, 2004) ACL, 2004, 23-30.

[5] Gaizauskas, R., Greenwood, M.A., Hepple, M., Roberts, I., and Saggion, H. The University of Sheffield's TREC 2004 Q&A experiments. In *Proceedings of The Thirteenth Text Retrieval Conference(TREC-2004)*, 2004.

[6] Grosz, B. *The Representation and Use of Focus in Dialogue Understanding.* Technical Report 151, SRI International, 333 Ravenswood Ave., Menlo Park, CA, 94025, 1977.

[7] Grosz, B. Focusing and description in natural language dialogue. In A. Joshi, B. Webber, and I. Sag (eds.), *Elements of Discourse Understanding.* Cambridge, England, Cambridge University Press, 1981, 85-105.

[8] Grosz, B. J., Joshi, A.K., and Weinstein, S. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA, 1983, 44-50.

[9] Grosz, B. J., Joshi, A. K., and Weinstein, S. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21, 2 (1995), 203-225.

[10] Grosz, B. J., and Sidner, C. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12, 3, (1986), 175-204.

[11] Gundel, J. *The Role of Topic and Comment in Linguistic Theory*. Distributed by Indiana University Linguistics Club. Bloomington, Indiana, 1976.

[12] Halliday, M. A. K., and Hasan, R. *Cohesion in English.* London: Longman, 1976.

[13] Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M., Mihalcea, R., Girju, R., Rus, V., Lacatusu, F., Morarescu, P., and Bunescu, R. Answering complex, list and context questions with LCC's Question-Answering Server. In TREC-10 Question-Answering. In E. M. Voorhees and D. K. Harman (eds.), *The Tenth Text Retrieval Conference (TREC 2001)*. NIST Special Publication 500-250. Gaithersburg, MD, 355-361.

[14] Hobbs, J. R. *On the Coherence and Structure of Discourse*. Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.

[15] Joshi, Aravind K., and Weinstein, S. Control of inference: role of some aspects of discourse structure- centering. In *Proceedings of international joint conference on artificial intelligence*, 1981, 385-387.

[16] Joshi, Aravind K., and Kuhn, S. Centered logic: the role of entity centered sentence representation in natural language inferencing. In *Proceedings of international joint conference on artificial intelligence* (Tokyo, Japan, August, 1979), 435-439.

[17] Jurafsky, D., and Martin, J. H. *Speech and Language Processing*. Prentice Hall, NJ, 2000.

[18] Kamp, H., and Reyle, U. *From Discourse to Logic.* Kluwer, Dordrecht, 1993.

[19] Liddy, E.D., Diekema, A.R., and Yilmazel, O. Context-based question answering evaluation. In *Proceedings of the 27th Annual ACM-SIGIR Conference*. Sheffield, England, 2004.

[20] Mann, W. C., and Thompson, S. A. *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.

[21] Milsakaki, E., and Kukich, K. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10, 1, (2004), 25-55.

[22] Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A., and Bolohan, O. LCC tools for question answering. In *Proceedings of the 11th Text Retrieval Conference (TREC-2002)*, (Gaithersburg, MD, November, 2002).

[23] Roberts, I. and Gaizauskas, R. Evaluating passage retrieval approaches for question answering. In *Proceedings of the 26th European conference on information retrieval*, 2004.

[24] Sidner, C. L. *Towards A Computational Theory of Definite anaphora Comprehension in English Discourse.* Ph.D. thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Technical Report 537. June, 1979.

[25] Voorhees, E. Overview of TREC 2001 question answering track. In *Proceedings of TREC.* (Gaithersburg, MD, November 13-16, 20.