Discourse Processing for Context Question Answering Based on Linguistic Knowledge

Mingyu Sun[a] , Joyce Y. Chai [b]

[a]Department of Linguistics
Michigan State University
East Lansing, MI 48824
sunmingy@msu.edu

[b]Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

Motivated by the recent effort on scenario-based context question answering (QA), this paper investigates the role of discourse processing and its implication on query expansion for a sequence of questions. Our view is that a question sequence is not random, but rather follows a coherent manner to serve some information goals. Therefore, this sequence of questions can be considered as a mini discourse with some characteristics of discourse cohesion. Understanding such a discourse will help QA systems better interpret questions and retrieve answers. Thus, we examine three models driven by Centering Theory for discourse processing: a reference model that resolves pronoun references for each question, a forward model that makes use of the forward looking centers from previous questions, and a transition model that takes into account the transition state between adjacent questions. 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. This paper provides a systematic evaluation of these models and discusses their potentials and limitations in processing coherent context questions.

1. INTRODUCTION

Question answering (QA) systems take users' natural language questions and automatically locate answers from large collections of documents. In the past few years, automated QA technologies have advanced tremendously, partly motivated by a series of evaluations conducted at the Text REtrieval Conference (TREC) [37]. To better facilitate user information needs, recent trends in QA research have shifted towards complex, context-based, and interactive question answering [37][35][20].

The National Institute of Standards and Technology (NIST) [1] initiated a special task on context question answering in TREC 10 [37], which later became a regular task in TREC 2004 [38] and 2005. The motivation is that users tend to ask a sequence of related questions rather than isolated single questions to satisfy their information needs. For example, suppose a user is interested in finding out 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.

[1] http://www.nist.gov/

We treat the question sequence as a mini discourse in which 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, which models local coherence of discourse [14]. Centering Theory describes how different linguistic devices (e.g., pronouns) are used to maintain the local coherence of a discourse and minimize the hearer's inference load. In coherent question answering, users also tend to maintain the coherence of discourse. This is evident by example (1) as well as the data provided in the TREC 2004 [38] (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.

In our current investigation, rather than a complete end-to-end study, this paper focuses on discourse processing on questions 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 the 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 the three models for discourse processing, and finally present our empirical evaluation and discuss the potentials and limitations of these models.

2. RELATED WORK

The term context in context question answering can refer to different context such as user context [28] and discourse context [37]. 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 [37]. The context task was designed to investigate the system capability to track context through a series of questions. However, as described in ([37], p50), 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.

Since 2004, scenario-based context QA has become a regular task at TREC evaluation [2]. Most work related to this task in TREC 2005 has focused on three aspects of question processing.

[2] http://trec.nist.gov/pubs/trec14/t14_proceedings.html

The first aspect is on question type analysis and categorization, which identifies the target types for context questions (such as whether a question is to ask "person", "number" or "location") [9][42]. This is similar to the processing for the isolated factoid questions, where the identified question type can help pin-point the expected answer strings. The second aspect emphasizes the processing of the words in the questions [9][31][4] and [42]. Parsing tools (which help to find verbal predicates), POS tagging tools, name entity recognizers (which tag name entities into categories), statistical analysis (unigram, bigram, and n-grams for question words), and knowledge bases (such as WordNet synsets [3], which provides synonyms for a particular word) are utilized to expand the queries. The third is to make use of the target (topic) provided by the TREC data sets for anaphora resolution such as [2]. However, except for [3] that applies Discourse Representation Structure (DRS) [29] in question processing, approaches based on discourse analysis have been limited. Therefore, the goal of our study is to investigate the role of discourse processing for context questions.

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 a 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 [19]. In our previous work [8], 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 a 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 represented for language interpretation and generation. Many theories have been developed for both texts (e.g., Hobbs theory [21] and Rhetorical Structure Theory [29]) and dialogues (e.g., Grosz and Sidner's conversation theory [16] and Discourse Representation Theory [26]). In this paper, our method is based on Centering Theory [14], a theory that relates the salience of entities in an utterance with the form of linguistic expressions 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 [18], 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 knowledge driven approach that aims to tie these two levels together based on Centering Theory [14]. 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 give a brief introduction to Centering Theory and then describe our models in detail.

[3] http://wordnet.princeton.edu/

Table 1
Extended transition states (Adapted from Brennan et al.([7], p157))

	$\mathrm{C}_b(\mathrm{U}_{n+1}) = \mathrm{C}_b\ (\mathrm{U}_n)$ or undefined $\mathrm{C}_b\ (\mathrm{U}_n)$	$\mathrm{C}_b\ (\mathrm{U}_{n+1}) \neq \mathrm{C}_b\ (\mathrm{U}_n)$
$\mathrm{C}_b\ (\mathrm{U}_{n+1}) = \mathrm{C}_p\ (\mathrm{U}_{n+1})$	Continue	Smooth shift
$\mathrm{C}_b\ (\mathrm{U}_{n+1}) \neq \mathrm{C}_p\ (\mathrm{U}_{n+1})$	Retain	Rough shift

3.1. Centering Theory

3.1.1. Background

Relying on situation semantics [6], Centering Theory and the centering framework discussed in [14] were developed from three sources: (1) the early centering theory [11][12][13], (2) the focusing theory and algorithm in capturing local discourse coherence [34], and (3) the relationship between the computational inference load and the change of focusing state [22][23]. 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. The *forward looking centers* $C_f(U_n)$ are defined as a set of ordered entities corresponding to entities mentioned in utterance U_n. They are 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 term *preferred center* C_p was introduced to represent the highest-ranked member of the forward looking centers ([7], p155). The term *backward looking center* and *forward looking center* correspond to Sidner's discourse focus and potential foci ([34], p222, p223).

We use the example (2) to help illustrate the notion of different centers in Centering Theory. Note that we intentionally use the definite noun phrases instead of pronouns in the questions to provide better explanation. In reality, many pronouns can be used in the question sequence and these pronouns will be resolved by the *reference model* described in Section 3.1.2.

(2)

Q1: Who is Tom Cruise?

Q2: What movies was Tom Cruise in?

Q3: When did Nicole Kidman marry Tom Cruise?

Q4: What was Nicole Kidman's Broadway debut?

Q5: What was the debut about?

Q6: What role did Nicole Kidman play in the debut?

Q7: Where did Tom Cruise wed Nicole Kidman?

In this example, the forward looking center of (2Q1) is a set with only one element Tom Cruise, the only entity in the question. There is no backward looking center for (2Q1) since it is the first question in the sequence. The preferred center is Tom Cruise since that is the only element in the set of forward looking centers.

C_f(2Q1):{Tom Cruise}
C_b(2Q1): undefined
C_p(2Q1): Tom Cruise

The forward looking centers of (2Q2) are entities Tom Cruise and movies. The entity Tom Cruise is mentioned again in (2Q2) so it is the backward looking center of (2Q2). The preferred center of (2Q2) is Tom Cruise, because the corresponding expression *Tom Cruise* is in the subject position and therefore is ranked higher than *movies* according to the grammatical role ranking (explained later).

C_f(2Q2):{Tom Cruise, movies}
C_b(2Q2): Tom Cruise
C_p(2Q2): Tom Cruise

In Centering Theory, three types of transition relations are defined across two adjacent utterances: *continuation*, *retaining* and *shifting*. Later work [7] extended shifting to smooth shifting and rough shifting. Two criteria are used to determine the type of transition: (1) whether $C_b(U_{n+1})$ is the same as $C_b(U_n)$; (2) whether $C_b(U_{n+1})$ is the most highly ranked entity of $C_f(U_{n+1})$, that is, the $C_p(U_{n+1})$. Table 1 shows how the transition types are defined. If both (1) and (2) hold, then the two utterances are related by a *continue* transition, which indicates that the speaker has been focusing on an entity and intends to continue talking about it. If (1) holds, but (2) does not, then the two utterances are related by a *retain* transition. In this case, the speaker intends to shift his focus onto a new entity and this entity is realized [4] as a center in a lower-ranked position other than the highest position. If (1) does not hold, then the two utterances are related by *smooth shift* or *rough shift*, depending on whether (2) holds or not. Both shifts occur when the speaker has shifted his focus from one entity to another entity. *Rough shift* is a rather extreme case where the new entity is not realized as the preferred center.

For example in (2), the following transitions can be identified according to Table 1:

- The backward looking center C_b(2Q2) is the same as C_p(2Q2) and C_b(2Q1) is undefined, therefore the transition between (2Q1) and (2Q2) is *continue*.

- The backward looking center C_b(2Q3) is Tom Cruise, which is the same as C_b(2Q2). However the C_p(2Q3) is Nicole Kidman, which is different from C_b(2Q3). The transition between (2Q2) and (2Q3) is therefore *retain*.

- The transition between (2Q4) and (2Q5) is *smooth shift* because the C_b(2Q4) (i.e. Nicole Kidman) is not the same as C_b(2Q5) (i.e. Broadway debut) while the C_b(2Q5) is the same as the C_p(2Q5)(i.e. Broadway debut).

- The transition between (2Q6) and (2Q7) is *rough-shift* because the C_b(2Q6) (i.e. Broadway debut) is different from the C_b(2Q7) (i.e. Nicole Kidman) while the C_b(2Q6) is also different from the C_p(2Q7) (i.e. Tom Cruise).

As shown in the example (2), the degree of coherence can 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 such as [30].

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 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 [14].

There have been various algorithms based on centering framework aiming to fulfill discourse processing tasks, such as [25],[7],[39],[5],[27],[40] and [36], etc. In particular, [7] proposed a centering algorithm to resolve third-person pronouns within the framework of Centering Theory. The two rules mentioned above and three constraints are specifically used in their work. The three constraints are stated as: (1) There is precisely one C_b; (2) Every element of $C_f(U_n)$ must be realized in U_n; (3) $C_b(U_{n+1})$ is the highest-ranked element of $C_f(U_n)$ that is realized in U_{n+1} [7]. 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*) ([14], p214). In this paper, we adopt [7]'s centering algorithm and a more detailed ranking hierarchy extended from ([7], also mentioned in

[4] The *realize* relation is defined as follows ([16], quoted in [41], p4): An utterance U realizes a center c if c is an element of the situation described by U, or c is the semantic interpretation of some subpart of U.

([24], p691) as follows: *subject >existential predicate nominal*[5] *>object >indirect object >demarcated adverbial PP*[6].

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 context questions. Given a question in a discourse, the first model forms query terms by resolving the pronouns (we name it the *reference model* in this paper) in the question. The second model incorporates the forward looking centers from the 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 references. The algorithm we implemented was based on [7]. 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 [14]), 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 [7] (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 ([24], p692). Once a pronoun is resolved, its antecedent is used in the query formation for the current question. For example:

(3)

Q1: When was Tom Cruise born?

Q2: When did he start acting?

The expression *Tom Cruise* will be added to the query terms for (3Q2) because the pronoun *he* in (3Q2) is resolved to the entity *Tom Cruise*.

Forward Model

In the forward model, query terms for a current question are formed by incorporating forward looking centers Cf 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 forward looking centers from the previous adjacent question form the local entity context for the current question. For example:,

(4)

Q1: How is Tom Cruise related to Nicole Kidman?

Q2: What movies was she in?

The expressions corresponding to the forward looking centers (i.e. *Tom Cruise*, *Nicole Kidman*) in (4Q1) are added for query expansion for (4Q2).

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 transitions of centers from one utterance to the next imply the degree of discourse coherence, which is captured by four types: *continue*, *retain*, *smooth shift*, and *rough shift*. The first two transitions mainly correspond to the situation where the user is continuing the

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

[6] 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.)

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 transition types are automatically identified by the reference resolution algorithm because of the transition preference rule in Centering Theory (see the algorithm in ([24], p692)). For questions that do not have pronouns, we use an entity-based algorithm that assumes the highest ranked entity is the centered 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 a transition type 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 [13].

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 NP heads 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* is assigned to a question pair if none of the four rules can be applied, for example, a question pair that does not 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 for the query expansion of the QA question in processing.

The strategy for the *continue* transition is motivated by the following two reasons. First, as pointed out in ([14], p216), there are cases where "the full definite noun phrases that realize the centers do more than just refer." Being part of a discourse, they contribute to the discourse coherence as well. 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 [17] on topic status and proper name's status in the definiteness hierarchy in [1], proper name should be given certain discourse prominence as it is 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.

(5)

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

In example (5), we identify the transition between (5Q2) and (5Q3) as *continue* because *it* in (5Q3) and *the state fish* in (5Q2) refer to the same entity (i.e., the state fish) and this entity is also the C_p of (5Q3). According to the strategy for *continue*, when processing (5Q3), in addition to the query term *the state fish* (corresponding to the antecedent for the pronoun *it* in (5Q3)), the proper name *Hawaii* from (5Q1) 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 en-

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

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.

vironment (such as time, location [7], 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.

(6)

Q1: Where was the 2nd presidential debate held in 2004?

Q2: Where was the 3rd debate held?

In example (6), the transition between (6Q1) and (6Q2) is identified as *retain* because according to Table 2, expressions realizing C_p(6Q1) and C_p(6Q2), that is, *the 2nd president debate* and *the 3rd debate* share the same NP head but different modifiers. The strategy for *retain* will allow (6Q2) to inherit its time constraint *2004* from (6Q1).

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.

(7)

Q1: When did Vesuvius destroy Pompeii the first time?

Q2: What civilization ruled at that time?

In example (7), the transition between (7Q1) and (7Q2) is identified as *rough shift* according to Table 3 because NPs realizing C_p(7Q1) (i.e., *Vesuvius*) and C_p(7Q2) (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 the preceding question. In this case, query terms *Vesuvius* and *Pompeii* will be added to (7Q2) 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 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

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

Pompeii; and (4) Tom Cruise. In total, 22 subjects participated in our study. These subjects were recruited from the undergraduate students who took an Introduction to Artificial Intelligence class at Michigan State University. 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 explicitly pointed out the following issues when they formed their questions:

- 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. Note that the sequences of questions were not specified "interactively", but rather all at once before the user saw any answers from the system. We consider this as *batch* QA, which represents a very practical scenario where a user has a set of questions in mind related to his/her information goal. In addition to our data, we also tested our models on the TREC 2004 data. The following is an example taken from TREC 2004.
(8)
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 (8)). Since the 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 the two data sets: our data and the TREC 2004 data. First of all, the TREC 2004 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 the TREC 2004 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 the TREC 2004 factoid questions are very similar in terms of the question length. However, the TREC 2004 data has a higher percentage of pronoun usage in the context questions. In our data, only questions with the topic *Tom Cruise* have a high percentage of pronouns, while the other topics have significantly lower percentage of pronouns. This variation will allow us to study the potential different impact of pronoun resolution in different data sets.

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

	Debate	Hawaii	Pompeii	TomCruise	Overall	TREC2004
Num. of topics	1	1	1	1	4	65
Num. of question sets	22	22	22	21	87	65
Total number of questions	132	131	134	125	522	230
Type of topics	Event	Entity	E/E*	Entity	E/E	Entity
Average question length	7.4	7.5	7.3	7.0	7.3	7.2
% of ques. with pronouns	14.5%	26.6%	25.0%	81.7%	36.3%	73.9%
% of ques. pron. refer to topics	56.3%	60.7%	25.0%	73.3%	61.1%	96.0%
Total number of transitions	110	109	112	104	435	165
Num. of *continue* transitions	21	19	26	69	135(30%)	105(64%)
Num. of *retain* transitions	42	31	27	18	118(27%)	30(18%)
Num. of *shift* transitions	47	59	59	17	182(43%)	30(18%)

* E/E refers to Event/Entity

Furthermore, the majority of the pronouns within each set in the TREC 2004 data (96%) refer to the topic/target which has been provided to the set. Therefore, incorporating target terms for query expansion will have the 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 data, the percentage of pronouns that refer to the topic is significantly lower, which indicates a higher demand on discourse processing.

In term of transitions, the majority of the TREC 2004 data has the *continuation* transition (64%), while our data exhibits more diverse behavior. By studying these different characteristics of the two data sets, we hope to learn their implications for specific strategies from our empirical evaluation.

5. EVALUATION

We conducted a series of experiments to compare the performance of the three models on both our data and the TREC 2004 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 2004 questions, we used the entire AQUAINT collection (about 3G). In all the experiments, we used the Lemur retrieval engine for document retrieval [8]. 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.

Our evaluation mainly addresses the following issues:

- How are the different models based on Centering Theory compared to each other in terms of document retrieval performance? Will the different models affect different types of questions? Are there any correlations between the characteristics of questions and the effectiveness of the strategies? To answer these questions, we compared the performance of each model on both data sets. We further provided detailed analysis of performance comparison based on different characteristics of questions such as the type of transitions and the pronoun usages.

- How sensitive is each model's response to performance limitation of automated dis-

[8] 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.

course 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)? To answer these questions, the evaluation was performed based on two configurations: 1) automated system where the pronoun resolution and transitions are all automatically identified by the computer system; 2) annotated system where the correct references to pronouns and transitions are annotated.

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 the TREC 2004 data) which exclude every first question in each set since processing the first question does not use any discourse information.

5.1. Overall Evaluation Results

Table 5 shows the overall performance of all three models on the two data sets 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 2004 data), the preceding question can simply provide a context for the current question.

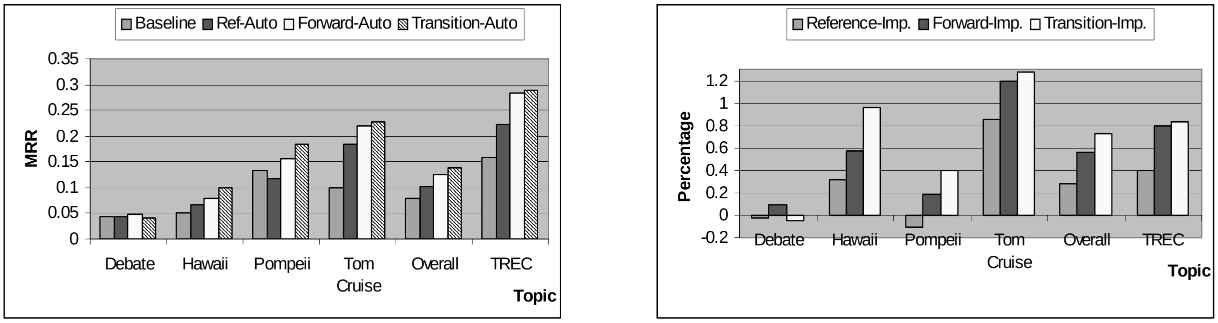
Since all three models based on Centering Theory rely on pronoun resolution, the performance of the automated pronoun resolution algorithm directly impacts the final performance of document retrieval. Therefore in Table 5, along with the performance resulting from automated processing (i.e., marked with “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(a), except for the *Debate* data the incremental increase in the complexity of discourse processing (e.g., from the reference model, to the forward model, to the 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 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 1(b) shows the percentage of improvement compared to the baseline model. First of all, it is possible that the 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 2004 data than that for the *Hawaii* and *Pompeii* data. A possible explanation is that both the *Tom Cruise* and the TREC 2004 data have higher percentage of pronouns (see Table 4). The specific transi-

Table 5
Overall performance of different models on document retrieval for our data and TREC 2004 data

Topic	Baseline	Ref	Ref	Ref	For	For	For.	Trans	Trans	Trans
		Auto	Key	%Diff	Auto	Key	%Diff	Auto	Key	%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 2004	0.158	0.221	0.265	20.0%	0.283	0.288	1.7%	0.289	0.296	2.4%



(a) Overall automated system

(b) Automated system compared to baseline

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

tions 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 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

Table 6
Document retrieval performance based on the transition model and passage retrieval performance from the University of Sheffield on TREC data in terms of the *coverage* measurement

Document Rank	Transition model	Sheffield's Lucene* [10]
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

*http://lucene.apache.org/java/docs/

defined as the percentage of questions for which a text returned at or below the given rank contains an answer ([10], p2). Figure 2 shows the coverage measurement for each model on different topics. Overall, we see that the transition model is consistently better than the other models. The entity topic resemblance between the *Tom Cruise* data and the TREC 2004 data again results in similar performance (i.e., they both have a 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 [10]. 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.

To further understand the effectiveness of these models on questions with different characteristics, we isolated two dimensions: 1) questions with different transition types and 2) questions with and without pronouns, and conducted a detailed performance analysis along these two dimensions. We report the results next.

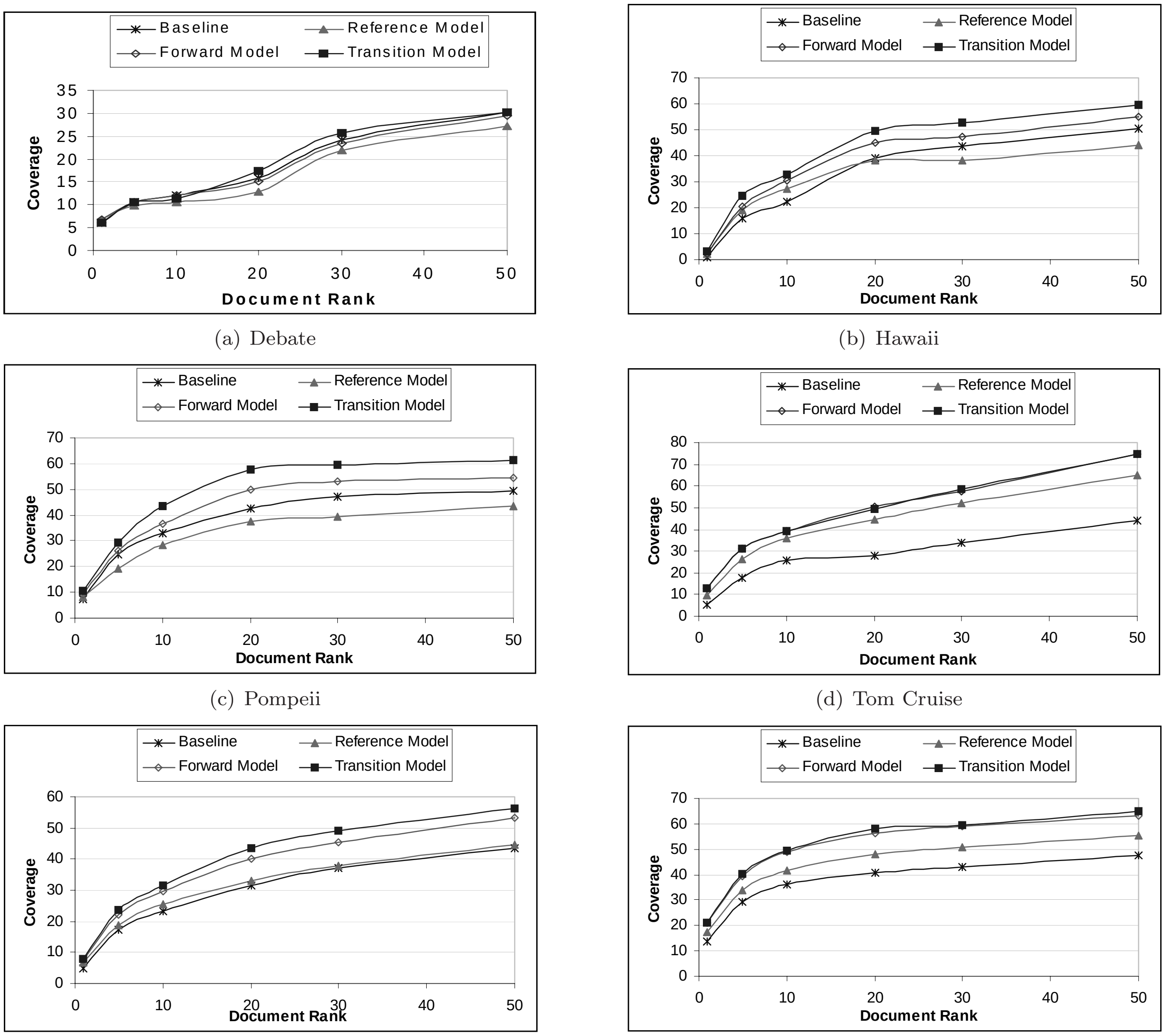
5.2. Evaluation and Analysis based on Transitions

In this section, we discuss the role of three models on question pairs with the transition type *continue*, *retain*, and *shift*, respectively.

5.2.1. Continue transition

Figure 3 shows the overall comparison of the three models on the question pairs with the transition type *continue*, with Figure 3(a) for the automated system and Figure 3(b) for the annotated system. In general, for *continue* pairs, the transition model works the best, then the forward model and the reference model, and the baseline is the worst. This implies that the transition model would work the best for the most coherent discourses, which, according to Centering Theory, have a higher percentage of *continue* pairs.

Figure 3(a) shows that the transition model performs consistently better than the forward model. This result indicates that the strategies used in the transition model for *continue* questions are adequate to provide appropriate context. The transition model provides more information than the forward model or the reference model, but at the same time lowers the risk of



(a) Debate

(b) Hawaii

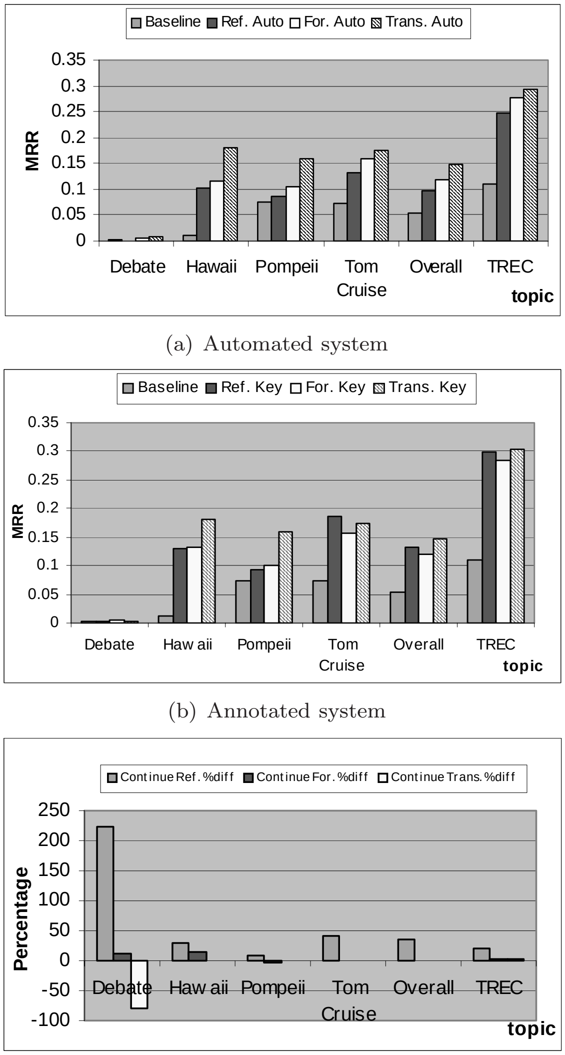
(c) Pompeii

(d) Tom Cruise

(e) Our overall data

(f) TREC2004 data

Figure 2. Coverage comparison between four models based on automated processing



(a) Automated system

(b) Annotated system

(c) Improvement of annotated system

Figure 3. Performance on CONTINUE pairs

introducing unnecessary forward looking centers into processing as in the forward model.

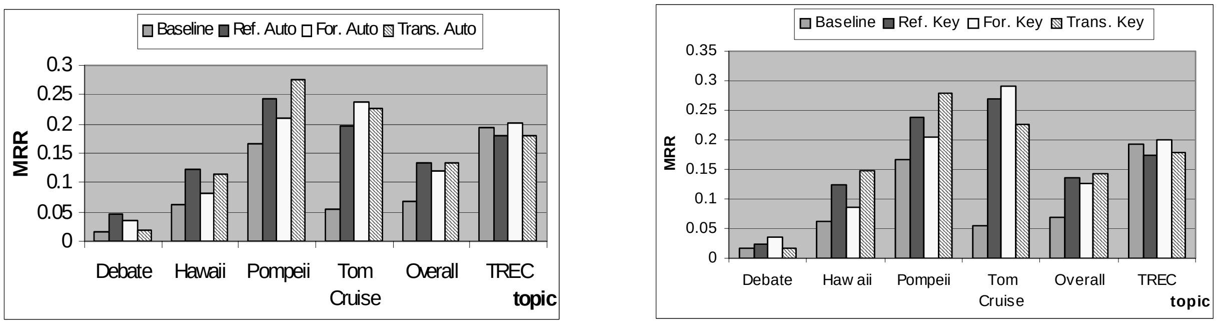
The forward model outperforms the reference model across all the topics, which is also shown in Figure 3(a). This result indicates that reference resolution alone is not enough for obtaining adequate context information for discourses marked with *continue* transitions. Meanwhile, we observed that the reference model outperforms the baseline model for all the topics except *Debate*. The reason is that the reference resolution error brings down the performance for the *Debate* data. This can be seen from Figure 3(b), which shows the performance on the *continue* pairs with all the pronouns correctly resolved. When all the pronouns are correctly resolved, the reference model actually outperforms the baseline model.

Table 7 shows the performance improvement of the transition model over the forward model and the reference model. The results indicate that, for *continue* pairs, the performance improvement of transition model is different across topics. The improvement is less for topics that have higher proportion of pronouns compared to other topics. For the *Tom Cruise* and the TREC 2004 data which have higher percentage of pronouns (i.e., 81.7% and 73.9% respectively), the transition model improves MRR modestly compared to other topics: 33.3% and 17.8% over the reference model, and 9.8% and 5.9% over the forward model respectively.

Figure 3(b) shows the overall performance for the three models based on the annotated pronoun resolution. We see that the transition model based on annotated references is consistently better than the forward model except for the *Debate* data. It seems that pronoun resolution does not help with the transition model for cases with the least number of pronouns. When annotated references are used, the reference model performs better than the forward model for *Tom Cruise* and TREC 2004, and also outperforms the transition model for *Tom Cruise*. These results indicate that when questions have higher percentage of pronouns (e.g., *Tom Cruise*), the reference model with pronouns properly resolved will achieve higher performance compared to other models.

Table 7
Transition model performance improvement for *continue*

Topic (continue)	Transition improvement over reference model (%)	Transition improvement over forward model (%)	Question (excl. the feeding) with pronoun (%)
Debate	707.0	70.1	14.5
Hawaii	78.6	55.6	26.6
Pompeii	86.4	53.4	25.0
Tom Cruise	33.3	9.8	81.7
Overall	50.0	23.9	36.3
TREC2004	17.8	5.9	73.9



(a) Automated system

(b) Annotated system

Figure 4. Performance on RETAIN pairs

Figure 3(c) shows the performance improvement of the annotated systems for the three models compared to the automated system. For the question pairs with the transition type *continue*, the performance of the annotated reference model increases across all the topics. This makes sense because if a question pair is on the same focused entity, according to rule 1 of Centering Theory, this entity would be pronominalized as the backward looking center of the second question. The annotated system avoids the mistakes that the automated system makes in terms of reference resolution. Figure 3(c) also shows that the performance improvement of the annotated forward model and the transition model is relatively small compared to the automated system. The implication from this result is that for *continue* pairs, the forward and the transition model are less sensitive to the accuracy of reference resolution than the reference model.

5.2.2. Retain transition

Next, we present the evaluation results for the *retain* pairs. Figure 4 shows the overall comparison of the three models on the question pairs with the transition type *retain*, with Figure 4(a) for the automated system and Figure 4(b) for the annotated system. Table 8 lists the performance improvement of the transition model over the other two models based on the automated system. We first compare the transition model and the reference model based on the automated processing. Figure 4(a) and Table 8 firstly show that the transitional model performs better than the reference model for *retain* pairs in *Pompeii* and *Tom Cruise*. One advantage of the transition model over the reference model is its capability

Table 8
Transition model performance improvement for *retain*

Topic (retain)	Transition improvement over reference model (%)	Transition improvement over forward model (%)	Question (excl. the feeding) with pronoun (%)
Debate	-57.29	-45.64	14.5
Hawaii	-6.01	38.14	26.6
Pompeii	13.76	30.85	25.0
Tom Cruise	15.65	-4.83	81.7
Overall	0.89	13.19	36.3
TREC2004	0	-10.82	73.9

of adding constraints from the context as in the example (9), where the year 1631 is inherited from Q_i to Q_{i+1} for the query expansion.

(9)

Q_i: In 1631 Vesuvius erupted again. This was the worst eruption since when?

Q_{i+1}: When was Vesuvius' last cycle?

Query terms for Q_{i+1}:
Transition model: {when, was, Vesuvius, last, cycle, 1631}
Reference model: {when, was, Vesuvius, last, cycle}

Secondly, Figure 4(a) and Table 8 show that the transitional model performs the same as the reference model for the TREC 2004 data. The TREC 2004 data does not have many constraints so the strategy for the transition model does not add more information given that the transition model is mostly used to resolve the references as the reference model. However, we would expect performance difference between the two models for longer questions with more constraints such as time phrases. Finally, Figure 4(a) and Table 8 show that the transition model performs worse than the reference model for *Debate* and *Hawaii*. What happened is that some constraints that do not carry much information, such as adverb *there*, actually introduce noise to the search process. Based on this result, we suggested excluding this kind of adverbs in QA processing.

Next, let us compare the transition model with the forward model. From Figure 4(a) and Table 8, we see that the transition model performs better than the forward model for the question pairs in *Pompeii* and *Hawaii*, worse for *Debate*, *Tom Cruise*, and TREC 2004. This result seems rather incidental. However, as we examine closely, we found that the transition model for *retain* pairs does not seem to work better than the forward model for questions that have a high percentage of pronouns (e.g., *Tom Cruise* and TREC 2004 questions). Note that this observation is similar to what has been noticed for *continue* pairs. The fact that the transition model does not work well for the *Debate* data, which does not have many pronouns, indicates that the high percentage of pronouns is not the necessary condition but the sufficient condition for worse transition model performance.

Another interesting observation from Figure 4(a) is that the baseline model outperforms the TREC 2004 data. The TREC 2004 data is more coherent than our user study data under the assumption that the more a discourse participant continues focusing on a discourse entity, the more coherent this discourse would be, and therefore the more *continue* pairs will be observed in this discourse. This is exactly the case for the TREC 2004 data as seen from Table 4. The TREC 2004 data has more *continue* pairs than our data (64% vs. 30%). Intuitively, a more coherent discourse would favor more context information for the purpose of discourse processing. However, the strategy we adopted for the transition model does not seem to help with the *retain* pairs, because the TREC 2004 data does not have many constraints such as time or location. The baseline instead is able to get more context information by simply concatenating the previous question to the question under processing.

Finally, we observed the low sensibility of the transition model to a system's capability of correctly resolving pronouns for the overall user data and especially for the TREC 2004 data.

5.2.3. Shift transition

Finally, we discuss the performance results for the *shift* pairs. Figure 5 shows the overall comparison of the three models on the question pairs with the transition type *shift*, with Figure 5(a) for the automated system and Figure 5(b) for the annotated system.

From Figure 5(a) and Figure 5(b), we see that the transition model performs the same as the forward model because the strategy for *shift* pairs in the transition model is simply to add the forward looking centers from the previous question to the current question, which is exactly the same as the forward model. The baseline model for questions with the *shift* type performs better than for question pairs with the other two types, which indicates that the questions with the least coherence may not need much processing or other processing techniques. It should be noted that, all the context questions within a sequence are somewhat related even if two adjacent ones are regarded as less coherent according to Centering Theory (e.g., identified as *shift*). This is why sometimes for *shift* pairs, by simply running the baseline, we can get pretty good performance (such as for the *Debate* data). The reason that the reference model does not work well is that the *shift* pairs normally do not have referring expressions.

The performance improvement based on the annotated system over the automated system for the *shift* pairs is not as significant as for the other two transition types. Since there are not many cases where pronoun resolution is involved in the *shift* pairs, it is hard to examine how pronoun resolution would impact the different models. We also observed the performance on the *Pompeii* data even drops a little for the annotated system. After examination of the processing, we found examples such as (10), which could provide a possible explanation.

(10)
Q_i: When did Pompeii vanish?
Q_{i+1}: When did Vesuvius erupt?
Q_{i+2}: How did people try to recover their possessions?

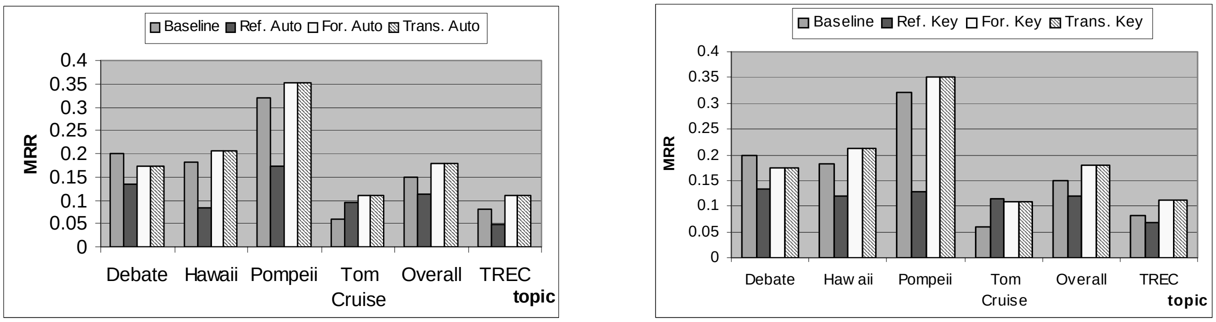
In ($10Q_{i+2}$), although the pronoun their is resolved to people, the reference resolution does not do much to the query terms. However, for the transition model, the proper name Vesuvius is added to the query terms for ($10Q_{i+2}$) because Vesuvius is the forward looking center of ($10Q_{i+1}$). By introducing an important discourse entity Vesuvius, this operation actually increases the chance of hitting the right document for ($10Q_{i+2}$).

To sum up, besides the individual performance characteristics, there are four major conclusions. First, for a context question discourse that has more *continue* pairs, the transition model works better than the forward model and the reference model. Second, for *retain* pairs, the transition model works the best for the data with constraints. Third, for the *shift* pairs, the baseline could be an effective alternative strategy; Fourth, the forward and the transition model are less sensitive to the accuracy of reference resolution than the reference model. In other words, the ability of correctly resolving pronouns affects the reference model the most and the transition model the least.

5.3. Evaluation and Analysis based on Pronouns

To further examine the performance of different models on questions with different pronoun usages, we separated questions into two categories for evaluation: questions with and without pronouns. Figure 6(a) and Figure 6(b) show the evaluation based on the pronoun dichotomy for the automated system and the annotated system.

When we compare Figure 6(a) and Figure 6(b), we notice that the performance of the transition model on the overall user data and the TREC 2004 data is better than the other two models both for the automated and for the annotated systems. This observation is similar to what was found when we separated the questions by transition types. Within individual user data, the per-



(a) Automated system

(b) Annotated system

Figure 5. Performance on SHIFT pairs

formance of the reference model on *Hawaii* and *Tom Cruise* gets increased more than that on the other two topics for the annotated system. A possible reason is that both the *Hawaii* and the *Tom Cruise* data have a high percentage of pronouns. The transition model stays comparatively stable between the automated and the annotated system.

Figure 7 shows the evaluation results for questions without pronouns, with Figure 7(a) for the automated system and 7(b) for the annotated system.

Figure 7(a) and 7(b) show that the transition model is still competitive with the other models even for the questions that do not have pronouns, although the advantage of the transition model for different topics is different. For example, the performance increase for the *Tom Cruise* data is not as big as for the *Pompeii* data.

Compared with Figure 6, the performance of the *Debate* data increases noticeably both for the automated and the annotated system. One possible explanation is that the majority of the *Debate* questions fall into this category. However there is no much difference within the three models for the *Debate* data. This indicates that centering-based models are more appropriate to process context questions focusing on entities rather than on events.

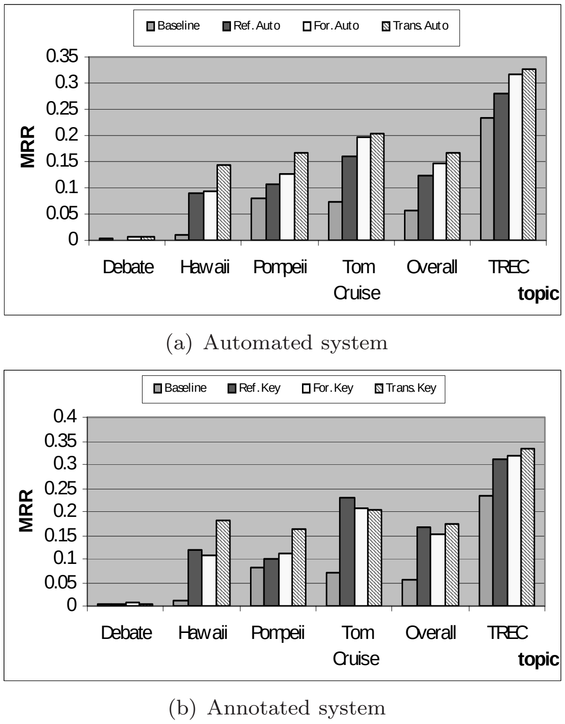
Figure 7(a) shows the automated reference model works better than the baseline model for the *Tom Cruise* and the TREC 2004 data, but not for the other topics. For the non-pronoun containing questions, the reference model just takes all the terms from the question itself. However the baseline model would add the whole previous question to the current question under processing. Comparing Figure 7(a) and 7(b), we see that for the baseline and the reference model, there is no performance improvement for the annotated system over the automated system since there are no pronouns to be resolved.

The performance improvement of the annotated system compared to the automated system for the transition model and the forward model is rather trivial since the difference is within 3%, better or worse depending on different topics.

In summary, there are two important messages conveyed from the analysis based on the pronoun separation. One, the transition model outperforms the forward model and the reference model for both the questions with and without pronouns. Second, there is no significant advantage of the transition model for the event type data.

6. CONCLUSION

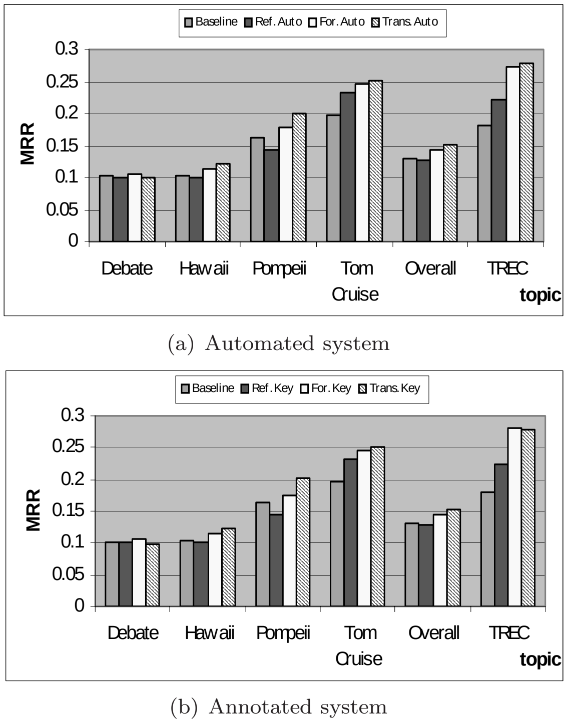
To support coherent information seeking, this paper explores the use of linguistic knowledge in discourse processing for a sequence of questions. A question sequence is considered as a coherent mini discourse and Centering Theory is applied to capture the discourse coherence. Three models



(a) Automated system

(b) Annotated system

Figure 6. Performance for questions with pronouns



(a) Automated system

(b) Annotated system

Figure 7. Performance for questions without pronouns

based on Centering Theory (the reference model, the forward model, and the transition model) were proposed, implemented, and evaluated on our user study data and the TREC 2004 data.

The empirical results indicate that the transition model outperforms the reference model as well as the forward model for the overall data, with or without pronouns. The transition model and the forward model are less sensitive to the accuracy of automated reference resolution of pronouns. More sophisticated processing based on discourse transitions and centers can significantly improve the performance of document retrieval compared to models that only resolve references. Since these models are based on discourse entities, the state-of-the-art natural language processing techniques are sufficient for discourse processing.

This paper presents our initial investigation on the role of discourse processing for context questions. There are many dimensions along which our future work will be pursued. For example, how to use linguistic knowledge and the existing linguistic theories to help process event-based context questions has become an interesting topic. We will also extend context question answering to fully interactive question answering and investigate the role of discourse processing in this new setting.

7. ACKNOWLEDGEMENT

The authors would like to thank Barbara Abbott, John Hale, Cecile Paris, Candy Sidner, and anonymous reviewers for their valuable comments and suggestions. This work was supported partially by grant IIS-0347548 from the National Science Foundation and by the Disruptive Technology Office (DTO) Phase III program for the Advanced Question and Answering for Intelligence (AQUAINT) through Broad Agency Announcement (BAA) N61339-06-R-0034 issued by the Program Executive Office for Simulation Training and Instrumentation (PEO STRI).

REFERENCES

1. B. Abbott, Definiteness and indefiniteness, in Handbook of pragmatics (L. R. Horn and G. Ward, eds.), Oxford, Blackwell, 2004.
2. T. Abou-Assaleh, N.Cercone, J. Doyle, V. Keselj, and C. Whidden, DalTREC 2005 QA system Jellyfish: Mark-and-match approach to question answering, In Proceedings of the 14th Text Retrieval Conference (TREC 2005), (Gaithersburg, MD), 2005.
3. K. Ahn, J. Bos, J. R. Curran, D. Kor, M. Nissim, and BonnieWebber, Question answering with QED at TREC-2005, in Proceedings of the 14th Text Retrieval Conference (Gaithersburg, MD), 2005.
4. L. Azzopardi, K. Balog, and M. D. Rijke, Language modeling approaches for enterprise tasks, in Proceedings of the 14th Text Retrieval Conference (TREC 2005), (Gaithersburg, MD), 2005.
5. B. Baldwin, Anaphora resolution with centering, in Workshop on Centering Theory in Naturally-Occurring Discourse, (Philadelphia, PA.), May 1993.
6. J. Barwise and J. Perry, Situations and attitudes. Cambridge, Mass.: Bradford Books, 1983.
7. S. Brennan, M. Friedman, and C. Polland, Centering approach to pronouns, in ACL87, (Standford, CA.), pp. 155–162, 1987.
8. J. Y. Chai and R. Jin, Discourse status for context questions, in Proceedings of HLT-NAACL 2004 workshop on pragmatics in question answering, (Boston, MA.), pp. 23-30, ACL, May 2004.
9. D. Ferres, S. Kanaan, D. Dominguez-Sal, E. Gonzalez, A. Ageno, M. Fuentes, H. Rodriguez, M. Surdeanu, and J. Turmo, TALP-UPC at TREC 2005: experiments using a voting scheme among three heterogeneous QA systems, in Proceedings of the 14th Text Retrieval Conference (TREC 2005), (Gaithersburg, MD), 2005.
10. R. Gaizauskas, M. A. Greenwood, M. Hepple, I. Roberts, and H. Saggion, The University of Sheffields TREC 2004 Q&A experiments, in Proceedings of the 13th Text Retrieval Conference (TREC-2004), 2004.
11. B. Grosz, The representation and use of focus in dialogue understanding, tech. rep., SRI International, 333 Ravenswood Ave., Menlo Park, CA, 94025, 1977.
12. B. J. Grosz, Focusing and description in natural language dialogue, in Elements of Discourse Understanding (A. Joshi, B. Webber, and I. Sag, eds.), pp. 85-105, Cambridge University Press, 1981.
13. B. J. Grosz, A. K. Joshi, and S.Weinstein, 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.), pp. 44-50, 1983.
14. B. J. Grosz, A. K. Joshi, and S. Weinstein, Centering: a framework for modeling the local coherence of discourse, Computational Linguistics, vol. 21, no. 2, pp. 203-225, 1995.
15. B. J. Grosz and C. Sidner, Attention, intention, and the structure of discourse, Computational Linguistics, vol. 12, no. 3, pp. 175-204, 1986.
16. B. J. Grosz, A. K. Joshi, and S.Weinstein, Towards a computational theory of discourse interpretation, Unpublished manuscript, 1986.
17. J. Gundel, The role of topic and comment in linguistic theory. Distributed by Indiana University Linguistics club, Bloomington, Indiana, 1976.
18. M. A. K. Halliday and R. Hasan, Cohesion in English. London: Longman, 1976.
19. S. Harabagiu, D. Moldovan, M. Pasca, M. Surdeanu, R. Mihalcea, R. Girju, V. Rus, F. Lacatusu, P. Morarescu, and R. Bunescu, Answering complex, list and context questions with LCCs question-answering server, in The Tenth Retrieval Conference (TREC-2001) (E. Voorhees and D. K. Harman, eds.), pp. 355-361, Gaithersburg, MD.: NIST special publi-

cation, 2001.

20. S. Harabagiu, A. Hickl, J. Lehmann, and D. Moldovan, Experiments with interactive question-answering, in Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL 05), (Ann Arbor, MI), pp. 205-214, 2005.
21. J. R. Hobbs, On the coherence and structure of discourse, Tech. Rep. Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.
22. A. K. Joshi and S. Weinstein, Control of inference: role of some aspects of discourse structure - centering, in Proceedings of the 7th international joint conference on artificial intelligence, (Vancouver), pp. 385-387, 1981.
23. A. K. Joshi and S. Kuhn, Centered logic: the role of entity centered sentence representation in natural language inferencing, in Proceedings of the 6th international joint conference on artificial intelligence, (Tokyo, Japan), pp. 435-439, August 1979.
24. D. Jurafsky and J. H. Martin, Speech and Language Processing. NJ.: Prentice Hall, 2000.
25. M. Kameyama, A property-sharing constraint in centering, in Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, (New York, NY), pp. 200-206, 1986.
26. H. Kamp and U. Reyle, From discourse to logic. Kluwer, Dordrecht, 1993.
27. A. Kehler, The effect of establishing coherence in ellipsis and anaphora resolution, in Proceedings of 31st Annual Meeting of the Association for Computational Linguistics, (Columbus, OH.), pp. 62-69, June 1993.
28. E. D. Liddy, A. R. Diekema, and O. Yilmazel, Context-based question answering evaluation, in Proceedings of the 27th Annual ACM-SIGIR Conference, (Sheffield, England), 2004.
29. W. C. Mann and S. A. Thompson, Rhetorical structure theory: A theory of text organization, tech. rep., Technical No. ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.
30. E. Milsakaki and K. Kukich, Evaluation of text coherence for electronic essay scoring systems, Natural Language Engineering, vol. 10, no. 1, pp. 25-55, 2004.
31. M. Mulcahy, K. White, I. Gabbay, and A. Ogorman, Question answering using the DLT system at TREC 2005, in Proceedings of the 14th Text Retrieval Conference (TREC-2005), (Gaithersburg, MD.), 2005.
32. D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu, and O. Bolohan, LCC tools for question answering, in Proceedings of the 11th Text Retrieval Conference (TREC-2002), (Gaithersburg, MD), November 2002.
33. I. Roberts and R. Gaizauskas, Evaluating passage retrieval approaches for question answering, in Proceedings of the 26th European conference on information retrieval, 2004.
34. C. L. Sidner, Towards a computational theory of definite anaphora comprehension in English discourse. Technical report 537, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, June 1979.
35. S. Small, T. Liu, N. Shimizu, and T. Strzalkowski, HITIQA: an interactive question answering system: a preliminary report., in Proceedings of the ACL 2003 workshop on multilingual summarization and question answering, 2003.
36. M. Strube and U. Hahn, Functional centering, in ACL-96, (Santa Cruz, CA), pp. 270- 277, 1996.
37. E. Voorhees, Overview of TREC 2001 question answering track, in proceedings of TREC, (Gaithersburg, MD), November 13–16 2001.
38. E. Voorhees, Overview of TREC 2004 question answering track, in Proceedings of the 13th Text REtrieval Conference (TREC 2004), (Gaithersburg, MD), 2004.
39. M. A. Walker, Evaluating discourse processing algorithms, in Proceedings of 27th Annual Meeting of the Association for Computational Linguistics, pp. 251-261, 1989.
40. M. A. Walker, M. Iida, and S. Cote, Japanese discourse and the process of centering, Computational Linguistics, vol. 20, no. 2, pp. 193-232, 1994.

41. M. A. Walker, A. K. Joshi, E. F. Prince, Centering in naturally occurring discourse: An overview, In Centering Theory in Discourse (M. A. Walker, A.K. Joshi and E.F. Prince, eds.), Clarendon Press, Oxford, 1998.
42. M. Wu, S. Duan, M.and Shaikh, S. Small, and T. Strzalkowski, ILQUA- an IE-driven question answering system, in Proceedings of the 14th Text Retrieval Conference (TREC 2005), Gaithersburg, MD, 2005.