

This is to certify that the
dissertation entitled

LINGUISTIC KNOWLEDGE BASED DISCOURSE MODELING FOR
CONTEXT QUESTION ANSWERING

presented by


MINGYU SUN

has been accepted towards fulfillment
of the requirements for the

PHD

degree in

LINGUISTICS


Major Professor's Signature

11/22/06

Date

MSU is an Affirmative Action/Equal Opportunity Institution

LIBRARY
Michigan State
University

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

**LINGUISTIC KNOWLEDGE BASED DISCOURSE
MODELING FOR CONTEXT QUESTION
ANSWERING**

By

Mingyu Sun

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Linguistics

2006

ABSTRACT

**LINGUISTIC KNOWLEDGE BASED DISCOURSE
MODELING FOR CONTEXT QUESTION
ANSWERING**

By

Mingyu Sun

This dissertation is motivated by the recent developments in scenario-based context question answering (QA). The role of discourse processing, and its implication on query expansion for a sequence of questions, is investigated. My view is that a question sequence is not random, but rather, follows a coherent manner to serve certain information goals of users. Therefore, this sequence of questions can be considered as a mini discourse with some characteristics of discourse cohesion and discourse coherence. Understanding such a discourse will help QA systems better interpret questions and retrieve answers. Thus, in the first part of my study, I propose three models driven by Centering Theory for discourse processing: an anaphora 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. The 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. In the second part of the study, the influence of the processing based on pronoun resolution and definite description resolution is investigated. Results show that a combined model that incorporates both approaches performs the best under the situation where no explicit target is given for

the context questions. The processing for the *event* type of context question answering is also investigated briefly. For different discourse models proposed in the dissertation, systematic evaluation is provided and the potentials and limitations of these models in processing coherent context questions are also discussed.

Copyright by
Mingyu Sun
2006

To my parents, Zhengmei Yao and Xian Sun, my husband, Chunlei Wu,
for their endless love.

ACKNOWLEDGMENTS

I am grateful for all the support I have received while researching and writing up this dissertation. I would like to thank the members of my dissertation committee: I am grateful to Dr. Barbara Abbott for her many years of encouragement and guidance. She has been showing me how to present my own ideas into scholarly writing. Her careful readings, detailed comments and insightful thoughts will never be forgotten. I would like to thank Dr. John Hale for being my committee chair, taking care of all the administrative issues, giving me quick replies to my questions, and especially making a long trip for my defense. Thanks especially to Joyce Chai, my co-chair, who has supported me every step of the way. She spent numerous hours sitting with me discussing the initial ideas of the dissertation, walking with me line by line on the drafts and guiding me to shape part of my work into two publications, giving comments on the implementation details and results. Without her support, I could not have done what I was able to do. I'd also like to thank Dr. Yen-Hwei Lin and Dr. Dennis Preston who have given me great comments on my proposal.

The acknowledgments would not be complete without a heartfelt thanks to my friends. I would like to thank Mike Kramizeh for his support, especially for standing by me at my defense to make sure the technology worked. I would like to thank Julie Delgado for her kindness and help during my stay at the Linguistic Department. I would like to thank Zheng Wang and Jun Ao for their long term support. Finally, I would like to thank Vineet Bansal, a dear friend who supported me through all of the challenges. He offered to trouble-shoot my programs, proofread draft copies and listened to my moanings. I cannot thank him enough.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xi
1 Introduction	1
1.1 Question answering	1
1.2 Problem statement	2
1.3 Overview of the thesis	5
2 TREC Question answering	7
2.1 TREC QA tasks	7
2.2 TREC systems	10
2.2.1 System architecture	10
2.2.2 Question processing	11
2.2.3 Further processing	13
2.3 TREC evaluation	15
2.4 Summary	17
3 QA discourse and discourse modeling	18
3.1 Introduction	18
3.2 QA context	18
3.3 QA discourse	19
3.4 Discourse modeling	21
3.4.1 Discourse structure	22
3.4.2 Discourse structure theories	22
3.4.3 Computational efforts	24
3.5 The hypotheses	26
3.6 Summary	29
4 Data collection and platforms	30
4.1 Introduction	30
4.2 System overview	30
4.3 Question collection	31
4.3.1 User data	31
4.3.2 TREC 2004 data	33
4.3.3 TREC 2005 data	34

4.4	Data format	36
4.4.1	Original TREC data format	39
4.4.2	Customized data format	40
4.4.3	Annotation validation	43
4.5	Document collections	44
4.5.1	User study documents	45
4.5.2	AQUAINT corpus	46
4.5.3	Document format	46
4.6	Document retrieval engine	47
4.7	Summary	49
5	Centering-based discourse modeling	50
5.1	Motivation	50
5.2	Centering Theory	55
5.2.1	Grosz, Joshi and Weinstein (1995)	56
5.2.2	Terminologies and definitions	57
5.2.3	Constraints and rules	59
5.2.4	Centering algorithms	63
5.3	Three discourse models for query expansion	65
5.3.1	Anaphora model	65
5.3.2	Forward model	72
5.3.3	Transition model	73
5.4	Data analysis	78
5.5	Evaluation	80
5.5.1	Overall Evaluation Results	81
5.5.2	Evaluation and Analysis based on Transitions	92
5.5.3	Evaluation and Analysis based on Pronouns	101
5.6	Summary	105
6	Discourse models based on definiteness hierarchy	106
6.1	Introduction	106
6.2	A new challenge in QA discourse	108
6.3	Related work	110
6.3.1	Computational work on definite descriptions	110
6.4	Theoretical background	112
6.4.1	Definiteness hierarchy	112
6.4.2	Familiarity theory	113
6.4.3	Uniqueness theory	115
6.5	Classification of definite description uses	117
6.5.1	Fine-grained classifications	117
6.5.2	Prince (1992)	121
6.5.3	Poesio and Vieira (1998)	122
6.5.4	Classifying definite descriptions in QA discourse	123
6.6	Discourse models focusing on definite descriptions	126

6.6.1	Data analysis	127
6.6.2	No-target situation	132
6.6.3	Target-given situation	136
6.7	Evaluation for definiteness hierarchy based models	140
6.7.1	Evaluation for the no-target situation	140
6.7.2	Evaluation for the target-given situation	146
6.7.3	Evaluation based on the definite description types	150
6.8	Summary	155
7	Conclusion	157
7.1	Summary	157
	APPENDICES	161
A	User data collection	161
A.1	Instruction	161
A.2	Documents collection on <i>Presidential Debates 2004</i> :	162
A.3	Documents collection on <i>Tom Cruise</i>	165
A.4	Documents collection on <i>Pompeii</i>	167
A.5	Documents collection on <i>Hawaii</i>	170
	BIBLIOGRAPHY	173

LIST OF TABLES

4.1	TREC 2004 targets	35
4.2	TREC 2005 targets	37
4.3	TREC 2005 targets (cont')	38
5.1	Extended transition states (Adapted from Brennan et al.[1])	62
5.2	Transition rules for questions without pronouns but with non-pronominal referring expressions	75
5.3	Examples of transition rules on non-pronominal referring expressions . . .	75
5.4	Query expansion strategies based on transition type	77
5.5	Characteristics comparison between my data and TREC 2004 data (including only factoid questions)	79
5.6	Overall performance of different models on document retrieval for my data and TREC 2004 data	85
5.7	Document retrieval performance based on the transition model and passage retrieval performance from the University of Sheffield on TREC data . . .	92
5.8	Transition model performance improvement for <i>continue</i>	93
5.9	Transition model performance improvement for <i>retain</i>	97
6.1	Definiteness hierarchy of noun phrases	113
6.2	Hearer and discourse status of a discourse entity*	122
6.3	Data analysis for the original TREC 2004 and TREC 2005 data	130
6.4	Data analysis for the modified TREC 2004 and TREC 2005 data	131
6.5	MRR performance based on definite hierarchy in no-target situation	143
6.6	Coverage performance on all TREC data* in no-target situation	144
6.7	MRR performance based on modified definite hierarchy in no-target situation	146
6.8	MRR Performance based on definite hierarchy in target-given situation . .	147
6.9	Coverage performance on all TREC data* in target-given situation	147
6.10	Question distribution based on definite description types	151

LIST OF FIGURES

2.1	Question answering system architecture	11
4.1	Flow chart of the centering-based automated system architecture	31
5.1	Overall comparison of four models based on automated processing	87
5.2	Coverage comparison between four models based on automated processing	91
5.3	Performance on CONTINUE pairs	94
5.4	Performance on RETAIN pairs	96
5.5	Performance on SHIFT pairs	100
5.6	Performance for questions with pronouns	103
5.7	Performance for questions without pronouns	104
6.1	MRR Performance comparison between the definiteness hierarchy based models and the modified version	146
6.2	MRR performance on all TREC data* in target-given situation	148
6.3	Overall comparison of four models based on automated processing on an- tecedents	151
6.4	MRR performance on questions with direct anaphoric definite description in no-target situation	152
6.5	MRR performance on questions with bridging definite description in no- target situation	153
6.6	MRR performance on questions with discourse-new definite description in no-target situation	154
6.7	MRR performance on questions without any definite descriptions in no- target situation	155

CHAPTER 1

Introduction

1.1 Question answering

With enormous electronic textual data available, it is important for human users to be able to find information through natural language questions. Research on Question Answering (QA) systems aims to provide such a capability. Given a natural language question from a user (such as a question in English, *What is the oldest sports trophy?*), an ideal computer system can consult a database, knowledge base, or go to the Web to locate the answer, *the America's Cup*. Natural language processing and Information Retrieval (IR) make it possible to provide informative, appropriate and non-misleading answers to simple questions. The Text REtrieval Conference (TREC^{1 2}) has been conducting evaluations for different QA systems since 1999. The submitted competing QA systems have been moving from the stage of dealing single fact-based questions to more complex questions such as opinion questions (like *What do people think about the 2004 presidential debate?*). Various issues in terms of system performance, uses, and techniques have invited international research interests and activities in QA.

Putting aside the problem of question complexity, once engaged in interaction with a system, a user may ask a series of questions. This sequence of questions is not arbitrary

¹<http://trec.nist.gov/>

²TREC is co-sponsored by the National Institute of Standards and Technology (NIST), Information Technology Laboratory's (ITL) Retrieval Group of the Information Access Division (IAD) and the Advanced Research and Development Activity (ARDA) of the U.S. Department of Defense.

but rather coherent, leading towards some information goal. In the TREC 2004 and TREC 2005 Question Answering system competitions, questions were grouped together by target (similar to topic). The participating systems usually have modules like question processing, document retrieval, passage selection and answer retrieval. In their system architectures, question processing was the first module in the pipeline. How to characterize a question sequence therefore becomes a very interesting problem in that it will affect query expansion in the question processing stage, answer extraction in the answer retrieval stage and the overall performance.

The purpose of this dissertation is to develop interpretive discourse models of question sequences for a computer system that takes textual questions as input. These models aim to provide proper interpretation of the questions and to provide efficient computation algorithms for question processing and query expansion. From a linguistic point of view, semantic and pragmatic information in the questions will be investigated in developing these models.

1.2 Problem statement

In this section I will introduce the notions of *context* and *context questions*, current methods used in processing context questions, and some relevant linguistic research.

In linguistic literature, the term *context* has been defined differently, for example, as “the discourse that surrounds a language unit and helps to determine its interpretation” by WordNet³, or as “the complete discourse record of a discourse at any given point, including both linguistic and situational information” (Geluykens, 1992)[2]. In Geluykens’ definition, the linguistic information usually refers to the surrounding words or sentences of a piece of text. The situational context refers to “the features of the non-linguistic world in relation to which linguistic units are systematically used” (Crystal, 1992) [3]. Recent work has given context a more dynamic interpretation. Roberts [4] pictures context at a given point of a discourse as an ordered set which includes five sets of linguistic elements

³See WordNet project website <http://wordnet.princeton.edu/>

(such as entities in a discourse, Common Ground propositions) or non-linguistic elements (such as intentional goals of speakers).

The term *context* in the field of Question Answering can refer to different contexts such as user context [5] and discourse context [6]. In this study, I will focus on the discourse context, in particular, the discourse of a sequence of questions. I will limit ourselves to a description of *context* as follows: given a series of questions, when one question is under processing, its preceding question or questions or constituents of these questions are all regarded as its *context*. A *context question* is a follow-up question that needs to be processed using its context information. *Context question answering* is the task of answering such context questions. Consider the following example as illustration:

(1)

Q1: Who was Tom Cruise married to?

(A1: Nicole Kidman)

Q2: What was her Broadway debut?

(A2: The Blue Room)

Q3: Who filed for divorce?

(A3: Tom Cruise)

It is impossible for a system to isolate (1Q2) from (1Q1) and (1A1), to retrieve correct answer (1A2). In this case, the pronoun *her* in (1Q2) has to be solved before (1Q2) is sent to an Information Retrieval (IR) engine(i.e. application that seeks information from various resources ⁴). For an IR engine, the resolution of pronoun references is necessary because otherwise pronouns would be treated as stop words⁵ and totally ignored in the retrieval process. (1Q2) therefore is the context question of (1Q1) and (1A1). One method that is currently used in QA systems is called reference-resolution. What this method does is to find what such referring expressions as pronoun *he*, demonstrative *this*, or definite noun phrase *the book* refer to in the context questions. For instance, in example (1), the possessive pronoun *her* will be resolved first using the extracted answer (1A1). Therefore

⁴such as the backend engine of <http://www.google.com> that retrieves textual information from the World Wide Web

⁵Please see the definition at http://en.wikipedia.org/wiki/Stop_words

the query terms⁶ that will be fed into an IR engine would be the terms from the current question (1Q2), {what, was, her, Broadway, debut} concatenated with the terms (words) from the answer to (1Q1), that is, {Nicole, Kidman}. Similar reference-based methods are computationally implemented. However, it is noticeable that this kind of approach is not able to solve (1Q3) since (1Q3) does not have any explicit reference whatsoever. Human being, however knows that the marriage (as being an event) involves two parties (i.e. two referents: Tom Cruise and Nicole Kidman) for (1Q3). The fact that there is not many systematic linguistic analyses on context question answering motivates the present study on how context questions are related not only lexically but also at a higher semantic and/or pragmatic level.

Real life experience has shown that questions raised in an information seeking session are often related, because one question forms a context for another. Presumably, the techniques used to process single question would not be an efficient means of gathering such context information for the system engaged in an interactive QA session. Thus, the context is critical in processing a question sequence. But how does a computer system determine what context to use and when to use it? What is the role of context in processing a series of question? The central hypothesis of this thesis is that context can be modeled, or captured by an entity-based linguistic theoretical framework.

This thesis is a linguistic effort to investigate the context question answering problem. The goal is to show that a system needs to take into account the linguistic knowledge of questions, that is, to properly represent discourse entities based on linguistic knowledge. This approach implies that discourse entities, represented by definite descriptions, pronouns or proper names play an important role in determining the relationship between/among context questions. This work also aims to show that successfully capturing the context in a question sequence requires computational algorithms that combine both lexical information and discourse information of the questions.

⁶Query terms here refer to a set of tokens obtained from the linguistic expressions in a natural language question. For instance, the query terms for the question *Who is Tom Cruise?* would be {who, is, Tom, Cruise}. They are treated as a bag of words by most retrieval engines.

1.3 Overview of the thesis

The thesis is organized as follows: in Chapter 2, I present the background information of the Text REtrieval Conferences(TREC), which provides essential material for a complete understanding of the context question answering task. This chapter introduces the TREC QA tasks, and includes discussions on the state-of-the-art TREC systems, the question processing component, and evaluation metrics for the TREC systems.

Chapter 3 continues the discussion of background material from a linguistic perspective. The context question problem is established as a discourse modeling problem, for a sequence of questions are treated as a mini discourse. Discourse structure and discourse structure theories are reviewed briefly. Relevant computational efforts are presented to show how previous work has led to the present state of the field. Moreover, two hypotheses are presented for the work discussed in Chapter 5 and Chapter 6.

Chapter 4 presents all the empirical elements involved in the current study. They include: the context question sets, the annotation scheme of the questions, the document collections, etc. Data collection as well as data formatting are described in detail.

Chapter 5 is motivated by recent efforts in scenario-based context question answering. I investigate the role of discourse processing and its implication on query expansion for a sequence of questions. The view is that a question sequence is not random, but rather follows a coherent manner to serve some information goals. 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, three models driven by Centering Theory for discourse processing are examined: an anaphora 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. The 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 pronoun references. This chapter provides

a systematic evaluation of these models and discusses their potentials and limitations in processing coherent context questions.

Chapter 6 presents another attempt at modeling context questions, especially the ones that contain definite descriptions as well as pronouns. Pronouns and definite descriptions are the focus for this part of the study. Definiteness hierarchy is adopted to help resolve pronouns in context questions. Definite descriptions are classified into three classes for the purpose of query expansion. Experiments on context question data are conducted under two situations: one is that a target/topic is given for each context set (target-given situation), and the other is that no target is given (no-target situation). Different models involving pronoun resolution, definite description resolution, and targets are developed. The results show that in the no-target situation, extra processing on the definite descriptions and the pronouns based on the definiteness hierarchy improves the performance significantly. In the target-given situation, a target-appending strategy shows its efficiency, which implies that identifying a topic or target for a context question set is a task as important as processing the context questions themselves.

Chapter 7 summarizes the work presented in the dissertation, and contains some hindsight recommendations for improvements and possible future work.

CHAPTER 2

TREC Question answering

This chapter provides the background material on the TREC QA tasks and the state-of-the-art Question Answering technology. It describes how context Question Answering enhances the current Question Answering systems.

2.1 TREC QA tasks

For the purpose of boosting research in information retrieval and facilitating technology transfer from the research community to commercial products, the Text REtrieval Conference (TREC) has been conducting tracks for different focus areas since 1992. For instance, TREC 2006 has 7 tracks: Blog Track, Enterprise Track, Genomics Track, Legal Track, Question Answering Track, SPAM Track and Terabyte Track. These tracks aim to solve different problems, for example, the Blog Track, a new track in TREC 2006 is designed to explore information seeking behavior in the blogosphere. Note that information retrieval is no longer confined to text retrieval, which provides users to access to natural language texts. It is possible to retrieve information from other type of media such as video clips. The various TREC tracks have ignited enormous research interests all over the world. The number of participating research groups has increased over the years ¹. These groups are mainly academic, commercial, and government institutions. Although some tracks

¹For example, there were 66 groups from 16 different countries participating in 1999 (TREC-8). But in 2004 (TREC-13), the number has increased to 103 groups from 21 different countries

are added and some are dropped over the years, the Question Answering track has been on the TREC task list since 1999, with some changes made gradually.

Information retrieval is concerned with how a user mines information to satisfy his/her information need. Traditionally, there are two types of information retrieval: *ad hoc* and *known item search*. Typical examples of *ad hoc* search would be a library search system or an online search engine, which takes the user's investigation topic of any kind, searches a set of documents (the library's holdings or the World Wide Web). The *known item search* is the scenario where the user is trying to find certain information that he/she knows exists. Both types of information retrieval focus on locating information on a document level, namely, document retrieval. What document retrieval does is to match a user query (such as a natural language question) against large textual collections and return a ranked document list. The text collections could be of any kind, such as newswire articles, book chapters, user manuals or a large-scale corpus. Document retrieval systems use different techniques to retrieve relevant documents from the collections. This practice reflects the characteristics inherited from the traditional library reference systems. The TREC QA tracks were initiated to take a step further from document retrieval to answer retrieval. At times, users would prefer the system to offer the answer instead of going through all the returned documents looking for the exact answer. The QA tracks aimed to address this challenge and each QA track consisted of one or more tasks trying to extract answers to different types of questions.

The first five tracks (from 1999's TREC-8 to 2003's TREC-12) focused on answering single *factoid* questions like "*what is the oldest sports trophy?*". The total of 2393 such questions were either created by the TREC staff or from search or query logs donated by different sources (Encarta, Microsoft, Ask Jeeves, AOL etc.). Since 2001, TREC added *list* type of questions in addition to the main factoid question task. List questions are questions such as *Name 32 countries Pope John Paul II has visited*. Answering such questions requires a system to submit an answer based on the information located in multiple documents. The list questions were created by the TREC assessors. TREC 2003 added *definition* type of questions such as *Who is Aaron Copland?*. A definition question

asks for relevant information on a person or an organization. Definite questions also require systems to be able to locate information from multiple documents. Comparing with the list questions, the information of interest for definition questions is much less incisively represented. The results of the TREC 2003 track demonstrated that the list and definition questions posed challenges not only for the QA systems but also for the evaluation of the systems. However, these two types of questions were kept in TREC 2004 and TREC 2005.

The main task in TREC 2004 was on *context questions*, a form of question first investigated in 2001's TREC-10. Instead of being independent individual questions, factoid questions and list questions are grouped into series each of which has a *target* associated with it. A question series consists of several factoid questions, zero to two list questions. To each series, TREC2004 also added a question of type *other*, which is equivalent to the TREC 2003 definition questions. It could also be interpreted as "tell me something interesting about this target". Each series has exactly one *other* question. The questions in a series ask for information about the target. The target and previous questions in the series provide the context for the question under processing. The series a question belongs to, the order of the question in the series, the question type (such as "list") and the associated target were all explicitly given in the XML format. Details of the TREC question sets and examples are given in Chapter 4. Question series were developed by NIST staff and TREC assessors. The questions were created in a scenario where an average native speaker of English asked a series of questions about a term (i.e. target) encountered while reading a US newspaper [7]. In practice, the TREC assessors created some questions related to the target and then searched the test document collections looking for answers to these questions. NIST staff then reviewed the information found in the related documents and reconstructed the final test series.

The development of the QA tracks and increased difficulty of QA tasks imply that QA technologies have advanced tremendously in response to the application demand in the area. The context QA task has become a regular task at TREC evaluation since 2004. For each series of context questions, participating systems were required to process each

question in the series independently from one another and in question order. Systems were allowed to use the questions or answers from earlier questions in a series to answer later questions but not the other way around. Next, I will introduce TREC QA systems and the technology involved in context Question Answering.

2.2 TREC systems

2.2.1 System architecture

Processing of the questions is supposed to be strictly automatic. The participating systems accomplish the TREC tasks by directing subtasks to different system components. There are four prototypical components common to most system architectures. In Figure 2.1, the pipeline consists of question processing, document retrieval, passage selection, and answer retrieval. Question processing components focus on the preprocessing of the questions and outputs query terms for the document retrieval component. Document retrieval component usually takes the query terms as input to a search engine (such as Lemur ² or Lucene ³), which is programmed to find information that matches with user's search criteria. Document retrieval component outputs a list of relevant documents for the passage selection component, which then further narrows down the search to paragraph level and outputs passages that may contain the potential answers. Within the targeted passages, the answer extraction component finally pinpoints an answer to the original question.

Based on the four components, different systems have their own system architectures tailored to fit their implementation purposes. Some could be as sophisticated as the ILQUA system in Wu 2005's work [8]. Note that many processing techniques were implemented in the ILQUA system consisting of subcomponents under the four logical components. For instance, there were four subcomponents for the question processing component. They were: syntactic chunking, type categorization, target classification and query generation.

²<http://www.lemurproject.org/>

³<http://lucene.apache.org/java/docs/>

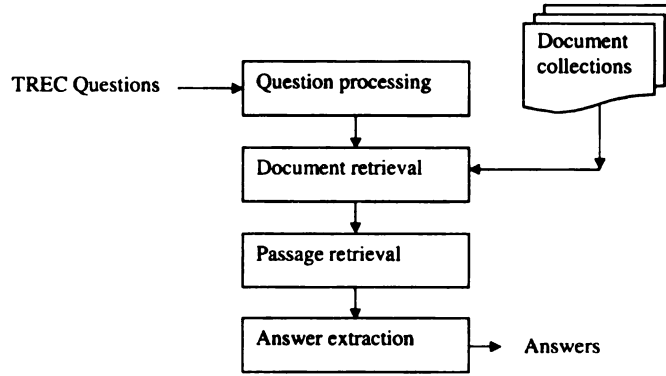


Figure 2.1. Question answering system architecture

It should be noted that the overall approach used in the participating systems stayed rather stable over the years.

For answering the factoid questions, systems usually first identify the expected answer type of the question (such as *date*, *name*, *location*, *size*, *speed*, *city*, *nationality*, etc.), and then use question words and/or related terms as the query to retrieve documents or passages that may contain answers that match with the answer type, and finally extract an answer from the retrieved passage. A system often uses the same architecture for processing both the factoid and the list questions. The difference is that they return a single answer for a factoid question but multiple answers for a list question. Next, I will briefly summarize the major techniques used in the TREC 2005 systems for the question processing component, which my discourse modeling is mostly concerned with.

2.2.2 Question processing

Traditional QA systems have question analysis as the first stage of processing. Research shows that the question analysis component included in the retrieval methods specialized for QA has effect on the overall system performance in a positive way [9]. Except for [10] that applies Discourse Representation Structure (DRS) in question processing, most work related to question processing has focused on three aspects. 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”). This

is similar to the processing for the isolated factoid questions. The goal is to help pin-point the expected answer strings. Several systems in TREC 2005 have adopted this technique such as [11] and [8].

The second aspect emphasizes the processing of the words (i.e. the tokens) in the questions. Parsing tools (which help to find verbal predicates), POS tagging tools, name-entity⁴ recognizers (which tag name-entities into categories), statistical analysis (unigram, bigram, n-grams for question words⁵) and knowledge bases (such as WordNet's Synset, which provides synonyms for a particular word) are utilized to expand the queries. Systems such as [11], [13],[14] and [8] used this method.

The third is to make use of the target (topic) provided by TREC datasets. TREC QA track provided a target for each context question series. A target is a linguistic expression denoting a person, a thing, an organization or an event. Although the target of each series is given, it is not necessarily mentioned in each of the factoid question and list question in a series. Therefore in TREC 2005 some systems chose to append the target to each question within the series such as [11]. Another approach was to replace all the pronouns that appear in the question set with the target such as [15]. The methods were effective because the target was the domain for the questions. However there were problems too. One problem was that sometimes pronouns do not refer to the target. Instead, some pronouns refer to entities appearing in the previous question(s). For example, the target in TREC 2005 question series 136 is *Shiite*, however the pronouns *his* and *he* do not refer to *Shiite*.

(1)

Q1: Who was the first Imam of the Shiite sect of Islam?

Q2: Where is *his* tomb?

Q3: What was this person's relationship to the Prophet Mohammad?

Q4: Who was the third Imam of Shiite Muslims?

⁴Name-entity could be person, location, and organization, as well as times, data, percentages, money amounts, etc.

⁵Interested readers may refer to Jurafsky and Martin 2000[12] for the definition of *unigram*, *bigram* and *n-grams*.

Q5: When did *he* die?

To solve this problem, some systems conducted true anaphora resolution for pronouns. However it was difficult to judge how much benefit was gained from the extensive anaphora processing⁶. After all, most pronouns do refer to the target. Another problem was that it is not necessarily the case that the target is referred only by the pronouns. Sometimes definite descriptions or demonstratives are used instead. For example, The target for question series 75 in TREC 2005 main task is *Merck & Co*.

(2)

Q1: Where is *the company* headquartered?

Q2: What is their symbol on the New York Stock Exchange?

Q3: Name products manufactured by *Merck*.

Some systems (such as U of Sheffield [16]) chose to replace nominals as well as the pronouns with the target. In TREC 2005, a new difficulty was introduced into the processing because of the *event* type of target. Events such as *Hindenburg disaster* and *Russian submarine Kursk sinks* were given as targets for question series. This new type of target added more complexity to the question processing, upon which I will have detailed discussion in Chapter 6.

In general, the overall approaches for question processing were question type analysis, lexical expansion on question words, simple target appending method, and statistical analysis. There was only one system in the TREC 2005 that used discourse analysis approach for context question processing. In general, the research on answering the questions in series from discourse analysis point of view has been limited.

2.2.3 Further processing

The main purpose of the document retrieval component is to search document candidates for further processing. Different software and search engines were chosen for this purpose in the TREC 2005 systems. Software such as MySQL was used in [17] to help with the full-text search. Some search engines are freely available for the research community,

⁶I will discuss this in Chapter 6

for instance, the Lemur⁷ tool kit used in [15] and [10], Lucene⁸ used in [16],[13] and [11] and PRISE⁹ used in [17]. Some were developed by the participating groups themselves, such as the INQUERY developed at University of Massachusetts at Amherst [8]. Search engines basically index the document collection in the way that they ensure the search to be quick and effective.

For the passage selection and answer extraction components, processing techniques are more sophisticated, varying from system to system. Statistical methods, heuristics as well as natural language processing tools allow the systems to finally extract answers for the original questions. Filtering of irrelevant documents is normally based on answer type, target or question terms. Passages were then selected among the candidate documents. Given the fact that state-of-the-art NLP techniques (e.g. natural language parsers, POS taggers etc.) are available, it is each individual system's choice as how to use these techniques to implement various ideas at this stage. For instance, lexical resources (WordNet, online dictionary *Wikipedia* ¹⁰) have been used to locate matched name entities in systems such as [16] and [15]; The technique of indexing (sentence level indexing) was also implemented in systems such as [13] to narrow down the search domain. Surface pattern matching[8] and logical form matching[16] were also seen in passage retrieval. Like in question processing, NLP software (parsers, POS taggers and name entity recognizers) were utilized at this stage too.

In the final stage of answer extraction, the techniques used in the TREC 2005 systems varied to the extent that no other stages have the same diversity as in this stage. Besides the techniques used in the previous stages, the Web was also used as a resource [10] [11]. Logical prover [18] and some ranking or scoring schemes (such as [10] and [19]) were also employed specifically for answer extraction.

Generally systems used different techniques to narrow down the search and reduced the returned information. The TREC participating groups submitted their retrieval results

⁷<http://www.lemurproject.org/>

⁸<http://lucene.apache.org/java/docs/index.html>

⁹<http://www-nlpir.nist.gov/works/papers/zp2/zp2.html>

¹⁰<http://www.wikipedia.org/>

within one week of receiving the test set. The TREC evaluations on the results were then conducted by the TREC staff. Next, I will discuss how the evaluations were done and what the current evaluation metrics are for the factoid questions of which the context question series consist in my data pool.

2.3 TREC evaluation

Since the QA track aims to obtain direct answers to the questions instead of documents containing the answers, the TREC evaluation is at the answer extraction level rather than the document retrieval level. After the answer string is returned by the answer extraction component, it will be independently judged by two human assessors. Only when there is a disagreement between these two assessors, will a third judge, normally a NIST staff member be introduced to make a final judgment. Each response will have the following 4 judgments: *correct*, *not exact*, *not supported*, and *incorrect*. The answer string is regarded as *correct* when it is exactly the right answer and it is supported by the document in which it appears. The response is *not exact* when the answer is contained in the answer string and the document returned supports the answer, but the string has more or less information than just the answer itself. The response is *not supported* when the answer string contains a right answer but the document does not clearly answer the question. The response is *incorrect* when the answer string does not contain the right answer. NIL responses are for the questions to which there are no answers in the document collection¹¹.

Considering the difference of question type and response format, the final score of the participating systems is decided as a weighted average of the scores computed according to question type. The scores for factoid questions were computed by *accuracy*, which is defined as the fraction of questions judged correct. For list questions, multiple answer strings were returned and the scores for list questions were computed using *instance precision (IP)* and *instance recall (IR)*: $IP = D/N$ and $IR = D/S$; where S is the

¹¹It was the NIST staff's ultimate goal to provide test sets where each question should have an answer in the document collections, however in both the TREC 2004 and the TREC 2005 data, there were some questions that did not have answers.

number of known instances for the list question, D is the number of correct, distinct responses returned by the system, and N is the total number of responses returned by the system. F score was then computed using the following formula: $F = \frac{2*IP*IR}{IP+IR}$.

Since the *other* questions are similar to *definition* questions, the returned answer strings should have information atomic about the target, and was not part of or an answer to an earlier question in the series [7]. Each answer string is defined as a *nugget* for the list question. The final score for an *other* question was computed using F measure: $F(\beta = 3)^{12} = \frac{10*precision*recall}{9*precision+recall}$. Next, I will see what nugget recall and nugget precision are. “Given the nugget list and the set of nuggets matched in a system’s response, the nugget recall of the response is the ratio of the number of matched nuggets to the total number of vital nuggets in the list.” Nugget precision is another measure which is more complicated than nugget recall. It was computed based on a length-based measure that “starts with an initial allowance of 100 characters for each (vital or non-vital) nugget matched.” “If the total system response is less than this number of characters, the value of the measure is 1.0. Otherwise, the measure’s value decreases as the length increases using the function $1 - \frac{length - allowance}{length}$ ” [7]. $\beta = 3$ indicates that the nugget recall is three times as important as nugget precision.

Based on the scores for each type of question, the final score for a QA run was computed as a weighted average of the three component scores: $FinalScore = .5 * FactoidAccuracy + .25 * ListAveF + .25 * OtherAveF$; where the *FactoidAccuracy* is the *accuracy* measure score for the *factoid* questions; the *ListAveF* is the average F score for the *list* questions; the *OtherAveF* is the average F score for the *other* type of questions. It should be noted that this score was question-based, which means that each question was treated equally no matter which series it is in. TREC also conducted series-based evaluation, which gives equal weight to each series. Interested readers may refer to [7] for more details. The evaluation details for the current study will be discussed in Chapter 5.

¹²where beta is a parameter signifying the relative importance of recall and precision. A value of 3 indicates that recall is 3 times as important as precision

2.4 Summary

This chapter has examined the TREC background for the task of question answering. It has introduced QA tasks in recent TREC conferences and presented the prototypical TREC system architecture. Based on the pipeline, relevant techniques used in TREC 2005 QA main task were introduced, especially the ones used in the question processing stage. I concluded with the evaluation metrics for the TREC QA track. The background is necessary to present the readers a complete picture in which context question answering resides.

The next chapter continues my discussion of background material by focusing on linguistic aspect of discourse processing and describing how the notion of discourse can be used to guide the current study. Theories of discourse coherence and discourse modeling will be the highlights of the next chapter.

CHAPTER 3

QA discourse and discourse modeling

3.1 Introduction

In this chapter, I continue the discussion on context question processing in a question answering system, but from a discourse point of view. Question series are assumed to be coherent and thus form a discourse. Some relevant linguistic theories of discourse coherence and discourse structure will be reviewed and investigated as to what extent they are useful in solving the current problem of context question answering.

3.2 QA context

Before I delve into solving the problem of context question answering, I would like to clarify some terminologies. The term *context* with regard to a question has been defined in Chapter 1 as previous questions or answers to the previous questions, to be more specific. Given that the state-of-the-art QA systems do not use the answers to the previous questions to help processing the current question, in the rest of the thesis, *context* will exclude the answers to the previous questions for two reasons: first, it is hard to guarantee that the correct answer of each question can be retrieved by the question answering system that I use; second, the focus of the study is on question processing rather than on answer extraction. As Chapter 2 indicated, question answering based on context was first

investigated in TREC 10 Question Answering Track [20]. The context task was designed to investigate systems' capability to track context through a series of questions. However, as described in [6], 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. Although context processing has not been incorporated into the TREC QA track, the results from TREC 10 motivate more systematic studies of discourse processing for context question answering.

3.3 QA discourse

Another term that I will be using throughout the thesis is *discourse*. It is desirable and necessary to clarify the definition of discourse because discourse processing of the context questions and discourse modeling would be the focus of this study. In discourse analysis literature, researchers reserve the term *discourse* with a small *d* for stretches of language. It was defined as “a stretch of language consisting of several sentences which are perceived as being related in some way” [21] or “a continuous stretch of (especially spoken) language larger than a sentence, often constituting a coherent unit, such as a sermon, argument, joke or narrative” [3]. *Discourse* was also defined as “the piece of communication in context” [22]. Grosz and Sidner [23] defined *discourse* as “a piece of language behavior that typically involves multiple utterances and multiple participants” whose definition is more conversation-oriented. Discourse analysts classify discourse types according to the communicative function that they perform. To name a few, there are conversations of social interactions, email exchanges between friends and interviews on TV shows etc. Following this classification, I may as well treat the conversation between users and a computer operated QA system as a special kind of discourse.

Carrying different communicative purposes, certain discourses can be distinguished from other kinds in terms of their recurring patterns. Setting aside most other types of discourses, I notice that discourses situated in a classroom, an oral exam, a courtroom or a question answering system all contain questions and answers. However, the ways they are combined or presented vary. The building block of a QA sequence is strictly alternating question-answer pairs. Here I would like to define a QA discourse.

A *QA discourse* is a textual dialogue between a human user and a question answering system with at least two question answer pairs.

Example (1) (repeated from Chapter 1) is such an instance of a QA discourse with three consecutive question answer pairs.

(1)

Q1: Who was Tom Cruise married to?

(A1: Nicole Kidman)

Q2: What was her Broadway debut?

(A2: The Blue Room)

Q3: Who filed for divorce?

(A3: Tom Cruise)

A QA discourse is different from other types of discourses not only in terms of its forming structure, but also in terms of the special participants that are involved. Unlike normal conversations, in a QA setting it is always the human users that initiate, direct and terminate the conversation. The traditional Gricean Cooperative Principle [24] is not necessarily observed in a QA discourse. In this sense the discourse is different from normal human-human conversation discourse in that both parties are not responsible for being cooperative.

With some unique characteristics, does a QA sequence resemble other types of discourse? Now consider the following example, a sequence whose questions are jumbled up from an originally coherent QA discourse as shown in example (1).

(2)

Q1: What was her Broadway debut?

(A1: ?)

Q2: Who was Tom Cruise married to?

(A2: Nicole Kidman)

Q3: Who filed for divorce?

(A3: Tom Cruise)

This sequence is not coherent at all. Although question (2Q1) and (2Q2) are grammatically perfect, they are not coherent in that it is impossible for a computer or even a human being to interpret what *her* in (2Q1) refers to. On the other hand, the sequence of (1Q1) and (1Q2) makes more sense than the reordered example (2), and thus is a piece of coherent text sequence. In the current study, I assume that the TREC test sets and the data collected from the user study (discussed in Chapter 4) are question series consisting of coherent QA discourse. I also call them mini discourses.

3.4 Discourse modeling

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 [25] and Rhetorical Structure Theory [26]) and dialogues (e.g., Grosz and Sidner’s conversation theory [23] and Discourse Representation Theory Kamp1993). The NLP community has seen successful applications of discourse modeling based on the discourse structures that these theories assume. For example, in TREC 2005, Ahn et. al. [10] applied the Discourse Representation Structure (DRS) in question processing. In particular, Chai and Jin’s work [27] presents a semantic rich discourse representation, which provides a motivation for the work reported in this dissertation. In order to provide some background information on the work that is done in this study, I will briefly go over some of these theories in this section. First, I will

discuss the relationship between discourse structure and discourse coherence; then I will look at some of the well-established discourse structure theories; Finally I will introduce some recent computational work in identifying discourse structures.

3.4.1 Discourse structure

A common assumption made in natural language discourse modeling is that the discourses under processing are coherent. But why is a discourse coherent? Researchers, especially linguists and computer scientists have been looking at discourse structure seeking answers to this question. Now that I have determined that QA series can be treated as a kind of discourse, it is necessary to explore its structure for the purpose of building discourse models for processing context questions.

Traditionally, with implicit structures, a discourse puts individual sentences together to form a discourse structure. To have a better understanding of discourse structure, its representation, and their role in context question answering, I now review some discourse structure theories in linguistic literature.

3.4.2 Discourse structure theories

There are various discourse structure theories trying to help accomplish such tasks as mentioned above. More importantly, based on different ways of representing discourse structure these theories offer different account for the coherence of a discourse.

Kamp and Reyle [28]’s DRT (Discourse Representation Theory) has discourse representation structure (DRS) as the constructing units for a discourse. Each sentence in a discourse to be processed is dealt with using the context of a structure (i.e. DRS) resulting from processing the previous sentence. Not until the DRS for the whole discourse is built, can one interpret the discourse content represented by compositional semantics. The interpretation of a discourse is thus a dynamic and incremental process. One of the applications of DRS is to resolve reference for pronouns, because the DRS determines possible antecedents to anaphora. Yet one problem with this approach of anaphora res-

olution is that it both overgenerates and undergenerates the possible interpretations of pronouns. The problem occurs due to the notion of accessibility, which is so defined that basically anything would be accessible for the target anaphora.

Grosz and Sidner [23]’s Intentional Discourse Model is based on three structures: segment structure, intentional structure and attentional structure. In this segment-hierarchically structured model, a discourse is considered as a built-up of small information chunks (i.e. discourse segments), which consist of a group of sentences. Each discourse segment conveys a communicative intention or a purpose that contributes to the overall goal of the discourse. Local coherence¹ within a discourse segment is maintained through the operations of *centers* ². In the meantime, the focus, which is the attentional state of this segment, depends on its intentional structure. This intention-based theory therefore has semantic elements built into the model. The core notion of intentions imposes structure on discourse and thus makes discourse coherence possible. The major concern for this theory is how to infer intentions and it requires strong cognitive evidence and for the same reason, it is hard to compute and verify the attentional states. Centering Theory [29] upon which the work in Chapter 5 is based, is part of the Intentional Discourse Model. It relates the salience of entities in an utterance with the form of linguistic expressions and the local coherence of discourse.

Another semantically driven discourse structure theory is Mann and Thompson’s RST (Rhetorical Structure Theory). Originally it was developed for the purpose of computer-based text generation, yet it has been established in linguistics too. Similar to Grosz and Sidner[23], RST represents discourse as hierarchically organized text based on rhetorical relations between the discourse parts often consisting of two spans of texts: the nucleus (the claim span) and the satellite (the evidence span). Relations are further classified into two types based on the meaning: the *subject-matter relations* expressing the content of the subject matter of the text (such as *elaboration* and *concession*) and the *presentational relations* facilitating the presentational process of two text spans (such as *list* and *join*).

¹this notion will be discussed in Chapter 5

²this notion will be discussed in Chapter 5

However the inferring of relations is not easy and the relations defined by RST seem ad hoc as well. Note that this theory is descriptive rather than generative so the inferring of relations assumes the coherence of discourse. A coherent discourse is built in such a way that any two text spans have at most one relation between them and the recursive application of schemes (abstract pattern for constituent text spans and a specification of the relation between the spans) strings nucleus and satellite together and thus gets bigger text spans.

With different discourse structures, the discourse structure theories have their own strength and weakness in dealing with linguistic phenomenon and tasks. Take anaphora resolution as an example. Techniques of anaphora resolution based on DRT have been widely spread in computational linguistics, because DRT addresses itself specifically to the problem of anaphora resolution. The structure of RST, however does not help with anaphora resolution, because it is descriptive and the basic constructing structure is on the sentence level. With the three structures modeled in Grosz and Sidner [23], it is also possible to conduct anaphora resolution because the attentional states make it possible to separate focus (entities) from text spans out of a discourse.

Essentially different discourse structure theories explain discourse coherence using various discourse structures. Discourse representations based on the structure thus can be classified into two categories: coherence relation based and discourse entity based. Theories such as RST and DRT are discourse relation based, which focus on the relations that link discourse pieces together. Centering Theory in Intentional Discourse Model, on the other hand, is entity-based. A discourse maintains its coherence through the entities that appear in it. Theories from both camps have made contributions to the field of computational linguistics.

3.4.3 Computational efforts

To properly capture the discourse coherence, researchers in computational linguistics have made efforts to automatically identify discourse structures.

One computational approach to discourse structure and discourse coherence was first

described in Hobbs' [30] and later in [31]. Discourse coherence relations between segments of a discourse are used to characterize discourse coherence. Such coherence relations as *result* and *explanation* are provided to establish coherence in a computational inference system. For example in (3) the relation *explanation* holds between (3S1) and (3S2).

(3)

S1. Max fell.

S2. John pushed him.

The inference made by abduction for (3) is that the state or event asserted by (3S2) causes or could cause the state or event asserted by (3S1). Abduction³ inferences are made according to certain interpretation algorithm. In order to interpret a sentence, first of all, the logical form (LF⁴) of the sentence has to be proved via abduction. Proving involves the logical form of the sentence that is derived from syntax and the constraints that predicates impose on their arguments. According to Hobbs, the best explanation to which the inference is made would be the one with the least cost based on the weight assigned to the predicates. However, how to correctly get the logical form and to make the right assumptions is not trivial. Probabilities and heuristics have been introduced into this approach to ensure best proof.

Marcu's [32] work is a computational effort for automatically identifying discourse structure. Theoretically this practice relies on the RST discussed above. The implementation of a decision-tree based machine learning approach relies on three components: corpora annotation, discourse segmenter and shift-reduce action identifier. *Elementary discourse units* (edus) are defined as small chunks of texts. With annotation, *edus* and rhetorical relations are annotated manually according to Marcu [33]'s annotation protocol. A discourse segmenter is a decision-based algorithm that partitions the input text into *edus*. Features that models both local and global contexts (POS, boundary identification such as comma, discourse marker, etc.) are represented in a binary format. The shift-reduce

³Abduction is the process by which, from $(\forall x)p(x) \supset q(x)$ and $q(A)$, one concludes $p(A)$. One can think of $q(A)$ as the observable evidence, of $(\forall x)p(x) \supset q(x)$ as a general principle that could explain $q(A)$'s occurrence, and of $p(A)$ as the inferred, underlying cause of $q(A)$ (p.98). [31]

⁴Usually derived from surface structure, logical form is the level of representation where linguistic expressions are assigned a representation of meaning.

action identifier is to derive the discourse tree. The basic idea is to add an input sentence under consideration to the previous discourse tree and decide what the relevant rhetorical relation is. This is a rule-based process which also requires the knowledge of structural (such as number of trees in the stack), lexical (cue-phrase-like words such as *although*, *but*, etc.), syntactic (position of these phrases), operational (last parsing operations) and semantic-similarity-based (WordNet-based word similarity) features. As a result, local rhetorical relations between text spans and then a global discourse structure will be determined. This approach captures a lot of features therefore its performance on segmentation and tree construction is better than on rhetorical relation identification, which heavily relies on manual annotation.

Stolcke et. al.'s work [34] presents another corpus-based approach for identification of discourse structure of a conversation, which is treated as a hidden Markov model. Similar to Marcu's work, lexical, collocational, prosodic cues are used as features in probabilistic machine learning. Dialogue acts (DA, such as *statement*, *opinion*, *agreement*, etc.) that represent the meaning of an utterance at the level of illocutionary force [35] are predicted. Constraints on the likely sequence of dialogue acts are modeled via a dialogue act n-gram⁵. Similar to Marcu's work, this study is based on the manual annotation of conversation corpus. The probabilistic modeling is based on the features that sometimes seem to be random and without linguistic motivation.

Having reviewed some computational work based on discourse structure and discourse structure theories, we noticed that there was no specific work addressing the context question problem using linguistic theories of discourse structure or discourse coherence.

3.5 The hypotheses

Halliday and Hasan (Halliday, 1976) have classified five types of so-called text-forming devices: reference, substitution, ellipsis, conjunction and lexical cohesion. I believe that

⁵*n-gram* usually refers to an *n-gram* model, which is often used in natural language processing. An *n-gram* model predicts x_i based on $x_{(i-1)}$, $x_{(i-2)}$, ..., $x_{(i-n)}$. For more information, please check <http://en.wikipedia.org/wiki/N-gram>

a QA discourse will present the same characteristic as having some of the “text-forming devices”. That is, lexical ties are presented between questions. For instance, reference *he* in (4Q1b), and ellipsis in (4Q2b) are such cases.

(4)

Q1a Who was Tom Cruise married to?

Q1b When was he born?

Q2a Where is the highest point on earth?

Q2b Where is the lowest?

These lexical ties form lexical cohesion relationships. Therefore I proposed Hypothesis I: similar to other types of discourses, a QA sequence presents the characteristics of having lexical cohesion relationships, which contribute to the first level of coherence hierarchy.

The question that rises is: are these devices good enough to handle context questions in QA? Let us look at example (5)?

(5)

Q1: What is the name of the volcano that destroyed Pompeii?

Q2: How many people died?

Q3: Any pictures?

Example (5) is a coherent QA discourse, however there is no apparent lexical cohesion relations between (5Q2) and (5Q3). In both Chapter 5 and Chapter 6, I will present discourse models for context questions with and without lexical cohesion relations.

Cohesion relations are only concerned with the pattern of relating linguistic expressions, not with the patterns of content. Discourse coherence has been studied from different perspectives, as mentioned in the previous section. Different coherent relationships have been identified [31], [32]. It is well acknowledged that a “felicitous discourse must meet a rather strong criterion, that of being coherent” ([36], p.241). I believe that a QA discourse also observes various coherence relationships between questions or even among questions. According to Kehler (2004)’s classification these relationships are resemblance, contiguity and cause-effect with each having different subcategories. For example the relationship cause-effect is further broken into result, explanation, violated expectation and denial

of preventer. Following the classifications of Rhetorical structure theory (Mann, 2003), (6Q1) and (6Q2) are coherent in that they form an evaluation relationship, that is (6Q2 assesses the event in (6Q1).

(6)

Q1: How often does the U.S. government conduct an official population census?

Q2: Is the census confidential?

Similarly, in example (7), (7Q2) being the context question of (7Q1), there is a causal relationship between the two questions (7Q1) and (7Q2), in addition to the explicit repetition of word *volcano*.

(7)

Q1: What is the name of the volcano that destroyed Pompeii?

Q2: Why did the volcano erupt?

However, the problem presented by example (1) is still not solved, that is, how to deal with context questions that do not have any reference clue or implicit rhetorical relationships, such as (1Q3) *Who filed for the divorce?*. In this thesis, I will not discuss such discourse coherence relations as the *evaluation* and the *causal* relations in example (6) and (7). Instead, I would like to look at the discourse coherence from a different angle. Empirical data shows that the TREC context questions are on the same target/topic. This information may be essential for the follow-up questions that cannot be processed otherwise. Therefore I would like to seek relevant linguistic theoretical frameworks to help with the processing on top of the discourse cohesion relations. Centering Theory adopted in Chapter 5 and the definiteness hierarchy used in Chapter 6 are the results of such efforts. Centering theory helps to explain the local coherence for a QA discourse, while the definiteness hierarchy also explains how a QA discourse organizes discourse entities in a way that coherence is maintained. In using them, I tried to capture the discourse coherence relations of the QA discourse. Therefore I proposed Hypothesis II which states as follows: coherence relations are necessary to capture more semantic information in a QA discourse and it is at a level higher than lexical cohesion.

3.6 Summary

In this section, I have first clarified the notion of QA context and introduced the notion of QA discourse. In order to build discourse models for context questions, I believe that a series of questions can be treated as a kind of discourse, which shares some common features of other types of discourses. Then, I have reviewed some of the discourse structure theories that have been used in NLP research, hoping to provide insights and background information for my study on QA discourse processing. In the following chapters, I will mainly present different discourse models based on the theoretical frames to process context QA discourses.

CHAPTER 4

Data collection and platforms

4.1 Introduction

This chapter presents the data formats and implementation platform that are used for the current study. Section 4.2 introduces the architecture of the automated QA discourse processing system. Section 4.3 presents the context questions collected from my user study and the data from TREC 2004 and TREC 2005. Section 4.4 explains how the data was annotated for computational implementation of discourse models. Section 4.5 introduces both the user document collection and the AQUAINT corpus, the document collection upon which the document retrieval engine runs. Section 4.6 introduces the document retrieval engine used for the entire implementation. Section 4.7 concludes the chapter.

4.2 System overview

The prototypical system architecture for QA systems as it was presented in Chapter 2 has four¹ basic modules. The basic idea is to run the natural language question through the system and extract the exact answer to the question. Since my focus of the study is on the first component, that is, the question processing, I will only evaluate the implemented system at document retrieval level instead of at answer extraction level or passage retrieval

¹Some systems incorporated *document retrieval* and *passage retrieval* into one module.

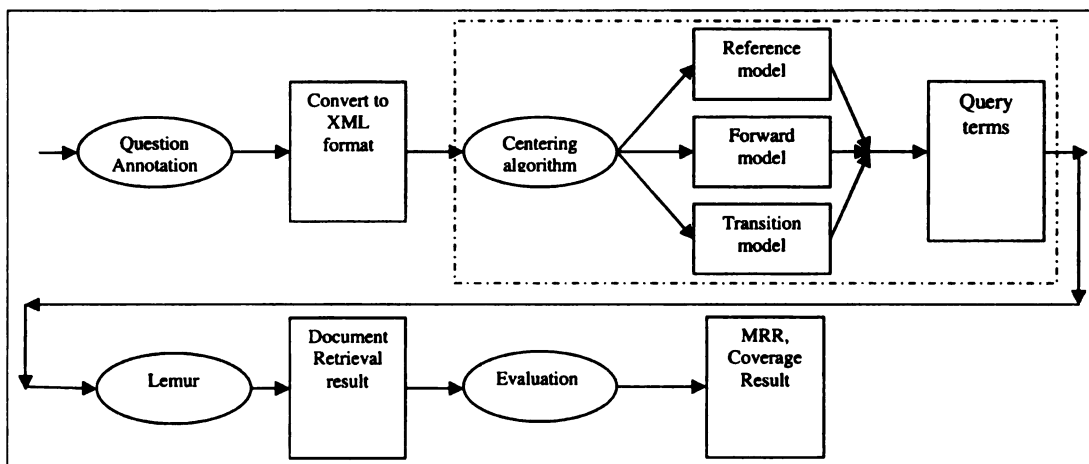


Figure 4.1. Flow chart of the centering-based automated system architecture

level. As is shown in figure 4.1, the dotted box represents the QA discourse modeling module² of the system. This module takes input in the form of context questions formatted in XML with relevant linguistic information annotated. The output of this module is the query terms resulting from the three models. Then, they become the input to the Lemur retrieval engine. The Lemur retrieval engine then returns a ranked list of documents. Based on the rank, evaluation is conducted by means of MRR and coverage, two evaluation metrics that will be discussed in Chapter 5.

4.3 Question collection

Two sets of data were used in the computational implementations in Chapter 5 and Chapter 6: context questions collected from a user study and factoid questions from the Text REtrieval Conference (TREC) 2004 and 2005.

4.3.1 User data

To support the investigation of different discourse models for processing context questions, a data collection effort was initiated through user studies. It was conducted through two

²which will be implemented in Chapter 5 and Chapter 6

steps: 1) design a user study and 2) administer a data collection session. In the first step, the following four topics were chosen for collecting context questions from subjects: (1) the presidential debate in 2004; (2) Hawaii; (3) the city of Pompeii; and (4) Tom Cruise. After the topics were decided, the key terms of these topics were entered to the *Google* search engine . A set of documents was manually selected from the search results. These documents contain relevant facts about each of these four topics.

In total, 22 subjects participated in my study. These subjects were recruited from the undergraduate students who took the Introduction to Artificial Intelligence class in Computer Science Department at Michigan State University. The data collecting session was conducted in a classroom environment and was finished within 20 minutes. Each subject was given both an instruction sheet and a copy of the document collection ³ prepared in the first step. On the instruction sheet, they were then asked to put him/herself in a position to acquire information about these topics. And they were asked to specify their information needs ⁴ and provide a sequence of questions (no less than 6 questions) to address that need. As a result of this effort, I collected 87 sets (i.e., sequences of questions) with a total of 522 questions. Example (1) shows a set of context questions collected on the topic *Tom Cruise*. One subject put down his information goal as to “know more about Tom Cruise’s personal life”.

(1)

Q1: What is Tom Cruise’s full name?

Q2: How long was he married to Nicole Kidman?

Q3: Where was Nicole born?

Q4: Who is Cruise dating now?

Q5: When was Penelope Cruz born?

Q6: How old was Kidman when she moved to Australia?

Specifically, the following issues during the data collection were emphasized:

³The document collection was provided because the subjects were instructed to ask questions, the answers to which were contained in the documents.

⁴i.e. what information do the users want to find about these topics? Please see the exact instruction in Appendix A.

- The answer to each question should come from a different document to enforce the use of the context for the subsequent questions. Users were provided with different paragraph segments, which originated with and were extracted from different documents on the same topic. Users were encouraged to ask questions to which these paragraph segments provide the exact answers. This way, the control that the answers to each question came from different document was guaranteed. This design is believed to be closer to a natural scenario 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.
- Since the goal of this research is to investigate discourse processing for coherent question answering, concise questions that depend on the discourse are of particular interest. Therefore users were asked to provide questions that are as natural and concise⁵ as possible.

This methodology of collecting context questions is motivated by TREC evaluation where the sequences of context questions were pre-defined by the National Institute of Standards and Technology (NIST).

4.3.2 TREC 2004 data

In addition to the data from the user study, TREC 2004 data and TREC 2005 data was also used for testing purposes on the discourse models developed in this research. As mentioned in Chapter 2, 65 sets of questions were finalized by NIST staff from the TREC assessors' original questions. The question types include factoid, list and other. For

⁵I only collected concise context questions rather than complicated questions, because I wanted to keep the user data consistent with the TREC data used in my research. The TREC 2004 and TREC 2005 data are mostly very concise and short. Please see the detailed data analysis in Chapter 5.

the purpose of evaluation, the final test sets aimed to ensure that there would be actual answers to the *factoid* and *list* questions and sufficient information for the *other* questions in the document collections⁶. Because of this criterion, a lot of original questions were eliminated. TREC 2004 question sets also avoided such questions that would specifically mention the answers of previous questions. Therefore strictly speaking, TREC 2004 question sets are not “true samples of naturally occurring user-system dialog” [7].

The main task in TREC 2004 QA track was to answer questions that were grouped by targets. In other words, each set of questions comes with a predefined target. Example (2) is taken from TREC 2004 where its target is *Jar Jar Binks*. The targets of the TREC 2004 test sets are listed in table 4.1. Of the 65 targets, 23 are about *person*, 25 are about *organization*, and 17 are about *thing*. There are a total of 230 factoid questions, 56 list questions, and 65 other questions. The number of questions per series ranges from 4 to 10.

(2)

Q1: What film introduced *Jar Jar Binks*?

Q2: What actor is used as his voice?

Q3: To what alien race does he belong?

Since TREC data was also designed to test system capability of answering list and definition questions, which are not the focus of this work, those questions will be omitted in the evaluation that is covered in Chapter 5 and 6. For TREC 2004 data, I only focus on the 230 factoid context questions in the analysis and evaluation. It should be noted that there were still 26 factoid questions that do not have answers in the corresponding document collection, that is, the AQUAINT corpus.

4.3.3 TREC 2005 data

Similarly, in TREC 2005’s QA main track, there were 75 sets of question series, targets of which are listed in table 4.2 and table 4.3. In addition to the target type *thing* (such as

⁶However, unfortunately there were still some questions that did not have answers in the document collections.

Table 4.1. TREC 2004 targets

S1 Crips	S2 Fred Durst	S3 Hale Bopp comet
S4 James Dean	S5 AARP	S6 Rhodes scholars
S7 agouti	S8 Black Panthers	S9 Insane Clown Posse
S10 Prions	S11 the band Nirvana	S12 Rohm and Haas
S13 Jar Jar Binks	S14 Horus	S15 Rat Pack
S16 cataract	S17 International Criminal Court	S18 boxer Floyd Patterson
S19 Kibbutz	S20 Concorde	S21 Club Med
S22 Franz Kafka	S23 Gordon Gekko	S24 architect Frank Gehry
S25 Harlem Globe Trotters	S26 Ice-T	S27 Jennifer Capriati
S28 Abercrombie and Fitch	S29 'Tale of Genji'	S30 minstrel Al Jolson
S31 Jean Harlow	S32 Wicca	S33 Florence Nightingale
S34 Amtrak	S35 Jack Welch	S36 Khmer Rouge
S37 Wiggles	S38 quarks	S39 The Clash
S40 Chester Nimitz	S41 Teapot Dome scandal	S42 USS Constitution
S43 Nobel prize	S44 Sacajawea	S45 International Finance Corporation
S46 Heaven's Gate	S47 Bashar Assad	S48 Abu Nidal
S49 Carlos the Jackal	S50 Cassini space probe	S51 Kurds
S52 Burger King	S53 Conde Nast	S54 Eileen Marie Collins
S55 Walter Mosley	S56 Good Friday Agreement	S57 Liberty Bell 7
S58 philanthropist Alberto Vilar	S59 Public Citizen	S60 senator Jim Inhofe
S61 Muslim Brotherhood	S62 Berkman Center for Internet and Society	S63 boll weevil
S64 Johnny Appleseed	S65 space shuttles	

F16, *Louvre Museum*, and *Virginia wine*), *person* (such as *Bing Crosby*) and *organization* (such as *United Parcel Service (UPS)*⁷ and *American Legion*), a new target type, *event* was introduced into the TREC 2005 data sets. Events could be presented as a nominal event such as *Preakness 1998* or a description event such as *1998 indictment and trial of Susan McDougal* or *France wins World Cup in soccer*. Of the 75 sets, 18 were *event*, 19 were *thing*, 19 were *organization* and 19 were *person*. The numbers show that NIST staff managed to keep the target types evenly distributed. There were total 362 factoid questions, 93 list questions and 76 other questions. For the same reason I mentioned above, only the 362 factoid questions will be included in the implementation and evaluation. Note that there were 30 factoid questions that do not have answers in the TREC 2005 data.

Therefore, if the TREC 2004 and TREC 2005 data are put together, there are 140 sets with total 592 factoid questions, 56 of which do not have actual answers. This no-answer factor will be considered in the evaluation for the work in Chapter 6.

4.4 Data format

Next, I will introduce the data format used in my research. NIST provided a single document for each year's QA test sets. They can be found at NIST's website⁸. Example (2) was taken from the TREC 2005⁹ question set document¹⁰, where all the information about the questions was encoded using XML tags.

What is XML then? XML stands for eXtensible Markup Language, which is a W3C¹¹ recommendation . XML is non-proprietary and extensible. Unlike natural languages, it is a computer language designed to describe data. In other words, it tells us what data it is by using tags. The tags are not predefined but defined at user's will. This language allows user to define his own tag, anything from a self-explanatory label in a natural language such as "this is a TREC question" to a sign such as "TREC-Q". Note that the purpose

⁷Most of the times, both the full name and the acronym of the organization, if there is one are given.

⁸<http://trec.nist.gov/data/qa.html>

⁹TREC 2005 questions can be found at <http://trec.nist.gov/data/qa/2005.qadata/QA2005.testset.xml>

¹⁰TREC 2004 questions can be found at <http://trec.nist.gov/data/qa/2004.qadata/QA2004.testset.xml>

¹¹World Wide Web Consortium

Table 4.2. TREC 2005 targets

S66 Russian submarine Kursk sinks	S67 Miss Universe 2000 crowned
S68 Port Arthur Massacre	S69 France wins World Cup in soccer
S70 Plane clips cable wires in Italian resort	S71 F16
S72 Bollywood	S73 Viagra
S74 DePaul University	S75 Merck & Co.
S76 Bing Crosby	S77 George Foreman
S78 Akira Kurosawa	S79 Kip Kinkel school shooting
S80 Crash of EgyptAir Flight 990	S81 Preakness 1998
S82 Howdy Doody Show	S83 Louvre Museum
S84 meteorites	S85 Norwegian Cruise Lines (NCL)
S86 Sani Abacha	S87 Enrico Fermi
S88 United Parcel Service (UPS)	S89 Little League Baseball
S90 Virginia wine	S91 Cliffs Notes
S92 Arnold Palmer	S93 first 2000 Bush-Gore presidential debate
S94 1998 indictment and trial of Susan McDougal	S95 return of Hong Kong to Chinese sovereignty
S96 1998 Nagano Olympic Games	S97 Counting Crows
S98 American Legion	S99 Woody Guthrie
S100 Sammy Sosa	S101 Michael Weiss
S102 Boston Big Dig	S103 Super Bowl XXXIV
S104 1999 North American International Auto Show	S105 1980 Mount St. Helens eruption

Table 4.3. TREC 2005 targets (cont')

S106	1998 Baseball World Series	S107	Chunnel
S108	Sony Pictures Entertainment (SPE)	S109	Telefonica of Spain
S110	Lions Club International	S111	AMWAY
S112	McDonald's Corporation	S113	Paul Newman
S114	Jesse Ventura	S115	Longwood Gardens
S116	Camp David	S117	kudzu
S118	U.S. Medal of Honor	S119	Harley-Davidson
S120	Rose Crumb	S121	Rachel Carson
S122	Paul Revere	S123	Vicente Fox
S124	Rocky Marciano	S125	Enrico Caruso
S126	Pope Pius XII	S127	U.S. Naval Academy
S128	OPEC	S129	NATO
S130	tsunami	S131	Hindenburg disaster
S132	Kim Jong Il	S133	Hurricane Mitch
S134	genome	S135	Food-for-Oil Agreement
S136	Shiite	S137	Kinmen Island
S138	International Bureau of Universal Postal Union (UPU)	S139	Organization of Islamic Conference (OIC)
S140	PBGC		

of XML is to describe textual data, therefore it is desirable to define meaningful names for the data of interest. Due to this feature of XML, the descriptive information stored by XML can be used across any platform. Another feature is that the data format of XML is reusable. Because XML is accepted and backed as an international standard, the format will be available for later access and processing.

With the evident advantages of XML, one would be eager to know how XML describes data. As a matter of fact, DTD (Document Type Definition) is used to define the building blocks of an XML document. DTD allows users to define the document structure with a list of legal elements. To avoid jumping too much into technical details, I now discuss the elements that were defined in the DTD for the TREC data and then I will present the customized DTD for the purpose of building my discourse models.

4.4.1 Original TREC data format

The DTD must reflect what part of the data is annotated. The first type of information being annotated is for the convenience of programming. Housekeeping information of questions such as question ID, question sequence boundary, answer to each question, etc. were annotated. The original data provided by TREC is shown in example (3).

(3)

```
<trecqa year="2005" task="main">
  <target id="66" text="Russian submarine Kursk sinks">
    <q><q id="66.1" type="FACTOID">When did the submarine
sink?</q></q>
    <q><q id="66.2" type="FACTOID">Who was the on-board commander of the
submarine? </q></q>
    <q><q id="66.3" type="FACTOID">The submarine was part of which Russian
fleet?</q></q>
    <q><q id="66.4" type="FACTOID">How many crewmen were lost in the
disaster?</q></q>
    <q><q id="66.5" type="LIST">Which countries expressed regret about the
loss?</q></q>
    <q><q id="66.6" type="FACTOID">In what sea did the submarine
```

```

sink?</q></qa>
      <qa><q id="66.7" type="LIST">Which U.S. submarines were reportedly in the
area?</q></qa>
      <qa><q id="66.8" type="OTHER">Other</q></qa>
</target>
.
.
.
</trecqa>

```

The information indicates the *year* of the TREC QA track (*2005*), the QA track task type (i.e. *main*), the question target (*Russian submarine Kursk sinks*), the target id(66), the individual question id(e.g. 66.1, 66.2 etc.), and the question type(*FACTOID*) in the series. The only information that is not encoded in the test sets is the target type. That information is given separately in another TREC document¹².

Another note on the data format is the numbering of the series. Since TREC 2004 has 65 sets, the TREC 2005 data sets continued the numbering from 66 till 140. Example (2) was actually the first set from the TREC 2005 data.

4.4.2 Customized data format

Besides the basic information (such as the target, the type, the ID etc.) of a question set, the second type of information is the linguistic information that needs to be encoded for the questions. There are two reasons to keep using XML as the annotation language to annotate such information as well. First, since the TREC 2004 and the TREC 2005 data were provided in the language of XML, it is wise to use XML to keep the consistency in that aspect. In doing so both datasets can be processed and evaluated using the same programs. Second, it is because of the self-explanatory, portable characteristics of XML discussed in the previous section. Next, I will present the customized data format for the user data and the TREC data. The goal of this study is to build discourse processing models for context questions based on linguistic knowledge. For the convenience of conducting evaluations on

¹²http://trec.nist.gov/data/qa/2005.qadata/05.target_type.txt

the models, I annotated relevant linguistic information manually so as to keep the research focused on modeling context questions rather than automatically identifying the linguistic information used for the modeling. In the following section, I will present the annotation schemes for the experiments discussed in Chapter 5 and Chapter 6, respectively.

In Chapter 5, the semantics of noun phrases is the focus. Special efforts were made to annotate noun phrase types, for example, pronoun, definite description, indefinite description etc. For the purpose of pronoun resolution, agreement constraints (i.e. number, gender and animacy) are annotated. The non-personal pronoun *it* is annotated to indicate whether it refers to an object, event, situation or other.

The grammatical roles of noun phrases were also annotated. The following grammatical roles are identified: possessor, subject, predicate nominal, object, indirect object, and adverbial prepositional phrase and other. In addition, the definiteness information of noun phrases was annotated. NP types include demonstrative, pronoun, definite description, possessive NP, proper names, indefinite NP. Finally adverbials of each question were annotated. The following adverbials are identified: location, time, reason, modifier and adverb.

Example (4) shows how an annotation of a question looks like. The annotation tag `q_id="q2"` shows that the question *why did Tom Cruise begin acting?* is the second question of a context question sequence. And the tag `prev_q="q1"` shows that its preceding question has the question ID `q1`. The answer tag `<a>` indicates that the answer to the question `q2` is in the document *QA880002-0003*. For some information, such as the topic, the annotation tells us that the topic of the original question starts from token 5 and ends at token 6, which points to the term "acting". The noun phrase *Tom Cruise*, which is annotated as starting from token 3 and ends at 4 is further annotated using tags *male*, *subject* and *proper name*. Moreover, in addition to the topic information, the question word *why* is annotated as focus. It is for the convenience of future work on question processing strategies that might use the information of question type.

(4)

`<q prev_q="q1" q_id="q2">`

```

<orig>Why did Tom Cruise begin acting ?</orig>
<a>QA880002-0003 </a>
<TOPIC start_tok="5" end_tok="6">
    <ENTITY sem_type="male" gram_role="subject">
        <PROPERTY prop_type="proper_name" start_tok="3" end_tok="4"/>
    </ENTITY>
</TOPIC >
<FOCUS start_tok="1" end_tok="1"/>
</q>

```

The 522 user study questions and the 230 TREC questions were annotated this way for the study in Chapter 5.

For the experiments conducted in Chapter 6, in addition to the semantic type, grammatical role and NP property information of an entity that were already annotated in the TREC 2004 and the user data, I added more detailed syntactic information to NPs, definite descriptions in particular. I annotated an entity ¹³in the way that its corresponding expression was split into premodifier, head and postmodifier. If any of the premodifier or postmodifier corresponds to an entity itself, that entity will be treated as a separate entity and thus annotated too. The discourse old/new status of definite descriptions is also annotated for evaluation purpose. Target was also annotated in the same fashion using the scheme. The semantic type of the target was annotated according to the information TREC provided. The following is an example:

(5)

```

<qa qaid="01" t="Crips" type="ORG">
<q_id="0">
<orig>Crips</orig>
<TOPIC start_tok="1" end_tok="1">
    <ENTITY sem_type="object-plural" gram_role="subject">
        <PROPERTY prop_type="proper_name" start_tok="1" end_tok="1"/>
    </ENTITY>
</TOPIC>

```

¹³Event is also decomposed into entities and actions (predicate verbs).

```

</q>
<q.id="q1">
<orig>When was the first Crip gang started ? </orig>
<TOPIC start_tok="7" end_tok="7">
  <ACTIVITY>
    <ENTITY sem.type="person-singular" gram_role="subject">
      <PROPERTY prop.type="definite_des" start_tok="6" end_tok="6"/>
      <CONSTRAINT constraint_type="premodifier" start_tok="4"
end_tok="5"/>
      <DEF_TYPE def.type="new"/>
    </ENTITY>
    <ENTITY sem.type="object-singular" gram_role="possessor">
      <PROPERTY prop.type="proper_name" start_tok="5" end_tok="5"/>
    </ENTITY>
  </ACTIVITY>
</TOPIC>
</q>

```

Based on the annotation scheme, the 230 TREC 2004 factoid questions and 362 TREC 2005 factoid questions were annotated accordingly.

4.4.3 Annotation validation

After the annotations were done, two validations were conducted. Firstly, the syntax of each annotation was checked. All the data were manually annotated using Oxygen¹⁴ XML editor 5.1. With the help of the validating tool in the software, the annotation was checked against the DTD to ensure that the tagging and hierarchical relations between tags were correct.

Secondly, Stanford Lexicalized Parser¹⁵(lexparser) version 1.4, a probabilistic lexicalized natural language context free grammar parser, was used. All the data was run through lexparser on a Unix system to validate the linguistic information that was annotated (i.e. the NP type and the grammatical role of each entity). Also some of the statistics presented

¹⁴<http://www.oxygenxml.com>

¹⁵<http://www-nlp.stanford.edu/software/lex-parser.shtml>

in the data analysis section was obtained by using the lexpaser (such as the number of the definite descriptions in the data). The following is an example of the resulting parse tree for the question *why did he start acting?*

(6)

```
Q: Why did he start acting?
(SBARQ
  (WHADVP (WRB Why))
    (SQ (VBD did)
      (NP (PRP he))
        (VP (VB start)
          (S
            (VP (VBG acting))))))
    (. ?))
```

The parse tree not only shows the syntactic structure of the question, but also indicates the NP types of noun phrases. The parsed tree tells us that *he* is an NP and also it is a pronoun (i.e. (NP (PRP he))). The main verb of the question is the VP at a higher level, that is, the action verb *start*. This information is important because the determination of question topic relies on the identification of the predicate verb of the main sentence for each question. Also the parser labels determiners, which helps to identify definite descriptions. This information is used in Chapter 6¹⁶

4.5 Document collections

In this section, I will introduce the document collections that were used in the document retrieval component in the study, first the user study documents, and then the AQUAINT corpus.

¹⁶Apparently, the parser tree outputs less rich than my annotation. In a fully automated system, the annotated information could be obtained by using other NLP software (such as taggers, name-entity recognizer etc.) and feature heuristics.

4.5.1 User study documents

As I have mentioned in section 4.3.1, the 22 subjects were provided with a document for each topic. Each paragraph of the document was chosen from a Web search through *Google*¹⁷. The following is a part extracted from the *Tom Cruise* document that was provided to the subjects.

(7) Tom Cruise

- That million megawatt smile has helped Tom Cruise reach the pinnacle of his profession and stay there. He's a down-to-earth movie star with huge box-office hits under his belt such as *Top Gun* and *Mission: Impossible*.
- Thomas Cruise Mapother IV was born on the 3rd of July, 1962 (eerily similar to his film *Born on the 4th of July*), in Syracuse, New York. He was the only boy of four children. Since his father was an electrical engineer, he must have inherited his love for acting from his mother, who was a teacher.
- His acting career really began because he injured his knee in high school and was forced to quit the amateur wrestling team.
- His popularity took a beating in movies like *All the Right Moves* in 1983, followed by *Legend* in 1985. Cruise's career began to solidify during his signature hit of the 1980s, *Top Gun*.

For each of these paragraphs, I created a pseudo document that contains it and filled up the document with more relevant texts on the topics using some of the *Google*¹⁸ searching results. I ended up having 52 such documents: 11 pseudo documents on *Hawaii*, 15 on *Pompeii*, 13 on *Tom Cruise*, and 13 for the *2004 presidential debate*. All these documents were combined into one file using the document format that will be introduced in Section 4.5.3.

¹⁷<http://www.google.com>

¹⁸I used the same topic as key word input to the Google search user interface (<http://www.google.com>).

4.5.2 AQUAINT corpus

The AQUAINT¹⁹ corpus was prepared by the Linguistic Data Consortium hosted at the University of Pennsylvania for the AQUAINT Project, and is used in official benchmark evaluations conducted by NIST. This corpus consists of newswire text data in English, drawn from three sources. The data is distributed on two CDs. CD1 (about 1.37GB) contains all the text data from the New York Times News Service. CD2 (about 1.67GB) contains all the text data from the other two sources, that is, the Xinhua News Service (from People's Republic of China), and the Associated Press Worldstream News Service. Within each source, the data files are organized by year, and within each year, there is one file per calendar day and the file name reflects the source and date (e.g. 19980605_NYT.html). Total amount of data is about 3.04GB with 3344 files.

4.5.3 Document format

There are two sets of document collections used in the research. The main document database is the AQUAINT corpus of English News Text. The other set is the document collection prepared for the user study. In order to run document retrieval engine on these data, both datasets have to use the same data format which will be discussed in the following paragraphs.

The format of each document file is rather simple. All data files contain a series of news stories as a concatenation of DOC elements (i.e. blocks of text bounded by <DOC> and </DOC> tags). Example (8) shows a segment of such a data file.

(8)

```
<DOC>
<DOCNO>NYT19980601.0001 </DOCNO>
<DOCTYPE>NEWS STORY </DOCTYPE>
<DATE_TIME>1998-06-01 00:02 </DATE_TIME>
<TEXT>
```

¹⁹<http://www ldc.upenn.edu/Catalog/docs/LDC2002T31/>

<P>

NEW YORK_ Kenneth Joseph Lenihan, a New York research sociologist who helped refine the scientific methods used in criminology, died May 25 at his home in Manhattan. He was 69.

</P>

</TEXT>

</DOC>

The user documents on the four topics were first collected without any special markup, therefore they were converted using the format of AQUAINT data. They were then concatenated to one file, which later was merged to the AQUAINT corpus for the Lemur retrieval engine to form a data index.

4.6 Document retrieval engine

The reason for introducing the document retrieval engine Lemur is that Lemur's indexing function is able to index the AQUAINT documents and the document collection prepared for the user study in TREC format. Lemur 2.0.3 was used as the document retrieval engine for this research. The Lemur Project is sponsored by the Advanced Research and Development Activity in Information Technology (ARDA) and by the National Science Foundation. The Lemur toolkit supports indexing of large-scale text database and the implementation of retrieval systems based on a variety of retrieval models. The Lemur system can run both under Unix/Linux operating systems and under Windows. In this research, the Lemur toolkit was installed on a Linux server where all the programs for different discourse models ran.

Although Lemur supports many useful features for researchers in language modeling and information retrieval, indexing and retrieval are the two most useful for the current research in Question Answering. Lemur is able to form a sequence of documents for indexing after tokenization, stemming and removal of stop words. The retrieval engine then takes converted query terms as input and returns ranked documents as output. The

query terms have to be converted to the TREC document format. Example (9) is an input in the converted format for the retrieval engine where the query terms are the same as the original question. In the case of modeling context questions, the query terms will be whatever terms that an original question results in, after it runs through a discourse processing algorithm.

(9)

```
<DOC>
<DOCNO>1394 </DOCNO>
<TEXT>
In what country did the game of croquet originate?
</TEXT>
</DOC>
```

Although Lemur supports passage retrieval, cross-language retrieval and other retrieval technologies, the document level ad hoc retrieval function was chosen for the discourse model evaluation and analysis in this work. Example (10) is an output for document level retrieval. The most important information of the output is the rank. The document that contains the answer to the question is supposed to be ranked higher than other documents which may contain relevant information about the topic but not the answer. In the later chapters, I will discuss the evaluation metrics used for the study, which involves the ranking of the returned documents.

(10)

query no.	document no.	rank	score
1394 Q0	NYT19980813.0364	1	109.842
1394 Q0	NYT19980812.0066	2	102.185
1394 Q0	NYT19980704.0174	3	87.2226
1394 Q0	NYT19980626.0194	4	70.6496
1394 Q0	NYT19980813.0462	5	64.5089
1394 Q0	NYT19980813.0457	6	64.5089

1394 Q0 NYT19980813.0460 7 64.5089
1394 Q0 NYT19980813.0459 8 64.5089
1394 Q0 NYT19980707.0070 9 47.3473
1394 Q0 NYT19980813.0240 10 46.6417

4.7 Summary

In this chapter, I first presented the data that was used in the context question discourse modeling. Then I proceeded to present the data formats and explain the question annotation details. Relating to data processing, other resources such as Stanford leparser, AQUAINT corpus, and the Lemur retrieval engine were also introduced to enable a better understanding of the technical implementation of the research.

CHAPTER 5

Centering-based discourse modeling

5.1 Motivation

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 TREC [6]. To better address user information needs, recent trends in QA research have shifted towards complex, context-based, and interactive question answering [6], [37] and [38].

Chapter 2 recounts how National Institute of Standards and Technology (NIST)¹ initiated a special task on context question answering in TREC 10 [6], which later became a regular task in TREC 2004 [7] 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?

¹<http://www.nist.gov/>

Q3: Is it endangered?

Q4: Any other endangered species?

This QA process is considered coherent because the questions are not isolated, but rather evolving and related to serve a specific information goal. I treat the question sequence as a mini discourse² in which each 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 description *the state fish* in (1Q2), pronoun *it* in (1Q3), and the indefinite noun phrase *any other endangered species* in (1Q4). In other words, there tend to be relationships between successive questions that users ask QA systems. However, it is still not known how best to make use of these relationships to facilitate answer retrieval. This chapter takes up this issue by considering how the relationships between successive questions can improve the query expansion step in QA.

Toward an answer, two linguistic levels of representation are relevant. These two levels would capture two kinds of relationships between questions in a QA mini-discourse.

The first level is called “lexical cohesion”. This level would capture the lexical cohesion relationships between questions, mainly, the repetition of same/similar words or reference to same entity. Halliday and Hasan [39] have classified five types of so-called text-forming devices: reference, substitution, ellipsis, conjunction and lexical cohesion. I believe that a QA discourse presents the same characteristic as having some of the “text-forming devices”. That is, lexical ties are presented between questions. For instance, in (2), pronoun reference *he*, repetition of *Dublin*, synonyms *film* and *movie*, and ellipsis in (2Q4b) are used in the context questions.

(2)

²See discussions in Chapter 3

Q1a: Who was Tom Cruise married to?

Q1b: When was *he* born?

Q2a: What river runs through *Dublin*?

Q2b: What is the most famous export for *Dublin*?

Q3a: Where was the *movie* “Somewhere in Time” filmed?

Q3b: What other *films* were shot there?

Q4a: Where is the highest point on earth?

Q4b: Where is the *lowest*?

Similar to other types of discourses, sequences of questions in QA have lexical cohesion relationships, which contribute to the first level of representation. This level is motivated by a psychological notion called *lexical entrainment*. Lexical entrainment refers to linguistic adaptation presented in natural speech where people repeat the same kinds of words, given a choice.

Lexical entrainment has been observed in natural conversations and psycholinguists have conducted various experiments trying to explain lexical entrainment. Pickering and Garrod’s [40] experiment results show that in conversations, speakers intend to use lexical entrainment for communication ease and efficiency. Historically, passing references during conversation is regarded as one of the explanations of lexical entrainment because referring is a collaborative process [41]. Brennan ([42], p.41) claims that in conversation, “people achieve conceptual pacts, or shared conceptualizations, which they mark by using the same terms.”. Lewis [43] points out in a dialogue that speakers play a game of cooperation, for instance, they use similar words referring to the same entities. Empirically, lexical entrainment results in the lexical ties that are exemplified in (2) (i.e. repetition for the “same word” use, synonyms for the “similar word” use, and reference for the “same entity” use). Moreover, Pickering and Garrod [40] claim that once dialogue is studied, it displays properties of entrainment at all levels (semantic, syntactic, phonetic, and others). Syntactic coordination/alignment examined by Branigan et. al. [44] is an example at the syntactic level. However, my study this part of my study only concerns with the lexical level. The discourse models that are proposed in this chapter should at least be able to

capture the information reflected by the *lexical cohesion* relationships.

The second level of linguistic representation “discourse coherence” is proposed as the other level of representation to capture the topic relationships between questions. I postulate that context questions are easier to process if a system can identify the key information that a discourse carries. From empirical experience, I notice that topic of a QA discourse may stay the same or shift between questions. The topic information may be essential for the follow-up questions that cannot be processed otherwise. For example in (3)

(3)

Q1: Who was Tom Cruise married to?

(A1: Nicole Kidman)

Q2: What was her Broadway debut?

(A2: The Blue Room)

Q3: Who filed for divorce?

(A3: Tom Cruise)

Tom Cruise and Nicole Kidman are two entities involved in the first two questions, and intuitively the information on them are the common ground knowledge, or potential topics for the third question, which can be used to reform (3Q3) to a more explicit question “Who filed for divorce, Tom Cruise or Nicole Kidman?”, which will greatly increase the chance of hitting the right answer for an IR system. Therefore the “discourse coherence” level attempts to identify the topicality of a discourse, which I believe is central to the deep understanding of a coherent QA discourse.

Given the two levels of representation, how do they help with query expansion part of IR systems? In answering this question, I find that Centering Theory is a very appropriate framework for solving the context QA problem. Firstly, based on lexical cohesion relationships, I hope to improve IR by adding resolved anaphors or terms used in earlier questions. Among the five types of lexical ties, pronominal reference will be my focus³. Centering Theory is, in part, a theory of pronominal reference uses, therefore it can pro-

³Other types of lexical relationships such as ellipsis will not be discussed in this study.

vide pronoun resolutions for context questions. Secondly, based on topic relationships, I hope to improve IR performance by selectively adding terms for query expansion by detecting topic change in the context QA discourse. Centering Theory, again, can fulfill the task because it also formalizes whether the topic stays the same or not.

In this chapter, I present a linguistic knowledge driven approach that aims to tie the lexical cohesion level with the discourse coherence level together based on Centering Theory [29]. This part of the study 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. The models are: (1) an anaphora 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 the current investigation, rather than a complete end-to-end study, I focus on discourse processing of questions for query expansion. Given a context question, the 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. 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, I evaluated the three models concerning their performance in document retrieval on two data sets: the data collected in the user studies and the data provided by the 2004's TREC. 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 the presidential election in 2004). The transition model and the forward model consistently outperform the reference resolution model.

This chapter focuses on modeling discourse processing for context questions and it is organized as follows. Section 5.2 presents a review on Centering Theory. Section 5.3

describes the three models for discourse processing based on Centering Theory. Finally section 5.5 presents the empirical evaluation and discusses the potentials and limitations of these models. Section 5.6 summarizes the chapter.

5.2 Centering Theory

In this section, I will spend some time reviewing CT basics. It is necessary because any possible misunderstandings of the theory would cause mistakes in the empirical implementations.

Centering Theory is an intellectual work that has been evolved and developed since the 1970s. It is well-established and continues to ignite various research interests. Centering Theory and the centering framework discussed in Grosz, Joshi and Weinstein 1995 [29] are developed from three sources:

- the early centering theory, which includes: the unpublished manuscript from Grosz, Joshi and Weinstein [45]), computational application in dialogue systems ([46], [47] [48], [49], and [50]);
- the computational theory of definite anaphora understanding and interpretation, and focusing algorithm in capturing local discourse coherence ([51], [52], and [53]);
- the relationship between the computational inference load and change of focusing state ([54],[55], [56], [45]).

Grosz, Joshi and Weinstein's work [29] integrates the three strands of work and presented the framework of centering. As mentioned in Chapter 3, Grosz and Sidner's work [23] provides a theory of discourse structure in which centering is a part. As a computational model for discourse interpretation, Centering Theory aims to identify the mechanism of how a discourse maintained its local coherence (within a discourse segment) using various referring expressions.

Centering Theory has been established as a linguistic theory and computational model that relates the local focusing of attention, inferencing complexity and the linguistic re-

ferring expressions in a discourse segment. It helps explain some linguistic phenomena that other semantic and inferential theories would not be able to explain. For instance, in example (4) (from [12], p.692)),

(4)

S1: John saw a beautiful Acura Integra at the dealership.

S2: He showed it to Bob.

S3: He bought it.

Centering Theory predicts that *he* in (4S3) prefers the interpretation of being John instead of Bob, because this would result in less inferential load and make this mini discourse more coherent. According to CT, which will be discussed more in a moment, John in (4S1) and the discourse entity co-specified by *he* in (4S2) are the centers of utterance (4S1) and (4S2), respectively. If the pronoun *he* in utterance (4S3) is interpreted as entity Bob instead of John, the center of (4S3) would shift to be Bob. This would result in a less coherent discourse. And according to the CT claims (to be discussed later) it will also end up with more inferencing and processing efforts for readers to understand the utterances. Intuitively the discourse would be more coherent if *he* in (4S3) is interpreted as John than if it is Bob, therefore CT explains why a different interpretation of the referring expression would affect the local coherence of discourse.

5.2.1 Grosz, Joshi and Weinstein (1995)

After the unpublished draft of Grosz, Joshi and Weinstein ([45], henceforth GJW86), Grosz, Joshi and Weinstein provided a series of work on centering in 1995 (henceforth GJW95). It summarizes some early work, clarified some of the early claims, and unfolds a whole framework of centering. Also it opens up various research directions for further exploration.

In GJW95, Centering Theory is formulated as a theory regarding focus of attention, referring expressions, and local coherence within a discourse([29]). The major claim that CT has is that discourse segments that keep mentioning the same discourse entities are more coherent than those in which different entities are mentioned. GJW95 formulates

this claim as follows: in a discourse segment, each utterance has a unique entity⁴ which functions as a linking device to the previous utterance. The mechanism built upon this core entity explains the local coherence through a set of hypotheses, rules and constraints.

5.2.2 Terminologies and definitions

In order for the readers to fully understand CT, it is necessary to clarify some of the terminologies and their definitions. These terms are *local coherence*, *focusing/attentional state*, *center* and *realization*.

The term *discourse coherence* in centering framework actually refers to *local coherence* as opposed to *global coherence*, therefore 2 levels in a 3-level discourse structure (Grosz, Joshi and Weinstein [56], henceforth GJW83). In addition to an attentional structure and an intentional structure, GJW83 specifically describes a linguistic structure which mainly concerns the actual organization of sentences. Within this linguistic structure, a discourse consists of smaller chunks of discourse, named *discourse segments*. The global coherence and local coherence are thus described in terms of discourse segments. Global coherence is the “coherence with other segments in the discourse”. Local coherence is the “coherence among the utterances in that segment” ([29], p.204).

“Local coherence refers to the ways in which individual sentences bind together to form larger discourse segments. It depends on such things as the syntactic structure of an utterance, ellipsis, and the use of pronominal referring expressions [Sidner, 1981]” ([56], p.44)

Corresponding to these two levels of coherence, there are two levels of focusing: global focusing and local focusing. It is local focusing that becomes centering. The notion of focusing has been associated with different terminologies. Focusing is proposed into CT by Sidner [51, 52, 53], later replaced by the term *attentional state* in Grosz and Sidner ([23]). GJW95 describes it as “an abstraction of the focus of attention of the discourse participants”, participants being either speakers/writers or hearers/readers for a discourse

⁴It is named as a backward looking center.

of either a dialogue or a written text. In other words, focusing tracked speakers/hearers' attention on discourse entities. Grosz and Sidner ([23],p.175) describe focusing as "inherently dynamic, recording the objects, properties, and relations that are salient at each point in the discourse". GJW95's description emphasizes more on the cognitive status of the discourse participants instead of on discourse objects focused on by the participants. I would like to clarify the notion of *local coherence* because in my work, a context question sequence is treated as a mini discourse and the discourse coherence hypothesized in Chapter 3 between/among context questions is local coherence.

Centering Theory claims that local coherence is captured by the operations on centers. A center is "regarded as an ascription of a property to a single individual" ([54], p.435). This description implies that a center is not a linguistic expression but an entity that the expression refers to. GJW95 describes centers as those entities serving to link utterances together coherently. In order to distinguish the linguistic expressions and the centers they refer to, I will italicize the linguistic expressions while discussing the examples in this chapter.

An essential term for understanding centers is *directly realizes*. It is closely related to the semantics of centers. In GJW95, the notion of *directly realizes* is stated as follows.

"U⁵ directly realizes c if U is an utterance of some phrase for which c is the semantic interpretation. *Realizes* is a generalization of *directly realizes*" ([29], p.209).

It is through *directly realizes* that centers relate to linguistic expressions. The realization relation makes it possible to have computational representations for centers in that the expressions are associated with the corresponding discourse entities for valid semantic interpretation. Simply, it allows computational processing on discourse entities. For us, it provides a way to represent entities from QA discourse.

Having discussed some of the important terminologies used in GJW95, I now switch to investigate the constraints and rules that GJW95 presents for the CT framework.

⁵U stands for an utterance, and c stands for a center.

5.2.3 Constraints and rules

In this section, the constraints and rules described in GJW95 will be discussed in detail. They are fundamental to centering algorithms and centering-based applications.

GJW95 differentiates *backward looking center* (represented as C_b) from *forward looking center* (represented as C_f) according to its discourse property. The backward looking center is the linking device between an utterance and its preceding utterance. Therefore, it is considered crucial in keeping local coherence. From a cognitive aspect, the repeated mention of this entity makes it easier for the discourse participants to access in memory, therefore more “retrievable into consciousness” in Chafe [57]’s term. In addition to the backward looking center, Centering Theory also provides the insight on the preferences for interpreting subsequent discourse through the speculation of forward looking center and preferred center.

GJW95 defines the forward looking centers⁶ as a set of entities mentioned in an utterance (represented as U_n). They are what the succeeding utterance U_{n+1} may be linked to. The term *preferred center*⁷ (represented as C_p) was introduced by Brennan, Friedman, and Pollard [1] to represent the highest-ranked member in this set. It is only when the preferred center of U_n is realized in U_{n+1} , is it defined as the backward looking center of U_{n+1} . Note that the preferred center only predicts the backward looking center in U_{n+1} in the sense that, it does not have to be realized in U_{n+1} . CT claims that local coherence is maintained if the preferred center of U_n is realized in U_{n+1} and thus becomes the backward looking center C_b of U_{n+1} .

To sum up, CT assigns different discourse status to all the entities among which the repeatedly mentioned one, i.e. the backward looking center is participants’ focus and most accessible in memory, therefore gets the most discourse prominence; all the candidate entities in U_n for the backward looking center are the forward looking centers; and the highest ranked forward looking center is preferred to be the backward looking center of

⁶Forward looking centers are initially defined as those entities that are arguments of the main predicate in JK79 and JW81.

⁷It is roughly corresponds to Sidner (1983)’s *expected focus*.

U_{n+1} , therefore is named the preferred center.

Walker, Joshi and Prince ([58],p.3) state that the “distinction between looking back to the previous discourse with the C_b and projecting preferences for interpretation in subsequent discourse with the C_p is a key aspect of centering theory”. After different centers are discussed, I turn to examine the specific constraints and rules that CT forces upon the operations of the centers.

The first constraint states that there is only one backward looking center in an utterance. Following JW81, GJW95 assumes that backward looking center of an utterance $C_b(U_n)$ is a singleton. Although there is no argument for this statement offered in either JW81 or GJW95, psycholinguistic evidence is shown that there is not more than one C_b ([59]; [60]).

Another constraint states that within the set of forward looking centers, every entity must be realized. This brings back the important notion of *realizes*. GJW95 mentions that the precise definition of U (utterance) *realizes* c (center) depends on specific semantic theory that one adopts. Relying on situational semantics [61] [62], Grosz, Joshi and Weinstein (1986)’s work [45] defines the *realize relation* as follows:

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.

By this definition, discourse entities could be realized as pronouns, zero pronouns, explicitly realized discourse entities (such as those directly realized) or implicitly via inferrables ([63]; [64]). It should be noted that in addition to discourse entities, discourse relations can also be realized ([56]). The relation *direct realizes* is a specialization of the relation *realizes*. I should point out that in this study, I will only focus on entities explicitly realized as pronouns, excluding those realized as zero pronouns or implicitly realized. Within the *direct realizes* relation, I will only emphasize on semantic reference, excluding pragmatic reference.

The last constraint states that forward looking centers could be ranked. The idea is that entities in U_n have different likelihood to be the backward looking center of U_{n+1} . This

constraint is critical to centering. Since the preferred center is the highest ranked entity among all the forward looking centers, the identification of the preferred center therefore depends on what ranking scheme a CT implementation employs. Treated as a parameter in CT framework ([65],[66]), the ranking of forward looking centers has received lots of attention when researchers apply CT cross-linguistically. The original CT proposals that lead to GJW95 are based on the observation of English and used grammatical role and pronominalization to rank the forward looking centers. In fact, various factors that may influence the ranking have been discussed from language to language, for instance, topic markers on NPs in Japanese ([67],[68]), features on the verbs in Italian ([69],[70]), word order in subordinate clauses in German ([71]), word order in modifier clauses in English ([60]), thematic relations in Turkish ([72], [73]), etc.

In addition to the three constraints, GJW95 also formulates two rules. Rule 1 is actually a pronominalization rule, which states as follows:

Rule 1: If some 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.

This rule implies that:

- this rule does not apply to entities that are realized in U_n but not in U_{n+1} ;
- this rule does not apply to utterance U_{n+1} that does not have any pronouns (this is possible because C_b could be realized by a proper name or a definite description);
- if in utterance U_{n+1} , there are multiple pronouns realizing entities from its preceding utterance U_n , one of them must realize the backward looking center C_b ;
- if there is only one pronoun, then that pronoun must realize the C_b .

The last two implications indicate that, as long as the C_b of utterance U_n is realized as a pronoun, the utterance U_{n+1} does not preclude using pronouns for other entities. In general, this rule explains how the use of pronouns could capture local coherence and at the same time, indicates that the C_b is often pronominalized.

Now, let us look at the second rule, which can be regarded as a transition rule. Originally, three types of transition relations are defined across two adjacent utterances: *continuation*, *retaining* and *shifting*. Brennan et al. [1] and later work extend the transitional state *shifting* to *smooth shifting* and *rough shifting*.

The definition of transition states mentioned in the rule is summarized in Table 5.1. Two factors are important in defining these states: (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 5.1. Extended transition states (Adapted from Brennan et al.[1])		
	$C_b(U_{n+1}) = C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough shift

If both (1) and (2) hold, then the two utterances are related by a *continue* transition, which indicates that the speaker has focused 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 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.

According to the summarization in Table 5.1, two utterances are most coherent if they share the same C_b and the C_b is the same as C_p , least coherent if neither they share the same C_b nor the C_b coincides with the C_p (rough shift). Meanwhile the last constraint (i.e. the forward looking centers are ranked.) in centering is critical, because the ranking of forward looking centers will determine the $C_b(U_{n+1})$ through identifying $C_p(U_n)$, thus influencing the transition relation that holds between two utterances.

Rule 2: Sequences of continuation are preferred over sequences of retaining;

and sequences of retaining are preferred over sequences of shifting.

Rule 2 is a preference constraint on the ordering of transition relations. This rule provides a coherence measurement through which one could explain why some discourse segments are more coherent than others. This preference rule has been used in measuring coherence degree in some applications ([74]). Unlike grammatical rules, constraint rules such as rule 2 can sometimes be violated resulting in an acceptable utterance sequence.

Based on the rules and constraints, various Centering Theory based algorithms have been developed over the years for different computational purposes. I will first briefly go over some of them.

5.2.4 Centering algorithms

There have been various algorithms based on centering framework aiming to fulfill discourse processing tasks, such as Kameyama [68], Brennan, et al. [1], Walker [75], Baldwin [76], Kehler [77], Walker et al. [78], Strube and Hahn [79] etc.

Kameyama [68] adds another structural parallelism constraint to GJW86. Briefly speaking, subject pronouns prefer subject antecedents and non-subject pronouns prefer non-subject antecedents. Brennan et al.[1] argue that this constraint is the consequence of the ranking scheme defined by grammatical functions and the preference for continuation over retaining. Walker [75] evaluates Brennan et al. [1]’s algorithm to find the referents for pronouns in naturally occurring texts, and also proposes rules to segment discourse since centering is intended to operate within a discourse segment. Kehler [77] suggests an anaphoric processing model that helps process ellipsis and anaphora resolution. In particular, this model provides explanation and evidence for ranking the forward looking centers according to their grammatical roles in the centering framework. Walker et al. [78] intend to investigate the role of centering in interpreting anaphoric expressions in Japanese. Agreeing with Walker et al. [78] that the ordering of the list of forward looking centers is the most important single construct of the centering model, Strube and Hahn [79] propose another ordering scheme that depends on the functional information

structure of utterances instead of the ranking principle depending on grammatical roles.

Next, I will summarize Brennan, et al. [1]’s work in which a centering algorithm to resolve third-person pronouns was proposed. This algorithm will be utilized for modeling the QA discourse. Brennan et al.’s work follows the two rules and three constraints mentioned above. Given the entity-based assumption in Centering Theory, the problem of how to rank the entities has been discussed extensively in the literature. The ranking scheme relying 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., $\text{subject} > \text{object(s)} > \text{other}$) [29].

Brennan et al.’s algorithm can be described as having four steps. They are listed as follows:

- 1. For each utterance, generate all the possible combinations of C_f and C_b in terms of reference assignment;
- 2. For each possible combination from step 1, apply various constraints to get the compatible combinations;
- 3. Rank the candidates by transitional orderings;
- 4. Assign the reference to the top candidate.

Generally, possible candidates for backward looking center will first be constructed, then be filtered and finally be classified and ranked. At the construction stage, referring expressions are identified and ordered by grammatical relations. Note that the ordering implies that the salience of discourse entities can be ranked with the grammatical function through which they are realized. After ordering the referring expressions, possible forward looking centers and backward looking centers are created. Finally the proposed candidates are created by the cross-product of the previous step. At the filtering stage, these candidates are filtered by a series of constraints such as syntactic coreference constraints, selectional restrictions and Rule 1 etc. At the final stage, the filtered candidates

are ranked by transitional orderings given in Rule 2. Examples and more implementation details of this algorithm will be discussed later.

In the following section I will present three interpretive models for processing context questions. Given a question in a discourse, the first model forms query terms by resolving the pronouns (I name it the anaphora model) 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).

5.3 Three discourse models for query expansion

5.3.1 Anaphora model

In the anaphora model, I use the centering algorithm based on Brennan et al. [1] to resolve pronoun references. 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 [29]), my approach keeps looking back at 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 Brennan et al. [1] (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 5.1. The details of the algorithm are reviewed in Jurafsky 2000 [12]. Once a pronoun is resolved, its antecedent is used in the query formation for the current question.

In this study, linguistic expressions will be used for query expansion based on discourse modeling. In example (5), the use of the expression such as *it* in (5Q3) to denote the entity the state fish is called reference. The linguistic expression used to perform reference is called referring expression, and the entity the state fish is called the referent. Both

expressions *it* and *the state fish* corefer to the same entity, one kind of fish that is named as the state fish of Hawaii. Referring expressions are very common in all kinds of discourse, including the mini discourse of context questions.

(5)

Q1: Where is Hawaii located?

Q2: What is the state fish?

Q3: Is it endangered?

For now, I will restrict the discussion on reference to entities as Centering Theory is an entity-based theory. Such linguistic expressions as indefinite noun phrases, definite descriptions, pronouns and demonstratives are often used in contextual questions to refer. However, in this chapter, I only focus on pronouns and investigate how they affect the processing of context questions within the framework of CT. As for reference resolution, there are some strict semantic (such as number, person, case, and gender agreement) and syntactic constraints that any successful resolution algorithm has to take into consideration. These hard rules help to reduce the referent candidates before further processing. Pronominalization is a form of definite reference as illustrated in example (5). Research on pronoun resolution has received tremendous attention from both linguistic and NLP communities. Various algorithms have been developed to address this problem. In the following sections, I will first discuss the pronoun resolution preference, then the pronoun resolution algorithms, finally the implementation details of the centering algorithm used in the anaphora model.

Pronoun resolution preferences

Different pronoun interpretation preferences have been applied in pronoun resolution algorithms to determine the potential referents. Jurafsky and Martin ([12], p.682) summarize these preferences as recency, grammatical role, repeated mention, parallelism and verb semantics. Briefly, the preference recency says that entities appearing in recent discourse are more salient than those from discourse further back. The preference grammatical role states that the grammatical role of the discourse entities determine the salience degree. In

other words, entities whose corresponding expressions are in subject position for example, are more likely to be the referents of pronouns than those appear in the object position, if the grammatical role preference decides that the salience degree of a subject is higher than that of an object. The preference repeated mention supports the idea that entities that have been focused on are more likely to keep their status, continuing to be focused on. This preference is another version of the transition preference constraint in CT, which favors the transition type *continue* instead of *retain* and *shift*. The preference parallelism is a syntactic constraint on potential referent candidates. Those entities whose realized expressions are structurally parallel with the pronouns are preferred to be the referents. The preference verb semantics is nonetheless a semantic-based constraint. The semantics of certain predicate verbs determines the pronoun interpretation and therefore determines the referents ⁸.

Unlike the agreement constraints, these preferences are more likely to be violated and one preference does not have to be ranked more salient than other preferences. For the same reason, a good pronoun resolution algorithm will have to decide which preference to implement and if more preferences are implemented, what is the difference computationally? Is the difference quantitative or qualitative?

Pronoun resolution algorithms

In this part, I will summarize some pronoun resolution algorithms implemented in the NLP literature, hoping to show how centering algorithm is different in terms of its mechanism.

Lappin and Leass ([80])'s algorithm singles out several influential factors each of which is assigned a numeric value or a weight which, as a matter of fact, takes a lot of empirical training to adjust. The sum of the weights then will be calculated to determine the referent of the pronoun in question. Besides the constraints required by agreement, this algorithm also considers other preferences. It ranks recency higher than grammatical role,

⁸The example ([12], p.683) shown below suggests that the semantic processing in pronoun resolution should be further investigated.

(1a) John telephoned Bill. (1b) He lost the pamphlet on Acuras.

(2a) John criticized Bill. (2b) He lost the pamphlet on Acuras.

which in turn is ranked higher than syntactic preference (e.g. head noun emphasis, which adds more weight to a referent if its corresponding expression is a head noun). In this way, pronoun resolution preferences are quantitativized and therefore calculable. In the meanwhile, the disadvantage of this approach is shown. It is experimental in terms of assigning the weights to each preference factor and this method does not guarantee optimal accuracy. In addition, the accuracy of the pronoun resolution will largely depend on the genre of the training corpus which means the weight assignment does not necessarily work best across all genres.

Hobbs'([81]) tree search algorithm basically is a syntactically-based algorithm. It fully relies on the correct and complete syntactic structure of the current sentence and its preceding sentences in the discourse since the search of the referents is performed entirely on the parsed trees of these sentences. In other words, the more accurate a parser is, the higher the accuracy this method may obtain. However, how a grammar is defined for a parse tree will affect the search results even if the same algorithm is adopted. Note that no explicit preferences are specified in the algorithm. Priorities in the search order implicitly reflect certain preferences. For example, recency is implicitly implemented because the search starts from the current sentence rather than its preceding sentence.

Note that the two algorithms just reviewed are different in several ways: 1) The first algorithm generates and processes a set of referent candidates while the second algorithm only proposes one. 2) The first algorithm explicitly specifies a set of preferences. More preferences could be added to this open set with ease in the sense that new operations are simply incorporated with existing components. The preferences in the second algorithm are implicitly implemented and any operation on the preferences, such as replacing or adjusting a preference, would increase the computation complexity tremendously in that the whole algorithm would be changed structurally.

Of course, none of the algorithms implement all of the preferences mentioned earlier. As for the performance of both algorithms it is hard to tell which is better than the other. The accuracy of both methods is below 90% to date. One has to admit the fact that automated pronoun resolution algorithms cannot resolve all the pronoun occurrences since sometimes

preferences contradict between themselves. And, it is obvious that as more preferences are introduced into an algorithm, the more complications it will encounter.

The third pronoun resolution algorithm I will describe is the centering-based algorithm that Brennan et al. ([1]) present. In section 5.2.3 I have described the two rules, three constraints and four intersentential transitions specified in this algorithm. The constraints include the coreference agreement constraints and the preference constraints discussed earlier.

Walker [75] reports an accuracy of 77.6% for Centering algorithm and 81.8% for Hobbs [81]; therefore the algorithm based on centering is a relatively efficient and competitive algorithm. The major reason that I did not use Hobbs's algorithm is that Hobbs' algorithm is purely syntactically-driven and in that sense it does not provide a mechanism to track how discourse entities are related semantically and how local coherence is maintained through the operations of discourse entities.

Implementation details

The algorithm implemented in the anaphora model is based on BFP87[1]. There are a few implementation details and modifications worth mentioning here.

First, I will consider the basic utterance unit to be a naturally occurring question in the context of QA. It has been an open issue as how to identify an utterance in centering. Does it have to be a sentence? Could it be a unit within a complex sentence? Multiclausal sentences in particular, introduce complication into the discussion. Poesio and Stevenson (2004) therefore treat utterance as a parameter in CT based on different arguments (see also [82], [83], and [84]) for discussions on identification of utterance with tensed clauses). GJW95 implicitly identifies utterances with sentences. When looking at the collected data for this study, I notice that the context questions have an average length of 7.28 words. Besides, in the data pool, there are no instances of questions with multiclausal sentences. Intuitively, it makes sense to identify an utterance with each individual question.

Secondly, instead of only dealing with the adjacent utterance (the notion of local coherence of Centering theory), this approach keeps looking back to all the previous questions

until an antecedent is found; Recall that rule 2 in GJW95 is written in terms of utterance sequences. As Grosz and Sidner [85] point out, starting from BFP87, “all the uses of this rule in language processing systems have adapted the rule by restricting it to pairs of utterances.” This results in missing “the essential intuition that what matters to coherence are centering transitions throughout a segment, not only between pairs of utterances” ([85],p.48). Taking the critique into consideration, I modified BFP87 to include more utterances/questions when searching for proper antecedents. Thirdly, the linguistic features used include gender, number and animacy. As mentioned in Chapter 3, these features are manually annotated according to my linguistic judgment.

Also, the ranking scheme is based on the same grammatical role hierarchy of the discourse entities as proposed in BFP87. It has been shown that the grammatical role functions as the primary determinant of discourse salience [86]. Miltsakaki [87] presents some evidence from corpus studies that entities in adjunct subordinate clauses have lower discourse salience than those in main clauses. In the implementation, I adopt Brennan et al [1]’s centering algorithm and a more detailed ranking hierarchy proposed as follows:

subject > existential predicate nominal ⁹ > object > indirect object >
demarcated adverbial PP ¹⁰

Note that entities in adjunct subordinate clauses are ranked lower than the entities in the main clause, but with the same grammatical role ranking hierarchy¹¹.

In the anaphora model, once a pronoun is resolved, its antecedent is used in the query formation for the current question. Let us first step through the algorithm for example (5), repeated below as (6).

(6)

Q1: Where is Hawaii located?

Q2: What is the state fish?

Q3: Is it endangered?

⁹A noun phrase that is used as a predicate in an existential sentence (e.g. *There is a cat.*).

¹⁰A 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.*).

¹¹The details of the algorithm are reviewed in Jurafsky 2000([12], p.691-693).

Since (6Q1) is the first question, or what is usually called a feeding question, it does not have a backward looking center according to CT. Since Hawaii is the only entity in (6Q1), it is the forward looking center as well as the preferred center which is defined as the highest ranked entity in C_f .

$C_f(Q1): \{\text{Hawaii}\}$

$C_p(Q1): \{\text{Hawaii}\}$

$C_b(Q1): \{\text{undefined}\}$

For (6Q2) there are no pronouns that need to be resolved, it is then treated as if it were a feeding question.

$C_f(Q2): \{\text{the state fish}\}$

$C_p(Q2): \{\text{the state fish}\}$

$C_b(Q2): \{\text{undefined}\}$

(6Q3) has one pronoun *it* that needs to be resolved using BFP87. According to Rule 1 and the coreference constraints (gender, number and animacy), C_p of (6Q2), the state fish, which is compatible with *it* is assigned as the referent. The linguistic expression *the state fish* corresponding to the C_p will then be added to (6Q3) to form query terms¹² for the document retrieval engine. To be more specific, the final query terms for (6Q3) would be: {is, it, endangered, the, state, fish}. Through this example, it is shown that the *realizes* relation makes it possible for a computer system to operate on discourse entities.

It should be noted that unless a pronominal referent gets assigned, the modified algorithm will go further back to a prior question. For example, in example (7) in order to resolve *him* in (7Q3), the algorithm will have to keep looking back to (7Q1) and locate Tom Cruise as the referent because the noun phrase *Nicole Kidman* in (7Q2) violates the gender constraint for *him*. Also when both *she* and *him* are resolved, the forward looking center for (7Q3) would be Nicole Kidman and Tom Cruise where subject (Nicole Kidman) is ranked higher than the object (Tom Cruise) according to the ranking hierarchy

¹²Query terms are the linguistic tokens that are input to a retrieval engine. The retrieval engine then sees them as a bag of words, which ignores the word order. When discussing the three computational models, by centers, I mean the realization of them, i.e. the corresponding linguistic expressions. Query terms in the brackets are separated by commas to indicate that they are a bag of words.

mentioned above.

(7)

Q1: When was Tom Cruise born?

Q2: Who was Nicole Kidman?

Q3: When did she marry him?

Of course, as expected, this algorithm cannot correctly resolve all the pronouns given that there are always complicated and unexpected cases that the algorithm cannot handle. Some complicated cases such as generics and cases that require more constraints are beyond the scope of this discussion. Since the efficiency of pronoun resolution algorithms are not the major concern of this work I will focus more on the discourse modeling. It is more of an implementation problem to add more constraints to the algorithm. As we will see, pronoun resolution is a common constituent of the processing models. Any extra efforts to improve the accuracy of pronoun resolution will benefit all three models.

5.3.2 Forward model

In the forward model, query terms for a current question are formed by incorporating forward looking centers C_f 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 anaphora model. The motivation for the forward model is based on my 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.

The motivation for the forward model is based on the assumption that a question discourse is coherent. The assumption behind this model is: the forward looking centers from the previous adjacent question provide more context information for the current question. Therefore this model would predict a better performance than the anaphora model. Consider example (8) that has two forward looking centers Tom Cruise and Nicole Kidman in (8Q1). The pronoun *she* in (8Q2) is resolved to Nicole Kidman using the anaphora model. The forward model will add the other forward looking center Tom

Cruise to the query terms when processing (8Q2). For the same reason, the forward looking center of (8Q2), i.e. Nicole Kidman will be added to (8Q3) for query expansion after *he* is resolved to Tom Cruise.

(8)

Q1: How is Tom Cruise related to Nicole Kidman?

Q2: What movies was she in?

Q3: What movies was he in?

$C_f(Q1)$: {Tom, Cruise, Nicole, Kidman}

$C_f(Q2)$: {Nicole, Kidman}

Forward model:

Query terms for Q2: {what, movies, was, she, in, Nicole, Kidman, Tom, Cruise}

Query terms for Q3: {what, movies, was, he, in, Tom, Cruise, Nicole, Kidman}

Anaphora model:

Query terms for Q2: {what, movies, was, she, in, Nicole, Kidman}

Query terms for Q3: {what, movies, was, he, in, Tom, Cruise}

From the resulting query terms, it is shown that in forward model both entities are incorporated into (8Q2) and (8Q3) to enrich the context of this mini discourse (which is different from anaphora model). In other words, not only the entities referred by pronouns are regarded as local context, but also the other forward-looking centers from previous utterances. Also, we notice that even if the order of (8Q2) and (8Q3) is switched, which is an equally possible discourse sequence, the fact that the resulting query terms would be the same shows that this algorithm is able to correctly capture the local context. Now that more entities are related to the context in processing question sequence, it is necessary to explore deeper to the semantic level as how these entities are related and how the relationship helps with the processing.

5.3.3 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 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 (see the algorithm in [12]). For questions that do not have pronouns, I used an entity-based algorithm, which extends the centering algorithm and assumes the highest ranked entity is the centered entity or most accessible in terms of interpretation and understanding. The same ranking scheme was used as in the anaphora model to assign a rank to each entity. Then the highest ranked entities from the adjacent question pair were compared and assigned a transition type according to Table 5.1.

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]. GJW95[29] mentioned that 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 the approach to identify transitions based on NP expressions, in particular, the definite noun phrases.

In the transition model, the extended algorithm is based on the following speculation. Intuitively, definite noun phrases that share the same NP head and modifier often refer to the same center, which results in a continuation using CT’s terminology. Similarly, attention will be retained if two similar entities referred to in two utterances have cor-

responding 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 5.2 shows the four rules that are used to identify different types of transitions. Table 5.3 shows the examples of how these transition rules would be applied to the non-pronominal referring expressions. 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, I 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 5.4 for the query expansion of the QA question in processing.

Table 5.2. 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

*Modifiers do not include the determiners *a*, *an* and *the*.

Table 5.3. Examples of transition rules on non-pronominal referring expressions

Transition	$Q_n(\text{NP})$	$Q_{n+1}(\text{NP})$
Continue	<i>a movie star</i>	<i>the movie star</i>
Retain	<i>the second debate</i>	<i>the third debate</i>
Smooth shift	<i>the best actor</i>	<i>the best actress</i>
Rough shift	<i>the space shuttle</i>	<i>the flight crew</i>

The strategy for the *continue* transition¹³ is motivated by the following two reasons. First, as pointed out in [29], 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 I conjecture that the highest ranked proper name in a question sequence carries more information than just for referring. In other words, I 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 [88] on topic status and proper name’s status in the definiteness hierarchy in [89], proper name should be given certain discourse prominence as it is an important definite noun phrase type. Since currently I do not resolve definite descriptions this strategy partially addresses the importance of definiteness status of other types of definite noun phrases besides pronouns.

(9)

Q1: Where is Hawaii located?

Q2: What is the state fish?

Q3: Is it endangered?

In my favorite example, repeated as (9), the transition between (9Q2) and (9Q3) is identified as *continue* because *it* in (9Q3) and *the state fish* in (9Q2) refer to the same entity (i.e., the state fish) and this entity is also the C_p of (9Q3). According to the strategy described in table 5.4 for *continue*, when processing (9Q3), in addition to the query term *the state fish* (corresponding to the antecedent for the pronoun *it* in (9Q3)), the proper name *Hawaii* from (9Q1) will also be inherited.

For the transition type *retain*, intuitively I 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, 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

¹³If there is no proper name, then do not expand query terms.

Table 5.4. 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 forward looking centers from the previous adjacent to the current question.

on this intuition.

(10)

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

Q2: Where was the 3rd debate held?

In example (10), the transition between (10Q1) and (10Q2) is identified as *retain* because according to table 5.2, expressions realizing $C_p(10Q1)$ and $C_p(10Q2)$, that is, *the 2nd president debate* and *the 3rd debate* share the same NP head but different modifiers. The strategy for *retain* will allow (10Q2) to inherit its time constraint *2004* from (10Q1).

For the transition type *shift*, currently I adopt the same 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.

(11)

Q1: When did Vesuvius destroy Pompeii the first time?

Q2: What civilization ruled at that time?

In example (11), the transition between (11Q1) and (11Q2) is identified as *rough shift* according to table 5.2 because NPs realizing $C_p(11Q1)$ (i.e., Vesuvius) and $C_p(11Q2)$ (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 (11Q2) 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.

5.4 Data analysis

The user data and the TREC 2004 factoid questions were used for the evaluation of the three models. As mentioned in Chapter 4, the user data includes four topics: (1) the presidential debate in 2004; (2) Hawaii; (3) the city of Pompeii; and (4) Tom Cruise. The basic information of the user data and the TREC 2004 data has been described in Chapter 4. Next, I will describe more information on these data in table 5.5.

Table 5.5 shows a comparison of the two data sets: my 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, the user 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 my 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*, my 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 5.5 we can see that, the surface characteristics across my 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 my 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 5.5. Characteristics comparison between my data and TREC 2004 data (including only factoid questions)

	Debate	Hawaii	Pompeii	Tom Cruise	Overall*	TREC2004
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	Eve/ent**	Entity	Eve/ent	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 <i>continue</i> transitions	21	19	26	69	135(30%)	105(64%)
Number of <i>retain</i> transitions	42	31	27	18	118(27%)	30(18%)
Number of <i>shift</i> transitions	47	59	59	17	182(43%)	30(18%)

* Overall user data ** 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 my 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 my data exhibits more diverse behavior. By studying these different characteristics of the two data sets, I hope to learn their implications for specific strategies from my empirical evaluation. Next, I will discuss the evaluations on the three models in detail.

5.5 Evaluation

A series of experiments were conducted to compare the performance of the three models on both the user study data and the TREC 2004 data. For the user study data, I 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, the entire AQUAINT collection (about 3G) was used. In all the experiments, the Lemur retrieval engine¹⁴ was used for document retrieval. Since the first occurrence of a correct answer is important, Mean Reciprocal Ranking (MRR) was taken as the first measurement. MRR is defined as:

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank_i} \quad (5.1)$$

where $rank_i$ is the rank of a retrieved document¹⁵ which provides the first correct answer for the i th question and N is the total number of questions evaluated. My evaluation

¹⁴ <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 a variety of retrieval models.

¹⁵ represented as DocRank in the following discussions.

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, I compared the performance of each model on both data sets. I 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 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)? 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 my 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 my 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.

5.5.1 Overall Evaluation Results

Before I present the overall evaluation results, it is worth pointing out the fact that the proposed three discourse models sometimes work, and sometimes fail for individual data questions. Now I pick some examples from the 230 TREC 2004 factoid questions.

The first example shows where the anaphora model works the best while the forward model and the transition model do not.

(12)

Q1: When was the first Crip gang started?

Q2: What does the name mean or come from?

Q3: What ethnic group/race are Crip members?

Q4: What is their gang color?

Query terms for Q4:

Anaphora model: {What, is, their, gang, color, *Crip, members*}

Forward model: {What, is, their, gang, color, *ethnic, group, race, Crip, members*}

Transition model: {What, is, their, gang, color, *Crip, members, first, Crip, gang*}

Anaphora model retrieval result: DocRank = 8, MRR = 0.125

Forward model retrieval result: DocRank = 10, MRR = 0.1

Transition model retrieval result: DocRank = 16, MRR = 0.1667

In example (12), the anaphora model identifies that the expression *Crip members* corefers with the pronoun *their*. The forward model adds the expressions (i.e. *ethnic group/race* and *Crip members*) that correspond to the forward looking centers for query expansion. The transition model identifies the transition between (12Q3) and (12Q4) as *continue* so it also adds the first proper name (i.e. *first Crip gang*¹⁶) in the discourse in addition to the resolution of *their*. The document retrieval results show that the anaphora model (MRR = 0.125) outperforms the other two models. The result indicates that more context information captured by the forward model and the transition model sometimes hurts the performance.

The next example shows that the forward model outperforms the anaphora model and the transition model.

(13)

Q1: Who is the lead singer / musician in Nirvana?

Q2: When was the band formed?

Q3: What is their biggest hit?

Q4: What style of music do they play?

¹⁶Since the definite description is not the concern of this chapter, if an expression that contains a proper name, the whole expression will be counted as a proper name in the implementation conducted in this chapter.

Query terms for Q4:

Anaphora model: {what, style, of, music, do, they, play, *Nirvana*}

Forward model: {what, style, of, music, do, they, play, *Nirvana*, *biggest*, *hit*}

Transition model: {what, style, of, music, do, they, play, *Nirvana*}

Anaphora model retrieval result: DocRank = 2, MRR = 0.5

Forward model retrieval result: DocRank = 1, MRR = 1.0

Transition model retrieval result: DocRank = 2, MRR = 0.5

In this example, the pronoun *they* is resolved to *Nirvana* for the anaphora model. For the transition model, the transition between (13Q3) and (13Q4) is identified as *continue* so the strategy for *continue* transition adds the first encountered proper name *Nirvana* to the query terms. Note that although the same term *Nirvana* is added for query expansion according to the anaphora model and the transition model, they are added for different reasons. The forward model, by simply adding the forward looking center (their biggest hit), provides more context for processing (13Q4). Again, the pronoun *their* in *their biggest hit* is resolved to *Nirvana*, but the other two terms *biggest* and *hit* actually help the document retrieval. Choosing forward looking centers, in this example works better (MRR = 1.0) than other techniques.

Example (14) shows the advantage of the transition model where it outperforms both the anaphora model and the forward model. The transition model gets the best MRR value (0.25).

(14)

Q1: What film introduced Jar Jar Binks?

Q2: What actor is used as his voice?

Q3: To what alien race does he belong?

Query terms for Q3:

Anaphora model: {To, what, alien, race, does, he, belong, *his*}

Forward model: {To, what, alien, race, does, he, belong, *his*, *actor*, *voice*}

Transition model: {To, what, alien, race, does, he, belong, *his*, *Jar*, *Jar*, *Binks*}

Anaphora model retrieval result: DocRank = 0, MRR = 0

Forward model retrieval result: DocRank = 17, MRR = 0.0588

Transition model retrieval result: DocRank = 4, MRR = 0.25

In example (14), the anaphora model fails to resolve the pronoun *he* in (14Q3), because the entity Jar Jar Binks is annotated as object singular instead of male singular. The centering program I designed therefore could not identify the antecedent of *he* being Jar Jar Binks. The forward model adds the expressions *his voice* and *actor* for query expansion. For the same reason, the pronouns *he* in (14Q3) and *his* in (14Q2) could not be resolved. However, the strategy used for *continue* between (14Q2) and (14Q3) adds the proper name *Jar Jar Binks* to the query input. The strategy actually compensates the reference resolution failure.

From the above examples, we see that if the document collection and/or the document retrieval engine stay the same, the proposed techniques both work and fail to work for individual cases.

Table 5.6 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 5.5, especially the TREC 2004 data), the preceding question can simply provide a context for the current question. Since all three models rely on pronoun resolution, the performance of the automated pronoun resolution algorithm directly impacts the final performance of document retrieval. Therefore in Table 5.6, along with the performance resulting from automated processing (i.e., marked with “auto” in the column title), I 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).

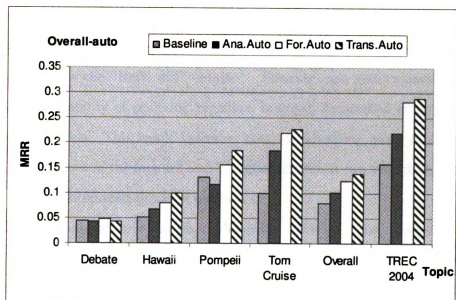
To better present the results, Figure 5.1 shows a detailed comparison between four models as a result of automated processing. As shown in Figure 5.1(a), except for the *Debate* data the incremental increase in the complexity of discourse processing (e.g., from the anaphora model, to the forward model, to the transition model) improves the overall

Table 5.6. Overall performance of different models on document retrieval for my data and TREC 2004 data

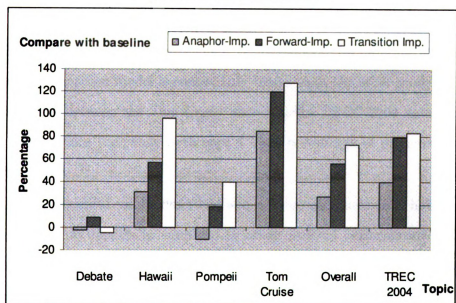
Topic	Baseline	Ana.Auto.	Ana.Key	Ana.%diff	For.Auto	For.Key	For.%diff	Trans.Auto.	Trans.Key	Trans.%diff
Debate	0.044	0.043	0.048	11.6%	0.048	0.048	0%	0.042	0.042	0%
Hawaii	0.051	0.067	0.085	26.9%	0.080	0.085	6.2%	0.100	0.110	10.0%
Pompeii	0.132	0.118	0.149	27.3%	0.156	0.163	4.5%	0.185	0.186	0.5%
Tom 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 2004	0.158	0.221	0.265	20.0%	0.283	0.288	1.7%	0.289	0.296	2.4%

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 my 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 5.1(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 anaphora 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 the 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 a higher percentage of pronouns (see Table 5.5). The specific transitions identified between two adjacent questions largely depend on the resolution of those pronouns. Therefore, the anaphora 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.,



(a) Overall automated system



(b) Automated system compared to baseline

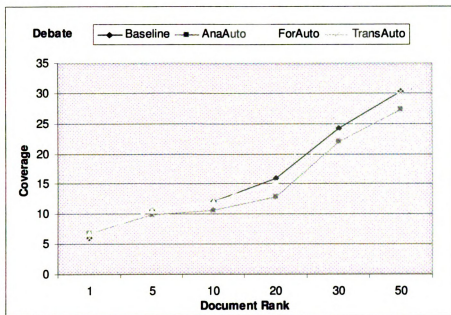
Figure 5.1. Overall comparison of four models based on automated processing

through the *continue* transition as discussed earlier).

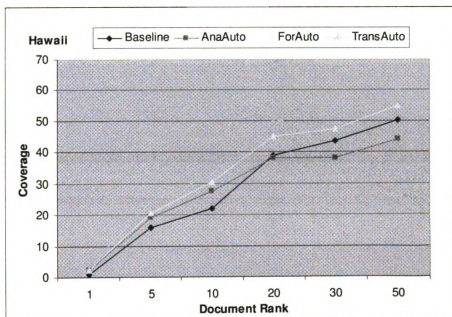
From the experimental results, it is interesting to point out that the sensitivity of each model varies in response to the accuracy of automated discourse processing. From Table 5.6, in the anaphora model, a perfect pronoun resolution makes a big difference compared to an imperfect automated pronoun 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, I 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 [90]. Figure 5.2 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).

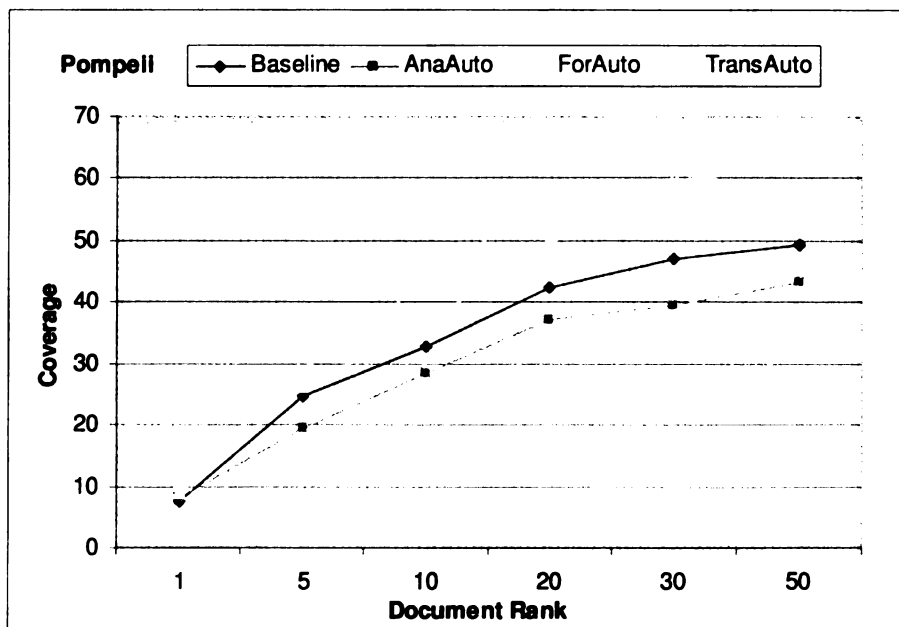
Given the experimental results described above, a natural question is how the retrieval performance from my models is compared to other retrieval performance. It is hard to achieve this kind of comparison because TREC 2004 did not provide document retrieval performance based on the context questions. The closest I can find is the “coverage” based on passage retrieval for TREC 2004 factoid questions provided by the University of Sheffield [90]. Table 5.7 shows the 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 my system was evaluated on document retrieval and Sheffield’s system was on passage retrieval, this is not a direct



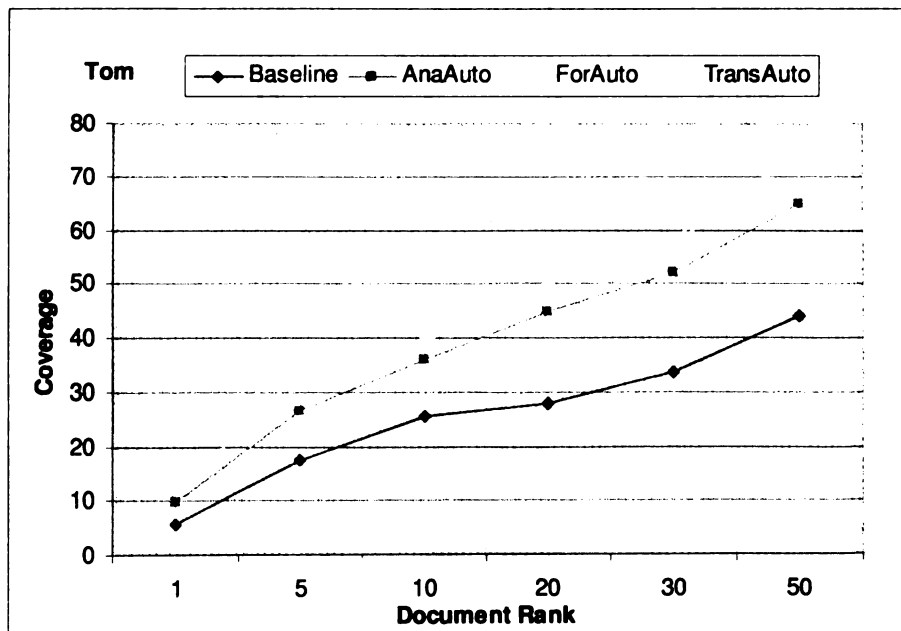
(a) Debate



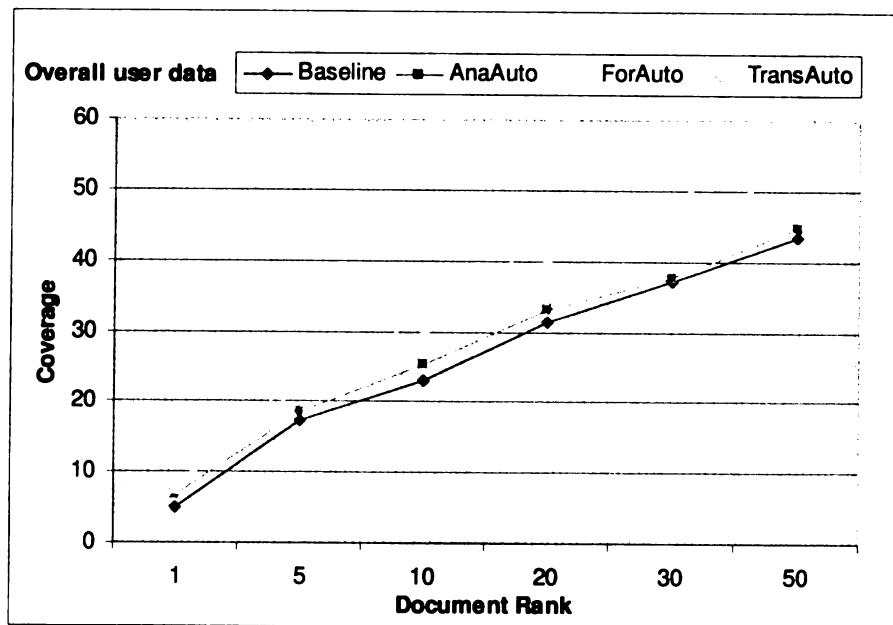
(b) Hawaii



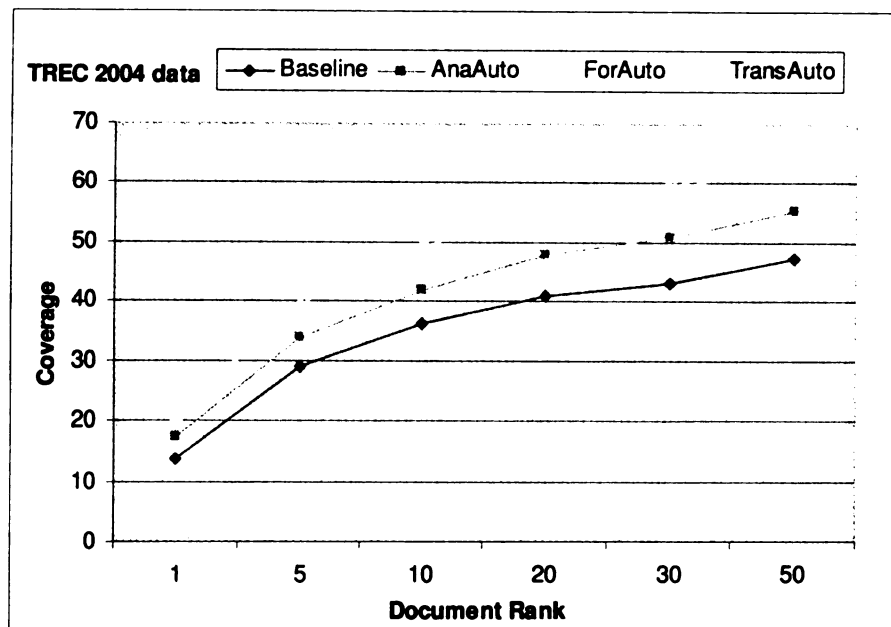
(c) Pompeii



(d) Tom Cruise



(e) My overall data



(f) TREC2004 data

Figure 5.2. Coverage comparison between four models based on automated processing

Table 5.7. 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* [90]
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/>

comparison. I list them together simply to have some sense about whether my 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, I 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. I report the results next.

5.5.2 Evaluation and Analysis based on Transitions

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

Continue transition

Figure 5.3 shows the overall comparison of the three models on the question pairs with the transition type *continue*, with Figure 5.3(a) for the automated system and Figure 5.3(b) for the annotated system. In general, for *continue* pairs, the transition model works the best, then the forward model and the anaphora 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.

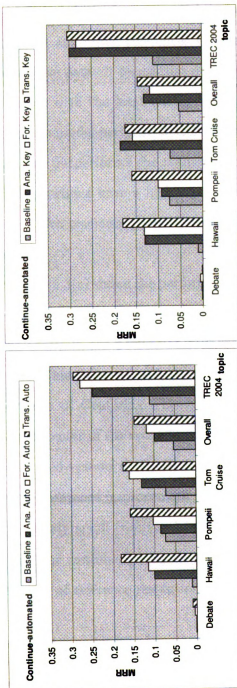
Figure 5.3(a) shows that the transition model performs consistently better than the forward model. This result indicates that the strategies used in the transition model for *continue* questions are adequate to provide appropriate context. The transition model provides more information than the forward model or the anaphora model, but at the same time lowers the risk of introducing unnecessary forward looking centers into processing as in the forward model.

The forward model outperforms the anaphora model across all the topics, which is also shown in Figure 5.3(a). This result indicates that reference resolution alone is not enough for obtaining adequate context information for discourses marked with *continue* transitions. Meanwhile, I observed that the anaphora 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 5.3(b), which shows the performance on the *continue* pairs with all the pronouns correctly resolved. When all the pronouns are correctly resolved, the anaphora model actually outperforms the baseline model.

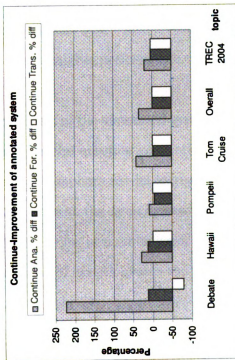
Table 5.8. Transition model performance improvement for *continue*

Topic	Increase over ana. model(%)	Increase over for. model(%)	Question (excl. the feeding) w/ 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

Table 5.8 shows the performance improvement of the transition model over the forward model and the anaphora 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 a higher proportion of pronouns compared to other topics. For the *Tom Cruise* and the TREC 2004 data which have a higher percentage of pronouns (i.e.,



(a) Automated system



(b) Annotated system

(c) Improvement of annotated system

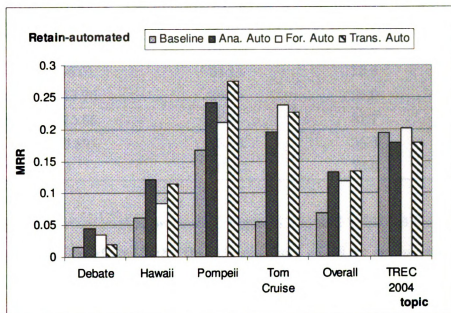
Figure 5.3. Performance on CONTINUE pairs

81.7% and 73.9% respectively), the transition model improves MRR modestly compared to other topics: 33.3% and 17.8% over the anaphora model, and 9.8% and 5.9% over the forward model respectively. Figure 5.3(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 anaphora 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 a higher percentage of pronouns (e.g. *Tom Cruise*), the anaphora model with pronouns properly resolved will achieve higher performance compared to other models.

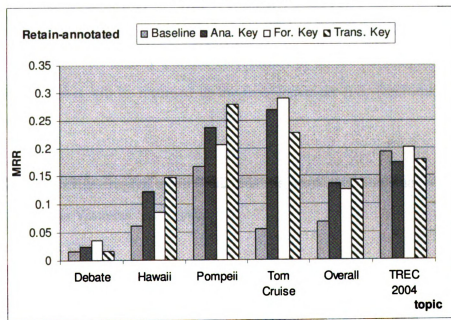
Figure 5.3(c) shows the performance improvement of the annotated systems for the three models compared to the automated system. For the question pairs with the transition type *continue*, the performance of the annotated anaphora model increases across all the topics. This makes sense because if a question 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 5.3(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 anaphora model.

Retain transition

Next, I present the evaluation results for the *retain* pairs. Figure 5.4 shows the overall comparison of the three models on the question pairs with the transition type *retain*, with Figure 5.4(a) for the automated system and Figure 5.4(b) for the annotated system. Table 5.9 lists the performance improvement of the transition model over the other two models



(a) Automated system



(b) Annotated system

Figure 5.4. Performance on RETAIN pairs

based on the automated system. I first compare the transition model and the anaphora

Table 5.9. Transition model performance improvement for *retain*

Topic	Increase over ana. model(%)	Increase over for. model(%)	Question (excl. the feeding) w/ 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

model based on the automated processing. Figure 5.4(a) and Table 5.9 firstly show that the transition model performs better than the anaphora model for *retain* pairs in *Pompeii* and *Tom Cruise*. One advantage of the transition model over the anaphora model is its capability of adding constraints from the context as in the example (16), where the year 1631 is inherited from Q_i to Q_{i+1} for the query expansion. The Lemur retrieval results based on different query expansion terms are also shown below. The document that contains the correct answer will not be returned at all for the anaphora model (i.e. DocRank=0) while it will be returned as the first document (i.e. DocRank = 1) for the transition model.

(15)

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}

Transition model retrieval result: DocRank =1, MRR = 1

Anaphora model: {when, was, Vesuvius, last, cycle}

Anaphora model retrieval result: DocRank =0, MRR = 0

Secondly, Figure 5.4(a) and Table 5.9 show that the transition model performs the same as the anaphora 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 anaphora model. However, I would expect performance difference between the two models for longer questions with more constraints such as time phrases. Finally, Figure 5.4(a) and Table 5.9 show that the transition model performs worse than the anaphora 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, I suggested excluding this kind of adverbs in QA processing.

Next, let us compare the transition model with the forward model. From Figure 5.4(a) and Table 5.9, 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 I examine closely, I 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 5.4(a) is that the baseline model outperforms the transition model for the TREC 2004 data. The TREC 2004 data is more coherent than my 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 5.5. The TREC 2004 data has more *continue* pairs than my data (64% vs. 30%). Intuitively, a more coherent discourse would favor more context information for the purpose of discourse processing. However, the strategy I 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, I 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.

Shift transition

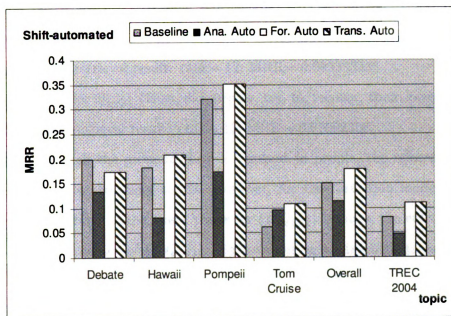
Finally, I discuss the performance results for the *shift* pairs. Figure 5.5 shows the overall comparison of the three models on the question pairs with the transition type *shift*, with Figure 5.5(a) for the automated system and Figure 5.5(b) for the annotated system.

From Figure 5.5(a) and Figure 5.5(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, I can get pretty good performance (such as for the *Debate* data). The reason that the anaphora model does not work well is that the *shift* pairs normally do not have referring expressions.

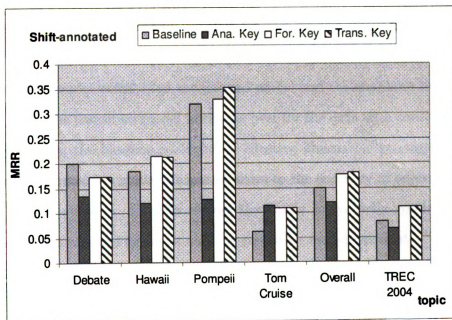
The performance improvement based on the annotated system over the automated system for the *shift* pairs is not as significant as for the other two transition types. Since there are not many cases where pronoun resolution is involved in the *shift* pairs, it is hard to examine how pronoun resolution would impact the different models. I also observed the performance on the *Pompeii* data even drops a little for the annotated system. After examination of the processing, I found examples such as (16), which could provide a possible explanation.

(16)

Q_i : When did Pompeii vanish?



(a) Automated system



(b) Annotated system

Figure 5.5. Performance on SHIFT pairs

Q_{i+1} : When did Vesuvius erupt?

Q_{i+2} : How did people try to recover their possessions?

Transition model for Q_{i+2} : {How, did, people, try, to, recover, their, possessions, Vesuvius}

Transition model retrieval result: rank= 12; MRR= 0.083333336

Anaphora model for Q_{i+2} : {How, did, people, try, to, recover, their, possessions}

Anaphora model retrieval result: rank= 13; MRR= 0.07692308

In ($16Q_{i+2}$), 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 ($16Q_{i+2}$) because the entity Vesuvius is the forward looking center of ($16Q_{i+1}$). By introducing an important discourse entity Vesuvius, this operation actually increases the chance of hitting the right document for ($16Q_{i+2}$). The retrieval result has shown that, the correct document returned for ($16Q_{i+2}$) is at a better rank for the transition model than for the anaphora model.

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 anaphora 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 anaphora model. In other words, the ability of correctly resolving pronouns affects the anaphora model the most and the transition model the least.

5.5.3 Evaluation and Analysis based on Pronouns

To further examine the performance of different models on questions with different pronoun usages, I separated questions into two categories for evaluation: questions with and without pronouns. Figure 5.6(a) and Figure 5.6(b) show the evaluation based on the pronoun dichotomy for the automated system and the annotated system.

When Figure 5.6(a) and Figure 5.6(b) are compared, it is noted 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 I separated the questions by transition types. Within individual user data, the performance of the anaphora 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.

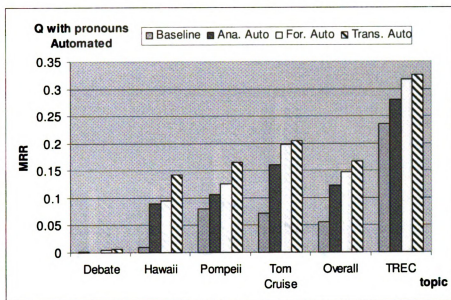
Figure 5.7 shows the evaluation results for questions without pronouns, with Figure 5.7(a) for the automated system and 5.7(b) for the annotated system.

Figure 5.7(a) and 5.7(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.

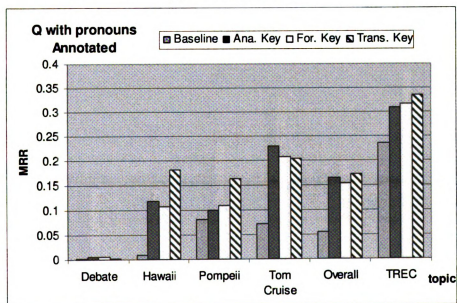
Compared with Figure 5.6, the performance of the *Debate* data increases noticeably both for the automated and the annotated system. One possible explanation is that the majority of the *Debate* questions fall into this category. However, there is no much difference within the three models for the *Debate* data. This indicates that centering-based models are more appropriate to process context questions focusing on entities rather than on events.

Figure 5.7(a) shows the automated anaphora 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 anaphora 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 5.7(a) and 5.7(b), we see that for the baseline and the anaphora 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

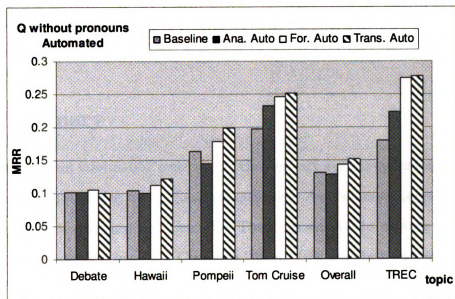


(a) Automated system

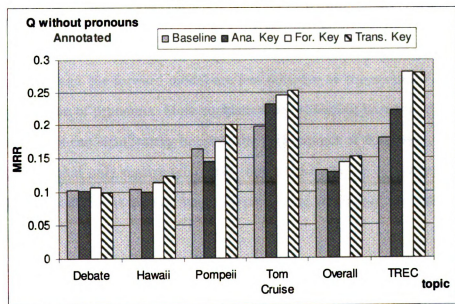


(b) Annotated system

Figure 5.6. Performance for questions with pronouns



(a) Automated system



(b) Annotated system

Figure 5.7. Performance for questions without pronouns

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 anaphora model for both the questions with and without pronouns. Second, there is no significant advantage of the transition model for the event type data.

5.6 Summary

To support coherent information seeking, this chapter 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 based on Centering Theory (the anaphora model, the forward model, and the transition model) were proposed, implemented, and evaluated on my user study data and the TREC 2004 data.

The empirical results indicate that the transition model outperforms the anaphora 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 chapter presents my initial investigation on the role of discourse processing for context questions. There are many dimensions along which my 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. I will also extend context question answering to fully interactive question answering and investigate the role of discourse processing in this new setting.

CHAPTER 6

Discourse models based on definiteness hierarchy

6.1 Introduction

In the previous chapter, I have investigated how to employ Centering Theory to process a context question using the context information from the previous question(s). Three discourse models are presented based on the framework of Centering Theory. In this chapter I continue the discussion on discourse modeling for context questions. In this part of the study, I adopt the definiteness hierarchy of noun phrases to help in the processing of context questions. The pronouns and the definite descriptions that appear in the questions will be my focus. The role of the target as providing context for all the questions in a set will also be investigated. I hope that, this empirical effort will lead to a solution to the problem of building an efficient discourse model for processing context questions which contain not only pronouns but also definite descriptions. The TREC 2004 and 2005 data are used in the evaluation, both of which consist of question sets where each set asks for information regarding a particular target.

A definite description is defined as a denoting phrase in the form of “the X” where X is a noun phrase or a singular common noun that describes a specific individual or object. The noun phrases *the book* or *the earth* are examples of definite descriptions.

Recall that in Chapter 5, the three models perform better on the *Tom Cruise* data than on the *Debate* data. Besides the topic difference (i.e. Tom Cruise is an entity but the 2004 presidential debate is an event), these two datasets are different in terms of the distribution for definite descriptions. A preliminary examination shows that 66.7% of the *Debate* questions are found to have the definite descriptions, more than 10 times of that for the *Tom Cruise* questions (6.4%). It is interesting to investigate how these definite descriptions are used and how they affect the processing of the context questions.

Secondly, this research is motivated by the new difficulty posed by the TREC 2005 data: 1) increased number of definite description occurrences and 2) the new *event* type of target. The TREC 2005 data contains more definite descriptions than the TREC 2004 data. TREC 2004 data consists of 65 sets of questions with 84 definite descriptions in total. The newly released TREC 2005 data consists of 75 sets of context questions with 239 definite descriptions in total, which is almost 3 times more than that of the TREC 2004 data. At the same time, the percentage of pronouns decreases for the TREC 2005 data. This suggests that the research on the definite descriptions should receive more attention than before. Since pronoun resolution is a well-researched topic, it is no longer my focus for context question processing. The existing QA systems do not provide more processing beyond resolving pronoun references, therefore they are not sufficient for handling the data with more complicated linguistic expressions.

As mentioned in Chapter 2, each question in a TREC question set is interpreted in the context of a TREC target, which could be of type *people*, *entity*, *organization* etc. The fact that all the TREC 2004 targets are in the form of noun phrases allows a system to use such an approach as replacing pronouns in a question with the corresponding target. However in TREC 2005, NIST added topic type *event* to the datasets. Targets like “Russian submarine Kursk sinks” or “Miss Universe 2000 crowned” are obviously not noun phrases. Problems arise immediately for the substitution strategy to process the questions with *event* type of targets. Therefore the role of target with respect to providing context for processing questions needs further examination.

Thirdly, unlike the study of pronoun resolution, the empirical studies in definite descrip-

tions from a computational perspective are rather limited. There is no well-established computer algorithm or approach specifically dedicated to processing them. It is particularly true for processing them in context questions. Motivated by these reasons, in this chapter I continue my study on context questions and initiate an investigation on the definite descriptions, and their role in developing efficient discourse models for processing mini QA discourses. In addition, I will discuss the role of pronouns in context questions from another perspective. In order to do that, there are two questions that need to be answered first: what would be a proper theoretical framework for me to work on? How do I come up with an appropriate taxonomy that is computationally possible for processing the definite descriptions in context questions? Once these questions are answered, I will be able to develop corresponding discourse models to process definite descriptions for query expansion.

The rest of the chapter is organized as follows: section 6.2 describes a new challenge that the current QA systems are facing; section 6.3 presents some related work on the computational efforts that involve targets, definite descriptions, and event in context question answering; section 6.4 presents the theoretical background information for the empirical study conducted in this chapter; section 6.5 discusses the uses of definite descriptions; section 6.6 presents the discourse models that were developed specifically for handling the pronouns and the definite descriptions in context questions; section 6.7 presents the evaluation results for the models discussed in section 6.6; and finally section 6.8 concludes the chapter.

6.2 A new challenge in QA discourse

The fact that the TREC 2005 QA test sets have more definite descriptions poses a new challenge for the current QA systems, because the state-of-the-art techniques have not provided adequate solutions to process them. In Chapter 2, I have mentioned that some of the TREC 2005 participating systems used the target-appending strategy to process context questions. This strategy has two drawbacks: 1) I observe that in the user data,

a pronoun used in a question does not necessarily refer to the target/topic, and (2) the target is not necessarily referred to only by a pronoun. The work in Chapter 5 has partially addressed the first issue. The centering-based models keep track of all the entities in the context questions, not merely focusing on the target entity, so a pronoun could refer to any of these entities. Next, I will look at the second problem. As a matter of fact, there are cases in the TREC 2005 data where definite descriptions are used to refer to the target. For instance, example (1)'s target is "Merck & Co.". However it is referred by the definite descriptions *the company* in (1Q1), (1Q2), (1Q4) and (1Q6).

(1)

- Q1: Where is *the company* headquartered?
- Q2: What does *the company* make?
- Q3: What is their symbol on the New York Stock Exchange?
- Q4: What is *the company's* web address?
- Q5: Name companies that are business competitors?
- Q6: Who was a chairman of *the company* in 1996?
- Q7: Name products manufactured by Merck.

This problem is addressed by the system developed at the University of Sheffield [16]. Their approach is to replace both the pronominal and the co-referential nominals with the target. The processed questions are shown in (2).

(2)

- Q1: Where is *Merck & Co.* headquartered?
- Q2: What does *Merck & Co.* make?
- Q3: What is *Merck & Co.'s* symbol on the New York Stock Exchange?
- Q4: What is the *Merck & Co.'s* web address?
- Q5: Name companies that are business competitors?
- Q6: Who was a chairman of *Merck & Co.* in 1996?
- Q7: Name products manufactured by *Merck & Co.*

This strategy seems working for example (1), however, another problem arises if there are definite descriptions that do not refer to the target, such as example (3) whose target is

“Shiite”. The definite descriptions *the first Imam of the Shiite sect of Islam* in (3Q1) and *the third Imam of Shiite Muslims* in (3Q4) do not refer to the target, but some entity related to the target.

(3)

Q1: Who was *the first Imam of the Shiite sect of Islam*?

Q2: Where is his tomb?

Q3: What was this person’s relationship to the Prophet Mohammad?

Q4: Who was *the third Imam of Shiite Muslims*?

Q5: When did he die?

Given the increasing trend of definite description occurrences in the TREC question sets, it is necessary to take a good look at the definite descriptions in the context questions. Next, I will look at some computational work concerning the processing of definite descriptions, hoping to give the readers some idea of where I am and how my research relates to others’ work.

6.3 Related work

I now review some of the computational work that involves definite descriptions.

6.3.1 Computational work on definite descriptions

Computational efforts on processing definite descriptions have focused on two aspects: definite description resolution and definite description classification. These two aspects are closely related to each other. The goal of definite description resolution is to develop approaches to identify the antecedents of definite descriptions automatically. The antecedent corefers to an entity/event¹ with a definite description (such as *Merck & Co.* is the antecedent of *the company* in example (1) because they corefer to the entity Merck & Co.

¹I only consider these two for the context questions.

company.). Message Understanding Conferences², MUC-6 [91] and MUC-7³ conducted subtasks to resolve definite descriptions. Systems such as Appelt et al.[92], Gaizaukas et al.[93], and Humphreys et al.[94] implemented specific modules to resolve definite descriptions. For example, the *Discourse Interpreter* module in Humphreys et al.[94]’s LaSIE-II system performs co-reference resolution between new and old concepts, a hierarchy of which represents the semantic net of the discourse under processing. Pronouns, proper names, definite descriptions and other type of noun phrases are considered in the co-reference mechanism. Searching for the antecedent of a definite description could be intrasentential or intersentential based on the co-reference rules that are specified by the system.

Definite description classification aims to automatically identify definite description uses. In 1990 Fraurud [95], two classes of uses are presented in the empirical study: *subsequent-mention* covers the cases where a definite description denotes the same entity as its antecedent denotes, and all other uses are classified into *first-mention*. Poesio and Vieira 1998’s work [96] presents a system for classifying definite descriptions in arbitrary domains. They conduct a corpus-based investigation on the definite description uses in written texts. The definite descriptions are classified into: *direct anaphoric*, *discourse new* and *bridging* (details to be discussed in section 6.5). The solution to the classification task often serves to help identify the antecedents of definite descriptions. The results show that, based on the different classification schemes, the agreement among human annotators varies in determining the classes assigned to the definite descriptions as well as the antecedents assigned to them. They also conclude that the majority of the definite description uses are not anaphoric.

²A major conference whose purpose is to facilitate information extraction and machine learning. More information can be found at <http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html>

³<http://www-nlpir.nist.gov/related-projects/muc/proceedings/muc.7.toc.html>

6.4 Theoretical background

In the upcoming sections, I will introduce the linguistic knowledge that is relevant in building QA discourse processing models. The theoretical issues include: 1) the definiteness hierarchy of noun phrases, which provides a ranked list of definite noun phrases. The ranking will be used in resolving pronouns and in helping resolve definite descriptions in the models developed for the study. 2) the Familiarity theory of definite descriptions, which provides us with a theoretical basis for categorizing the definite description uses in context questions; 3) some classic categorizations of the definite description uses, which cast helpful insights for us to classify the definite descriptions that appear in QA discourses. I now examine these issues in order.

6.4.1 Definiteness hierarchy

In addition to *definite description*, it is not uncommon to find that researchers include more NP types as definite NPs. For example, based on the work of Prince [97], Birner & Ward [98] and Ariel [99][100], Abbott[89] presents two lists of NP types which are ranked roughly according to the *definiteness* and *indefiniteness* of these NPs. Definiteness is considered as a property of noun phrases used to identify “particular and determinate entity or group of entities” [89]. Such noun phrases as proper names, personal pronouns, NPs with various quantifiers etc. have been treated as having the property of definiteness. Meanwhile, there are NP types that have the characteristics opposite to the definite NPs, that is, when they are used, no particular entity or entities can be referred to. In that sense, they have the property of *indefiniteness*.

Adapted from the lists presented in Abbott[89], table 6.1 shows a ranking hierarchy that I used in the study. In the list, *pronoun* is treated as the most definite NP type, while *indefinite NP* (represented in the list as “A/An”) is ranked as the least definite NP type. In my implementations, I actually combine *demonstratives*, *definite descriptions* and *possessive NPs* together and label them as one category *definite description*. The reason is that there are only a few occurrences of *demonstratives* and *possessive NPs* in

the TREC 2004 and TREC 2005 data. The incorporation of these three types will help me focus on the study of the definite descriptions. For the same reason, I combine *bare NPs* (represented as $[DET\emptyset]$ in the list) with *indefinite NPs* to one category relabeled as *indefinite*. Other NP types mentioned in Abbott[89] are not included in the study because those NP types rarely occur in the QA discourses that I am concerned with.

Table 6.1. Definiteness hierarchy of noun phrases

NP type	Note	Example
Pronoun*	personal pronoun	<i>he, them</i>
Demonstrative	NP with demonstrative determiner	<i>this book</i>
Definite description	NP with determiner <i>the</i>	<i>the book</i>
Possessive NP	NP with genitive NPs as determiner	<i>his book</i>
Proper name	Full name or partial name	<i>Bing Crosby</i>
$[DET\emptyset]$	all bare NPs	<i>books</i>
A/An**	Indefinite NP	<i>a book</i>

* most definite

** least definite

After the rearrangement of the definite noun phrases, I now have the ranked definiteness hierarchy as shown below:

(4)

Pronoun > Definite description > Proper name > Indefinite ⁴

I will return to this hierarchy in section 6.6 for more discussion and present its application in the QA discourse modeling.

6.4.2 Familiarity theory

In my study, I categorize the definite description uses based on the Familiarity theory in linguistic literature. Next, I will cover the important essentials of this theory, and the “Uniqueness” theory for comparison.

Christophersen 1939’s work[101] describes what has come to be known as Familiarity theory of definite description. Essentially, he describes the usage of definite description

⁴“>” means the NP type on the left side is more definite than the one on the right side.

based on an intuitive observation, that is, speakers tend to use the definite article (i.e. *the*) to refer to a referent that is mutually familiar to the hearer as well as to the speaker.

Now the speaker must always be supposed to know which individual he is thinking of; the interesting thing is that the *the*-form supposes that the hearer knows it too. For the proper use of the form it is necessary that it should call up in the hearer's mind the image of the *exact individual* [italics added] that the speaker is thinking of. If it does not do that, the form will not be understood. (Christophersen 1939 [101], p.28)

The article *the* brings about that to the potential meaning (the idea) of the word is attached a certain association with **previously** [bold added] acquired knowledge, by which it can be inferred that *only one* [italics added] definite individual is meant. That is what is understood by *familiarity*. (Christophersen 1939 [101], p.72)

While Christophersen emphasizes that both the speaker and hearer share knowledge of an individual through previous communication, he does not deny the implied uniqueness of this *exact* and *only one* individual.

A similar familiarity assumption is also stated in Heim's work [102] [103]. Heim develops the famous File Change semantics attempting to address the difference between definite and indefinite noun phrases. In her file card metaphor, a discourse has various file cards each representing a discourse entity. The definites and indefinites differ in terms of how the file cards are kept. If an indefinite is used in the discourse, it is seen as starting a new card. On the other hand, if a definite is used, it is seen as updating an old card. On Heim's analysis, definite descriptions denote discourse-old entities (whose existence is presupposed) while indefinites introduce discourse-new entities in Prince [97]'s term.

However, there are many cases where definite descriptions do not denote entities that have been mentioned in the discourse context or not assumed to be familiar to the addressee. Consider *the Miami debate* in example (5). The definite description denotes an event that has not been introduced in this mini QA discourse.

(5)

Q1: How many debates were there in the 2004 presidential election?

Q2: What was the focus of the Miami debate?

The familiarity theory thus faces the problem presented in such empirical data as (5). David Lewis proposes the *accommodation principle*, which is meant to rescue the theory, in his classic paper “Scorekeeping in a language game” [104]. In his analogy, conversations are regarded as language games keeping scores in an evolving way. Therefore definite descriptions do not necessarily denote entities “in existence” or “in some contextually determined domain of discourse” ([104], p.348). The problematic cases such as (5) are explained by his rule of accommodation for presupposition.

If at time t something is said that requires presupposition P to be acceptable, and if P is not presupposed just before t , then - *ceteris paribus* and within certain limits- presupposition P comes into existence at t . (Lewis [104], p.340)

In other words, referents may come into being automatically at the moment the speaker utters the denoting definite description. However Lewis does not specify any “limits” that would constrain the so-called *accommodation*. Gazdar([105], p.107) and Abbott [106] raise similar questions as to what degree the rule would allow a presupposition allegeable. Abbott([106],p.1426) points out that “the theory would become almost vacuous, since no counterexamples would be raised against it”.

6.4.3 Uniqueness theory

Given the problems of familiarity theory, one would want to seek an account for the problematic examples. Uniqueness theory is another theory on definite descriptions. Unlike familiarity theory, the notion of uniqueness has been established more from a semantic rather than a pragmatic perspective. Here, I cannot skip one person, Bertrand Russell who contributed significantly to the philosophy of language, and especially to the understanding of definite descriptions. Although his classic work came out more than a century

ago, it has been and still is an inspiration for the study of the semantics of denoting phrases. The famous definite description *the king of France* used in example (6) is such a denoting phrase.

(6)

S1. The King of France is bald.

S2. The King of France is not bald.

On Russell [107]’s analysis, neither the statement (6S1) nor the statement (6S2) is true.

The definite description *the king of France* is treated as compound statements in (7).

(7)

S1. There is an x such that x is being King of France.

S2. There is no y , y not equal to x , such that y is being King of France.

S3. x is bald.

S4. x is not bald.

Note that (7S1) describes the existence of the entity the King of France while (7S2) states the uniqueness of it. The fact that France is a republic and there is no present King existing makes (7S1) false and therefore (6S1), whose truth value is decided by the conjunction of three propositions (7S1), (7S2) and (7S3) is false; so is (6S2), whose truth value is decided by (7S1), (7S2) and (7S4). In plain English, the logic form of (6S1) says: “there is an individual/entity who has the property of being King of France, and there is only one such individual/entity, and this individual is bald”. On logical analysis, (6S1) is described as (8).

(8)

$$\exists(\text{King of France}(x) \ \& \ (y)(\text{King of France}(y) \rightarrow (y = x)) \ \& \ \text{Bald}(x))$$

Therefore, Russell’s analysis assigns the uniqueness for an entity that fits into the descriptive content of the NP. However his theory of definite description has received criticisms in the philosophical, logical and linguistic literature. Interested readers are recommended to find more discussions on definite descriptions in Löbner [108], Kadmon 1990[109], and Hawkins [110]. The discussion of the uniqueness theory is presented here to provide a complementary view of the familiarity theory. In the current study of modeling the QA

discourse, I adopt a view that is closer to the familiarity theory. In the next section, I will elaborate how I classify the definite description uses in QA discourses conforming to this view.

6.5 Classification of definite description uses

As mentioned in section 6.1, the TREC 2005 data has more occurrences of definite descriptions than the TREC 2004 data. In this section, I will look closely at how these definite descriptions are used in the QA discourse and provide a computationally feasible categorization of the uses.

Parallel to the study of the definiteness and indefiniteness of noun phrases, several classification schemes have been described in literature, either for the purpose of theoretical study or for the purpose of natural language processing. Just to name a few, they are Hawkins [111], Prince [97], Fraurud [95] and Poesio and Vieira [96]. The definite descriptions in my data are classified mainly based on these four works.

Hawkins' classification is the most descriptive one and fundamentally important to the later works. Hawkins (1978) presents a very fine-grained enumeration of definite description classification. By reviewing his work, I intend to explore the feasibility of applying the classification to the definite descriptions in my data. Along with the brief review, some examples of definite description uses will be presented. They are mostly drawn from the TREC 2004 data and my user study data.

6.5.1 Fine-grained classifications

Based on the work of Christophersen [101] and Jespersen [112], Hawkins [111] discusses 6 significant uses of the definite article *the*. These uses are: anaphoric use, immediate situation use, larger situation use, associative anaphoric use, unfamiliar use and unexplanatory modifier use. From the detailed classification to be discussed below, we will see that some of the definite description uses are anaphoric, as the familiarity theory predicts, and some are not. Besides, the identification of the usage type sometimes relies on the

semantic and pragmatic interpretation of the definite description, and sometimes on its syntactic features.

The first significant use is the *anaphoric use*. It refers to the use where a definite description refers to a discourse entity that is already in the discourse. Following Sidner [51], the term *cospecify* is intended to mean that both the definite description and its antecedent denote the same object. This use is common in the TREC data. The occurrences of definite descriptions in (9Q3), (10Q2) and (11Q3) (i.e. *the court*, *the organization*, *the awards* and *the prize*) indeed cospecify with discourse entities introduced in the early discourse. (i.e. *the international criminal court*, *AARP* and *the Nobel prize awards* respectively.)

(9)

Q1: When was *the international criminal court* established?

Q2: What kind of cases does it try?

Q3: Who is the sponsor of *the court*?

(10)

Q1: What does *AARP* stand for?

Q2: When was *the organization* started?

(11)

Q1: Who established *the Nobel prize awards*?

Q2: When were *the awards* first given?

Q3: What is the monetary value of *the prize*?

The examples show the typical ways in which an anaphoric use is usually characterized:

- the definite description could share the same descriptive predicate as in the antecedent (e.g. *the international criminal **court*** with *the **court***);
- use hypernym (e.g. *organization* is the hypernym of *AARP*);
- use synonym (e.g. *prize* is synonymous to *awards*) etc. to indicate the same referent.

The second significant type has two subtypes: *visible situation use* and *immediate situation use*. These two uses do not have any occurrences in my data pool, because these two

uses involve spoken language instead of written language. Basically, the QA discourses that I am concerned with are in a written text form.

The third use is the *larger situation use* (with two subtypes: larger situation uses with specific knowledge about the referent, and with general knowledge about the referent). Speakers/writers are sometimes in anticipation of the hearer/reader's knowledge of entities that exist in the non-immediate or larger situation of utterance. Definite descriptions are thus used to indicate that both the speaker/writer and hearer/reader know about the existence of the referents, or in other words, share knowledge of the referents. For example, the referent could be an entity/object known to community members (such as the definite description *the president in 2006* referring to George W. Bush is known knowledge for the people in the U.S.) or an object known to everyone on earth, such as *the sun* or *the moon*. Consider (12) as an example for a larger situation use.

(12)

Q1: When were the 2004 presidential debates?

Q2: What did the first debate cover?

Q3: Who won?

Q4: What was covered in the 3rd debate?

In the feeding question, by using *the 2004 presidential debates*, the user expects mutual knowledge of the presidential debates taking place in the U.S. in 2004.

The fourth significant use of a definite description is the *associative anaphoric use*. It indicates that the speaker and hearer share knowledge of the relations between entities, or properties/attributes, or the components of the entities. For example in (13), *the crew* is associated with *the Challenger space shuttle* because normally a space shuttle is supposed to have a crew that performs various services on the flight. This semantic relation is common sense knowledge shared by the speaker/writer and the hearer/reader. In other words, *the Challenger space shuttle* triggers the use of *the crew*⁵.

(13)

⁵In Sidner[51], this type of usage is also discussed and such triggering phrases as *the Challenger space shuttle* are named *associators*.

Q1: Which was *the first space shuttle*?

Q2: When was *the Challenger space shuttle* disaster?

Q3: How many members were in *the crew*?

The other two uses that I am about to present are not as commonly seen as the first four. These uses are “unfamiliar” in the sense that the entities denoted by the definite descriptions are not previously introduced in the discourse, neither is the knowledge on them shared by the speaker and the hearer in one way or another, thus unfamiliar to the hearer.

The fifth significant type the *unfamiliar use* is further classified into 4 subtypes. These four subtypes are characterized as having certain syntactic features associated with the definite descriptions. These features could be one of the following:

- referent-establishing relative clauses: the referent is introduced into a discourse by a relative clause whose content helps to identify the referent. In example (14), the relative clause *that destroyed Pompeii* helps to establish a volcano as the referent.

(14)

Q1: What is the name of the volcano *that destroyed Pompeii*?

Q2: What month did this first happen?

- associative clause: this use occurs when an associative relationship, especially genitive relationship holds between the expressions in question. *Of the last debate*⁶ in (15) helps the hearer to identify and locate the referent in question.

(15)

Q1: What was the format *of the last debate*?

- nominal modifiers: in example (16), the modifier *the band* refers to the class to which the head noun *Clash* belongs. In other words, the modifier is a hypernym of the head noun.

(16)

Q1: what kind of music does *the band* Clash play?

⁶*the format of the last debate* could be converted to *the last debate's format*.

- NP-complement: the definite description appears with a complement to the head noun, as shown in (17).

(17)

S1. Bill is amazed by the fact *that there is so much life on earth*[italics added].
(Hawkins[111], p.140)

Finally, the sixth significant use is the *unexplanatory modifiers use*. In (18), *the largest*, *the same* and *the first contribution in Hawaii* are classified into such a category because they do not function to establish any definite referent to the hearer.

(18)

Q1: Is tourism *the largest contribution in Hawaii*?

Q2: Was the winner of this debate *the same as the first*?

Clearly, the classification employed by Hawkins is quite sophisticated covering many detailed uses of definite descriptions. The basic idea of his classification aligns itself with the familiarity theory on the semantics of definite descriptions.

6.5.2 Prince (1992)

Different from Hawkins, Prince (1992)[97] discusses the uses of noun phrases from a new angle. Only two aspects are taken into consideration: hearer status and discourse status. It should be noted that, this classification is also along the lines of the familiarity analysis of definite descriptions described in Christophersen [101] and Heim[102]. A discourse entity could be *familiar* with respect to a discourse as well as to a hearer.

Table 6.2 shows the taxonomy of classifying discourse entities. Discourse entities may be *hearer-new* or *hearer-old* depending on the speaker/writer's beliefs about the hearer/reader's belief. If the speaker assumes that a discourse entity is "brand-new", not to be known to the hearer, it is then classified as *hearer-new*. On the other hand, if in a situation, the speaker takes an entity to be known to the hearer, he/she would most likely use a proper name or a definite description to represent this so-called *hearer-old* information. However, the definiteness of a noun phrase does not necessarily reflect the

hearer-status. In addition, as Prince pointed out, this classification highly relies on the actual situation, which means a *hearer-new* reference for one may turn out to be *hearer-old* for another.

Table 6.2. Hearer and discourse status of a discourse entity*

	Discourse-new	Discourse-old
Hearer-new	Brand new	NA
Hearer-old	Unused	Evoked

*reproduced from Prince 1992[97]

Discourse-new and discourse-old are the other pair of names used to describe discourse entities. As the name indicates, discourse entities could be classified to be new or old with respect to their discourse status. This idea echoes with Heim's familiarity theory in that discourse-old entities are the ones that have been previously evoked. Discourse-new entities are the ones that have not been mentioned in the previous discourse.

From the discussion, we see that Hawkins's anaphoric uses of definite descriptions are for the *discourse-old* entities, while his larger situation and immediate situation uses are often for the *discourse-new* and *hearer-old* entities. In addition, in Prince's theory, the *inferred* can be described by these two information schemes as well. That is, the *inferred* are technically *discourse-new* and *hearer-new*, but are not entirely new. Those entities are related to the *hearer-old* entities, therefore they could be inferred. This use is similar to Hawkins' associative anaphoric uses (such as *the door* and its triggering noun phrase *a building*). Finally, Prince's *containing inferred* are *inferred* yet the entities are identified with the help of the descriptive content specified in the noun phrases. Definite descriptions involving NP complements, referent-establishing relative clauses and associative clauses could be examples of the *containing inferred*, which is similar to Hawkins' unfamiliar use.

6.5.3 Poesio and Vieira (1998)

Another coarse-grained classification of definite description uses is presented in Fraurud [95]. As mentioned in section 6.3, Fraurud only uses two classes: *subsequent-mention* and

first-mention. Fraurud's *subsequent-mention* is similar to Hawkins' anaphoric use and Prince's discourse-old.

Based on the taxonomy modified from Hawkins and Prince's work, three uses are classified in Poesio and Vieira [96]: *direct anaphoric*, *discourse new* and *bridging*. *Direct anaphora* covers the cases where subsequently mentioned definite descriptions corefer with an antecedent that has the same head noun as the description; It differs from the anaphoric use classified by Hawkins or Prince's textually evoked classes in that it only allows the cases where the antecedent shares the same head noun with the definite description. To be more specific, this use does not include more complicated anaphoric relations such as synonyms or hypernyms etc. *Discourse new* includes definite descriptions that are used to "denote entities not related by shared associative knowledge to entities already introduced in the discourse" ([96],p.542), which is similar to Prince's discourse-new. *Bridging* covers two cases: 1) definite description having an antecedent denoting the same discourse entity, but using a different head noun; 2) Definite descriptions are similar to Hawkins' associative, and to the *indirect reference by association* described by Clark 1977 [113]. According to Clark, the associated information has different levels of predictability from the entity, event or situation mentioned earlier. The most typical associations are *Necessary part*, *probable part* and *inducible part*. Examples of *necessary part* would be the *ceiling* of a *room* and the *author* of a *book*. However, the *window* may be a probable part of a *room*. The inducible part is an associated part that one would not normally predict but has to be inferred. For example, the use of *chandelier* is associated to a *room* by hearer's inference, given that a room does not necessarily have a chandelier.

Having reviewed the classifications of the definite description uses in literature, we now look at how the definite descriptions are categorized in my study.

6.5.4 Classifying definite descriptions in QA discourse

After the preliminary examination of my data pool, I find that it is hard to adopt any of the classifications mentioned above without making any changes. If I adopted Hawkins' classification scheme, the number of occurrences for each use varies widely. For example,

in the TREC04 data, there are no occurrences of visible situation use and immediate situation use at all. There are only 3 cases of referent-establishing relative clauses out of 88 definite descriptions, while there are 46 cases of anaphoric use. Given that Hawkins' classification has 6 major classes (8 if including the subtypes) and the distribution from the TREC data and my data, it is not computationally beneficial for a computer system to identify each of these classes. Therefore I do not use this fine-grained classifications for my implementation. The study in Poesio and Vieira [96] also shows that the more classes a classification scheme has, the more difficult it is for human to agree on the annotation of the definite description uses. Therefore complicated classification schemes may cause a problem in evaluation.

Prince's[97] classification takes into account the hearer status, which is not relevant in my case. As I have mentioned, I focus on the written QA discourse modeling. It would have been a good categorization if the study were on the user interactive QA. Fraurud's [95] classification is much simpler than Hawkins', yet I feel it is way too coarse grained for my research purpose. In order to examine the behavior and the role of definite descriptions in QA discourse processing, I would prefer a classification that is not too simple, yet not too complicated.

The three-class taxonomy in Poesio and Vieira [96] is adopted in my study, with some modifications made to suit my purpose. Based on the three classes in Poesio and Vieira [96], I categorize the definite descriptions in my data pool into three classes: discourse-new, direct anaphoric and bridging. Based on the three classes, I then manually annotate all the definite descriptions in the data for further processing. The classification is based on the discourse status rather than the hearer status. *Discourse-new* conveys that the entity that the definite description refers to has never occurred in previous QA discourse. The classification is rather strict. For example, the definite descriptions that are part of common knowledge and normally would be categories as hearer-old (such as *the U.S.* or *the U.N.*) are treated as discourse-new if they never appear in the questions before. That is, all the *larger situation uses* in Hawkins' term will be annotated as discourse-new if they have not been mentioned in the previous QA discourse.

Bridging means that the entity is associated with an entity that appears in the previous discourse. Now consider example (19) where the definite description *the company's headquarters* in (19Q2) is used as *bridging* because the entities that *the company's headquarters* refer to are associated with the entity Conde Nast. Headquarters are common to a big company, in other words. *The publishing company Conde Nast* in (19Q1) is used as discourse-new in this example. Different from the bridging use that is described in Poesio and Vieira [96], I restrict the annotation of the bridging uses based on a mutual-containing relationship, which will be discussed in section 6.6.2. Briefly, I will identify the use of *the company's headquarters* as *bridging*. Based on this usage, the expression *the publishing company Conde Nast* for the associated entity in (19Q1) will then be used for query expansion in a definite model, which will be described in section 6.6.2.

(19)

Q1: Where is the publishing company Conde Nast?

Q2: Where are the company's headquarters?

Direct anaphoric means that the entity is the same entity that has been mentioned in the earlier discourse. For the ease of annotation, I follow the strict interpretation mentioned in Poesio [96], only considering the expressions with the same head and referring to the same entity. For example in (20), *the center* is used as anaphorically because it refers to the same entity that *the Berkman Center for Internet and Society* in (20Q1) refers to. Both expressions share the same head *center*.

(20)

Q1: Where is the Berkman Center for Internet and Society located?

Q2: When was the center formed?

Next, I present the discourse models developed to process the QA discourses.

6.6 Discourse models focusing on definite descriptions

Readers may have noticed that the discourse models presented in Chapter 5 have not used the *target(topic)* information of the question set, although the targets are explicitly provided in the TREC 2004 data. In this section, I present two sets of experiments that are conducted under two different conditions: 1) the QA system is provided with the targets; 2) the QA system does not know the targets. The reasons I want to examine the two situations are that on the one hand, I would like to see how the presence of a known target would affect the discourse modeling; on the other hand, I would like to see how well a QA discourse would benefit from extra processing based on linguistic knowledge beyond known targets.

For the convenience of presentation, I call these two conditions as the no-target situation and the target-given situation. Although TREC data provides the *target* information, in practice, the target information may not be provided in the first place, but rather needs to be inferred from the context. The design of the no-target situation is motivated by such circumstances. In the target-given situation, before actually asking questions, the users explicitly inform the computer system: “Hey, this is what I want to know about.” Before I elaborate the models developed for these two situations, it is necessary to discuss the questions that I aim to answer:

- What is the role of a target? We have seen that most participating TREC systems made use of the targets (discussed in Chapter 2), but we also see the potential problems mentioned at the beginning of this chapter. What is the difference that a discourse processing model would make in the no-target and target-given situations if everything else remains the same? What would be the most efficient way for a system to intelligently store the relevant context for processing later questions?
- What is the role of the definiteness hierarchy in processing QA discourse? How can it be used? I implemented the anaphora model in Chapter 5 using the centering

algorithm. To resolve pronouns, the discourse entities were ranked and then the antecedent was picked according to the ranking of the forward looking centers. Is it possible to pick the antecedent according to the NP type ranking in the definiteness hierarchy?

- How does the classification of the definite descriptions help with the processing?
What are the potential implications for my system?

In order to build discourse models to explore these questions, the TREC 2004 and TREC 2005 data are annotated accordingly. Next I will describe the data separately.

6.6.1 Data analysis

Data annotation

In addition to the annotation entries introduced in Chapter 4, I add more linguistic information to the XML annotation of the TREC 2004 and 2005 data. For the target-given situation, I add the target to the question sets and make them the first question of each set. I label it as *q0* as other questions are incrementally labeled as *q1*, *q2*, etc. The targets are annotated the same way as other questions. The example (5) in Chapter 4, repeated as (21) below shows the basic question information and linguistic information annotated for the questions. Because this part of the research is also entity-based, my major concern is the noun phrases, definite noun phrases in particular. The semantic information (such as the *NP type*) as well as the syntactic information (such as the *grammatical role*) of these noun phrases are also annotated. For a definite description, the head noun, its pre-modifiers⁷ and post-modifiers⁸ are annotated as well. In online processing, these types of information can be automatically acquired through parsing. Based on the three classes introduced above, the use of the definite description is also annotated. For example, for the definite description *the first Crip gang*, *gang* is annotated as the head, *first Crip* as the premodifier, *new* as to indicate *discourse-new*. Entities that

⁷Modifiers appearing before the head noun.

⁸Modifiers appearing after the head noun, including relative clauses, NP complements etc.

appear as modifiers for other entities are also annotated. The only annotation difference for the TREC data in the two situations is that the targets are removed for the no-target situation, leaving everything else (i.e. the annotation tags) the same.

(21)

```

<qa qaid="01" t="Crips" type="ORG">
  <q_id="q0">
    <orig>Crips</orig>
    <TOPIC start_tok="1" end_tok="1">
      <ENTITY sem_type="object-plural" gram_role="subject">
        <PROPERTY prop_type="proper_name" start_tok="1" end_tok="1"/>
      </ENTITY>
    </TOPIC>
  </q>
  <q_id="q1">
    <orig>When was the first Crip gang started ? </orig>
    <TOPIC start_tok="7" end_tok="7">
      <ACTIVITY>
        <ENTITY sem_type="person-singular" gram_role="subject">
          <PROPERTY prop_type="definite_des" start_tok="6" end_tok="6"/>
          <CONSTRAINT constraint_type="premodifier" start_tok="4"
end_tok="5"/>
          <DEF_TYPE def_type="new"/>
        </ENTITY>
        <ENTITY sem_type="object-singular" gram_role="possessor">
          <PROPERTY prop_type="proper_name" start_tok="5" end_tok="5"/>
        </ENTITY>
      </ACTIVITY>
    </TOPIC>
  </q>

```

Example (22) shows a context question set from the TREC data. We see that (22Q1) uses the pronoun *it* to refer to the target *Camp David*. In order to process (22Q1), a system has to know the target first. There are 7 such cases in the TREC 2004 data and 12 in the TREC 2005 data where the pronouns in the feeding questions refer to the

targets.

(22) Target: Camp David

Q1: Where is it?

Q2: How large is it?

Q3: What was it originally called?

Q4: When was it first used?

Q5: What U.S. President first used it?

There are also cases like (23) where the pronouns appearing in the subsequent questions refer to the target *Longwood Gardens*.

(23) Target: Longwood Gardens

Q1: When was the initial land purchased?

Q2: Where is it?

Q3: How large is it?

Q4: Who created it?

Q5: How many visitors does it get per year?

Q6: When is the best month to visit the gardens?

Considering examples like (22) and (23), some of the original TREC feeding questions are modified so as to provide necessary information. The modifications could be one of the followings:

- substitute pronouns with the target:

Q1: When was he born?

=>When was the architect Frank Gehry born?

- complete proper names:

Q1: When was Guthrie born?

=>When was Woody Guthrie born?

- complete the event situation

Q1: When did the school shooting occur?

=>When did the Kip Kinkel school shooting occur?

For example (23), the feeding question (23Q1) was modified to:

Q1: When was the initial land of Longwood Gardens purchased ?

From my study, I find that users may not explicitly provide a target and then specify a sequence of questions. The first question specified in a sequence tends to provide the context and implicitly indicate the target/topic of the sequence. The main purpose of the modifications is to provide enough information for the feeding questions, because without an explicitly provided target, the processing of the subsequent questions will have to rely on the feeding question.

Data analysis

Table 6.3. Data analysis for the original TREC 2004 and TREC 2005 data

Number	TREC 2004	TREC2005	Diff%*	Total
Question set	65	75	15.4	140
Factoid questions	230	362	57.4	592
Question that has no answer	26	30	15.4	56
Pronoun	132	102	-22.7	234
Question that has pronoun	127	97	-23.6	224
Pronoun per question	0.574	0.282	-50.9	0.395
Definite description	84	239	185	323
Question that has DD**	74	188	154	262
DD per question	0.365	0.66	81	0.546
Discourse new	58(69.1%)	113(47.3%)	94.8	171(52.9%)
Direct anaphoric	19(22.6%)	69(28.9%)	263	88(27.3%)
Bridging	7(8.3%)	57(23.8%)	714	64(19.8%)

* the increase of TREC 2005 data compared with the TREC 2004 data

** Definite description

Now, let us look at the data characteristics of the TREC 2004 and TREC 2005. Table 6.3 shows the data distribution statistics for the TREC questions. There are two major differences between the TREC 2004 and TREC 2005 data: the decrease of the pronoun occurrences and the increase of the definite description occurrences in the TREC 2005 data compared with the TREC 2004 data.

The number of pronoun per question of the TREC 2005 (0.282) decreases 50.9% compared with the TREC 2004 data (0.574). This poses challenges for the pronoun resolution oriented approaches. The number of definite descriptions per question in the TREC 2005 data (0.66) increases 81% compared with the TREC 2004 data (0.365).

The total number of questions that have definite descriptions in the TREC 2005 data increases 154%, while the total number of questions only increases 57.4%. Also, the major use of the definite descriptions is *discourse-new*, 69.1% for the TREC 2004 data and 47.3% for the TREC 2005 data. To be more specific, more than half of them (52.9%) are discourse-new for the total TREC data. Only 27.3% definite descriptions are anaphoric and 19.8% are bridging.

The definite description distribution of the TREC data is very similar to what is reported in Poesio and Vieira [96]. They have 50% of discourse-new, 30% anaphoric and 18% bridging/associative. Poesio and Vieira reveal that in certain genre of texts, “the definite descriptions are not primarily anaphoric”. Therefore I hope my investigation of the definite descriptions in context questions will cast new insights into generalizing their behavior. Table 6.4 shows the data distribution of the modified data sets for the no-target

Table 6.4. Data analysis for the modified TREC 2004 and TREC 2005 data

Number	TREC 2004	TREC2005	Diff%*	Total
Question set	65	75	15.4	140
Factoid questions	230	362	57.4	592
Question that has no answer	26	30	15.4	56
Pronoun	125	90	-28	215
Question that has pronoun	120	85	-29.2	205
Pronoun per question	0.543	0.249	-54.1	0.346
Definite description	99	248	151	347
Question that has DD**	83	193	133	276
DD per question	0.43	0.685	59.3	0.586
Discourse new	73(73.7%)	121(48.8%)	65.8	194(55.9%)
Direct anaphoric	19(19.2%)	69(27.8%)	263	88(25.4%)
Bridging	7(7.1%)	58(23.4%)	729	65(18.7%)

* the increase of TREC 2005 data compared with the TREC 2004 data

** Definite description

situation. Because I substitute some pronouns in the original feeding questions with the corresponding targets, I then have more definite descriptions (i.e. total 99 instead of 84 for the TREC 2004 data, 248 instead of 239 for the TREC 2005 data). The fact that more definite descriptions are added to the feeding question can explain the increase of the *discourse-new* occurrences(55.9%) compared with the original TREC data (52.9%). In general, I observe consistent pronoun and definite description distribution for the modified TREC data.

In the coming sections I describe my discourse modeling in detail. First I present the experiments on the modified TREC data where the targets are not explicitly specified, and then the experiments conducted on the original TREC data with the targets explicitly specified.

6.6.2 No-target situation

First I explore the situation where no targets are explicitly given for the context question sets. Under this condition, the processing on all the later questions will have to solely rely on the context from the previous question(s). I design three models to investigate the issues I have discussed at the beginning of section 6.6: 1) a pronoun-resolution model(I also rename it a *pronoun model*); 2)a definite description resolution model (I rename it a *definite model*); 3) a combined-resolution model (I rename it a *combined model*); The pronoun model aims to resolve all the pronouns appearing in the questions. Unlike the centering-based approaches described in Chapter 5 for the anaphora model, here the pronouns are resolved based on the definiteness hierarchy. The definite model aims to resolve the definite descriptions in the questions and locate the antecedent that each definite description refers to. The combined model resolves both the pronouns and the definite descriptions, therefore it is an incorporation of the pronoun model and the definite model. The separation of the resolution components (the pronoun model and the definite model) enables us to examine the individual effect that they have on the entire processing. Meanwhile, the combined model allows us to see the overall effect on pronouns and definite descriptions has on the QA discourse modeling.

Pronoun model

Recall that in Chapter 5, I have presented an anaphor model that resolves pronouns using a Centering algorithm [1]. However, here I employ a different mechanism to resolve pronouns for the pronoun model. I use some semantic constraints and the definiteness hierarchy instead. The semantic constraints include the number, person, case, and gender agreement between the pronouns and their potential antecedents. They are used to filter out the unqualified discourse entities. If a pronoun appears in question Q_n , then this filtering is conducted in a backward fashion, that is, first to filter the entities in Q_{n-1} , then Q_{n-2} , so on and so forth until a proper candidate is found. If more than one candidates are left after the semantic constraint filtering, then the definiteness hierarchy is used to rank the candidates. Whichever entity that is ranked higher will be assigned as the antecedent of the pronoun. The modified definiteness hierarchy is shown as follows, where major NP types are ranked according to their definiteness.

(24)

Pronoun > Definite description > Proper name > Indefinite > Other

It should be noted that if a potential candidate is a definite description and this definite description has to be resolved further, the pronoun model will not further resolve the candidate definite description. For example, *its* in (25Q3) will only be resolved to the definite description *the center* in (25Q2) instead of being resolved to its real antecedent *the Berkman Center for Internet and Society*.

(25)

Q1: Where is *the Berkman Center for Internet and Society* located?

Q2: When was *the center* formed?

Q3: What is *its* mission?

By doing so, I try to separate the influence of resolving definite descriptions, which is my focus in the definite model.

Definite model

The definite (resolution) model intends to resolve all the definite descriptions in the questions. In other words, for each definite description, it attempts to find an antecedent for it in the QA discourse if there is any. The discourse-new entities are defined as the entities that never occur in the previous discourse, therefore they do not have antecedents. In my study I mostly focus on the processing of the definite descriptions of the other two types: direct anaphoric and bridging. The basic idea is to locate the antecedent and then inherit relevant information of the antecedent entity to help with query expansion.

Now I describe in detail the steps that are taken to build the model: 1) the identification of the direct anaphoric use of the definite descriptions; 2) the identification of the bridging use of the definite descriptions; 3) the identification of the discourse-new use of the definite descriptions; 4) the strategy for query expansion.

The identification of the direct anaphoric definite descriptions is based on a mechanism called NP head matching. The following situations are allowed or regarded as *NP head matched* between a definite description and its antecedent:

- head-matching for the premodified definite description containing less information than the antecedent that has premodifiers; For example, the definite *the first debate* has less information than its antecedent *the first 2000 presidential debate*;
- head-matching for the premodified definite description containing additional information than the antecedent that does not have premodifiers; For example, the premodified definite *the first Auto Show* has more information than its antecedent *the show*.

These two rules are mentioned in Poesio and Vieira [96]'s work, and they also mention about the other two rules that do not constitute as a head match.

- head-matching for cases where both the definite and the antecedent are postmodified and the modifications are not the same; For example, *the Bollywood equivalent of Beverly Hills* and *Bollywood 's equivalent of the Oscars* both have the same head noun *equivalent* but different postmodifiers *of Beverly Hills* and *of the Oscars*.

- head-matching for the premodified definite containing more information than the antecedent that has premodifiers; For example, the premodified definite *his second successful career* has more information than the antecedent *his primary career*.

I relax the criteria so the two situations above do constitute a head match, because after the preliminary examinations, I found that these two head-matching rules would mostly result in correct resolution. This can be explained by the fact that the context question sets are all about specific topics. Intuitively, it is unlikely for a series to explore two different entities whose corresponding expressions have the same head.

The bridging uses are identified by a mutual-containing relationship between the definite description and its antecedent. Mutual-containing is a rule that I define to identify different but related entities. Entities may relate to each other by sharing certain properties. Expressions that refer to bridging entities were related through the following ways:

- mutual-containing occurs when the antecedent appears as either the premodifier or the postmodifier of the definite; For example, the antecedent *space shuttle* appears as the premodifier in the definite *the space shuttle disaster*.
- mutual-containing occurs when the definite appears as either the premodifier or the postmodifier of the antecedent; For example, the definite *the Challenger* appears as the postmodifier of the antecedent *the crew of the Challenger*.
- mutual-containing occurs when the definite and the antecedent share the same premodifier or postmodifier. For example, the definite *the first flight* and the antecedent *first shuttle* share the same premodifier *first*.

In general, the mutual-containing rules attempt to establish a relationship between the definite description and its antecedent by their lexical ties. Definite descriptions that are of the traditional bridging use will not be considered as bridging if the mutual-containing rule cannot be applied, for instance, the definite description *the knob* and its antecedent *door*. *The knob* will then be identified as *discourse-new* if it appears for the first time in the discourse. This kind of bridging, however, is possible for a system to identify with the help of an extra semantic knowledge resource, such as *WordNet*.

Based on the head-matching and the mutual-containing rules, if a definite description is neither a *direct-anaphoric* nor a *bridging*, and if it appears in the discourse for the first time, it will then be identified as *discourse-new*.

Once the definite descriptions are identified⁹, I then use the heading-matching and the mutual-containing rules to find the antecedents of the *direct anaphoric* and *bridging* definites¹⁰. To look for the antecedent for the definite in question Q_n , the definite model will process the entities in its previous questions Q_{n-1} , then Q_{n-2} until an antecedent is found according to the rules. In the models I implement, the strategy for query expansion is to inherit all the terms from the antecedent (of the definites of *direct anaphoric* use or *bridging* use) and add them to the current question for query expansion.

Combined model

The combined model incorporates the pronoun resolution element with the definite resolution element. It is a joint effort of providing more linguistic knowledge to the processing of the context questions.

6.6.3 Target-given situation

Now I turn to discuss the situation where the targets are provided. As I have mentioned in Chapter 2, most participating 2005 TREC QA systems directly append the corresponding target to the questions that contain pronouns for query expansion. Under the assumption that a target is given for a question set, I conduct a series of experiments aiming to address the following issues:

- What is the role of an explicit target for a context question set? Is it all that is needed to process the QA discourse?
- How does the target-given situation affect the pronoun resolution and the definite description resolution? Are the resolutions still necessary for the purpose of query expansion?

⁹The definite description uses are manually annotated in the XML according to the rules.

¹⁰The identification of the antecedents are implemented automatically in the system.

- What is the difference on the situations where targets are explicitly specified or implicitly inferred?

In order to answer these questions, I design several experiments that aim to separate the influence of the target from the influence of the extra processing for pronouns and definite descriptions. The following models are examined in detail: a baseline model, a pronoun model, a definite model, a combined model, a target model, and a combined target model.

Baseline

The baseline takes the whole previous question and concatenates with the current question for query expansion. This model assumes that the previous question provides context for the processing of the current question.

Pronoun model

The pronoun model is the same as the pronoun model in the no-target situation except that if a pronoun cannot be resolved either by the semantic constraints or the definiteness hierarchy, it will be resolved to be the target. Extra efforts are made by this model to resolve all the pronouns which include both the target-referring pronouns and other pronouns such as *it* in (26Q4).

(26) Target: Rose Crumb

Q1: What was her occupation?

Q2: Where was she from?

Q3: What organization did she found?

Q4: When did she found it ?

Q4': Query for Q4 in the pronoun model: {When, did, she, found, it, Rose, Crumb, organization}

Retrieval result for Q4 in the baseline model: DocRank= 0; MRR= 0.0

Retrieval result for Q4 in the pronoun model: DocRank= 2; MRR= 0.5

Therefore the query expansion for (26Q4) would look like (26Q4'), where *she* is resolved

to the target eventually, after the pronoun resolution algorithm finds out that it cannot be resolved to any entities from any preceding questions of (26Q4). Meanwhile, *it* in (26Q4) will be resolved to *organization* through the semantic constraint rules, because only the entity organization matches with *it* in terms of number and gender. Again, the definiteness hierarchy is used to rank the candidate entities if there are more than one entities left after the semantic constraint filtering. The retrieval result shown in (26) indicates that the pronoun model performs better than the baseline model in terms of MRR.

Definite model

The definite model is to resolve the definite descriptions in the question sets and pick corresponding terms from the antecedents for the purpose of query expansion. This is the same as what I have discussed in section 6.6.2 except that targets are used to help resolving the definites. The definite descriptions are resolved to their antecedents in the preceding questions. For different uses (i.e. *discourse-new*, *direct anaphoric* and *bridging*), I select the terms from the antecedents to expand the question under processing. In section 6.6.2, I do not process any expressions that correspond to the discourse entities used as *discourse-new*. However, now that the target is given, it is more likely that the *discourse-new* entity is the entity that the target refers to. Therefore, I match the head noun of the *discourse-new* entity to the target entities. If there is a head-matching between the *discourse-new* entity and any of the target entities, then I identify the matched target entity to be the antecedent of the *discourse-new* definite. For example, *Berkman Center for Internet and Society* is the target for the questions in example (27). The definite description *the center* is annotated as *discourse-new* because the entity appears for the first time in the mini QA discourse. However, the definite description *the center* head-matched ¹¹with the target noun phrase *Berkman Center for Internet and Society*, for both of which share the same head *center*. The identification rules for the antecedents of the *direct anaphoric* and *bridging* definites in this model are the same as what I employ

¹¹See the discussion of head-matching in section 6.6.2

in the no-target situation.

(27)

Q1: Where is *the center* located?

Q2: When was the center formed ?

Q3: What is its mission?

After the antecedents of the definites are identified, the linguistic expressions corresponding to the antecedents are used to form queries for the questions currently being processed. Namely, for example (27), the final query terms for (27Q1) would be {Where, is, the, center, located, Berkman, Center, for, Internet, and, Society}. Recall that the query terms are input into the retrieval engine as a bag of words, so the repeated terms such as *center* are handled by the retrieval engine.

Combined model

The combined model is to resolve both the pronouns and the definite descriptions described above. This incorporated model attempts to resolve different types of references that may occur in QA discourses.

Target model

The target model is to expand the questions by adding the target terms to each individual question; For example, in example (28) the target is provided as *Rose Crumb*, so the query expansion for the target model would give us the input as shown in (29).

(28)

Q1: What was her occupation?

Q2: Where was she from?

Q3: What organization did she found?

Q4: When did she found it?

Q5: How old was she when she won the awards?

(29)

Query for Q1: {What, was, her, occupation, Rose, Crumb}

Query for Q2: {Where, was, she, from, Rose, Crumb}

Query for Q3: {What, organization, did, she, found, Rose, Crumb}

Query for Q4: {When, did, she, found, it, Rose, Crumb}

Query for Q5: {How, old, was, she, when, she, won, the, awards, Rose, Crumb}

In this example, the pronouns in (28) are mostly resolved to *Rose Crumb* implicitly by concatenating the target terms. However, *It* in (28Q4) is not resolved in this model.

Combined target model

Finally, the combined target model is to resolve both the pronouns and the definite descriptions with the help of the target information. In addition to the combined model where the unresolved pronouns and the discourse-new definites are processed using the information from the targets, this model simply concatenates the target terms to each question as an enhancement to provide more information even if there is no pronoun or definite in the question.

6.7 Evaluation for definiteness hierarchy based models

6.7.1 Evaluation for the no-target situation

The evaluations are conducted using the same metrics as I have used in Chapter 5. The MRR (Mean Reciprocal Ranking) and *coverage* are used as major evaluation methods. Results on the TREC 2004 and 2005 data are also presented separately to show the effects that different models have on them. The performance results under the assumption that no specific targets are given will be presented first.

Six models are implemented for the no-target situation: 1) a baseline model that takes the whole previous question to concatenate with the current question for query expansion; 2) a pronoun model that resolves the pronouns in the TREC questions; 3) an extensive pronoun model (to be discussed below); 4) a definite model that resolves the definite

descriptions in the TREC questions; 5) a combined model that resolves both the pronouns and the definite descriptions;

When I implement the pronoun model, I have found that 15 out of the total 215 pronouns could not be resolved according to the semantic constraints. There are no qualified potential antecedents for the pronouns based on the constraints such as gender, number etc. I now show one example of each type that occurs in the data.

(30)

Q1: What causes *tsunamis*?

Q2: Where does *it* commonly occur?

Q3: What is *its* maximum height?

Q4: How fast can *it* travel?

Q5: What language does the term “ tsunami “ come from?

This example shows the case where a common noun is used in a plural form but referred to by a singular pronoun.

(31)

Q1: What is the primary symptom of *a cataract*?

Q2: How are *they* treated?

Example (31) is a different case where a common noun is used in a singular form but referred to by a plural pronoun.

(32)

Q1: When was *the company Harley-Davidson* founded?

Q2: Where is it based?

Q3: *They* are best known for making what product?

Example (32) shows a case where plural pronoun is used to refer to an organization.

(33)

Q1: What is Jesse Ventura 's political party affiliation?

Q2: What is his birth name?

Q3: What is his wife 's name?

Q4: How many children do *they* have?

Example (33) shows a situation where two entities in the discourse are referred to together.

(34)

Q1: What was *the Louvre Museum* before *it* was a museum?

In example (34), the antecedent of a pronoun appears in the same question rather than in a previous question. The next example shows a case where the pronoun *it* is not used for referring. *It* is used as an expletive.

(35)

Q1: When was Cassini launched for space probe?

Q2: How much did *it* cost to build?

For the cases mentioned above, again, the definiteness hierarchy is used in my implementation. I modify the algorithm to resolve the pronouns according to the ranking of the entities in the feeding question based on the definiteness hierarchy. I name it the *extensive pronoun model*. This model selects the highest ranked entity from the feeding question to be the antecedent. By doing so, the unresolved pronoun is forced to be resolved to some entity in the QA discourse, given the assumption it is related to the most prominent entity according to the definiteness hierarchy.

Table 6.5 shows the MRR performance of the 4 models that are run on the TREC data: the baseline model, the pronoun model, the extensive pronoun model, and the combined model that adopts both the extensive pronoun resolution and the definite description resolution (i.e. the definite model)¹² Table 6.6 shows the coverage information of the 4 models. Considering there are questions that have no answers¹³(sometimes the feeding questions), I separate the evaluation into two groups. The *Incl* row in table 6.5 shows the results on all the questions in my data pool and the *Excl* row shows the results on the questions excluding the ones that do not have answers. Note that for the definite model used for the combined model in Table 6.5, a strategy that inherits the whole noun

¹²The performance of the definite model itself is not shown because it is not comparable with other models.

¹³Although it was NIST staff's intension to provide test questions that had answers in the test document collections, there were still such questions existing in the TREC 2004 and TREC 2005 tracks.

Table 6.5. MRR performance based on definite hierarchy in no-target situation

MRR	Baseline	Pronoun	Diff%	ExtensivePronoun	Diff%	Combined***	Diff%
TREC04 Incl*	0.188	0.270	43.8	0.275	46.6	0.274	45.9
TREC04 Excl**	0.212	0.304	43.8	0.310	46.6	0.309	45.9
TREC05 Incl	0.220	0.242	9.7	0.247	12.1	0.266	20.8
TREC05 Excl	0.240	0.263	9.7	0.269	12.1	0.290	20.8
Total Incl	0.217	0.252	16.3	0.258	18.8	0.269	24.0
Total Excl	0.240	0.279	16.3	0.285	18.8	0.297	24.0

* Include all the questions

** Exclude the questions that do not have answers

***Combined model implements the extensive pronoun resolution

Table 6.6. Coverage performance on all TREC data* in no-target situation

Doc rank	Baseline	Pronoun	PronEx**	Combined***
1	14.0	17.0	17.0	18.0
5	30.0	34.0	35.0	36.0
10	38.0	41.0	42.0	45.0
20	43.0	48.0	49.0	52.0
30	46.0	52.0	53.0	56.0
50	53.0	58.0	60.0	62.0

* Based on the total TREC data, i.e. including all the questions

** PronEx is the extensive pronoun model.

*** Combined model implemented the extensive pronoun resolution.

phrase that corresponds to the antecedent entity to the current question in processing is applied. Different strategies could be implemented but the basic idea is to inherit relevant information of the antecedent entity to help with the query expansion. The *Diff%* columns in the table 6.5 show the MRR increasement (in percentage) of the models (the ones represented by the columns to the left of the *Diff%* columns) compared with the baseline model.

From table 6.5 and table 6.6, we see that: 1) the combined model works the best for the overall TREC data; This suggests that the extra processing of the pronouns and the definite descriptions does benefit the performance; 2) The extensive pronoun model has more advantage on the TREC 2004 data than on the TREC 2005 data. This can be explained by the fact that the TREC 2004 data has a higher percentage of pronouns and the TREC 2005 data has a higher percentage of definite descriptions; 3) The pronoun model, the extensive pronoun model, and the combined model work better than the baseline model. The result indicates that if both the pronouns and the definite descriptions are resolved, more context information is obtained than that provided by the baseline model, which takes the whole previous question for query expansion. 4) when I compare the extensive pronoun model which makes use of the definiteness hierarchy with the centering-based anaphora model in Chapter 5, I find that the extensive pronoun model works better than the centering-based anaphora model. The MRR (for all the questions) for the extensive pronoun model for the TREC 2004 is 0.275, while the MRR for the

anaphor model on the TREC 2004 data is 0.221 (see table 5.6). The performance increases 24.5%. I also notice that the transition model (MRR was 0.289) in Chapter 5 works better than the extensive pronoun model for the TREC 2004 data; 5) Finally, I notice a significant improvement if I exclude the questions that do not have answers in the evaluation.

Proper name is another important type of definite noun phrase in context questions. I have mentioned in Chapter 5 that proper names should be given a certain discourse prominence in the processing and also the strategy that I use for the transition model emphasizes its significance. Now I would like to investigate how important it is when it is put in the definiteness hierarchy. Intuitively, in information retrieval, it seems that proper name should be ranked higher than definite descriptions, for that proper names often are name entities ¹⁴ and name entity recognition has been widely used in QA processing. Therefore I switch the position of definite description with proper name in the definiteness hierarchy, and run the same models once again keeping all other elements the same. The modified hierarchy is shown below:

Pronoun > Proper name > Definite description > Indefinite > Other ¹⁵

Table 6.7 shows the average MRR for the pronoun model, the extensive pronoun model and the combined model after the modification of the ranking. Figure 6.1 shows the performance comparison between the modified models and the original definiteness hierarchy based models.

From table 6.7 and figure 6.1, we see that the performance of the modified models is worse than the original definiteness hierarchy based models on all the TREC data, especially so for the TREC 2005 data. This result indicates that the definiteness hierarchy correctly captures the discourse prominence of the noun phrases in the QA discourse and the models based on it show consistent performance advantages than otherwise.

¹⁴Name entity could be person, location, and organization, as well as times, data, percentages, money amounts, etc.

¹⁵">" in this hierarchy indicates that the NP type on the left side is more important than the one on the right side in processing QA discourse.

Table 6.7. MRR performance based on modified definite hierarchy in no-target situation

MRR	Baseline	ModPronoun	ModExtensivePronoun	ModCombined
TREC04 Incl*	0.188	0.270	0.275	0.274
TREC04 Excl**	0.212	0.304	0.310	0.309
TREC05 Incl	0.220	0.236	0.241	0.260
TREC05 Excl	0.240	0.257	0.263	0.284
Total Incl	0.217	0.249	0.254	0.266
Total Excl	0.240	0.275	0.281	0.293

* Include all the questions

** Exclude the questions that do not have answers

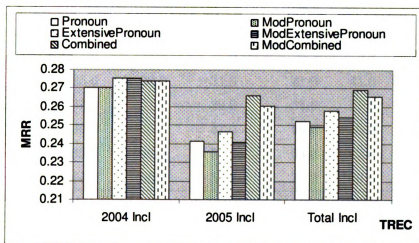


Figure 6.1. MRR Performance comparison between the definiteness hierarchy based models and the modified version

6.7.2 Evaluation for the target-given situation

Table 6.8 shows the average MRR for the models¹⁶ that are built for the original TREC data whose targets are provided along with the question sets. Table 6.9 shows the coverage information of the models for the total TREC data on all questions. Figure 6.2 shows the MRR performance of all the models based on the assumption that the targets are provided for the context question sets. It shows the MRR for all the TREC questions including the questions that do not have answers.

From table 6.8, table 6.9 and figure 6.2, I have the following observations:

¹⁶For the same reason, the performance of the definite model is not listed here.

Table 6.8. MRR Performance based on definite hierarchy in target-given situation

MRR	Baseline	Pronoun	Combined	Target	CombinedTarget
TREC04 Incl*	0.188	0.244	0.249	0.314	0.315
TREC04 Excl**	0.212	0.275	0.280	0.354	0.355
TREC05 Incl	0.176	0.217	0.226	0.288	0.287
TREC05 Excl	0.192	0.236	0.246	0.314	0.313
Total Incl	0.180	0.227	0.235	0.298	0.298
Total Excl	0.199	0.251	0.259	0.329	0.329

* Include all the questions

** Exclude the questions that do not have answers

Table 6.9. Coverage performance on all TREC data* in target-given situation

Doc rank*	Baseline	Pronoun	Combined	Target	CombinedTarget
1	11.0	15.0	15.0	19.0	19.0
5	25.0	30.0	31.0	41.0	41.0
10	31.0	38.0	39.0	51.0	51.0
20	36.0	45.0	47.0	59.0	59.0
30	39.0	49.0	51.0	63.0	63.0
50	44.0	55.0	57.0	70.0	69.0

* total TREC data, i.e. including all the questions

- The pronoun model performs better than the baseline and also it works better for the TREC 2004 data than for the TREC 2005 data. Meanwhile, the combined model does show a big performance improvement for both the TREC 2004 and the TREC 2005 data compared with the baseline.
- The target model and the combined target model perform better than the other models. This result supports the empirical practice of the most TREC QA systems. As I have mentioned in Chapter 2, appending a target to each of the questions for query expansion was one of the major methods that some TREC QA systems used in the 2004 and 2005 TREC.
- The result also shows that the performance of the combined target model is almost the same as the target model as shown in figure 6.2. It seems that the performance improvement converges even when I incorporate the target-appending approach with the extra processing of the pronouns and the definite descriptions. One possible rea-

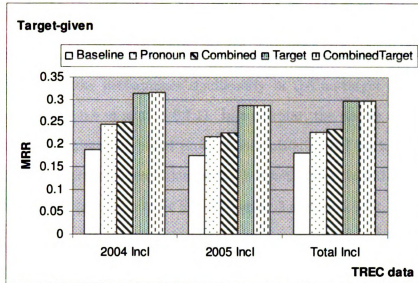


Figure 6.2. MRR performance on all TREC data* in target-given situation

* Including all the questions in the TREC 2004 and TREC 2005 data

son could be that most pronouns and the definite descriptions are actually resolved to the targets. Out of the total 592 TREC questions, only 51 questions in which the pronouns or the definite descriptions are not resolved to the targets. In other words, these cases do not affect the performance in a significant way. An interesting question arises immediately: does the target-appending method work the best for the data that is more diverse in terms of discourse entities? Obviously, a TREC question set is mostly about one target entity, sometimes about an event, and rarely about multiple entities. Because of the lack of data, I reserve this question as one of my future research topics.

Under the two conditions (i.e. no-target and target-given), I have built different discourse models to process the mini QA discourses, mostly involving pronouns, definite descriptions and targets. From figure 6.2, we see that that the two models that explicitly involved the targets (i.e. the target model and the combined target model) have a significant performance improvement over the other models (i.e. the models that do not explicitly append the targets to each question) in the target-given situation.

The results, thus far have answered the questions that I set out to investigate. In

summary, the conclusions are:

- The linguistic knowledge based models (the pronoun model and the combined model) improve the performance significantly in the no-target situation (see table 6.5 and table 6.6 in section 6.7.1). In particular, the results also indicate that the definiteness hierarchy based pronoun resolution model works better than the centering-based pronoun resolution model, but not better than the centering-based transition model (see table 5.6 in Chapter 5 section 5.5.1);
- Exactly as what the hierarchy presents, definite descriptions in QA discourses do have more important status than proper names, therefore the processing on them has been justified (see table 6.7 and figure 6.1);
- The extra processing of pronouns and definite descriptions (i.e. the pronoun model and the combined model) profits even under the target-given situation if the target-appending strategy is not used (see table 6.8 and table 6.9). This means that the strategy of appending the whole previous question for query expansion is not as efficient as to resolve the pronouns and the definites in the questions.
- The experiments show that target-appending is an efficient strategy to surpass other models (see table 6.8 and figure 6.2 in section 6.7.2).

The conclusions are meaningful because on the one hand, the belief that the linguistic knowledge is important in QA processing has been confirmed, especially the use of the definiteness hierarchy in processing the QA discourse has shown to be efficient; on the other hand, if somehow a proper target could be identified from the feeding question, then it would be easier to build a simple processing model that only appends the target to each question. However, in practice, it is not easy to identify a target only based on the first question.

6.7.3 Evaluation based on the definite description types

In order to examine the effect that different definite description types have on the performance in the no-target situation, I conduct another evaluation based on my definite description classification.

The questions are separated according to what type of a definite description they contain. If a question contains more than one definite description, the question is grouped by the ranking illustrated below: *direct anaphoric* > *bridging* > *discourse-new*. That is to say, if a question has both *bridging* and *discourse-new* definite descriptions, it will be grouped with questions that contain *bridging* definite descriptions, because *bridging* is ranked higher than *discourse-new*. Likewise, if a question has both *direct anaphoric* and *bridging* definite descriptions, it will be grouped with questions that contain *direct anaphoric* definite descriptions. The motivation of the grouping is due to the intuition that a *direct anaphoric* definite is more familiar with respect to a discourse than a *bridging* definite, which in turn is more familiar than a *discourse-new* definite. Consider example (36).

(36)

Q1: When was the *Challenger* space shuttle disaster?

Q2: How many members were in the crew of the *Challenger*?

There are two definite descriptions in (36Q2): *the crew of the Challenger* and *the Challenger*. The former is identified as a *bridging* because its postmodifier *Challenger* has mutual-containing relation with its antecedent *Challenger* in (36Q1). The later *the Challenger* is identified as *direct anaphoric* because it head-matches with its antecedent *Challenger* in (36Q1). According to my grouping scheme, question (36Q2) is identified as a question containing *direct anaphoric*.

Accordingly, I have four groups of questions: questions containing *direct anaphoric* definite descriptions, *bridging* definite descriptions, *discourse-new* definite descriptions, and questions containing no definite descriptions at all (represented as *other*). Table 6.10 shows the question distribution for the TREC 2004 and TREC 2005 data based on such a

grouping. Note that the TREC 2005 data has more questions that involve *direct anaphoric* definite descriptions (17.1%) than the TREC 2004 data (only 8.3%). Also there are more questions in the TREC 2005 data involving *bridging* (9.1%) than in the TREC 2004 data (only 2.6%). More than a quarter of the questions in both the TREC 2004 (25.2%) and the TREC 2005 (27.1%) data contain *discourse-new* definite descriptions. The number of questions that do not contain any definite descriptions in the TREC 2004 data decreases compared with the TREC 2005 data.

Table 6.10. Question distribution based on definite description types

Question number	direct anaphoric	bridging	discourse-new	other	Total
Q in TREC 2004	19(8.3%)	6(2.6%)	58(25.2%)	147(63.9%)	230
Q in TREC 2005	62(17.1%)	33(9.1%)	98(27.1%)	169(46.7%)	362
Q in TREC Total	81(13.7%)	39(6.6%)	156(26.3%)	316(53.4%)	592

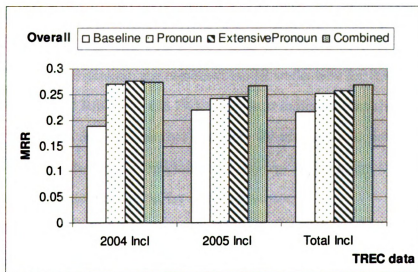


Figure 6.3. Overall comparison of four models based on automated processing on antecedents

For the convenience of the discussion that follows, I now present the results from table 6.5 using figure 6.3. Figure 6.3 shows the MRR performance on all the questions in the no-target situation. Again, in the no-target situation, the combined model works the best for the total TREC data and the TREC 2005 data. The extensive pronoun model works the best for the TREC 2004 data. Figure 6.4 shows the MRR performance for

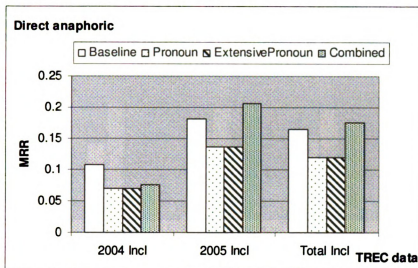


Figure 6.4. MRR performance on questions with direct anaphoric definite description in no-target situation

the questions that contain *direct anaphoric* definite descriptions. It also shows that the baseline works the best for the TREC 2004 data (8.3% of questions containing *direct anaphoric* definite descriptions), the combined model works best for the TREC 2005 data (17.1%) and the total TREC data (13.7%). The result indicates that the combined model shows more advantage on the data that has more *direct anaphoric* definite descriptions. In other words, it is effective to implement the strategy of including the antecedent of a *direct anaphoric* definite description for query expansion.

Figure 6.5 shows the MRR performance for the questions that contain the *bridging* definite descriptions. The combined model outperforms other models for the TREC 2005 and the total TREC data. This result suggests that definite description resolution is a very important element to improve the performance for the questions that contain *bridging* definite descriptions. For the TREC 2004 data, both the pronoun model and the extensive pronoun model work better than other models. This is similar to the result on the overall data shown in figure 6.3. In other words, for data that has a higher percentage of pronouns, the advantage of pronoun resolution based models is bigger than the advantage of the definite description resolution based models.

Figure 6.6 shows the MRR performance for the questions that contain *discourse-new*

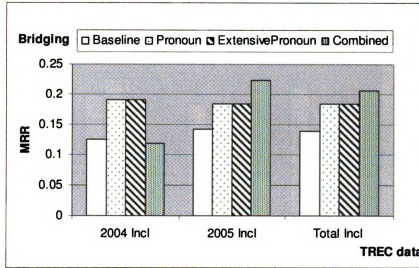


Figure 6.5. MRR performance on questions with bridging definite description in no-target situation

definite descriptions. It also shows that the pronoun model, the extensive pronoun model, and the combined model perform the same for the TREC 2004 data. This could be explained by the fact that all three models resolve pronouns. And, this element is necessary and effective for the data that has a higher percentage of pronouns. For the questions that contain *discourse-new* definite descriptions, the baseline model works the best for the TREC 2005 data and the total TREC 2005 data. This result is not surprising because in my implementation, I do not process the *discourse-new* definite descriptions for query expansion. The baseline, however, by using the whole previous question for query expansion, provides more context information than the other models.

Figure 6.7 shows the MRR performance for the questions that do not contain any definite descriptions. The extensive pronoun model and the combined model actually perform the same. This is because the definite description resolution element in the combined model does not take any effect due to the lack of presence of any definite descriptions. Besides, both the extensive pronoun model and the combined model perform better than the other models for all the datasets. This result indicates that the pronoun resolution strategy used in the extensive pronoun model is efficient, that is, to resolve some unresolvable pronoun references to the highest ranked entities (according to the

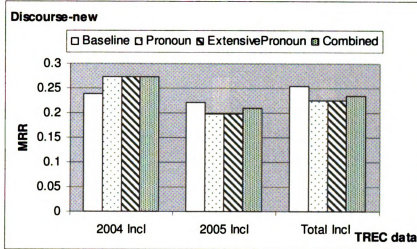


Figure 6.6. MRR performance on questions with discourse-new definite description in no-target situation

definiteness hierarchy) in the feeding questions.

In summary, I separate the questions by definite description types in this evaluation and come to the following conclusions: 1) for the *direct anaphoric* question group, the combined model outperforms other models for the TREC 2005 data and the total TREC data; For the TREC 2004 data, which has a very low percentage of *direct anaphoric* uses, the baseline works the best; 2) for the *bridging* question group, the combined model works better than other models for the TREC 2005 data and the total TREC data; For the TREC 2004 data, both the pronoun model and the extensive pronoun model work better than the other models; 3) for the *discourse-new* question group, the baseline works the best for the TREC 2005 data and the total TREC data; For the TREC 2004 data, all the three models that involve pronoun resolution work better than the baseline; 4) for the *other* question group, both the extensive pronoun model and the combined model work the best for all the datasets.

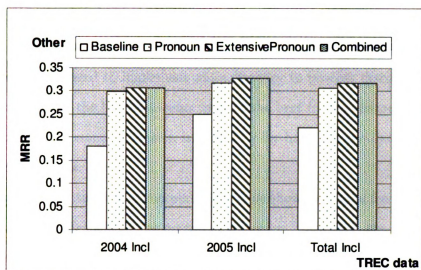


Figure 6.7. MRR performance on questions without any definite descriptions in no-target situation

6.8 Summary

This chapter discusses how to employ the linguistic knowledge of definiteness hierarchy in processing mini QA discourses. I have introduced the definiteness hierarchy of definite descriptions, the Familiarity theory, and briefly the Uniqueness theory. In order to classify the definite descriptions appearing in context questions, I then have reviewed some major classifications in literature. Based on those work, three classes of definite descriptions are presented: *discourse-new*, *direct anaphoric* and *bridging*. In the implementation part, I have separated the experiments into two groups: one for the situation where no explicit targets are provided, and the other has explicit targets. Different models have been proposed for each group.

In the no-target situation, the influence of the extra processing based on pronoun resolution and definite description resolution has been investigated. The conclusion is that the combined model which incorporates both efforts performs the best. Another evaluation is conducted based on definite description types. Questions are separated into groups to isolate the influence of different models. The conclusion is that, although definite description resolution in the definite model does not show advantage over the baseline on

the overall data, it does show big performance improvement on the questions that either have *direct anaphoric* definite descriptions or *bridging* definite descriptions. Also, the results also show the need to implement more strategy to process *discourse-new* definite descriptions. For now, the baseline seems to work the best for the *discourse-new* question group. Finally, the pronoun resolution is necessary for handling the data that has a high percentage of pronouns.

In the target-given situation, the experiments have shown that the feeding questions are not as important as they are in the no-target situation. Moreover, the target-appending strategy was shown to be very efficient.

CHAPTER 7

Conclusion

7.1 Summary

In context question answering, questions are asked in a sequence. The TREC QA tasks provide targets as to indicate what all the questions in the sequence are about. The TREC 2005 QA systems employed approaches such as *target-appending* to relate the questions with the target. However, the questions in the sequence are related to each other in a certain way. This seems to be intuitive, however little work has been done to examine the relationship between the questions, except for the effort of resolving pronoun references. This is particularly true for the TREC QA tracks, where most participating systems did not employ discourse processing for the question series. This thesis is therefore motivated to investigate the role of discourse processing and its implication on query expansion for a sequence of questions.

Chapter 1 introduces the context question problem that this dissertation aims to address. Chapter 2, Chapter 3, and Chapter 4 have introduced the background material for the study. The TREC QA tasks which inspire the current research are introduced in Chapter 2. State-of-the-art context question answering techniques used in the TREC 2005 context QA systems are also presented. Chapter 3 has presented the hypotheses for the discourse modeling presented in the dissertation. Context question sequence is treated as a mini discourse where it observes similar characteristics as other types of discourse.

A two-level representation of the discourse is hypothesized for the purpose of providing query terms for query expansion at the stage of processing context questions. The lexical cohesion representation is targeted to capture lexical relationships for the expressions that correspond to discourse entities. The discourse coherence level is to capture the topic change between questions. The idea is that a discourse model that captures both levels of information will provide adequate context information to process context questions.

Centering Theory is adopted in the work reported in Chapter 5, because it provides an excellent theoretical framework upon which a two-level discourse modeling is conducted. The main objective of Chapter 5 is to introduce centering-based models that relate discourse entities by the operations of *centers* to process context questions. At lexical cohesion level, questions are often linked to each other by such lexical ties as pronoun references. At discourse coherence level, a QA discourse is coherent in terms of their topicality relationships. Adjacent questions may have the same topic, similar topic or different topics. These topics are believed to be represented by the backward looking center of each question. Therefore in Chapter 5, three discourse models were proposed for discourse processing based on these two levels of speculations, especially driven by Centering Theory. The reference model aims to capture the lexical relation between questions, to be more specific, to resolve the pronoun references in the questions. The forward model aims to provide more context information on top of the pronoun resolution by adding the forward looking centers from the previous question. The transition model is to selectively pick context information from the previous question based on the topic transition between two adjacent questions. The results show that the transition model based on discourse transitions and centers can significantly improve the performance of document retrieval. The results from this chapter emphasize the need to properly capture more semantic information, rather than only focusing on lexical links, which only reflect discourse cohesion relations.

Chapter 6 looks at the context question problem based on another two-level representation of semantic information in the QA discourse. It aims to study definite descriptions in context questions due to the frequent occurrences of the definite descriptions in the TREC

2005 QA test sets and limited research on this topic. The lower level representation is similar to the one used in Chapter 5. It intends to capture the lexical relations between expressions for discourse entities, pronoun references and definite description references in particular. The other level of representation is hypothesized as follows: the more familiar an entity is to the discourse, the more definite an NP form will be used to refer to the entity. On the familiarity assumption, discourse entities are related to each other at a higher level. In other words, the linguistic expressions used to refer to the entities and their related entities would have different definiteness statuses in the QA discourse. Definiteness hierarchy is therefore employed to relate these entities together. Based on a definiteness hierarchy modified from Abbott 2004[89], different discourse models are proposed, mainly focusing on the processing of the pronouns and the definite descriptions. The typical uses of the definite descriptions in the context questions are classified into *discourse-new*, *direct anaphoric* and *bridging*. In general, it is found that the extra processing based on the definiteness hierarchy improves the performance significantly for the situation where no targets are explicitly provided to the question sets. In the target-given situation, on the other hand, the target-appending strategy is efficient compared with other processing models. Moreover, the definiteness hierarchy based pronoun resolution model is shown to perform better than the anaphora model in Chapter 5¹.

The results in Chapter 5 and Chapter 6 indicate that QA discourse is coherent in that discourse entities are related to each other not only at a lexical level but also at a discourse level where more semantic information can be captured.

The result in Chapter 6 also indicates the need to develop discourse processing models that can successfully predict a target. However this seems rather unrealistic. The TREC test sets were prepared specifically for the QA task in that every question was explicitly about the target. However, sometimes in reality, users may change the target gradually over the questions (e.g. resulting in *shift* transition in the transition model). For my future study, it is interesting to examine two things: 1) explore what the linguistic cues in context questions are to indicate a topic change; 2) how to build a discourse model

¹The state-of-the-art NLP technology, however is sufficient to implement both models.

to capture this topic change. Once these two questions are answered, it is expected that a discourse model that makes use of the system-detected target would outperform the models proposed for the no-target situation.

The major contribution of the research presented in this dissertation is that it provides us working discourse models for the context question problem, based on linguistic knowledge. Based on the data characteristics (e.g. percentage of pronouns, definite description distribution etc.), different models could be incorporated into the state-of-the-art QA systems.

There are also other dimensions along which my future work will be pursued. For example, how to use linguistic knowledge and the existing linguistic theories to process event-based context questions has become an interesting topic. I will also extend context Question Answering to fully interactive question answering and investigate the role of discourse processing in this new setting.

APPENDIX A

User data collection

A.1 Instruction

Assume there is an intelligent agent, who understands natural language (English) quite well. This agent has read thousands of web pages and can answer your natural language questions by searching through those web pages. For example, if a user is interested in finding out information about volcanoes and eruption history, he/she may ask the following questions in a coherent manner (i.e., all questions asked are related and connected and serving a particular information goal), and then the agent will find answers to each of these questions.

Q1. What's the name of the volcano that destroyed the ancient city of Pompeii?

Q2. When did it happen?

Q3. How many people were killed?

Q4. Any pictures?

Q5. Where is Mount Rainier?

Now suppose you are interested in finding out some information about "Presidential Debates 2004" or "Tom Cruise", you want to ask the agent a sequence of coherent questions, however with the following constraints:

1. The answer to each question must be found in the attached articles (if your question is indexed Q_{ij} , please underline your answer and mark it A_{ij} in the articles)

2. Each paragraph in the articles can only answer no more than one question
3. Ask at least 6 consecutive and coherent questions
4. Any question type is ok (e.g., what, when, where, how many, yes/no, why etc).

Please write down your questions and mark their answers:

Presidential Debates 2004

Q11

Q12

Q13

Q14

Q15

Q16

Q17(optional)

Q18(optional)

Read your questions carefully. Do those questions make sense (i.e., you would have ask the same questions if you were talking to a human agent)? What information goal did you have when you asked those questions? Are all the questions centered around this information goal? Refine your questions if necessary.

My information Goal is:

A.2 Documents collection on *Presidential Debates* 2004:

Presidential Candidates Conclude 2004 Debates

- President George W. Bush and Senator John F. Kerry faced off three times during the 2004 election: September 30, October 8, and October 13

September 30, the University of Miami continues its exciting lineup of debate-related events

- The September 30 presidential debate at the University of Miami, which attracted a record viewing audience of 63 million people and put the University in the international spotlight, is now history – but the series of debate-related activities Celebrating American Democracy and Diversity continues.
- The debate covered Foreign Policy & Homeland Security primarily, which is believed to be the single most important issue to voters in this election.
- In the first debate, President Bush' scowls were thought to play a large part in viewers' belief that Kerry won, based on polls by CBS News and others.

Oct 8, Washington University ready for Debate

- Although voters cite Iraq as a major concern, the economy consistently ranks at the top.
- Though Mr. Bush was more composed than in last week's first presidential debate, all agreed his tone was sometimes antagonistic and he again appeared uncomfortable being challenged. Kerry, on the other hand, was viewed as measured and articulate. However, none of the experts touted a clear winner.
- But what is clear after two Bush-Kerry debates is that the candidates don't care for each other. They may shake hands, but resentment runs deep. Kerry infers Mr. Bush is simplistic to the point of dishonesty; Mr. Bush infers Kerry is complicated to the point of ineffectiveness
- Thomson thought the debate was "much more even" than the first one. Like Thurber, Thompson said, "I suspect that this is not a debate in which there will be a substantial change one way or another."

Oct 13, ASU shines as debate host:

- Arizona State University hosts the third and final debate between President Bush and Sen. John Kerry Wednesday. The debate, to focus on domestic policy, is the only meeting between the Republican and Democratic candidates to be held in the Far West
- More than 2,500 local, national, and international media are leaving Arizona State University with a positive and lasting impression following their experience covering the final 2004 Presidential Debate October 13 at Gammage Auditorium.

Oct 13, Bush and Kerry Clash Over Jobs and Taxes at Last Debate

- Sen. John Kerry said Wednesday night that President Bush bears responsibility for a misguided war in Iraq, lost jobs at home and mounting millions without health care. Bush tagged his Democratic rival as a lifelong liberal bent on raising taxes and government spending.
- Kerry and the president also debated abortion, gay rights, immigration and more in a 90-minute debate that underscored deep differences only 19 campaign days before Election Day.
- This debate was similar in format to the first - the two rivals standing behind identical lecterns set precisely 10 feet apart. Bush was on better behavior, though, and there was no grimacing and scowling this time when it was Kerry's turn to speak.

Please write down your questions and mark their answers:

Tom Cruise Q21

Q22

Q23

Q24

Q25

Q26

Q27(optional)

Q28(optional)

Read your questions carefully. Do those questions make sense (i.e., you would have ask the same questions if you were talking to a human agent)? What information goal did you have when you asked those questions? Are all the questions centered around this information goal? Refine your questions if necessary.

My information Goal is:

A.3 Documents collection on *Tom Cruise*

Tom Cruise

- That million megawatt smile has helped Tom Cruise reach the pinnacle of his profession and stay there. He's a down-to-earth movie star with huge box-office hits under his belt such as Top Gun and Mission: Impossible.
- Thomas Cruise Mapother IV was born on the 3rd of July, 1962 (eerily similar to his film Born on the 4th of July), in Syracuse, New York. He was the only boy of four children. Since his father was an electrical engineer, he must have inherited his love for acting from his mother, who was a teacher.
- His acting career really began because he injured his knee in high school and was forced to quit the amateur wrestling team.
- His popularity took a beating in movies like All the Right Moves in 1983, followed by Legend in 1985. Cruise's career began to solidify during his signature hit of the 1980s, Top Gun.
- He proved his dramatic talents in the 1988 drama Rain Man, where he co-starred with Oscar-winner Dustin Hoffman. Oliver Stone's Born on the Fourth of July

(1989) earned him a Best Actor Oscar nomination for his hard-hitting portrayal of anti-war activist Ron Kovic.

- Cruise fell short for his role in *Far and Away* (1992) with co-star Nicole Kidman, who he later married in 1990 after sharing the screen once again in *Days of Thunder*.
- Cruise and his wife of 10 years, Nicole Kidman, filed for divorce in February 2001.
- He has since been seen with Spanish beauty Penelope Cruz, his *Vanilla Sky* co-star. Keeping with the science-fiction theme of *Vanilla Sky*, Cruise's next starring role is Steven Spielberg's *Minority Report*. A third *Mission Impossible* is rumored to be in the works.

Nicole Kidman

- She starred in *Days Of Thunder* and the kinky thriller *Eyes Wide Shut*, opposite then husband, Tom Cruise.
- Many would be fooled into thinking that Nicole was born down under because of her noticeable Australian accent, but they would be wrong. Nicole Mary Kidman was born in Honolulu, Hawaii, on June 20, 1967. The Kidmans lived in the U.S. because Nicole's biochemist father was conducting research on breast cancer.
- Once Nicole was 4 years old, her family moved to Australia, where Nicole and her younger sister were raised under strict rules. Anthony and Janelle Kidman were extremely politically active, and instilled certain values into their daughters. The Kidman girls were even required to discuss a political issue or current affair with their parents at the end of every day.

Penelope Cruz

- There's a reason why Penelope Cruz is nicknamed the "Spanish Enchantress" and the "Madonna of Madrid" outside of her native land. Gorgeous, sultry and gifted, she has paid her dues and climbed her way from supporting actress to feature player

Penelope Cruz Sanchez was born in Madrid, Spain, on the 28th of April 1974. Raised along with brother Eduardo and sister Monica in Madrid, Penelope was always fond of the arts, particularly ballet and jazz. Her passion for dance led to her decision to abandon traditional schooling, focusing her time and energy instead on the graceful art of ballet. Please write down your questions and mark their answers:

Pompeii

Q31

Q32

Q33

Q34

Q35

Q36

Q37(optional)

Q38(optional)

Read your questions carefully. Do those questions make sense (i.e., you would have ask the same questions if you were talking to a human agent)? What information goal did you have when you asked those questions? Are all the questions centered around this information goal? Refine your questions if necessary.

My information Goal is:

A.4 Documents collection on *Pompeii*

Pompeii

- On August 23, 79 AD, Pompeii looked like any other busy, prosperous city. People were moving about, trading goods, news, and friendly talk.

- Three days later, on August 26, all of these sounds had fallen silent, and the place itself had vanished. Almost nothing was seen of Pompeii for more than 1500 years. Now, more than 1900 years later, we are learning more and more about the last days of Pompeii.
- What happened to Pompeii preserved a treasury of information about life in the ancient Roman Empire. You can begin your exploration of the mystery of Pompeii and the life of people in the Roman Empire by clicking on enter below. Once you see the map, you can choose any place to start, but Vesuvius might make the best beginning!
- 62 February 5–A major earthquake almost destroys the cities of Pompeii and Herculaneum. Many buildings are damaged or destroyed, and the aqueducts that bring water into Pompeii are ruined.
- 79 August 24 and 25–Vesuvius erupts, burying Pompeii in ash and cinders and covering Herculaneum in mud as hard as rock. Ash, rock and cinders fall over a large area, damaging houses in many faraway places and blackening the sky over what is now known as Naples for three days.
- 79 Late August–People who escaped from Pompeii return to the city to try to find their houses and possessions. Many people dig shafts down into the town and recover some of their valuables. This effort is finally abandoned.
- 202 -533 Vesuvius erupts at least five times.
- 1594 Workers digging a tunnel to supply water to a nearby village find a stone that says decurio Pompeiis. The city has been so long forgotten that most people think Pompeiis refers to a famous Roman ruler named Pompey.
- 1631 Vesuvius erupts again. This is by far the worst eruption since 79 AD. Lava flows from the volcano in seven different streams, destroying nearly all of the towns below.

- 1707 Prince d'Elboeuf, hearing of some interesting finds during the digging of a well, starts digging for treasure. At this time he does not know the name of the city he is digging up.
- 1860 Giuseppe Fiorelli is appointed as director of the Pompeiian dig. Fiorelli wants to uncover the entire Roman city. Up until this time, most people have been digging single holes or opening up small areas to look for treasure. Fiorelli wants to share the riches of the lost city with the entire world.
- 1860-1875 Under Fiorelli's watchful eye, digging is continued. Modern archeological techniques and recording systems begin to be used to catalog and identify all of the objects uncovered.
- 1875-Present Digging continues under the guidance of many different directors. Many of the objects uncovered are placed in a museum in nearby Naples, where they show the world of today much about ancient Roman times.
- 1913-1944 Vesuvius erupts several times, finishing the eruption cycle that started in 1631.
- Today all we can see of Pompeii are ruins. But the ruins tell us many things about the ancient world. The disaster that destroyed the city of Pompeii in 79 AD preserved forever a treasury of the past. Careful excavation and exploration of the ruins continues to add to my knowledge of ancient Roman times.

Please write down your questions and mark their answers:

Hawaii

Q41 _____

Q42 _____

Q43 _____

Q44 _____

Q45 _____

Q46

Q47(optional)

Q48(optional)

Read your questions carefully. Do those questions make sense (i.e., you would have asked the same questions if you were talking to a human agent)? What information goal did you have when you asked those questions? Are all the questions centered around this information goal? Refine your questions if necessary.

My information Goal is:

A.5 Documents collection on *Hawaii*

BIOLOGY

- The Hawaiian Islands have a wide variety of plant, marine and animal life. Vegetation zones include: coastal, dryland forest, mixed open forest, rain forest, subalpine and alpine. More than 90 percent of the native plants and animals living in Hawaii are found nowhere else in the world, and a greater variety of fish exist in Hawaiian waters than elsewhere. The humuhumunukunukuapuaa is the unofficial state fish.
- Hawaii is sometimes called the Endangered Species Capital of the World. At least one third of all the endangered species in the United States are found in Hawaii including the Nene Goose (official state bird), the Humpback Whale (official state marine mammal), the Pacific Green Sea Turtle and the Pueo (Hawaiian owl). The exotic species, man, poses a greater threat than nature to Hawaii's native flora and fauna.

CLIMATE

- The Hawaiian Islands have only two seasons: “summer” between May and October and “winter” between October and April.
- The climate is subtropical, with a normal annual temperature of 77F, making these islands “- the peace fullest, rest fullest, balmiest, dreamiest haven of refuge for a worn and weary spirit the surface of the earth can offer.”———— Mark Twain

CULTURAL HISTORY

- The Hawaiian Islands are stepping-stones linking East to West. Here Polynesian sensuality, American pragmatism, and Oriental exoticism weave a tapestry of cultural extremes.
- Hawaii’s multi-cultural society has had major immigration from: Polynesia, United States, China, Japan, Portugal, Puerto Rico, Korea, Philippines

ECONOMY

- Hawaii’s cost of living is one of America’s highest, its per capita personal income below average. In fact, Hawaii’s cost of living for a family of four is estimated to be approximately 27% higher than the U.S. average for a comparable standard of living. In 1999, Hawaii’s average per capita personal income of \$27,544 was 3.5% below the U.S. average - the price of living in Paradise!
- Hawaii’s major sources of annual income include (1998/1999): Tourism - \$10.3 billion, Federal Defense Spending - \$4.2 billion, Sugar - \$133.1 million, Pineapple - \$145.1 million
- The 1990’s has been the worst decade in Hawaii’s economic history since World War II due, in large measure, to the decline in tourism from the East and the demise of the sugar and pineapple industries. To meet the challenges of the 21st Century, Hawaii is working to diversify its economy with a focus on industries such as science and technology, health and wellness tourism, diversified agriculture, ocean research and development, and film and television production

GEOGRAPHY

- Hawaii is the most remote island chain in the world, over 2,000 miles from the nearest landfall. Distance makes for splendid isolation - these Polynesian islands are removed from all else but one another.
- Hawaii consists of eight major islands plus 124 minor islands, reefs and shoals, strung like a necklace across the Pacific for over 1,500 miles. The eight major islands (which make up over 99% of the total land area) are Oahu, Maui, Hawaii (known as Big Island), Kauai, Molokai, Lanai, Kahoolawe (uninhabited) and Niihau (privately owned).

BIBLIOGRAPHY

- [1] S. E. Brennan, M. Friedman, and C. Pollard, "Centering approach to pronouns," in *ACL '87*, (Standford, CA.), pp. 155–162, 1987.
- [2] R. Geluykens, "Studies in discourse and grammar," in *From Discourse Process to Grammatical Construction: on left-dislocation in English*, vol. 1, Amsterdam/Philadelphia: John Benjamins Publishing Company, 1992.
- [3] D. Crystal, *Introducing Linguistics*. London: Penguin, 1992.
- [4] I. Roberts and R. Gaizauskas, "Evaluating passage retrieval approaches for question answering," in *Proceedings of the 26th European conference on information retrieval*, 2004.
- [5] 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.
- [6] E. Voorhees, "Overview of trec 2001 question answering track," in *proceedings of TREC*, (Gaithersburg, MD), November 13-16 2001.
- [7] E. Voorhees, "Overview of trec 2004 question answering track," in *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, (Gaithersburg, MD), 2004.
- [8] 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.
- [9] K. Collins-Thompson, J. Callan, E. Terra, and C. L. Clarke, "The effect of document retrieval quality on factoid question answering performance," in *SIGIR*, 2004.
- [10] 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 (TREC 2005)*, (Gaithersburg, MD), 2005.
- [11] 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.

- [12] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. NJ.: Prentice Hall, 2000.
- [13] M. Mulcahy, K. White, I. Gabbay, and A. O’gorman, “Question answering using the dlt system at trec 2005,” in *Proceedings of the 14th Text Retrieval Conference (TREC-2005)*, (Gaithersburg, MD.), 2005.
- [14] 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.
- [15] P. Schone, G. Ciany, R. Cutts, P. McNamee, J. Mayfield, and T. Smith, “Qactis-based question answering at trec-2005,” in *Proceedings of the 14th Text Retrieval Conference (TREC 2005)*, Gaithersburg, MD, 2005.
- [16] R. Gaizauskas, M. A. Greenwood, H. Harkema, M. Hepple, H. Saggion, and S. A., “The university of sheffield’s trec 2005 q&a experiments,” in *Proceedings of the 14th Text Retrieval Conference (TREC-2005)*, 2005.
- [17] 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.
- [18] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang, “Employing two question answering systems in trec-2005,” in *Proceedings of the 14th Text Retrieval Conference (TREC 2005)*, 2005.
- [19] D. Ahn, S. Fissaha, V. Jijkoun, K. Mller, M. d. Rijke, and E. T. K. Sang, “Towards a multi-stream question answering-as-xml-retrieval strategy,” in *Proceedings of the 14th Text Retrieval Conference (TREC 2005)*, 2005.
- [20] 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 lcc’s question-answering server,” in *The Tenth Retrieval Conference (TREC-2001)* (E. Voorhees and D. K. Harman, eds.), pp. 355–361, Gaithersburg, MD.: NIST special publication, 2001.
- [21] M. Hoey, *On the Surface of Discourse*. London: Allen and Unwin, 1983.
- [22] D. Nunan, *Introducing Discourse Analysis*. Australia: Penguin English, 1993.
- [23] B. J. Grosz and C. Sidner, “Attention, intention, and the structure of discourse,” *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.
- [24] H. P. Grice, “Logic and conversation,” in *Syntax and Semantics* (P. Cole and J. Morgan, eds.), pp. 41–58, Academic Press, 1975.

- [25] 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.
- [26] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: A theory of text organization," Tech. Rep. Technical report ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.
- [27] 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.
- [28] H. Kamp and U. Reyle, *From discourse to logic: introduction to modeltheoretic semantics of natural language, formal logic, and discourse representation theory*. Kluwer, Dordrecht, 1993.
- [29] 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.
- [30] J. Hobbs, "Coherence and coreference," *Cognitive Science*, vol. 3, pp. 67–90, 1979.
- [31] J. R. Hobbs, M. Stickel, D. Appelt, and P. Martin, "Interpretation as abduction," Tech. Rep. Note 499, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, December 1990.
- [32] D. Marcu, *The Theory and practice of discourse parsing and summarization*. Cambridge, Massachusetts. London, England: the MIT Press, 2000.
- [33] D. Marcu, "A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts," in *The COLING/ACL'98 Workshop on Discourse Relations and Discourse Markers*, (Montreal, Canada), pp. 1–7, August 1998.
- [34] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.
- [35] J. L. Austin, *How to Do Things with Words*. Cambridge, MA: Harvard University Press, 2nd ed. ed., 1962.
- [36] A. Kehler, "Discourse coherence," in *Handbook of Pragmatics* (L. R. Horn and G. Ward, eds.), Basil Blackwell, 2004.
- [37] 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.

- [38] 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.
- [39] M. A. K. Halliday and R. Hasan, *Cohesion in English*. London: Longman, 1976.
- [40] M. Pickering and S. Garrod, "Routinization in the interactive-alignment model of dialogue." Unpublished manuscript, 2004.
- [41] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," in H. H. Clark, M. J. Pickering, & A. A. Cleland (eds.), *Syntactic coordination in dialogue*. *Cognition*, vol. 22, pp. 1–39, 1986.
- [42] S. E. Brennan, "Lexical entrainment in spontaneous dialog," in *Proceedings of International Symposium on Spoken Dialogue, ISSD-96*, (Philadelphia, PA), pp. 41–44, 1996.
- [43] D. Lewis, *Convention: A philosophical study*. Basil Blackwell/Harvard University Press, 1969.
- [44] H. Branigan, M. Pickering, and A. Cleland, "Syntactic coordination in dialogue," *Cognition*, vol. 75, pp. B13–B25, 2000.
- [45] B. J. Grosz, A. Joshi, and S. Weinstein, "Towards a computational theory of discourse interpretation." Unpublished manuscript, 1986.
- [46] B. J. Grosz, "The representation and use of focus in a system for understanding dialogs," in *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, (Cambridge, Mass.), 1977.
- [47] B. Grosz, "The representation and use of focus in dialogue understanding," tech. rep., SRI International, 333 Ravenswood Ave., Menlo Park, CA, 94025, 1977.
- [48] B. J. Grosz, "Discourse analysis," in *Understanding Spoken Language* (D. Walker, ed.), pp. 235–268, Elsevier North-Holland, 1978.
- [49] B. J. Grosz, "Focusing in dialog," in *Theoretical Issues in Natural Language Processing-2*, pp. 96–103, University of Illinois at Urbana-Champaign, Champaign, Illinois, 1978.
- [50] 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.
- [51] 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.

- [52] C. L. Sidner, "Focusing for interpretation of pronouns," *American Journal of Computational Linguistics*, vol. 7, no. 4, pp. 217–231, 1981.
- [53] C. L. Sidner, "Focusing in the comprehension of definite anaphora," in *Computational Models of Discourse* (M. Brady and R. C. Berwick, eds.), pp. 267–330, Cambridge, Mass.: MIT Press, 1983. reprinted in Grosz et al. (1986a:363-94).
- [54] 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.
- [55] 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.
- [56] 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.
- [57] W. Chafe, "Givenness, contrastiveness, definiteness, subjects, and topics," in *Subject and Topic* (C. Li, ed.), pp. 25–55, New York: Academic Press, 1976.
- [58] M. A. Walker, A. Joshi, and E. Prince, *Centering Theory in Discourse*. Clarendon Press / Oxford, 1998.
- [59] S. B. Hudson-DZmura, *The Structure of Discourse and Anaphor Resolution: The Discourse Center and the Roles of Nouns and Pronouns*. PhD thesis, University of Rochester, 1988.
- [60] P. C. Gordon, B. J. Grosz, and L. A. Gillom, "Pronouns, names and the centering of attention in discourse," *Cognitive Science*, vol. 17, no. 3, pp. 311–347, 1993.
- [61] J. Barwise and J. Perry, *Situations and attitudes*. Cambridge, Mass.: Bradford Books, 1983.
- [62] J. Barwise, "The situation in logic-iv: on the model theory of common knowledge," Tech. Rep. Technical report No. 122, Stanford, California, CSLI, 1988.
- [63] E. F. Prince, "On the function of existential presupposition in discourse," in *14th regional meeting of the Chicago Linguistic Society*, (Chicago), pp. 362–376, 1978.
- [64] H. H. Clark and C. R. Marshall, "Definite reference and mutual knowledge," in *Elements of discourse understanding* (A. Joshi, B. Webber, and I. Sag, eds.), Cambridge: Cambridge University Press, 1981.
- [65] M. Poesio, R. Stevenson, B. di Eugenio, and J. Hitzeman, "Centering: A parametric theory and its instantiations," *Computational Linguistics*, vol. 30, pp. 309–363, Sept 2004.

- [66] S. Cote, "Ranking forward-looking centers," in *Centering Theory in Discourse* (M. Walker, A. Joshi, and E. Prince, eds.), pp. 55–69, Clarendon Press / Oxford, 1998.
- [67] M. Kameyama, *Zero anaphora: the case of Japanese*. PhD thesis, Stanford University, 1985.
- [68] 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.
- [69] B. Di Eugenio, "Centering and the italian pronominal system," in *COLING 90: the 13th International Conference on Computational Linguistics*, (Helsinki), pp. 20–25, Aug. 1990.
- [70] B. Di Eugenio, "Centering in italian," in *Centering Theory in Discourse* (M. Walker, A. Joshi, and E. Prince, eds.), pp. 115–137, Clarendon Press / Oxford, 1998.
- [71] O. Rambow, "Pragmatic aspects of scrambling and topicalization in german: a centering approach," in *the workshop on centering theory in naturally-occurring discourse, IRCS*, (University of Pennsylvania), May 1993.
- [72] m. D. Turan, *Null vs. overt subjects in Turkish discourse: a centering analysis*. PhD thesis, University of Pennsylvania, 1995.
- [73] m. D. Turan, "Ranking forward-looking centers in turkish: universal and language-specific properties," in *Centering Theory in Discourse* (M. Walker, A. Joshi, and E. Prince, eds.), pp. 139–160, Clarendon Press / Oxford, 1998.
- [74] 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.
- [75] M. A. Walker, "Evaluating discourse processing algorithms," in *Proceedings of 27th Annual Meeting of the Association for Computational Linguistics*, pp. 251–261, 1989.
- [76] B. Baldwin, "Anaphora resolution with centering," in *Workshop on Centering Theory in Naturally-Occurring Discourse*, (Philadelphia, PA.), May 1993.
- [77] 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.
- [78] 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.
- [79] M. Strube and U. Hahn, "Functional centering," in *ACL-96*, (Santa Cruz, CA), pp. 270–277, 1996.

- [80] S. Lappin and H. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, no. 4, pp. 535–561, 1994.
- [81] J. R. Hobbs, "Resolving pronoun references," *Lingua*, vol. 44, pp. 311–338, 1978. Reprinted in Grosz1986b.
- [82] M. Kameyama, "Intrasentential centering: A case study," in *Centering Theory in Discourse* (M. Walker, A. Joshi, and E. Prince, eds.), pp. 89–112, Clarendon Press / Oxford, 1998.
- [83] M. Strube, "Never look back: an alternative to centering," in *Proc. Of COLING-ACL*, (Montreal), pp. 1251–1257, 1998.
- [84] E. Miltsakaki, "Locating topics in text processing," in *Proceeding of CLIN*, 1999.
- [85] B. J. Grosz and C. L. Sidner, "Lost intuitions and forgotten intentions," in *Centering Theory in Discourse* (M. A. Walker, A. K. Joshi, and E. F. Prince, eds.), pp. 39–51, Clarendon, Oxford, 1998.
- [86] R. Prasad, *Constraints on the generation of referring expressions, with special reference to Hindi*. PhD thesis, University of Pennsylvania, 2003.
- [87] E. Miltsakaki, *The Syntax-Discourse Interface: Effects of the Main-Subordinate Distinction on Attention Structure*. PhD thesis, University of Pennsylvania, 2003.
- [88] J. Gundel, "The role of topic and comment in linguistic theory. distributed by indiana university linguistics club.," Bloomington, Indiana, 1976.
- [89] B. Abbott, "Definiteness and indefiniteness," in *Handbook of pragmatics* (L. R. Horn and G. Ward, eds.), Oxford, Blackwell, 2004.
- [90] R. Gaizauskas, M. A. Greenwood, M. Hepple, I. Roberts, and H. Saggion, "The university of sheffield's trec 2004 q&a experiments," in *Proceedings of the 13th Text Retrieval Conference (TREC-2004)*, 2004.
- [91] B. M. Sundheim, "Overview of the results of the muc-6 evaluation," in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, (Columbia, MD), pp. 13–31, November 6-8 1995.
- [92] T. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson, "Sri international fastus system muc-6 test results and analysis," in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, (Columbia, MD), NIST, Morgan-Kaufmann Publishers., 1995.
- [93] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks, "University of sheffield: Description of lasie system as used," in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, (Columbia, Maryland), Morgan Kaufman Publishers Inc., 1995.

- [94] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks, "University of sheffield: description of the lasie-ii system as used for muc-7," in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.
- [95] K. Fraurud, "Definiteness and the processing of noun phrases in natural discourse," *Journal of Semantics*, vol. 7, pp. 395–433, 1990.
- [96] M. Poesio and R. Vieira, "A corpus-based investigation of definite description use," *Computational linguistics*, vol. 24, pp. 183–216, 1998.
- [97] E. F. Prince, "The zpg letter: Subjects, definiteness, and information status," in *Discourse description: Diverse linguistic analyses of a fund-raising text* (W. C. Mann and S. A. Thompson, eds.), pp. 295–325, Philadelphia: John Benjamins, 1992.
- [98] B. Birner and G. Ward, *Informational status and noncanonical word order*. Philadelphia: John Benjamins, 1998.
- [99] M. Ariel, "Referring and accessibility," *Journal of Linguistics*, vol. 24, pp. 65–87, 1988.
- [100] M. Ariel, *Accessing noun-phrase antecedents*. London: Routledge, 1990.
- [101] P. Christophersen, *The articles: A study of their theory and use in English*. Oxford University Press, 1939.
- [102] I. Heim, *The semantics of definite and indefinite noun phrases*. PhD thesis, University of Massachusetts Amherst, MA, 1982.
- [103] I. Heim, "File change semantics and the familiarity theory of definiteness," in *Meaning, use and the interpretation of language* (R. Bauerle, C. Schwarze, and A. von Stechow, eds.), pp. 164–189, Berlin: Walter de Gruyter, 1983.
- [104] D. Lewis, "Scorekeeping in a language game," *Journal of Philosophical Logic*, vol. 8, pp. 339–59, 1979.
- [105] G. Gazdar, *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic Press, 1979.
- [106] B. Abbott, "Presuppositions as nonassertions," *Journal of Pragmatics*, vol. 32, pp. 1419–1437, 2000.
- [107] B. Russell, "On denoting," *Mind*, vol. 14, pp. 479–493, 1905.
- [108] S. Lbner, "Definites," *Journal of Semantics*, vol. 4, pp. 279–326, 1985.
- [109] N. Kadmon, "Uniqueness," *Linguistics and Philosophy*, vol. 13, pp. 273–324, 1990.

- [110] J. A. Hawkins, "On (in)definite articles: implicatures and (un)grammaticality prediction," *Journal of Linguistics*, vol. 27, pp. 405–442, 1991.
- [111] J. A. Hawkins, *Definiteness and Indefiniteness*. London: Croom Helm, 1978.
- [112] O. Jespersen, *A modern English grammar on historical principles*, vol. III. London: George Allen and Unwin, 1949.
- [113] H. H. Clark, "Bridging," in *Thinking: Readings in cognitive science* (P. N. Johnson-Laird and P. C. Wason, eds.), pp. 411–20, Cambridge: Cambridge University Press, 1977.

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 02845 6261