Lite-GISTexter at DUC 2005

Finley Lacatusu, Andrew Hickl, Paul Aarseth
Language Computer Corporation
1701 N. Collins Blvd.
Richardson, TX 75080
finley@languagecomputer.com

Lara Taylor
Department of Linguistics
University of California, San Diego
La Jolla, CA 92037
ltaylor@ling.ucsd.edu

Abstract

This paper describes how Language Computer Corporation's Lite-GISTexter multi-document summarization system addresses the challenge of providing summary-length answers to the types of complex questions asked in DUC 2005. We show that techniques first used in the automatic answering of "relationship" questions can be used in multi-document summarization in order to provide accurate and responsive summary answers to a wide range of complex questions. Although we have found that traditional multi-document summarization techniques do remain effective in producing summary answers to questions, we argue that the best results should be obtained by systems that focus summary answers by approximating the information need of complex questions.

1 Introduction

This paper argues that the task of producing summary answers in response to questions requires a hybrid approach to multi-document summarization (MDS) that models not only the information contained in a relevant set of documents but also the information requirements of complex research scenarios as well. Although we expect traditional MDS techniques can be used to produce summaries from the set of relevant documents assembled for each of this year's topics, we believe that the best summarization results for this task will be obtained by systems that try to provide informative answers to each of the questions and subquestions contained in a research scenario. Without an adequate representation of the information need(s) of a scenario, systems may produce summaries that are coherent and informative, yet do not provide users with the information that they really want to know. In this paper, we discuss how we extended Language Computer Corporation's MDS system, Lite-GISTexter, to provide coherent summary-length answers to questions by employing question decomposition and representation techniques first developed for complex question-answering.

In DUC 2005, participants were asked to provide summary answers in response to research scenarios consisting of one or more complex questions. Although this task, which we refer to as question-directed summarization (QDS), represents a significant departure from many of the previous DUC tasks, it is very similar to the problem of complex (or "relationship") question-answering, a task which has recently received much attention from the automatic question-answering (Q/A) community ((Harabagiu et al., 2005), (Hickl et al., 2004), (Small et al., 2004)). Unlike traditional approaches to factoid Q/A, which have depended on the recognition of a simple semantic answer type, current approaches to complex Q/A have assumed that no single answer type can adequately represent all of the semantic requirements of a complex question. In order to provide relevant answers to complex questions, systems must represent the information need of a question itself, either by decomposing it into a series of simpler questions, or by associating it with conceptual or topic-based representations that can be used to retrieve potentially relevant passages from a corpus. We expect that question-directed summarization systems could benefit from a similar approach: by representing the information need of complex questions, QDS systems can produce summaries that are not just topic-relevant, but also responsive to the user's needs.

LCC's Lite-GISTexter combines information taken from the topic representation of a relevant set of documents with representations of the information need(s) of complex research scenarios to build question-directed summaries. After syntactically decomposing complex questions into a series of simpler subquestions, we extract a set of keywords, predicate-argument structures, and syntactic frames from each sub-question to provide a rough approximation of its information need. We then compute both topic signatures (Lin and Hovy, 2000) and topic relations (Harabagiu, 2004) using the relevant set of documents given for each scenario. These topic-based terms and relations are then combined with features extracted from each subquestion to extract sentences from the document collection for inclusion in the summary. Although we have employed topic-based approaches to MDS in past DUC evaluations (e.g. (Lacatusu et al., 2003), (Lacatusu et al., 2004)), we believe the combination of question-derived features with topic-based features enables us to create summaries that are much more responsive than in previous systems.

Unlike complex Q/A systems, which return an unstructured list of candidate answers, QDS systems must return summary answers that are more than just a list of text snippets. After sentences are selected for inclusion in a summary, Lite-GISTexter re-assmbles the summary in order to reduce redundancy, clarify referring expressions, and improve the overall coherence of the text. The high marks our system receives for "Non-Redundancy" suggests that our method for eliminiating redundant sentences works well, while still maintaining excellent information coverage (as reflected by the Pyramid, Modified Pyramid, and Responsiveness scores). In addition, Lite-GISTexter's very competitive score for "Referential Clarity" suggests that a simple technique that expands the included context for sentences containing unresolved referring expressions may be effective for QDS.

In the rest of this paper, we describe how Lite-GISTexter produces summary answers to complex questions. Section 2 discusses some of the challenges of representing the information need of complex questions for both automatic question-answering and multi-document summarization. Section 3 provides an overview of the current Lite-GISTexter system, while Section 4 discusses the performance of Lite-GISTexter as evaluated by both human and automatic evaluation metrics. Finally, Section 5 includes our conclusions and plans for future enhancements to our system.

2 Answering Complex Questions

With the accuracy of today's best factoid Q/A systems nearing (and in some cases, exceeding) the 70% F threshhold, work in automatic Q/A has begun to focus on the answering of complex questions. Although researchers have not yet agreed upon a standard definition of exactly what constitutes a complex (or "relationship") question, for the purposes of this paper, we claim that a complex question can be defined as a natural language question whose information need cannot be associated with a single semantic answer type from an idealized ontology of semantic entity or event types.

Unlike factoid questions, which presuppose that a single correct answer can be found that completely satisfies all of the information requirements of the question, complex questions often seek multiple different types of information and do not presuppose that one single answer could meet all of its information needs simultaneously. For example, with a factoid question like *What companies are planning to build mini-mills?*, we assume that a user is looking for a list of organizations (specifically, companies) which intend to build mini-mills. In this case, users do not expect systems to return additional related information (such as the names of companies who have already built mini-mills), as the answer itself is sufficient to meet the information need of the question. In fact, returning additional information here is an undesirable result, as the system would be more informative than is necessary and violate the Gricean Maxim of Quantity. In contrast, with a complex question like *What has been the impact of new steel producing technologies?*, the wider focus of this question indicates that users may not have a clearly defined (or pragmatically restrictive) information need, and therefore would be amenable to receiving additional supporting information that was relevant to their overall goal.

Question: What are new technologies for producing steel? What are the technical and cost benefits of these new technologies over older methods of steel production? What companies are producing steel using these new technologies at so-called mini-mills? What technical and financial problems have they encountered? What companies are planning to build mini-mills which will employ these newer methods?

Syntactic Decomposition

Q_1: What are new technologies for producing steel?
Q_2: What are the technical benefits of these new technologies over older methods of steel production?
Q_3: What are the cost benefits of these new technologies over older methods of steel production?
Q_4: What companies are producing steel using these new technologies at so-called mini-mills?
Q_5: What technical problems have they encountered?
Q_6: What financial problems have they encountered?
Q_7: What companies are planning to build mini-mills which will employ these newer methods?

Semantic Decomposition for Q_1

Q_{11}: What new sources of steel are being explored?
Q_{12}: How are new gasses being used in the production of steel?
Q_{13}: What is being done to reduce the environmental impact of steel mills?
Q_{14}: How are U.S. steel mills changing their steel production techniques?
Q_{15}: What trade organizations sponsor steel research conferences?
Q_{16}: What companies are selling new steel production equipment?

Semantic Decomposition for Q_4

Q_{41}: What steel companies have mini-mills?
Q_{42}: What countries have the most mini-mills?
Q_{43}: What steel production facilites are there other than mini-mills?
Q_{44}: What steel products are being made at mini-mills?
Q_{45}: What steel companies do not have mini-mills?
Q_{46}: What companies have sold technologies to mini-mills?
Q_{47}: What companies have started greenfield mini-mills?
Q_{48}: What companies have used sources of direct reduced iron at their mini-mills?

Figure 1: Question Decomposition

Question decomposition is an important step in representing the information need of any complex question. In

Nucor, which has pioneered the use of a cost-effective new steel-making technology called thin-stab casting, is one of the lowest cost manufacturers of steel in the world. It is the leader of the US 'mini-mills'-producers which have undercut the large, integrated steel manufacturers in many markets through the use of scrap metal and electric-arc furnace technology. Nucor's two-year-old plant at Crawfordsville is special because it is the first mini-mill to make flat-rolled products-albeit for the less sophisticated end of that market-and it does so with a new technology, called thin-slab casting, which sharply cuts the time and capital costs required to make sheet steel. For the thin-slab process to work in the most cost-effective way, it needs to be fed with relatively inexpensive supplies of molten metal. That in turn depends on the local electricity costs and the price and availability locally of scrap metal. The success of Nucor, the US company that is making hot-rolled coil at mini-mills in Indiana and Arkansas, has prompted integrated steel producers elsewhere to question whether their large, labour-intensive plants can survive against smaller, nimbler competitors. Industry has already suggested cuts of up to 25.8m tonnes in crude steel, and 17.9m tonnes in rolled products, but the Commission will consult again with steel producers over the next two months to see if further capacity cuts are possible. Brussels says the industry should be aiming to cut 30m tonnes of crude steel capacity and 20m in rolled products.

Lite-GISTexter summary

Q_1 Big old-fashioned steel companies, which have spent much of the past decade reeling from falling demand and prices, are trying to regain the lead in new production technology.
Q_2 Mr Busse acknowledged that Europe still had overcapacity in steelmaking but said Nucor's success in the US had been based on applying new technology in markets where there was also overcapacity of older equipment.
Q_3 Now put these labour advantages together with Nucor's relatively cheap electricity costs (thanks again to a rural, greenfield location) and the operating advantages of its new technology.
Q_4 Nucor, which has pioneered the use of a cost-effective new steel-making technology called thin-stab casting, is one of the lowest cost manufacturers of steel in the world.
Q_5 This problem is likely to recede gradually as the US and global economies emerge from recession, and American prices could get a further fillip if the industry eventually wins some form of protection through its anti-dumping actions.
Q_6 In Europe the problem has been created by subsidised public sector producers who cut prices to maintain market share and keep plants running.
Q_7 LARMCO, the US steel group, is to build a flat-rolled steel mini-mill at its facility in Mansfield, Ohio, the first to be announced by a US integrated steelmaker, AP-DJ reports.

Answers provided by LCC's PALANTIR Q/A system

Figure 2: Comparison of Question-Directed Summarization and Complex Q/A.

previous work (Hickl et al., 2004), we have suggested that complex questions need to be decomposed into a series of simpler questions before they can be submitted to a Q/A or QDS system. Although today's Q/A systems can use techniques based on keyword density and topic information to find relevant answers to even the most complex of questions, we have shown that by decomposing complex questions into sets of sub-questions, systems can improve the average quality of answers returned and achieve better overall coverage for the question as a whole. We suggest that complex questions can be decomposed both syntactically and/or semantically. With syntactic decomposition, sub-questions are extracted from complex questions by separating conjoined phrases, recognizing embedded questions, and by removing speech act and psych verbs from the text of the question. Results from the syntactic decomposition of the questions contained in DUC 2005 topic 413 are presented in Figure 1. In contrast, semantic decomposition is the process by which complex questions are broken down into hierarchies of related sub-questions which represent one or more aspects of the information need of the main question. Although we have not implemented techniques for semantic question decomposition into our 2005 DUC system, we feel that the automatic generation of semantic subquestions like the ones presented in Figure 1 for questions (Q_1) and (Q_4) could greatly enhance the performance of future QDS systems.

Recent work also has underlined the importance of providing answers that respond not just to single questions, but also to entire research scenarios. Equivalent to DUC topics, we define research scenarios as the collections of questions and topics that are assembled by users gathering information for a specific task. Since Q/A systems can only answer a single question at a time, most Q/A-based approaches to providing answers to research scenarios have focused on maintaining a dialogue-style interaction with a user or providing research suggestions from a specially-created database. Regardless of the mode of interaction, users of Q/A must manually compile information from the system's responses into a coherent answer. QDS systems like Lite-GISTexter do not face this limitation, however. By representing the sub-questions contained in a research scenario, QDS systems can produce a single summary that provides users with a coherent presentation of the information they need. As we can see in Figure 2, which contrasts the summary output of Lite-GISTexter with the top answers returned by LCC's PALANTIR Q/A system, the presumption of coherence in the summary text seems to provide a more complete response to the research scenario than the individual answers returned for each sub-question.

3 System Overview

The creation of question-directed summaries depends on the extraction of a set of sentences whose content both describes the information content of a set of relevant documents and also meets the information needs of the questions (and sub-questions) contained in each topic. By modeling both the topical structure of relevant documents and the information requirements of questions, we expect to create summaries that are informative, coherent, and responsive.

Lite-GISTexter uses two different topic representations to identify the set of relevant sentences that should be included in a summary: topic signatures (TS_1) and enhanced topic signatures (TS_2) . As with our DUC 2004 system, we represented the topic of a document collection by adapting the topic signature method first proposed in (Lin and Hovy, 2000). Originally developed for single-document summarization, this algorithm computes a weight for each term in a document cluster based on its relative frequency in a relevant set of documents. (Complete details of our topic signature implementation are provided in (Lacatusu et al., 2004).) In DUC 2005, we also experimented with the enhanced topic signatures first described by (Harabagiu, 2004). Unlike Lin and Hovy's topic signatures (which are limited to sets of individual terms), Harabagiu's enhanced topic signatures can be used to discover a set of relevant relations that exist between topic signature terms and to provide each relation with a weight depending on its overall significance to the topic being modeled.

With enhanced topic signatures, topics are represented as the set of relevant relations that exist between topic signature terms: $TS_2 = \{topic, < (r_1, w_1)...(r_m, w_m) >\}$, where r_i is a binary relation between two topic concepts. Two different forms of topic relations are considered by this approach: (1) syntax-based relations that exist between the verbs and their arguments; and (2) context-based relations (C-relations) that exist between entities and events that occur in the same context, but are otherwise not syntactically or semantically dependent on each other. C-relations are motivated by: (a) frequent collocations of certain nouns with the topic verbs or topic nominalizations, and (b) an approximation of the intra-sentential centering, as introduced in (Kameyama, 1997). After a set of topic signature terms has been generated for a particular set of documents, topic relations are then generated for each noun and verb in TS_1.

Enhanced topic signatures are calculated in the following way:
Step 1: Generate candidate relations: in each document relevant to the seed relation, all syntax-based and C-relations are identified. To discover topic relations we have used a very large corpus of texts: the AQUAINT corpus (LDC Catalog # LDC2002T31) which contains 375 million words corresponding to about 3GB of data. The seed relation becomes a query q that is used by the SMART IR system (Buckley et al., 1998) to generate the set of relevant documents. These documents are processed to identify Verb-Subject, Verb-Object, and Verb-PP relations. Document processing starts with the identification of named entities. Part of speech (PoS) tags and non-recursive, or basic, noun phrases are identified using the transformation-based learning method reported in (Ngai and Florian, 2001). Simple verb phrases (VP) and prepositional phrases (PP) are identified with finite state automata (FSA) grammars. Syntactic relations, such as Verb-Subject, Verb-Object, and Verb-Prepositional Attachment are recognized by another FSA. The C-relations are discovered by creating a salience window for each verb in the document. The NPs of each salience window are extracted and ordered with an ordering relation introduced in (Kameyama, 1997). Both syntax-based relations and C-relations are expanded by replacing names with their semantic classes or by replacing words with concepts from a large, hand-crafted ontology of over 200,000 English words.

Step 2: The candidate topic relations are ranked following a method introduced in (Riloff, 1996). Each relation is ranked based on its *Relevance-Rate* and its *Frequency*. The *Frequency* counts the number of times a relation is identified in the set of relevant documents. *Relevance-Rate = Frequency/Count*, where *Count* measures the number of times an extracted relation is recognized in any document, relevant or not.
Step 3: Select a new topic relation based on the ranking in Step 2.
Step 4: Restart the discovery by using the latest discovered relation for classifying relevant documents. The discovery procedure stops after N=100 iterations or when no new relations are discovered.

(Examples of the output of TS_1 and TS_2 for DUC 2005 topic 413 are presented in Table 1.) Once the set of TS_1 terms and TS_2 relations have been identified, Lite-GISTexter assigns each sentence in the document collection a score equal to the sum of the weights of all of the terms and/or relations it contains. Sentences are ranked based on this composite score.

Topic Siganture

steel(n)	851	nucor(pn)	721
mill(n)	400	plant(n)	225
crawfordsville(pn)	211	furnace(n)	202
slab(n)	181	integrate(v)	144
british_steel(pn)	137	ravenscraig(pn)	132
big_steel(pn)	125	us_steel(pn)	116
busse(pn)	108	producer(n)	107
industry(n)	104	tonne(n)	99
capacity(n)	94	steelmaker(n)	93
roll(v)	90	ton(n)	90
steelmaking(n)	84	sheet(n)	84
cost(n)	83	overcapacity(n)	69
mr_busse(pn)	61	cast(v)	59
scrap(n)	57	greenfield(n)	55
arbed(pn)	53	iron(n)	52
production(n)	50	msa(pn)	50
technology(n)	49	subsidise(v)	45
product(n)	45	method(n)	42
labour(n)	41	sms(pn)	40
madrid(pn)	40	ore(n)	39
europe(pn)	37	minimill(n)	37
basque(pn)	36	blast(n)	36
manufacturer(n)	36	subsidy(n)	32

Enhanced Topic Siganture

sheet - steel	49	steel - producer	35
make - steel	32	NE:OTHER - plant	28
NE:LOCATION - steel company	24	produce - steel	20
steel - market	20	sheet - market	20
mini mill - group	20	steel - manufacturer	19
Basque - country	18	NE:MEASURE - ton	17
Ensidesa - NE:OTHER	17	NE:OTHER - AHV	17
Ensidesa - AHV	17	begin - operation	16
expand - capacity	15	greenfield - site	15
NE:MEASURE - tonne	14	steel - maker	14
Nucor - pioneer	13	cast - steel	13
Nucor - have	13	steel - mini mill	13
Crawfordsville - plant	13	report - dollar	13
thin slab - cast	13	announce - plan	12
NE:NUMBER - job	12	cut - job	12
NE:OTHER - process	12	industry - minister	11
NE:LOCATION - producer	11	EC - official	11

Table 1: Signatures for topic 413 - *Steel production*

Lite-GISTexter uses techniques implemented in LCC's question-answering systems to approximate the information need of each DUC topic. First, complex questions in each topic were syntactically decomposed: conjoined NPs and lists were split into separate sub-questions. For example, a question like, *What technical and financial problems have they encountered?* was split into 2 sub-questions: *What technical problems have they encountered?* and *What financial problems have they encountered?*. (As was mentioned in Section 2, no semantic decomposition was performed for this year's topics.) Next, keywords were extracted automatically from each sub-question, and stopwords were removed. These keywords were then associated with sets of alternations originally developed for LCC's PALANTIR automatic Q/A system. (No manual selection of keywords was performed, however.) A sample of these alternations for two different terms is provided in Figure 3. Question keywords (and their corresponding alternations) are then used to re-rank the topic-weighted sentences for each sub-question; the top-ranked sentences are then sent to a summary generation module.

benefit: advance, advantage, aid, ameliorate, assist, avail, better, build, contribute to, favor, further, improve, make it, pay, pay off, profit, promote, relieve, serve, succor, work for, acquire, derive, come by, receive, find, collect, obtain, help, payment

problem: state, resolve, difficulty, condition, affairs, effort, overcome, grapple, bear, bitch, predicament, quandry, plight, extricate, difficult, unplesant, trying, awkward, entangle, pinch, fix, holem, jam, mess, muddle, pickle, situation, hard, rough, stress, strain, job, trouble, hindrance, wrinkle, interfere, question, matter, issue

Figure 3: Alternations for the question keywords *benefit* and *problem*

Summary answers were generated by merging the top-ranked sentences selected from each subquestion into a single paragraph. Two simple optimizations were then performed to improve the overall quality and legibility of summaries. In order to reduce redundancy, we used a semantic parser to create predicate-argument structures for each sentence included in the summary; lower-ranked sentences that featured the same predicate-argument structure as higher-ranked sentences were dropped from the summary. In addition, we attempted to resolve pronouns in summary sentences by including the immediately preceding sentence from the sentence's original document; if this immediately preceding sentence also contained a pronoun, both sentences were dropped from the summary. This process was repeated until the summary reached a total of 250 words.

4 Results

In this section, we discuss the results of the 2005 DUC evaluations. Although Lite-GISTexter received very competitive scores for both ROUGE and the DUC "Quality Questions", it received highest marks from the metrics which evaluate the quality of information returned by a summarization system: Pyramids and Responsiveness.

Lite-GISTexter ranked first overall on the standard Pyramid scores and third overall on the "modified" Pyramid scores. We believe this to be strong evidence that our hybrid approach to QDS resulted in an ability to provide summary answers that approximated the information content of human summaries. Results from Pyramid and "modified" Pyramid scores are presented in Table 2.

peerid	average score	rank score	average modifi edscore	rank modifi edscore
14	**0.2477**	**1**	**0.1874**	**3**
17	0.2398	2	0.1972	2
10	0.2340	3	0.2000	1
15	0.2322	4	0.1793	5
7	0.2307	5	0.1840	4
4	0.2197	6	0.1722	6
16	0.2170	7	0.1706	7
32	0.2134	8	0.1607	12
6	0.2110	9	0.1639	11
19	0.2089	10	0.1672	9
12	0.2086	11	0.1645	10
11	0.2085	12	0.1691	8
21	0.2063	13	0.1589	13
26	0.1970	14	0.1413	15
28	0.1944	15	0.1400	17
3	0.1894	16	0.1459	14
13	0.1855	17	0.1412	16
25	0.1691	18	0.1395	18
1	0.1666	19	0.1258	20
27	0.1631	20	0.1306	19
31	0.1587	21	0.1215	21
24	0.1491	22	0.1140	22
20	0.1446	23	0.0937	24
30	0.1376	24	0.1131	23
23	0.1216	25	0.0609	25

Table 2: Results for the Pyramid metric.

Since details were not forthcoming, we are not able to provide any better analysis of why our system dropped from first place under the Pyramid score to third place under the "modified" Pyramid score. It is our expectation that with advance details of the "modified" Pyramid score, we could have optimized our summary output for this somewhat different method of scoring.

It is our contention that the Pyramid scoring system can be used to provide an objective measure of how responsive machine-generated summaries are to individual questions. As described in (Nenkova and Passonneau, 2004), Pyramid scores depend on the collection of most frequently occurring semantic content units (SCUs) found in the human summaries prepared for a given scenario. Taken together, these SCUs can be seen as representing the information that should be contained in a perfectly responsive answer summary. Since Pyramid scores are derived from the percentage of SCUs that human annotators believe to be represented in a machine-generated summary, we argue that systems that receive higher Pyramid scores are better able to meet the information needs of each scenario. In addition, we recommend that Pyramid scoring be used to evaluate the results of future TREC Relationship Tasks: by increasing the rigor of the evaluation process, we feel that Q/A systems could receive a more representative evaluation of their answers than they do currently.

In contrast, our responsiveness score was 18.77, which was good enough for tenth overall (seventh among systems participating in the Pyramid evaluations). Although human summaries have generally received high marks for this metric, we believe that Responsiveness crucially depends on an evaluator's own conception of what constitutes a sufficiently responsive summary answer. Unlike Pyramid scores, which depend on common elements extracted from multiple different summaries, Responsiveness scores rely on one person's interpretation of what should – and what should not – be in a summary. We believe that the difference between our Pyramid and Responsiveness scores (in terms of overall rank) stems not from issues necessarily with the information content of our summaries, but other factors such as discourse coherence and topic structure which could lower a user's satisfaction with an otherwise contentful summary.

Lite-GISTexter received high marks for both ROUGE and several of the DUC "Quality Questions". For the ROUGE-1 metric we obtained an F-score of 0.36066, as presented in Table 3. While our high score (4.34, on average) for "Grammaticality" is not surprising, as Lite-GISTexter extracts only complete sentences from text, our 4.5 (average) score for "Non-redundancy" suggests our approach based on overlapping predicate-argument structures removes duplicate passages without exacting a high content penalty. Although we would have preferred to make use of a pronoun resolution system in Lite-GISTexter, our 3.3 score for "Referential Clarity" suggests that our heuristics for dealing with pronouns may be an adequate stop-gap solution.

SYSID	ROUGE-1	95% CI Lower	95% CI Upper
C	0.45853	0.44324	0.47210
A	0.45567	0.44369	0.46759
E	0.44885	0.43254	0.46450
I	0.44816	0.43067	0.46759
J	0.44330	0.43089	0.45781
B	0.44232	0.42814	0.45673
D	0.43973	0.42724	0.45305
G	0.43959	0.42694	0.45186
F	0.42580	0.40723	0.44368
H	0.41501	0.39971	0.43161
15	0.37978	0.37468	0.38470
4	0.37517	0.37020	0.38047
17	0.36925	0.36324	0.37478
10	0.36146	0.35647	0.36626
19	0.36128	0.35607	0.36656
6	0.36114	0.35588	0.36602
14	**0.36066**	**0.35386**	**0.36710**
7	0.35782	0.35261	0.36312
11	0.35715	0.35187	0.36249
24	0.35402	0.34883	0.35878
25	0.35297	0.34788	0.35840
16	0.35212	0.34728	0.35700
3	0.34833	0.34316	0.35351
21	0.34538	0.34040	0.35032
13	0.34278	0.33747	0.34837
12	0.33997	0.33465	0.34494
28	0.33778	0.33208	0.34350
27	0.33403	0.32871	0.33908
32	0.33302	0.32813	0.33816
26	0.32411	0.31303	0.33369
30	0.32313	0.31775	0.32836
20	0.31782	0.30867	0.32617
31	0.30155	0.29048	0.31175
1	0.29243	0.28329	0.30122
23	0.24151	0.23238	0.24971

Table 3: Results for the ROUGE-1 metric.

5 Conclusions and Future Work

In this paper, we have shown that question-directed summarization requires a hybrid approach that combines representations of the topic of a document collection along with representations of the information need of complex research scenarios. We believe that Lite-GISTexter's top marks in Pyramid score (and competitive marks for ROUGE and Responsiveness) stems from its ability to retrieve sentences that are both topical and also relevant to the user's overall information need.

In future work, we plan to experiment with more sophisticated ways of combining question representations and topic representations for sentence selection and ranking. As we learn more about what makes information topical (as well as responsive), we expect the performance of both our QDS and our complex Q/A systems to improve. Finally, although we have focused primarily on maximizing the relevance and responsiveness of our summaries, we acknowledge that we now need to start employing better summary organization techniques in order to increase the topic consistency and overall coherence of our summary answers.

References

Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. 1998. SMART High Precision: TREC 7. In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, pages 285–298.

Sanda Harabagiu, Andrew Hickl, and John Lehmann. 2005. Experiments with Interactive Question-Answering. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics*, Ann Arbor, MI.

Sanda Harabagiu. 2004. Incremental Topic Representations. In *Proceedings of the 20th COLING Conference*, Geneva, Switzerland.

Andrew Hickl, John Lehmann, John Williams, and Sanda Harabagiu. 2004. Experiments with Interactive Question-Answering in Complex Scenarios. In *Proceedings of the Workshop on the Pragmatics of Question Answering at HLT-NAACL 2004*, Boston, MA.

M. Kameyama. 1997. Recognizing referential links: An information extraction perspective. In *Proceedings of the*

ACL/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, pages 46–53.

Finley Lacatusu, Paul Parker, and Sanda Harabagiu. 2003. Lite-GISTexter: Generating short summaries with minimal resources. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1. Association for Computational Linguistics.

Finley Lacatusu, Andrew Hickl, Sanda Harabagiu, and Luke Nezda. 2004. Lite-GISTexter at DUC 2004. In *DUC 2004*, Boston, MA.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th COLING Conference*, Saarbrücken, Germany.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: the Pyramid Method. In *HLT-NAACL 2004*, Boston, MA.

Grace Ngai and Radu Florian. 2001. Transformation-Based Learning in the Fast Lane. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*, pages 1044–1049.

Sharon Small, Tomek Strzalkowski, Tracy Janack, Ting Liu, Sean Ryan, Robert Salkin, Nobuyuki Shimizu, Paul Kantor, Diane Kelly, Robert Rittman, Nina Wacholder, and Boris Yamrom. 2004. HITIQA: Scenario Based Question Answering. In *Proceedings of the Workshop on the Pragmatics of Question Answering at HLT-NAACL 2004*, Boston, MA.