Answer Extraction System by Question Type from Query-Biased Summary for Newspaper Articles

Yohei SEKI
Dept. of Informatics, The Graduate University for Advanced Studies (Sokendai)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Dept. of Integrated Information Technology, Aoyama Gakuin University
6-16-1 Chitosedai Setagaya-ku, Tokyo 157-8572, Japan
seki@grad.nii.ac.jp/seki@it.aoyama.ac.jp

Abstract

Recently, many researchers are focusing on the application of Natural Language Processing (NLP) techniques such as summarization, information extraction, and text mining. One of the challenges with these technologies is developing an accurate Question and Answering System [1]. In this paper, we will discuss Japanese Q&A problematic issues that have appeared in my experimental system. My system is implemented with query-biased summarization techniques to mine from a number of documents.
Keywords: NTCIR, Japanese Q&A System, multi-document summarization technique, information fusion from multiple newspaper articles, and QAC (Question and Answering Challenge).

1 Introduction

There is a year long workshop being held by the National Institute of Informatics in Japan called NTCIR-3. We participated in the 'Question and Answering Challenge' (QAC) formalrun [2] in the spring of 2002: Japanese Q&A tasks. We created an experimental system for the Japanese Q&A to detect problems specific to the Japanese language. Our input data was Mainichi Newspaper articles from 1998 and 1999 Year. This included about 230,000 articles. In this paper, we implement and test one approach for Q& A to take answers from a query-biased summarization such as in Okumura([6]) to mine from multiple document sources. We also discuss some Japanese related problematic issues.

This paper consists of seven sections. We explain the tasks of QAC in Section 2, and discuss details of our system design and approach in Section 3. Section 4 provides an overview of our system user interface. Section 5 contains a brief evaluation of our system with QAC problems. In Section 6, some problematic issues are discussed. Finally, we present our conclusions in Section 7.

2 Question and Answering Tasks in QAC

The Question and Answering Challenge (QAC) [2] consisted of three tasks. The first and second task contained the same 200 questions. A list of five accurate answers was the goal in the first task; The goal of the second task was to extract the correct answer set. The third task had 40 problems and each problem had one follow-up question. The formalrun with these three tasks was held on four consecutive days in May, 2002.

The Answers were to be noun phrases which indicated a person's name, organization names, money, size, date and so on. The source documents were a two-year-period of Japanese newspaper articles.

3 Query-Biased Multi-Document Summarization Based Approach for the Q&A System

My approach for the Q&A System consisted of three procedures: question analysis, summarization of questions from various articles, and answer formation.

3.1 Question Analysis

The Question analysis process is basically divided in two parts. One is the detection of question type, and the other is the extraction of keywords with a numeric score that summarizes documents. We use the Japanese part-of-speech tagger, 'Chasen'[1] in order to break the question sentences into morphemes. Question types are categorized with keywords as follows:

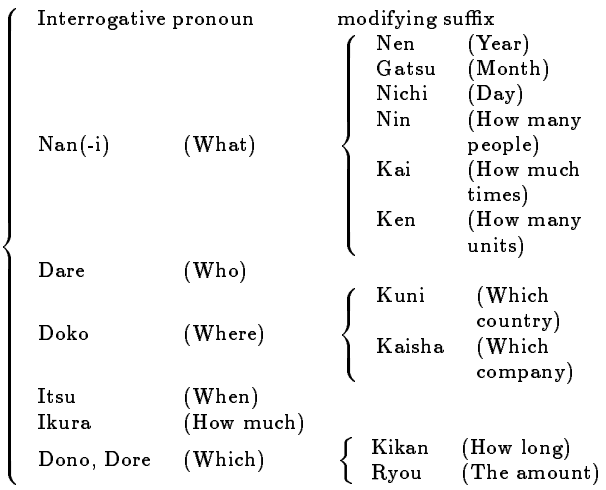

Figure 1. Japanese Question Taxonomy

The question taxonomy above shows that Japanese question types are determined by a combination of an interrogative pronoun and a modifying suffix.

Another process is keyword detecting and scoring. We score keywords in each question as follows:

1. Each matching noun morpheme receives 1 point.
2. The proper noun or phrase containing the proper noun receives 3 points.
3. A time related adverb/noun receives 0.5 points.
4. Each verb or adjective morpheme (except some basic elements) receives 1 point.

3.2 Sentence Extraction with Multi-Document Summarization Technique

Next, we extracted sentences related to each question keywords from a two year supply of newspaper articles. The question keyword scores determine these individual sentence scores.

If a sentence contains a keyword, the keyword score is added to the sentence score, then the sentence score is divided by the sum of all the keyword scores in that question. Therefore, a maximum score of a sentence is 1. If a score of any sentence is more than 0.4, the sentence is extracted and stored into the answer file for that question. This is a kind of cut and paste summarization technique [3] from a wide source of newspaper articles [5]. In order to accelerate our system's performance, some multi-document summarization techniques [4] with text segmenting and clustering [7] were also needed. When this MDS approach is adopted, the Q&A accuracy performance must be kept in mind. MDS has some information fusion or aggregation processes to avoid overlapping information. If this process was applied wrongly, the correct answer would be removed from summary. We did not implement this process at this stage but implemented a similar process at the answer formation stage.

3.3 Answer Formation from Summary Sentences

Answer Formation is the process of extracting answers from summary sentences using question types. We implemented this step as pattern matching according to question type information with Perl. We use question type information like

[1] http://chasen.aist-nara.ac.jp/

Nan-Nen Nan-Gatsu (In what year and month did the event happen?), and encode that information in regular expressions like $/(0-9)\{1,2\}gatsu(0-9)\{1,2\}nichi/$ in order to detect answer candidates.

Some question types were needed to extract distance patterns or make answers with a parsing technique. We implemented noun formation functions according to question types with a recursive function about part-of-speech information (concerned with noun morpheme type). The noun phrase formation process was different according to question types and was localized with Perl functions. Some examples are as follows:

1. Who (Dare) Questions
   'Chasen' tagged personal names as 'noun-proper noun-personal name'. When 'Chasen' tagged a personal name correctly, the personal name is extracted based on the noun formation. In addition, an abbreviated name like 'J.F.K.' or some hard to place place noun needs to be extracted with an answer formation process. This type answer was not tagged correctly with the morpheme tagger. Therefore, we need some parsing technique to look before and after the part-of-speech information.
2. When (Itsu) Questions
   'When' questions' difficulties mainly stemmed from unknown details: What year, month, day, or time? We extracted answers from 'when' questions with time-related number extraction and formation. When some time-related suffixes were matched, this pattern was formed following Japanese conventional time-expressing order; year, month, and day. When time information was expressed with 'of' or other modifying terms, there might be gaps between some time expressions. For example, 'In Keicho 5 (1600), the war of Sekigahara started on the 15th September.' The year and the date are separated in the sentence but both are necessary in an answer. If that information together was expressed in one sentence, my system would have no problem extracting the correct answer to form one time expression.
3. Where (Doko) Questions
   'Where' questions also varied in their answers according to the details. To find a specific location of an event such as a war in East Timor in Indonesia, the initial input question might not be able to place 'Daerah Istimewa Aceh' province without wider geographic information. The morpheme tagger tagged a place noun as 'noun-...-place' and a country name noun as 'noun-...-place-country'. In my system, this distinction is judged mainly based on question keyword information. When the question was judged to be concerned with country name, the corresponding function was called.
4. Amount Questions
   In the Japanese language, amount information is characterized with a modifying suffix like 'liter' or 'cubic meter'. Therefore, this suffix information is key in extracting an answer. Number information was tagged correctly as 'noun-number' or 'prefix-auxiliary-number'. My system formed these elements to make quantity noun phrases.

Extracted answers were scored with their source sentence score and their occurring frequencies. Some answer candidates with same meanings were merged to a single answer with information fusion or aggregation techniques to avoid overlapping answers.

3.4 Detecting Answers for Follow-up Questions

In Task 3, we employed a different approach because follow-up questions often contain pronouns instead of nouns and don't contain specific keywords. To extract an answer in a follow-up question, we use a summary from the first question and the question type pattern in the follow-up question. Thif formal run, I cannnot submit the result because of time-consuming problem.

4 System User Interface

The Q&A system produced summaries including sentence weights and source article ID numbers. They were tagged in XML-style formats. When the answer formation process was executed, answers were provided with their occurring articles by using summary information. This system is shown in Figure 2.

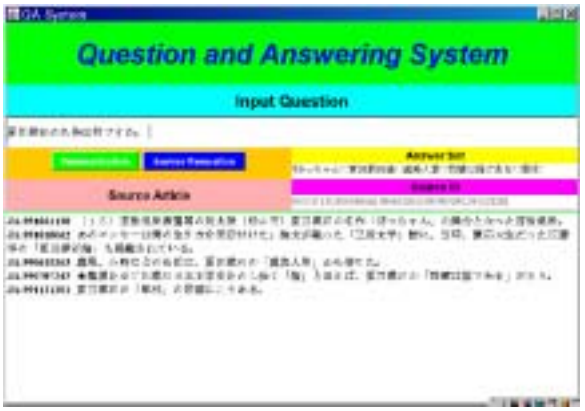

Figure 2. Q&A System

5 Evaluation

QAC results were evaluated with MRR (Mean Reciprocal Rank) and F-score (or F-measure) [8] metrics. Some bugs in our system were removed after the dryrun was finished. The results of our present system are shown as follows.

1. Task 1 (Top five Q&A)
   Task 1 had 200 questions. My system score The total score ranges are shown in Table 1. More accurate result according to each question is shown in Table 2.

SysID	S10012
Task	TASK1
Points	19.5
Question	200
Answer	305
Output	634
Correct	45
Recall	14.754
Precision	7.098
F-measure	9.585
MRR	0.1

Table 1. Scoring in Task 1

2. Task 2 (Answer Set)
   Task 2 had the same questions as Task 1. The goal of Task 2 was to extract the correct answer set. Our system answered this task as the best 20 answers. F-score ($\frac{2 \times Precision \times Recall}{Precision + Recall}$) ranges are shown in Table 3.

SysID	S20008
Task	TASK2
Points	9.2
Question	200
Answer	305
Output	2414
Correct	63
Recall	20.656
Precision	2.61
F-measure	4.634
MRR	0.046

Table 3. Scoring in Task 2

3. Task 3 (Follow-up Q&A)
   Task 3 had 40 follow-up questions to each of the original questions. I cannot submit my results for Task 3 before deadline bacause of time-consuming problem of my system.

6 Some Problematic Issues

In this research, we only used surface information and didn't use deeper semantic information

1001	×	×	×	×	×
1002	11 月 11 日 (○)	×	×	11 日 (○)	×
1003					
1004	×	×	×	×	×
1005	×	×	×	×	×
1006	×	来年 10 月 (○)	×	×	10 月 1 日 (○)
1007	×	×	×	×	×
1008	×	×	×	×	×
1009	×	×	×	×	×
1010					
1011	×	×	×	×	
1012	×	×	×	×	×
1013	×	スピルバーグ (○)	スティーブン・スピルバーグ (○)	×	×
1014	×	×	×	×	×
1015	×	×	×	×	×
1016 〜 39	timeout				
1040					
1041	×	×	×	×	×
1042	×	×	×	×	×
1043					
1044					
1045	×	和同開珎 (○)	×	×	×
1046	×	ベルリン (○)	×	×	×
1047	×	×	×	×	×
1048	×	×	×	×	×
1049					
1050	×	×			
1051	×	×	×	×	×
1052					
1053	×	×	×	×	×
1054	×	加藤紘一 (○)	小泉純一郎 (○)	×	×
1055	×	×	×	×	×
1056	11 月 28 日 (○)	先月 28 日 (○)	×	28 日午前 2 時 35 分 (○)	×
1057	×	×	×	×	×
1058	×	湯川秀樹 (○)	江崎玲於奈 (○)	×	×
1059	×	サントリー (○)			
1060	×	×	伊東千秋 (○)	×	×
1061	×	千葉県浦安市 (○)	×	千葉県 (○)	×
1062	×	×	×	×	×
1063	×	×	×	義経 (○)	×
1064	青森 (○)	×	青森県大間町 (○)	×	×
1065	×	×	×	×	×
1066	×	×	×	×	×
1067	×	×	×	×	×
1068	×	×	×	×	×
1069					
1070					
1071	×	×	×	×	×
1072	×	×	×	×	×
1073					
1074	×	桜田慧 (○)	×	×	×
1075	×	×	×	×	×
1076	×	×	×	×	×
1077	×	×	×	×	×
1078	×	×	中国 (○)	×	×
1079	×	×	ガル (○)	×	×
1080					
1081	×	×	×	×	×
1082	×	×	×	×	×
1083					
1084					
1085	×	×	小渕 (○)	×	小渕恵三 (○)
1086	×	×	×	×	×
1087					
1088	×	×	×	×	×
1089					
1090 〜 99	miss				
1100	×	29 日 (○)	×	×	×

Table 2. Answer Situation in Task 1

1101	×	×	エルサレム (○)	×	×
1102	×	×	×	×	×
1103	×				
1104	×				
1105					
1106					
1107					
1108	***NTT*** データ通信 (○)	***NTT*** データ (○)	×	×	×
1109	×	×	×	×	×
1110	×	×	×	×	×
1111	×	×	×	×	×
1112	×	×	×	×	×
1113					
1114	×	×	×	×	×
1115	×	佐々木 (○)	イチロー (○)	×	佐々木主浩 (○)
1116	×	×	×		
1117	和同開珎 (○)	×	×	×	×
1118	×	×	×	×	×
1119	×	×	×	×	ロシア (○)
1120	×	×	×		
1121					
1122	×	×	×	×	×
1123	×	×	×	×	×
1124					
1125	日韓 (○)	日本 (○)	韓国 (○)	×	×
1126					
1127	×	×	×	×	×
1128	×	×	×	×	シュテフィ・グラフ (○)
1129	オーストラリア (○)	×	×	×	×
1130	×	×	×	×	×
1131	中国 (○)	×	インド (○)	ロシア (○)	×
1132					
1133	さつき (○)	×	×	×	×
1134	×	×	×	×	×
1135	×	×	×	×	×
1136	2001 年 (○)	×	×	×	×
1137	×	×	×	×	×
1138	×	×	×	×	×
1139	×	×	×	秋野不矩 (○)	阿川弘之 (○)
1140	×	×	×	×	×
1141	×	×	×	×	×
1142	×	×	×	×	×
1143	×	×	×	×	×
1144	×	キリバス (○)	×	×	×
1145	×	×	×	×	×
1146	×	秋元康 (○)	×	×	×
1147	×	×			
1148	×	BINGOBONGO・サンタマリア (×)	×		
1149	×	×	×	×	×
1150					
1151	×	×	×	×	×
1152	×	×	×	×	×
1153	×	×	×	×	×
1154	×	×	×	×	×
1155					
1156	×	×	×	×	×
1157	×	×	×	×	×
1158	×	×	×	×	×
1159	×	×	2003 年 (○)	×	×
1160	×	×			
1161	×	×	×	×	ポルトガル (○)
1162					
1163	×	×	×	×	×
1164	×	×	×	×	小泉純一郎 (○)
1165	×	×	×	×	×
1166	×	×	×	×	×
1167	×	×	×	×	×
1168	×	×	×	×	×
1169					
1170	×	×	×	×	×
1171	×	×	×	×	×
1172	×	×	×	×	×
1173	×	×	×	×	×
1174	×	×	黒沢明 (○)	×	黒沢 (○)
1175	×	×	×	×	×
1176	×	×	×	×	×
1177	×	×	×	×	×
1178 ～ 1200	timeout				

Table 2. Answer Situation in Task 1 (continuing from the previous page)

like a thesaurus would provide. Our result set contained erroneous elements, but in Task 2, $\frac{1}{5}$ of the correct answers were found. There are two reasons why correct answers were not found: there was too much erroneous information extracted and the correct answers were not extracted and put in the initial summary.

The source input data of QAC contained a very large (about 230,000) amount of articles. Our system caused some time-consuming problems because our system extracted summaries with common weighing values for every question type. Some questions extracted too many summary and others didn't extract enough summaries. In fact, the assigned threshold 0.4 was very sensitive according to question types. When this threshold was set as '> 0.4' (not equal), some questions contained more accurate answers in the best 10 answer candidates, but other questions' answers were missed. Although our threshold , of course, can be changed easily according to question type, some explicit criteria between threshold values and question types were hard to establish. In addition, when commonly used and polysemous question keywords were detected, many sentences with erroneous elements were extracted.

On the other hand, answer quality problems mainly stemmed from the question analysis quality. Questions which extracted too much erroneous information were mainly concerned with unique personal names or too specific place names. Other questions which did not contain correct answers were relatively unique-patterned questions. In order to increase the accuracy, we need to use a more semantic sensitive program.

We explained our improvement strategy for the Japanese Q&A problematic issues. In Japanese, there are two ways to say 'in the second place': "Dai-ni-i" and "ni-i". In the latter, the prefix "Dai" is omitted. We implemented a noun phrase formation to detect an answer with a parsing technique, but the two Japanese examples above came up with two different answers. A technique in detecting same meanings to make a single answer is also needed. This technique is a kind of multi-document summarization technique [4], especially for information fusion from multiple sources.

In addition, our system has a time-consuming problem. According to question type, our system made a big summary from source text, then answer formation process took a long time to extract. I divide my answer formation process to five parts of 40 questions, and executed parallel process. But from 200 questions, my system cannot make answer lots of questions because of this problem.

7 Conclusions and Future Direction

We tested our experimental Q&A System mainly using morpheme type information and the multi-document summarization based technique. Our results contained some correct answers and each answer was provided with its occurring article ID number. Therefore, our system is useful for checking results with people.

In Japanese, question analysis process is a little more complex than English because question type is determined with the combination of interrogative pronoun and modifying suffix. A parsing and information fusion techniques regarding Japanese morphemes are needed in implementing the answer formation process.

In order to improve our results, some semantic information for the question category or taxonomy of inquiries [1] may be needed to reduce the amount of incorrect answers from a large summary source. In addition, if the assigned threshold for summarization is changed according to question type information, better results will follow. In order to determine precise thresholds according to question types, we will try more Q& A tasks and adjust our system.

Acknowledgements

We thank the National Institute of Informatics and members held for NTCIR-3 QAC, and also thank Robert B. Whitehead for re-reading the English.

References

[1] J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, and S. H. et al. Issues, tasks and program structures to roadmap researh in question & answering (q & a). http://www-nlpir.nist.gov/projects/duc/roadmapping.html, 2001.

[2] J. Fukumoto and T. Kato. An overview of question and answering challenge (qac) of the next ntcir workshop. http://www.nlp.cs.ritsumei.ac.jp/qac/qac-ntcirWS2.pdf, 2001.

[3] H. Jing and K. McKeown. Cut and past based text summarization. In ANLP-NAACL 2000, Seattle, WA USA, May 2000.

[4] I. Mani. Automatic Summarization, volume 3 of Natural Language Processing. John Benjamins, Amsterdam, Philadelphia, first edition, 2001.

[5] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In the 18th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 74–82, Seattle, WA USA, July 1995.

[6] M. Okumura and H. Mochizuki. Query-biased summarization based on lexical chaining. Computational Intelligence, 16(4):578–585, 2000.

[7] G. C. Stein, T. Strzalkowski, and G. B. Wise. Summarizing multiple documents using text extraction and interactive clustering. In Pacific Association for Computational Linguistics (PACLING-1999), 1999.

[8] G. C. Stein, T. Strzalkowski, G. B. Wise, and A. Bagga. Evaluating summaries for multiple documents in an interactive environment. In 2nd Int. Conf. on Language Resources & Evaluation (LREC2000), 2000.