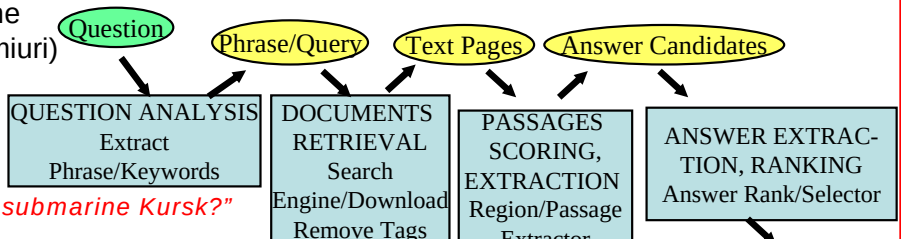
Question Answering System with Fine Grain Answer Types and Search Refinement

Iwate Prefectural University (Japan) / Faculty of Software and Information Science / David Ramamonjisoa

Inputs: a question in Japanese; a set of keywords

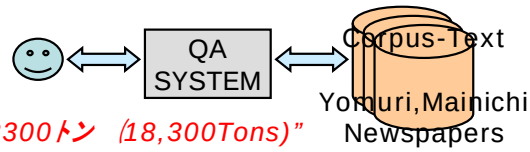
Output: a set of possible answers drawn from the corpus (newspaper resources: Mainichi and Yomiuri)

"A simple factoid question"



沈没したロシアの原子力潜水艦クルスク号はどれくらいの大きさでしたか。

"How big was the size of the sank Russian atomic submarine Kursk?"

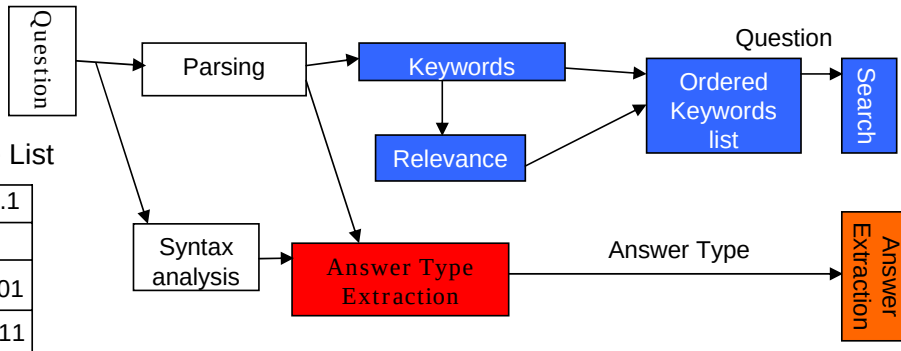


"1万8300トン (18,300Tons)"

Answer: "A word/number likely to be the desired answer"

Question Processing

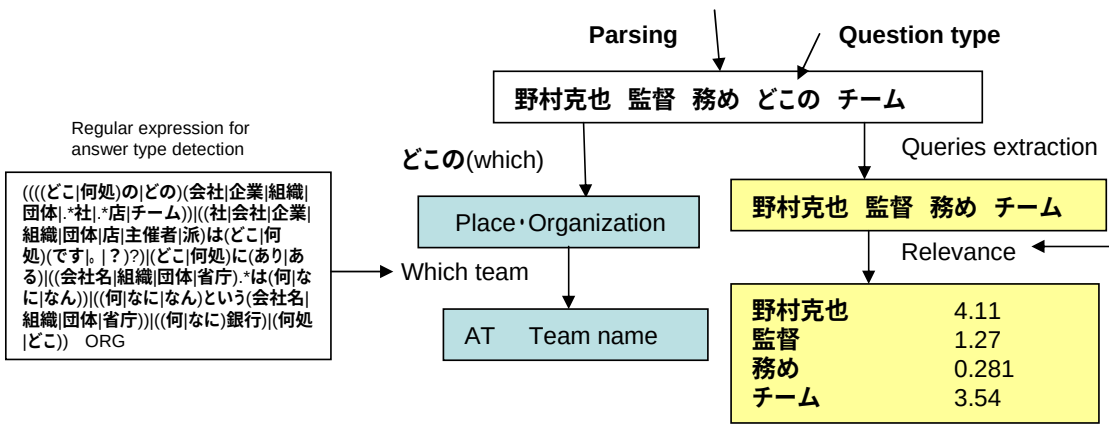
Fine Grain Answer Type and Keyword Relevance



Example of Keywords and Relevance List

野村克也	4morphemes	+1.1
	Proper noun	+3
	First word (question)	+0.01
Total Relevance		+4.11

野村克也氏が監督を務めたのはどこのチームですか。



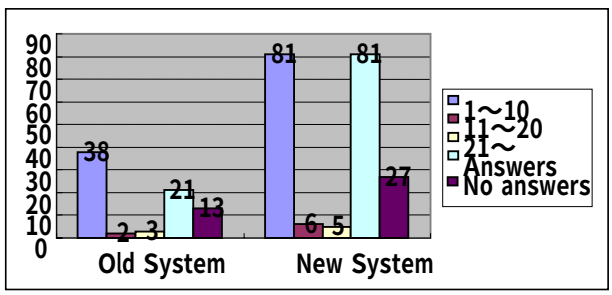
Regular expression for answer type detection

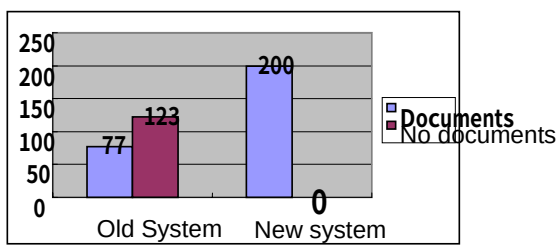
((((どこ|何処)の|どの)(会社|企業|組織|団体|.*社|.*店|チーム))|((社|会社|企業|組織|団体|店|主催者|派)は(どこ|何処)(です|。|？)?)|(どこ|何処)に(あり|ある)|((会社名|組織|団体|省庁).*は(何|なに|なん))|((何|なに|なん)という(会社名|組織|団体|省庁))|((何|なに)銀行)|(何処|どこ)) ORG

Word pattern	weight
カタカナが含まれている	+2
「」に囲まれている	+3
数値が含まれている	+3
文字数が１文字	+0．2
文字数が２文字	+0．２５
文字数が３文字	+0．5
文字数が４文字	+1．1
文字数が５文字以上	+1．2
品詞が一般名詞	+1
品詞が固有名詞-人名	+3
品詞が自立動詞	+0．００１
N番目の検索質問	N×0.01

Experiment results

Keyword Relevance effects in answers extraction





Comparison of the found documents by the two systems

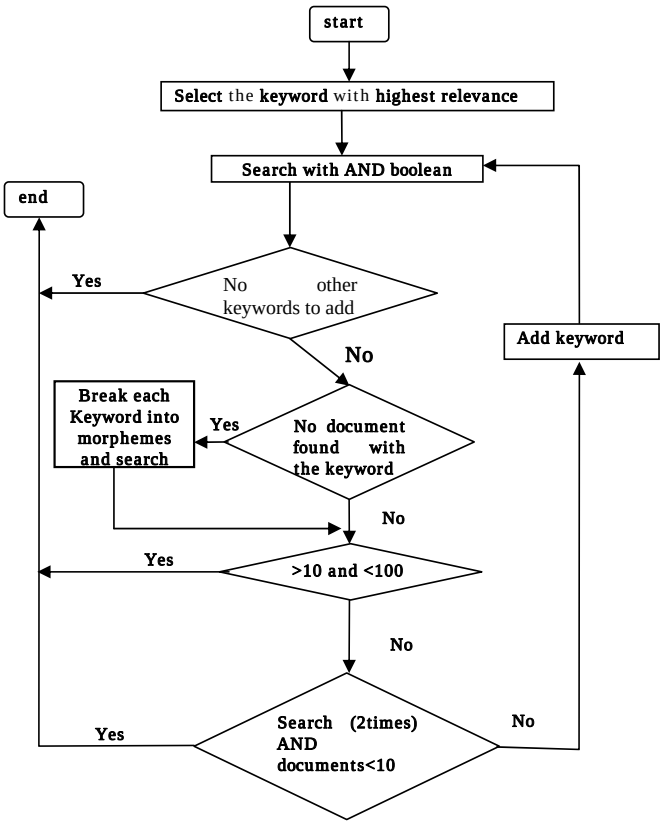
Discussion

For 200 questions, the result shows that only 38 answers are provided by the old system compared to 81 answers for the new system.
The new system has better performance in the top 11 to 20 and top 21 to 50. The new system could retrieve 81 documents related to the 81 questions, in contrast to only 21 questions retrieved by the previous system. This proves that document search was improved to extract documents for each question
Correct Candidate Answers are found in 173 questions for 200 questions test.

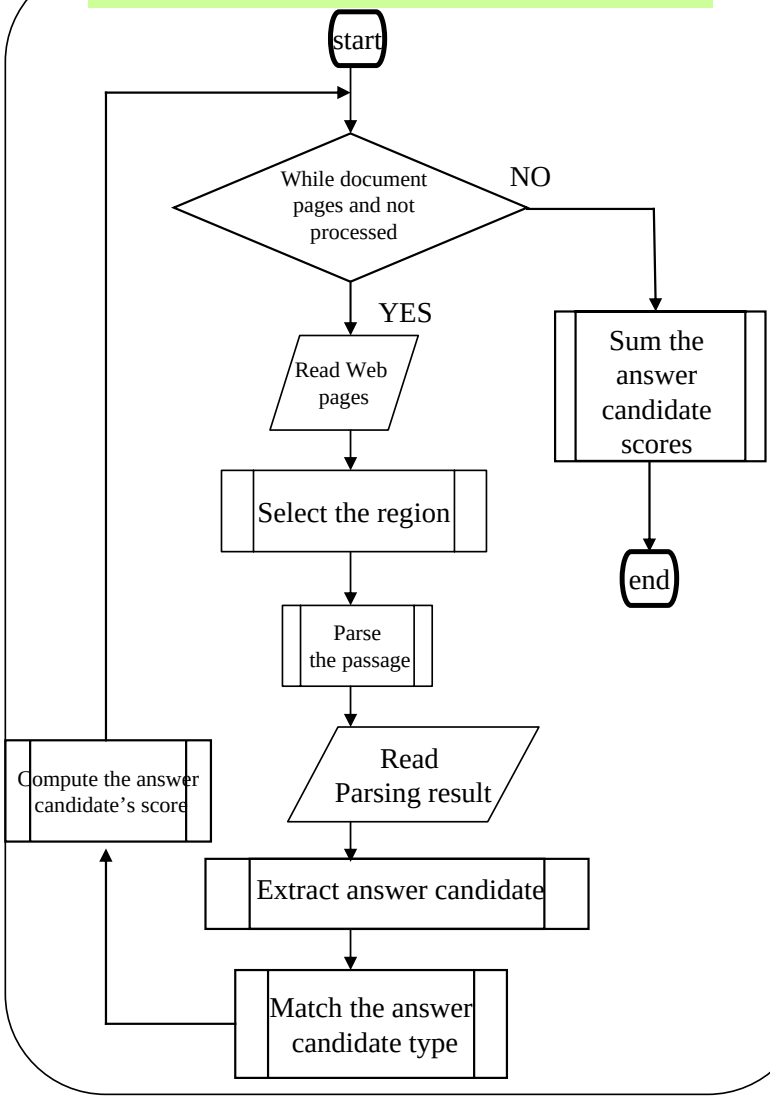
Iterative search and passage selection

The search module looks for relevant articles by using a Boolean search program for a corpus. The search is performed iteratively by adding or removing terms in the Boolean search query in order to obtain a sufficient number of articles for the answer extraction. The result of search is a ranked list according to the similarity measure *sim(q,D).* The algorithm for passage selection is described below. The passage retrieval sub-module selects the best passage from the page having similar keywords.

Flowchart for Document Search



Answer Extraction Module



Scoring

$$S1(a)=\sum_{i=1}^{n}\frac{1}{d_i+1}\ (1),\quad S1'(a)=S1(a)+\frac{1}{r(a)}\ (2),\quad S2(a)=S1'(a)*(1+\frac{1}{n_{types}})\ (3),\quad TotalScore(a_i)=\sum_{j=1}^{num\_passages}S2(a_i)\ \ (4)$$

Conclusion and Future Work

A new question analysis and passage search technique were introduced to improve an existing Question Answering System. This system is then tested with the NTCIR-QAC2 corpus and then applied to NTCIR-QAC3. The system has shown a better performance in documents retrieval and question analysis in comparing to the older version. However, we realize that the answers extraction modules must be redesigned and reimplemented in order to make the system competitive. In the future, we are aiming to develop a cross-lingual question answering system.