| | |
|---|---|
| Title | 多様な質問を受け付ける質問応答システムの回答選択に関する研究 |
| Author(s) | 横川，裕太 |
| Citation | |
| Issue Date | 2015-03 |
| Type | Thesis or Dissertation |
| Text version | author |
| URL | http://hdl.handle.net/10119/12626 |
| Rights | |
| Description | Supervisor: 白井　清昭，情報科学研究科，修士 |

# Study on Answer Selection in Question Answering System that Accepts Various Questions

Yuuta Yokogawa (1310075)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 12, 2015

**Keywords:** Question Answering, Answer Selection, Answer Type, Machine Learning.

Question answering (QA) is a task to retrieve an answer to a question given by a user from a collection of documents. Such a system is called question answering system (QA system). The early studies on question answering mainly focus on factoid QA, which is a task to answer named entities or short sentences. Recently, it is more expected to develop a QA system that can answer various questions that ask a reason or method for example (non-factoid questions). A general QA system analyzes what kinds of information (e.g. person name, location, reason) are asked by the user, retrieves candidates of answers from the documents, and chooses only the answer candidates that agree with the user's request. The information type asked in the user's question is called 'answer type' (or 'question type'). However, it is difficult to retrieve the correct answer if the answer type of the question is not pre-defined or the ambiguous question is given. Furthermore, it is also difficult to define a comprehensive set of the answer types in advance.

This research proposes a question answering system that can accept a wide variety of questions, although it does not classify an answer type of a given question. In the proposed method, the score of the answer candidate is measured from two points of view: 'content relevance' and 'matching degree of answer type'. 'Content relevance' evaluates similarity between the contents or topics of the question and the candidate of the answer. While, 'matching degree of answer type' evaluates degree of agreement of the answer types between the question and answer candidate. The proposed system does not explicitly classify the answer type of the question and answer candidate, but just check if the answer types of them agree or not implicitly. To measure it, a classifier to judge the agreement of the answer types is trained by supervised machine learning, where styles or characteristics of the question and answer candidate are used as features. The way how to create a training data of the classifier is also examined. Furthermore, this research proposes two methods to combine the above two kinds of scores for answer selection.

Web documents are used as the source of the proposed QA system. Content relevance between the question and answer candidate is defined by cosine similarity between content word vectors of them. To improve preciseness of them, following methods are also proposed: (1) putting more weights for important words in the question, (2) adding words appearing in a context of the answer candidate to the word vector and (3) a rank of the search engine is considered to the content relevance score.

To estimate the matching degree of answer type, the binary classifier, which judges if the answer types of the given pair of the question and answer candidate agree, is trained by supervised machine learning, namely L2-regularized logistic regression. The public tool LIBLINEAR is used for training the classifier. The probability of the positive class provided by LIBLINEAR is used as the score of the matching degree of answer type. Six kinds of features are used for training. Three are extracted from the question: interrogative word, 3-gram including interrogative word and expression in the end of the question sentence. The other three are extracted from the answer candidate: expression in the end of the clause, combination of them and sequence of functional words. The training data is constructed by using a dataset of Yahoo! Answers Japan, a community-driven QA site. The positive examples are the pairs in the Yahoo! Answers dataset. While, the negative samples are pairs of questions and answers of another question. Preventing from accidentally choosing the question and answer whose answer types are same as the negative sample, a question $q'$ that are dissimilar to the question $q$ is searched, then $q$ and the answer $a'$

of $q'$ is chosen as the negative sample. Similarity between the questions is measured by cosine similarity of the feature vectors only consisting of the features extracted from the questions.

It is desirable that the ratio of the positive and negative samples in the training data is same as that in the set of the candidates of answers retrieved by the QA system. In this research, the ratio is estimated by simulation of answer candidate retrieval in the QA system on Yahoo! Answers dataset. First, the pairs of questions and answers in the dataset are subdivided into several clusters by a clustering algorithm. Next, for each question (called 'query question'), the system searches answers relevant to it and retrieve the top $N_d$ answers from a collection of the answers. The same document retrieval module in the proposed QA system is used for answer retrieval, and $N_d$ is set to the same number of the documents retrieved by it. If the question of the retrieved answer is classified in the same cluster of the query question, we suppose it is a positive sample (i.e. answer types are same), otherwise negative sample. The ratio of the positive and negative sample is estimated as the ratio in the set of answers retrieved for all questions in the dataset. In our experiment, it is estimated as 1:5.9.

Two methods to combine the scores of 'content relevance' and 'matching degree of answer type' are proposed: 'filtering method' and 'addition method'. In the filtering method, the answer candidate is discarded when the answer type classifier judges the answer type of it disagree with the question. The score of each remained answer candidate is defined as the content relevance score. In the addition method, the relative scores, proportion to the highest score in the set of the answer candidates, of both are added with the equal weights.

We present experiments to evaluate the proposed methods. First, the performance of the answer type classifier was evaluated. 10% of the data (the positive and negative samples constructed by the above method) was used as a test data, and the rest 90% was used as a training data. Accuracy etc. of the judgment were measured. The accuracy was 0.86 when the training data with 1:5.9 positive and negative samples was used. Through investigation of effectiveness of the features, it was found that the features extracted from the question were more effective than that from the answer.

Next, the performance of the proposed QA system was evaluated. A set of 35 questions of 7 types, 5 for each type, is prepared as the evaluation data. The answers to these questions obtained by the system was evaluated by mean reciprocal rank ($MRR$), average precision of top 10 answers ($AP'$) and precision of top 10 answers ($P_{top10}$). Two proposed systems outperformed a baseline, a system that ranks the answer candidates only with the content relevance score. Therefore, it was found that agreement of the answer types of the question and answer was important for answer selection. The QA system developed by the training data where the ratio of positive and negative samples was 1:5.9 outperformed the systems using 1:1 or 1:10 training data. We can conclude that our method to guess the ratio of the positive and negative samples is useful. Comparing the filtering method and addition method for answer selection, the latter was mostly better. Its $MRR$, $AP'$ and $P_{top10}$ were 0.63, 0.53 and 0.27, respectively. However, the filtering method was better for the questions that ask definition of an entity or a fact.