Title	質問応答システムにおける詳細な質問タイプの同定手法の実装と評価［課題研究報告書］
Author(s)	若山，龍太
Citation	
Issue Date	2014-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/12041
Rights	
Description	Supervisor:白井清昭，情報科学研究科，修士

An Implementation and Evaluation of Fine-grained Question Type Identification for Question Answering System

Ryota Wakayama (1110702)

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

February 8, 2014

Keywords: Question Answering system, Machine learning, Question type, SVM, k-NN.

A general Question Answering system is known as a man-machine interactive system which receives a question represented as a natural language and responds an answer of it. Question Answering system contains several sub-modules: analysis of question sentence, identification of question type, document retrieval, extraction of answer candidates and selection of the answer. Each technology used in the sub-modules is so important for our society, and it can be applied for many natural language processing (NLP) applications. In this paper, we focus on identification of question type. Usually, question types represent the categories of the question. "PERSON", "LOCATION", "COMPANY" and so on are examples of question types. The question type plays an important role in a traditional Question Answering system. Question Answering system first identifies the question type of the question, then extracts named entities which are same kinds of the question type as answer candidates. Therefore, identifying the question type is important to improve the performance of the overall Question Answering system.

The goal of this report is to develop a question type identification module toward a practical QA system. There are three sub-goals to achieve it. First, we define fine-grained question types. Eight types of question defined by IREX project are often used in many previous researches. In this report, we define question types following “Sekine’s Extended Named Entity Hierarchy”. It is a hierarchy of types of named entities consisting of 200 fine-grained types. We assume that these named entities can be utilized as fine-grained question types. It enables QA system to distinguish the type of the question more sophisticatedly. To the best of our knowledge, this is the first attempt to automatically identify the question type defined by Sekine’s Extended Named Entitity Hierarchy.

Second, we implement question type identification module based on machine learning. Furthermore, we empirically evaluate what features contribute to improve the accuracy of identification of question type.

Finally, we try to use two different kinds of training data. For supervised learning of a classifier for question type identification, a corpus of question sentences annotated with correct types of questions is necessity. However, there is no corpus with gold fine-grained question types derived from Sekine’s Extended Named Entity Hierarchy. Construction of such a corpus requires much human labor. On the other hand, many corpora annotated with named entities, most of them are general text such as newspaper articles, are available. In this report, a large amount of named entity tagged corpus as well as a small collection of questions with question types are used to train the classifier of question type identification.

As described above, the supervised machine learning is applied for identification of question type. We used following four features to classify question types, (1)Content word, (2)Word bi-gram, (3)Interrogative and (4)Word dependency. To obtain these features, the sentences in the training data are analyzed by morphological tool and dependency parser. In general, words can be classified into “content word” or “ancillary words”. Content word is a word which has its own meaning. While words which do not have its own meaning but have grammatical function are called ancillary words. For example, noun, verb, adjective are classified as content word, while postposition, auxiliary verb are classified as ancillary word. The first feature is content words in the question sentence. The second feature is word bi-gram. It is a N-gram of words where N = 2. In other words, word bi-gram is two adjacent words in the question sentence. For example, if a question is “夏目/漱石/の/代表作/は/何/です/か/? (what is the most representative work of Soseki Natsume?)” (‘/’ indicates a word boundary), 夏目+漱石, 漱石+の, の+代表作 and so on are word bi-gram features. Word bi-gram may represent a meaning of the sentence more precisely than word uni-gram (or content word, i.e. the first feature). The third feature, interrogative is one of the important keyword to interpret what is asked by the question. Interrogatives “what”, “where”, “which”, “who”, “when”, “how” and “why” are used as features in this research. The last feature, word dependency means a dependency relation between words. It can be extracted by a dependency parser.

As mentioned above, no public corpus with fine-grained question types is available. It makes question type identification difficult because of lack of the training data. To tackle this problem, the following two corpora are used as the training data QAC corpus and Newspaper corpus. QAC corpus is a collection of 1,218 questions for QA system provided by QAC–1 (Question Answering Challenge–1; a workshop of question answering system evaluation). We manually annotate each question with its fine-grained question type. Newspaper corpus is a collection of Mainichi-shimbun newspaper articles annotated with named entity tags. It contains about 290,000 named entities. Usually only the first corpus is used as the training data for identification of question type, but data sparseness is serious because the number of sentences in it is too small. However, we consider that second corpus might also be useful for obtaining significant information to identify question types. We compare the accuracy of question type identification when either or both of two corpora are used in our experiment.

We use two algorithms of machine learning to classify types of questions: one is SVM(Support Vector Machine), the other is k-NN method. SVM is one of the non-parametric supervised learning method suggested by V.Vapnik in 1995. The one of the outstanding features of SVM is known as a “maximum margin classifier”. SVM finds a ‘separate hyperplane’ that split positive and negative samples in high-dimensional feature space and ‘support vector’ that is the nearest positive or negative sample to the hyperplane. The separate hyperplane is chosen so that Euclidean distance (margin) between hyperplane and support vector becomes maximum. In NLP research fields, SVM performs better than other machine learning algorithms in many previous studies. In k-NN(nearest neighbor), means a training sample similar to the target data. The target data is classified by majority voting of the labels of the k nearest neighbors. Each data is represented as feature vector to measure the distance or similarity. There're several ways to measure similarity of two vectors. Jaccard coefficient, Dice coefficient and Simpson coefficient are often used to calculate similarity in k-NN method. In this report Dice coefficient is used.

The several experiments have been conducted to evaluated the methods implemented by this report. We used QAC corpus for the evaluation. By five-fold cross validation, we train the classifiers by SVM or k-NN and calculate the accuracy of the classifiers on the test data. When Newspaper corpus is used as the training data, all QAC corpus is used as the test data. When both QAC and Newspaper corpus is used as training data, five-fold cross validation is also applied: QAC corpus is split into test and training data and Newspaper corpus is added to the training data.

According to the results of the experiment, the best accuracy was 60.3% where the classifier is SVM with features except for word bi-gram trained from QAC corpus only. Among k-NN classifiers, the best accuracy was 52.0% where $k = 5$ and the feature set and training corpus are same as SVM. Hereafter we discuss the results of the experiments from following four viewpoints. First, difference of definition of question type is considered. The best accuracy 60.3% was significantly worse than the accuracy reported by the privous work. For example, Sasaki et al. reported that 88.0% accuracy was achieved by a supervised classifier. However, number of question types is only 8 in their study, while we used two hundred question types based on Sekine's Extended Named Entity Hierarchy. Since identification of fine-grained question types is more difficult than coarse-grained ones, it is natural that our result was worse. However, there is much room to improve the accuracy for practical Question Answering system. Second, difference of the machine learning algorithm is discussed. The best accuracy of SVM was 63.0%, while k-NN 52.0%. It seems that SVM is more effective algorithm for question type identification. The third discussion is effectiveness of the use of Newspaper corpus as the training data. The best accuracy of the SVM classifier trained from both QAC and Newspaper corpus was 56.1%, while 60.3% when only QAC corpus is used.

Although the size of Newspaper corpus is much greater than QAC corpus, it could not improve the accuracy. This may be because that sentences in Newspaper corpus are declarative, not questions. Finally, effectiveness of individual features are discussed. We used the following four features: content word, word bi-gram, interrogative and word dependency. Validity of each feature is examined as follows: we remove one feature from the feature set, train the classifier and compare the accuracy of it with the classifier trained with all 4 features. When SVM classifier is trained from only QAC corpus, the effective features were content word and interrogative. However, the accuracy was dropped when word bi-gram or word dependency is used as a feature, indicating that they are ineffective. On the other hand, using k-NN method($k = 5$), the most effective feature was interrogative, followed by content word and word dependency. Word bi-gram did not contribute the gain of the accuracy due to the lack of the training data. When using Newspaper corpus, however, the most effective feature was word bi-gram. Word bi-gram seemed to work effectively since the size of Newspaper corpus is large.