Development of the Natural Language Processing Tools for the Information Retrieval

Koji Sekiguchi (1210905)

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

March, 2014

Keywords: Information Retrieval, Morphological Analysis, Named Entity Extraction

Nowadays, information retrieval has become an essential tool for computer users. However, there are problems in the current information retrieval. It could not fulfill the following two expectations of the users: (1) it is unable to return the information that the user wants to know, (2) it incorrectly return the information that the user do not wants to know. The former problem is referred to as "retrieval omission", the latter is referred to as "retrieval error". If there is a retrieval omission, users must enter queries again until they can obtain necessary results. If there is a retrieval error, users must take a long time to find the necessary documents in the huge set of returned documents. In general, there is a trade-off between these two problems. When increasing the output of a search engine to decrease the retrieval omission, the retrieval error will increase. When reducing the output of a search engine to decrease the retrieval error, the retrieval omission will increase. In this report, I had proposed a method to overcome both problems not simultaneously but sequentially. In the proposed method,

to reduce the retrieval omission first, then to reduce the retrieval error by refining the search results interactively.

In order to reduce the retrieval omission, I developed the following three tools that utilize the existing research on the natural language processing: (1) a program to extract the word and its acronym, (2) a program to extract knowledge of the orthographical variants ('kanji okurigana' in Japanese), (3) a morphological analysis tool for N-best path search.

In "the program to extract the word and its acronym", I implemented the method proposed by Sakai et al. with some modification so that it can be applied for a dictionary as knowledge source. By applying the developed program to Wikipedia, the recall of the extracted words and their acronyms was more than 90%, but the precision was below 70%. However, when relaxing the evaluation criteria where not only acronym but also related words are correct, the precision was more than 80% with little loss of the recall. "The program to extract knowledge of the orthographical variants ('kanji okurigana' in Japanese)", where 'kanji okurigana' is kana character written after kanji character to complete the full reading of the word, is a tool that automatically extracts the set of groups of the orthographical variants of the okurigana from a word dictionary. By applying the program to IPAdic consisting of 392,126 word entries, 9,449 groups of the orthographical variants were obtained. "The morphological analysis tool for N-best path search" is a morphological analyzer that can handle the ambiguity of word segmentation in Japanese text. It tokenizes Japanese texts and outputs all tokens on the path of the minimum cost (the best path) first, then outputs noun tokens only on the 2nd to N-th ranked paths. Since it would not output useless tokens such as duplicated and non-noun tokens on the 2nd and lower ranked paths, the computational cost can be reduced in the information retrieval. Furthermore, by outputting noun tokens on the 2nd and lower ranked paths, we can expect that search omission due to mismatch of the word segmentation could be reduced. This morphological analyzer is based on Nagata's method to search the N-best paths for a given input sentence. In Nagata's method, the program will analyze the input sentence forward one character at a time to build the word lattice. Then by searching the built word lattice, it outputs tokens on the N-best paths in ascending order of costs. As data structure of the word dictionary of the morphological analyzer, I implemented a double array trie module that requires only a short storage of the memory. In addition, the output from both of "the program to extract the word and its acronym" and "the program to extract knowledge of the orthographical variants" are incorporated to the word dictionary in the developed morphological analyzer. By experiment, it is confirmed that the morphological analyzer can output tokens on the N-best paths and also output tokens obtained by these two additional programs. When building an inverted index in information retrieval system with this morphological analyzer, improvement of the recall could be expected.

By techniques against the retrieval omission problems described so far, the precision will decrease. To solve this problem, an interactive search to refine search results is considered. Here the interactive search refers to the procedure to repeat search by adding new search criteria or queries to the first query entered by a user. To ‘refine’ the search results, newly added queries are concatenated with AND operator with the original query. This operation enables us to increase the precision gradually. In order to implement the interactive search, the documents to be searched must have structured fields other than the text. As most of documents such as newspapers do not have such structure, the interactive search is rather difficult to be performed. Applying the technique of the named entity extraction on those documents, however, structured fields including named entities, which can be used for the interactive search, are automatically annotated. Similar to part-of-speech (POS) tagging, named entity extraction is a kind of the sequential labeling problem. From the named entity tagged corpus, the model to determine a named entity tag for each individual tokens is automatically obtained by machine learning. In this report, Conditional Random Field (CRF) is chosen as the machine learning algorithm. An annotated corpus with Sekine’s extended named entity hierarchy tags is used as the training data of the machine learning. From the points of view of business use, the tags of Sekine’s extended named entity hierarchy are

mapped to 9 newly defined coarse-grained NE tags. Using 12,000 sentences excerpted from Sekine's extended named entity tagged corpus and evaluating the implemented named entity extraction tool by 3-fold cross-validation, I found the great gap of F-measures among different named entity tag types. To reveal causes of the differences, the frequencies of POSs of named entities are examined. It is found that the general POS was the most frequent for the low performance named entity type, while POSs that highly related to the meaning of the named entity were frequently appeared for the high performance named entity type. Furthermore, adding more effective features for machine learning will be good for improving the F-measure on the named entity extraction.