Automatically Building the lexicon of Variant in Japanese that belongs to Attribute of Product for Opinion Mining

LIU CHAOYU(1410047)

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

February 10, 2016

キーワード: Knowledge Acquisititon, Attribute of Product, Variant in Japanese, Lexicon, Opinion Mining .

There are some phenomenons in Japanese, such as the part of words which appears in the Okurigana, the word which has long sound or not, the difference of type of characters, which describe the same object by the different strings. These phenomenons are called variant in Japanese. The variant in Japanese is an important problem in the process of language.

In one side, Online Shopping is widely used in recent years. The users often contribute the opinions after using the products to the shopping websites. These opinions are valuable for the makers and the online shopping users. For the makers, these opinions are useful for the improving the design of products and changing marketing strategy in future. For the online shopping users, before buying the products, they can know that the products whether fit themselves or not by looking through these opinions which were writen by the other online shopping users. However, there is such a huge amount of opinions that the users cannot look through all of them.Thus, it is necessary to study a technique of processing the informations of opinions for the products.

In this research, I aim at building a lexicon in Japanese about attribute variant of the product automaticly. The data for the processing are collected from the spec tables and the informations of opinions of the product. There are two charaeristics in this research. One is that it extract attributes from both table and review text. Two is that pattern toextract variant of attribute is automatically acquired from review text.

The outline of the proposed method is as follows. The first step is what extracts attribute variants from the spec tables which collected from kakaku.com and the makers 'webpages. The second step is what extracts attribute variants from the review texts which have been collected from kakaku.com. The third step is what combines attribute variants extracted

from the previous two steps to build the lexicon of products attribute variant. In the first step, firstly, I collect the html source files from kakaku.com by the Python language and the BeautifulSoup library. In kakaku.com, the spec tables are designed in the same format so that it is easy to extract. Then it accesses into the makers 'webpage by the links appear in the products 'webpage of kakaku.com, to collect html source files.I create five rules to extract spec tables from the the makers 'html source files. These five rules is summarized by analysing many spec tables from the maker. The first rule is that the class attribute of table tag contains ' spec '. The second rule is that before the table tag, the class attribute of brother 's node contains ' spec '. The third rule is that before the table tag, the brother 's node contains keywords which means it is a spec table. The fouth rule is that the table is the only one in the webpage.The fifth rule is that the tables with the same class attribute which appear more than ten times. Then it extracts attributes and values from the spec tables. There are some rules to identify positions of attribute and value in the table. For the table in kakaku.com, the cell of th tag as attribute is extracted and the cell of td tag as value is extracted. For the table in maker 's webpage, the first rule is that in every line, the number of th tag is zero, the number of td tag is more than two, in this occasion, the cell of the first td tag as attribute is extracted and the cells of the rest of td tags as value are extracted. The second rule is that in every line, the number of th tag is one, the number of td tag is more than one, in this occasion, the cell of th tag as attribute is extracted and the cells of all of the td tags as value are extracted. The third rule is that in every line, the number of th tag is two, the number of td tag is one, in this occasion, the cells of th tags as attribute are extracted and the cell of the td tag as value is extracted. The last rule is that in every line, the number of th tag is two, the number of td tag is zero, in this occasion, the cell of the first th tag as attribute is extracted and the cell of the second th tag as value is extracted. Next, I make the feature vector of a set of attribute and value by using the rules I created. At last, processing these feature vectors by using CLUTO which is aggromerative clustering. This process is to gather the attributes which means the same meaning. In the processing of clustering, I set 90 percent of the number of attributes as the number of cluster 's group. The early stage of building the lexicon of products attribute variant is finished. In the second step, I first collect html soruce file from kakaku.com and then extract the reviews from the html soruce files. Next, I cut the review texts into some sentences by the stopwords such as a period or the space. Then, it extracts the sentence which contain the attribute in the lexicon that builded in the former step. After processing these sentences, it gets some patterns which is used for extract attibute variant. In this process, I first do the Morphological analysis by Mecab on these chosen sentences. I extract patterns from part of sentences which the number of words between an attribute and an evaluation word

is less than five. The evaluation word is from Japanese Sentiment Polarity Lexicon. Next I score these patterns. The scoring is based on two rules. The first one is that the number of sentences which are matching with the pattern, is more than three. The second one is that the value of the number of sentences which are matching with the pattern and the attribute in the early stage of building the lexicon divide the number of sentences which are matching with the pattern is more than 0.5. Using the patterns which meet the rules as aboved to make matching process with all of the review texts. The word A appears in the same position with the attribute in the pattern is the variant of the attribute if A is a Noun. Thirdly, I combine attribute variants extracted from the previous two steps to build the lexicon of variant in Japanese of products attribute. In this step, the new attribute is thought to be the varient of the old attribute which is appears in the early stage of lexicon and is used for make the pattern for extract the new one.

The proposed methods are evaluated by the experiments. The purity of extracting the spec table from makers 'webpage is 0.89. The recall of extracting the spec table from makers 'webpage is 0.82. And the purity of extracting the set of attribute and value from the maker 's spec table is 0.90. The purity of clusting is 0.829. I know that it is a good result in the early stage of building lexicon of the products 'attribute variant. The evaluation of extracting attribute from review text are as follows. The purity of extracting for the notecomputer category is 0.34. The purity of extracting for the camara category is 0.067. The purity of extracting for the TV category is 0.12. In future, I will search for a better method for choosing extract pattern.

In this research, when extracting the spec table from the makers 'web pages, I only extracted the ones which table tags meet the rules I created. But there are many styles which are not created by table tag in the spec table. In future, I decide to extract spec table which is not table tag from the makers 'web pages. In the process of clustering, I set 90 percent of the number of attributes as the number of cluster 's group. In future, I will search a better method for decide the number of cluster 's groups. In addition, the lexicon of the products 'attribute variant which builded by the approving method is used for the real anaylizing the informations of the evaluations. To check how much the attribute variant lexicon can contribute to the anaylizing the informations of the evaluations.