Adding Unknown words to a Thesaurus by Using Principal Components Analysis

Katsuhito Suzuki

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

February 13, 1997

Keywords: principal components analysis, data sparseness problem ,polysemy problem, co-occurrences.

For most natural language processing(NLP) systems, thesauruses comprise indispensable linguistic knowledge. Thesauruses have been widely used in past NLP research (e.g., word sense disambiguation, acquisition of synonym). They are handcrafted, machine-readable and have fairly broad coverage. From the viewpoint of NLP systems dealing with a particular domain, however, these thesauruses include many unnecessary (general) words and do not include necessary domain-specific word. Also, after a thesaurus is built, a word is never added newly, and there is a problem which can't cope with a change in the vocabulary.

To solve these problem, we propose a method that positioning new words in a thesaurus. In this paper, words that are not contained in the thesaurus but that appeared in the corpus more than once are called *unknown words*.

As the approach of positioning an unknown word to thesaurus, there is a thing based on the character infomation. A Kanji character is an ideograms, and it becomes the useful infomation concerned with the meaning of th word. When Japanese thesaurus is made the target, it can get semantic infomation comparatively easily by using infomation on the charcter(kanji). For example, a "natural language" belongs to the node which is the same as the "language", and a "economic problem" belongs to the node which is the same as the "problem". However, this character information isn't effective because it can't be applied in the case of the word that an unknown word is a katakana or an abbreviation.

As other approch, there is a thing based on the word co-occurrence data that extracted from a large corpus. For example, by the co-occurrences data of noun and verb in the one related to case, a noun in the thesaurus and an unknown word can be expressed in the term (case and verb).

Than whether to be similar of the term of those nouns and the unknown word, in the condition, the place of the arrangement can be decided. For the research based on the co-occurrences, there is Uramoto's, Tokunaga's, and Nakayama's method.

By Uramoto's research, when a person built a thesaurus, because used information (a standard for classification) isn't written, the standard for classification is extracted from the co-occurrences and by using the standard for classification and the cosine distance, how to arrange an unknown word in the upper lower thesaurus of ISAMAP was proposed. By Uramoto's technique, An unknown word is arranged in how many or the inside of node meeting that it adjoined it. And by Tokunaga's research, co-occurrences are used to estimate the probability for an unknown word that to belong to node of the thesaurus. An unknown word is arranged in node of the classified thesaurus of the classified vocabulary table based on the probability. By Tokunaga's technique, the place of some arrangement is presented in the unknown word. And by Nakayama's research, a standard for classification is extracted in the method which is different from Uramoto's one from the co-occurrences, how to arrange an unknown word in the classified vocabulary table was proposed by using the standard. Even Nakayama's method, the place of some arrangement is presented in the unknown word. A resemblance occasion calculation between the words becomes possible by using co-occurrences, and the place of the arrangement of the unknown word can be estimated. But, when a degree of a resemblance is calculated by using the co-occurrences, there are following two fundamental problems.

Data sparseness problem : There are usually many ways to express agiven concept, so the literal terms in an unknown word may not match those of a relevant word.

Polysemy probrem : Most words [1] have multiple meanings, so it will match the term of the word that are not of interest to term of the unknown word.

In the sparseness problem, it considers that the sparseness problem is easy to bring about by Uramoto's research to calculate a degree of a resemblance with a word to belong to node directly, and the unknown word. By Tkunaga's research, to calculate not a degree of a resemblance of the unknown word and the word but a degree of a resemblance with the unknown word and the word meeting, though the one for the deficiency of the co-occurrences can be made up for, it is weak when usual sparseness is compared with the way of dissolving it. As for the way of solving sparseness, it is the thing that some verbs which look alike semantically are collected by using the structure of the thesaurus. This way of dissolving it is being used by Nakayama's research. The drawback for this method is that some added terms may have different meaning(polysemy effect).

In the problem of the polysemy, it isn't being taken into consideration by the usual research. As for the method of the ambiguity solution of the word, polysemy are classified by the dictionary for more than one meaning and an appropriate thing is chosen from that meaning, however, so it is not satisfactory precision, it is not an effective method.

[1] The polysemy of the noun isn't taken into consideration, and only verbal polysemy is taken into consideration by this research.

Because there is a problem, the way of solving usual sparseness and polysemy, this research in, by using a statistical method that is called principal components analysis(PCA), it copes with the problem of sparseness and the problem of the polysemy which aren't handled by the usual research, and a degree of a resemblance in consideration of the nature of PCA is proposed.

This research in, by the character of PCA, because the same term as the way that before solves it is collected and it is divided a term and operation is possible. It thinks that the solution of sparseness and polysemy can be dissolved.

In the experiment, we used the partial tree about the "vehicle" of the EDR thesaurs, and got co-occurrence data from the EDR co-occurrence dictionary. It was prepared the 235 experiment data by using 10-fold cross validation in the experiment, and after dividing those experiment data into the sparseness problem and polysemy problem, it was registered and evaluated to thesaurus. As for the polysemy problem to use in the experiment, it was made the case that there was more than one meaning in the term in a restricted domain of the vehicle. Then, 213 questions become the polysemy problems. The sparseness problem is defined that when it calculates cosine between a experiment data and correct answer node that were prepared beforehand and the value of 0 cosine shall become. Then, 77 questions become the sparseness problems. It examines whether there is a correct answer by No.N place how to evaluate it as a result of 32 kinds of resemblance occasion calculations with middle node and the unknown word(N=1,5,10). A resemblance degree that an effect was able to obtain to the polysemy problem and the sparseness problem by PCA, as a result of an experiment is a proposal method and be biggest to, each problem, when it compares it with before PCA and 32.5% and 9.4% of improvements were observed. The accuracy of a resemblance degree that contributes a correct answer ranking to the sparseness problem most, when I admit to about 10 is 41.6% of Euclid distance and next was 32.5% of proposal method. Also, the accuracy of a resemblance degree that contributes to the polysemy problem most is 73.7% of a proposal method and next, was 66.2% of Tokunaga's method. By conventional research, the place of some arrangement is presented in the unknown word. There were about 10% of improvements, compared with a resemblance degree that was proposed with conventional research, if this method is complied with even this research.