Title	決定木学習を応用した日本語要約文の自動作成
Author(s)	原口, 良胤
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1223
Rights	
Description	Supervisor:奥村 学, 情報科学研究科, 修士

Japanese Text Summarization Using Applied Decision Tree Learning

Yoshitsugu Haraguchi

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

February 15, 1999

Keywords: automatic summarization, decision tree learning, machine learning.

These days, the number of on-line text is increasing, and it is not easy to obtain the information we need effectively. Therefore, the need of automatic sammarization technique is also increasing as well.

To apply decision tree learning to automatic summarization, it is neccesary to consider a sammary as a set of important sentences. Important sentences are selected from texts by a set of subjects manually. The characteristics of each sentence are caliculated as its properties, and with these properties, a decision tree is obtained by tarining. This decision tree is expected to choose the important sentences from texts.

A decision treee is constructed by training using a set of training data, which are classified into more than two classes. The constructed decision tree by training is able to classify new data into correct classes. However, when the number of data in one class is extremely smaller than those of the other classes, those data are considered as noise. As a result, the constructed decision tree is unable to classify minor data correctly. Automatic summarization using decision tree learning has this problem. Obviously, the number of important sentences in a text is much smaller than that of non-important sentences. Therefore, important sentences are not necessarily extracted correctly from texts.

In this paper, two level decision tree construction (TLDT) technique is introduced. By applying TLDT to automatic summarization, the problem described above is expected to be eased. First, 'near miss' sentences are extracted from a set of training texts. 'Near miss' sentences are non important sentences which are similar to important sentences in attributes. 'Near miss' sentences are coupled with important sentences, and classified as mix class. Other non important sentences which are not in mix class, are classified as far class. In the first decision tree, training is done with these two classes. As a result, the first decision tree is able to classify these two classes. After the first decision tree construction is done, the *mix* class is re-classified into important sentences and 'near miss' classes. With these re-classified data, the second decision tree is trained, and this tree is able to classify important sentences and non important sentences.

An experiment is done as follows.

In the experiment, 75 texts, which are articles, editorial and columns from 1995 Nikei newspaper, are used. Important sentences in the texts are selected manually.

First, attributes of each sentence are calculated. Attributes are as follows.

- sentence location in text
  normalized value of the number of previous sentences in a text
- sentence length
  normalized value of the number of characters in a sentence.
- sentence location in paragraph
  normalized value of the number of previous sentences in a paragraph
- $tf * idf$
  normalized value of $tf * idf$
- attitude of sentence
  discrete value describing a type of attitude
- lexical chain
  normalized value of the number of words contained in a lexical chain in a text
- adjunction
  discrete value describing a type of adjunction
- conjunction
  discrete value describing a type ofconjunction
- anaphora
  discrete value describing a type of anaphora

These attributes data are devided into a test set and a training set. Training set are classified into two, *mix* and *far* calsses. We tried 3 ways to calculate similarity of 'near miss' sentences as follows.

- similarity calculation using all attributes in sentences
- similarity calculation using a part of attributes in sentences
- similarity calculation using discrete attribute only

And also, we tried 2 ways of constructing decision trees. In one case, in extracting 'near miss' sentences and constructing decision trees, continuous values are convertied into discrete values. The other case, continuous values are used as they are.

Table 1: Comparison between C4.5 and TLDT(mix:far=1:2)

	recall	precision	f-measure
C4.5	0.222	0.461	0.300
TLDT	0.280	0.454	0.347

At the same time, to reduce the imbalance of the number of mix class and far class, we tried 2 ways. One is in constructing a training set, we choose the same number of far class data as that of mix class. And also, we tried the ratio of mix class to far class 1:2. To choose the data from far class, we ranked them according to the number of attributes similarity to mix. And, we select the data the lower rank.

The first decision tree is constructed with training set. And second decision tree is obtained by using re-classified data set. After these steps, test set are evaluated. For cross validation, we repeat this sequence 10 times.

As a result, comparing to C4.5 itself, TLDT shows relatively better performance in recall, precision and f-measure. The following table shows the comparison between the results of C4.5 and TLDT.

Where, f-measure is: $f - measure = \frac{2 \times recall \times precision}{recall + precision}$

According to the results, there seems a tendency that when the raitio of $mix : far$ is increased, recall reduces and on the contrary, precision increases. As a result, f-measure increases slightly.

In futre work, firstly, it may be possible to increase the overall recall by trying some more variation on 'near miss' extraction. And secondly, we are prepared to consider a summary ratio. A summary ratio is a value which describes the number of impoertant sentences extracted from a text. To construct training set according to summary ratio, we may be able to control the number of important sentences which the system extracts.