Modeling (In)Variability of Human Judgments for Text Summarization

Tadashi Nomoto
National Institute of Japanese Literature
1-16-10,Yutaka Shinagawa
Tokyo, 142-8585, Japan
nomoto@acm.org

Yuji Matsumoto
Nara Institute of Science and Technology
8916-5 Takayama Ikoma
Nara,630-0101, Japan
matsu@is.aist-nara.ac.jp

ABSTRACT

The paper proposes and empirically motivates an integration of supervised learning with unsupervised learning to deal with human biases in summarization. In particular, we explore the use of probabilistic decision tree within the clustering framework to account for the variation as well as regularity in human created summaries.

1. INTRODUCTION

In our recent study [1], we observed that an unsupervised method based on clustering sometimes better approximates human created extracts than a supervised approach. That appears somewhat contradictory given that a supervised approach should be able to exploit information about whether or not to include a given sentence in an extract, whereas an unsupervised approach blindly chooses sentences according to some selection scheme. A question is, why this should be the case.

The reason, we speculate, may have to do with the variation among judges in selecting sentences for a summary. Prior work on summarization informs us that judgments on extraction can vary widely among humans. Curiously, however, there is also a finding in [1] that a supervised system performs much better on data for which there is high agreement among humans than an unsupervised system, suggesting that there are indeed some regularities to be found and exploited to the system's advantage. So we might conclude that there are two, apparently conflicting aspects to human judgments on sentence extraction.

In the paper, we will explore an integration of supervised and unsupervised paradigms as a potential approach to accounting for the (in)variability of human judgments. More specifically, we will be concerned with extending the summarization framework in [1] by embedding a probabilistic decision tree within the clustering framework. The idea is to call upon clustering to deal with the variability and decision tree to deal with the invariability or significant biases.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2. METHOD

We basically follow our previous work [1] in building a generic summarizer which we call the diversity based summarizer (DBS). DBS is a two step summarizer which first brings up clustering to identify diverse topical areas in text and then refers to a tfidf ranking model to choose the best sentence from each topic area identified. In the present paper, we take a step further by extending DBS to incorporate supervised learning, in particular probabilistic decision tree (ProbDT). ProbDT works like a regular decision tree except that instead of generating a class label for a given instance, it generates the probability that it belongs to a particular class. The modification allows us to directly exploit human supplied information on which sentence to extract for a summary, which would not be possible with the clustering/tfidf-based-ranking framework. ProbDT also allows a ranking of sentences, which one would need to deal with variable length summarization.

Combining ProbDT and DBS can be done quite straightforwardly by replacing the tfidf based ranking part with ProbDT. Thus instead of picking up a sentence with the highest tfdif score, DBS/ProbDT strives to find a sentences with the highest score for $P(Select \mid \vec{u}, \mathrm{DT})$, the probability that a given sentence u is selected as one to be included in a summary, using a decision tree DT. A specific value of the probability is found by using the following equation:

$$P(Select \mid \vec{u}, \mathrm{DT}) = \alpha\big(\tfrac{\text{the number of "Select" sentences at } t(\vec{u})}{\text{the total number of sentences at } t(\vec{u})}\big).$$

$\vec{u}$ is a vector representation of sentence u. $t(\vec{u})$ denotes a leaf node in DT assigned to $\vec{u}$, and α some smoothing function.

Moreover, it would be interesting to examine whether performance of DBS coupled with ProbDT depends on a choice of decision tree algorithm. To see this, we consider three decision tree algorithms of different flavors: C4.5, MDL-DT, and SSDT. MDL-DT is like C4.5 except that it makes use of an optimized pruning based on Minimum Description Length Principle (MDL) in place of reduced error pruning used in C4.5. Another approach called SSDT aims at discovering recurring patterns in highly biased data, where a target class accounts for a tiny fraction of the whole data [2]. Note that the issue of biased data distribution is particularly relevant for summarization, as a set of sentences to be identified as summary-worthy usually account for a very small portion of the data.

Attributes for the decision trees include location, length

Table 1: Performance on JFD-1995 at varying compression rates. 'V' indicates that the relevant classifier is coupled with clustering.

cmpr. rate	C4.5*	C4.5*/V	MDL-DT*	MDL-DT*/V	SSDT*	SSDT*/V	Z	DBS(Z/V)
0.2	0.371	**0.459**	0.353	**0.418**	0.437	**0.454**	0.231	**0.429**
0.3	0.478	**0.507**	0.453	**0.491**	**0.527**	0.517	0.340	**0.491**
0.4	0.549	**0.554**	0.535	**0.545**	**0.605**	0.553	0.435	**0.529**
0.5	**0.614**	0.600	0.585	**0.593**	**0.639**	0.606	0.510	**0.582**

Table 2: Performance on synthetic data generated from the DUC-2001 corpus.

cmpr. rate	C4.5*	C4.5*/V	MDL-DT*	MDL-DT*/V	SSDT*	SSDT*/V	Z	DBS(Z/V)
0.2	**0.503**	0.263	0.215	**0.238**	**0.385**	0.232	0.093	**0.240**
0.3	**0.501**	0.295	0.262	**0.289**	**0.402**	0.274	0.110	**0.289**
0.4	**0.491**	0.354	0.297	**0.310**	**0.416**	0.305	0.127	**0.322**
0.5	**0.491**	0.335	**0.342**	0.338	**0.426**	0.318	0.150	**0.334**

and (tfidf) weight of a sentence.[1] Classes include 'Select' for those to be included in a summary and 'Don't Select' for those not. As a naming convention, we superscript a classifier's name with an asterisk to mean a probabilistic version of the associated classifier. Thus a probabilistic version of C4.5 is referred to as 'C4.5*.'

3. TEST DATA

We asked 112 Japanese subjects (students at graduate and undergraduate level) to extract 10% sentences in a text which they consider most important in making a summary. The number of sentences to extract varied from two to four, depending on the length of a text. We used 75 texts from three different categories (25 for each category); column, editorial and news report. Texts were selected randomly from articles that appeared in a Japanese financial daily *Nihon Keizai Shimbun* published in 1995. We assigned about 7 subjects to each article for the extraction task. The agreement among subjects turned out to be modest, scored 0.25 on the kappa scale, indicating the considerable variability in subjects' decisions. We assigned sentences with one or more votes to 'Select' class and those without to 'Don't Select' class. Of the total of 1424 sentences in the data, 707 were assigned to 'Select' and the remaining 717 were assigned to 'Don't Select.' Let us call the data set supplemented with human judgments 'JFD-1995.'

Another set of test data was *artificially* created from the DUC-2001 corpus (www-nlpir.nist.gov/projects/duc).[2] In contract to JFD-1995, which carries sentence by sentence judgments on extraction, the corpus comes with *abstracts*, along with the source texts. The problem is of course, an abstract often involves thorough restatements of its sources in the text so that it is no longer possible to identify sentences the abstract is actually based on. As a way out of the problem, we artificially assigned a 'Select' label to three sentences (in the source text) lexically most similar to each sentence in the abstract. The data set contained the total of 3775 sentences, of which 704 instances were labeled as 'Select' and the remaining cases were labeled as 'Don't Select.' We note one important difference from the Japanese corpus: while several people were involved in creating abstracts, each abstract was created by one judge.

[1] We define the tfidf weight of a sentence as the sum of tfidf's of terms it contains.

[2] The corpus consists of new wire articles from several sources such as Wall Street Journal, Associated Press, San Jose Mercury News, etc .

4. RESULTS AND DISCUSSION

Table 1 and 2 summarize results for the JFD-1995 and DUC-2001 data, respectively.[3] All the figures are in F-measure, i.e., $F = \frac{2*P*R}{P+R}$, and determined by 10-fold cross-validation, where one divides test data in 10 blocks, allocates nine for training and the remaining one for testing and averages scores over 10 folds. In either table, 'Z' refers to a baseline summarizer based on the tfidf based ranking model, which simply selects sentences whose sum of tfidf's for terms they contain ranks highest. Either table compares performance of C4.5*, MDL-DT*, and SSDT* against their performance when combined with clustering. Furthermore, we note that DBS is in fact Z/V, i.e, Z coupled with clustering.

Results in Table 1 and 2 strike one as rather 'similar' for the most part, even though they come from sources in different languages and domains: in either JFD-1995 or DUC-2001, clustering appears to boost performance of a summarizer whose ranking model is weak, e.g., Z, MDL-DT*; on the other hand, for a summarizer with a strong ranking model such as SSDT* and also C4.5* (for DUC-2001), clustering tends to hurt its performance. The fact that clustering is generally ineffective in DUC-2001 compared to JFD-1995 may imply that there are significant regularities to be exploited by supervised ranking models such as C4.5* and SSDT*, perhaps because each abstract in DUC-2001 is created singlehandedly.

It is curious to note that MDL-DT* is not performing as well as C4.5* and SSDT* in either corpus. The reason has to do with the property of MDL-DT* that it produces as small a tree as possible, covering the entire data space with a few regions. This could cause problems since points (sentences) would be assigned to the same probability even when they are separated far apart in the same region.

5. REFERENCES

[1] T. Nomoto and Y. Matsumoto. An experimental comparison of supervised and unsupervised approaches to text summarization. In *Proceedings of 2001 IEEE International Conference on Data Mining*, pages 630–632, San Jose, 2001. IEEE Computer Society.

[2] H. Wang and P. Yu. SSDT: A scalable subspace-splitting classifier for biased data. In *Proceedings of 2001 IEEE International Conference on Data Mining*, pages 542–549, San Jose, December 2001. IEEE Computer Society.

[3] Note that compressing a text by α% means picking up α% of sentences in the text.