Multi-document Summarization Using Support Vector Regression

Sujian Li, You Ouyang, Wei Wang, Bin Sun
Inst. of Computational Linguistics, Peking University
{lisujian, oyangu, bswen}@pku.edu.cn

Abstract

Most multi-document summarization systems follow the extractive framework based on various features. While more and more sophisticated features are designed, the reasonable combination of features becomes a challenge. Usually the features are combined by a linear function whose weights are tuned manually. In this task, Support Vector Regression (SVR) model is used for automatically combining the features and scoring the sentences. Two important problems are inevitably involved. The first one is how to acquire the training data. Several automatic generation methods are introduced based on the standard reference summaries generated by human. Another indispensable problem in SVR application is feature selection, where various features will be picked out and combined into different feature sets to be tested. With the aid of DUC 2005 and 2006 data sets, comprehensive experiments are conducted with consideration of various SVR kernels and feature sets. Then the trained SVR model is used in the main task of DUC 2007 to get the extractive summaries.

1. Introduction

The main task in DUC (Document Understanding Conference)[1] 2007 is similar to that in DUC 2005 and DUC 2006, which aims at generating a brief, well-organized, fluent summary for multiple documents with topic query guided. Due to the immaturity of the text generation techniques, most summarization systems are still designed with a summary extractive framework. The key of such a system is sentence extraction, to extract important sentences which can both represent the content of the documents and answer the questions users are interested in. A typical example is MEAD [Radev 2003], a framework for multi-document summarization, which got competitive performance. PolyU extended the features of MEAD for topic-based summarization task and performed well in DUC 2005 [Li 2005].

For most feature-based summarization systems, the feature weights are usually assigned manually by experience due to lack of training data. As corpus accumulates, machine learning approaches have been introduced to summarization task. FDU used a maximum entropy classification approach to extract sentences based on the key sentences corpus in DUC 2003[Zhao 2005]. OHSU [Fisher 2006] used human reference summaries for training corpus via ROUGE [Lin 2004] package and a supervised perceptron model for sentence scoring in DUC2006. In this paper, we will introduce the widely used Support Vector Machine (SVM) model for feature selection and weighting in extractive summarization. To solve the lack of corpus, we designed several automatic methods of generating training data which depend on the human reference summaries of DUC2006.

[1] http://duc.nist.gov

2. System overview

Our summarization system is designed with the extractive framework. Important sentences are extracted and re-organized to form a summary. Thus, the whole system is divided into three modules: text preprocessing, sentence scoring and post-processing. In text preprocessing, query and documents are segmented into sentences and news heads of the documents are removed. The goal of sentence scoring module is to evaluate the importance of each sentence with reference to their features. During post-processing, sentences with higher scores are extracted to compose the summary. To avoid information redundancy, a simple rule-based method is applied for removing redundant phrases Sentences are then reordered by chronological order. This year the key of our system is how to use the machine learning approach to combine features for scoring the sentences.

3. SVR-based Sentence Scoring

3.1 Support Vector Regression

SVM (Support Vector Machine)[Vapnik 1995, Vapnik 1998, Gunn 1998] is a widely-used and promising technique for pattern classification and regression estimation. The theory of SVM is based on the structural risk minimization (SRM) principle. For nonlinear case, SVMs map input data into a high dimension space, which can solve the problems of nonlinear classification.

Considering the problem of approximating the set of training data[Gunn 1998],

$$D = \{(x^1, y^1), ..., (x^l, y^l)\},$$

Where (x^1, y^1) represents a sentence with x^1 as its feature vector and y^1 as its score. Here a linear function $f(x) = w \cdot x + b$ is satisfied.

The SVR regression function is optimized through minimizing the following functional,

$$\Phi(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_i L(y_i - (w \cdot x_i + b)) \quad (1)$$

where C is a pre-specified value, and L is the loss function of the system. By introducing a normalization factor $\frac{1}{2} \| w \|^2$, SVR outperformed traditional regression approaches in both theory and practice, especially for small-scale samples. In this task, we use the LibSVM package to conduct regression.

3.2 Generating Training Corpus

Some feature-based sentences scoring systems [Li 2005, Li 2006] performed well with human-specified weights. However, their shortcomings are also obvious, e.g. lack of theoretical support, uncontrollable performance. In our system, to explore the potential of a fixed feature set, we use machine learning approaches to obtain the optimal scoring function. However, the main constraint of using machine learning methods is the lack of training data. To overcome this problem, we use reference summaries manually written to automatically generate data to train our SVM model.

The test data in DUC 2006 has been used in our work. There are 50 topics and each topic has 25 news documents and a topic query description containing a title and a narrative. For each topic, the organizers give 4 reference summaries written by different persons. Here we simply hypothesize that the more similar a sentence is to the four summaries, the larger its score must be. We designed two different strategies for sentence scoring based on sentence similarity measure. Given a sentences *s* and a standard summary *S*, the sentence similarity of *s* to *S* is defined as

$$Sim(s,S)=\frac{\sum_{t_i\in s}\sum_{t_j\in S} same(t_i,t_j)}{|s|} \quad (2)$$

where $same(t_i,t_j)=1$ if $t_i=t_j$ and 0 otherwise.

Given the definition of the similarity function, two sentence scoring strategies are described as follows.

(1) **Average:** The score of a sentence *s* is computed by summing the similarity of this sentence to each summary of the four human summaries.

$$Score(s)=\sum_i sim(s,S_i) \quad (3)$$

(2) **Maximum:** The score of a sentence *s* is computed by assigning the maximum of the similarity of the sentence to each summary of the four human summaries.

$$Score(s)=\max_i\{sim(s,S_i)\} \quad (1)$$

The strategies are of different purpose, **Average** strategy computes the similarity sum of all sentences in the four summaries, and the extracted sentences tend to be similar to each other. With this strategy, the quality of the extracted sentences is stably good, but the coverage rate of the document content may be low. **Maximum** strategy computes the maximum similarity to different summary so that the extracted sentences are more diversified. In the experiments, the maximum strategy performs better thus we choose the maximum strategy as the training data generation method in our system.

3.3 SVR-based sentence scoring

The features used in our SVR-based system are similar to those in our DUC 2006 system [Li 2006].

(1) Word-based Feature

$$V_{word}(s)=\frac{1}{|s|}\sum_{t_j\in s}\sum_{t_i\in q} same(t_i,t_j)$$

where *V* is the feature value, *q* is the topic description.

(2) Phrase-based Name Entity Feature

$$V_{entity}(s)=\frac{|entity(s)\cap entity(q)|}{|s|}$$

where $|entity(s)\cap entity(q)|$ is the number of the named entities in *s* and *q*.

(3) Semantic-based WordNet Feature

$$V_{wordnet}(s)=\frac{1}{|q|}\sum_{t_j\in s}\sum_{t_i\in q} similarity(t_i,t_j)$$

where the computation of $similarity(t_i,t_j)$ adopts the lesk similarity proposed in [Christiane 1998]

(4) Centroid Feature

$$V_{centroid}(s)=\sum_{t_j\in s} tfidf(t_j)$$

Where $tfidf(t_j)$ is the *tf-idf* score of t_j in the whole data set.

(5) Named Entity Number Feature

$$V_{entitynum}(s)=\frac{|entity(s)|}{|s|}$$

where $|entity(s)|$ is the number of named entities in *s*.

(6) Sentence Position Feature

$$V_{position}(s)=1-\frac{i-1}{n}$$

Firstly we use the data sets with human summaries to generate the training data. For each sentence in the document sets, its features are extracted and the corresponding score is computed with the maximum strategy mentioned in section 3.2 according to four human summaries. Then all sentences are converted to feature vectors with corresponding

scores, and the training data $D = \{< V_s, score(s) >\}$ is formed. Then the regression function $f : V_s \rightarrow score(s), s \in D'$ is learned by SVR model. To the sentences in the test data sets without human summaries, they can be scored by the feature vectors and the SVR regression function:

$$score(s) = f(V_s) \quad . \quad (6)$$

4 Evaluations

4.1 Test Data and Metrics

DUC 2007 provides 45 document sets for evaluation. Each document set includes a fixed number – 25 documents and its query. Each query contains a query title and a query narrative. A query title is usually a phrase which describes briefly the topic. A query narrative is usually composed of several factoid or definition questions, which need answers given in the summary. NIST assessors created 4 reference summary for each topic. All submitted systems are either manually or automatically evaluated, including linguistic quality, responsiveness, ROUGE-2, ROUGE-SU4, and Pyramid.

In our SVR-based system, we used 6 topics and their corresponding human summaries from DUC2006 to generate the training data.

4.2 Results

Among the 30 submitted systems, our system ranks about 5th in the evaluations of both ROUGE-2 and ROUGE-SU4 evaluations, 5th in the BE evaluation.

The result of our system and the best submitted system is listed in Table 1.

Evaluation	Our system	Best
ROUGE-2	0.11172	0.12448
ROUGE-SU4	0.16280	0.17711
Responsiveness	2.933	3.311
BE	0.06230	0.06632

Table 1 Performances in DUC2007

We also applied our system in DUC2005 and DUC2006, to show the advantage of the SVR system, a baseline system with the same features and manually-assigned weights is given also. In these systems, no sentence simplification is done and sentences with higher scores are selected without redundant removal method. The result is listed in table 2 and 3:

Submission	Rouge-2
Best submitted system	0.09558
SVR-based system	0.09057
Baseline system	0.08012

Figure2 Performances in DUC2006

Submission	Rouge-2
Best submitted system	0.07250
SVR-based system	0.07242
Baseline system	0.06310

Figure3 Performances in DUC2005

4.3 Analysis

Our system performed comparably in the automatic evaluations. The main reason is that appropriate lexical and syntactic features are adopted and the weight parameters are assigned suitably by SVR. The performance in responsiveness evaluations is not so good, and the reason may be that not much sentence simplification and reordering methods are introduced.

In the data set of DUC2005 and DUC2006, the SVR-based system also performed comparably well to the best submitted systems.

Compared to the baseline system with manually weights, the superiority of the SVR-based system is significant, which shows the SVR-based scoring method is robust and reliable.

5 Conclusions and Future Work

Our system adopts the traditional framework of extractive summarization. That is, some sentences are extracted from original text and reorganized into a summary with consideration of the query. The process of sentence extraction depends on various features. And we propose the state-of-art machine learning approach to combine the features. In future work, we will focus on how to generate better training data and design more elaborate features. Also, the relation between the training data generation method and the features will be studied.

Acknowledgements

This work is supported by IBM University Joint research program, NSFC program (60603093), and 973 National Basic Research Program of China (2004CB318102).

References

Christiane, F., 1998. WordNet: an Electronic Lexical Database. MIT Press.

Fisher, S. and Roark, B., 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. In Proceedings of DUC2006.

Gunn, S., 1998. Support Vector Machines for Classification and Regression, ISIS Tech. Rep., May 1998.

Li, S.J. Ouyang, Y., Sun, B, Guo, Z.L., 2006. Peking University at DUC 2006. In Proceedings of DUC2006.

Li, W.J., Li, W., Li, B.L., Chen, Q., Wu, M.L., 2005. The Hong Kong Polytechnic University at DUC2005. In Proceedings of DUC2005.

Lin.C.Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization, Barcelona. ACL.

Radev, D. R., Jing, H.Y., and Budzikowska, M., 2000. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, April 2000. Workshop on Summarization.

Radev, D.R., Otterbacher, J., Qi, H., Tam, D., 2003. MEAD ReDUCs: Michigan at DUC 2003. In Proceedings of DUC2003.

Vapnik, V.N., 1995. The Nature Of Statistical Learning Theory. New York: Springer. 1995

Vapnik, V.N., 1998. Statistical learning theory, John Wiley and Sons, New York..

Zhao, L., Huang, X.J., Wu, L.D., 2005. Fudan University at DUC 2005. In Proceedings of DUC 2005. (2005)