A SVM-Based Ensemble Approach to Multi-Document Summarization

Yllias Chali[1], Sadid A. Hasan[2], and Shafiq R. Joty[3]

[1] University of Lethbridge
Lethbridge, AB, Canada
chali@cs.uleth.ca
[2] University of Lethbridge
Lethbridge, AB, Canada
hasan@cs.uleth.ca
[3] University of British Columbia
Vancouver, BC, Canada
rjoty@cs.ubc.ca

Abstract. In this paper, we present a Support Vector Machine (SVM) based ensemble approach to combat the extractive multi-document summarization problem. Although SVM can have a good generalization ability, it may experience a performance degradation through wrong classifications. We use a committee of several SVMs, i.e. Cross-Validation Committees (CVC), to form an ensemble of classifiers where the strategy is to improve the performance by correcting errors of one classifier using the accurate output of others. The practicality and effectiveness of this technique is demonstrated using the experimental results.

Keywords: Multi-Document Summarization, Support Vector Machines, Ensemble, Cross-Validation Committees.

1 Introduction

Although SVMs achieve high generalization with training data of a very high dimension, it may degrade the classification performance by making some false predictions. To overcome this drawback, a SVM ensemble is clearly suitable. The main idea of an ensemble is to construct a set of SVM classifiers and then classify unseen data points by taking a majority voting scheme. In this paper, we concentrate on performing query relevant and extractive multi-document summarization task as it is defined by DUC-2007 (Document Understanding Conference). We use the cross-validation committees [1] approach of constructing an ensemble to inject differences into several SVM classifiers. We then compare the ensemble system's performance with a single SVM system and a baseline system. The evaluation result shows the efficiency of ensemble approaches in this problem domain.

Y. Gao and N. Japkowicz (Eds.): Canadian AI 2009, LNAI 5549, pp. 199–202, 2009.

1. Divide whole training data set D into $v-$fractions $d_1, \ldots, d_v$
2. Leave one fraction d_k and train classifier c_k with the rest of the data $(D - d_k)$
3. Build a committee from the classifiers using a simple averaging procedure.

Algorithm 1: Cross-Validation Committees Method

2 Cross-Validation Committees (CVC)

In this research, we use the *cross-validation committees* approach to build a SVM ensemble. CVC is a training set sampling method where the strategy is to construct the training sets by leaving out disjoint subsets of the training data [2]. The typical algorithm of the CVC approach [1] is presented in Algorithm 1.

3 Experimental Setup

3.1 Problem Definition

The task at DUC-2007 is defined as: *"Given a complex question (topic description) and a collection of relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic"*. We consider this task and employ a SVM-based ensemble approach to generate topic-oriented 250-word extract summaries for 25 topics of DUC-2007 document collection using DUC-2006 data set for training.

3.2 Data Labeling

We use an automatic labeling method to label our large data sets (DUC-2006 data) using ROUGE [3]. For each sentence in a topic, we calculate its ROUGE score corresponding to the given abstract summaries from DUC-2006. Then based on these scores, we choose the top N sentences to have the label $+1$ (summary) and the rest to have the label -1 (non-summary).

3.3 Feature Extraction

Each of the document-sentences is represented as a vector of feature-values. We extract several query-related features and some other important features from each sentence. The features we use are: n-gram overlap, Longest Common Subsequence (LCS), Weighted LCS (WLCS), skip-bigram, exact word overlap, synonym overlap, hypernym/hyponym overlap, gloss overlap, Basic Element (BE) overlap, syntactic tree similarity measure, position of sentences, length of sentences, Named Entity (NE), cue word match and title match [4,5,6].

3.4 SVM Ensemble

We use the cross-validation committees (CVC) approach to build a SVM ensemble. We divide the training data set (DUC-2006 data) into 4 equal-sized fractions and according to the CVC algorithm, each time we leave separate 25%

data out and use the rest 75% data for training. Thus, we generate 4 different SVM models. Next, we present the test data (25 topics of DUC-2007 data) before each of the generated SVM models which produces individual predictions to those unseen data. Then, we create the SVM ensemble by combining their predictions by simple weighted averaging. We increment a particular classifier's decision value, the normalized distance from the hyperplane to a sample by 1 (giving more weight) if it predicts a sentence as positive and decrement by 1 (imposing penalty), if the case is opposite. The resulting prediction values are used later for ranking the sentences. During training steps, we use the third-order polynomial kernel keeping the value of the trade-off parameter C as default. For our SVM experiments, we use SVM^{light}[1] package [7]. We perform the training experiments in WestGrid[2].

3.5 Sentence Ranking

In the Multi-Document Summarization task at DUC-2007, the word limit was 250 words. To meet this criteria, we rank sentences in a document set, then select the top N sentences. We use the combined decision values of the 4 different SVM classifiers to rank the sentences. Then, we choose the top N sentences until the summary length is reached.

3.6 Evaluation Results

In DUC-2007, each topic and its document cluster were given to 4 different NIST assessors, including the developer of the topic. The assessor created a 250-word summary of the document cluster that satisfies the information need expressed in the topic statement. These multiple "reference summaries" are used in the evaluation of our summary content. We evaluate the system generated summaries using the automatic evaluation toolkit ROUGE [3]. We generate summaries for the first 25 topics of the DUC-2007 data and tested our SVM ensemble's performance with a single SVM system and a baseline system.

Table 1. ROUGE measures for SVM Ensemble

Measures	R-1	R-L	R-W	R-SU
Precision	0.4081	0.3359	0.1791	0.1621
Recall	0.3705	0.3051	0.0877	0.1334
F-score	0.3883	0.3197	0.1177	0.1463

In Table 1, we present the ROUGE scores of our SVM ensemble system in terms of Precision, Recall and F-scores. We show the four important ROUGE metrics in the results: ROUGE-1 (unigram), ROUGE-L (LCS), ROUGE-W (weighted LCS with weight=1.2) and ROUGE-SU (skip bi-gram). Table 2 shows

[1] http://svmlight.joachims.org/

[2] http://westgrid.ca/

Table 2. ROUGE F-Scores for Diff. Systems

Systems	R-1	R-L	R-W	R-SU
Baseline	0.3347	0.3107	0.1138	0.1127
Single	0.3708	0.3035	0.1113	0.1329
Ensemble	0.3883	0.3197	0.1177	0.1463

the F-scores for a single SVM system, one baseline system and the SVM ensemble system. The single SVM system is trained on the full data set of DUC-2006 and the baseline system's approach is to select the lead sentences (up to 250 words) from each topic's document set. Table 2 clearly suggests that the SVM ensemble system outperforms both the single SVM system and the baseline system with a decent margin in all the ROUGE measures.

4 Conclusion

In this paper, we applied an effective supervised framework to confront query-focused text summarization problem based on Support Vector Machine (SVM) ensemble. The experimental results on the 25 document sets of DUC-2007 show that the SVM ensemble technique outperforms the conventional single SVM system and the baseline system which proves the effectiveness of the ensemble approach in the domain of supervised text summarization.

References

1. Parmanto, B., Munro, P.W., Doyle, H.R.: Improving committee diagnosis with resampling techniques. In: Advances in Neural Information Processing Systems, vol. 8, pp. 882–888 (1996)
2. Dietterich, T.G.: Ensemble methods in machine learning, pp. 1–15. Springer, Heidelberg (2000)
3. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of Association for Computational Linguistics, Barcelona, Spain, pp. 74–81 (2004)
4. Chali, Y., Joty, S.R.: Selecting sentences for answering complex questions. In: Proceedings of EMNLP (2008)
5. Edmundson, H.P.: New methods in automatic extracting. Journal of the ACM 16(2), 264–285 (1969)
6. Sekine, S., Nobata, C.A.: Sentence extraction with information extraction technique. In: Proceedings of the Document Understanding Conference (2001)
7. Joachims, T.: Making large-Scale SVM Learning Practical. In: Advances in Kernel Methods - Support Vector Learning (1999)