Supervised Approaches to Complex Question Answering

Yllias Chali
University of Lethbridge
Lethbridge, AB, Canada
chali@cs.uleth.ca

Sadid A. Hasan
University of Lethbridge
Lethbridge, AB, Canada
hasan@cs.uleth.ca

Shafiq R. Joty
University of British Columbia
Vancouver, BC, Canada
rjoty@cs.ubc.ca

Abstract

Unlike simple questions, complex questions cannot be answered easily as they often require inferencing and synthesizing information from multiple documents. Hence, this task is accomplished by the query-focused multi-document summarization systems. In this paper, we consider the problem definition given at the DUC-2007 main task and experiment with different supervised learning techniques to confront the complex question answering problem. As representative supervised methods, we use Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), and MaxEnt (Maximum Entropy). We also experiment with an ensemble based approach combining the individual decisions of these classifiers. We use DUC-2006 data to train our systems, whereas 25 topics of DUC-2007 data set are used as test data. The evaluation results reveal the effectiveness of these approaches in the problem domain.

1 Introduction

In real-world complex question answering domain, a question cannot be answered by simply stating a name, date, quantity, etc. For instance, the question: *"Describe steps taken and worldwide reaction prior to the introduction of the Euro on January 1, 1999. Include predictions and expectations reported in the press."* require inferencing and synthesizing information from multiple documents. The information synthesis in NLP can be seen as a kind of topic-oriented, informative multi-document summarization. Here, the main goal is to produce a single text as a compressed version of a set of documents with a minimum loss of relevant information (Mani, 2001).

Supervised classifiers are typically trained on data pairs, defined by feature vectors and corresponding class labels. On the other hand, unsupervised approaches rely on heuristic rules that are pretty difficult to generalize (Shen et al., 2007). Supervised extractive summarization can often be regarded as a two-class classification problem treating summary sentences as positive samples and non-summary sentences as negative samples. Given the features of a sentence, a machine-learning based classification model can judge how likely the sentence is important to be in the summary (Wong et al., 2008). Hidden Markov Model (HMM) requires a careful feature selection to achieve high accuracy bearing fewer assumptions of independence (Conroy and O'Leary, 2001). The statistical technique such as Maximum Entropy (MaxEnt) works in a way assuming nothing about the information of which it has no prior knowledge (Ferrier, 2001). Conditional Random Fields (CRF) tend to carry out the summarization task in a discriminative manner (Shen et al., 2007). On the other hand, Support Vector Machines (SVM) take a strategy that maximizes the margin between critical samples and the separating hyperplane for efficient classification (Vapnik, 1998). At present, one of the most active research areas in supervised learning is the methods for constructing good ensemble of classifiers which needs the sub-classifiers to differentiate greatly (Qi and Huang, 2007).

In this paper, we consider the complex question answering problem defined in the DUC-2007 main task[1]. We focus on an extractive approach of summarization to answer complex questions where a subset of the sentences in the original documents are chosen. We accomplish the task by applying the supervised approaches: SVM, HMM, CRF and MaxEnt. We present an extensive experimental

[1] http://www-nlpir.nist.gov/projects/duc/duc2007/

comparison to evaluate their performances. We also experiment with a heterogeneous ensemble approach by taking a weighted voting of the individual classifier's predictions. A further analysis shows the effectiveness of this ensemble approach. The remainder of the paper is organized as follows: Section 2 describes the related work, Section 3 presents a brief literature review on the SVM, HMM, CRF, MaxEnt and ensemble methods' theory, Section 4 discusses the experimental settings and shows the evaluation results, and finally, Section 5 concludes the paper by cuing some future works.

2 Related Work

Single document summarization systems using SVMs demonstrated good performance for both Japanese (Hirao et al., 2002a) and English documents (Hirao et al., 2002b). Hirao et al. (2003) showed effectiveness of their multiple document summarization system employing SVMs for sentence extraction. Conroy and O'Leary (2001) used the HMM method denoting two kinds of states, where one kind corresponds to the summary states and the other corresponds to the non-summary states. Given a new cluster of documents, they calculated the probability of a sentence to be in a summary state. Finally, the trained model can be used to select the most likely summary sentences. The motivation of applying CRF in text summarization came from observations on how humans summarize a document by posing the problem as a sequence labeling problem. Shen et al. (2007) showed the effectiveness of CRF by applying it to a generic single-document extraction task. Ferrier (2001) applied the MaxEnt technique to single document text summarization and found that the maximum entropy classifier produces better results than the naive Bayes technique. Ensemble techniques are in the focus of the researchers over the years as different methods for constructing good ensembles are developed (Dietterich, 2000). A boosting based support vector ensemble was proposed to achieve good performance in summarizing text from a Vietnamese corpus (Nguyen et al., 2005).

3 Theory

3.1 Support Vector Machines (SVM)

SVM is a powerful methodology for solving machine learning problems introduced by Cortes and Vapnik (1995) based on the Structural Risk Minimization principle. Training samples each of which belongs either to positive or negative class can be denoted by:

$$(x_1,y_1),\ldots,(x_u,y_u),\ x_j \in R^n,\ y_j \in \{+1,-1\}$$

Here, x_j is a feature vector of the j-th sample represented by an n dimensional vector; y_j is its class label. u is the number of the given training samples. SVM separates positive and negative examples by a hyperplane defined by:

$$w \cdot x + b = 0,\ w \in R^n, b \in R \tag{1}$$

where "$\cdot$" stands for the inner product. The examples on $w \cdot x + b = \pm 1$ are called the Support Vectors which represent both positive or negative examples. The hyperplane must satisfy the following constraints:

$$y_i \left(w \cdot x_j + b\right) - 1 \geq 0$$

Hence, the size of the margin is $2/||w||$. We assume the following objective function to maximize the margin:

$$Minimize_{w,b,\xi}\ J(w,\xi) = \frac{1}{2}||w||^2 + C\sum_{j=1}^{u}\xi_j \tag{2}$$

$$s.t.\ y_j \left(w \cdot x_j + b\right) - \left(1 - \xi_j\right) \geq 0$$

Here, $||w||/2$ indicates the size of the margin, $\sum_{j=1}^{u}\xi_j$ indicates the penalty for misclassification, and C is the cost parameter that determines the trade-off for these two arguments. The decision function depends only on support vectors $(\lambda_i \neq 0)$. SVMs can handle non-linear decision surfaces with kernel function $K(x_i \cdot x)$. Therefore, the decision function can be rewritten as follows:

$$g(x) = \sum_{i=1}^{u} \lambda_i y_i K(x_i,x) + b \tag{3}$$

In this paper, we use polynomial kernel functions, which have been found to be very effective in the study of other tasks in natural language processing (Joachims, 1998):

$$K(x,y) = (x \cdot y + 1)^d \tag{4}$$

3.2 Hidden Markov Models (HMM)

HMMs are a form of generative model, that assign a joint probability $p(x,y)$ to pairs of observation and label sequences, x and y respectively (Wallach, 2002). Formally, an HMM is fully defined by

- A finite set of states S.
- A finite output alphabet X.
- A conditional distribution $P\left(s^{'}|s\right)$ representing the probability of moving from state s to state $s^{'}$, where $s,s^{'} \in S$

- An observation probability distribution $P(x|s)$ representing the probability of emitting observation x when in state s, where $x \in X$ and $s \in S$.

- An initial state distribution $P(s), s \in S$.

A HMM may be represented as a directed graph G with nodes S_t and X_t representing the state of the HMM (or label) at time t and the observation at time t, respectively. The probability of the state at time t depends only on the state at time $t-1$. Similarly, the observation generated at time t only depends on the state of the model at time t.

Finding the optimal state sequence given the observation sequence and the model is most efficiently performed using a dynamic programming technique known as Viterbi alignment (Rabiner, 1989).

3.3 Conditional Random Fields (CRF)

To reap the benefits of using a conditional probabilistic framework for labeling sequential data and simultaneously overcome the label bias problem, Lafferty et al. (2001) introduced CRFs.

CRF allows the specification of a single joint probability distribution over the entire label sequence given the observation sequence, rather than defining per-state distributions over the next states given the current state. Given an observation sequence (sentence sequence here) $X = (x_1, \cdots, x_T)$ and the corresponding state sequence $Y = (y_1, \cdots, y_T)$, the probability of Y conditioned on X defined in CRFs, $P(Y|X)$, is as follows:

$$\frac{1}{Z_X} exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l} \mu_l g_l(y_i, X)\right) \tag{5}$$

where Z_X is the normalization constant that makes the probability of all state sequences sum to one; $f_k(y_{i-1}, y_i, X)$ is an arbitrary feature function over the entire observation sequence and the states at positions i and $i-1$ while $g_l(y_i, X)$ is a feature function of state at position i and the observation sequence; λ_k and μ_l are the weights learned for the feature functions f_k and g_l, reflecting the confidence of the feature functions (Shen et al., 2007).

3.4 Maximum Entropy (MaxEnt)

The main principle of the MaxEnt method is to model all that is known and assume nothing about that which is unknown. In other words, given a collection of facts, the model must be consistent with all the facts, but otherwise act as uniformly as possible (Berger et al., 1996). MaxEnt models can be termed as multinomial logistic regression if they are to classify the observations into more than two classes (Jurafsky and Martin, 2008). However, in this research, we used the MaxEnt model to classify the sentences into two classes: summary or non-summary. The parametric form for the maximum entropy model is as follows (Nigam et al., 1999):

$$P(c|s) = \frac{1}{Z(s)} exp\left(\sum_i \lambda_i f_i\right) \tag{6}$$

$$Z(s) = \sum_c exp\left(\sum_i \lambda_i f_i\right) \tag{7}$$

Here, c is the class label and s is the item we are interested in labeling that is the sentences here. Z is the normalization factor that is just used to make the exponential into a true probability. Each f_i is a feature with the associated weight λ_i which can be determined by numerical optimization techniques in absence of a closed form solution.

3.5 Ensemble Method

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by voting) to classify new examples. The main idea is to work together for the same goal and while doing this, minimize errors of each other. Many methods for constructing ensembles have been developed over the years which consider Bayesian voting, manipulation of the training examples, input features and output targets, injecting randomness and so on (Dietterich, 2000). Differences among the classifiers can be realized by using separate training samples on the same learning method (Qi and Huang, 2007). However, in this paper we experiment with an ensemble method that uses the same training set on different learning methods. Thus, we consider it as a heterogeneous ensemble that joins the above four classifiers which are somehow different in accomplishing the classification task. We combine the individual decisions of the classifiers by taking a weighted voting and then the combined decision values are used to label the unseen data set.

4 Experiments

4.1 Task Description

The problem definition at DUC-2007 was: *"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 use the DUC-2006 data to train all the systems. Then, we produce extract summaries for the 25 documents of the DUC-2007 data according to the task description.

4.2 Training Data Preparation

For supervised learning techniques, huge amount of annotated or labeled data sets are obviously required as a precondition. When humans are employed in the process, producing such a large labeled corpora becomes time consuming and expensive. Hence, we implement an automatic labeling methodology to label our large data sets (DUC-2006 data) using ROUGE (Lin, 2004). 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 sentences) and the rest to have the label -1 (non-summary sentences). We obtain a balanced set of training samples by annotating 50% sentences of each document set as positive and the rest as negative.

4.3 Features

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 (Edmundson, 1969).

4.4 Settings

For SVM we use second order polynomial kernel for the ROUGE and ESSK labeled training. For the BE, syntactic, and semantic labeled training third order polynomial kernel is used. The use of kernel is based on the accuracy we achieved during training. We apply 3-fold cross validation with randomized local-grid search for estimating the value of the trade-off parameter C. We try the value of C in 2^i following heuristics, where $i \in \{-5, -4, \cdots, 4, 5\}$ and set C as the best performed value 0.125 for second order polynomial kernel and default value is used for third order kernel. We use SVM^{light} package[2] for training and testing in this research. In case of HMM, we apply the Maximum Likelihood Estimation (MLE) technique by frequency counts with add-one smoothing to estimate the three HMM parameters: initial state probabilities, transition probabilities and emission probabilities. We use Dr. Dekang Lin's HMM package[3] to generate the most probable label sequence given the model parameters and the observation sequence (unlabeled DUC-2007 test data). We use MALLET-0.4 NLP toolkit[4] to implement the CRF. We formulate our problem in terms of MALLET's SimpleTagger class which is a command line interface to the MALLET CRF class. We modify the SimpleTagger class in order to include the provision for producing corresponding posterior probabilities of the predicted labels which are used later for ranking sentences. We build the MaxEnt system using Dr. Dekang Lin's MaxEnt package[5]. To define the exponential prior of the λ values in MaxEnt models, an extra parameter α is used in the package during training. We keep the value of α as default.

4.5 Sentence Ranking

Forcing summaries to obey a certain length constraint is a common set-up in summarization, as in the multi-document summarization task at DUC-2007, the word limit was 250 words. A simple solution to combat this problem is to rank the sentences in a document, then select the top N sentences. In SVM systems, we use $g(x)$, the normalized distance from the hyperplane to x to rank the sentences. Then, we choose the top N sentences until the summary length is reached. For HMM systems, we use the Maximal Marginal Relevance (MMR) based method to rank the sentences (Carbonell et al., 1997). In CRF systems, we generate posterior probabilities corresponding to each predicted label in the label sequence to

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

[3] http://www.cs.ualberta.ca/ l̃indek/hmm.htm

[4] http://mallet.cs.umass.edu/

[5] http://www.cs.ualberta.ca/l̃indek/downloads.htm

Systems	R-1	R-2	R-SU
Base	0.3347	0.0649	0.1127
Ensemble	0.3950	0.0885	0.1530
HMM	0.3945	0.0898	0.1499
MaxEnt	0.3938	0.0871	0.1490
CRF	0.3725	0.0742	0.1329
SVM	0.3708	0.0672	0.1328

Table 1: ROUGE F-Scores for Diff. Systems

measure the confidence of each sentence for summary inclusion. Similarly for MaxEnt, the corresponding probability values of the predicted labels were used to rank the sentences. The combined weighted votes of all the classifiers are used to rank the sentences in the ensemble approach.

4.6 Evaluation

The multiple "reference summaries" given by DUC-2007 are used in the evaluation of our summary content. We evaluate the system generated summaries using the automatic evaluation toolkit ROUGE (Lin, 2004). We report the three widely adopted important ROUGE metrics in the results: ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-SU (skip bi-gram). We generate summaries for the 25 topics of the DUC-2007 data set and analyze all the supervised systems' performance with one baseline system of DUC-2007. The baseline system generates summaries by returning all the leading sentences (up to 250 words) in the $\langle TEXT \rangle$ field of the most recent document(s). Table 1 shows the ROUGE F-measures for SVM, HMM, CRF, MaxEnt and the baseline system.

The table clearly suggests that the most of the supervised systems outperform the baseline system with a high margin and the ensemble system is typically the best performer. This is because, the individual classifier decisions are combined together to judge the sentence labels correctly. Comparison of the four supervised methods individually, reveals the fact that HMM is performing the best and MaxEnt, CRF and SVM are the next in the performance ranking, respectively. The reason for this is, HMM treats the task as a sequence labeling problem and the scores are improved further as we used the MMR method for sentence ranking which is a proved effective way of reducing the redundancy. We can also understand that the MaxEnt method is performing better than the SVM, yielding the matter that high-order kernels do not suit well for the problem domain.

Confidence Intervals In table 2, We show the 95% confidence intervals of the important evaluation metrics (for F-measures of ROUGE-1 and ROUGE-SU) for the baseline system, and our supervised systems to report the significance for doing meaningful comparison. We use the ROUGE tool for this purpose. ROUGE uses a randomized method named bootstrap resampling to compute the confidence interval. We use 1000 sampling points in the bootstrap resampling.

Systems	ROUGE-1	ROUGE-SU
Baseline	0.3266 - 0.3423	0.1084 - 0.1167
Ensemble	0.3739 - 0.4127	0.1386 - 0.1663
HMM	0.3745 - 0.4107	0.1363 - 0.1613
MaxEnt	0.3763 - 0.4109	0.1367 - 0.1618
CRF	0.3553 - 0.3892	0.1205 - 0.1457
SVM	0.3558 - 0.3865	0.1217 - 0.1444

Table 2: 95% Confidence Intervals

If we analyze table 2, we find that all the supervised systems perform significantly better than the baseline system. On the other hand, for the ensemble system, HMM, and MaxEnt, we get a high overlap in terms of all the ROUGE measures.

5 Conclusion and Future Work

In the work reported in this paper, we have performed an extensive experimental comparison of different supervised machine learning techniques in confronting the complex question answering problem defined in the DUC-2007 main task. Experimental results on the 25 document sets of DUC-2007 show that supervised approaches outperform the conventional systems. We also find that the heterogeneous ensemble method works well for this domain whereas the high-order kernels are not very suitable since SVM performs badly. Our experiments reveal that HMM works best for the given task, on the other hand, MaxEnt and CRF show decent performance. In future, We plan to apply Integer Linear Programming (ILP) approaches in order to see whether it works well or not in this problem domain.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Jaime Carbonell, Yibing Geng, and Jade Goldstein. 1997. Automated query-relevant summarization and diversity-based reranking. In *IJCAI-97 Workshop on AI in Digital Libraries*, pages 12–19. Japan.

John M. Conroy and Dianne P. O'Leary. 2001. Text summarization via hidden markov models. In *SIGIR*, pages 406–407.

Corinna Cortes and Vladimir N. Vapnik. 1995. Support Vector Networks. *Machine Learning*, 20:273–297.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. pages 1–15. Springer-Verlag.

Harold P. Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Louisa Ferrier. 2001. *A Maximum Entropy Approach to Text Summarization*. M.Sc. thesis, School of Artificial Intelligence, Division of Informatics, University of Edinburgh.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002a. Extracting important sentences with support vector machines. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7.

Tsutomu Hirao, Yutaka Sasaki, Hideki Isozaki, and Eisaku Maeda. 2002b. NTT's text summarization system for DUC2002. In *Proceedings of the Document Understanding Conference*, pages 104–107.

Tsutomu Hirao, Jun Suzuki, Hideki Isozaki, and Eisaku Maeda. 2003. NTT's multiple document summarization system for DUC2003. In *Proceedings of the Document Understanding Conference*.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML)*.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

John Lafferty, Andrew K. McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Chin-Yew Lin. 2004. 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*, pages 74–81. Barcelona, Spain.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Co, Amsterdam/Philadelphia.

Minh Le Nguyen, Akira Shimazu, Xuan Hieu Phan, Tu Bao Ho, and Susumu Horiguchi. 2005. Sentence extraction with support vector machine ensemble. In *First World Congress of the International Federation for Systems Research (IFSR'05), Symposium on Data/Text Mining from Large Databases*. Kobe.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classication. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*.

Hengnian Qi and Meili Huang. 2007. Research on SVM ensemble and its application to remote sensing classification. In *ISKE'07*.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–285.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document Summarization Using Conditional Random Fields. In *IJCAI*, pages 2862–2867.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley Interscience.

Hanna Wallach. 2002. *Efficient Training of Conditional Random Fields*. M.Sc. thesis, Division of Informatics, University of Edinburgh.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 985–992. Manchester.