Complex Question Answering: Homogeneous or Heterogeneous, Which Ensemble Is Better?

Yllias Chali[1], Sadid A. Hasan[2], and Mustapha Mojahid[3]

[1] University of Lethbridge, Lethbridge, AB, Canada
chali@cs.uleth.ca
[2] Philips Research North America, Briarcliff Manor, NY, USA
sadid.hasan@philips.com
[3] IRIT, Toulouse, France
mustapha.mojahid@irit.fr

Abstract. This paper applies homogeneous and heterogeneous ensembles to perform the complex question answering task. For the homogeneous ensemble, we employ Support Vector Machines (SVM) as the learning algorithm and use a Cross-Validation Committees (CVC) approach to form several base models. We use SVM, Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Maximum Entropy (MaxEnt) techniques to build different base models for the heterogeneous ensemble. Experimental analyses demonstrate that both ensemble methods outperform conventional systems and heterogeneous ensemble is better.

Keywords: Complex Question Answering, Homogeneous Ensemble, Heterogeneous Ensemble.

1 Introduction

This paper is concerned with the application of ensemble based methods for the complex question answering task. We use query-focused supervised extractive multi-document summarization technique for this purpose [1–3]. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [4]. Generation of ensembles can be categorized into two types: 1) homogeneous, if the base learning model is built from the same learning algorithm, and 2) heterogeneous, where different learning algorithms are combined to generate the base learning models [12]. 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 [2, 6, 14]. The next section presents our experimental design and evaluation framework, and then we conclude the paper with future directions.

E. Métais, M. Roche, and M. Teisseire (Eds.): NLDB 2014, LNCS 8455, pp. 160–163, 2014.

2 Experimental Settings and Evaluation

We use the query-focused summarization task proposed in DUC[1] (2005-2007) to simulate our complex question answering experiments. We use the DUC-2006 data to train all the systems and then produce extract summaries for the DUC-2007 data. Supervised classifiers are typically trained on data pairs, defined by feature vectors and corresponding class labels. We use an automatic labeling approach to annotate the training data using ROUGE [1, 3, 9]. From each sentence of the training (and testing) data, we extract different query-related features and importance-oriented features such as: 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) match, cue word match and title match [1, 3, 5, 13].

For homogeneous ensemble, we divide the training data into 4 equal-sized fractions. Then, according to the CVC algorithm [2, 4, 11, 12], 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 feed the test data to each of the generated SVM models which produces individual predictions (decision scores along with a label $+1$ or -1). The decision scores are the normalized distance from the separating hyperplane to each sample. To create the SVM ensemble, we combine the predictions by simple weighted averaging. We increment a particular classifier's decision value 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 for the SVM keeping the value of the trade-off parameter C as default. For our SVM experiments, we use the SVM^{light} package[2] [7]. The individual classifier settings for the heterogeneous ensemble formation are as follows. For SVM, we use the same setup as homogeneous ensemble. We implement the HMM model by Lin's HMM package[3]. We use the MALLET NLP toolkit [10] to implement the CRF. We modify its SimpleTagger class in order to include the provision for producing corresponding posterior probabilities of the predicted labels which were used later to rank the sentences. We build the MaxEnt system using Lin's MaxEnt package[4]. We combine the decision values of the four different classifiers by a weighted voting to build an ensemble. We impose a positive weight (ranging from 1 to 5 depending on the individual classifier's performance, more weight if it is declared positive by a better performer based on scores) to each positively classified sentence. We take no action for the negatively classified sentences so that they could fall back during ranking. The combined weighted votes of all the classifiers are used to rank the sentences to produce 250-word summaries [1].

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

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

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

[4] http://www.cs.ualberta.ca/~lindek/downloads.htm

We consider the multiple "reference summaries" of DUC-2007 to automatically evaluate our summaries using the ROUGE toolkit [9]. We compare the ensemble systems' performance with a baseline system. The baseline system's approach is to select the lead sentences (up to 250 words) from each topic's document set. In table 1, we present the ROUGE F-scores of different systems. We can see that the homogeneous ensemble improves the ROUGE-1, ROUGE-2 and ROUGE-SU scores over the baseline system by 16.2%, 26.6% and 30.3% respectively. The heterogeneous ensemble improves the ROUGE-1, ROUGE-2 and ROUGE-SU scores over the baseline system by 18.3%, 37.5% and 36.6% and over the homogeneous system by 1.80%, 8.64% and 4.79% respectively. Three native English speaking university graduate students judged[5] all the system generated summaries for readability (fluency) and overall responsiveness according to the TAC 2010 summary evaluation guidelines[6]. Table 2 presents the average readability and overall responsive scores of all the systems. The results again show that the ensemble systems perform better than the baseline system and heterogeneous ensemble performs the best in terms of overall responsiveness.

Table 1. ROUGE F-Scores for different systems

Systems	ROUGE-1	ROUGE-2	ROUGE-SU
Baseline	0.334	0.064	0.112
Homogeneous	0.388	0.081	0.146
Heterogeneous	0.395	0.088	0.153

Table 2. Readability and overall responsiveness scores for all systems

Systems	Readability	Overall Responsiveness
Baseline	4.24	1.80
Homogeneous	3.41	3.30
Heterogeneous	3.85	3.63

3 Conclusion and Future Work

In this paper, we presented the use of two ensemble methods: homogeneous and heterogeneous to perform the complex question answering task. Our experiments suggested the following: (a) ensemble methods outperform the conventional systems, and (b) heterogeneous ensemble performs the best for this problem. Future work is foreseen to use different learning algorithms for homogeneous ensemble and to improve the base classifiers' performance for both ensemble methods.

[5] The inter-annotator agreement of Fleiss' $\kappa = 0.63$ is computed for the three judges indicating a substantial degree of agreement [8].

[6] `http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html`

Acknowledgments. The research reported in this paper was conducted at the University of Lethbridge and supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada-discovery grant, and the University of Lethbridge.

References

1. Chali, Y., Hasan, S.A.: Query-focused Multi-document Summarization: Automatic Data Annotations and Supervised Learning Approaches. Journal of Natural Language Engineering 18(1), 109–145 (2012)
2. Chali, Y., Hasan, S.A., Joty, S.R.: A SVM-Based Ensemble Approach to Multi-Document Summarization. In: Gao, Y., Japkowicz, N. (eds.) Canadian AI 2009. LNCS (LNAI), vol. 5549, pp. 199–202. Springer, Heidelberg (2009)
3. Chali, Y., Hasan, S.A., Joty, S.R.: Do Automatic Annotation Techniques Have Any Impact on Supervised Complex Question Answering? In: Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2009), Suntec, Singapore, pp. 329–332 (2009)
4. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
5. Edmundson, H.P.: New methods in automatic extracting. Journal of the ACM 16(2), 264–285 (1969)
6. Gashler, M., Giraud-Carrier, C.G., Martinez, T.R.: Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. In: ICMLA, pp. 900–905 (2008)
7. Joachims, T.: Making large-Scale SVM Learning Practical. In: Advances in Kernel Methods - Support Vector Learning (1999)
8. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1), 159–174 (1977)
9. 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)
10. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002)
11. 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)
12. Rooney, N., Patterson, D.W., Anand, S.S., Tsymbal, A.: Random subspacing for regression ensembles. In: FLAIRS Conference (2004)
13. Sekine, S., Nobata, C.A.: Sentence extraction with information extraction technique. In: Proceedings of the Document Understanding Conference (2001)
14. Silva, C., Ribeiro, B.: Rare class text categorization with SVM ensemble. Journal of Electrotechnical Review (Przeglad Elektrotechniczny) 1, 28–31 (2006)