Domain Adaptation for Summarizing Conversations

Oana Sandu, Giuseppe Carenini, Gabriel Murray, and Raymond Ng

Introduction

Much of natural language processing research involves developing a computational method to perform some task, and testing it on some data set by comparing its predictions on a data set with gold-standard labels assigned by human annotators. However, it is difficult to achieve good performance on a domain lacking labeled data. The aim of domain adaptation is to remedy this problem by using readily available labeled data from a different source domain along with potentially a small amount of labeled data and a larger amount of unlabeled data in the target domain. We are interested in using domain adaptation to improve extractive summarization of conversations. Since very few email corpora have been annotated for summarization, we attempt to leverage the labeled data available in the multi-party meetings domain for the summarization of email threads. We present here our ongoing work in investigating promising domain adaptation methods for our task. We apply several approaches to supervised domain adaptation, which require in-domain labeled data, and semi-supervised domain adaptation, which use unlabeled in-domain data. We also the compare their performance when the data is represented with a small set of conversational features versus a large set of lexical features. From our results, we conclude that both supervised and semi-supervised domain adaptation methods can improve summarization when using the lexical features, although no method improves over the baseline with conversational features.

Domain Adaptation Methods

We first consider a common scenario requiring supervised domain adaptation, in which a large amount of labeled data is available in the source domain, and a small amount of labels exists for the target. The baseline *indomain* is to just train on in-domain target data, and several methods have been proposed to try to improve performance by also using out-of-domain data. As some simple ones, *merge* trains on the combination of the two labeled data sets, *ensemble* takes the linear combination of the predictions of a classifier trained on the source data and another trained on the target data, and *pred* trains a classifier on source data, runs it on the target data, and uses its predicted label probability as an extra input feature to the final target-trained model (1). A more sophisticated supervised domain adaptation method is *featcopy* introduced by Daume (2), which takes the existing features and extends the feature space by making a general, a source-specific, and a target-specific version of each feature.

Semi-supervised domain adaptation methods have the advantage of using labeled out-of-domain data and unlabeled in-domain data, without the requirement of human-labeled training data in the target domain. In self-training, a training set is used composed initially of the labeled source data, and repeatedly incremented with the highest confidence predictions on unlabeled target data; in Blitzer's structural correspondence learning (SCL) (3), a subset of informative features is identified using labeled source data and then correlated with the rest of the features using unlabeled data to identify correspondences between the two domains. We implement these methods as *selftrain*, Blitzer's original *scl* which uses all the initial features with the addition of augmented features, and *sclsmall*, a modification of SCL suggested in (4) that only uses the new features. The baseline we use for the semi-supervised methods is *transfer*, which trains on the labeled source data only.

Experiments

As source data, we use the AMI meeting corpus (5), for which 115000 dialogue acts were annotated for extractive summarization. As labeled target training and test data, we use the BC3 corpus (6), a set of 40 email threads from with annotations similar to AMI. As unlabeled email data, we use 39 threads from the Enron corpus (7). We extract two sets of features from the data: one is the set of 24 conversational features introduced by Murray in (8), and the second is the large set of lexical features proposed in (9), of which we select the top 10,000 by mutual information. We run all the domain adaptation methods described above on this data, using the conversational features in one experiment and the lexical features in another experiment. In each of the experiments, we do 5-fold cross validation on BC3, repeated three

times, and for each run we subsample 10,000 sentences from AMI, as it is a much larger set of data than BC3 and Enron. In one run, all methods are given the same training data and tested on the same testing data. To measure performance, we run a logistic regression classifier and compute the resulting area under the receiver operator curve (auROC). We also perform pairwise t-tests to test for significant differences between the auROCs obtained for different methods, and report significance for $p < 0.005$.

Results

Table 1. Comparison of the performance of domain adaptation methods with the two feature sets

method	indomain	merge	ensemble	featcopy	pred	transfer	selftrain	scl	sclsmall
using conversational features									
auROC	0.838	0.747	0.751	0.839	0.838	0.677	0.678	0.663	0.646
time(s)	0.79	2.42	2.64	8.44	5.38	2.08	100.2	52.85	66.74
using lexical features									
auROC	0.623	0.638	0.667	0.615	0.625	0.636	0.636	0.651	0.742
time(s)	4.87	13.64	13.77	78.63	30.99	9.73	448.8	813.7	828.3

For each method, table 1 presents the average auROC value and average processing time it achieved using each of the two sets of features. The best performance overall is achieved by *indomain, merge*, and *ensemble* when using conversational features. With lexical features, *merge* and *ensemble* improve significantly on the supervised baseline of *indomain*, whereas no domain adaptation method beats *indomain* with conversational features. Also, no semi-supervised method improves significantly over *transfer* when using the conversational features. The lexical features turn out to be more useful in the semi-supervised scenario, with *scl* and *sclsmall* both improving significantly over *transfer*. *sclsmall* achieves the best auROC of all the methods in the experiment using lexical features. As for time, using the small set of conversational features is also more efficient than the lexical features, and the semi-supervised methods are markedly more expensive than the supervised methods since they are iterative.

Conclusion and Future Work

Our findings indicate that for summarizing of conversations, with the set of sophisticated conversational features, it is better to train in-domain on email data because using meetings data decreases performance. However, using lexical features, both in the supervised case where there is some labeled in-domain data and in the semi-supervised case where only unlabeled in-domain data is available, domain adaptation helps improve summarization. In our future work, we will combine lexical and conversational features and repeat the comparison. We will also test different ratios of data from each domain in order to understand when each of the domain adaptation methods works best for extractive summarization.

References

1. Florian, R, Hassan, H, Ittycheriah, A, Jing, H, Kambhatla, N, Luo, X, et al. A statistical model for multilingual entity detection and tracking. In *Proc. HLT-NAACL 2004*.

2. Daume H. Frustratingly easy domain adaptation. In *Proc. of ACL 2007*.

3. Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In *Proc. of EMNLP 2006*.

4. Blitzer, J. Domain Adaptation of Natural Language Processing Systems. PhD Thesis.

5. Carletta J, Ashby S, Bourban S, et al. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005*.

6. Ulrich J, Murray G, Carenini G. A publicly available annotated corpus for supervised email summarization. In *Proc. of AAAI EMAIL-2008*.

7. Klimt, B, & Yang, Y. Introducing the enron corpus. In *Proc. of CEAS 2004*.

8. Murray G, Carenini G. Summarizing spoken and written conversations. In *Proc. of EMNLP 2008*.

9. Murray, G, Carenini, G, & Ng, R. Interpretation and transformation for abstracting conversations. *HLT NAACL 2010*.