Vol.2011-NL-204 No.16
2011/11/22

A Comparative Study of Similarity Measures for Cross-Domain Sentiment Classification

DANUSHKA BOLLEGALA[†1]

1. Introduction

Users express opinions about products or services they consume in blog posts, shopping sites, or review sites. It is useful for both consumers as well as for producers to know what general public think about a particular product or service. Automatic document level sentiment classification[9),11)] is the task of classifying a given review with respect to the sentiment expressed by the author of the review. For example, a sentiment classifier might classify a user review about a movie as *positive* or *negative* depending on the sentiment expressed in the review. Sentiment classification has been applied in numerous tasks such as opinion mining[8)], opinion summarization[6)], contextual advertising[3)], and market analysis[4)].

Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a specific domain[9)]. However, sentiment is expressed differently in different domains, and it is costly to annotate data for each new domain in which we would like to apply a sentiment classifier. For example, in the domain of reviews about *electronics* products, the words "durable" and "light" are used to express positive sentiment, whereas "expensive" and "short battery life" often indicate negative sentiment. On the other hand, if we consider the *books* domain the words "exciting" and "thriller" express positive sentiment, whereas the words "boring" and "lengthy" usually express negative sentiment. A classifier trained on one domain might not perform well on a different domain because it would fail to learn the sentiment of the unseen words.

Work in *cross-domain sentiment classification*[1)] focuses on the challenge of training a classifier from one or more domains (source domains) and applying the trained classifier in a different domain (target domain). A cross-domain sentiment classification system must overcome two main challenges. First, it must identify which source domain features are related to which target domain features. Second, it requires a learning framework to incorporate the information regarding the relatedness of source and target domain features. Following previous work, we define cross-domain sentiment classification as the problem of learning a binary classifier (i.e. positive or negative sentiment) given a small set of labeled data for the source domain, and unlabeled data for both source and target domains. In particular, no labeled data is provided for the target domain.

In our previous work, we proposed a cross-domain sentiment classification method that uses an automatically generated sentiment sensitive thesaurus. We used an asymmetric relatedness measure that captures the phrases that express similar sentiments in different domains to build the sentiment sensitive thesaurus. In this follow-up paper, we study the sensitivity of this method against two other symmetric relatedness measures. Namely, cosine similarity and Lin's similarity measure[5)]. We would like to direct the interested reader to Bollegala et al.[2)] for the details of the cross-domain sentiment classification method that we study in this paper.

2. Relatedness Measures for Cross-Domain Sentiment Classification

In our previous work[2)], we represented a phrase u (both unigrams are bigrams of words are considered as phrases) using a feature vector $\boldsymbol{u}$. Here, three types of features are generated to represent a phrase. First, all the other phrases that co-occur with u in review sentences are selected as features. Second, the corresponding part-of-speech tags of those phrases are also selected as features to be included in $\boldsymbol{u}$. Those two types of features are called *lexical elements*. Third, from each labeled review we generate *sentiment elements* by appending the sentiment label (i.e. positive or negative) to all the lexical elements generated from that review. Both lexical and sentiment elements are used as features to represent a particular phrase. Moreover, the features are aggregated over all instances of occurrences of a particular phrase in a corpus to construct its feature vector.

We denote the value of a feature w in the feature vector $\boldsymbol{u}$ by $f(\boldsymbol{u}, w)$. We use pointwise mutual information between u and w as $f(\boldsymbol{u}, w)$. Pointwise mutual information is

†1 The University of Tokyo

known to be biased (over-weighting rare co-occurrences). To correct this bias we used the discounting scheme first proposed by Pantel and Ravichandran[10]. Moreover, negative pointwise mutual information values are set to zero following the recommendation by Lin[5].

Next, for two phrases u and v (represented by feature vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively), we compute the relatedness $\tau(v,u)$ of phrase v to phrase u as follows:

$$\tau(v,u) = \frac{\sum_{w \in \{x | f(\boldsymbol{v},x)>0\}} f(\boldsymbol{u},w)}{\sum_{w \in \{x | f(\boldsymbol{u},x)>0\}} f(\boldsymbol{u},w)}. \quad (1)$$

Note that relatedness is an asymmetric measure according the definition given in Equation 1, and the relatedness $\tau(v,u)$ of an element v to another element u is not necessarily equal to $\tau(u,v)$, the relatedness of u to v. Moreover, the relatedness measure defined in Equation 1 is in the range $[0,1]$. This relatedness measure can be seen as the additive recall measure proposed by Weeds and Weir[12] under their co-occurrence retrieval model.

In our previous work[2], we used this relatedness measure to build a thesaurus and subsequently used that thesaurus to expand feature vectors to train and test with a binary classifier. We omit the details of this method here. Our main objective of this follow-up work is to study the sensitivity of that method to the relatedness measure that is used to construct the sentiment sensitive thesaurus. For this purpose we construct a sentiment sensitive thesauri using two other popularly used relatedness measures. Namely the cosine similarity and Lin's similarity measure[5]. Next, we describe those two relatedness measures.

- **Cosine Similarity**: This is the cosine of the angle between the two vectors that represent two lexical elements u and v. Using the notation introduced above, it can be computed as follows:

$$\tau_{cos}(v,u) = \frac{\sum_{w \in \Gamma(v)} f(\boldsymbol{u},w)}{||u||\,||v||}, \quad (2)$$

$$||v|| = \sqrt{\sum_{w \in \Gamma(v)} (f(\boldsymbol{v},w))^2}, \quad (3)$$

$$||u|| = \sqrt{\sum_{w \in \Gamma(u)} (f(\boldsymbol{u},w))^2}.$$

Domain	positive	negative	unlabeled
kitchen	1000	1000	16746
DVDs	1000	1000	34377
electronics	1000	1000	13116
books	1000	1000	5947

Table 1 Number of reviews in the benchmark dataset.

Cosine similarity is widely used as a measure of relatedness in numerous tasks in natural language processing[7].

- **Lin's Similarity Measure**: We use the similarity measure proposed by Lin[5] for clustering similar words. This measure has shown to outperform numerous other similarity measures for word clustering tasks. It is computed as follows:

$$\tau_{Lin}(v,u) = \frac{\sum_{w \in \Gamma(v) \cap \Gamma(u)} (f(\boldsymbol{v},w) + f(\boldsymbol{u},w))}{\sum_{w \in \Gamma(v)} f(\boldsymbol{v},w) + \sum_{w \in \Gamma(u)} f(\boldsymbol{u},w)}. \quad (4)$$

All other components of the cross-domain sentiment classification method proposed in our previous work[2] such as feature vector construction procedure, pointwise mutual information calculation procedure, feature vector expansion procedure, and the binary classifier training procedure are held fixed in the experiments described in this paper. Therefore any difference in performance can be directly attributable to the relatedness measure used to build the sentiment sensitive thesaurus.

3. Dataset

We use the cross-domain sentiment classification dataset[⋆1] prepared by Blitzer et al.[1] to compare the proposed method against previous work on cross-domain sentiment classification. This dataset consists of Amazon product reviews for four different product types: books, DVDs, electronics and kitchen appliances. Each review is assigned with a rating (0-5 stars), a reviewer name and location, a product name, a review title and date, and the review text. Reviews with rating > 3 are labeled as positive, whereas those with rating < 3 are labeled as negative. The overall structure of this benchmark dataset is shown in Table 1. For each domain, there are 1000 positive and 1000 negative examples, the same balanced composition as the polarity dataset constructed by Pang et al.[9].

⋆1 http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

The dataset also contains some unlabeled reviews for the four domains. This benchmark dataset has been used in much previous work on cross-domain sentiment classification and by evaluating on it we can directly compare the proposed method against existing approaches.

Following previous work, we randomly select 800 positive and 800 negative labeled reviews from each domain as training instances (total number of training instances are $1600 \times 4 = 6400$), and the remainder is used for testing (total number of test instances are $400 \times 4 = 1600$). In our experiments, we select each domain in turn as the target domain, with one or more other domains as sources. Note that when we combine more than one source domain we limit the total number of source domain labeled reviews to 1600, balanced between the domains. For example, if we combine two source domains, then we select 400 positive and 400 negative labeled reviews from each domain giving $(400 + 400) \times 2 = 1600$. This enables us to perform a fair evaluation when combining multiple source domains.

4. Experiments and Results

Table 2 shows the performance of the proposed method when different relatedness measures are used to build the sentiment sensitive thesaurus that is used for domain adaptation. Because the relatedness measure defined in Equation 1 and to test whether this asymmetry is of any significance for the current task, we reversed the two arguments u and v in this relatedness measure and used this **reversed** relatedness measure as a baseline. The overall column in Table 2 is the average classification accuracy over the four target domains.

From Table 2, we see that the relatedness measure defined in Equation 1 outperforms all other relatedness measures compared in the table from an overall point-of-view. However, the differences in performance among all four relatedness measures compared in Table 2 are not significant. Therefore, we conclude that the cross-domain sentiment classification method that was proposed in our previous work[2] is *not sensitive* to the relatedness measure that is being used to build the sentiment sensitive thesaurus.

We have identified three reasons as to why the proposed method is insensitive to the relatedness measure being used. Next, we describe those reasons in detail.

First, note that the proposed method only uses the relatedness scores to rank the candidate expansions and do not use the absolute value of the relatedness scores. Therefore, as long as two different relatedness measures produce identical rankings for candidate expansions, those relatedness measure will obtain the same classification accuracy under the proposed method.

Table 2 Comparison of different relatedness measures.

Relatedness Measure	kitchen	dvd	electronics	books	overall
Cosine Similarity	0.8342	0.7826	0.8363	0.7657	0.8047
Lin's Similarity Measure[5]	0.8367	0.7826	0.8438	0.7632	0.8065
Proposed $\tau(v, u)$	0.8518	0.7826	0.8386	0.7632	0.8095
Reversed $\tau(u, v)$	0.8342	0.7852	0.8463	0.7632	0.8072

Second, we train a binary classifier after we perform feature expansion in the proposed cross-domain sentiment classification method. Therefore, if a particular relatedness measure introduces some incorrect expansion candidates, those candidates can get pruned out in the final model because the binary classifier will assign low weights to incorrect expansion candidates. Therefore, the binary classifier training procedure can be considered as a safe guard against any disfluencies inherent in a particular relatedness measure. However, it must be emphasized that although the binary classifier can prune out incorrect expansion candidates it *cannot* introduce the correct expansion candidates. Therefore, if a relatedness measure does not bring in the correct expansion candidates during the feature expansion step it will hurt the performance of cross-domain sentiment classification despite having a binary classifier.

Third, we observed that the asymmetry in the relatedness measure defined in Equation 1 is very small. In other words, although we defined an asymmetric relatedness measure for the purpose of cross-domain sentiment classification, the level of asymmetry that can be observed in actual data is very small. We arrive at this conclusion from the results of the following experiment.

We select pairs of phrases u and v, where v is listed as a neighbor of the base entry u as well as u is listed as a neighbor of the base entry v. From a sentiment sensitive thesaurus that has 1000 base entries, we were able to generate one-million $(1,000,0000)$ such pairs. This shows that all base entries in the sentiment sensitive thesaurus also appear as a neighbor for some other base entry. Next, we plot the relatedness score $\tau(u, v)$ against the relatedness score $\tau(v, u)$ to produce a correlation plot as shown in Figure 1. From Figure 1, we see that there is a high correlation between the proposed relatedness measure and its argument-reversed version. In fact, the Pearson correlation for

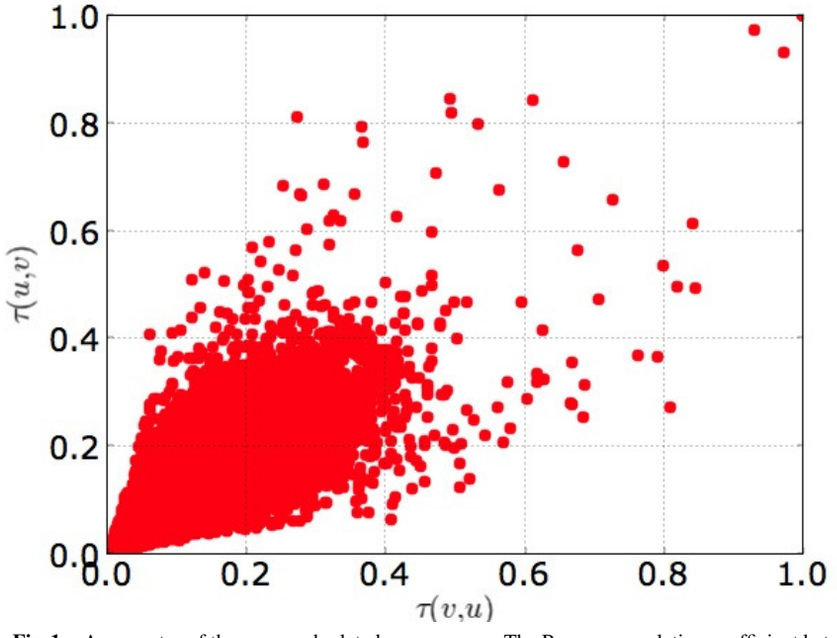

Fig. 1 Asymmetry of the proposed relatedness measure. The Pearson correlation coefficient between $\tau(v,u)$ and $\tau(u,v)$ is 0.8839.

this plot is as high as 0.8839 with a tight confidence interval of $[0.8835, 0.8844]$. Therefore, as mentioned above, the asymmetry of the proposed relatedness measure cannot be observed for the dataset being used in the experiments. This result explains why symmetric relatedness measures such as cosine similarity and Lin's similarity measure are performing at the same level as the proposed asymmetric relatedness measure on this benchmark dataset.

5. Conclusions

In this follow-up paper, we studied the effect of the relatedness measure on the cross-domain sentiment classification method proposed in our previous work[2]. We experimented with two other symmetric relatedness measures and a argument-reversed baseline using the multi-domain sentiment dataset. Our experimental results show that there is no significant difference among classification accuracy with the different relatedness measures used to build the sentiment sensitive thesaurus. Further investigations revealed three reasons for this insensitivity of the proposed method to the relatedness measure being used.

References

1) Blitzer, J., Dredze, M. and Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification, *ACL 2007*, pp.440–447 (2007).
2) Bollegala, D., Weir, D. and Carroll, J.: Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification, *ACL'2011*, pp. 132 – 141 (2011).
3) Fan, T.-K. and Chang, C.-H.: Sentiment-oriented contextual advertising, *Knowledge and Information Systems*, Vol.23, No.3, pp.321–344 (2010).
4) Hu, M. and Liu, B.: Mining and Summarizing Customer Reviews, *KDD 2004*, pp.168–177 (2004).
5) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *ACL 1998*, pp. 768–774 (1998).
6) Lu, Y., Zhai, C. and Sundaresan, N.: Rated aspect summarization of short comments, *WWW 2009*, pp.131–140 (2009).
7) Manning, C.D. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts (2002).
8) Pang, B. and Lee, L.: Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1–135 (2008).
9) Pang, B., Lee, L. and Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques, *EMNLP 2002*, pp.79–86 (2002).
10) Pantel, P. and Ravichandran, D.: Automatically Labeling Semantic Classes, *NAACL-HLT'04*, pp.321 – 328 (2004).
11) Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *ACL 2002*, pp.417–424 (2002).
12) Weeds, J. and Weir, D.: Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity, *Computational Linguistics*, Vol.31, No.4, pp.439–475 (2006).