Analysis of Asymmetric Measures for Performance Estimation of a Sentiment Classifier

Diego Uribe, Arturo Urquizo, Enrique Cuan

Instituto Tecnológico de la Laguna,
División de Estudios de Posgrado e Investigación,
Revolución y Cuauhtémoc, Torreón, Coah., Mexico
{diego,aurquizo,kcuan}@itlalaguna.edu.mx

Abstract. The development of a sentiment classifier experiences two problems to cope with: the demand of large amounts of labelled training data and a decrease in performance when the classifier is applied to a different domain. In this paper, we attempt to address this problem by exploring a number of metrics that try to predict the cross-domain performance of a sentiment classifier through the analysis of divergence between several probability distributions. In particular, we apply similarity measures to compare different domains and investigate the implications of using non-symmetric measures for contrasting feature distributions. We find that quantifying the difference between domains is useful to predict which domain has a feature distribution most similar to the target domain.

Keywords: Sentiment classifier, performance estimation, asymmetric measures.

1 Introduction

Domain adaptation is a common problem in several computational linguistic tasks. Information extraction is a task that takes unseen texts as input and produces structured-unambiguous data as output. However, it is a domain dependent task since when we need to extract information from a new domain, a new ad-hoc system is demanded. But building an information extraction system is difficult and time consuming [7], [5]. Similar challenges are also addressed by an open domain question answering system, a system to obtain concise answers to questions stated in natural language, that needs to be adaptable to play a crucial role in business intelligence applications [17].

Opinion mining is another very interesting computational linguistic task concerned with the classification of the reviews posted by the users, as well as the identification of the aspects of an object that people like or dislike [10], [6]. Since the object might be a product, a service, an organization, etc., opinion mining is also a domain dependent computational linguistic task. As reviews in different domains may be expressed in very different ways, training a classifier using data from one domain may fail when testing against data from another

one. In other words, we have to cope with a harder problem when the available training instances are dissimilar to the target domain. Aue and Gamon illustrate how the accuracy of a classifier trained on a different domain drops significantly compared to the performance of a classifier trained on its own native domain [2]. Thus, to determine which subset of outside domains has a feature distribution most similar to the target domain is of paramount importance.

In this study, we focus our attention in the analysis of the distributions corresponding to different domains in order to look for similarities that allow us to optimize the use of the available data. Said in another way, our aim is to look for differences between domains by using divergence measures. We show in this paper how two unannotated and different datasets are used to measure the contrast between their corresponding feature distributions. Once the contrast is determined, we can estimate the performance on the target domain B of an opinion classifier trained on the domain A. Since by using a non-symmetric measure we can obtain two similarity scores (AB and BA), it is also possible to estimate the performance on the target domain A of an opinion classifier trained on the domain B. As we analyse several domains, we may decide to implement a generic classifier depending on the similarity between one domain and other distinct domains.

We evaluate our approach with a data collection of several domains [15]. The results of the experimentation conducted show how the quantification of the divergence among domains is worthwhile to predict the domain with a feature distribution similar to a new target data.

The description of our work is organized as follows. The next section 2 makes a brief review of previous work on the domain adaptation problem in sentiment analysis. Section 3 describes in detail the divergence measures used in our analysis. Section 4 defines the dataset used in our experimentation as well as the pre-processing task for the extraction of the linguistic features to which we submitted our data collection. Then, the results of the experimentation are exhibited and discussed in section 5. Finally, conclusions are given in section 6.

2 Related Work

In this section we briefly describe some of the substantial works dealing with the problem of domain adaptation in sentiment classification. One of the first works in this specific topic was carried out by Aue and Gamon [2]. Their work is based on multiple ways of fusion of labeled data from other domains and unlabeled data from the target domain. The best results were obtained with an approach based on bootstrapping techniques. Shou-Shan et al. [14] propose an interesting algorithm for multi-domain sentiment classification based on the combination of learners for specific domains called *member classifiers*. The combination of these member classifiers is done according to two types of rules: fixed and trained rules. The purpose of the combination process is to obtain and to make available global information to the final classifier. Likewise, Blitzer et al. [3] cope with the domain adaptation problem by extending an algorithm for sentiment classifier

by making use of *pivot features* that relate the source and target domains. This relationship is defined in terms of frequency and mutual information estimation.

Also, there are significant works about comparing corpora to explore how corpus properties can affect the performance of natural language processing (NLP) tools. Sekine studied the effect of the use of different corpora on parsing tools [13]. Another interesting work to predict the cross-domain performance of an NLP tool was carried out by Asch and Daelemans [1]. Their work makes use of six similarity metrics to measure the difference between two corpora. Once the similarity is calculated, they investigate the correlation between similarity and the performance of an NLP tool such as a part-of-speech (POS) tagger. Other investigation concerned with the adaptation problem was carried out by Mansour et al. [11]. In their work, they make use of the Rényi divergence measure [12] to estimate the distance between diverse distributions.

3 Approach

In this section, we describe in detail our approach to estimate the subset of different domains with a feature distribution similar to the target domain. Unlike traditional supervised learning, adaptive learning entails the necessity to extract and exploit metaknowledge to assist the user in the task of selecting a suitable predictive model while taking into account the domain of application. One form of metaknowledge is to obtain insight about the data distribution [4]. We analyze in this work two ways to look for differences between domains by using non-symmetric divergence measures such as: Kullback-Leibler divergence [8] and cross-entropy.

3.1 Kullback-Leibler (KL) Divergence

The KL divergence (also known as relative entropy) is a measure of how different two probability distributions are. To be more specific, the KL divergence of q from p, denoted by $D(p \parallel q)$, is a measure of the information lost when q is used to approximate p.

$$D(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

As the relative entropy of a target dataset, given a source dataset, is the data required to reconstruct the target, our interest in this divergence measure consists in to observe the behavior of a learning tool for the domain B that has been trained in terms of the domain A. In other words, we are interested in the performance estimation of a learning tool for B when using a feature set corresponding to A, rather than to B. And since KL is a non-symmetric measure, we can also observe the behavior of a learning tool in the opposite direction: the performance for A when using a feature set corresponding to B.

3.2 Cross Entropy (CE)

Cross entropy is also a measure to compare different probability distributions that bears close relation to the KL divergence (relative entropy). However, the purpose of the cross entropy between a random variable X with probability distribution p and another probability distribution q, denoted by $H(X, q)$, is to observe the role of q as model of the real distribution p.

$$H(X, q) = H(X) + D(p \parallel q) = -\sum_{x} p(x) \log q(x) \tag{2}$$

Our interest in this similarity measure consists in to observe a model q of the real distribution p. According to the expression 2, it is noticed that we want to minimize $D(p \parallel q)$: an optimum probabilistic model q is obtained as long as the divergence between p and q can be minimized. In this way, we are concerned in a learning tool for B trained on a model dataset A, in order to assess how accurate the model is in predicting B. Moreover, cross entropy is also a non-symmetric measure, so we can observe the behavior of a learning tool by measuring the cross entropy between A and B, and vice versa.

4 Experimental Setup and Results

We use for the experimentation conducted in this work a collection of Epinions reviews developed by Taboada et al. [15]. Such dataset consists of eight different categories: books, cars computers, cookware, hotels, movies, music and phones. There are 50 opinions per category, giving a total of 400 reviews in the collection, which contains a grand total of 279,761 words. And since within each category there are 25 reviews per polarity, the baseline accuracy for each domain is 50%.

The set of sentences corresponding to each review in the dataset used in our experimentation was submitted to a tagger based on a broad use of lexical features: The Stanford Tagger with a remarkable degree of accuracy [16]. We model each review as a feature vector. The granularity of the feature sets used in our experiments consisted of unigrams and bigrams. As the frequency of the ngram is required by the similarity measures, frequency features rather than binary features represent our content vector.

To be able to observe reliable regularities in the information provided by our divergence measures, we have only considered those domains with at least 4,000 features. Thus, we have discarded two domains: Cookware and Phones. These domains have been discarded because we can consider them as *outliers*: their number of features is clearly separated from the rest of the domains.

Once the data representation model (datasets and the content vector) has been defined, we estimate the subset of domains with a feature distribution similar to the target domain by making use of the similarity measures previously mentioned. Thus, from the feature sets corresponding to each domain we produce the matrix of the KL divergence (relative entropy) of the unigrams across domains shown in Table 1. It can be noticed how the entries on the main diagonal are zero, that is, when $p = q$, the KL divergence is 0.

Table 1. KL scores across domains

	Books	**Cars**	**Compu**	**Hotels**	**Movies**	**Music**
Books		0.64	0.54	0.56	0.42	0.70
Cars	0.42		0.46	0.50	0.50	0.67
Compu	0.24	0.49		0.44	0.46	0.59
Hotels	0.44	0.63	0.51		0.50	0.82
Movies	0.25	0.49	0.52	0.39		0.64
Music	0.21	0.51	0.42	0.42	0.40	

In the same way, we generate the cross entropy of the unigrams across domains shown in Table 2. In this case, the entries on the main diagonal represent the entropy of p, that is, when $p = q$, the cross entropy is H(p).

Table 2. CE scores across domains

	Books	**Cars**	**Compu**	**Hotels**	**Movies**	**Music**
Books		6.81	6.51	6.56	7.32	7.42
Cars	6.00		7.12	6.83	6.30	6.81
Compu	5.66	7.41		6.56	6.36	6.64
Hotels	6.17	7.24	6.70		6.56	6.96
Movies	6.67	6.64	6.60	6.35		7.50
Music	5.63	6.36	6.00	5.80	6.37	

5 Analysis and Discussion

Taking into account that the higher the cross entropy, the more similar the two domains, cross entropy is a guide as to how well a classifier trained on one domain will work when tested on another target domain.In the case of the KL divergence: the lower the relative entropy, the more similar the two domains, we can also make use of the KL divergence to estimate the performance of a classifier that has been trained on a foreign domain. For example, when the target domain is Books, Table 1 suggests the Movies domain as the best option to train a classifier. However, Table 2 proposes the Music domain as the best option.

Thus, once we obtained the similarity distributions for each target domain, we want to corroborate such distributions. In other words, we want to observe if, for example, the feature set of Movies represents a better option to classify Books reviews rather than the Music's features. In order to carry out this corroboration, we evaluate the performance for each target domain using a classifier based on the common features between the unseen reviews of the target domain and each of the foreign domains.

The method to evaluate the performance is based on 5-fold cross-validation and the use of support vector machines (SVM). As we know, SVM is a hyperplane

classifier that has proved to be a useful approach to cope with natural text affairs [12]. Table 3 shows the SVM classifier accuracy for each target domain. Taking into account that the baseline accuracy for each domain is 50%, Table 3 exhibits how the use of the linguistic features corresponding to Music (77%) represent a better option to classify Books rather than the Movies' features (73%).

Table 3. SVM scores across domains

	Books	**Cars**	**Compu**	**Hotels**	**Movies**	**Music**
Books		73%	75%	69%	73%	77%
Cars	74%		82%	78%	76%	78%
Compu	80%	74%		78%	80%	78%
Hotels	76%	76%	72%		76%	74%
Movies	82%	76%	78%	84%		84%
Music	80%	74%	74%	68%	74%	

Additionally, we make use of the TP rate and FP rate values to show an alternative perspective of the performance estimation. Since the TP rate and FP rate values of different classifiers on the same test dataset are often represented diagrammatically by a ROC graph, Figure 1 shows the ROC graph corresponding to different classifiers tested on Books and trained on each of the foreign domains. As Figure 1 shows, the use of Music as model for Books has outperformed the rest of the domains. Therefore, the information provided by cross entropy has been more useful to identify a feature distribution most similar to the target domain.

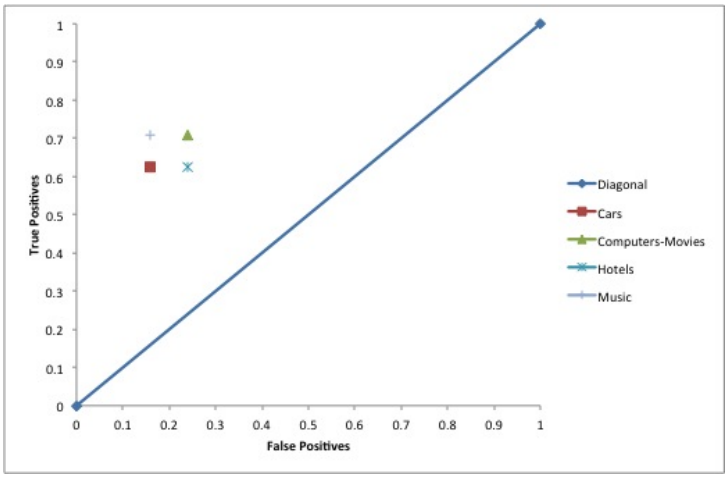

Fig. 1. ROC graph corresponding to Books

The use of non-symmetric measures is also exhibited by the obtained results. As we can see in Tables 1 and 2, the divergence value obtained between any two domains, A and B, is not the same as the one obtained between B and A. For example, both tables show how the divergence value obtained between Books and Movies is not the same discrepancy value obtained between Movies and Books. More specifically, the KL divergence in Table 1 shows how Books diverge less from Movies than the opposite case. On the other hand, the cross entropy in Table 2 shows how Movies is a more useful model to predict Books than Books to classify Movies. Taking into account the results shown in Table 3, the use of the linguistic features corresponding to Books represent a better option to classify Movies (82%) rather than the use of Movies to classify Books (73%).

Also, as an alternative perspective of the performance estimation, Figure 2 shows the ROC graph corresponding to different classifiers tested on Movies and trained on each of the foreign domains. Now, by comparing Figure 1 and Figure 2, we can see how the use of Books as model for Movies has outperformed the rest of the domains. Thus, the information provided in this case by relative entropy has been more useful to identify a feature distribution most similar to the target domain.

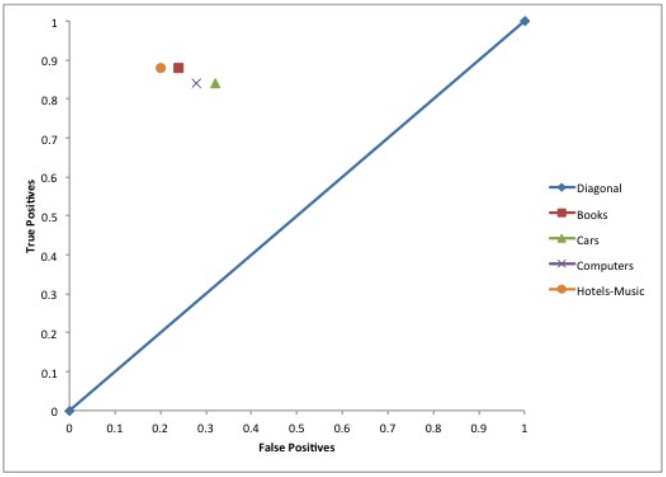

Fig. 2. ROC graph corresponding to Movies

By analyzing the information provided by our divergence measures (Tables 1 and 2), the most important point is to observe regularities that allow us to make reliable predictions when training a classifier for a domain for which no annotated data is available. According to the obtained results in our experimentation, we are able to observe the relationship between the performance (Table 3) and a divergence measure (Tables 1 and 2): in most of the cases, when the difference

between the cross entropy values is close (less than one), KL is the best guide to predict the domain with a feature distribution similar to the target domain (i.e. see the divergence values between Books and Movies and vice versa). Otherwise, CE is an alternative to make such predictions (i.e. see the divergence values between Movies and Music and vice versa).

6 Conclusions and Future Work

In this paper, we focus our attention in the analysis of the distributions corresponding to different domains in order to determining the subset of domains with a feature distribution similar to the unlabeled target domain. By making use of non-symmetric divergence measures, we estimate the performance on the unlabeled target domain B of an opinion classifier trained on the domain A and vice versa: the performance on the unlabeled target domain A of an opinion classifier trained on the domain B. We find that quantifying the difference between domains is useful not only to predict which domain has a feature distribution most similar to the target domain but also to optimize the use of the available data.

As part of our future work, we intend to extend our distributional analysis by including measures that allow us to cope with feature-distribution vectors that are quite sparse. In particular, we intend to explore the implications of the use of the Jenson-Shannon divergence [9].

Also, we are interested in the analysis of more datasets collections. For example, the dataset collected by Blitzer et al. [3] is an interesting collection of product reviews from four domains: books, DVDs, electronics, and kitchen appliances. We think this collection is worth our attention to improve and optimize our analysis.

References

1. Asch, V., Daelemans, W.: Using domain similarity for performance estimation. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing. pp. 31–36 (2010)
2. Aue, A., Gamon, M.: Customizing sentiment classifiers to new domains: a case study. In: Proceedings of Recent Advances in Natural Language Processing (RANLP) (2005)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) (2007)
4. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning Applications to Data Mining. Springer-Verlag (2009)
5. Cardie, C.: Empirical methods in information extraction. AI Magazine 39(1), 65–79 (1997)
6. Feldman, R.: Techniques and applications for sentiment analysis. Communications of the ACM 56(4), 82–89 (2013)

7. Glickman, O., Jones, R.: Examining machine learning for adaptable end-to-end information extraction systems. In: AAAI 1999 Workshop on Machine Learning for Information Extraction (1999)
8. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
9. Lee, L.: Measures of distributional similarity. pp. 25–32 (1999)
10. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag (2007)
11. Mansour, Y., Mohri, M., Rostamizadeh, A.: Multiple source adaptation and the rényi divergence. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 367–374 (2009)
12. Rényi, A.: On measures of information and entropy. In: Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability. vol. 1, pp. 547–561 (1961)
13. Sekine, S.: The domain independence of parsing. In: Proceedings of the 5th Conference on Applied Natural Language Processing (1997)
14. Shou-Shan, L., Chu-Ren, H., Cheng-Qing, Z.: Multi-domain sentiment classification with classifier combination. Journal of Computer Science and Technology 26(1), 25–33 (2011)
15. Taboada, M., Anthony, C., Voll, K.: Creating semantic orientation dictionaries. In: Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC). pp. 427–432 (2006)
16. Toutanova, K.and Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). pp. 252–259 (2003)
17. Vila, K., Ferrández, A.: Model-driven restricted-domain adaptation of question answering systems for business intelligence. In: Proceedings of the 2nd International Workshop on Business Intelligence and the WEB. pp. 36–43 (2011)