Evaluation of Different Strategies for Domain Adaptation in Opinion Mining

Anne Garcia-Fernandez[*], **Olivier Ferret**[†], **Marco Dinarelli**[‡]

[†]CEA, LIST, Vision and and Content Engineering Laboratory
Gif-sur-Yvette, F-91191 France
olivier.ferret@cea.fr

[*]LAS - EHESS
52 rue du cardinal Lemoine, Paris, F-75005 France
annegf@college-de-france.fr

[‡]LaTTiCe
1 rue Maurice Arnoux, Montrouge, F-92120 France
marco.dinarelli@ens.fr

Abstract

The work presented in this article takes place in the field of opinion mining and aims more particularly at finding the polarity of a text by relying on machine learning methods. In this context, it focuses on studying various strategies for adapting a statistical classifier to a new domain when training data only exist for one or several other domains. This study shows more precisely that a self-training procedure consisting in enlarging the initial training corpus with unannotated texts from the target domain that were reliably classified by the classifier is the most successful and stable strategy for the tested domains. Moreover, this strategy gets better results in most cases than (Blitzer et al., 2007)'s method on the same evaluation corpus while it is more simple.

Keywords: Opinion mining, domain adaptation, self-training

1. Introduction

The work we present in this article takes place in the field of opinion mining and focuses more particularly on the global classification of texts as positive or negative. A large set of approaches have been developed for performing this task, relying on various resources such as manually annotated corpora or lexicons dedicated to opinion mining or sentiment analysis. Whatever the kind of approach, the problem of domain adaptation is particularly significant in the field of opinion mining because this task is not supposed to be topic dependent whereas the resources used for achieving it generally depend on a particular context. This is especially true for corpora built for training statistical classifiers, but the problem also arises for lexicon-based approaches as exploiting only a general opinion lexicon leads to poor results, even when the context in which these opinion terms occur is taken into account (Taboada et al., 2011; Choi and Cardie, 2009).

(Jijkoun et al., 2010) and (Gindl et al., 2010) are examples of work about adapting a general opinion lexicon to a specific domain without supervision, with two different strategies: (Jijkoun et al., 2010) identifies the terms of the domain that are similar to terms whose polarity is already known from a general lexicon while (Gindl et al., 2010) discards the terms whose polarity can change from one domain to another. However, a general lexicon can also be used to compensate the dependence of a training corpus on a particular domain, as in (Denecke, 2009), which exploits the SentiWordNet lexicon (Esuli and Sebastiani, 2006). Finally, another trend of work focuses on domain adaptation without exploiting a reference opinion lexicon. While (Aue and Gamon, 2005) tested different strategies inspired by work on text classification, (Blitzer et al., 2007) proposed a method called *Structural Correspondence Learning* (SCL) for identifying pivots between different domains and (Li and Zong, 2008) explored two kinds of fusion: fusion of training corpora coming from different domains and fusion of the results of classifiers trained on different domains. The option of the unsupervised annotation of a training corpus by the means of self-training was considered in (Drury et al., 2011) for one domain. Its transposition to the context of several domains is one of the main approaches we have evaluated in this work.

2. Strategies for domain adaptation

The work we consider in this article focuses more specifically on the classification of texts in terms of positive or negative polarity by relying on a supervised approach exploiting a manually annotated training corpus but without using an opinion lexicon built *a priori*. In this context, similar to (Blitzer et al., 2007) or (Li and Zong, 2008) for instance, our objective is to find, starting from a training corpus for one or several source domains, called *source corpus*, and an unannotated corpus for a target domain, called *development corpus*, the best strategy for building a new training corpus, called *target training corpus*, from these existing corpora to get the best possible results on the target domain. The strategies we have considered for tackling this problem can be differentiated according to two main factors:

- the target training corpus is built from only one or several source domains;

```
<review>
 (...)
 <rating>5.0</rating>
 <review_text>
  I read Les Misérables after I saw the opera, and it has inspired
  in me more than any book I've ever read. I don't believe one
  could ever find a better novel anywhere. For everyone (...)
 </review_text>
</review>
```

Figure 1: Example of a review of the MDSD corpus for the BOO domain

	Training corpus		Development corpus	Test corpus
Domains	#reviews	#words/reviews	# reviews	# reviews
Kitchen appliances (KIT)	2,000	96	2,000	3,945
Books (BOO)	2,000	174	2,000	2,465
DVD (DVD)	2,000	189	2,000	3,945
Electronics appliances (ELE)	2,000	113	2,000	3,945

Table 1: General description of our evaluation corpus

- the building of the target training corpus makes use or not of a development corpus.

We have more particularly tested the following strategies, each one exploiting the same amount of manually annotated data for building the training corpus for the target domain:

One source corpus [BASELINE] This baseline strategy uses the training corpus of only one source domain as a training corpus for the target domain.

Iterative learning from a development corpus [ITE-FIXED, ITE-THRE] In this strategy, we first train a model with one source corpus, as in the BASELINE strategy, then classify all the documents of the development corpus, select the documents with the higher classification confidence score and add them to the target training corpus (initially made of the documents of the considered source corpus). This set of operations forms a basic step that is repeated several times, each iteration starting with a larger training corpus than the preceding one. The overall process corresponds to a kind of self-training procedure. In our case, the process is repeated until all the development corpus has been integrated into the target training corpus. The selection of the best documents at each step can be done according to a fixed threshold applied to the confidence score (ITE-THRE) or according to a fixed number of documents to add to the target training corpus at each step (ITE-FIXED).

Several source corpora [MULTI-DOMAIN] In this configuration, the target training corpus is built by joining the corpora of several source domains and more particularly in this case, of all source domains. This represents the baseline for approaches using multiple corpora.

Vote [MULTI-VOTE] This strategy consists in training a model from each source corpus and performing a final classification by the vote of the resulting classifiers: each document is classified according to the majority decision of the classifiers dedicated to each source domain.

Iterative learning from several source corpora [ITE-MULTI-VOTE] This strategy is a hybrid of the vote and the iterative learning strategies. Starting from a set of N source domains, N classifiers are trained from the training corpus associated with each of these domains. Then, the documents of the development corpus are processed by these classifiers and the documents that are classified with the highest confidence by a classifier are added to its training corpus. Finally, the classification of a document results from the vote of these classifiers.

3. Implementation and results

Our experiments were done with the reference Multi-domain Sentiment Dataset (MDSD) corpus of (Blitzer et al., 2007), which is made of reviews of various products available on the Amazon website. These reviews cover four types of products: DVDs (DVD), books (BOO), electronics (ELE) and kitchen appliances (KIT). Each review is composed of a short text – only few sentences – a title and a rating ranging from 1 to 5, 1 being the most negative rating while 5 being the most positive one. Figure 1 gives an example of such review for the subcorpus "Books".

As (Blitzer et al., 2007), we consider a review as positive if its rating is strictly higher than 3 and negative otherwise. Each type of product is considered as a separate domain and for each of these domains, the same amount of training data – 2,000 reviews, split in a balanced way into positive and negative examples – is available. The only difference with (Blitzer et al., 2007) is our splitting of the test corpus for each of these domains into two subsets: one subset of 2,000 reviews used as part of the development corpus and the remaining part taken as our test corpus for this domain. This is summarized by Table 1 which also gives the average size of reviews for the training corpus.

For evaluating our different strategies, we adopted as (Torres-Moreno et al., 2007) a classification model based

TRN	TEST	baseline	ite-fixed	ite-thre	multi-domain	multi-vote	ite-multi-vote
DVD	BOO	79.7	48.8	84.4$*$	68.1	69.6	75.6†
ELE	BOO	75.4	41.6	79.3$*$			
KIT	BOO	70.9	38.1	81.8$*$			
BOO	DVD	77.2	69.5	82.0$*$	72.2	70.3	81.4†
ELE	DVD	76.2	54.3	73.4			
KIT	DVD	76.9	54.0	78.2$*$			
BOO	ELE	77.5$*$	64.5	65.5	69.1	64.7	77.5†
DVD	ELE	74.1$*$	69.7	62.2			
KIT	ELE	86.8$*$	60.4	75.7			
BOO	KIT	78.9	68.4	82.7$*$	75.4	71.0	81.4†
DVD	KIT	81.4	61.1	82.3$*$			
ELE	KIT	85.9	65.6	78.7			

Table 2: Results for our different strategies

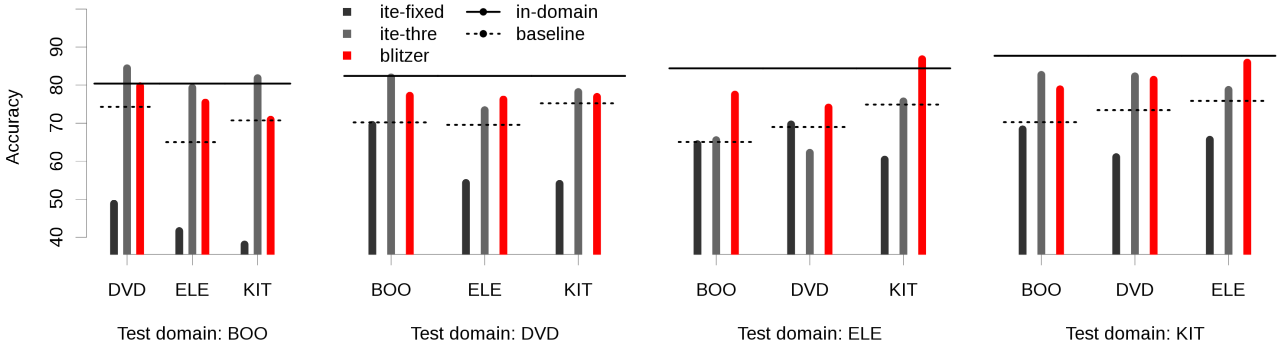

Figure 2: Accuracy of the strategies only based on one source domain, compared to (Blitzer et al., 2007)[1]

on boosting, implemented with BoosTexter (E. and Y., 2000). This model is made of a set of binary rules, called *i.e.* decision stumps, used as weak learners in this context. Each rule tests the presence of a particular n-gram and is associated with a probability in relation to each considered class, two in our case. For the classification of a document, a score is computed for each class from the rules of this class triggered by the document and the class with the highest score is assigned to the document. The values of BoosTexter's parameters were set experimentally by optimizing results for our `baseline` approach: no use of lemmatisation, rules based on unigrams and 50 rounds of weak learning.

Table 2 shows the results of the evaluation of our different strategies in this evaluation framework, expressed in terms of classification accuracy to make the direct comparison with the results of (Blitzer et al., 2007) possible. If we look at strategies relying on only one source domain, we can first observe that the accuracy of the `baseline` strategy is higher than 70% for all possible pairs of a training domain (TRN) and a test domain (TEST). We can also note that the `ite-fixed` strategy is our worst strategy, with results consistently lower than our `baseline` strategy. This result can be explained by the fact that this strategy does not take into account the confidence score given by BoosTexter. As a consequence, it has no means to stop adding new examples to the training set and finally add all the development corpus to the training corpus without trying to limit the number of negative examples for the target domain. By contrast, exploiting the confidence score of the classifier, which is done by the `ite-thre` strategy, gives the best accuracy (marked with $*$) for most of pairs (source domain, target domain). The results of this strategy are lower than the results of `baseline` only in some cases where ELE is the source or the target domain. Finally, it is also interesting to note that results are not symmetric: training on domain A and testing on domain B does not lead to the same accuracy as training on domain B and testing on domain A. The difference can even be significant, as for the pair (DVD,KIT): the accuracy is equal to 81.4 with DVD as the source domain while it is only equal to 76.9 with KIT as the source domain.

The results of the strategies based on several source domains are presented in the last three column of Table 2. The `multi-domain` strategy is our baseline in this context. The comparison of its results with those of `baseline`, the baseline for one source domain strategies, is clearly in favor of `baseline` in all cases. This observation tends to show that using a training corpus covering several domains (`multi-domain` strategy) is a worse option for classifying documents in terms of opinion in another domain than using a homogeneous training corpus, provided that the size of the training corpus is the same in the two cases. However, it is possible that such conclusion can depend on the size of the corpora and the number of domains. The introduction of a mechanism of vote as it is done by the `multi-vote` strategy does not lead to exceed the results of `multi-domain`. As `ite-fixed`, multi-vote does not exploit the confidence score given by the classifier and the fact that a document is assigned to a class by a majority of source domain classifiers does not seem to be sufficient to compensate this problem. But once again, the number of source domains may have an influence that is not visible in these experiments. Finally, the ite-multi-vote strategy appears as the best strategy among those relying on several source domains as it gets the best results (marked with †) for all domains. The average accuracy of ite-multi-vote, equal to 79.0, is even slightly higher than the average accuracy of ite-thre, equal to 77.2, especially thanks to better results for the target domain ELE. More generally, the iterative learning method seems to be particularly interesting, as it gets the best results both for one or several source domains.

Figure 2 shows a comparison of our results with those of (Blitzer et al., 2007). The horizontal solid lines give the accuracy of a purely intra-domain strategy according to (Blitzer et al., 2007): the training and test corpora belong to the same domain. The horizontal dotted lines correspond the accuracy of our baseline strategy. For each pair of a source and a target domain, our results are compared to those of (Blitzer et al., 2007). The main observation is that our best strategy, ite-thre, has globally higher results than (Blitzer et al., 2007), except for the ELE domain (either as source or target domain). The method of (Blitzer et al., 2007) proposed a way to find terms that are likely to be equivalent in two domains. These terms are n-grams such as *must read* (BOO) or *excellent product* (KIT). However, these terms are mainly made of plain words. One possible explanation of the difference between our results and those of (Blitzer et al., 2007) is the fact that our classifiers do not focus on a particular type of lexical units. As a consequence, they also exploit grammatical words, which conjugates two interesting properties: they are more likely to be found in all domains than plain words and they are particularly useful for expressing the negation or the intensity of an opinion (Taboada et al., 2011).

4. Conclusion and perspectives

In this article, we have presented a study about different possible strategies for building a statistical classifier for a target domain with only annotated data for one or several other domains. We have more particularly showed the efficiency for this task of a kind of self-training learning procedure based on the incremental annotation of an unlabeled corpus of the target domain. In a significant number of cases, this approach even exceeds the results of (Blitzer et al., 2007) whereas it is much simpler. The most direct extension to this work is to apply its method to other kinds of statistical classifiers. Furthermore, we are interested in developing more complex self-training strategies by using different kinds of classifiers, as in (Wiebe and Riloff, 2005), or different configurations of the same kind of classifiers.

5. References

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. In 5th *International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.

Choi, Y. and Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 590–598, Singapore.

Denecke, K. (2009). Are SentiWordNet scores suited for multi-domain sentiment classification? In *Fourth International Conference on Digital Information Management (ICDIM 2009)*, pages 1–6, Ann Arbor, Michigan.

Drury, B., Torgo, L., and Almeida, J. J. (2011). Guided Self Training for Sentiment Classification. In *RANLP 2011 Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing (ROBUS 2011)*, pages 18–25, Hissar, Bulgaria.

E., S. R. and Y., S. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(1):135–168.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, Genova, Italy.

Gindl, S., Weichselbraun, A., and Scharl, A. (2010). Cross-domain contextualization of sentiment lexicons. In Coelho, H., Studer, R., and Wooldridge, M., editors, *19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 771–776, Lisbon, Portugal.

Jijkoun, V., de Rijke, M., and Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 585–594, Uppsala, Sweden.

Li, S. and Zong, C. (2008). Multi-domain sentiment classification. In *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (ACL-08: HLT)*, pages 257–260, Columbus, Ohio.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Torres-Moreno, J.-M., El-Bèze, M., Béchet, F., and Camelin, N. (2007). Comment faire pour que l'opinion forgée la sortie des urnes soit la bonne ? Application au défi DEFT 2007. In *Atelier DEFT'07, Plate-forme AFIA 2007*, Grenoble, France.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Sixth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, pages 486–497, Mexico City, Mexico. Springer.

[1] The *in-domain* approach corresponds to the same configuration as our baseline approach, except that the training and the test domains are part of the same domain.