Improving Citation Polarity Classification with Product Reviews

Charles Jochim*
IBM Research – Ireland
charlesj@ie.ibm.com

Hinrich Schütze
Center for Information & Language Processing
University of Munich

Abstract

Recent work classifying citations in scientific literature has shown that it is possible to improve classification results with extensive feature engineering. While this result confirms that citation classification is feasible, there are two drawbacks to this approach: (i) it requires a large annotated corpus for supervised classification, which in the case of scientific literature is quite expensive; and (ii) feature engineering that is too specific to one area of scientific literature may not be portable to other domains, even within scientific literature. In this paper we address these two drawbacks. First, we frame citation classification as a domain adaptation task and leverage the abundant labeled data available in other domains. Then, to avoid over-engineering specific citation features for a particular scientific domain, we explore a deep learning neural network approach that has shown to generalize well across domains using unigram and bigram features. We achieve better citation classification results with this cross-domain approach than using in-domain classification.

1 Introduction

Citations have been categorized and studied for a half-century (Garfield, 1955) to better understand when and how citations are used, and to record and measure how information is exchanged (e.g., networks of co-cited papers or authors (Small and Griffith, 1974)). Recently, the value of this information has been shown in practical applications such as information retrieval (IR) (Ritchie et al., 2008), summarization (Qazvinian and Radev, 2008), and even identifying scientific breakthroughs (Small and Klavans, 2011). We expect that by identifying and labeling the *function* of citations we can improve the effectiveness of these applications.

There has been no consensus on what aspects or functions of a citation should be annotated and how. Early citation classification focused more on *citation motivation* (Garfield, 1964), while later classification considered more the *citation function* (Chubin and Moitra, 1975). Recent studies using automatic classification have continued this tradition of introducing a new classification scheme with each new investigation into the use of citations (Nanba and Okumura, 1999; Teufel et al., 2006a; Dong and Schäfer, 2011; Abu-Jbara et al., 2013). One distinction that has been more consistently annotated across recent citation classification studies is between *positive* and *negative* citations (Athar, 2011; Athar and Teufel, 2012; Abu-Jbara et al., 2013).[1] The popularity of this distinction likely owes to the prominence of sentiment analysis in NLP (Liu, 2010). We follow much of the recent work on citation classification and concentrate on citation polarity.

2 Domain Adaptation

By concentrating on citation polarity we are able to compare our classification to previous citation polarity work. This choice also allows us to access the wealth of existing data containing polarity annotation and then frame the task as a domain adaptation problem. Of course the risk in approaching the problem as domain adaptation is that the domains are so different that the representation of a positive instance of a movie or product review, for example, will not coincide with that of a posi-

*This work was primarily conducted at the IMS – University of Stuttgart.

[1] Dong and Schäfer (2011) also annotate polarity, which can be found in their dataset (described later), but this is not discussed in their paper.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 42–48,
Baltimore, Maryland, USA, June 23-25 2014.

tive scientific citation. On the other hand, because there is a limited amount of annotated citation data available, by leveraging large amounts of annotated polarity data we could potentially even improve citation classification.

We treat citation polarity classification as a sentiment analysis domain adaptation task and therefore must be careful not to define features that are too domain specific. Previous work in citation polarity classification focuses on finding new citation features to improve classification, borrowing a few from text classification in general (e.g., n-grams), and perhaps others from sentiment analysis problems (e.g., the polarity lexicon from Wilson et al. (2005)). We would like to do as little feature engineering as possible to ensure that the features we use are meaningful across domains. However, we do still want features that somehow capture the inherent positivity or negativity of our labeled instances, i.e., citations or Amazon product reviews. Currently a popular approach for accomplishing this is to use deep learning neural networks (Bengio, 2009), which have been shown to perform well on a variety of NLP tasks using only bag-of-word features (Collobert et al., 2011). More specifically related to our work, deep learning neural networks have been successfully employed for sentiment analysis (Socher et al., 2011) and for sentiment domain adaptation (Glorot et al., 2011). In this paper we examine one of these approaches, marginalized stacked denoising autoencoders (mSDA) from Chen et al. (2012), which has been successful in classifying the polarity of Amazon product reviews across product domains. Since mSDA achieved state-of-the-art performance in Amazon product domain adaptation, we are hopeful it will also be effective when switching to a more distant domain like scientific citations.

3 Experimental Setup

3.1 Corpora

We are interested in domain adaptation for citation classification and therefore need a target dataset of citations and a non-citation source dataset. There are two corpora available that contain citation function annotation, the DFKI Citation Corpus (Dong and Schäfer, 2011) and the IMS Citation Corpus (Jochim and Schütze, 2012). Both corpora have only about 2000 instances; unfortunately, there are no larger corpora available with citation annotation and this task would benefit from more annotated data. Due to the infrequent use of negative citations, a substantial annotation effort (annotating over 5 times more data) would be necessary to reach 1000 negative citation instances, which is the number of negative instances in a single domain in the multi-domain corpus described below.

The DFKI Citation Corpus[2] has been used for classifying citation function (Dong and Schäfer, 2011), but the dataset also includes polarity annotation. The dataset has 1768 citation sentences with polarity annotation: 190 are labeled as *positive*, 57 as *negative*, and the vast majority, 1521, are left *neutral*. The second citation corpus, the IMS Citation Corpus[3] contains 2008 annotated citations: 1836 are labeled *positive* and 172 are labeled *negative*. Jochim and Schütze (2012) use annotation labels from Moravcsik and Murugesan (1975) where positive instances are labeled *confirmative*, negative instances are labeled *negational*, and there is no neutral class. Because each of the citation corpora is of modest size we combine them to form one citation dataset, which we will refer to as CITD. The two citation corpora comprising CITD both come from the ACL Anthology (Bird et al., 2008): the IMS corpus uses the ACL proceedings from 2004 and the DFKI corpus uses parts of the proceedings from 2007 and 2008. Since mSDA also makes use of large amounts of unlabeled data, we extend our CITD corpus with citations from the proceedings of the remaining years of the ACL, 1979–2003, 2005–2006, and 2009.

There are a number of non-citation corpora available that contain polarity annotation. For these experiments we use the Multi-Domain Sentiment Dataset[4] (henceforth MDSD), introduced by Blitzer et al. (2007). We use the version of the MDSD that includes *positive* and *negative* labels for product reviews taken from Amazon.com in the following domains: books, dvd, electronics, and kitchen. For each domain there are 1000 positive reviews and 1000 negative reviews that comprise the "labeled" data, and then roughly 4000 more reviews in the "unlabeled"[5] data. Reviews

[2] https://aclbib.opendfki.de/repos/trunk/citation_classification_dataset/

[3] http://www.ims.uni-stuttgart.de/~jochimcs/citation-classification/

[4] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[5] It is usually treated as unlabeled data even though it ac-

Corpus	Instances	Pos.	Neg.	Neut.
DFKI	1768	190	57	1521
IMS	2008	1836	172	–
MDSD	27,677	13,882	13,795	–

Table 1: Polarity corpora.

were preprocessed so that for each review you find a list of unigrams and bigrams with their frequency within the review. Unigrams from a stop list of 55 stop words are removed, but stop words in bigrams remain.

Table 1 shows the distribution of polarity labels in the corpora we use for our experiments. We combine the DFKI and IMS corpora into the CITD corpus. We omit the citations labeled *neutral* from the DFKI corpus because the IMS corpus does not contain neutral annotation nor does the MDSD. It is the case in many sentiment analysis corpora that only positive and negative instances are included, e.g., (Pang et al., 2002).

The citation corpora presented above are both unbalanced and both have a highly skewed distribution. The MDSD on the other hand is evenly balanced and an effort was even made to keep the data treated as "unlabeled" rather balanced. For this reason, in line with previous work using MDSD, we balance the labeled portion of the CITD corpus. This is done by taking 179 unique negative sentences in the DFKI and IMS corpora and randomly selecting an equal number of positive sentences. The IMS corpus can have multiple labeled citations per sentence: there are 122 sentences containing the 172 negative citations from Table 1. The final CITD corpus comprises this balanced corpus of 358 labeled citation sentences plus another 22,093 unlabeled citation sentences.

3.2 Features

In our experiments, we restrict our features to unigrams and bigrams from the product review or citation context (i.e., the sentence containing the citation). This follows previous studies in domain adaptation (Blitzer et al., 2007; Glorot et al., 2011). Chen et al. (2012) achieve state-of-the-art results on MDSD by testing the 5000 and 30,000 most frequent unigram and bigram features.

Previous work in citation classification has largely focused on identifying new features for improving classification accuracy. A significant amount of effort goes into engineering new features, in particular for identifying cue phrases, e.g., (Teufel et al., 2006b; Dong and Schäfer, 2011). However, there seems to be little consensus on which features help most for this task. For example, Abu-Jbara et al. (2013) and Jochim and Schütze (2012) find the list of polar words from Wilson et al. (2005) to be useful, and neither study lists dependency relations as significant features. Athar (2011) on the other hand reported significant improvement using dependency relation features and found that the same list of polar words slightly hurt classification accuracy. The classifiers and implementation of features varies between these studies, but the problem remains that there seems to be no clear set of features for citation polarity classification.

The lack of consensus on the most useful citation polarity features coupled with the recent success of deep learning neural networks (Collobert et al., 2011) further motivate our choice to limit our features to the n-grams available in the product review or citation context and not rely on external resources or tools for additional features.

3.3 Classification with mSDA

For classification we use marginalized stacked denoising autoencoders (mSDA) from Chen et al. (2012)[6] plus a linear SVM. mSDA takes the concept of *denoising* – introducing noise to make the autoencoder more robust – from Vincent et al. (2008), but does the optimization in closed form, thereby avoiding iterating over the input vector to stochastically introduce noise. The result of this is faster run times and currently state-of-the-art performance on MDSD, which makes it a good choice for our domain adaptation task. The mSDA implementation comes with LIBSVM, which we replace with LIBLINEAR (Fan et al., 2008) for faster run times with no decrease in accuracy. LIBLINEAR, with default settings, also serves as our baseline.

3.4 Outline of Experiments

Our initial experiments simply extend those of Chen et al. (2012) (and others who have used MDSD) by adding another domain, citations. We train on each of the domains from the MDSD –

tually contains positive and negative labels, which have been used, e.g., in (Chen et al., 2012).

[6]We use their MATLAB implementation available at `http://www.cse.wustl.edu/~mchen/code/mSDA.tar`.

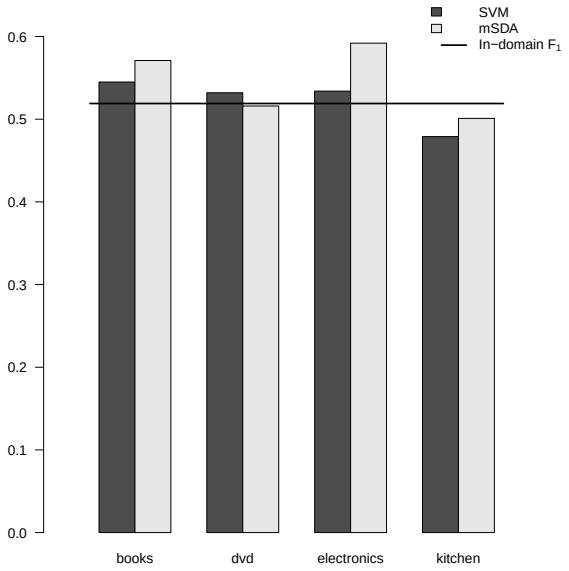

Figure 1: Cross domain macro-F_1 results training on Multi-Domain Sentiment Dataset and testing on citation dataset (CITD). The horizontal line indicates macro-F_1 for in-domain citation classification.

books, dvd, electronics, and kitchen – and test on the citation data. We split the labeled data 80/20 following Blitzer et al. (2007) (cf. Chen et al. (2012) train on all "labeled" data and test on the "unlabeled" data). These experiments should help answer two questions: does a larger amount of training data, even if out of domain, improve citation classification; and how well do the different product domains generalize to citations (i.e., which domains are most similar to citations)?

In contrast to previous work using MDSD, a lot of the work in domain adaptation also leverages a small amount of labeled target data. In our second set of experiments, we follow the domain adaptation approaches described in (Daumé III, 2007) and train on product review and citation data before testing on citations.

4 Results and Discussion

4.1 Citation mSDA

Our initial results show that using mSDA for domain adaptation to citations actually outperforms in-domain classification. In Figure 1 we compare citation classification with mSDA to the SVM baseline. Each pair of vertical bars represents training on a domain from MDSD (e.g., books) and testing on CITD. The dark gray bar indicates the F_1 scores for the SVM baseline using the 30,000 features and the lighter gray bar shows the mSDA results. The black horizontal line indicates the F_1 score for in-domain citation classification, which sometimes represents the goal for domain adaptation. We can see that using a larger dataset, even if out of domain, does improve citation classification. For books, dvd, and electronics, even the SVM baseline improves on in-domain classification. mSDA does better than the baseline for all domains except dvd. Using a larger training set, along with mSDA, which makes use of the unlabeled data, leads to the best results for citation classification.

In domain adaptation we would expect the domains most similar to the target to lead to the highest results. Like Dai et al. (2007), we measure the Kullback-Leibler divergence between the source and target domains' distributions. According to this measure, citations are most similar to the books domain. Therefore, it is not surprising that training on books performs well on citations, and intuitively, among the domains in the Amazon dataset, a book review is most similar to a scientific citation. This makes the good mSDA results for electronics a bit more surprising.

4.2 Easy Domain Adaptation

The results in Section 4.1 are for *semi-supervised* domain adaptation: the case where we have some large annotated corpus (Amazon product reviews) and a large unannotated corpus (citations). There have been a number of other successful attempts at *fully supervised* domain adaptation, where it is assumed that some small amount of data is annotated in the target domain (Chelba and Acero, 2004; Daumé III, 2007; Jiang and Zhai, 2007). To see how mSDA compares to supervised domain adaptation we take the various approaches presented by Daumé III (2007). The results of this comparison can be seen in Table 2. Briefly, "All" trains on source and target data; "Weight" is the same as "All" except that instances may be weighted differently based on their domain (weights are chosen on a development set); "Pred" trains on the source data, makes predictions on the target data, and then trains on the target data with the predictions; "LinInt" linearly interpolates predictions using the source-only and target-only models (the interpolation parameter is chosen on a development set); "Augment" uses a larger feature set with source-specific and target-specific copies of features; see

Domain	Baseline	All	Weight	Pred	LinInt	Augment	mSDA
books	54.5	54.8	52.0	51.9	53.4	53.4	**57.1**
dvd	53.2	50.9	**56.0**	53.4	51.9	47.5	51.6
electronics	53.4	49.0	50.5	53.4	54.8	51.9	**59.2**
kitchen	47.9	48.8	50.7	**53.4**	52.6	49.2	50.1
citations	51.9	–	–	–	–	–	**54.9**

Table 2: Macro-F_1 results on CITD using different domain adaptation approaches.

(Daumé III, 2007) for further details.

We are only interested in citations as the target domain. Daumé's source-only baseline corresponds to the "Baseline" column for domains: books, dvd, electronics, and kitchen; while his target-only baseline can be seen for citations in the last row of the "Baseline" column in Table 2.

The semi-supervised mSDA performs quite well with respect to the fully supervised approaches, obtaining the best results for books and electronics, which are also the highest scores overall. Weight and Pred have the highest F_1 scores for dvd and kitchen respectively. Daumé III (2007) noted that the "Augment" algorithm performed best when the target-only results were better than the source-only results. When this was not the case in his experiments, i.e., for the treebank chunking task, both Weight and Pred were among the best approaches. In our experiments, training on source-only outperforms target-only, with the exception of the kitchen domain.

We have included the line for citations to see the results training only on the target data ($F_1 = 51.9$) and to see the improvement when using all of the unlabeled data with mSDA ($F_1 = 54.9$).

4.3 Discussion

These results are very promising. Although they are not quite as high as other published results for citation polarity (Abu-Jbara et al., 2013)[7], we have shown that you can improve citation polarity classification by leveraging large amounts of annotated data from other domains and using a simple set of features.

mSDA and fully supervised approaches can also be straightforwardly combined. We do not present those results here due to space constraints. The combination led to mixed results: adding mSDA to the supervised approaches tended to improve F_1 over those approaches but results never exceeded the top mSDA numbers in Table 2.

[7]Their work included a CRF model to identify the citation context that gave them an increase of 9.2 percent F_1 over a single sentence citation context. Our approach achieves similar macro-F_1 on only the citation sentence, but using a different corpus.

5 Related Work

Teufel et al. (2006b) introduced automatic citation function classification, with classes that could be grouped as positive, negative, and neutral. They relied in part on a manually compiled list of cue phrases that cannot easily be transferred to other classification schemes or other scientific domains. Athar (2011) followed this and was the first to specifically target polarity classification on scientific citations. He found that dependency tuples contributed the most significant improvement in results. Abu-Jbara et al. (2013) also looks at both citation function and citation polarity. A big contribution of this work is that they also train a CRF sequence tagger to find the citation context, which significantly improves results over using only the citing sentence. Their feature analysis indicates that lexicons for negation, speculation, and polarity were most important for improving polarity classification.

6 Conclusion

Robust citation classification has been hindered by the relative lack of annotated data. In this paper we successfully use a large, out-of-domain, annotated corpus to improve the citation polarity classification. Our approach uses a deep learning neural network for domain adaptation with labeled out-of-domain data and unlabeled in-domain data. This semi-supervised domain adaptation approach outperforms the in-domain citation polarity classification and other fully supervised domain adaptation approaches.

Acknowledgments. We thank the DFG for funding this work (SPP 1335 *Scalable Visual Analytics*).

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of NAACL-HLT*, pages 596–606.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of NAACL-HLT*, pages 597–601.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of ACL Student Session*, pages 81–87.

Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of LREC*, pages 1755–1759.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, pages 440–447.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy capitalizer: Little data can help a lot. In *Proceedings of EMNLP*, pages 285–292.

Minmin Chen, Zhixiang Eddie Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of ICML*, pages 767–774.

Daryl E. Chubin and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5:423–441.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Transferring naive bayes classifiers for text classification. In *AAAI*, pages 540–545.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263.

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of IJCNLP*, pages 623–631.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Eugene Garfield. 1955. Citation indexes to science: A new dimension in documentation through association of ideas. *Science*, 122:108–111.

Eugene Garfield. 1964. Can citation indexing be automated? In *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings*, pages 189–192.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of ICML*, pages 513–520.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL*, pages 264–271.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING*, pages 1343–1358.

Bing Liu. 2010. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group.

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *Proceedings of IJCAI*, pages 926–931.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of COLING*, pages 689–696.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceedings of CIKM*, pages 213–222.

Henry G. Small and Belver C. Griffith. 1974. The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, 4(1):17–40.

Henry Small and Richard Klavans. 2011. Identifying scientific breakthroughs by combining co-citation analysis and citation context. In *Proceedings of International Society for Scientometrics and Informetrics*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, pages 151–161.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006a. An annotation scheme for citation function. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, pages 80–87.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006b. Automatic classification of citation function. In *Proceedings of EMNLP*, pages 103–110.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML*, pages 1096–1103.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354.