Sample Selection Guided by Domain and Task for Cross-Domain Targeted Sentiment Analysis

Kasturi Bhattacharjee[1], **Rashmi Gangadharaiah**[1], and **Smaranda Muresan**[1, 2]

[1]AWS AI
[2]Columbia University
{kastb,rgangad,smaranm}@amazon.com

Abstract

Building supervised targeted sentiment analysis models for a new target domain requires substantial annotation effort since most datasets for this task are domain-specific. Domain adaptation for this task has two dimensions: the nature of targets and the opinion words used to describe sentiment towards the target. We present a data sampling strategy informed by domain differences across these two dimensions with the goal of selecting a small number of examples, thereby minimizing annotation effort. This obtains performance in the 86-100% range compared to the full supervised model using only ∼4-15% of the full training data.

1 Introduction

Targeted sentiment analysis aims to detect sentiments towards specific targets in a given document. For instance, in a sentence from a Restaurant review *"Yes, they use fancy **ingredients**, but even fancy ingredients don't make for good **pizza** unless someone knows how to get the **crust** right."*, the target **ingredients** has a Positive sentiment while **pizza** and **crust** have Negative sentiments. A key challenge for this task is that domain differences manifest themselves in terms of target types as well as the choice of opinion words used to express the sentiments towards those targets. For Restaurant reviews, one is likely to find targets related to *food*, *ambience* or *service* with corresponding sentiment expressions such as *delicious*, *fancy*, *attentive*, etc. A Laptop review, on the other hand, is likely to have references to *software* and sentiment expressions such as *fast*, *easy*, etc. Obtaining fine-grained sentiment annotations for specific spans of text is often time-consuming, expensive and requires domain expertise. Thus, we often encounter scenarios where we have labeled data from one or more *source* domains but none or very little labeled data from a new and *different target* domain of interest. In this paper, we propose a novel data sampling strategy for cross-domain targeted sentiment analysis that takes advantage of the two dimensions of domain differences for this task: *targets* and *sentiment expressions*. We experiment on three labeled datasets in English that vary in domain - SemEval 2016 Task 5 (Pontiki et al., 2016) containing restaurant reviews **(R)**, SemEval 2014 Task 4 (Pontiki et al., 2014) containing laptop reviews **(L)** and a Twitter dataset **(T)** (Dong et al., 2014), which contains tweets about celebrities (*Lady Gaga*), products (*Windows 7*), and companies (*Google*). Positive, Negative and Neutral are the sentiment labels considered. Table 1 presents the data statistics.

2 Methodology & Experiments

Our proposed selection strategy aims to pick examples that are *informative* and *representative* of the target domain. Informativeness is a commonly used criteria for active-learning techniques (Wang et al., 2017; Settles and Craven, 2008; McCallum and Nigam, 1998), for which we apply an uncertainty-based sampling method that uses **entropy** (Wang et al., 2017; Wang and Shang, 2014; Settles, 2009). Let D_s and D_t represent the training data for the *source* and *target* domains respectively. For each document in D_t, we predict the probability distribution over the 3 sentiment labels for *each target*, using a model trained on D_s, and compute the entropy per target prediction. The *average entropy* across all targets of the document indicates the *overall uncertainty* for the document, thereby discovering hard-to-classify documents for the model.

Split	# Docs	# Pos, Neg, Neu spans
R-Train	1103	1107 397 61
R-Val	131	129 41 8
R-Test	420	468 114 30
L-Train	1320	884 786 434
L-Val	146	110 84 30
L-Test	411	341 128 169
T-Train	5588	1420 1392 2776
T-Val	659	141 168 350
T-Test	691	173 173 345

Table 1: Dataset stats. **R**=SemEval 2016 Restaurant Reviews, **L**=SemEval 2014 Laptop Reviews, **T**=Twitter. **Pos**=Positive, **Neg**=Negative and **Neu**=Neutral sentiments.

Setting	Highest RS scoring words
R→L	*easy*, *new*, *other*, *same*, *many*, *perfect*
L→R	*good*, *delicious*, *friendly*, *attentive*, *romantic*
L→T	*new*, *real*, *bad*, *last*, *famous*, *dead*

Table 2: Words with highest Relative Salience (RS) scores for each cross-domain setting.

Setting	Sampling Strategy	Sample Documents Picked
L→R	Relative Salience	Be sure to try the seasonal, and always *delicious*, specials.
	Entropy	I had **Lobster Bisque** it has 2 oz. of Maine Lobster in it.
R→L	Relative Salience	I like how the Mac OS is so simple and *easy* to use.
	Entropy	pros: the macbook pro notebook has a large **battery life** and you wont have to worry to charge your laptop every five hours or so.
L→T	Relative Salience	"Sonny helped me grow, and become more aware of the media, and paparazzi, and the *famous* life. It makes me think twice." - demi lovato.
	Entropy	Gorbachev's 80th birthday was a huge success! among the guests were **arnold schwarzenegger** , Sharon Stone and Kevin Spacey. Exciting!

Table 3: Examples selected by RS-based and Entropy-based sampling for various cross-domain settings. *Italics* shows sentiment expressions for which RS is computed, while **bold** shows the targets picked by the Entropy-based method.

Further, we use **Relative Salience (RS)** (Mohammad, 2011) to detect sentiment expressions that are more *representative* of the target domain when compared to source domain. Based on the simplifying assumption that sentiments towards targets are expressed through adjectives, we first extract all adjectives for each dataset using a POS tagger. For each cross-domain experiment, we compute the RS of an adjective w as, $RS(w|D_s, D_t) = f_t/N_t - f_s/N_s$, where, f represents the frequency of occurrence of w in the training data, while N represents the total number of words in the training data. The subscripts s and t stand for *source* and *target* respectively. Note that labels are not considered for this, just the raw documents. Higher the RS score, more the importance of the corresponding sentiment expression in the target domain w.r.t. the source domain. Documents containing any of the top 10 adjectives with the highest RS score from the target training set are selected to form the RS set.

Our proposed method of sampling involves selecting documents collected from *both* the Relative Salience and Entropy-based methods in different proportions for model training. Given the number of documents (*n*) we wish to sample, the various combinations we experiment with include selecting 50%-50%, 30%-70% and 20%-80% from RS and entropy-based strategies, respectively. Depending on the combination, we first pick the top k documents ordered from highest to lowest entropy score, followed by the remaining number of documents picked from the RS set. We experiment with a varying number *n* of sampled documents, starting with a small value (25 documents for Laptops and Restaurants, and 50 for Twitter) and going up to $\sim$15% of the training data for our experiments.

SemEval datasets both consist of reviews in two different domains (restaurants and laptops). For our experiments, we explore both (**R→L**) and (**L→R**) as cross-domain settings. Further, we use the Twitter dataset that is different in genre to both L and R, and choose **L→T** as the cross-domain setting. Table 3 provides a few document samples picked by RS and Entropy. As expected, the RS method picks examples containing sentiment expressions that are more relevant to the target domain. With **L** $\rightarrow$ **R**, we see sentiment expressions such as *friendly*, *delicious* and *romantic* that are more representative of the Restaurant domain (Table 2). Meanwhile, the Entropy-based approach selects examples that the model is most uncertain about. For example, targets such as **Lobster bisque** are unlikely to be present in the Laptops domain and result in the model's uncertainty in predictions.

Setting	Samples	Entropy	RS+Entropy
R→L	**Price** was higher when purchased on MAC when compared to price showing on PC when I bought this product.	Neutral	Negative
L→R	Nice **ambience**, but highly overrated place.	Neutral	Positive
L→T	Quality night , amazing costumes but got ta say **lady gaga** was the best though.. poor gaga left shoes and phone in my car ha	Negative	Positive

Table 4: **Targets** from test set that were *incorrectly* labeled by model trained using entropy-based sampled data, but were *correctly* predicted by model trained using the RS+Entropy sampled data.

The underlying BERT-based (Devlin et al., 2019) model for targeted sentiment classification accepts as input the entire document and target spans with boundaries, that are used to pool tokens to form a span representation. Using span representation and document as context, we perform multi-class classification to predict sentiment for each target, by minimizing cross-entropy loss across sentiment labels. We first train the model on labeled training data of the *source* domain. Target domain documents are then sampled using our proposed sampling method, and are used to further train the model. We experiment with a varying number of sampled documents going up to ∼15% of the training data. Model performance on target domain is reported using Macro F1. **Random** baseline includes selecting documents at random, while **Entropy** baseline involves selecting top *n* documents using only entropy-based sampling.

3 Results

As shown in Table 5, RS+Entropy strategy outperforms both baselines for each sentiment class, across all cross-domain settings. Additionally, as shown in Table 6 we are able to achieve performance in the 86-100% range compared to the fully supervised model using only ∼4-15% of the full training data. In Table 4, we show examples of targets for which model trained on Entropy-based sampled data makes errors in prediction, while model trained on RS+Entropy sampled data predicts correctly.

Setting	Sampling Strategy	Pos F1	Neg F1	Neu F1
R→L	RS+Entropy	**85.03**	**72.30**	**52.08**
	Entropy	83.92	71.97	48.20
	Random	82.66	71.92	39.84
L→R	RS+Entropy	**94.34**	**77.64**	**28.00**
	Entropy	94.27	78.71	20.00
	Random	92.04	71.89	00.00
L→T	RS+Entropy	**58.39**	**62.91**	**71.37**
	Entropy	53.51	60.06	70.26
	Random	55.11	59.51	69.85

Table 5: F1 across each sentiment class using proposed method and baselines

Setting	% of Supervised Model Macro F1	%Training Data
R→L	100	∼4
L→T	92.26	∼11
L→R	86.68	∼15

Table 6: Comparison with fully supervised setting

4 Conclusion

We propose a data sampling strategy for cross-domain targeted sentiment analysis that selects examples based on the two dimensions of domain differences for the task - targets and sentiment expressions. The proposed method combining Relative Salience and Entropy based sampling, when applied to three different cross-domain settings, is able to extract samples that are both informative and representative of the target domain. This helps the model achieve 86-100% of fully supervised performance using only 4-15% of the full training data, thus helping to reduce annotation cost. Further, it outperforms random and entropy-based baselines both in label-wise and overall model performance.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland, June. Association for Computational Linguistics.

A. McCallum and K. Nigam. 1998. Employing EM and Pool-Based Active Learning for Text Classification. In *ICML*.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119.

Gaoang Wang, Jenq-Neng Hwang, Craig Rose, and Farron Wallace. 2017. Uncertainty sampling based active learning with diversity constraint by sparse selection. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6.