Domain and Task-Informed Sample Selection for Cross-Domain Target-based Sentiment Analysis

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

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

Abstract

A challenge for target-based sentiment analysis is that most datasets are domain-specific and thus building supervised models for a new target domain requires substantial annotation effort. Domain adaptation for this task has two dimensions: the nature of the targets (e.g., entity types, properties associated with entities, or arbitrary spans) and the opinion words used to describe the sentiment towards the target. We present a data sampling strategy informed by the difference between the target and source domains across these two dimensions (i.e., targets and opinion words) with the goal of selecting a small number of examples that would be hard to learn in the new target domain compared to the source domain, and thus good candidates for annotation. 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

Target-based sentiment analysis aims to detect sentiments associated with specific targets in a given document. For instance, in Table 1, the targets ***service***, ***decor***, ***food***, ***portions*** have positive sentiment whereas ***operating system*** and ***kim kardashian*** have a negative sentiment. 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. Current datasets vary in their types of targets such as entities of various types (e.g., *Person*, *Location*, *Organization*, *Food*), predefined aspect/property categories (e.g., *quality* and *price*) or arbitrary spans that can denote an event ("The *opening night* was a success"). For instance, as shown in Table 1, for Restaurant reviews, one is likely to find target spans that are related to *food* (***food***, ***portions***), *ambience* (***decor***) or ***service***. Tweets might contain *celebrity* references (***kim kardashian***) as targets, while a Laptop review is likely to have references to *software* (***operating system***). Moreover, sentiment expressions vary from domain-to-domain as well. As shown in Table 1, we encounter sentiment expressions such as *delicious* for Restaurants domain, *older* for Laptops domain, and *famous* for Twitter that contains sentiment towards people.

Domain	Examples
Restaurants	The **service** is *excellent*, the **decor** is *great*, and the **food** is *delicious* and comes in *large* **portions**.
Laptops	I have had another Mac, but it got slow due to an *older* **operating system**.
Twitter	No, twitter, I don't want to follow **kim kardashian** - why is she *famous* btw or Chris Brown.

Table 1: Target spans (in **bold**) and sentiment expressions (*italicized*) from Restaurant review (Pontiki et al., 2016), Laptop review (Pontiki et al., 2014), and Twitter dataset (Dong et al., 2014).

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 domains (*source domains*) but none or very little labeled data from a new and *different* domain of interest (*target domain*). In this paper, we focus on a novel data sampling strategy for cross-domain target-based sentiment analysis that does not require sentiment labels but just the targets. It takes advantage of the two dimensions of domain differences for this task: targets and sentiment expressions. Our goal is complementary to work on transfer learning for domain adaptation for

this task (Rietzler et al., 2020).

Our proposed selection strategy aims to pick examples that are *informative* and *representative* of the target domain. To capture informativeness, a commonly used criteria in active learning settings (Settles and Craven, 2008; McCallum and Nigam, 1998), we use entropy-based sampling (Wang et al., 2017; Wang and Shang, 2014; Settles, 2009). This helps us sample examples that the model is most uncertain about in its sentiment predictions for given *targets*. Although entropy-based sampling is popular in active learning settings, to the best of our knowledge, it has not been applied to the task of sample selection for cross-domain targeted sentiment analysis. Further, we use Relative Salience (Mohammad, 2011) to pick examples containing *sentiment expressions* that are more *representative* of the target domain w.r.t the source domain. The efficacy of our data sampling strategy is tested by comparing the performance of the trained models on the sampled data against models trained on strong baselines such as entropy-based sampling (Section 3). Our proposed sampling strategy achieves performance in the 86-100% range compared to the full supervised model using only ∼4-15% of the full training data.

2 Datasets

We use three labeled datasets in English for target-based sentiment analysis 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)** introduced by Dong et al., which contains tweets about celebrities (*Britney Spears*, *Lady Gaga*), products (*xbox*, *Windows 7*), and companies (*Google*). A *document* for R and L refers to a sentence of a review, with most documents containing a single target, and some containing multiple targets as well (30% of R-train, 38% of L-train). A tweet is a document for T, with each of them containing a single target. **R** and **T** contain Positive, Negative and Neutral sentiment labels for the target spans while **L** contains *Conflict* as a sentiment label. To maintain parity with **R** and **T**, we drop the conflict label from **L**. We retain the original train-test splits for all 3 datasets. Additionally, we sample 10% of the training data at random for a validation set.

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 2: 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 3: Words with highest Relative Salience (RS) scores for each cross-domain setting.

3 Methodology

Entropy-based Sampling. In order to sample documents that contain hard-to-classify spans from the target domain, we use an uncertainty-based sampling method, that uses entropy (Shannon, 1948) to discover documents containing *targets* the model is uncertain about. 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. This aims to select documents based on informativeness.

Relative Salience (RS) based Sampling. We use Relative Salience (Mohammad, 2011) as a way to extract sentiment expressions that are more *representative* of the target domain when compared to the source domain. Based on the simplifying assumption that sentiment towards target spans are expressed through adjectives, we first extract all adjectives for each dataset using a Parts-of-Speech 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

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

t stand for *source* and *target* respectively. Note that labels are not considered for this, just the raw documents. Thus, RS score of a sentiment expression captures its importance in the target domain, w.r.t the source domain (see examples in Table 3). For each cross-domain scenario, we select documents from the target training set that contain any of the top 10 adjectives with the highest RS score.

RS+Entropy Sampling. 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 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. In Table 4, we provide 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 (source) → R (target), we see sentiment expressions such as *friendly*, *delicious* and *romantic* that are more representative of the Restaurant domain (see Table 3). 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. A similar behavior is observed with R→L and L→T.

4 Model & Experimental Setup

The underlying model we use for target-based sentiment classification is a BERT model (Devlin et al., 2019). The model accepts as input the entire document and target spans with boundaries. The document is first encoded by BERT and span boundaries are used to pool tokens to form a span representation. Using span representation and the document as context, we perform multi-class classification to predict the sentiment for each span, by minimizing cross-entropy loss across sentiment labels.

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

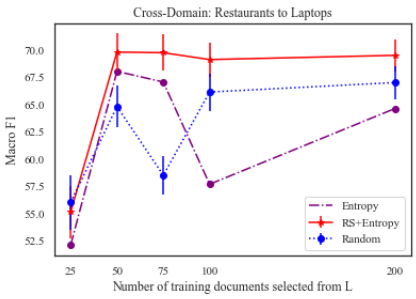
Table 5: F1 for each sentiment class obtained using various sampling strategies.

Experimental Setup & Baselines. SemEval datasets both consists 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.

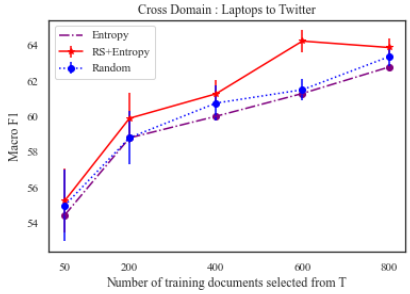
We first train the BERT model on *labeled* training data of the source domain. Documents from the target domain are then sampled using our proposed sampling method which is used to train the model. Model performance on target domain is reported using Macro F1. We experiment with a varying number n of sampled documents, starting with a small value (25 documents for Laptops and Restaurants, and 50 for Twitter) and going up to ∼15% of the training data for our experiments. Our baselines includes selecting a subset of n documents from the target domain at random as well as selecting the top n using entropy-based sampling only. For each experiment, we use the corresponding validation set for hyper-parameter optimization.

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

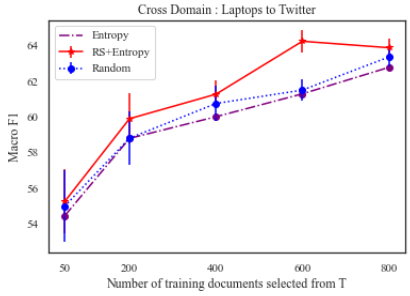
Table 6: **Targets** from test set that were *incorrectly* labeled by model trained using entropy-based sampled data, but were *correctly* predicted by model trained using the RS+Entropy sampled data.



(a)



(b)



(c)

Figure 1: F1 on the corresponding test sets (a) Laptops for R→L (b) Restaurants for L→R (c) Twitter for L→T.

5 Results

Figure 1 shows the mean Macro F1 scores (with standard deviation over 3 runs) for all three cross-domain settings with various sizes of sampled data. We find our proposed method to outperform both baselines for each cross-domain setting. In addition, Table 7 represents the amount of sampled data used by the model for training in these cross-domain settings and corresponding Macro F1 achieved as compared to a model trained with the full labeled training data. For R→L, we achieve 100% of Macro F1 as compared to the fully supervised case with only ∼4% of the training documents (4% of training instances). For L→T, we obtain 92.26% of the supervised setting with ∼11% of the training documents (∼11% of training instances). For L→R, our proposed method achieves within ∼86.68% of the fully supervised setting with ∼15% of the training documents (∼15% of training instances). Further, as shown in Table 5, RS+Entropy strategy outperforms both Entropy and Random baselines for each class, across all cross-domain settings.

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

Table 7: Comparison with fully supervised setting.

Error Analysis In Table 6, we show examples of targets for each cross-domain setting for which the model trained on Entropy-based sampled data makes errors in prediction, while model trained on RS+Entropy sampled data predicts correctly.

6 Conclusion

We propose a data sampling strategy for cross-domain target-based 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. 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. 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. 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. 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. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

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

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, Honolulu, Hawaii. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

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.