Is Machine Translation Ripe for Cross-lingual Sentiment Classification?

Kevin Duh and **Akinori Fujino** and **Masaaki Nagata**
NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Kyoto 619-0237, JAPAN
{kevin.duh,fujino.akinori,nagata.masaaki}@lab.ntt.co.jp

Abstract

Recent advances in Machine Translation (MT) have brought forth a new paradigm for building NLP applications in low-resource scenarios. To build a sentiment classifier for a language with no labeled resources, one can translate labeled data from another language, then train a classifier on the translated text. This can be viewed as a domain adaptation problem, where labeled translations and test data have some mismatch. Various prior work have achieved positive results using this approach.

In this opinion piece, we take a step back and make some general statements about cross-lingual adaptation problems. First, we claim that domain mismatch is not caused by MT errors, and accuracy degradation will occur even in the case of perfect MT. Second, we argue that the cross-lingual adaptation problem is qualitatively different from other (monolingual) adaptation problems in NLP; thus new adaptation algorithms ought to be considered. This paper will describe a series of carefully-designed experiments that led us to these conclusions.

1 Summary

Question 1: If MT gave perfect translations (semantically), do we still have a domain adaptation challenge in cross-lingual sentiment classification?

Answer: Yes. The reason is that while many translations of a word may be valid, the MT system might have a systematic bias. For example, the word "awesome" might be prevalent in English reviews, but in translated reviews, the word "excellent" is generated instead. From the perspective of MT, this translation is correct and preserves sentiment polarity. But from the perspective of a classifier, there is a domain mismatch due to differences in word distributions.

Question 2: Can we apply standard adaptation algorithms developed for other (monolingual) adaptation problems to cross-lingual adaptation?

Answer: No. It appears that the interaction between target unlabeled data and source data can be rather unexpected in the case of cross-lingual adaptation. We do not know the reason, but our experiments show that the accuracy of adaptation algorithms in cross-lingual scenarios have much higher variance than monolingual scenarios.

The goal of this opinion piece is to argue the need to better understand the characteristics of domain adaptation in cross-lingual problems. We invite the reader to disagree with our conclusion (that the true barrier to good performance is not insufficient MT quality, but inappropriate domain adaptation methods). Here we present a series of experiments that led us to this conclusion. First we describe the experiment design (§2) and baselines (§3), before answering Question 1 (§4) and Question 2 (§5).

2 Experiment Design

The cross-lingual setup is this: we have labeled data from source domain S and wish to build a sentiment classifier for target domain T. Domain mismatch can arise from *language differences* (e.g. English vs. translated text) or *market differences* (e.g. DVD vs. Book reviews). Our experiments will involve fixing

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 429–433,
Portland, Oregon, June 19-24, 2011.

T to a common testset and varying S. This allows us to experiment with different settings for adaptation.

We use the Amazon review dataset of Prettenhofer (2010)[1], due to its wide range of languages (English [EN], Japanese [JP], French [FR], German [DE]) and markets (music, DVD, books). Unlike Prettenhofer (2010), we reverse the direction of cross-lingual adaptation and consider English as target. English is not a low-resource language, but this setting allows for more comparisons. Each source dataset has 2000 reviews, equally balanced between positive and negative. The target has 2000 test samples, large unlabeled data (25k, 30k, 50k samples respectively for Music, DVD, and Books), and an additional 2000 labeled data reserved for oracle experiments. Texts in JP, FR, and DE are translated word-by-word into English with Google Translate.[2]

We perform three sets of experiments, shown in Table 1. Table 2 lists all the results; we will interpret them in the following sections.

	Target (T)	Source (S)
1	Music-EN	Music-JP, Music-FR, Music-DE, DVD-EN, Book-EN
2	DVD-EN	DVD-JP, DVD-FR, DVD-DE, Music-EN, Book-EN
3	Book-EN	Book-JP, Book-FR, Book-DE, Music-EN, DVD-EN

Table 1: Experiment setups: Fix T, vary S.

3 How much performance degradation occurs in cross-lingual adaptation?

First, we need to quantify the accuracy degradation under different source data, *without* consideration of domain adaptation methods. So we train a SVM classifier on labeled source data[3], and directly apply it on test data. The oracle setting, which has no domain-mismatch (e.g. train on Music-EN, test on Music-EN), achieves an average test accuracy of $(81.6 + 80.9 + 80.0)/3$ = 80.8%[4]. Average cross-lingual accuracies are: 69.4% (JP), 75.6% (FR), 77.0% (DE), so degradations compared to oracle are: -11% (JP), -5% (FR), -4% (DE).[5] Cross-market degradations are around -6%[6].

Observation 1: Degradations due to market and language mismatch are comparable in several cases (e.g. MUSIC-DE and DVD-EN perform similarly for target MUSIC-EN). **Observation 2:** The ranking of source language by decreasing accuracy is DE $>$ FR $>$ JP. Does this mean JP-EN is a more difficult language pair for MT? The next section will show that this is not necessarily the case. Certainly, the domain mismatch for JP is larger than DE, but this could be due to phenomenon other than MT errors.

4 Where exactly is the domain mismatch?

4.1 Theory of Domain Adaptation

We analyze domain adaptation by the concepts of labeling and instance mismatch (Jiang and Zhai, 2007). Let $p_t(x, y) = p_t(y|x)p_t(x)$ be the target distribution of samples x (e.g. unigram feature vector) and labels y (positive / negative). Let $p_s(x, y) = p_s(y|x)p_s(x)$ be the corresponding source distribution. We assume that one (or both) of the following distributions differ between source and target:

- Instance mismatch: $p_s(x) \neq p_t(x)$.
- Labeling mismatch: $p_s(y|x) \neq p_t(y|x)$.

Instance mismatch implies that the input feature vectors have different distribution (e.g. one dataset uses the word "excellent" often, while the other uses the word "awesome"). This degrades performance because classifiers trained on "excellent" might not know how to classify texts with the word "awesome." The solution is to tie together these features (Blitzer et al., 2006) or re-weight the input distribution (Sugiyama et al., 2008). Under some assumptions (i.e. covariate shift), oracle accuracy can be achieved theoretically (Shimodaira, 2000).

Labeling mismatch implies the same input has different labels in different domains. For example, the JP word meaning "excellent" may be mistranslated as "bad" in English. Then, positive JP

[1] http://www.webis.de/research/corpora/webis-cls-10

[2] This is done by querying foreign words to build a bilingual dictionary. The words are converted to tfidf unigram features.

[3] For all methods we try here, 5% of the 2000 labeled source samples are held-out for parameter tuning.

[4] See column EN of Table 2, Supervised SVM results.

[5] See "Adapt by Language" columns of Table 2. Note JP+FR+DE condition has 6000 labeled samples, so is not directly comparable to other adaptation scenarios (2000 samples). Nevertheless, mixing languages seem to give good results.

[6] See "Adapt by Market" columns of Table 2.

Target	Classifier	Oracle	Adapt by Language				Adapt by Market		
		EN	JP	FR	DE	JP+FR+DE	MUSIC	DVD	BOOK
MUSIC-EN	Supervised SVM	**81.6**	68.5	**75.2**	76.3	**80.3**	-	76.8	74.1
	Adapted TSVM	79.6	**73.0**	74.6	**77.9**	78.6	-	**78.4**	**75.6**
DVD-EN	Supervised SVM	80.9	70.1	**76.4**	**77.4**	**79.7**	**75.2**	-	74.5
	Adapted TSVM	**81.0**	**71.4**	75.5	76.3	78.4	74.8	-	**76.7**
BOOK-EN	Supervised SVM	80.0	69.6	75.4	**77.4**	**79.9**	73.4	76.2	-
	Adapted TSVM	**81.2**	**73.8**	**77.6**	76.7	79.5	**75.1**	**77.4**	-

Table 2: Test accuracies (%) for English Music/DVD/Book reviews. Each column is an adaptation scenario using different source data. The source data may vary by language or by market. For example, the first row shows that for the target of Music-EN, the accuracy of a SVM trained on translated JP reviews (in the same market) is 68.5, while the accuracy of a SVM trained on DVD reviews (in the same language) is 76.8. "Oracle" indicates training on the same market *and* same language domain as the target. "JP+FR+DE" indicates the concatenation of JP, FR, DE as source data. Boldface shows the winner of Supervised vs. Adapted.

reviews will be associated with the word "bad": $p_s(y = +1|x = bad)$ will be high, whereas the true conditional distribution should have high $p_t(y = -1|x = bad)$ instead. There are several cases for labeling mismatch, depending on how the polarity changes (Table 3). The solution is to filter out these noisy samples (Jiang and Zhai, 2007) or optimize loosely-linked objectives through shared parameters or Bayesian priors (Finkel and Manning, 2009).

Which mismatch is responsible for accuracy degradations in cross-lingual adaptation?

- Instance mismatch: Systematic MT bias generates word distributions different from naturally-occurring English. (Translation may be valid.)
- Label mismatch: MT error mis-translates a word into something with different polarity.

Conclusion from §4.2 and §4.3: Instance mismatch occurs often; MT error appears minimal.

Mis-translated polarity	Effect
$\pm \rightarrow 0$ e.g. ("good" $\rightarrow$ "the")	Loose a discriminative feature
$0 \rightarrow \pm$ e.g. ("the" $\rightarrow$ "good")	Increased overlap in positive/negative data
$+ \rightarrow -$ and $- \rightarrow +$ e.g. ("good" $\rightarrow$ "bad")	Association with opposite label

Table 3: Label mismatch: mis-translating positive ($+$), negative ($-$), or neutral (0) words have different effects. We think the first two cases have graceful degradation, but the third case may be catastrophic.

4.2 Analysis of Instance Mismatch

To measure instance mismatch, we compute statistics between $p_s(x)$ and $p_t(x)$, or approximations thereof: First, we calculate a (normalized) average feature from all samples of source S, which represents the unigram distribution of MT output. Similarly, the average feature vector for target T approximates the unigram distribution of English reviews $p_t(x)$. Then we measure:

- KL Divergence between Avg(S) and Avg(T), where Avg() is the average vector.
- Set Coverage of Avg(T) on Avg(S): how many word (type) in T appears at least once in S.

Both measures correlate strongly with final accuracy, as seen in Figure 1. The correlation coefficients are $r = -0.78$ for KL Divergence and $r = 0.71$ for Coverage, both statistically significant ($p < 0.05$). This implies that instance mismatch is an important reason for the degradations seen in Section 3.[7]

4.3 Analysis of Labeling Mismatch

We measure labeling mismatch by looking at differences in the weight vectors of oracle SVM and adapted SVM. Intuitively, if a feature has positive weight in the oracle SVM, but negative weight in the adapted SVM, then it is likely a MT mis-translation

[7]The observant reader may notice that cross-market points exhibit higher coverage but equal accuracy (74-78%) to some cross-lingual points. This suggests that MT output may be more constrained in vocabulary than naturally-occurring English.

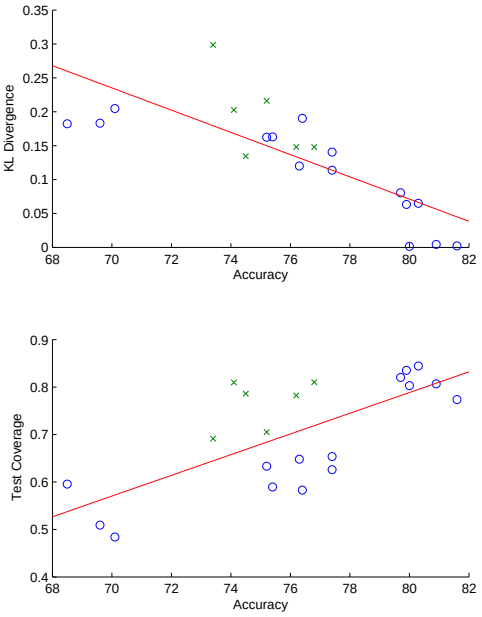

Figure 1: KL Divergence and Coverage vs. accuracy. (o) are cross-lingual and (x) are cross-market data points.

is causing the polarity flip. Algorithm 1 (with K=2000) shows how we compute polarity flip rate.[8]

We found that the polarity flip rate does not correlate well with accuracy at all ($r = 0.04$). **Conclusion:** Labeling mismatch is *not* a factor in performance degradation. Nevertheless, we note there is a surprising large number of flips (24% on average). A manual check of the flipped words in BOOK-JP revealed few MT mistakes. Only 3.7% of 450 random EN-JP word pairs checked can be judged as blatantly incorrect (without sentence context). The majority of flipped words do not have a clear sentiment orientation (e.g. "amazon", "human", "moreover").

5 Are standard adaptation algorithms applicable to cross-lingual problems?

One of the breakthroughs in cross-lingual text classification is the realization that it can be cast as domain adaptation. This makes available a host of pre-existing adaptation algorithms for improving over supervised results. However, we argue that it may be

[8]The feature normalization in Step 1 is important to ensure that the weight magnitudes are comparable.

Algorithm 1 Measuring labeling mismatch

Input: Weight vectors for source w_s and target w_t
Input: Target data average sample vector avg(T)
Output: Polarity flip rate f

1: Normalize: w_s = avg(T) * w_s; w_t = avg(T) * w_t
2: Set S_+ = { K most positive features in w_s}
3: Set S_- = { K most negative features in w_s}
4: Set T_+ = { K most positive features in w_t}
5: Set T_- = { K most negative features in w_t}
6: **for** each feature $i \in T_+$ **do**
7: if $i \in S_-$ then $f = f + 1$
8: **end for**
9: **for** each feature $j \in T_-$ **do**
10: if $j \in S_+$ then $f = f + 1$
11: **end for**
12: $f = \frac{f}{2K}$

better to "adapt" the standard adaptation algorithm to the cross-lingual setting. We arrived at this conclusion by trying the adapted counterpart of SVMs off-the-shelf. Recently, (Bergamo and Torresani, 2010) showed that Transductive SVMs (TSVM), originally developed for semi-supervised learning, are also strong adaptation methods. The idea is to train on source data like a SVM, but encourage the classification boundary to divide through low density regions in the unlabeled target data.

Table 2 shows that TSVM outperforms SVM in all but one case for cross-market adaptation, but gives mixed results for cross-lingual adaptation. This is a puzzling result considering that both use the *same* unlabeled data. Why does TSVM exhibit such a large variance on cross-lingual problems, but not on cross-market problems? Is unlabeled target data interacting with source data in some unexpected way?

Certainly there are several successful studies (Wan, 2009; Wei and Pal, 2010; Banea et al., 2008), but we think it is important to consider the possibility that cross-lingual adaptation has some fundamental differences. We conjecture that adapting from artificially-generated text (e.g. MT output) is a different story than adapting from naturally-occurring text (e.g. cross-market). In short, MT *is* ripe for cross-lingual adaptation; what is not ripe is probably our understanding of the special characteristics of the adaptation problem.

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alessandro Bergamo and Lorenzo Torresani. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in Neural Information Processing Systems (NIPS)*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jenny Rose Finkel and Chris Manning. 2009. Hierarchical Bayesian domain adaptation. In *Proc. of NAACL Human Language Technologies (HLT)*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proc. of the Association for Computational Linguistics (ACL)*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proc. of the Association for Computational Linguistics (ACL)*.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inferenc*, 90.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4).

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proc. of the Association for Computational Linguistics (ACL)*.

Bin Wei and Chris Pal. 2010. Cross lingual adaptation: an experiment on sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*.