Cross-Cultural Transfer Learning for Text Classification

Dor Ringel1  Gal Lavee2  Ido Guy2  Kira Radinsky1
1 Technion - Israel Institute of Technology
2 eBay Research
{dorringel,kirar}@cs.technion.ac.il
idoguy@acm.org glavee@eBay.com

Abstract
Large training datasets are required to achieve competitive performance in most natural language tasks. The acquisition process for these datasets is labor intensive, expensive, and time consuming. This process is also prone to human errors. In this work, we show that cross-cultural differences can be harnessed for natural language text classification. We present a transfer-learning framework that leverages widely-available unaligned bilingual corpora for classification tasks, using no task-specific data. Our empirical evaluation on two tasks – formality classification and sarcasm detection – shows that the cross-cultural difference between German and American English, as manifested in product review text, can be applied to achieve good performance for formality classification, while the difference between Japanese and American English can be applied to achieve good performance for sarcasm detection – both without any task-specific labeled data.

1 Introduction
The collection of large, task-specific labeled datasets poses a major challenge to the application of supervised text classification in many domains. Acquiring these datasets is expensive, time-consuming, and error-prone.

In this work, we propose leveraging bilingual datasets for classification tasks. The large-scale availability of such datasets counteracts the low availability of task-specific labeled data. The cross-cultural differences expressed in these bilingual datasets can be learned and then transferred for text classification tasks with no labeled data. Our work, thus, extends the idea of distant supervision (Mintz et al., 2009) to tasks for which no relevant large-scale curated datasets exist.

Culture was defined by Hofstede et al. (2010) as “the collective programming of the mind, which distinguishes the members of one group of people from another.” Shweder et al. (2006) defines membership in a cultural group as “thinking and acting in a certain way, in the light of particular goals, values, pictures of the world.” Based on these definitions we can reasonably expect members of a specific cultural group to think and act in ways that are distinct from others. For example, Hedderich (2010), who investigated cross-cultural differences at the workplace, found that the overall interaction of employees in Germany is more formal than their American counterparts. House (1997) found that “German subjects tended to interact in ways that were more direct, more explicit, more self-referenced, and more content-oriented”.

Similarly, several works have studied cross-cultural differences between Japanese and English (both American and British variants) (Koga and Pearson, 1992; Minami, 1994; Martinsons, 2001; Adachi, 2010). Specifically, Adachi (1996) and Ziv (1988) investigated sarcasm in Japanese and the latter showed that American students are more sarcastic than their Japanese counterparts.

The main thesis underlying this work is that texts composed by members of a specific cultural group are distinct from those composed by members of another. Observational studies on cultural differences, such as those described in the examples above, allow us to identify the high-level semantics of these differences.

We present a transfer learning algorithm which learns a model encapsulating cross-cultural differences from bilingual data, and applies the learned model to text classification tasks. We study this idea using two natural language classification tasks: (1) Formality classification – by learning the difference between the writing style of Germans compared to Americans; and (2) Sarcasm detection – by learning the difference between the writing style of Japanese compared to Americans.
Our approach requires only a bilingual dataset without the need for any alignments, special labels, or task-specific training data. Our empirical evaluation demonstrates that such cross-cultural distinction can be successfully transferred to those tasks. We present an algorithm that, given two unaligned corpora of texts written in two languages, transforms the two document classes into a common representation. A binary classifier is then trained to distinguish between representation of documents originating from one language as compared to the other. The classifier can be subsequently applied to a binary text classification task. For example, if the document is deemed by the American-Japanese classifier as American, we will infer that the text is sarcastic.

We study various representations of both words and documents and present a novel document representation algorithm adapted to our task. Our empirical results suggest that our transfer-learning approach based on cross-cultural differences achieves comparable performance to direct learning approaches trained on task-specific labeled data. Additional results demonstrate the contribution of our proposed representation approach.

The contributions of this work can be summarized as follows: (1) We propose a transfer learning framework to enable text classification using cross-cultural differences learned on bilingual data; (2) We propose a representation scheme for documents, designed for the task of text classification based on bilingual data; (3) We perform an empirical evaluation of our approach and contribute our labeled datasets to the community.

2 Related Work

Cross-cultural differences have been studied extensively in the social science literature (Shweder, 1991; Weber and Hsee, 1998; Liu and McClure, 2001; Boroditsky et al., 2003; Yin et al., 2011; Hong et al., 2012). Specifically, Paul and Girju (2009); Garimella et al. (2016) explored differences through language analysis.

Text classification approaches have traditionally used distributed representation of texts (e.g. TF-IDF) and applied supervised models to these representations (Joachims, 1998; Wang and Manning, 2012). More recent work has pursued improved representations (Joulin et al., 2017) and novel neural architectures (Conneau et al., 2017).

Transfer learning in natural language processing has recently moved beyond pre-trained word embeddings (Howard and Ruder, 2018) to more advanced approaches (Peters et al., 2018; Radford et al., 2018).

Cross-lingual word embedding seeks a mapping between embedding spaces representing different languages (Ruder et al., 2019). This goal is most often achieved by training monolingual word embeddings for multiple languages independently, and then learning a transformation between them using either supervised or unsupervised methods.

Cross-lingual language understanding (XLU) is an area of research that has recently gained much attention. Related methods seek to learn a joint-embedding space for multiple languages (Conneau et al., 2018; Devlin et al., 2019).

This work, to the best of our knowledge, is the first to leverage cross-cultural differences in bilingual text data to perform inference on monolingual data.

3 Cross-Cultural Transfer Learning

In this section we describe our approach for text classification tasks such as formality classification and sarcasm detection, using transfer learning from bilingual data. A supervised learning model is trained to differentiate between a pair of languages based on their cross-cultural differences, as manifested in the available text data. Specifically, the training data contains a corpus of text documents collected from two distinct languages, with the assumption that the language of each document is known. Note that, the bilingual data needed for this approach is coarsely taken from the same domain. However, the texts need not be aligned beyond this coarse level. That is, full alignment between specific documents across the two languages is not required. Algorithm 1 summarizes the steps of our method.

<table>
<thead>
<tr>
<th>Algorithm 1 Cross-Cultural Transfer Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: $C_0$ (documents from language $L_0$)</td>
</tr>
<tr>
<td>Input: $C_1$ (documents from language $L_1$)</td>
</tr>
<tr>
<td>Input: $T$ (transformation method)</td>
</tr>
<tr>
<td>Input: $D_t$ (binary classification task dataset)</td>
</tr>
<tr>
<td>Output: $y_t$ (predictions)</td>
</tr>
<tr>
<td>1: $C^T_0$, $C^T_1$ ← TRANSFORM/Documents($T$, $C_0$, $C_1$)</td>
</tr>
<tr>
<td>2: $h$ ← TRAIN_CLASSIFIER($C^T_0$, $C^T_1$)</td>
</tr>
<tr>
<td>3: $D^T_t$ ← TRANSFORM/Documents($T$, $D_t$)</td>
</tr>
<tr>
<td>4: $\hat{y}_t$ ← TRANSFER_LEARNING($h$, $D^T_t$)</td>
</tr>
</tbody>
</table>

3872
3.1 Formal Framework

Given two collections of documents from languages $L_0$ and $L_1$ denoted as $C_0$ and $C_1$, respectively, the following are the steps to applying our framework:

1. Using a transformation, denoted $T$, transform the two document classes into a common representation. The resulting transformed collections of documents are denoted $C^T_0$ and $C^T_1$, respectively.

2. Train a binary classifier, denoted $h$, to discriminate document representations in $C^T_0$ from those in $C^T_1$. That is, the classifier is trained to classify representations of documents as originating from either $L_0$ or $L_1$.

3. Given a binary classification task dataset, denoted $D_t$, which consists of text documents (with a known language), transform all texts in $D_t$ using transformation $T$, to obtain a set of document representations denoted $D^T_t$.

4. Finally, apply the classifier $h$, trained on the bilingual data, to the document representations in $D^T_t$ to obtain predictions for each member of this set, denoted $\hat{y}_t$.

Figure 1 summarizes the process.

3.2 Transformations

The effectiveness of the framework described above depends in large part on the choice of transformation $T$. In general, transformation $T$ transforms documents from languages $L_0$ and $L_1$ into a common representation. A well chosen transformation should enable the supervised learning in subsequent steps to focus on cross-cultural differences between the original documents rather than superficial differences between the texts.

In our construction, we used compound transformations composed of the following parts: (1) Translation of the texts to a common language (Section 3.2.1); (2) Mapping the tokens in the text to word embeddings (Section 3.2.2); (3) Combining the word embeddings to form document-level vector representations (Section 3.2.3).

3.2.1 Translation

The first part of our document transformation involves translating the documents. We chose to translate documents from $L_0$ to $L_1$, so that documents authored in language $L_1$ (those documents denoted $C_1$ above), require no translation.

The motivation for choosing $L_1$ as the “target” language may involve the reliability of the translation to this language, the availability of high quality word embeddings in this language, or the fore-knowledge that downstream tasks will involve datasets written in this language.

Under this framework we evaluate two different translation approaches: (1) A state of the art machine translation system we denote as MT. Following preliminary experiments with both the method by Vaswani et al. (2017) and with Google cloud translation\footnote{https://cloud.google.com/translate/docs}, we focused on the latter as it achieved better results. (2) A word-by-word nearest-neighbor search, denoted as NN, where we use the method of Lample et al. (2018), adapted to our task. These choices represent a complexity-accuracy trade-off. We expect the MT system to give high-quality translation, but at the cost of higher complexity. Conversely, the NN approach is simple to implement, but may produce lower-quality translations.

In order to implement NN some notion of distance between words in language $L_0$ and words in $L_1$ is needed. We used cross-lingual word embeddings for this purpose. For a given language pair, the embedding of the joint vector space representation of many words from both languages is known. For implementing nearest neighbor search for a particular word in $L_0$, we simply compare the embedding vector representation of this word
with all known embedding vectors of words in language \( L_1 \), and select the one that maximizes their similarity, as suggested by Lample et al. (2018) (and done efficiently by employing the method of Johnson et al. (2019)). For translating German to English such an embedding dataset is available for download.\(^2\) For Japanese to English, we trained our own such model (following Lample et al. (2018)), as this language pair is not available for download. Note that the embeddings used in this step may or may not be related to the embeddings described in the next section.

### 3.2.2 Word Embeddings

Once our documents are converted to a single language using one of the translation methods above, we represent each word in the document as an embedding vector, which is a dense distributed version of the corresponding word.

We experiment with several types of word embeddings, including: pre-trained embedding models trained on large corpora of general English language text, publicly available for download (Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2018)).

### 3.2.3 Document Representation

After applying translation and mapping tokens to word embeddings, our documents are represented as a variable-length list of vectors (each corresponding to word). In the following we examine various methods for combining these vectors into a single uniform length representation, and derive a document representation scheme that uses label information to achieve better discrimination, which we conjecture increases the ability to encode the cross-cultural difference.

The methods we explore produce weighted combinations of the word vectors in the document, normalized to unit length. A model that captures the intuition that text is composed of both local and global context is proposed by Arora et al. (2017). Applying maximum likelihood estimation to this model obtains the representation:

\[
c_d = \sum_i \frac{a}{U(w_i)} + a \mathbf{v}_{w_i} \tag{1}
\]

where \(c_d \in \mathbb{R}^k\) is the resulting document representation vector and \(\mathbf{v}_{w_i} \in \mathbb{R}^k\) is the embedding vector corresponding to the \(i\)-th of \(n\) words in document \(d\); \(U(w)\) denotes the unigram distribution; and \(a > 0\) is a free parameter that controls smoothing. This equation expresses the intuitive approach of averaging the word embeddings, each weighted inversely to the word’s frequency. Thus, very popular words are down-weighted. This is also related to an inverse document frequency (IDF) weighting proposed in De Boom et al. (2016).

In our application we seek a document embedding to assist in solving a binary classification task. In other words, we seek a context vector \(c_d\) that is most discriminative, taking advantage of additional information about some binary label. More specifically, we assume that we have a conditional model for the probability of each token given the class.

Formalizing this idea, we seek an embedding that will maximize the expected log-probability of a document:

\[
\mathbb{E} \left[ \log \left( P(w_1, w_2, \ldots, w_n) \right) \right] = \mathbb{E} \left[ \log \prod_{i=1}^n P(w_i | c_d, L) \right] + C
\]

where we have used Bayes rule and \(C\) is constant w.r.t \(c_d\). As the notation suggests, we now have a generative model for each class:

\[
P(w_i | c_d, L_0) = \alpha \cdot P(w_i | L_0) + (1 - \alpha) \cdot \left( \frac{\exp \left( \mathbf{v}_{w_i}^\top c_d \right)}{Z} \right)
\]

\[
P(w_i | c_d, L_1) = \alpha \cdot P(w_i | L_1) + (1 - \alpha) \cdot \left( \frac{\exp \left( \mathbf{v}_{w_i}^\top c_d \right)}{Z} \right)
\]

where \(P(w_i | L_0)\) and \(P(w_i | L_1)\) are the conditional unigram models, and \(Z\) denotes the normalization constant.

Focusing on the probability of a single token:

\[
f(c_d) = \mathbb{E} \left[ \log P(w_i | c_d, L) \right] = \frac{1}{2} \left( \log P(w_i | c_d, L_0) + \log P(w_i | c_d, L_1) \right)
\]

under the assumption that \(P(L_0) = P(L_1) = \frac{1}{2}\).

This expression has the following gradient (w.r.t. \(c_d\)):

\[
\nabla f(c_d) = \frac{1}{2} \left( \frac{P(w_i | L_0) + P(w_i | L_1) + 2a \epsilon_i}{\left( P(w_i | L_0) + a \epsilon_i \right) \cdot \left( P(w_i | L_1) + a \epsilon_i \right)} \right) \mathbf{v}_{w_i} \tag{3}
\]

where \(a = \frac{(1 - \alpha)}{\alpha Z} \) and \(\epsilon_{i,d} = \exp \left( \mathbf{v}_{w_i}^\top c_d \right)\). Note that when we evaluate Equation (3) with \(c_d = 0\) the expression becomes:

\[
\nabla f(0) = \frac{1}{2a} \left[ \frac{P(w_i | 0) + P(w_i | 1) + 2a}{(P(w_i | 0) + a) \cdot (P(w_i | 1) + a)} \right] \mathbf{v}_{w_i}
\]
Thus, maximizing the Taylor approximation $f(c_d) \approx f(0) + \nabla f(0)^T c_d$ w.r.t. $c_d$ (following Arora et al. (2017)) yields the following estimator:

$$c_d = \frac{1}{2} \sum_{i=1}^{n} a \left[ \frac{P(w_i|L_0) + P(w_i|L_1) + 2a}{(P(w_i|L_0) + a)(P(w_i|L_1) + a)} \right] v_{w_i} \tag{4}$$

Examining the expression above, we can see that for a fixed value of $a$, the numerator of the ratio grows with the frequency of $w_i$ in either language $L_0$ or $L_1$. The denominator of the ratio grows when $w_i$ is frequent in both languages. Hence, the weighting scheme gives more weight to vectors corresponding to words that are frequent in either language, but not both. Further, similarly to Equation 1, the expression gives a larger weight to words that have low frequency in either language.

Note that the representations above define an unnormalized quantity, so that the absolute values of the vector weights are less important than their relative values. The final document representation is given by $c_d / ||c_d||$.

4 Evaluation

In order to evaluate our setup, we consider the application of the above framework to two binary text classification tasks: formality classification and sarcasm detection. For each such task we require a bilingual dataset containing texts drawn from two languages for training and another (one or more) dataset to use for evaluation. Table 1 summarizes the characteristics of both the bilingual and evaluation datasets.

4.1 Formality Classification Task

The formality classification task (Heylighen and Dewaele, 1999) aims to classify a document as either formal or informal. For evaluating this task we used the following datasets:

**Amazon Product Descriptions and Reviews**

Product descriptions and their corresponding user reviews from the Motors and Fashion categories of Amazon e-commerce website (He and McAuley, 2016). We consider the description texts to be examples of formal writing and the reviews to be examples of informal writing (Novgorodov et al., 2019). We filter documents that are not in English or shorter than 10 terms. We then sample one review per description document to result with a balanced dataset in terms of class labels.

New York Times Article snippets and Comments

Article-snippets and their corresponding user comments from the New York Times (Keswani, 2018). We consider article-snippet texts to be examples of formal writing and the comment texts to be examples of informal writing. We focus on the news section, and filter documents shorter than 10 terms. We then sample one comment per article-snippet to result with a balanced dataset in terms of class labels.

**Formality in Online Communication**

Texts annotated for formality, originating from four types of online communication: News, Blog, Email, and community question answering forums (denoted as Answers) (Lahiri, 2015; Pavlick and Tetreault, 2016). The mean formality score for each document (across all annotators) ranges from -3 (very informal) and 3 (very formal). To make this dataset suitable for our binary classification framework, we only consider documents with a mean formality score in the highest and lowest 20th percentile per communication type as formal and informal, respectively.

**Surrogate Bilingual Dataset for Formality**

Based on prior research by Hedderich (2010), who investigated cross-cultural differences between Germans and Americans and identified the former to be more formal, we selected German and American English (Ger-Am) as the language pair to serve as a surrogate for the formality task. The German language is chosen to represent the formal class and the American English language is chosen to represent the informal class. We use the eBay Fashion product review bilingual data to learn the cross-culture differences. The data is extracted from the German and American sites of eBay e-commerce website, and the reviews originate from 24 sub-categories (e.g., “Jewelry”).

4.2 Sarcasm Detection Task

The sarcasm detection task (Tepperman et al., 2006) aims to classify a document as either non-sarcastic or sarcastic. For evaluating the sarcasm detection task, we used the following dataset:

**Sarcasm Corpus V2**

The Sarcasm Corpus version 2 (Oraby et al., 2016) includes response text from quote-response pairs annotated for sarcasm. Each sample consists of the form of sarcasm (Gen for general sarcasm, RQ for rhetorical question or...
Table 1: The size of the datasets and the characteristics of the length of documents in the datasets. The bilingual datasets used to learn the cross-cultural difference during training for each of the tasks are marked in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>size</th>
<th>mean</th>
<th>std.</th>
<th>min.</th>
<th>median</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Motors</td>
<td>786</td>
<td>78.48</td>
<td>42.98</td>
<td>10</td>
<td>72</td>
<td>197</td>
</tr>
<tr>
<td>Amazon Fashion</td>
<td>866</td>
<td>68.98</td>
<td>45.07</td>
<td>10</td>
<td>58</td>
<td>258</td>
</tr>
<tr>
<td>New York Times</td>
<td>3,226</td>
<td>27.38</td>
<td>5.370</td>
<td>10</td>
<td>27</td>
<td>49</td>
</tr>
<tr>
<td>Answers</td>
<td>858</td>
<td>12.91</td>
<td>2.585</td>
<td>10</td>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td>Blog</td>
<td>480</td>
<td>29.05</td>
<td>16.72</td>
<td>11</td>
<td>25</td>
<td>201</td>
</tr>
<tr>
<td>Email</td>
<td>366</td>
<td>26.46</td>
<td>18.92</td>
<td>10</td>
<td>22</td>
<td>230</td>
</tr>
<tr>
<td>News</td>
<td>622</td>
<td>28.73</td>
<td>14.82</td>
<td>11</td>
<td>26</td>
<td>248</td>
</tr>
<tr>
<td>eBay Ger-Am</td>
<td>11,576</td>
<td>33.71</td>
<td>25.47</td>
<td>10</td>
<td>26</td>
<td>188</td>
</tr>
<tr>
<td>Amazon Jp-Am</td>
<td>37,048</td>
<td>57.61</td>
<td>36.13</td>
<td>10</td>
<td>45</td>
<td>224</td>
</tr>
</tbody>
</table>

Sarcasm detection

<table>
<thead>
<tr>
<th>Dataset</th>
<th>size</th>
<th>mean</th>
<th>std.</th>
<th>min.</th>
<th>median</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarcasm Gen</td>
<td>3,260</td>
<td>47.54</td>
<td>34.72</td>
<td>10</td>
<td>35</td>
<td>241</td>
</tr>
<tr>
<td>Sarcasm RQ</td>
<td>786</td>
<td>79.82</td>
<td>42.45</td>
<td>13</td>
<td>76</td>
<td>188</td>
</tr>
<tr>
<td>Sarcasm Hyp</td>
<td>582</td>
<td>62.76</td>
<td>40.87</td>
<td>11</td>
<td>53</td>
<td>195</td>
</tr>
</tbody>
</table>

Surrogate Bilingual Dataset for Sarcasm

Based on prior research on cross-cultural differences between Japanese and American (Koga and Pearson, 1992; Minami, 1994; Martinsons, 2001; Adachi, 2010), and specifically Adachi (1996) and Ziv (1988) who addressed sarcasm in Japanese, we selected Japanese and American English (Jp-Am) as the language pair to serve as a surrogate for the sarcasm task. The Japanese language is chosen to represent the non-sarcastic class and the American English language is chosen to represent the sarcastic class. We use the Amazon Japanese and American reviews as the bilingual data to learn the cross-culture difference. The data originates from the Japanese and American marketplaces of the Amazon Customer Reviews Dataset\(^4\), and the reviews originate from 39 categories (e.g. “Books”).

4.3 Surrogate Bilingual Datasets Pre-processing

To allow the classifier to learn cross-cultural differences during training, and before transferring to the specific task, we seek to remove from the bilingual datasets used for training any other sources of information that may incidentally distinguish between the two language corpora: (1) We filter documents containing fewer than 10 or more than 250 terms; (2) We apply language detection (Shuyo, 2010) to remove documents in languages outside the surrogate languages; (3) We sample the same number of documents per each sub-category, document length and review-rating. This enables us to remove any bias towards a topic; (4) We sample the same number of documents from both languages to result with a balanced dataset in terms of class labels. We draw the reader’s attention to the fact that the documents in the bilingual datasets are plain reviews which have not gone through any formality classification or sarcasm detection as part of their pre-processing. We publicly release the surrogate datasets.\(^5\)

4.4 Classification Algorithms

We experiment with both Logistic Regression and XGBoost classifiers (McCullagh, 2018; Chen and Guestrin, 2016), and report the results of the former as it yielded superior performance. For Logistic Regression, we tune both the regularization norm ($L_1$ or $L_2$), and the regularization coefficient ($C$). For XGBoost, we tune the learning-rate, maximum tree depth, number of rounds (trees), minimum loss reduction required for partitioning, sub-sample ratio of features per tree, and both the $L_1$ and $L_2$ regularization coefficients. The configurations empirically chosen (based on grid search experiments on the validation set) were (1) $L_2$ regularization with $C = 10^4$ for models trained on bilingual data; (2) $L_2$ regularization with $C = 1$ for models trained on task-specific data. We also experimented with the following neural network architectures for the classification: LSTM-RNN (Hochreiter and Schmidhuber, 1997), HAN (Yang et al., 2016), QRNN (Bradbury et al., 2017), and VDCNN (Conneau et al., 2017). However, these models did not achieve any substantial performance gain to justify their additional complexity.

4.5 Evaluation Metrics

We report the AUC scores of stratified, five-fold cross-validation experiments on the formality classification and sarcasm detection tasks. For each fold of each experiment, the evaluation dataset is partitioned into train, validation and test splits. Task-specific models use all three splits for training, hyper-parameter tuning and evaluation, respectively, while models trained on bilingual data only use the test split for evaluation (so

\(^4\)https://s3.amazonaws.com/amazon-reviews-pds/readme.html

\(^5\)https://github.com/dorringel/CCTL
Table 2: AUC performance for our approach in comparison to the task-specific models directly trained on the evaluation datasets (results are comparable to supervised state-of-the-art). The final column reports the percent difference between the two methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>task-specific</th>
<th>our-approach</th>
<th>∆(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>formality classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Motors</td>
<td>98.75</td>
<td>95.61</td>
<td>-3.14</td>
</tr>
<tr>
<td>Amazon Fashion</td>
<td>97.97</td>
<td>91.67</td>
<td>-6.30</td>
</tr>
<tr>
<td>New York Times</td>
<td>90.57</td>
<td>84.30</td>
<td>-6.27</td>
</tr>
<tr>
<td>Answers</td>
<td>95.11</td>
<td>91.16</td>
<td>-3.95</td>
</tr>
<tr>
<td>Blog</td>
<td>89.24</td>
<td>79.29</td>
<td>-9.95</td>
</tr>
<tr>
<td>Email</td>
<td>96.16</td>
<td>88.91</td>
<td>-7.25</td>
</tr>
<tr>
<td>News</td>
<td>85.42</td>
<td>78.04</td>
<td>-7.38</td>
</tr>
</tbody>
</table>

Table 3: AUC performance for MT translation method in comparison to NN translation method. The best result per dataset is marked in bold. “*” indicates statistically significant difference of the leading method from the other method. The final column reports the percent difference between the two methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NN</th>
<th>MT</th>
<th>∆(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>formality classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Motors</td>
<td>91.76</td>
<td>95.61*</td>
<td>3.85</td>
</tr>
<tr>
<td>Amazon Fashion</td>
<td>84.59</td>
<td>91.67*</td>
<td>7.08</td>
</tr>
<tr>
<td>New York Times</td>
<td>81.02</td>
<td>84.30*</td>
<td>3.38</td>
</tr>
<tr>
<td>Answers</td>
<td>88.60</td>
<td>91.16*</td>
<td>2.56</td>
</tr>
<tr>
<td>Blog</td>
<td>78.01</td>
<td>79.29*</td>
<td>1.28</td>
</tr>
<tr>
<td>Email</td>
<td>88.30</td>
<td>88.91*</td>
<td>0.61</td>
</tr>
<tr>
<td>News</td>
<td>77.70</td>
<td>78.04*</td>
<td>0.34</td>
</tr>
</tbody>
</table>

5 Results and Discussion

Our empirical evaluation explores the following main research questions: (1) How effective is the approach of using bilingual data as described in Section 3 on our chosen tasks? (2) How does the choice of transformation impact the performance of the approach? Specifically, we consider the choice of machine translation and document representation.

Comparison to Models Trained on Task-specific Data In order to determine the overall effectiveness of our approach as described in Section 3, we consider the result of training a model using task-specific labeled data. We expect that if our approach is effective, we can achieve performance close to that of a model trained on this task-specific data. Table 2 reports the cross-validation AUC comparing our approach to the model trained on task-specific data. The table considers both tasks across the various evaluation datasets described in Section 4. The final column reports the percent difference between the two methods.

Examine the table, we observe that our method is within 10% of the AUC of the model trained on task-specific data, on all datasets and tasks, and as close as 4% on four of the datasets. Our model was trained entirely on bilingual data with no examples from the task it was evaluated on. The comparison with a model that was trained on data specific to the target task yields only a small improvement over our method. The comparable performance supports our thesis that cross-cultural information which exists within the bilingual data can be leveraged to achieve performance that is nearly equal to that achieved by collecting expensive task-specific labels.

Effect of Translation Component Table 3 considers the effect of the translation component of the compound transformation described in Section 3.2.1. Recall we considered two approaches for translation which represent a complexity-accuracy trade-off. We expect the MT approach to provide a higher quality translation, but it is substantially more complex to train and deploy. Conversely, the NN approach is simple to deploy at the cost of overall translation quality. The table illustrates that the difference in translation quality does have an impact on the down-stream task performance. However, the magnitude of the impact varies widely across the datasets in our experiments. For the formality task, the <8% gap between the methods may justify the reduced complexity of the NN approach. It is interesting to note that the model performs well even with a very simple word-by-word translation scheme.

Effect of Document Representation Table 4 considers the effect of the document representation component of the transformation discussed in Section 3.2.3 across the tasks, datasets, and translation methods. The first result column shows the...
amazing
be
so
you’re
a
perfect
clone
of
one
your
parents
with
zero
copying
...sell
for
$50.00
.give
me
the
money
and
wait
here!

my
dad
loved
watching
my
palmer
play,
and
I
loved
sharing
the
watching
with
him.
i
treasure
seeing
the
moments
of
joy
arnie
put
on
my
father’s
face.

great
replica!

awesome
item
for
the
price!
the
box
looks
fake
but
the
belt
inside
looks
...watching
with
him.
i
treasure
seeing
the
moments
of
joy
arnie
put
on
my
father’s
face.

(a) Formality classification: Two documents from the Amazon Fashion and the New York Times evaluation datasets classified as informal, respectively. Orange color indicates contribution to the document being classified as informal and blue color otherwise.

(b) Sarcasm detection: Two documents from the Sarcasm Corpus evaluation dataset classified as sarcastic. Green color indicates contribution to the document being classified as sarcastic and red color otherwise.

Figure 2: Documents from the evaluation datasets of both the formality classification, and sarcasm detection tasks and their corresponding LIME interpretation (Ribeiro et al., 2016). Color intensity is in proportion with the word’s contribution to the classification according to the LIME algorithm.

Table 4: AUC performance for various representation methods. AVG refers to a simple unweighted average of word vectors, IDF refers to a document-frequency-based weighting according to equation 1. LANG refers to a weighting scheme that takes the language of origin into consideration, based on Equation 4. The best result per dataset is marked in bold. “*” indicates statistically significant difference of the leading method from both other methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AVG</th>
<th>IDF</th>
<th>LANG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Motors$_{MT}$</td>
<td>95.26</td>
<td>95.47</td>
<td>95.61</td>
</tr>
<tr>
<td>Amazon Motors$_{NN}$</td>
<td>76.93</td>
<td>86.46</td>
<td>91.76*</td>
</tr>
<tr>
<td>Amazon Fashion$_{MT}$</td>
<td>90.31</td>
<td>91.19</td>
<td>91.67</td>
</tr>
<tr>
<td>Amazon Fashion$_{NN}$</td>
<td>61.30</td>
<td>75.83</td>
<td>84.59*</td>
</tr>
<tr>
<td>New York Times$_{MT}$</td>
<td>82.45</td>
<td>82.27</td>
<td>84.30</td>
</tr>
<tr>
<td>New York Times$_{NN}$</td>
<td>70.64</td>
<td>75.14</td>
<td>81.02*</td>
</tr>
<tr>
<td>Answers$_{MT}$</td>
<td>87.88</td>
<td>87.82</td>
<td>91.16*</td>
</tr>
<tr>
<td>Answers$_{NN}$</td>
<td>80.10</td>
<td>80.60</td>
<td>88.60*</td>
</tr>
<tr>
<td>Blog$_{MT}$</td>
<td>78.63</td>
<td>77.89</td>
<td>79.29</td>
</tr>
<tr>
<td>Blog$_{NN}$</td>
<td>65.56</td>
<td>68.80</td>
<td>78.01*</td>
</tr>
<tr>
<td>Email$_{MT}$</td>
<td>87.22</td>
<td>88.21</td>
<td>88.91</td>
</tr>
<tr>
<td>Email$_{NN}$</td>
<td>73.80</td>
<td>77.70</td>
<td>88.30*</td>
</tr>
<tr>
<td>News$_{MT}$</td>
<td>77.71</td>
<td>77.84</td>
<td>78.04</td>
</tr>
<tr>
<td>News$_{NN}$</td>
<td>65.30</td>
<td>69.40</td>
<td>77.70*</td>
</tr>
</tbody>
</table>

We also studied the effect of the choice of word embeddings as discussed in Section 3.2.2, but there were no statistical significant differences.

performance of a simple unweighted average of the word vectors. The second result column shows a document-frequency-based weighting according to equation 1. The third result column shows the performance of a weighting scheme that takes the language of origin into consideration, based on Equation 4.

Examining the table, we observe that using a non-uniform weighting scheme generally gives improved performance over the naive unweighted baseline. The effect of the document representation method is significant when paired with the NN translation approach. Thus, the translation and document representation components of the compound transformation are complementary in the sense that when translation is of high quality, a naive document representation suffices. Conversely, when translation quality is sub-optimal, the choice of document representation can significantly impact performance.

Qualitative Examples To better understand what the models trained on bilingual data actually learn, we utilize the LIME algorithm (Ribeiro et al., 2016) to attain each word’s gravity to the classification of the document. Figure 2 provides visual explanations of documents from the evaluation datasets of both the formality classification, and sarcasm detection tasks and their corresponding LIME interpretation. Specifically, Figure 2a showcases documents from the Amazon Fashion and the New York Times datasets classified as informal, respectively, while Figure 2b showcases documents from the Sarcasm Corpus dataset classified as sarcastic. Examining the formality visualization (2a), we observe that words that affect the formality of the document, such as awesome in the first example and treasure in the second one, are indeed assigned with a higher weight by the LIME algorithm. Similarly for the sarcasm visualization (2b), we observe that words that contribute to the document being sarcastic, such as perfect and amazing in the first example, and rolex and right followed immediately by back in the second example, are indeed assigned with a higher weight by the LIME algorithm.

Most Discriminating Words To further demonstrate qualitatively the formality discriminating in-
In this work, we show that cross-cultural differences can be harnessed for natural language text classification. We present a transfer-learning framework that leverages bilingual corpora for classification tasks using no task-specific data, and evaluate its performance on formality classification and sarcasm detection tasks. We show that our approach achieves comparable performance to task-specific methods directly trained on the two tasks, and propose a document representation scheme designed for bilingual training data. Such a representation can improve performance substantially when a low-quality translation method is used. In future work, we would like to generalize our approach to the multilingual case of multiple languages and a multi-class target task, explore applications beyond text classification, and study transformations that eliminate the need for a translation system.

**References**


<table>
<thead>
<tr>
<th>German (Translated)</th>
<th>American English</th>
</tr>
</thead>
<tbody>
<tr>
<td>long</td>
<td>love</td>
</tr>
<tr>
<td>high</td>
<td>one</td>
</tr>
<tr>
<td>well</td>
<td>loved</td>
</tr>
<tr>
<td>good</td>
<td>perfect</td>
</tr>
<tr>
<td>even</td>
<td>amazing</td>
</tr>
<tr>
<td>quality</td>
<td>wearing</td>
</tr>
<tr>
<td>cheap</td>
<td>favorite</td>
</tr>
<tr>
<td>chic</td>
<td>like</td>
</tr>
<tr>
<td>hand</td>
<td>comfortable</td>
</tr>
<tr>
<td>alternative</td>
<td>new</td>
</tr>
<tr>
<td>woody</td>
<td>terrible</td>
</tr>
</tbody>
</table>

Table 5: Most discriminative unigrams between German and American English according to their $wc(w_i, L_j)$ scores, based on the eBay Ger-Am product reviews dataset.

In this work, we show that cross-cultural differences can be harnessed for natural language text classification. We present a transfer-learning framework that leverages bilingual corpora for classification tasks using no task-specific data, and evaluate its performance on formality classification and sarcasm detection tasks. We show that our approach achieves comparable performance to task-specific methods directly trained on the two tasks, and propose a document representation scheme designed for bilingual training data. Such a representation can improve performance substantially when a low-quality translation method is used. In future work, we would like to generalize our approach to the multilingual case of multiple languages and a multi-class target task, explore applications beyond text classification, and study transformations that eliminate the need for a translation system.


Nakatani Shuyo. 2010. Language detection library for java.


Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In Proceedings of the 20th international conference on World wide web, pages 247–256. Association for Computing Machinery.