Cross-Cultural Transfer Learning for Text Classification

Dor Ringel[1] **Gal Lavee**[2] **Ido Guy**[2] **Kira Radinsky**[1]
[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.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3873–3883, Hong Kong, China, November 3–7, 2019.

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).

Algorithm 1 Cross-Cultural Transfer Learning

Input: C_0 (documents from language L_0)
Input: C_1 (documents from language L_1)
Input: T (transformation method)
Input: D_t (binary classification task dataset)
Output: $\hat{\mathbf{y}}_t$ (predictions)

$C_0^T, C_1^T \leftarrow \text{TRANSFORM\_DOCUMENTS}(T, C_0, C_1)$
$h \leftarrow \text{TRAIN\_CLASSIFIER}(C_0^T, C_1^T)$
$D_t^T \leftarrow \text{TRANSFORM\_DOCUMENTS}(T, D_t)$
$\hat{\mathbf{y}}_t \leftarrow \text{TRANSFER\_LEARNING}(h, D_t^T)$

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.

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:

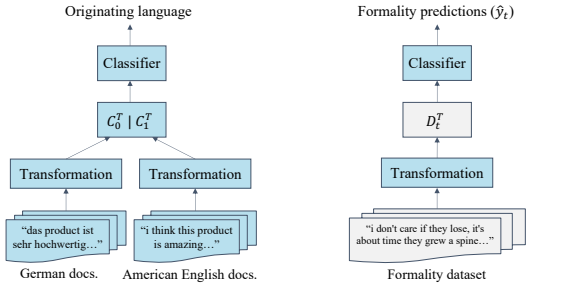
1. Using a transformation, denoted T, transform the two document classes into a common representation. The resulting transformed collections of documents are denoted C_0^T and C_1^T, respectively.

2. Train a binary classifier, denoted h, to discriminate document representations in C_0^T from those in C_1^T. 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 $\hat{\mathbf{y}}_t$.

Figure 1 summarizes the process.



(a) Train on transformed bilingual data (b) Apply to a binary classification task

Figure 1: Illustration of the cross-cultural transfer learning framework through the example of the formality classification task. (a) The classifier is trained on document representations of the German-American bilingual data ($C_0^T \mid C_1^T$) to distinguish documents as originating from either German or American English (L_0 or L_1). (b) The trained classifier is applied to a formality task.

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[1], 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

[1] https://cloud.google.com/translate/docs

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., 2017)).

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:

$$\mathbf{c}_d = \sum_i^n \frac{a}{\mathcal{U}(w_i) + a} \mathbf{v}_{w_i} \tag{1}$$

where $\mathbf{c}_d \in \mathbb{R}^k$ 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; $\mathcal{U}(w)$ 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 $\mathbf{c}_d$ 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*:

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

where we have used Bayes rule and C is constant w.r.t $\mathbf{c}_d$. As the notation suggests, we now have a generative model for each class:

$$P\left(w_i \mid \mathbf{c}_d, L_0\right) = \alpha \cdot P\left(w_i \mid L_0\right) + (1-\alpha) \cdot \left(\frac{\exp\left(\mathbf{v}_{w_i}^\top \mathbf{c}_d\right)}{Z}\right)$$
$$P\left(w_i \mid \mathbf{c}_d, L_1\right) = \alpha \cdot P\left(w_i \mid L_1\right) + (1-\alpha) \cdot \left(\frac{\exp\left(\mathbf{v}_{w_i}^\top \mathbf{c}_d\right)}{Z}\right)$$

where $P\left(w_i \mid L_0\right)$ and $P\left(w_i \mid L_1\right)$ are the conditional unigram models, and Z denotes a normalization constant.

Focusing on the probability of a single token:

$$\begin{aligned} f\left(\mathbf{c}_d\right) &= \mathbb{E}\left[\log P\left(w_i \mid \mathbf{c}_d, L\right)\right] \\ &= \frac{1}{2}\left[\log P\left(w_i \mid \mathbf{c}_d, L_0\right) + \log P\left(w_i \mid \mathbf{c}_d, L_1\right)\right] \end{aligned} \tag{2}$$

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

This expression has the following gradient (w.r.t. $\mathbf{c}_d$):

$$\begin{aligned} &\nabla f\left(\mathbf{c}_d\right) = \\ &\frac{1}{2} a \cdot e_{i,d} \left[\frac{P\left(w_i \mid L_0\right) + P\left(w_i \mid L_1\right) + 2a e_{i,d}}{\left(P\left(w_i \mid L_0\right) + a e_{i,d}\right) \cdot \left(P\left(w_i \mid L_1\right) + a e_{i,d}\right)}\right] \mathbf{v}_{w_i} \end{aligned} \tag{3}$$

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

$$\nabla f\left(0\right) = \frac{1}{2} a \left[\frac{P(w_i|y) + P(w_i|L_1) + 2a}{(P(w_i|L_0)+a)\cdot(P(w_i|L_1)+a)}\right] \mathbf{v}_{w_i}$$

[2] https://github.com/facebookresearch/MUSE

Thus, maximizing the Taylor approximation $f(\mathbf{c}_d) \approxeq f(0) + \nabla f(0)^\top \mathbf{c}_d$ w.r.t. $\mathbf{c}_d$ (following Arora et al. (2017)) yields the following estimator:

$$\mathbf{c}_d = \frac{1}{2} \sum_{i=1}^{n} a \left[\frac{P(w_i|L_0)+P(w_i|L_1)+2a}{(P(w_i|L_0)+a)\cdot(P(w_i|L_1)+a)} \right] \mathbf{v}_{w_i} \quad (4)$$

Examining the expression above, we can see that for a fixed value of a, the numerator of the 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 $\frac{\mathbf{c}_d}{||\mathbf{c}_d||}$.

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 (Kesarwani, 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 20^{th} 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)[3] 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

[3] https://nlds.soe.ucsc.edu/sarcasm2

Dataset	size	mean	std.	min.	median	max
formality classification						
Amazon Motors	786	78.48	42.98	10	72	197
Amazon Fashion	866	68.98	45.07	10	58	258
New York Times	3,226	27.38	5.370	10	27	49
Answers	858	12.91	2.585	10	13	48
Blog	480	29.05	16.72	11	25	201
Email	366	26.46	18.92	10	22	230
News	622	28.73	14.82	11	26	248
eBay Ger-Am	11,576	33.71	25.47	10	26	188
Sarcasm detection						
Sarcasm Gen	3,260	47.54	34.72	10	35	241
Sarcasm RQ	786	79.82	42.45	13	76	188
Sarcasm Hyp	582	62.76	40.87	11	53	195
Amazon Jp-Am	37,048	57.61	36.13	10	45	224

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.

Hyp for hyperbole), a class label (either sarcastic or non-sarcastic), a quote text, and a response text (the text annotated for sarcasm).

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 ($L1$ or $L2$), 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 $L1$ and $L2$ regularization coefficients. The configurations empirically chosen (based on grid search experiments on the validation set) were (1) $L2$ regularization with $C = 10^4$ for models trained on bilingual data; (2) $L2$ 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

Dataset	task-specific	our-approach	Δ(%)
formality classification			
Amazon Motors	98.75	95.61	-3.14
Amazon Fashion	97.97	91.67	-6.30
New York Times	90.57	84.30	-6.27
Answers	95.11	91.16	-3.95
Blog	89.24	79.29	-9.95
Email	96.16	88.91	-7.25
News	85.42	78.04	-7.38
Sarcasm detection			
Sarcasm Gen	80.38	76.99	-3.39
Sarcasm RQ	81.04	77.34	-3.70
Sarcasm Hyp	71.03	65.32	-5.71

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.

Dataset	NN	MT	Δ(%)
formality classification			
Amazon Motors	91.76	**95.61***	3.85
Amazon Fashion	84.59	**91.67***	7.08
New York Times	81.02	**84.30***	3.28
Answers	88.60	**91.16***	2.56
Blog	78.01	**79.29***	1.28
Email	88.30	**88.91***	0.61
News	77.70	**78.04***	0.34
Sarcasm detection			
Sarcasm Gen	65.00	**76.99***	11.99
Sarcasm RQ	63.70	**77.34***	13.64
Sarcasm Hyp	54.13	**65.32***	11.19

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.

that both types of models are evaluated on the same test set). Statistical significance is measured using one-way paired t-test (Casella and Berger, 2002) with $p<0.01$.

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.

Examining 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

Dataset	AVG	IDF	LANG
formality classification			
Amazon Motors$_{MT}$	95.26	95.47	**95.61**
Amazon Motors$_{NN}$	76.93	86.46	**91.76***
Amazon Fashion$_{MT}$	90.31	91.19	**91.67**
Amazon Fashion$_{NN}$	61.30	75.83	**84.59***
New York Times $_{MT}$	82.45	82.27	**84.30**
New York Times $_{NN}$	70.64	75.14	**81.02***
Answers $_{MT}$	87.88	87.82	**91.16***
Answers $_{NN}$	80.10	80.60	**88.60***
Blog $_{MT}$	78.63	77.89	**79.29**
Blog $_{NN}$	65.56	68.80	**78.01***
Email $_{MT}$	87.22	88.21	**88.91**
Email $_{NN}$	73.80	77.70	**88.30***
News $_{MT}$	77.71	77.84	**78.04**
News $_{NN}$	65.30	69.40	**77.70***
Sarcasm detection			
Sarcasm Gen $_{MT}$	76.68	76.82	**76.99**
Sarcasm Gen $_{NN}$	61.40	**65.00***	62.10
Sarcasm RQ $_{MT}$	76.81	**77.34**	77.16
Sarcasm RQ $_{NN}$	61.13	**63.70***	60.85
Sarcasm Hyp $_{MT}$	64.31	65.12	**65.32**
Sarcasm Hyp $_{NN}$	50.94	53.32	**54.13**

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.

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.

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.

great replica ! ! awesome item for the price ! the box looks fake but the belt inside looks real !

my dad loved watching mr palmer play , and i loved sharing the 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.

so you're a perfect clone of one your parents with zero copying errors ? amazing .

tell ya what ; i got this rolex i'm wiling to sell ya for $50 . 00 . give me the money and wait here . i'll be right back !

(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.

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-

German (Translated)		American English	
long	original	love	cute
high	broken	one	another
well	material	loved	going
good	handy	excellent	lovely
even	narrow	amazing	overpowering
cut	time	perfect	wanted
quality	clear	wearing	wonderful
cheap	ok	favorite	incredible
chic	flowery	like	overwhelming
hand	fake	comfortable	back
alternative	optimal	new	tiny
woody	waterproof	terrible	glad

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.

formation latent in German and American data, we seek the words that are most helpful in discriminating between the two languages. This notion is made concrete using the relative contribution of word w_i to the Kullback-Leibler divergence (Berger and Lafferty, 2017) between the languages, applied to all words, separately for each language:

$$wc\left(w_i, L_j\right) = P\left(w_i \mid L_j\right) \log\left(\frac{P(w_i|L_j)}{P\left(w_i|L_{(1-j)}\right)}\right)$$

Table 5 presents the unigrams that achieve the highest values for the expression above as computed on the eBay bilingual dataset of German and American English reviews. We can see that terms conveying information, such as `long`, `high`, `quality`, `cheap`, `woody`, and `clear` are more common in translated German documents, while terms conveying emotion, such as `love`, `amazing`, `favorite`, `wonderful`, `terrible`, and `overwhelming` are more common in American English documents.

6 Conclusions

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

Takanori Adachi. 1996. Sarcasm in japanese. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 20(1):1–36.

Yumi Adachi. 2010. Business negotiations between the americans and the japanese. *Global Business Languages*, 2(1):4.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

Adam Berger and John Lafferty. 2017. Information retrieval as statistical translation. In *ACM SIGIR Forum*, pages 219–226. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, pages 61–79.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-recurrent neural networks. *International Conference on Learning Representations (ICLR 2017)*.

George Casella and Roger L Berger. 2002. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for natural language processing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:*

Volume 1, Long Papers, pages 1107–1116. Association for Computational Linguistics.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

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. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517. International World Wide Web Conferences Steering Committee.

Norbert Hedderich. 2010. German-american intercultural differences at the workplace: A survey. *Global Business Languages*, 2(1):14.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations - Software of the Mind: Intercultural Cooperation and its Importance for Survival (3. ed.)*. McGraw-Hill.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. Association for Computing Machinery.

Juliane House. 1997. *Translation quality assessment: A model revisited*, volume 410. Gunter Narr Verlag.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

J. Johnson, M. Douze, and H. Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Aashita Kesarwani. 2018. New York Times Comments. Online; accessed 27 August 2018.

Yago Koga and Bethyl A Pearson. 1992. Cross-cultural advertising strategies in japanese vs. american women's magazines. *Intercultural Communication Studies II*, 1(1):1–18.

Shibamouli Lahiri. 2015. SQUINKY! A corpus of sentence-level formality, informativeness, and implicature. *arXiv preprint*, abs/1506.02306.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Raymond R Liu and Peter McClure. 2001. Recognizing cross-cultural differences in consumer complaint behavior and intentions: an empirical examination. *Journal of consumer marketing*, 18(1):54–75.

Maris G Martinsons. 2001. Comparing the decision styles of american, chinese and japanese business leaders. In *Best Paper Proceedings of Academy of Management Meetings, Washington, DC*.

Peter McCullagh. 2018. *Generalized linear models*. Routledge.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119. Curran Associates Inc.

Masahiko Minami. 1994. English and japanese: A cross-cultural comparison of parental styles of narrative elicitation. *Issues in Applied Linguistics*, 5(2):383–407.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2019. Generating product descriptions from user reviews. In *The World Wide Web Conference*, pages 1354–1364. ACM.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41. Association for Computational Linguistics.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, pages 1408–1417. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished manuscript accessible via the OpenAI Blog.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*.

Nakatani Shuyo. 2010. Language detection library for java.

Richard A. Shweder. 1991. *Thinking through cultures: Expeditions in cultural psychology*. Harvard University Press.

Richard A Shweder, Jacqueline J Goodnow, Giyoo Hatano, Robert A LeVine, Hazel R Markus, and Peggy J Miller. 2006. The cultural psychology of development: One mind, many mentalities. pages 716–792.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. "yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth International Conference on Spoken Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Elke U Weber and Christopher Hsee. 1998. Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Management science*, 44(9):1205–1217.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

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.

Avner Ziv. 1988. Teaching and learning with humor: Experiment and replication. *The Journal of Experimental Education*, 57(1):4–15.