Mach Translat (2018) 32:143–165
https://doi.org/10.1007/s10590-017-9202-6

# Cross-lingual sentiment transfer with limited resources

**Mohammad Sadegh Rasooli**[1] · **Noura Farra**[1] · **Axinia Radeva**[1] · **Tao Yu**[2] · **Kathleen McKeown**[1]

Received: 31 May 2017 / Accepted: 2 October 2017 / Published online: 15 November 2017


**Abstract** We describe two transfer approaches for building sentiment analysis systems without having gold labeled data in the target language. Unlike previous work that is focused on using only English as the source language and a small number of target languages, we use multiple source languages to learn a more robust sentiment transfer model for 16 languages from different language families. Our approaches explore the potential of using an *annotation projection* approach and a *direct transfer* approach using cross-lingual word representations and neural networks. Whereas most previous work relies on machine translation, we show that we can build cross-lingual sentiment analysis systems without machine translation or even high quality parallel data. We have conducted experiments assessing the availability of different resources such as in-domain parallel data, out-of-domain parallel data, and in-domain comparable data.

This work was done while the author was at Columbia.

✉ Noura Farra
noura@cs.columbia.edu

Mohammad Sadegh Rasooli
rasooli@cs.columbia.edu

Axinia Radeva
aradeva@columbia.edu

Tao Yu
tao.yu@yale.edu

Kathleen McKeown
kathy@cs.columbia.edu

[1] Department of Computer Science, Columbia University, New York, NY 10027, USA

[2] Department of Computer Science, Yale University, New Haven, USA

Our experiments show that we can build a robust transfer system whose performance can in some cases approach that of a supervised system.

**Keywords** Cross-lingual sentiment · Direct transfer · Annotation projection · Low-resource

## 1 Introduction

Online discussion forums and social media are often used to express subjective views. They are frequently exploited to express political viewpoints or to post reviews, but they can also be used in serious situations such as disasters where people post about their experiences. In such cases, negative sentiment can be an indication of continued problems while positive sentiment can be an indication that the situation has been happily resolved. Sentiment analysis involves assigning a sentiment label, usually positive, negative, or neutral, to a given text. While there has been quite a bit of work on identifying sentiment in English (e.g. Socher et al. 2013; Zhang et al. 2016), and some in languages such as Chinese (e.g. Wan 2008), Arabic (e.g. Abdul-Mageed and Diab 2011) and Spanish (e.g. Brooke et al. 2009), there has been much less work in identifying sentiment for low-resource languages where sentiment-labeled data or even machine translation systems do not exist. Yet emergency situations requiring humanitarian assistance can occur in areas where such languages are spoken; for example, the remote mountainous Xinjiang Autonomous Region in China where the Uyghur language is spoken was the site of a major earthquake. When such disasters occur, the ability to quickly develop systems for recognizing the sentiment posted by locals in their native languages can help humanitarian organizations quickly identify and respond to the areas of greatest need.

Most previous work in cross-lingual sentiment analysis (e.g. Duh et al. 2011; Balahur and Turchi 2014; Salameh et al. 2015; Zhou et al. 2016a, b) has typically assumed the availability of a full machine translation system, manually created lexicons, or in-domain parallel corpora. In the event of the occurrence of an incident situation, such resources may not be available for the target language. Furthermore, previous work has focused on a small number of languages and has solely used English as the source language. However, if labeled training data is available in languages in the same family as the target language, it would be useful to utilize such data from multiple source language families to build better sentiment transfer models.

In this work, our goal is to create robust sentiment analysis systems for languages or settings where no labeled sentiment training data exists and where machine translation capabilities are minimal. We create cross-lingual sentiment systems for multiple languages, leveraging parallel and comparable data from different genres. We consider the standard source-to-target cross-lingual transfer scenario as well as a multilingual scenario in which multiple source languages are used to transfer information to a target language. Our target languages come from a range of Indo-European, Turkic, Afro-Asiatic, Uralic, and Sino-Tibetan language families.

We consider two main approaches for transferring sentiment: *annotation projection* and *direct transfer*. We apply the classical annotation projection in a new setting

with a large number of different sources to project supervised labels from the source languages to the target language. Our direct transfer approach directly trains a sentiment analysis system on the labeled sentences of the source language(s) and applies the trained model directly on the target language. We were motivated to experiment with these methods based on our experience in the LoReHLT Situation Frame evaluation, where we developed an approach based on annotation projection to aid in filling the urgency slot. Since the evaluation, we have improved our annotation projection approach and developed a new direct transfer approach. Our goal in this paper is to show through experimentation the value of these two approaches and the impact of the availability of different resources.

The main contributions of this paper are the following:

- We introduce a novel direct transfer method using deep learning with a partial lexicalization strategy. The approach relies on translating some input words into the target languages while keeping the structure of the source language. Therefore, our model does not rely on having fully translated text.[1]
- We introduce a simple but effective annotation projection strategy which uses a state of the art sentiment system that is easily and quickly applied to any new language. Our projection model works particularly well for non-Indo-European languages even when in-domain data is not available.
- We systematically experiment with different methods based on direct transfer and projection to develop sentiment analysis systems in the absence of rich resources. Our experiments on a set of 16 different languages show that we can create a robust model. Our experiments study the effect of parallel data from different genres, single and multi-source transfer, and manual dictionaries. Our use of the Bible and Quran for parallel data simulate a low resource scenario for languages where more resources do exist.
- We introduce a cross-lingual sentiment scenario which does not rely on parallel data at all, but uses comparable topical corpora along with the direct transfer model to predict sentiment labels in the low-resource language. We show that this is a plausible alternative to using out-of-domain parallel corpora. To our knowledge, comparable corpora have not been previously utilized to create word representations for cross-lingual sentiment analysis.

Our experiments show that both our approaches benefit from having multiple source languages and in some cases the performance of the transfer method is close to the supervised model. We show that the direct transfer approach does particularly well with out-of-domain data and when multiple source languages are available, while projection is more suited for in-domain data and for non-European language families. An ensemble of the two approaches results in further performance boosts in the multi-source setting. Our error analysis treating English as a low-resource language concludes that our transfer model makes very reasonable errors on tweets with out-of-vocabulary words.

The rest of this paper is structured as follows: Sect. 2 overviews related work. Sections 3 and 4 explain our annotation projection and direct transfer approaches.

[1] The source code for the direct transfer model is available here: https://github.com/rasoolims/senti-lstm.

Experimental settings are described in Sect. 5, followed by discussion and error analysis in Sects. 6 and 7.

## 2 Related work

While the vast majority of work in sentiment analysis has focused on English (Hu and Liu 2004; Liu 2012; Socher et al. 2013), there have been many studies on cross-lingual analysis and transfer across languages. Cross-lingual sentiment systems have mostly focused on projection methods using in-domain corpora (Mihalcea et al. 2007), and on machine translation methods which either translate the target language text into a resource-rich language such as English and apply an English sentiment model (Wan 2008; Salameh et al. 2015), or translate the English labeled data into the target language and build a target-language system (Balahur and Turchi 2014). Recent neural network approaches include that of Zhou et al. (2016a), who developed a hierarchical attention model by translating the training data into the target language and modeling both the source and target side using a bidirectional LSTM, and Zhou et al. (2015) who also translated the source training data and used denoising autoencoders to create bilingual embeddings incorporating sentiment information from labeled data and their translations. These approaches rely on machine translation of in-domain text and would unlikely be applicable in a low resource setting where in-domain parallel data or high quality MT are not available. Previous work has also focused on developing language-specific systems for a new language, e.g. Mukund and Srihari (2010). for Urdu and and Joshi et al. (2010) for Hindi; these methods when available are beneficial to augment cross-lingual techniques.

Duh et al. (2011) has argued that even if perfect machine translation were available, it would not be a complete solution for the cross-lingual sentiment task, for if that were viewed as a domain adaptation task, machine translation systems would produce "domain-mismatch" between vocabulary distributions in the original and translated data, with limited sentiment vocabulary in the translated data. Meng et al. (2012) avoided the use of an MT system by instead developing a cross-lingual mixture model which maximizes the likelihood of generating a bilingual Chinese-English parallel corpus and determining word generation probabilities for the sentiment classes, where labeled data in the target language need not be available. This approach is complementary to our direct transfer method. Instead of using a cross-lingual mixture model, we train systems directly on source language data using cross-lingual word vector representations. As with Meng et al. (2012), we utilize parallel data. In our case, the parallel data is used either to create automatic dictionaries and cross-lingual embeddings, or for annotation projection. Moreover, we explore many source-target language pairs as well as parallel corpora from religious text that is out-of-domain but more likely to be available for under-resourced languages.

It is worth noting that there are a few complementary methods that are similar to the direct transfer model without using machine translation. For example Zhou et al. (2014) used a bilingual corpus and stacked autoencoders to create shared sentence representations for Chinese and English sentences from which sentiment models were trained directly on English labels. Chen et al. (2016) trained a direct transfer model using

an adversarial deep averaging network, whereby the model tries to create language-invariant bilingual representations that are not recognizable by a language predictor. Unlike these approaches, our strategy leverages data from different source languages and parallel data from both in-domain and religious text, and our projection method applies a new configuration of majority voting when more than one translation is available for a target language sentence.

Our direct transfer approach retains some similarities to the machine translation approach with the following differences: first, instead of fully translating the training data into the target language, we use the parallel corpora to create word alignment dictionaries and then only translate words with available dictionary entries. Moreover, even without any translation into the target language, it is sufficient to provide our model with cross-lingual word representations (word embeddings and word clusters) and train our model directly using those representations. Second, our method does not apply any reordering or structural changes to the source sentences, which makes the translation less likely to alter the sentiment of the sentence. Our model also works with comparable cross-lingual data that is not aligned on sentence or document level. As far as we are aware, the use of comparable corpora for cross-lingual sentiment analysis is very limited, whereby applications of comparable embeddings have been typically restricted to cross-lingual document classification or lexicon creation (Vulić and Moens 2016).

There has been a great deal of work on inducing cross-lingual word representations (e.g. Täckström et al. 2012; Hermann and Blunsom 2013; Faruqui and Dyer 2014; Ammar et al. 2016). In this paper, we use the method of Rasooli and Collins (2017), which is also similar to the method of Gouws and Søgaard (2015) and Wick et al. (2015). Our source translation approach with dictionaries is inspired by Rasooli and Collins (2017).

## 3 Annotation projection

Annotation projection is a classic but effective cross-lingual transfer technique, especially when parallel corpora from multiple sources are available. In *single-source* annotation projection, it is assumed that a parallel corpus is available where sentences $\{s_1, s_2, \ldots, s_m\}$ from the source language $L_s$ are aligned to sentences $\{t_1, t_2, \ldots, t_m\}$ in a low-resource language $L_t$; i.e. $s_i$ is a manual translation for $t_i$ for $i = 1, \cdots, m$. This work assumes that labeled training examples $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ exist for the source language $L_s$. Thus, a supervised system $\mathcal{M}_{sup}$ is first trained on $\mathcal{X}$ and afterward, the trained system $\mathcal{M}_{sup}$ is used to predict the labels of the source-side parallel translated text $\{l_{s_1}, l_{s_2}, \ldots, l_{s_m}\}$. Those labels are projected to the target-side translation by assuming $l_{s_i} \rightarrow l_{t_i}$. After applying projection, the target-side translation text $\{t_1, t_2, \ldots, t_m\}$ is used with the projected labels $\{l_{t_1}, l_{t_2}, \ldots, l_{t_m}\}$ to train the target model $\mathcal{M}_{proj}$. The trained model $\mathcal{M}_{proj}$ is the final model for analyzing sentiment in the target language.

*Multi-source transfer* is an extension we propose to single-source transfer. In this setting, there are $k$ source languages, each with separate labeled training data. It is also assumed that there exists multi-parallel corpora where every target language

sentence has $k$ source language translations. In this setting, one can apply the same projection technique to create a supervised model $\mathcal{M}^i_{sup}$ for each source language $L_i$ ($i = 1, \cdots k$) and project labels to the target language sentence. Then majority voting is applied on the $k$ projected labels for each target sentence to get the most reliable projected label. As with single-source projection, a supervised system is then trained on the projected data to get the final model $\mathcal{M}_{proj}$.

### 3.1 The learning model

Annotation projection requires only a simple, fast and easily transferable model. For this we choose to use Naive Bayes Logistic Regression with word embedding features (NBLR + POSwemb) (Yu et al. 2017) to train systems for both *source-side* labeled data and *target-side* projected data. This is an extension of Wang and Manning (2012). The advantage of this model is that it can be quickly trained whenever a new language is introduced, which makes it potentially more efficient for the LoReHLT scenario of developing a system within one day as training time is less than that of a deep learning model.

The model uses two types of features: sparse and dense. The sparse features include ngrams up to length 3. For each ngram $t$ in sentence $x$, the model counts the number of times that the ngram occurs with each possible label $l$ in the training data and stores it as $f^l(t)$ . It defines two vectors $p_l$ and $q_l$ for each label $l$, where each ngram $t$ in sentence $x$ has the following values:

$$\begin{cases} p_l(t) = \alpha + f^l(t) \\ q_l(t) = \alpha + \sum_{y \neq l} f^y(t) \end{cases}$$

where $\alpha$ is a smoothing constant. Finally the following log-count ratio is defined:

$$r_l = \log \left( \frac{p_l / ||p_l||_1}{q_l / ||q_l||_1} \right)$$

where $||p_l||_1$ and $||q_l||_1$ are the $L-1$ norm values of the corresponding vectors. The value $r_l(t)$ for any ngram $t$ not appearing in sentence $x$ is zero, leading to a very sparse vector. Hence, as opposed to having a sparse count vector (as in traditional machine learning methods), we show the features by the concatenation of its log-count ratios:

$$r(x) = r_{negative} \circ r_{neutral} \circ r_{positive}$$

where $\circ$ is the concatenation operator. Thus, if the number of seen ngrams in the training data is $m$, the feature vector $r$ will be a sparse vector in $\mathbb{R}^{3m}$.[2]

For the dense feature representation, we group words according to their part-of-speech (POS) tags in the set $P$ = {NOUN, VERB, ADJECTIVE} and average their pre-trained word embedding vectors. The concatenation of the three averaged word

[2] In the case of using English as the supervised source language, we also append additional positive and negative indicator features as additional unigrams and calculate their log ratio counts. These indicators are extracted from the sentiment lexicons of Wilson et al. (2005) and Hu and Liu (2004).

vectors and the sparse vector $r(x)$ is the final input feature to the logistic regression classifier.

## 4 Direct transfer

In direct transfer, a system can be trained on a source language dataset directly, with an unrestricted number of source languages. In other words, a supervised sentiment analysis system can be trained on the concatenation of $k$ labeled datasets $\cup_{i=1}^{k} \mathcal{X}_i$ from $k$ source languages or it can use the outcome of an ensemble of $k$ single-source direct transfer models. The advantage of direct transfer is the ability to directly use the gold labels of the source data in a single model without having to train separate source-language and target-language models. With multi-source, it can directly combine training data from different source languages in a single model. Moreover, the dependence on parallel data is reduced; it is possible to train a direct transfer model without any parallel data, namely if a bilingual dictionary or comparable data is available.

The main problem with direct transfer is that most of the common features do not generalize beyond each source language: for example, lexical features in one language are unlikely to appear in other languages. We apply the following techniques to address this problem:

- *Word representation features* We train cross-lingual word representations such that words with similar meanings in different languages have similar representations. We use the method of Rasooli and Collins (2017) to train cross-lingual word embeddings and word clusters. The method of Rasooli and Collins (2017) uses cross-lingual dictionaries to apply random code-switching on monolingual texts (e.g. Wikipedia) in all source and target languages. This leads to a concatenation of monolingual texts in different languages for which each sentence has words from different languages. The code-switched text is trained to derive word clusters and embeddings.
  For the comparable corpora setting, where we do not have cross-lingual dictionaries, we merge monolingual texts from topically comparable documents in the source and target languages and create cross-lingual embeddings on the combined text.
- *Lexicalization with code-switching* This approach allows us to enchance the direct transfer model by using lexicalized features directly from the target language while at the same time preserving the source language structure. It assumes that either a parallel corpus or bilingual dictionary is available.
  We use a simple deterministic partial translation strategy inspired from Rasooli and Collins (2017). First, we use the parallel corpora to create source-target word alignments using GIZA++. The word alignments are used to create bilingual dictionaries. During training, the bilingual dictionaries are used to translate words from the labeled source training data to the target language.
  We translate all words that have an entry in the dictionaries. Therefore, since many words do not exist in the bilingual dictionaries, the resulting training data looks like code-switched text. The following sentence is an example English sentence par-

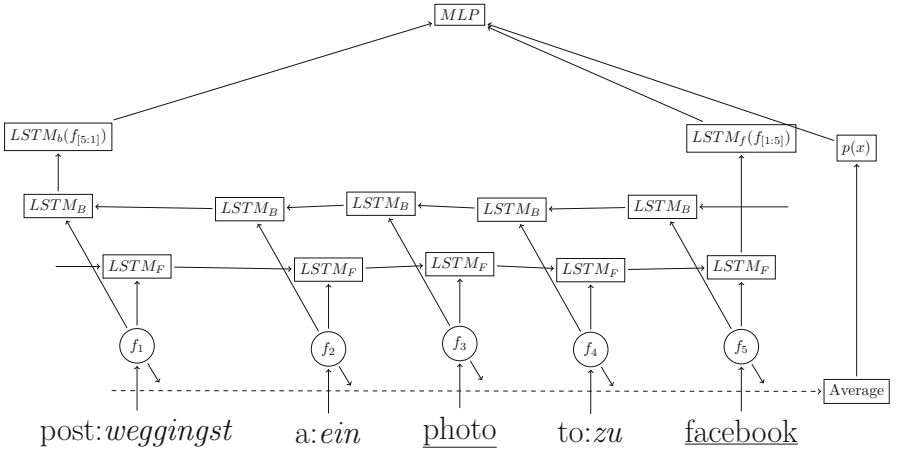


**Fig. 1** A graphical depiction of the neural network model in our direct transfer approach. This is an example of an English tweet translated to German. The words are shown at the bottom with their translation appearing after colon (":") for which underlined words are not translated. The circles ($f_i$) show the embedding layer, the LSTMs are shown on top of the embedding layer and the average layer on the right side of the words. All the intermediate layers are concatenated and fed to a multi-layer perceptron (MLP). More details are given in Sect. 4

tially translated to German via the dictionary extracted from Quran and Bible (the underlined words are not translated): **post:*weggingst*** a:***ein*** **photo** to:***zu*** **facebook**.

### 4.1 Deep learning model

Direct transfer requires a powerful model that can fully benefit from cross-lingual word representations and utilize the context and structure of the input text in order to determine sentiment labels. For this we choose to model direct transfer using long short-term memory (LSTM) networks (Hochreiter and Schmidhuber 1997). The input to our model is a sequence of $n$ words in the sentence $x = \{x_1, x_2, \ldots, x_n\}$. We assume the lexicalization step is applied on the words in the training data if a dictionary exists. The model uses two representations for the input: a recurrent representation based LSTM and the other based on averaging over all words. The latter representation is mainly useful when the source and target languages do not have similar word orders. A graphical depiction of the underlying model is shown in Fig. 1.

*Embedding layer* We use the following features for every word in a sentence. For all the following features, if a source training word can be translated into the target language and its translation has an entry in the word embeddings dictionary, we use that as the feature; otherwise we use the source word feature as the input feature:

- A *fixed* pre-trained cross-lingual word embedding $x_{ce} \in \mathbb{R}^{d_{ce}}$ extracted from the method of Rasooli and Collins (2017).
- A randomly initialized word embedding $x_e \in \mathbb{R}^{d_e}$ for every word in the sentence.
- The cross-lingual word cluster embedding $x_{cc} \in \mathbb{R}^{d_{cc}}$ to represent the word cluster identity of each word.

– In the case of single source transfer from English, we also use the fixed two-dimensional Sentiwordnet (Baccianella et al. 2010) score $x_{sw} \in \mathbb{R}^2$ that represents the likelihood of a word being positive or negative. We simply translate Sentiwordnet lexicon to the target language by using the translation dictionaries.

*Intermediate layer* The input for every word $x_i$ to the intermediate layer is the concatenation of the embedding features: $f \in \mathbb{R}^{d_f} = [x^i_{ce}; x^i_e; x^i_{cc}; x^i_{sw}]$. The intermediate layer is composed of the following structures:

– **Recurrent layer** We use a Bidirectional LSTM (BiLSTM) for representing the sequence of words in the sentence: a forward pass $LSTM_f(f_{[1:n]}) \in \mathbb{R}^{d_{rec} \times d_f}$ gives the final representation of the sentence by looking from the beginning to the end, and the backward pass $LSTM_b(f_{[n:1]}) \in \mathbb{R}^{d_{rec} \times d_f}$ looks from the end to the beginning. Then the output of the two LSTMs are concatenated as $r(x) \in \mathbb{R}^{2 \cdot d_{rec}}$.
– **Average layer** The average layer $p(x) \in \mathbb{R}^{d_f}$ is the average over all the input features $f$ for all words in the sentence. This layer represents the bag-of-words information of the sentence without taking the sequence information into account.

*Output layer* The two intermediate layers $r(x)$ and $p(x)$ are concatenated and fed to a hidden layer $H \in \mathbb{R}^{d_h \times (2 \cdot d_{rec} + d_f)}$ activated by rectified linear units (ReLU) (Nair and Hinton 2010):

$$H(x) = ReLU(H([r(x); p(x)]))$$

Finally the hidden layer output is fed to the output softmax layer. We use the log-likelihood objective function with the Adam optimizer (Kingma and Ba 2014) to learn the model parameters.

## 5 Experimental settings

### 5.1 Transfer scenarios

We have conducted a diverse set of experiments in order to evaluate our methods and determine the impact of the availability of different resources. We experiment with three scenarios:

– *Parallel data* In this setting, we experiment with both the annotation projection and direct transfer approaches. In order to determine the effect of domain and genre, we use different types of parallel corpora: religious text, contemporary political text and relatively in-domain and in-genre data from the Linguistic Data Consortium.
– *Manual translation dictionaries* Instead of automatically creating dictionaries from word alignments, we use the manual translation dictionaries extracted from Wiktionary[3] and use them in the direct transfer model. The Wiktionary entries are noisy and for some languages, the coverage of the lexicon is very low.
– *Comparable corpora* In this setting, we do not use any aligned corpus or bilingual dictionary, but instead we rely on cross-lingual word embeddings extracted

[3] https://www.wiktionary.org/.

from comparable corpora. The embeddings are created by merging together topically similar English and target-language documents using the *length-ratio-shuffle* method of Vulić and Moens (2016). However, instead of preserving the order of the sentences in the original documents, we first rank the sentences by their average TF-IDF scores and then merge the documents. This makes it more likely that frequently occurring topical words in the source language will align to the corresponding topic words in the target language. In this setting, since we do not have access to any dictionaries, we only use the fixed pre-trained word embedding features $x_{ce}$ for the direct transfer method.

### 5.2 Datasets, tools and settings

*Labeled sentiment data* We downloaded tweets labeled with sentiment for 12 languages from Mozetič et al. (2016)[4] as well as SentiPers data (Hosseini et al. 2015), a set of digital products reviews for Persian, as our source languages. All labeled data is annotated for positive, negative, and neutral. The evaluation languages are Arabic (ar), Bulgarian (bg), German (de), English (en), Spanish (es), Croatian (hr), Hungarian (hu), Polish (pl), Portuguese (pt), Russian (ru), Slovak (sk), Slovene (sl), Swedish (sv), Uyghur (ug) and Chinese (zh). We use 80% of the data as training, 10% for development and 10% of the data for testing. We use all development data sets to train the supervised models. We set the number of training epochs for the transfer model by looking at the performance on the Persian development data. As labeled training data for Uyghur, Chinese and Arabic is either unavailable or much smaller (in the case of Arabic) than that of Mozetič et al. (2016)'s Twitter data, we have used these three languages only as target languages. For evaluation data, we employed a native informant for Uyghur and Chinese to annotate a small number of sentences, and for Arabic, we used the newly released evaluation data of SemEval 2017 Task 4[5]. Table 1 shows the data sizes.

*Parallel data* We use the following parallel corpora:

– *Bible and Quran* The motivation for experimenting with religious data is that it is a lot more likely to be available for many languages, even very low-resource languages. We use the Quran and Bible translations as out-of-domain parallel datasets, from a very different genre than our test data, and thus create a low resource scenario for languages that have larger resources available. We use the corpus of Christodouloupoulos and Steedman (2014) for the Bible dataset and the Tanzil translations[6] for the Quran. This dataset has multiple translations for some languages.[7] When using the annotation projection approach, in cases where we have more than one Quran English translation for a target-language sentence, we

[4] Not all tweets from Mozetič et al. (2016) are available anymore (some of them were deleted).

[5] http://alt.qcri.org/semeval2017/task4/.

[6] http://tanzil.net/trans/.

[7] We excluded a subset of the translations from Russian and English that are interpretations as opposed to translations. For Russian, we use the Krachkovsky, Kuliev, Osmanov, Porokhova, and Sablukov translations

**Table 1** Training and evaluation sizes for different languages

| Dataset | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | #sen | #tok | #types | #sen | #tok | #types |
| Arabic | – | – | – | 6100 | 115401 | 19474 |
| Bulgarian | 23739 | 313003 | 52922 | 2958 | 38685 | 12545 |
| Chinese | – | – | – | 487 | 10243 | 3213 |
| Croatian | 56212 | 726037 | 88966 | 7025 | 90544 | 23827 |
| English | 46623 | 601284 | 39294 | 5828 | 75147 | 10857 |
| German | 63669 | 822605 | 78353 | 7961 | 102822 | 19907 |
| Hungarian | 36167 | 427951 | 83931 | 4520 | 53588 | 17750 |
| Persian | 15000 | 331431 | 13840 | 3027 | 67926 | 6886 |
| Polish | 116105 | 1455273 | 136089 | 14517 | 182194 | 37018 |
| Portuguese | 62989 | 712852 | 41982 | 7872 | 89553 | 12440 |
| Russian | 44757 | 522218 | 84463 | 5594 | 64779 | 18823 |
| Slovak | 40470 | 576611 | 82281 | 5058 | 71131 | 20751 |
| Slovene | 74238 | 1106211 | 108735 | 9277 | 137371 | 29899 |
| Spanish | 137106 | 2007046 | 91419 | 17133 | 249690 | 27942 |
| Swedish | 32600 | 494139 | 42446 | 4074 | 61423 | 11698 |
| Uyghur | – | – | – | 346 | 6805 | 3560 |

We ignored the hash tags, urls and name mentions in the data

run the model on all translations and apply majority voting to get the most frequent label.

- *Europarl* We use the Europarl data (Koehn 2005) as contemporary political text. We restricted ourselves to those sentences that are translated to all of the 10 languages (Bulgarian, German, English, Spanish, Hungarian, Polish, Portuguese, Slovak, Slovene, and Swedish). That comprised a total of 294738 sentences for all languages.
- *Linguistic data consortium (LDC) parallel data* The LDC packages are produced under the Low Resource Languages for Emergent Incidents (LORELEI) program and consist of a combination of the following genres: news, discussion forums, and social networks such as Twitter. We use seven English to target parallel translations from the LDC data: Chinese (16440 sentences)[8], Persian (57087 sentences)[9], Hungarian (157931 sentences)[10], Arabic (49446 sentences)[11], Russian (193967

Footnote 7 continued
and for English, we use the Ahmedali, Arberry, Daryabadi, Itani, Mubarakpuri, Pickthall, Qarai, Qaribullah, Sahih, Sarwar, Shakir, Wahiduddin, and Yusufali translations.

[8] LDC2016E30_LORELEI_Mandarin.

[9] LDC2016E93_LORELEI_Farsi.

[10] LDC2016E99_LORELEI_Hungarian.

[11] LDC2016E89_LORELEI_Arabic.

sentences)[12], Spanish (345940 sentences)[13], and Uyghur (99272 sentences)[14]. Since our evaluation set consists of Twitter data except for Persian, LDC has the most similar genres to our evaluation datasets.

*Automatic dictionaries* We use GIZA++ (Och and Ney 2003) to create automatic word alignments between source and target words from the parallel sentence-aligned corpora. After getting the intersected alignments from the two alignment directions, we choose the most frequent translation for each word in order to prune alignment noise.

*Comparable data* Comparable data is an alternative when parallel corpora are not available. To create a comparable data corpus, we collected Wikipedia articles about similar topics in each of the 16 languages. We picked a set of broad pre-defined topics and used the Wikipedia API[15] to automatically retrieve articles about these topics in each language. For maximum coverage, the articles are topic-aligned but not document-aligned in the different languages. We chose 60 pre-defined topics (e.g politics, science, sports), including named political and geopolitical entities relevant to the 16 languages (e.g Xinjiang, Russia, Arabs, Barack Obama), and translated them either manually (if we speak the language), using an online dictionary (in the case of Uyghur, where Google Translate is not available), or using Google Translate for the remaining languages. The translation of these topic words simulates 15 min of the availability of a native informant who speaks or knows the target language; they can be thought of as seed words for creating a cross-lingual linguistic resource. The extracted corpora contain 1M sentences (Arabic), 938K (Bulgarian), 1.6M (German), 2.1M (English), 1.6M (Spanish), 790K (Persian), 1.5M (Hungarian), 1.2M (Croatian), 1.1M (Polish), 1M (Portuguese), 1.6M (Russian), 973K (Slovak), 1.4M (Slovene), 671K (Swedish), 1M (Chinese), and 236K (Uyghur).

*POS tagging and tokenization* OpenNLP is used to split sentences. To tokenize words, we use the Stanford Chinese segmenter (Chang et al. 2008), Madamira Arabic tokenizer (Pasha et al. 2014)[16], Hazm Persian tokenizer[17], European tokenizers in the Europarl package, and OpenNLP[18] for all other languages. We use the POS information in the Universal Dependencies[19] to train POS taggers for the NBLR+POSwemb (annotation projection) model.

[12] LDC2016E95_LORELEI_Russian.

[13] LDC2016E97_LORELEI_Spanish.

[14] LDC2016E57_LORELEI_IL3_Incident_Language_Pack_for_Year_1_Eval .

[15] https://pypi.python.org/pypi/wikipedia.

[16] Madamira is used in low-resource mode with the form-based ATB_BWFORM tokenization scheme.

[17] https://github.com/sobhe/hazm.

[18] https://opennlp.apache.org/.

[19] http://universaldependencies.org/.

*Embeddings and brown clusters* We use the Wikipedia dump data set to train the monolingual and cross-lingual word embeddings and Brown clusters. We use the Word2vec tool[20] with its default setting and dimension of 300 for all of our embeddings (monolingual and cross-lingual), except for the comparable embeddings where we used a window size of 10. We trained monolingual and cross-lingual word clusters with the method of Stratos et al. (2014) with 500 clusters.

*Model settings* We set the following parameters for the NBLR +POSwemb model: $\alpha = 1$, $C = 0.01$ or 0.05. We use the LIBLINEAR package (Fan et al. 2008) implemented in the sklearn software[21]. The Dynet library (Neubig et al. 2017) with its default settings is used for the neural network. We use 7 epochs for the single-source transfer and 2 epochs for the multi-source transfer with concatenation. The following parameters are used: $d_{ce} = 300$, $d_e = 400$, $d_{cc} = 50$, $d_{rec} = 400$, $d_h = 400$, and batch size of 10 K.

All our models were tuned only when supervised data was assumed available or projected. When we applied our direct transfer models to the new language, we did not do any additional tuning on the target language, except in the case of Persian, which we used as a development language.

### 5.3 Baseline approaches

We use two simple baselines. The first translates the Sentiwordnet lexicon (Baccianella et al. 2010) to each target language by using the dictionaries extracted from the Bible and Quran parallel data. It decides to give positive/negative sentiment, if the average positive/negative score is at least 0.1 higher than the average negative/positive score. Otherwise it assigns the neutral label. The second baseline assigns all sentences *neutral* majority label of the English training data.

## 6 Results and discussion

This section reports and discusses our results showing the performance of the annotation projection and direct transfer models under the availability of different resources: in-domain parallel corpora (LDC), contemporary European political text (EP), and out-of-domain parallel corpora (BQ), which is available for every language. For direct transfer, we additionally have comparable data (CP) and Wiktionary (WK). Models in all languages apply three-way sentiment classification (positive, negative, and neutral). Since accuracy can be misleading if the majority label from the source language data is dominantly predicted, we use F1 score macro-averaged over the three classes as our evaluation metric. Significance is computed using the bootstrap significance method (Berg-Kirkpatrick et al. 2012).

[20] https://code.google.com/archive/p/word2vec/.

[21] http://scikit-learn.org/stable/.

### 6.1 English to target transfer

In the first set of experiments, we conduct single-source transfer by using only English as the source language. The results for projection and direct transfer are depicted in Table 2 for difference resources. The single-source transfer approaches perform significantly better ($p < 0.05$) than the *Neut* majority baseline for all languages and significantly better than the *Senti* baseline in all cases except with Ugyhur, where our evaluation set is quite small, and in some languages with CP and WK, where the transfer is dependent on the quality of the Wikipedia resources.

We see some trends appearing in terms of the impact of different resources. First, and unsurprisingly, using in-domain data (LDC) enables significant increases in performance for most languages. For projection, using LDC outperforms both BQ and EP in all cases. For direct transfer, LDC outperforms BQ and EP in all cases except for Persian, Arabic, and Spanish (where both closely approach the supervised bound). Interestingly, BQ actually ends up being a better choice for Persian and Arabic because the BQ corpus is larger or has more unique Quran translations than LDC parallel translations, leading to a dictionary with higher coverage. The quality of direct transfer is affected by the coverage of the dictionary that was extracted, particularly for languages which have morphologically complex words.

Second, when comparing automatic vs. manual (Wiktionary) dictionaries, the result depends heavily on the quality of Wiktionary for that language. Wiktionary significantly outperforms BQ in 4 out of 14 languages, such as Chinese, where the manual dictionary is of exceptionally high quality; in the remaining cases, BQ is either better or the two are indistinguishable. The quality of dictionary is an important factor in direct transfer because both the cross-lingual embeddings and code-switching are the result of those dictionaries. Wiktionary is better for languages such as German and Chinese but less so in Arabic and Persian, where many of the Wiktionary entries occur in their uninflected forms, but we have large BQ corpora for both Arabic and Persian and so better BQ performance.

In comparing projection with direct transfer, direct transfer does significantly better than projection in the low resource BQ setting. This applies to all cases except Chinese and Uyghur and the results are indistinguishable for two other languages, Slovene and Swedish. Thus, for low resource settings without in-domain data, we can conclude that direct transfer is a better than projection; this makes sense because it relies on dictionaries and embeddings rather than projecting sentences which are not relevant for sentiment analysis. On the other hand, with EP, projection does better in five out of nine cases and with LDC, projection does better than direct transfer in five out of seven cases. These results imply that if we do have high-quality in-domain parallel data, projecting annotations from English seems to be preferable.

Our results for English–Chinese are comparable in magnitude to those of previous work, e.g Meng et al. (2012); note we do 3-way classification instead of 2-way.

*Effect of code-switching* Table 3 shows the results with and without applying target-language lexicalization. The results with BQ and EP shows that the process of translating lexical items (code-switching) is essential to improving performance. This effect is not as pronounced or as consistent when using LDC, where the quality of

**Table 2** F1 macro-averaged scores of different methods in the single-source (English to target) transfer setting

| Language | Baselines | | Projection | | | Direct | | | | | Sup. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Neut* | *Senti* | *BQ* | *EP* | *LDC* | *CP* | *WK* | *BQ* | *EP* | *LDC* | |
| Arabic | 18.6 | 25.6 | 29.8 | – | 37.2 | 21.1 | 31.0 | **37.3** | – | 30.0 | – |
| Bulgarian | 22.4 | 30.6 | 29.4 | 38.7 | – | 28.6 | **45.3**‡ | 33.0 | 43.5 | – | 54.5 |
| Chinese | 19.7 | 35.6 | 39.8 | – | 52.7 | 32.7 | **66.8**‡ | 30.3 | – | 56.3 | – |
| Croatian | 12.8 | 19.8 | 24.2 | – | – | 13.9 | – | **30.8**‡ | – | – | 61.6 |
| German | 23.9 | 32.0 | 41.0 | 47.3 | – | 37.7 | **51.0**‡ | 43.5 | 45.4 | – | 59.9 |
| Hungarian | 16.5 | 29.2 | 29.9 | 38.1 | **47.0**‡ | 22.4 | 40.8 | 41.1 | 31.8 | 42.3 | 60.4 |
| Persian | 17.9 | 25.3 | 26.5 | – | 33.1 | 20.7 | 31.7 | **40.1**‡ | – | 26.5 | 67.8 |
| Polish | 13.8 | 26.0 | 27.9 | 38.8 | – | 30.2 | 32.9 | 41.7 | **43.3**‡ | – | 64.5 |
| Portuguese | 17.3 | 22.9 | 36.4 | **39.3** | – | 22.2 | 35.4 | 38.6 | **39.3** | – | 51.1 |
| Russian | 20.0 | 29.5 | 36.1 | – | 48.0 | 25.3 | 43.8 | 44.8 | – | **48.1** | 69.2 |
| Slovak | 11.9 | 19.2 | 23.3 | 30.0 | – | 24.6 | **36.6**‡ | 22.6 | 20.4 | – | 70.1 |
| Slovene | 20.6 | 28.1 | 31.8 | **44.6**‡ | – | 25.3 | 32.1 | 32.2 | 40.1 | – | 58.6 |
| Spanish | 19.3 | 25.3 | 36.0 | 41.8 | **42.7** | 27.7 | 37.7 | 42.6 | 39.4 | 42.2 | 45.4 |
| Swedish | 16.1 | 22.7 | 35.0 | 44.6 | – | 31.1 | 40.4 | 39.1 | **49.0**‡ | – | 62.5 |
| Uyghur | 23.6 | 36.7 | 37.4 | – | **38.6** | 25.7 | 28.0 | 30.0 | – | 37.5 | – |
| Average | 18.3 | 27.2 | 32.3 | 40.4 | 42.8 | 25.9 | 39.5 | 36.5 | 39.1 | 41.1 | 60.5 |

"*Neut*" predicts the majority label of the source data (English) and "*Senti*" refers to the baseline Sentiwordnet method. The "sup." column refers to the supervised LSTM model. "CP" refers to the use of embeddings extracted from comparable corpora, "WK" refers to the full direct model using the Wiktionary lexicon. "BQ", "EP" and "LDC" refer to the different parallel datasets used in our experiments: Bible and Quran as "BQ", Europarl as "EP" and the LDC in-domain parallel datasets.
‡ Indicates that the best result is significantly better than the second-best (and also other) numbers

**Table 3** Experimental results on the English to target direct transfer with and without code-switching ("− cs", "+ cs") on the labeled data

| Language | CP | BQ | | EP | | LDC | |
|---|---|---|---|---|---|---|---|
| | | −cs | +cs | −cs | +cs | −cs | +cs |
| Arabic | 21.1 | 26.4 | 37.3 | – | – | 36.7 | 30.0 |
| Bulgarian | 28.6 | 24.7 | 33.0 | 24.6 | 43.5 | – | – |
| Chinese | 32.7 | 16.5 | 30.3 | – | – | 55.9 | 56.3 |
| Croatian | 13.9 | 18.5 | 30.8 | – | – | – | – |
| German | 22.4 | 36.0 | 43.5 | 40.5 | 45.4 | – | – |
| Hungarian | 37.7 | 30.2 | 41.1 | 28.2 | 31.8 | 40.6 | 42.3 |
| Persian | 20.7 | 18.4 | 40.1 | – | – | 34.9 | 26.5 |
| Polish | 30.2 | 23.6 | 41.7 | 24.6 | 43.3 | – | – |
| Portuguese | 22.2 | 20.2 | 38.6 | 26.5 | 39.3 | – | – |
| Russian | 25.3 | 24.1 | 44.8 | – | – | 43.4 | 48.1 |
| Slovak | 24.6 | 15.1 | 22.6 | 17.9 | 20.4 | – | – |
| Slovene | 25.3 | 35.2 | 32.2 | 31.8 | 40.1 | – | – |
| Spanish | 27.7 | 27.0 | 42.6 | 27.3 | 39.4 | 40.9 | 42.2 |
| Swedish | 31.1 | 23.5 | 39.1 | 29.4 | 49.0 | – | – |
| Uyghur | 25.7 | 16.1 | 30.0 | – | – | 25.4 | 37.5 |
| Average | 25.9 | 23.7 | 36.5 | 27.9 | 39.1 | 39.7 | 41.1 |

"CP" is the results with comparable corpora, "BQ" is the Bible and Quran parallel data, "EP" is the Europarl data and "LDC" is the LDC parallel data. The underlined numbers are not significantly better than the all-neutral baseline

bilingual embeddings is already high enough that it suffices for training the direct model without much additional lexicalization.

*Comparable Corpora* We also compare the results without lexicalization to the results obtained using comparable corpora, which is a similar setting in which the direct model relies only on cross-lingual embeddings without any bilingual dictionary access. Both comparable and no-lexicalization perform significantly better (except where underlined) than the baseline. Moreover, we see that while the F-measure is not high, the comparable corpora still tend to perform better than BQ in the no code-switch configuration. Thus, when considering only the bilingual embeddings and not relying on any translation, in-domain comparable corpora tend to outperform out-of-domain parallel corpora. In this work we have built simple comparable corpora-derived bilingual embeddings by merging topical documents together, but with specialized comparable embeddings and lexicon bootstrapping, we can expect to observe more gains for sentiment analysis using in-domain comparable corpora. This is an interesting and still largely unexplored direction for future work.

### 6.2 Multi-source transfer

In this set of experiments, we use training datasets from all languages (except that of the target language) as source languages, with English now included as a target low-resource language. For projection, as described in Sect. 3, we applied majority voting on sentences aligned to translations in more than one language. For direct transfer,

**Table 4** F1 macro-average scores with different models (Projection, Direct, Ensemble) using different multi-parallel resources: Bible and Quran (BQ) and Europarl (EP) in the multi-source experiments

| Language | Projection | | Direct | | Ensemble | | Sup. |
|---|---|---|---|---|---|---|---|
| | BQ | EP | BQ | EP | BQ | EP | |
| Arabic | **45.8**‡ | – | 26.6 | – | 38.8 | – | – |
| Bulgarian | 38.4 | 42.7 | 41.6 | 48.4 | 43.4 | **49.1** | 54.5 |
| Chinese | **42.0** | – | 31.1 | – | 38.1 | – | – |
| Croatian | 42.5 | – | **43.4** | – | **43.4** | – | 61.6 |
| English | 45.8 | 42.1 | 50.6 | **54.0** | 51.9 | 53.9 | 65.3 |
| German | 47.5 | 43.2 | 51.3 | 50.3 | 50.0 | **54.7**‡ | 59.9 |
| Hungarian | 33.8 | 48.5 | 50.7 | **53.8**‡ | 51.0 | 51.6 | 60.4 |
| Persian | 37.2 | – | **43.7**‡ | – | 37.2 | – | 67.8 |
| Polish | 41.7 | 49.3 | 38.3 | **54.6**‡ | 45.1 | 52.1 | 64.5 |
| Portuguese | 38.5 | 34.0 | 34.3 | 37.8 | **38.9** | 37.9 | 51.1 |
| Russian | 41.6 | – | 46.2 | – | **48.3**‡ | – | 69.2 |
| Slovak | 37.9 | 45.7 | 48.0 | **48.2**‡ | 37.8 | 46.7 | 70.1 |
| Slovene | 39.6 | 44.7 | 42.3 | 46.5 | 45.6 | **48.8**‡ | 58.6 |
| Spanish | 33.7 | 40.9 | 42.9 | 43.5 | 44.0 | **45.2**‡ | 45.4 |
| Swedish | 45.0 | 47.7 | 44.6 | 46.5 | 43.9 | **49.8**‡ | 62.5 |
| Uyghur | **33.4** | – | 27.5 | – | 29.0 | – | – |
| Average | 40.3 | 43.9 | 41.4 | 48.4 | 42.9 | 49.0 | 60.8 |

‡ Indicates that the best number is significantly better than the second-best number

we experimented with the two techniques described in Sect. 4: (1) Concatenation, where we concatenate all the training sets from different languages and use that as the training data for a single model, and (2) Ensemble, where we train a separate direct model for each source language and then use the most frequent label assignment on the test data. We found that the concatenation approach performs best on average, so we show those results for the direct transfer model. Finally, we apply an ensemble of the projection and direct transfer approaches: (1) projection, (2) concatenation, and (3) single-source ensembles. If the results are tied, we back up to concatenation, projection and the English label respectively. Table 4 shows the results for different settings.

Comparing the single-source with the multi-source experiments, we see that having multiple source languages results in a clear improvement for both projection and direct transfer approaches, with significant improvements observed in most languages. The exceptions here are Chinese, where the single-source English–Chinese Wiktionary is especially good, Portuguese, where single-source is the same or slightly better, and Uyghur, where the English projection model with LDC gives better results than using multiple source languages; but this result is not statistically significant. In addition, direct transfer now almost always outperforms projection (10 out of 16 languages for BQ, 9 out of 10 languages for EP). For Arabic, Chinese, and Uyghur, projection is

better overall. We note that in these three languages, word order and syntactic structure can be quite different from the source and this may result in lower F-scores for direct transfer. [22]

The ensemble of approaches also enables some gains over the direct transfer model, particularly for the European languages, where the direct model is most effective. In fact, we see that for some of these languages, especially Spanish, the performance of the transfer method approaches that of the supervised model: this in particular is very interesting because the supervised models use gold labeled in-domain data with careful tuning while the transfer model is blindly trained on noisy or crafted datasets from other languages.

*Effect of language families* We notice that European languages (e.g German, English, Swedish, Spanish, Hungarian) tend to benefit a lot from the direct transfer model and in particular the multi-source direct transfer and ensemble models. On the other hand, languages like Arabic, Chinese, Uyghur benefit less from the direct transfer model and have larger gains with projection. The non-Indo-European families are less syntactically and semantically similar to the Indo-European source language families and are more likely to incur changes in structure and word ordering when moving from train to test. We note that Hungarian, which is a Uralic language (also not Indo-European) does not follow this pattern. This may be because it has some similarities with European and notably the Balto-Slavic languages.

To investigate whether certain languages transfer sentiment better from each other, we ran our direct transfer model separately using each source language and determined for each target language the source language with highest performance. The results are shown in Table 5 with BQ and EP.

For Arabic, Chinese, and Uyghur, which are the most different from the other languages, Slovak, Swedish and German are the best source languages. For the European languages, we notice some interesting trends; for example, the Germanic families (English, Swedish and German) transfer well from each other and additionally are good source languages in general. Slovene transfers well from its Southern Balto-Slavic siblings (Croatian and Bulgarian), and Slovak gets its best performance (48.7 with EP) with Polish (its Western Slavic sibling) as a source language. The Balto-Slavic languages (Slovak, Polish, Croatian, Slovene, Bulgarian, and Russian) generally transfer well to each other and to Hungarian. Surprisingly the Romance families (Portuguese and Spanish) are not the best source languages for each other. There are of course other factors involved, such as the quality and size of the training data in the different languages; more extensive experiments would be required to do a full language family analysis.

## 7 Error analysis

In order to better understand how the transfer model is doing and what kind of errors it makes, we conducted an error analysis where we considered English to be a target

[22] We note that the best scoring system for Arabic on the SemEval 2017 test set had a macro-average recall and accuracy of 58 when training with supervised Arabic data (Rosenthal et al. 2017).

**Table 5** F1 macro-average scores for best source language in the direct transfer model with Bible and Quran (BQ) and Europarl (EP) parallel corpora

| Language | BQ | EP |
|---|---|---|
| Arabic | 39.1 (Slovak) | – |
| Bulgarian | 44.3 (Swedish) | 44.8 (Swedish) |
| Chinese | 37.6 (Swedish) | – |
| Croatian | 40.4 (Slovak) | – |
| English | 51.7 (Swedish) | 49.0 (German) |
| German | 46.5 (Swedish) | 49.6 (Bulgarian) |
| Hungarian | 44.4 (Russian) | 48.8 (Polish) |
| Persian | 42.6 (English) | – |
| Polish | 42.2 (Bulgarian) | 50.7 (Hungarian) |
| Portuguese | 39.1 (Croatian) | 39.5 (Swedish) |
| Russian | 44.8 (English) | – |
| Slovak | 42.8 (Swedish) | 48.7 (Polish) |
| Slovene | 45.5 (Croatian) | 45.7 (Bulgarian) |
| Spanish | 42.6 (English) | 41.0 (German) |
| Swedish | 46.7 (German) | 49.0 (English) |
| Uyghur | 31.5 (German) | – |

low-resource language. We sampled 100 errors of the best performing English models for direct and projection, where at least one of the two models makes an error.

In our sample, the supervised model agrees with the gold label 63% of the time. Considering the 100 error samples, in 34% of cases, the direct model agrees with the gold label, in 26% of cases, the projection model agrees with the gold label, and in the remaining 40% of cases, both models make an error. Since the direct model performs better than projection when English is a target language, we analyzed the 66 samples from this model and compared them with the predictions of the supervised model in order to understand whether the errors come from the model itself or from the sentiment transfer.

Generally, we found that the source of sentiment errors comes from the following reasons: a key sentiment indicator was missed (e.g., “love,” “excited”, “bored”), there were misleading sentiment words (e.g., “super” in context of “getting up super early”, “handsome” in context of a question), the tweet contained mispelled/rare words (e.g., “bff,”, “bae”, “puta”), inference was required (e.g., “i need to seriously come raid your closet” is positive without containing positive words), the correct answer was not clear or not easily determined for a human annotator (e.g., “a mother’s job is forever”) , or the gold label was clearly wrong (e.g “thanks for joining us tonight! we kept it as spoiler free as possible!” has a neutral instead of positive gold label). There are thus many tweets in this error sample where the sentiment is not clear cut.

The samples analyzing the transfer are divided into four groups with examples provided for each:

1. In the first group (48.5% of cases), the supervised model makes a correct prediction, but the transfer model results in an error. Looking at examples in this group, we

found that this often occurs when the English target data contains rare, mispelled, or informal language words which are unlikely to have been learned using either cross-lingual representations or word alignments from parallel corpora.
   - "***fck na*** *! ! marshall ! bear nation hopes your* ***aight*** ! ! !" (negative, transfer predicts positive)
   - *"eagles might get* ***doored*** *tonight :'("* (negative, transfer predicts positive)
2. In the second group (26% of cases), the supervised model and the transfer model make the same error and thus the cause for the error likely comes from the model rather than the transfer. We determined that 6 of these cases have an incorrect gold label, and 11 result from errors of the supervised model where the answer was unclear, key sentiment was missed, or inference was required.
   - *"don 't let anyone discourage you from following your dreams ! it was one of the best decisions i made because it changed my lif..."* (gold negative [wrong], transfer predicts positive)
   - *"can 't wait to be an uncle again a wee boy this time , surely his names got to be jack if no , at least make it his middle name"* (gold positive [requires inference], transfer predicts neutral)
3. In the third group (16.6% of cases), the supervised and transfer model make different kinds of errors and thus the source of the error is likely from both the model itself and the transfer. We determined that three of these cases have an incorrect gold label, and the remaining eight are an error of the supervised model where the answer was unclear, key sentiment was missed, inference was required, or the sentence contained misleading sentiment words.
   - *"mount gambier that was rad and sweaty as hell , just one show left on the tour for us tomorrow in adelaide "* (gold positive [misleading sentiment], supervised neutral, transfer negative)
4. In the fourth group (9% of cases), the gold and supervised models agree, but the transfer model, which was trained on different data, actually makes a better prediction.
   - *"this photo taken on 9th September with high quality one of my bday gifts from my friend thank you brother"*(gold and supervised predict neutral, transfer predicts positive)

A large number of errors are clearly because of the transfer, but there are also several cases where even the supervised model makes the same error as the transfer model. The errors the transfer models make are reasonable because of the difficult nature of the Twitter data and the vocabulary it learns from out of domain (BQ or EP) parallel data.

## 8 Conclusion

We have developed and experimented with two transfer methods for sentiment analysis in a diverse set of scenarios. We explored the impact of the availability of different resources including multiple gold labeled datasets in rich-resource languages, parallel data of different genres and domains, comparable corpora, and manually and automatically generated dictionaries.

Our experiments show that using multiple source languages yields the best results for most target languages. Naturally, we saw that having similar genre and domain as the training data produces better results than out-of-domain and dissimilar genres; however, with our partial lexicalization strategy, out-of-domain parallel corpora still prove to be an effective low-resource setting.

We have shown that both projection and direct transfer are effective approaches for transferring sentiment in these low-resource scenarios. Direct transfer is the best performer with multiple source languages from related language families, and when high-quality parallel data is not available. Projection is the best performer when parallel in-domain data is available and with target languages which differ structurally from the source languages. We also found that the ensemble of projection and direct transfer approaches can lead to higher and more robust performance for several of the European languages.

In the situation where comparable data is available but a dictionary is not, our method using comparable data is just as good or better than using an out-of-domain parallel corpus in the same setting; bootstrapping dictionaries from comparable corpora is thus a direction for future work. We also demonstrated that having a high quality dictionary is essential for the direct transfer. We have seen that in different settings, different models do the best; thus future work should also explore strategies for model selection.

**Acknowledgements** Noura Farra, Kathleen McKeown and Axinia Radeva were supported by DARPA LORELEI Grant HR0011-15-2-0041. Kathleen McKeown and Tao Yu were supported by DARPA DEFT Grant FA8750-12-2-0347. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S government. We thank the reviewers for their detailed and helpful comments. We thank the Uyghur native informant and Appen for arranging the annotation. We thank Zixiaofan (Brenda) Yang for preparing the Uyghur evaluation data and meeting the informant.

## References

Abdul-Mageed M, Diab MT (2011) Subjectivity and sentiment annotation of modern standard Arabic newswire. In: Proceedings of the 5th linguistic annotation workshop, Association for Computational Linguistics, pp 110–118

Ammar W, Mulcaire G, Tsvetkov Y, Lample G, Dyer C, Smith NA (2016) Massively multilingual word embeddings. arXiv:1602.01925

Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC 10:2200–2204

Balahur A, Turchi M (2014) Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput Speech Lang 28(1):56–75

Berg-Kirkpatrick T, Burkett D, Klein D (2012) An empirical investigation of statistical significance in NLP. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, Association for Computational Linguistics, pp 995–1005. http://aclweb.org/anthology/D12-1091

Brooke J, Tofiloski M, Taboada M (2009) Cross-linguistic sentiment analysis: from english to spanish. In: RANLP, pp 50–54

Chang PC, Galley M, Manning CD (2008) Optimizing Chinese word segmentation for machine translation performance. In: Proceedings of the third workshop on statistical machine translation (StatMT '08), Association for Computational Linguistics, Stroudsburg, PA, pp 224–232. http://dl.acm.org/citation.cfm?id=1626394.1626430

Chen X, Sun Y, Athiwaratkun B, Cardie C, Weinberger K (2016) Adversarial deep averaging networks for cross-lingual sentiment classification. arXiv:1606.01614

Christodouloupoulos C, Steedman M (2014) A massively parallel corpus: the bible in 100 languages. Language resources and evaluation, pp 1–21

Duh K, Fujino A, Nagata M (2011) Is machine translation ripe for cross-lingual sentiment classification? In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies: short papers—volume 2, Association for Computational Linguistics (HLT '11), Stroudsburg, PA, pp 429–433. http://dl.acm.org/citation.cfm?id=2002736.2002823

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874

Faruqui M, Dyer C (2014) Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th conference of the european chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, pp 462–471. http://www.aclweb.org/anthology/E14-1049

Gouws S, Søgaard A (2015) Simple task-specific bilingual word embeddings. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, Association for Computational Linguistics, Denver, Colorado, pp 1386–1390. http://www.aclweb.org/anthology/N15-1157

Hermann KM, Blunsom P (2013) Multilingual distributed representations without word alignment. arXiv:1312.6173v4

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hosseini P, Ahmadian Ramaki A, Maleki H, Anvari M, Mirroshandel SA (2015) SentiPers: a sentiment analysis corpus for Persian

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 168–177

Joshi A, Balamurali A, Bhattacharyya P (2010) A fall-back strategy for sentiment analysis in Hindi: a case study. In: Proceedings of the 8th ICON

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. CoRR abs/1412.6980. arXiv:1412.6980v9

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. MT Summit 5:79–86

Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167

Meng X, Wei F, Liu X, Zhou M, Xu G, Wang H (2012) Cross-lingual mixture model for sentiment classification. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics (volume 1: long papers), Association for Computational Linguistics, Jeju Island, pp 572–581. http://www.aclweb.org/anthology/P12-1060

Mihalcea R, Banea C, Wiebe J (2007) Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th annual meeting of the association of computational linguistics, Association for Computational Linguistics, Prague, Czech Republic, pp 976–983. http://www.aclweb.org/anthology/P07-1123

Mozetič I, Grčar M, Smailović J (2016) Twitter sentiment for 15 European languages. http://hdl.handle.net/11356/1054, Slovenian language resource repository CLARIN.SI

Mukund S, Srihari RK (2010) A vector space model for subjectivity classification in Urdu aided by co-training. In: Proceedings of the 23rd international conference on computational linguistics: posters, Association for Computational Linguistics, pp 860–868

Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814

Neubig G, Dyer C, Goldberg Y, Matthews A, Ammar W, Anastasopoulos A, Ballesteros M, Chiang D, Clothiaux D, Cohn T, et al (2017) Dynet: the dynamic neural network toolkit. arXiv:1701.03980

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Pasha A, Al-Badrashiny M, Diab MT, El Kholy A, Eskander R, Habash N, Pooleery M, Rambow O, Roth R (2014) Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. LREC 14:1094–1101

Rasooli MS, Collins M (2017) Cross-lingual syntactic transfer with limited resources. Trans Assoc Comput Linguist 5:279–293. https://transacl.org/ojs/index.php/tacl/article/view/922

Rosenthal S, Farra N, Nakov P (2017) Semeval-2017 task 4: sentiment analysis in Twitter. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, pp 502–518. http://www.aclweb.org/anthology/S17-2088

Salameh M, Mohammad SM, Kiritchenko S (2015) Sentiment after translation: a case-study on Arabic social media posts. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, pp 767–777

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Citeseer, vol 1631, p 1642

Stratos K, Kim Dk, Collins M, Hsu D (2014) A spectral algorithm for learning class-based n-gram models of natural language. In: Proceedings of the association for uncertainty in artificial intelligence

Täckström O, McDonald R, Uszkoreit J (2012) Cross-lingual word clusters for direct transfer of linguistic structure. In: Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, Association for Computational Linguistics, pp 477–487

Vulić I, Moens MF (2016) Bilingual distributed word representations from document-aligned comparable data. J Artif Intell Res 55:953–994

Wan X (2008) Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the 2008 conference on empirical methods in natural language processing, Association for Computational Linguistics, Honolulu, Hawaii, pp 553–561. http://www.aclweb.org/anthology/D08-1058

Wang S, Manning CD (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, Association for Computational Linguistics, pp 90–94

Wick M, Kanani P, Pocock A (2015) Minimally-constrained multilingual embeddings via artificial code-switching. In: Workshop on transfer and multi-task learning: trends and new perspectives, Montreal, Canada

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp 347–354

Yu T, Hidey C, Rambow O, McKeown K (2017) Leveraging sparse and dense feature combinations for sentiment classification. arXiv:1708.03940

Zhang R, Lee H, Radev D (2016) Dependency sensitive convolutional neural networks for modeling sentences and documents. arXiv:1611.02361

Zhou G, He T, Zhao J (2014) Bridging the language gap: learning distributed semantics for cross-lingual sentiment classification. In: Zong C, Nie JY, Zhao D, Feng Y (eds) Nat Lang Process Chin Comput. Springer, Berlin, pp 138–149

Zhou H, Chen L, Shi F, Huang D (2015) Learning bilingual sentiment word embeddings for cross-language sentiment classification. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), Association for Computational Linguistics, Beijing, China, pp 430–440. http://www.aclweb.org/anthology/P15-1042

Zhou X, Wan X, Xiao J (2016a) Attention-based lstm network for cross-lingual sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Association for Computational Linguistics, Austin, Texas, pp 247–256. https://aclweb.org/anthology/D16-1024

Zhou X, Wan X, Xiao J (2016b) Cross-lingual sentiment classification with bilingual document representation learning. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers), Association for Computational Linguistics, Berlin, pp 1403–1412, http://www.aclweb.org/anthology/P16-1133