Cross-Lingual Emotion Lexicon Induction using Representation Alignment in Low-Resource Settings

Arun Ramachandran
Microsoft
Hyderabad, India
ramachandran.arun@outlook.com

Gerard de Melo
Hasso Plattner Institute, University of Potsdam
Potsdam, Germany
gdm@demelo.org

Abstract

Emotion lexicons provide information about associations between words and emotions. They have proven useful in analyses of reviews, literary texts, and posts on social media, among other things. We evaluate the feasibility of deriving emotion lexicons cross-lingually for over 350 languages, many of them resource-poor, from existing emotion lexicons in resource-rich languages. For this, we start out from very small corpora to induce cross-lingually aligned vector spaces. Our study empirically analyses the effectiveness of the induced emotion lexicons by measuring translation precision and correlations with existing emotion lexicons, along with measurements on a downstream task of sentence emotion prediction.

1 Introduction

Two main forms of classifying emotions are often distinguished: representing them along continuous dimensions, or breaking them into discrete categories (Stevenson et al., 2007; Calvo and Kim, 2013). A prominent instance of the former approach is the PAD model by Russell and Mehrabian (1977), which represents affect along 3 dimensions: *pleasure*, *arousal*, and *dominance*. An example of the latter is the Wheel of Emotions by Plutchik (1980), who argued that most emotions can be derived from a set of eight basic ones – *anger*, *fear*, *sadness*, *disgust*, *surprise*, *anticipation*, *trust*, and *joy*.

There have been efforts to create emotion lexicons, where each word is assigned either scores or discrete classes reflecting the associated emotions. Such lexicons are useful in emotional analyses of product reviews, literary texts, or posts on social media, inter alia. Bradley et al. (1999) solicited human affective norm ratings to create such a dataset for English based on the PAD model. Mohammad and Turney (2013) relied on crowdsourcing to annotate words with Plutchik's 8 basic emotions, providing binary labels. The recent NRC Emotion Intensity Lexicon (Mohammad, 2018) reconciles the notion of discrete emotions, corresponding to commonly invoked emotion names, with the benefits of continuous scoring in accounting for degrees of emotion intensity. Again relying on crowdsourcing, the lexicon provides intensity scores for Plutchik's eight basic emotions.

Affective norm ratings have as well been procured for certain other languages. An alternative route is to draw on automated techniques such as machine translation, as has been done for the NRC Emotion Intensity lexicon, where the English words are translated to other languages using Google Translate while retaining the original scores. Buechel et al. (2020) used Google Translate to translate a source emotion lexicon to a target lexicon that serves as training data, based on which valence/arousal/dominance or 5 basic emotions are predicted for a range of resource-rich languages. However, at the time of writing this paper, Google Translate serves around 100 languages. This raises the question of whether similar resources can be induced for resource-poor languages using minuscule amounts of data.

In this paper, we investigate simple means of deriving emotion ratings for resource-poor languages. In particular, we consider the case of drawing on very small corpora, focusing on partial translations of the Bible. We explore different cross-lingual embedding alignment techniques that allow us to transfer

Proceedings of the 28th International Conference on Computational Linguistics, pages 5879–5890
Barcelona, Spain (Online), December 8-13, 2020

English emotion ratings to over 350 languages, assessing the accuracy of translations and of our induced emotion ratings. We have made the resulting induced emotion lexicons freely available[1].

2 Related Work

2.1 Monolingual Emotion Lexicon Construction

Ground-truth emotion lexicons are typically constructed by manually annotating words with associated emotions. Bradley et al. (1999) aggregated results of a questionnaire to create an emotion lexicon with ratings for the PAD model, Warriner et al. (2013) compiled a similar dataset with larger coverage, and Shoeb and de Melo (2020) solicited emotion ratings for emojis. Crowd-sourcing platforms such as Amazon's Mechanical Turk can be used to expedite the annotation process (Mohammad and Turney, 2013), with techniques such as best-worst scaling to better account for the variance between crowd workers (Kiritchenko and Mohammad, 2016).

Apart from manual compilation, different strategies can be invoked to construct monolingual emotion lexicons automatically. For instance, the DepecheMood lexicon (Staiano and Guerini, 2014) was derived using statistical measures based on emotionally tagged text crawled from specific Web sites. Raji and de Melo (2020) revealed that unsupervised distributional semantics can outperform such supervised techniques.

2.2 Cross-Lingual Emotion Lexicon Induction

Leveau et al. (2012) showed that word translations across languages are strongly correlated in emotion. As machine translation gradually increased in accuracy, inducing affect-related resources cross-lingually become more feasible (Mihalcea et al., 2007). Lexicons for sentiment polarity have been induced cross-lingually using various forms of supervision (Chen and Skiena, 2014; Abdalla and Hirst, 2017; Barnes et al., 2018; Dong and de Melo, 2018b; Dong and de Melo, 2018a). In terms of emotion, Buechel et al. (2020) induced fine-grained emotion lexicons for the 91 languages for which Google Translate was available. However, machine translation tools are limited by the amount of available training data.

In recent years, induction has thus often been achieved by means of cross-lingual word embeddings. While numerous approaches for bilingual embedding training (Gouws and Søgaard, 2015) have been explored, it can be more convenient to draw on potentially larger amounts of monolingual data for embedding training and then achieve a post-hoc alignment of the embedding spaces. Mikolov et al. (2013) showed that word vectors in different languages can often be aligned with reasonably high accuracy using simple linear transformations. Xing et al. (2015) showed that enforcing orthogonality on the linear transformation matrix may result in better translation accuracy. There are now also several unsupervised alignment algorithms seeking to identify orthogonal transformations of embedding vector spaces (Lample et al., 2018; Artetxe et al., 2018; Grave et al., 2019). In this paper, we investigate such approaches for cross-lingual emotion lexicon induction.

Work so far has been limited in at least one of the following ways: 1) polarity lexicon induction as opposed to fine-grained emotion lexicons, 2) induction dependent on supervised data, or 3) unsupervised induction but with languages for which resources like Google Translate or pre-trained fastText embeddings are available. In the following sections, we present a method of emotion lexicon induction that works with resource-poor languages for which such tooling is unavailable.

3 Proposed Method

In this section, we introduce some pertinent definitions and provide a brief overview of our methodology to induce emotion ratings for resource-poor languages.

We consider a target language L_{T} that is typically a resource-poor one, for which no emotion ratings are available, and a source language L_{S}, for which emotion ratings are available. We define an emotion rating $\sigma_e(w) \in [0,1]$ as an emotion intensity score, i.e., the degree of emotional association of word w with emotion $e \in \mathcal{E}$ for a set of target emotions $\mathcal{E}$. Accordingly, an emotion lexicon E can be regarded

[1] http://emotionlexicon.org/

as a function of the form $\mathcal{V} \times \mathcal{E} \rightarrow [0, 1]$ that maps words w from a vocabulary $\mathcal{V}$ paired with emotions $e \in \mathcal{E}$ to word–emotion ratings $\sigma_e(w)$.

Our method requires three resources: a monolingual text corpus for each of L_T and L_S, and an emotion lexicon E_S for the source language. We induce a target-language E_T in a two-step process: First, we induce a cross-lingual word embeddings space covering both L_T and L_S, by drawing on the monolingual corpora as well as unsupervised cross-lingual alignment (Section 4). Subsequently, we derive emotion ratings for L_T using this vector space, based on the source lexicon E_S (Section 5).

Our empirical investigations focus mainly on the first step. We evaluate three algorithms to induce cross-lingual word embeddings in such low-resource settings. We also explore additional supervision when the input corpora possess sentence-level alignments, that is, information about which sentences in L_T are translations of which sentences in L_S. This is the case for the Bible translations considered in this study, due to the presence of verse identifiers.

4 Cross-Lingual Embedding Induction

In this section, we explain our overall approach to obtain cross-lingually aligned word embeddings, and then briefly outline three of the algorithms we use for alignment along with our modifications.

4.1 Approach

For each input corpus, we first invoke the fastText skip-gram algorithm to learn monolingual word embeddings (see Section 6.2 for details). The text in each monolingual corpus is preprocessed to eliminate all Unicode punctuation and converted to lower case. We obtain two embedding matrices $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$ for the source and target languages, respectively, with corresponding vocabularies $\mathcal{V}_\mathrm{S}$ and $\mathcal{V}_\mathrm{T}$.

Our goal is to induce a single cross-lingual embedding matrix $\mathbf{X}_\mathrm{C}$ that covers both $\mathcal{V}_\mathrm{S}$ and $\mathcal{V}_\mathrm{T}$ in a single space. For this, we explore three algorithms to align $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$: Wasserstein-Procrustes (Grave et al., 2019), Unsupervised Orthogonal Refinement (Artetxe et al., 2018), and a neural language model (Wada et al., 2019). We also consider modifications of the latter two algorithms and evaluate these modified variants alongside the original ones. Note that the neural language model does not require word embeddings to have already been trained on monolingual corpora, as it jointly trains on two corpora to produce embeddings that already reside in a common space. Thus, only the preprocessing steps are performed for it. In the following, we describe each of these techniques in more detail.

4.2 Wasserstein-Procrustes

Given two matrices $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$ containing word embeddings in two different languages, the Wasserstein-Procrustes technique by Grave et al. (2019) calculates a projection matrix such that the Euclidean distances of the projected embeddings are minimized:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}}\, ||\mathbf{X}_\mathrm{S}\mathbf{W} - \mathbf{X}_\mathrm{T}||_2$$

This is done in an iterative fashion by alternatively a) finding a permutation of $\mathbf{X}_\mathrm{T}$ that minimizes the above equation, then b) using stochastic gradient descent to move to a more optimal value of $\mathbf{W}$ and then using singular value decomposition to obtain the nearest orthogonal matrix. Grave et al. also use an initialization wherein they employ a convex relaxation of the equation they try to optimize in the iterative phase, allowing them to solve for an approximation of the orthogonal matrix $\mathbf{W}$ in the above equation. Ultimately, we obtain the final cross-lingual embedding matrix $\mathbf{X}_\mathrm{C} = \begin{bmatrix} \mathbf{X}_\mathrm{S}\mathbf{W} \\ \mathbf{X}_\mathrm{T} \end{bmatrix}$.

4.3 Unsupervised Orthogonal Refinement

Artetxe et al. (2018) presented another algorithm for unsupervised alignment. The goal is to compute orthogonal transformation matrices $\mathbf{W}_\mathrm{S}$ and $\mathbf{W}_\mathrm{T}$ to align embedding matrices $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$ in the same embedding space, while also building a bidirectional translation mapping between the words in either language. This is achieved in four steps:

1. **Normalization.** Embedding matrices are length normalized, then centered around the mean dimension-wise, then normalized again (to obtain unit vectors for each embedding).
2. **Unsupervised initialization.** In this step, $\pi(\sqrt{\mathbf{M}_\mathrm{S}})$ and $\pi(\sqrt{\mathbf{M}_\mathrm{T}})$ are computed, where $\mathbf{M}_\mathrm{S} = \mathbf{U}\mathbf{S}^2\mathbf{U}^\intercal$ for $\mathbf{U}\mathbf{S}\mathbf{V}^\intercal = \mathrm{SVD}(\mathbf{X}_\mathrm{S})$ (making $\mathbf{M}_\mathrm{S}$ the SVD of $\mathbf{X}_\mathrm{S}\mathbf{X}_\mathrm{S}^\intercal$), and similarly for $\mathbf{M}_\mathrm{T}$. Here, π sorts each row of its operand in descending order. The idea is that $\pi(\mathbf{X}_\mathrm{S}\mathbf{X}_\mathrm{S}^\intercal)$ and $\pi(\mathbf{X}_\mathrm{T}\mathbf{X}_\mathrm{T}^\intercal)$, unlike $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$, are approximately identical up to a permutation of their rows (an assumption that has already been made in the form of assuming the embedding spaces for different languages are at least approximately isometric, as otherwise without it, attempting to find orthogonal mapping matrices is a futile effort). These sorted matrices are then used to compute an initial bilingual dictionary using step b) of the next phase.
3. **Iterative refinement.** The orthogonal mapping matrices and the bilingual dictionary are iteratively refined by repeating two steps until convergence: a) Compute the optimal orthogonal mapping matrices $\mathbf{W}_\mathrm{S}$ and $\mathbf{W}_\mathrm{T}$ such that similarities for words that translate to each other in the bilingual dictionary are maximized. b) Compute the optimal bilingual dictionary by using a variation of nearest neighbors to identify words in the other language that are closest in the aligned embedding space. The exact scoring mechanism for computing the nearest neighbors is discussed later in Section 5. This phase employs an annealing dropout-like mechanism that randomly deletes entries from the bilingual dictionary to help escape poor local optima.
4. **Final refinement.** After the previous iterative phase converges on a solution, the mapping matrices are re-weighted according to the cross-correlation in each component, increasing the relevance of those dimensions that best match across languages.

4.4 Orthogonal Refinement with Sentence Alignment Initialization

We modified the technique from Section 4.3 for the setting of sentence-level alignments being available, as is the case for the Bible translations that we consider in this study. To exploit this auxiliary source of supervision, we modified the unsupervised initialization phase, the second of the four phases described in Section 4.3. Normally, this step hinges on the assumption that words that are translations of each other have similar statistical distributions. Starting from matrices $\mathbf{X}_\mathrm{S}$, $\mathbf{X}_\mathrm{T}$ whose rows contain word embeddings trained on monolingual corpora, an initial bilingual dictionary is induced. This is then iteratively refined in the subsequent phase.

Rather than use word embedding matrices, we modified this phase to align term–sentence matrices $\mathbf{D}_\mathrm{S}$, $\mathbf{D}_\mathrm{T}$. These are sparse matrices whose rows correspond to words and columns correspond to sentences. Each entry reflects the count of words in that sentence. Thus, we compute $\mathbf{U}\mathbf{S}\mathbf{V}^\intercal = \mathrm{SVD}(\mathbf{D}_\mathrm{S})$, such that $\mathbf{M}_\mathrm{S} = \mathrm{SVD}(\mathbf{D}_\mathrm{S}\mathbf{D}_\mathrm{S}^\intercal)$, and likewise for $\mathbf{M}_\mathrm{T}$ based on $\mathbf{D}_\mathrm{T}$. In our experiments described in Section 7.1, we find that this greatly enhances the robustness of the approach.

4.5 Neural Language Model

Finally, we consider a neural language model for unsupervised joint representation induction, as proposed by Wada et al. (2019). The idea is to use jointly-trained forwards and backwards LSTMs trained on monolingual corpora from multiple languages. Different word embedding layers and decoders are used for each language, but weights in the hidden layers are shared, along with the embeddings for the beginning and end-of-sentence tokens, and the weights for calculating the probability of the end-of-sentence token. The shared weights encourage the word embeddings across different languages to be encoded in roughly the same space. After training, the initial word embedding layer weights are used to project word tokens into the same aligned embedding space.

We also investigated a variant of this technique, replacing the LSTMs in the model with QRNNs (Bradbury et al., 2017), and adopting one-cycle learning rate scheduling (Smith, 2018) to reduce the training time and improve the model's precision.

Method	SPA	HIN	NLD	ELL	RUS	YOR	GLA	SIN	MRI
Procrustes	34.7	32.8	48.9	0.0	0.1	25.2	36.6	0.0	36.1
	10.9	6.5	11.2	0.0	0.02	**1.0**	5.0	0.0	7.1
Orth. Ref.	**38.6**	**34.8**	52.3	0.1	1.0	0.0	**39.5**	**24.1**	0.2
	12.1	**6.9**	12.0	0.02	0.2	0.0	**5.4**	**1.2**	0.04
NLM	23.0	2.4	35.8	11.7	4.7	2.4	3.6	0.4	6.1
	7.6	0.51	8.5	2.4	1.0	0.1	0.5	0.02	1.3
Mod. Orth. Ref.	38.2	34.1	**53.6**	**36.0**	**34.1**	**27.1**	**39.5**	**24.1**	**38.4**
	12.0	6.8	**12.3**	**7.2**	**7.0**	**1.0**	**5.4**	**1.2**	**7.6**
Mod. NLM	36.6	5.4	47.2	24.7	10.4	6.9	15.6	2.1	11.2
	12.1	1.1	11.2	5.1	2.2	0.3	2.3	0.1	2.4

Table 1: Precision@3 for nine languages. The top row for each method considers the subset of the gold bilingual dictionary excluding out-of-vocabulary words. The bottom row considers the entire gold dictionary, treating out-of-vocabulary words as incorrect. Top precision scores are marked in bold.

5 Cross-Lingual Emotion Rating Induction

Equipped with our cross-lingual embedding space $\mathbf{X}_{\mathrm{C}}$, we are now able to induce emotion ratings cross-lingually based on the source language emotion lexicon E_{S}. For each target language word $w \in \mathcal{V}_{\mathrm{T}}$ and each emotion $e \in \mathcal{E}$, we compute a score

$$\sigma_e(w) = \frac{1}{|T_w|} \sum_{w' \in T_w} \sigma_e(w'), \tag{1}$$

where $\sigma_e(w')$ is the emotion rating of a word w' from L_{S} according to the source emotion lexicon E_{S}, and

$$T_w = \underset{W \subset \mathcal{V}_{\mathrm{S}}, |W|=k}{\operatorname{argmax}} \sum_{w' \in W} \mu(\mathbf{v}_w, \mathbf{v}'_w), \tag{2}$$

i.e., the set of $k = 3$ words w' from the source language vocabulary $\mathcal{V}_{\mathrm{S}}$ that are most related to w in terms of the corresponding cross-lingual word vectors $\mathbf{v}_w$, $\mathbf{v}_{w'}$ from $\mathbf{X}_{\mathrm{C}}$.

To compute the relatedness $\mu(\mathbf{v}_w, \mathbf{v}'_w)$, we adopt Cross-Domain Similarity Local Scaling (CSLS) scores. CSLS assesses the relatedness between two word embeddings $\mathbf{v}_1$ and $\mathbf{v}_2$ from two different languages L_1 and L_2 as follows:

$$\mu(\mathbf{v}_1, \mathbf{v}_2) = 2\frac{\mathbf{v}_1^\intercal \mathbf{v}_2}{||\mathbf{v}_1||\,||\mathbf{v}_2||} - R_{L_1}(\mathbf{v}_2) - R_{L_2}(\mathbf{v}_1) \tag{3}$$

$$R_{L_i}(\mathbf{v}) = \frac{1}{K} \sum_{\mathbf{v}' \in N_{L_i}(\mathbf{v})} \frac{\mathbf{v}^\intercal \mathbf{v}'}{||\mathbf{v}||\,||\mathbf{v}'||} \tag{4}$$

The advantage of CSLS over simple cosine similarities is that it compensates for *hubness*, the property that some vectors in an embedding space reside near overly many other vectors (Lazaridou et al., 2015). It achieves this by subtracting hubness factors $R_{L_1}(\mathbf{v}_2)$ and $R_{L_2}(\mathbf{v}_1)$ for $\mathbf{v}_1$, $\mathbf{v}_2$, where $R_{L_i}(\mathbf{v})$ yields the average cosine similarity of the $K = 10$ nearest neighbors of $\mathbf{v}$ in the other language L_i.

6 Experimental Setup

In the following, we present an empirical analysis of the feasibility of inducing emotion ratings using the above methods when drawing on very small monolingual corpora. We first present our data sources (Section 6.1) and algorithmic parameters (Section 6.2), and then discuss various methods of measurement to verify the effectiveness of our methods (Section 6.3). The results follow in Section 7.

Method	SPA	HIN	COS	EST	KIR	LTZ
	7.9K	8K	3.8K	9.4K	11K	8K
Procrustes	1.1	0.8	0.6	0.5	0.2	0.0
	0.3	0.1	0.02	0.04	0.02	0.0
Orth. Ref.	0.4	0.0	0.0	0.2	0.2	0.0
	0.1	0.0	0.0	0.02	0.02	0.0
NLM	8.0	0.7	1.1	**2.1**	0.2	0.0
	2.4	0.1	0.06	**0.2**	0.02	0.0
Mod. Orth. Ref.	4.7	**7.7**	0.5	0.8	**1.1**	**2.8**
	1.3	**1.2**	0.02	0.06	**0.1**	0.1
Mod. NLM	**16.1**	0.9	**1.9**	1.3	0.0	2.0
	4.8	0.2	**0.1**	0.1	0.0	**0.2**

Table 2: Precision@3 for six languages. The training data for these languages was much smaller in size. The number of sentences in each language is in the table header, below each language code. The rest of the layout is similar to Table 1.

6.1 Data Sources

Languages and Corpora. For data to train and align word embeddings, we crawled Bible texts for around 1,600 languages from several sources.[2] Each of these languages differ in the number of Bible verses available. Around 350 of these languages have at least 30K verses available (for comparison, the English King James Version has 31,102 verses).

We used English as our resource-rich language L_{S}. We selected our resource-poor languages L_{T} in two groups. We picked nine languages that had the full 31K verses present in one group. In this group, Spanish, Hindi, Dutch, Greek, and Russian are present. While these are not actually resource-poor, we included these to have a useful point of reference against which to compare the performance of our methods with other languages. This group also includes Yoruba, Scots Gaelic, Sinhala, and Maori, languages that have fewer speakers and less data available on the Internet.

In our second group, we picked six languages that had around 10K or fewer verses available. We picked Spanish and Hindi as reference languages again, this time with Bible translations including only the New Testament (around 8K verses). We also picked Corsican, Estonian, Kyrgyz, and Luxembourgish, for which the only Bibles we obtained were ones with around 10K or fewer verses.

Source Lexicon. For the emotion lexicon in English (E_{S}), we used the NRC Emotion Intensity Lexicon (EIL) by Mohammad (2018). The NRC EIL contains English words with real-valued intensity scores for eight basic emotions – *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*.

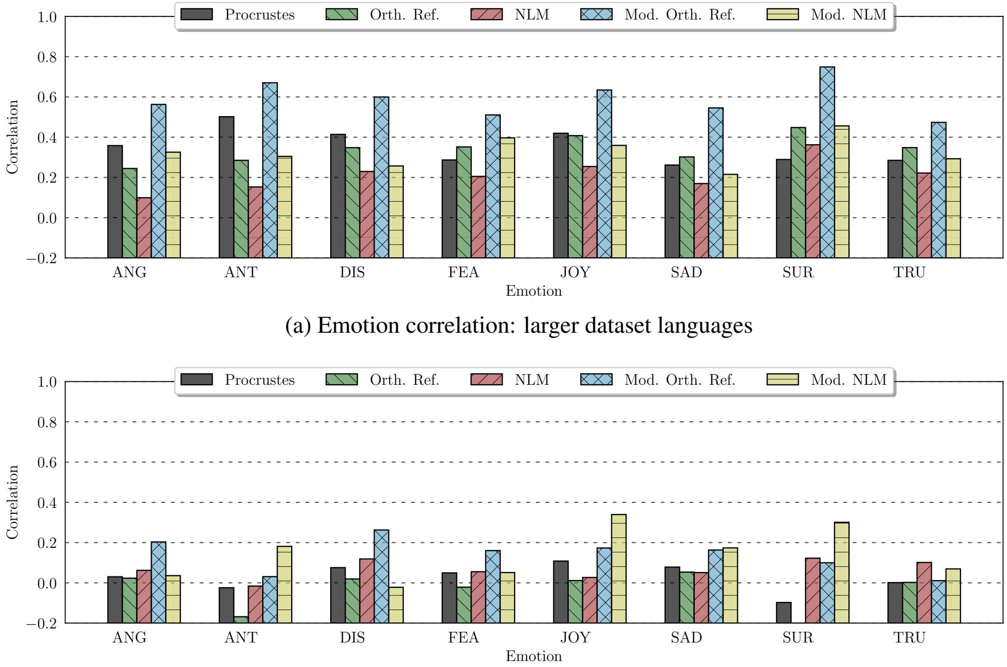
Ground Truth. The NRC EIL also includes emotion lexicons for around 100 other languages obtained by translating the English words using Google Translate (note that we have fully translated Bibles for over 350 languages, so we are able to cover many more languages than the NRC EIL does). The NRC EIL's machine-translated emotion lexicons serve as a silver standard ground truth against which we compare the emotion ratings we induce using our methods.

6.2 Settings and Parameters

When creating fastText skip-gram embeddings for Wasserstein-Procrustes and Orthogonal Refinement, for each language, we trained for 25 epochs with a learning rate of 0.1 and learned 100-dimensional embeddings. These were created only for words with a frequency count of 5 or greater when training on Bibles with 31k sentences, while the frequency cutoff was set to 2 for smaller Bibles with fewer translated sentences.

For Orthogonal Refinement, we used the same settings as the original version by Artetxe et al. (2018). For our variant from Section 4.4, we modified the initialization phase. We picked the common verses

[2] We considered digitalbibleplatform.com, png.bible, and bible.com



(a) Emotion correlation: larger dataset languages

(b) Emotion correlation: smaller dataset languages

Figure 1: Pearson Correlation of induced emotion ratings with the NRC EIL. The top figure shows ratings for each emotion averaged across the larger dataset languages (from Table 1). The bottom one shows the same, except with ratings averaged across the smaller dataset languages (from Table 2).

from Bibles in L_{S} and L_{T} and used those to create the term–sentence frequency matrix. We also trained the initial fastText embeddings only on the common verses. The remaining hyperparameters for the alignment were the same as for the original version.

For the neural language model, we used the same settings as in the original paper by Wada et al. (2019), except for an increase in the number of epochs from 10 to 20. This was to match the number of epochs used in our modified model, so as to provide a fair comparison. For our modified variant of the neural language model, we used SGD optimization and set the maximum learning rate for the one-cycle scheduling to 0.2, training the model for 20 epochs. We used similar frequency count cutoffs as with the fastText embeddings, except for setting the threshold for English to 3, as this worked better empirically. We trained a model for each language pair L_{S}, L_{T}.

6.3 Measurement Methods

Cross-Lingual Embedding Quality. To assess the quality of the cross-lingual embeddings, we used the bilingual dictionaries with 5k word translations from Lample et al. (2018) for the languages for which they are available as a gold standard. For others, we used the NRC EIL, as it contains English words that are machine-translated to other languages to assign them emotion ratings. We report two metrics for each language L_{T}:

a) We take each word in L_{T} present in the gold standard dictionary, but we remove out-of-vocabulary words not present in our corpus vocabulary $\mathcal{V}_{\mathrm{T}}$, as these are irrelevant for our later downstream emotion ratings task. On this set, we calculate the fraction of words for which our cross-lingual embeddings $\mathbf{X}_{\mathrm{C}}$ yield the correct translations according to Eq. 3, in terms of precision at $k = 3$.
b) For comparison, we also report the same precision at $k = 3$ scores as above, but without eliminating out-of-vocabulary words. Here, if a word in the gold standard dictionary is not present in our induced dictionary, we simply count it as incorrectly translated.

Emotion Rating Induction. To evaluate the accuracy of our emotion ratings for each language L_{T}, we take the intersection of words in the NRC EIL and in the respective target corpus vocabulary $\mathcal{V}_{\mathrm{T}}$, and

Emotion	SPA	HIN	NLD	ELL	RUS	YOR	GLA	SIN	MRI
	9.3K	7.1K	7.8K	10.7K	11.5K	4.2K	8.0K	12.9K	3.4K
Anger	0.635	0.421	0.672	0.628	0.672	0.676	0.438	0.688	0.234
Anticipation	0.794	0.589	0.791	0.649	0.831	0.735	0.690	0.415	0.538
Disgust	0.657	0.225	0.740	0.554	0.766	0.682	0.844	0.635	0.293
Fear	0.484	0.434	0.623	0.517	0.467	0.589	0.549	0.620	0.313
Joy	0.636	0.551	0.796	0.564	0.733	0.504	0.729	0.659	0.536
Sadness	0.577	0.592	0.603	0.471	0.729	0.639	0.455	0.271	0.572
Surprise	0.918	0.652	0.915	0.806	0.898	0.614	0.751	0.933	0.253
Trust	0.518	0.329	0.688	0.439	0.398	0.485	0.630	0.432	0.342

Table 3: Induced emotion ratings using our variant of Orthogonal Refinement. These ratings are for the nine large dataset languages. The bottom row of the header is the size of the induced emotion lexicon, calculated by counting the number of word–emotion pairs for a given language.

calculate the Pearson correlation coefficient for each language and each emotion. Unlike with translation precision, we do not also consider results without eliminating out-of-vocabulary words, as very few words per emotion (typically less than 100) are shared by both our induced dictionary and the NRC EIL, while thousands of words per emotion are often present in either the NRC EIL or in our induced emotion ratings individually. Thus, calculating the correlation on the entire set does not yield meaningful results.

7 Results

We present the evaluation of cross-lingual embeddings in Section 7.1 and of the resulting emotion ratings in Section 7.2. Additionally, we conduct a case study on sentence-level emotion ratings in Section 7.4.

7.1 Cross-Lingual Embedding Induction

In Tables 1 and 2, we provide the evaluation of our cross-lingual embedding induction phase in terms of translation precision. Table 1 considers the set of languages with the full 31K verses of translated Bible text. We observe that Wasserstein-Procrustes is frequently outperformed by Orthogonal Refinement, although the latter fails entirely for a greater number languages.

Exploiting parallel information, our modified Orthogonal Refinement is substantially more robust and obtains the best results for most of the languages, losing out on just a few to the original Orthogonal Refinement. However, the original method is not as robust, failing to arrive at embedding alignments for 4 out of 9 languages. Our initialization procedure, while not affecting precision much where alignments could already be found without it, appears to aid in bootstrapping the alignment process. While our procedure is clearly less scalable than operating on the word embedding matrices, on our datasets with just 31k sentences or fewer, the computations could be performed on a single GPU in just a few minutes. Hence, we conclude that our variant is best-suited for small aligned corpora, whereas for large corpora the original method is likely to work well enough.

Our variant of the Neural Language Model (NLM) performs significantly better than the original by Wada et al. (2019), and also is more robust than the original Orthogonal Refinement. However, it does not prevail over our variant of Orthogonal Refinement.

Table 2 provides the results for languages with around 10K or fewer verses translated. Across the board, all algorithms fail to achieve satisfactory results. Our algorithm variants show slightly better results than the original methods, but the absolute precision remains low. It appears that such neural representation learning methods require more data in order to start arriving at robust embeddings suitable for accurate translation induction.

7.2 Emotion Ratings

The correlation of induced emotion ratings with those in the ground truth are presented in the graphs in Figure 1, reported separately for each method and emotion. Figure 1a considers the set of nine languages with large datasets (as listed in Table 1), and each reading was obtained by averaging the coefficients across the nine languages. Our variant of Orthogonal Refinement attains the best scores across all emotions. Figure 1b is similar to Figure 1a but presents the correlations for the smaller dataset languages (those listed in Table 2). As expected, the correlation scores are generally lower here compared to the languages with larger datasets, confirming that such minuscule amounts of training data are insufficient to induce emotion ratings using our methods.

7.3 Qualitative Analysis

To better understand in what ways our induced emotion ratings deviated from the NRC EIL, we performed a qualitative analysis. We took the 50 Spanish, 30 Yoruba and 30 Sinhala words whose induced emotion rating deviated the most from the NRC EIL's and labeled each of them with the cause of error based on our inspection of the nearest source language neighbors and their corresponding emotion ratings. The results of this analysis are presented in Table 4. The following discussion focuses on Spanish, as the error categories are essentially the same for Yoruba and Sinhala.

Error Category	Frequency		
	SPA 50	YOR 30	SIN 30
Random Mistranslation	21	18	21
Exaggeration	10	2	2
Ambiguity	6	0	0
Antonym Mistranslation	4	0	3
Adjacent Mistranslation	4	3	2
Questionable NRC Ratings	3	7	1
Random Addition	2	0	1

Table 4: Emotion error category frequencies. The number below the language code in the header indicates the number of words analyzed.

While 21 of the Spanish word errors appeared to be due to random mistranslations with no identifiable patterns, we were able to categorize the remaining 29. 10 of these seemed to be exaggerated translations (*falso* to *murderer* instead of just *false*, *engañar* to *evil* instead of just *cheat*, *cambio* to *turmoil* instead of *change*). An interesting theme here is the exaggeration of words for deceit, cunning, and falsehoods (*astucia*, supposed to be *cunning* or *craftiness*, was translated to *evil*, *hatred*, and *slander*). Such shifts may stem from the biblical narrative in our source corpora, which may diverge from common use.

The next frequent issue is ambiguity, where a word has multiple meanings and the NRC EIL picked one, while our methods picked another. We also observed words being translated to their antonyms (*calma* to *madness* instead of *calm*, *contento* to *ruin* and *sad* instead of *happiness*) and adjacent ideas (*ayuda* to *distress* instead of *aid*, *médico* to *disease* instead of *doctor*). This is an expected result when drawing on distributional semantics, as antonyms and adjacent concepts appear in similar contexts as the original words. Finally, we encounter the issue of correct translations for a word being the top ones, but incorrect translations also getting included at the end of the list, which ends up skewing the final emotion rating.

A notable deviation from the error category frequency pattern is that of questionable NRC ratings for Yoruba. Inspecting Yoruba literature corpus searches yielding translations in context, we noticed that the NRC translations appeared to be incorrect surprisingly frequently, which led to our emotion ratings deviating significantly from those of the NRC for these mistranslated words.

7.4 Sentence-Level Evaluation

As an additional case study, we also evaluate our induced emotion ratings on the downstream task of predicting the emotion of sentences in an unsupervised manner.

Data. Due to the scarcity of emotion-labeled corpora for low-resource languages, we here rely on the Spanish language LiSSS corpus (Torres-Moreno and Moreno-Jiménez, 2020), but again induce our Spanish ratings using our corpora of just 31k / 7.9k Bible verses. LiSSS provides around 500 sentences from the literary domain, each manually annotated with one or more of five emotions – *love*, *fear*, *happiness*,

Method	Precision	
	Unweighted	IDF
NRC EIL	0.598	0.607
Procrustes	0.477	0.470
Orth. Ref.	0.551	0.540
NLM	0.292	0.262
Mod. Orth. Ref.	0.467	0.491
Mod. NLM	0.287	0.329

(a) Alignment methods trained on 31K Spanish Bible.

Method	Precision	
	Unweighted	IDF
Procrustes	0.306	0.301
Orth. Ref.	0.273	0.266
NLM	0.283	0.301
Mod. Orth. Ref.	0.290	0.292
Mod. NLM	0.262	0.252

(b) Alignment methods trained on 7.9K Spanish Bible.

Table 5: Precision of induced emotion ratings from various alignment methods on the LiSSS corpus.

anger, and *sadness*. We dropped the sentences labeled exclusively with *love*, as that is not an emotion present in the NRC EIL. We were left with 428 sentences, which we used for evaluation.

Method. Given a sentence S, we predict its emotion as

$$\underset{e \in \mathcal{E}}{\operatorname{argmax}} \sum_{w \in S} \lambda_w \sigma_e(w), \tag{5}$$

where $\mathcal{E}$ is the set of four candidate emotions. We consider two different weighting schemes: The first simply sets $\lambda_w = 1$, while the second sets it to the the IDF score of w in the Spanish Bible corpus. For sentences labeled with a single emotion, the predicted emotion must match it to be counted as correct. For sentences labeled with multiple emotions, the predicted emotion must be among the true emotions.

Results. Table 5a shows the results of the evaluation against the LiSSS corpus when our method is trained on the full 31K verse Spanish Bible. For reference, the expected precision that random guessing would achieve is 0.271. The NRC EIL, as expected, does the best, as it used Google Translate, while our methods had only small Bible corpora as training data. The NRC EIL also shows a slight improvement upon adding IDF weighting. Interestingly, the three unmodified alignment methods produce emotion ratings that actually do better without IDF weighting. We conjecture that this is because the translation precision for rarely seen words is too low in these methods for IDF to be effective. Another interesting observation is that while the our modified Orthogonal Refinement and NLM methods attained a comparable translation precision, this does not correlate with comparable precision on the LiSSS corpus. In fact, NLM does hardly better than chance, while our modified Orthogonal Refinement does almost twice as well as chance. Table 5b shows the results of the LiSSS evaluation when training only on the 7.9K verses Bible version. Here, none of the methods do much better than chance.

8 Conclusion

In this paper, we investigate approaches to cross-lingually induce emotion ratings based on very small training corpora. This is achieved by taking an emotion lexicon for a resource-rich language and inducing a cross-lingual embedding space to transfer the source language emotion ratings to words in the resource-poor target languages. We compare several strategies to achieve this and evaluate them in terms of both translation precision and the final correlation of the induced emotion ratings with existing emotion lexicons. Generally, we find that our modified variants of the original algorithms yield important gains in such low-resource settings. We also evaluate them on the downstream task of unsupervised sentence-level emotion prediction on a human-annotated literary corpus. Overall, while the methods do not work sufficiently well with 10K or fewer verses, we attained satisfactory results on languages for which translated Bibles with at least 31K verses exist. This still leaves us with the ability to induce cross-lingual emotion ratings for over 350 languages, available online at `http://emotionlexicon.org/`.

References

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 1, pages 506–515.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 789–798.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2483–2493.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-Recurrent Neural Networks. In *International Conference on Learning Representations*.

Margaret M. Bradley, Peter J. Lang, Margaret M. Bradley, and Peter J. Lang. 1999. Affective norms for english words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217.

R. Calvo and S. Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 383–389.

Xin Dong and Gerard de Melo. 2018a. Cross-lingual propagation for deep sentiment analysis. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 5771–5778. AAAI Press.

Xin Dong and Gerard de Melo. 2018b. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of ACL 2018*, pages 2524–2534.

Stephan Gouws and Anders Søgaard. 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*, pages 1386–1390.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of Machine Learning Research*, volume 89, pages 1880–1890.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817.

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

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. 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, pages 270–280.

Nicolas Leveau, Sandra Jhean-Larose, Guy Denhiere, and Ba-Linh Nguyen. 2012. Validating an interlingual metanorm for emotional analysis of texts. *Behavior research methods*, 44, 07.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv:1309.4168 [cs.CL].

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33.

Shahab Raji and Gerard de Melo. 2020. What sparks joy: The AffectVec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997, New York, NY, USA. ACM.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Abu Awal Md Shoeb and Gerard de Melo. 2020. Emotag1200: Understanding the association between emojis and emotions. In *Proceedings of EMNLP 2020*.

Leslie N. Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv:1803.09820 [cs.LG].

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 427–433.

Ryan Stevenson, Joseph Mikels, and Thomas James. 2007. Characterization of the affective norms for english words by discrete emotional categories. *Behavior research methods*, 39:1020–4.

Juan-Manuel Torres-Moreno and Luis-Gil Moreno-Jiménez. 2020. Lisss: A toy corpus of spanish literary sentences for emotions detection. arXiv:2005.08223 [cs.CL].

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13, 915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.