A Multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction

Mengjie Zhao and **Hinrich Schütze**
CIS, LMU Munich, Germany
mzhao@cis.lmu.de

Abstract

We present a new method for sentiment lexicon induction that is designed to be applicable to the entire range of typological diversity of the world's languages. We evaluate our method on Parallel Bible Corpus+ (PBC+), a parallel corpus of 1593 languages. The key idea is to use Byte Pair Encodings (BPEs) as basic units for multilingual embeddings. Through zero-shot transfer from English sentiment, we learn a seed lexicon for each language in the domain of PBC+. Through domain adaptation, we then generalize the domain-specific lexicon to a general one. We show – across typologically diverse languages in PBC+ – good quality of seed and general-domain sentiment lexicons by intrinsic and extrinsic and by automatic and human evaluation. We make freely available our code, seed sentiment lexicons for all 1593 languages and induced general-domain sentiment lexicons for 200 languages.[1]

1 Introduction

Lexicons play an important role in sentiment analysis. Sentiment lexicons are available for high-resource languages like English (Pang et al., 2008; Baccianella et al., 2010; Mohammad and Turney, 2013), but not for many low-resource languages. Researchers are trying to fill this gap by inducing lexicons monolingually (Badaro et al., 2014; Eskander and Rambow, 2015; Rouces et al., 2018) as well as multilingually (Chen and Skiena, 2014), often by transfer from high-resource to low-resource languages.

The world's languages are heterogeneous – of particular relevance for us is heterogeneity with respect to morphology and with respect to marking token boundaries. This heterogeneity poses difficulties when designing a universal approach to lexicon induction that works for all languages – implementing a high quality tokenizer and morphological analyzer for each language is not feasible short-term. Given the small number of native speakers in low-resource languages (Goldhahn et al., 2016), crowdsourcing cannot easily be carried out either.

To overcome this heterogeneity and provide sentiment resources for low-resource languages, we present a new approach to sentiment lexicon induction that is universal – that is, it is applicable to the full range of typologically different languages – and apply it to 1593 languages. Our method first takes a parallel corpus as input and applies BPE (Gage, 1994) segmentation to it. We then create a multilingual BPE embedding space, from which a ZS (zero-shot) lexicon for each language $\mathcal{L}$ is extracted by zero-shot transfer from English sentiment to $\mathcal{L}$. We use PBC+, an expansion of the Parallel Bible Corpus (Mayer and Cysouw, 2014), as our parallel corpus. The ZS lexicons show high quality, but are specific to the domain of PBC+ (the Bible). We then adapt them to the general domain. For brevity, we also use *generic* to refer to *general-domain*.

Our method is universal and language-agnostic – it does not require language-dependent preprocessing. We carry out intrinsic and extrinsic, automatic and human evaluations on 95 languages. Intrinsic evaluation shows that our approach produces word ratings that strongly correlate with gold standard lexicons and human judgments. Extrinsic evaluation on Twitter sentiment classification demonstrates that our lexicons perform comparably or better than existing lexicons derived in multilingual settings.

We chose an **approach to sentiment analysis based on lexicons** in this paper because it is transparent and meets high standards of explainability. A classification decision can easily be traced

[1] cistern.cis.lmu.de

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3506–3517
Florence, Italy, July 28 - August 2, 2019.

back to the lexicon entries in the document that are responsible. Many more complex methods, e.g., many deep learning approaches, do not meet this standard. Transparency is of particular importance for low-resource languages because error analysis and verification are paramount when working with small and noisy resources that are typical of low-resource languages.

Our **contributions**: (i) We propose a new method for inducing sentiment lexicons for a broad range of typologically diverse languages. We use BPEs as basic units and show that they work well across languages. (ii) We carry out extensive evaluation to confirm correctness and high quality of the created lexicons. (iii) We make our code, the 1593 ZS seed sentiment lexicons and 200 generic sentiment lexicons freely available to the community. This is the up-to-now largest sentiment resource in terms of language coverage that has been published.

2 Related Work

Monolingual Lexicon Induction. Sentiment lexicons for many languages have been induced. Eskander and Rambow (2015), Wang and Ku (2016), and Rouces et al. (2018) create Arabic, Chinese, and Swedish sentiment lexicons, respectively. Monolingually induced sentiment lexicons for specific domains like Twitter and finance are also devised (Mohammad et al., 2013; Hamilton et al., 2016). These methods are specialized such that applying them to other languages is non-trivial. For example, Eskander and Rambow (2015) link AraMorph (Buckwalter, 2004) with SentiWordNet by additionally considering part-of-speech information, which may not be available in lexical resources in other languages. Inducing Chinese sentiment lexicons (Wang and Ku, 2016) needs properly tokenized corpora, which is not a hard requirement in Swedish. In contrast, we aim to design a method applicable to typologically diverse languages and we apply it to 1500+ languages.

Bi/Multi-Lingual Lexicon Induction. Gao et al. (2015) propose a graph based method for learning sentiment lexicons in target language by leveraging English sentiment lexicons. They rely on a high-quality word alignment, which is difficult to produce if languages are typologically diverse and the size of the parallel corpus is small. Chen and Skiena (2014) devise a knowledge graph

eng	The book of the history of Jesus Christ , son of David , son of Abraham :
fra	Le livre de l'histoire de Jésus Christ , fils de David , fils d'Abraham :
jpn	アブラハムの子，ダビデの子，イエス･キリストについての歴史の書 :

Table 1: PBC+ verse 40001001 in three languages

based method to build sentiment lexicons for 136 major languages. Several linguistic resources such as Google Translate and Wiktionary are used to link words across languages. In contrast, our approach uses *BPE embeddings* to extract alignment signals from the parallel corpus, an approach that is better applicable across diverse languages. We do not require resources like Wiktionary. We cover more languages than Chen and Skiena (2014) and more words (e.g., 300K for Amharic).

Language-Agnostic NLP. Language-agnostic NLP has demonstrated strong performance in areas such as *neural machine translation* (NMT) and universal *representation learning*. A particular difficulty is languages that do not mark token boundaries by whitespace such as Japanese. We refer to them as *non-segmented languages*. Sennrich et al. (2016) show the strength of BPE in translating rare words. Kudo (2018) introduces subword regularization that utilizes multiple subword sequences to improve the robustness of NMT models. Sennrich et al. (2016)'s `subword-nmt`[2] requires preprocessing (specifically, tokenization) for non-segmented languages, however, `sentencepiece`[3] (Kudo and Richardson, 2018) used by Kudo (2018) requires no preprocessing even for non-segmented languages. This research indicates the potential of language-agnostic *NMT*.

Effective representations of *words* (Schütze, 1993), e.g., word embeddings (Mikolov et al., 2013; Pennington et al., 2014), have been extended to be bilingual (Ruder, 2017; Artetxe et al., 2017) or multilingual (Dufter et al., 2018), with (Artetxe et al., 2018) and without (Conneau et al., 2017) supervision. Artetxe and Schwenk (2018) train a language-agnostic BiLSTM encoder creating universal *sentence* representations of 93 languages, and performing strongly in crosslingual tasks. Lample and Conneau (2019) show that pretraining the encoders with a crosslingual language model objective helps in achieving state-

[2] `github.com/rsennrich/subword-nmt`
[3] `github.com/google/sentencepiece`

of-the-art results in crosslingual classification and NMT. This research demonstrates the strength of language-agnostic methods for *representation learning* in NLP. Language-agnostic NLP models can generalize across languages without requiring language-dependent preprocessing. These advantages motivate us to design a universal approach for sentiment lexicon induction for 1500+ languages.

3 Method

Figure 1 shows the four steps of our method: (i) BPE segmentation. (ii) Multilingual embedding space creation. (iii) ZS lexicon induction. (iv) Domain adaptation to the general domain. We work with the parallel corpus PBC+. PBC+ extends the Parallel Bible Corpus by adding[4] 500 translations of the New Testament in 334 languages, resulting in a sentence-aligned parallel corpus containing New Testament verses in 2164 translations of 1593 languages. Many languages have several translations of the New Testament in PBC+. We use the term "edition" to refer to a single translation. Table 1 shows a verse in three languages. As shown, the Japanese (jpn) verse is not tokenized.

3.1 BPE Segmentation

Given the linguistic heterogeneity of the world's languages, it is crucial to first decide which type of linguistic unit to use to represent a language $\mathcal{L}$ in the multilingual space. The word, the linguistic unit typically generated from whitespace tokenization, is not ideal for universal approaches because non-segmented languages require carefully designed tokenizers. Character (or byte) n-gram is an alternative unit (Wieting et al., 2016; Gillick et al., 2016; Schütze, 2017; Dufter et al., 2018), but the optimum length n varies across languages, e.g., $n = 2$ may be suitable for Chinese (Foo and Li, 2004), but clearly not for English.

In our desire to design a universal approach, we use `sentencepiece` to segment PBC+ editions in all 1593 languages into sequences of *BPE segments*. We will show that this segmentation works across languages.

The widely used BPE segmentation algorithm `subword-nmt` only considers BPE segments within words (Sennrich et al., 2016) and some frequent BPEs are essentially valid *words*. `sentencepiece` adopts this setting for segmented languages like English (Kudo, 2018). But for non-segmented languages, `sentencepiece` does not require any language-dependent preprocessing – it learns a data-driven "tokenizer" on-the-fly from raw text. Hence, `sentencepiece` BPE segments can be larger linguistic units than say, English words, e.g., phrases. Examples for Japanese BPE segments in PBC+ are: "愛のうちに" (in love) and "何と言えばよいでしょうか" (what should I say).

We will use the term "BPE" to refer to all BPE segments produced by `sentencepiece`, including subwords, words and cross-token units like phrases. Figure 1 (a) shows some sample units. As shown, the English segments can be words or subwords (underlined). Dominant contexts of shown subwords – insp: inspiration, inspired; crim: crime, criminals; blasphe: blasphemy, blasphemed; hest: highest, richest.

3.2 Multilingual Space Creation

We next create the multilingual space hosting BPEs in 1593 languages of PBC+. We use the Sentence ID (S-ID) method (Levy et al. (2017), cf. also Le and Mikolov (2014)), a strong baseline in multilingual embedding learning.

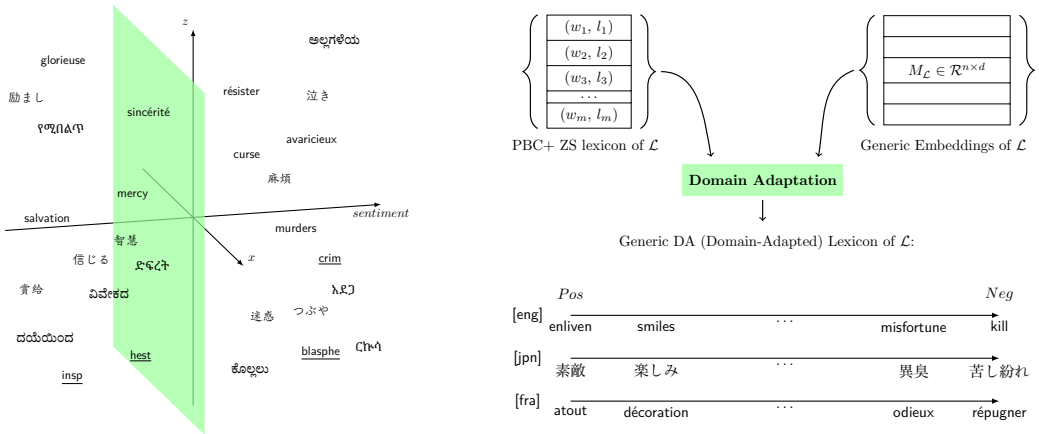
Given a sentence-aligned parallel corpus, the S-ID method first creates an embedding training corpus by recording co-occurrences between the sentence ID and the sentence's words (the New Testament verse ID and BPEs in our case) in all languages. Figure 2 shows examples from the training corpus; each BPE is associated with a 3-digit ISO 639-3 language code. After that, an embedding learner is applied to the created corpus to learn the multilingual space. We use *word2vec-skipgram* (Mikolov et al., 2013) as our embedding learner.

3.3 Zero-Shot Transfer of English Sentiment

Embeddings encode sentiment information (Pennington et al., 2014; Tang et al., 2014; Amir et al., 2015; Rothe et al., 2016). We exploit this for zero-shot transfer of English sentiment to the other 1592 languages. We train two linear SVMs to classify sentiment of English BPE embeddings as positive vs. non-positive (POS) and as negative vs. non-negative (NEG).

We use this setup – as opposed to binary classification positive vs. negative – to address the fact that some long BPE segments in non-segmented

[4] We use `github.com/ehsanasgari/1000Langs`



(a) **PBC+ ZS (zero-shot) lexicons**: Created by zero-shot crosslingual transfer

(b) **Generic DA (domain-adapted) lexicons**: Created by PBC-to-general-domain adaptation

Figure 1: Universal sentiment lexicon induction. (a): S-ID multilingual space of BPEs and sentiment classification hyperplanes (only the positive vs. non-positive plane is shown) learned from English. Underlined units are English BPEs with strong sentiment. (b): Creating generic DA lexicons using PBC+ ZS lexicons and generic embeddings.

languages may encode both sentiments. Using two SVMs allows us to identify then filter out segments with compositional sentiments during zero-shot transfer. This setup also enables direct comparison with Dufter et al. (2018) in Table 2.

The two SVMs are then applied to all embedding vectors in the multilingual space to yield a ZS lexicon for each of the 1593 languages.

3.4 PBC+ to General Domain Adaptation

Our ZS lexicons show high quality (see §5.2), but are specific to the PBC+ domain, i.e., the Bible. We adapt them to the general domain by obtaining generic embeddings and using ZS lexicon BPEs as labels to predict the sentiment of each generic embedding.

We assume that we have access to generic embeddings or, alternatively, that we can learn them from a generic corpus. We now describe how we predict the sentiment of generic embeddings. Given the PBC+ ZS lexicon $\mathcal{B}$ and the generic embedding matrix $M_\mathcal{L} \in \mathbb{R}^{n \times d}$ of language $\mathcal{L}$, we train a matrix $Q_\mathcal{L} \in \mathbb{R}^{d \times d}$ such that BPE pairs with same sentiment ($G_s \subset \mathcal{B} \times \mathcal{B}$) have small $l2$ distance while BPE pairs with different sentiment ($G_d \subset \mathcal{B} \times \mathcal{B}$) have large $l2$ distance, i.e., $\forall w, v \in \mathcal{B}, w \neq v$:

$$\arg\min_{Q_\mathcal{L}} \sum_{(w,v) \in G_d} -\alpha \|PQ_\mathcal{L}(e_w - e_v)\|_2 \quad + \sum_{(w,v) \in G_s} (1-\alpha)\|PQ_\mathcal{L}(e_w - e_v)\|_2 + \frac{\lambda}{2}\|PQ_\mathcal{L}\|_F^2$$

where $e_w, e_v \in \mathbb{R}^d$ are embeddings of BPEs w, v. d is embedding dimension. n is vocabulary size. $\alpha \in [0, 1]$ is the hyperparameter balancing the two sub-objectives. λ is a regularization weight. $P \in \mathbb{R}^{d \times d}$ is an identity matrix in the first dimension, i.e., a selector. This objective concentrates sentiment information in an embedding vector to a 1-dimensional ultradense sentiment space, resulting in a real-valued generic sentiment score. We minimize the objective using stochastic gradient descent (SGD).

After training, the generic sentiment score of BPE w in language $\mathcal{L}$ is computed as $s_w = PQ_\mathcal{L}e_w$. We refer to this method as *REG* and we call a lexicon computed by REG a *generic DA (domain-adapted) lexicon* since we always adapt from the Bible to the general domain in this paper.

REG is inspired by Densifier (Rothe et al., 2016), which is state of the art on SemEval2015 10E (Rosenthal et al., 2015) – determining

40001002	@Jesus:eng
40001002	@አብርሃም:amh
40001002	@ಏಸೊ:kan
40001002	@雅各:zho
66002003	བཟོད་བསྲན་བྱས:bod
...	...

Figure 2: Samples of S-ID embedding training corpus. 40001002 and 66002003: S-ID, i.e., IDs of New Testament verses. amh=Amharic, kan=Kannada, zho=Chinese, bod=Tibetan.

strength of association of Twitter terms with sentiment. Rothe et al. (2016) show that Densifier induces high quality and coverage sentiment lexicons in a *domain adaptation* setup. Densifier forces $Q_{\mathcal{L}}$ to be orthogonal to preserve the structure of the embedding space. As we are only interested in accurate sentiment prediction, we replace the orthogonality with $l2$ regularization: $\frac{\lambda}{2}\|PQ_{\mathcal{L}}\|_F^2$. The orthogonal constraint in Densifier – computing an SVD after each batch update – is expensive ($\mathcal{O}(d^3)$) and requires non-trivial training regime (Rothe et al., 2016). We will show that our formalization delivers comparable results.

In our experiments, we can use the generic *word embeddings* provided by Bojanowski et al. (2017) for 157 languages. Additionally, Heinzerling and Strube (2018) create generic *BPE embeddings* for 257 languages by segmenting Wikipedia articles using `sentencepiece` then running GloVe on the segmented corpora. As discussed above (§3.1), some BPEs in the PBC+ ZS lexicons are words, some are subwords – so we can utilize both sets.

4 Experiments

4.1 Datasets and Settings

We use the 7958 New Testament verses in PBC+ that were also used by Dufter et al. (2018) to create the multilingual BPE embedding space. To cover as many BPEs as we can, we segment each PBC+ edition three times with vocabulary sizes 2000, 4000 and 8000 using `sentencepiece`. S-ID generates a 31GB embedding training corpus including 7,414,810 BPEs in 1593 languages.

English training set. We employ VADER, a simple but widely used rule-based model for general sentiment analysis (Hutto and Gilbert, 2014), to create sentiment labels for English BPEs. We consider BPEs with sentiment score $\geqslant$ +0.1 (resp. $\leqslant$ -0.1) as positive (resp. negative). BPEs with score 0 are treated as neutral. As a result, we have 851 positive, 906 negative and 13,861 neutral training BPEs in English. We uniformly sample 878 = floor$((851 + 906)/2)$ neutral BPEs to speed up training.

Zero-shot transfer. The two SVMs for POS and NEG (§3.3) are trained on English training set (see above), then applied to all vectors in the multilingual BPE embedding space to create ZS lexicons for 1593 languages. We only keep high-confidence BPEs – those with a predicted probability for either POS or NEG of ≥ 0.7 (Platt et al., 1999) – to ensure ZS lexicons encode clear sentiment signals. The PBC+ ZS lexicon of language $\mathcal{L}$ is then the set of all high-confidence sentiment-bearing BPEs from $\mathcal{L}$.

Evaluation. Following Abdaoui et al. (2017), Bar-Haim et al. (2017), Rouces et al. (2018), we evaluate the quality of *PBC+ ZS lexicons* based on gold sentiment lexicons in Japanese (JA) (concatenation of Kobayashi et al. (2005); Higashiyama et al. (2008)), Czech (CZ) (Veselovská and Bojar, 2013), German (DE) (Waltinger, 2010), Spanish (ES) (Perez-Rosas et al., 2012), French (FR) (Abdaoui et al., 2017) and English (EN) (*WHM* lexicon, the concatenation of Wilson et al. (2005), Hu and Liu (2004) and Mohammad and Turney (2013), created by Rothe et al. (2016)). F1 is evaluation metric. We always compute F1 on the intersection of our and gold lexicon. Gold lexicons are also used in intrinsic evaluation of generic DA lexicons (Table 6). Additionally, the English WHM lexicon is also used in the evaluation of the universality of our approach (Table 8).

For *intrinsic evaluation* of *generic DA lexicons*, we compare our results with Densifier. Rothe et al. (2016) provide embeddings and train/validation splits of gold standard lexicons in CZ, DE, ES, FR and EN – we also use them in our experiments. We show (i) using GEN (the same training words as Densifier), REG (§3.4) induces generic lexicons in comparable quality; (ii) using PBC+ ZS lexicons, the induced generic DA lexicons are also in high quality. Kendall's τ (Kendall, 1938) is evaluation metric. As Densifier is implemented in *MATLAB*, we implement our model in *NumPy* (Oliphant, 2006) which is more accessible to the community.

For *extrinsic evaluation* of *generic DA lexicons*, we carry out Twitter sentiment classification in 13 languages. For each language, we retrieve $\approx$12,000 tweets from the human annotated dataset devised by Mozetič et al. (2016), and sample balanced number of positive and negative tweets (for clearer comparisons and descriptions) which are then randomly split 80/20 into train/test. We compare our lexicons with Chen and Skiena (2014)'s work. Two classification models are used (§5.3) – COUNT (count-based, Chen and Skiena (2014)) and ML (machine-learning-based, Eskander and Rambow (2015)). Accuracy is evaluation metric.

4.2 Hyperparameter Tuning

We train the multilingual BPE embedding space using *word2vec-skipgram* with default parameters except: 25 negative samples, 10^{-4} occurrence threshold, 200 dimensions and 10 iterations.

We tune the two linear SVMs for POS and NEG by 5-fold cross validation on English training set.

Following Rothe et al. (2016), when inducing generic DA lexicons, we run a grid search on their train/validation sets to find α and λ. With the same settings, we additionally conduct an experiment on Japanese (JA Wiki), a non-segmented language, to show the universality of our approach. For EN Twitter (SemEval2015 10E), we tune our model on the trial (dev) set and report results on the test set. In all experiments, we search $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, $\lambda \in \{0.01, 0.1, 1\}$. Learning rate is 0.1, batch size 100, and the maximum number of updating steps 30,000.

Following Eskander and Rambow (2015), in machine-learning-based Twitter sentiment classification for each of the 13 languages, we find the optimum SVM (positive vs. negative tweet) hyperparameters (C and kernel) by running 5-fold cross validation on the training set.

5 Results and Discussion

5.1 Multilingual BPE Space Evaluation

We first evaluate the multilingual BPE space by carrying out the crosslingual *verse sentiment* classification experiment in Dufter et al. (2018). Two linear SVMs are trained on 2147 English training verses to classify the verse sentiment (positive vs. non-positive, i.e., POS, and negative vs. non-negative, i.e., NEG). A verse is represented as the TF-IDF weighted sum of the embeddings of its BPEs. We then conduct the crosslingual verse sentiment analysis – using the SVMs to classify 476 test verses of Dufter et al. (2018)'s 1664 editions in 1259 languages. Table 2 gives results averaged over 1664 editions. *Word* and *Char* are two multilingual spaces created by Dufter et al. (2018). For *Word*, whitespace tokenization is used to segment all editions. For *Char*, all editions are segmented to sequences of overlapping byte-ngrams (length n varies across languages, see Dufter et al. (2018)). Next, the S-ID method is utilized to create the two multilingual spaces.

The S-ID BPE space outperforms both S-ID Word and S-ID Char spaces. This observation meets our expectation – the data-driven BPE segmentation is superior to splitting on whitespace (Word) or overlapping byte-ngram segmentation (Char), for non-segmented languages like Japanese whose PBC+ editions are not tokenized.

	Word		Char		BPE	
	POS	NEG	POS	NEG	POS	NEG
S-ID	.79	.88	.65	.86	**.81**	**.89**

Table 2: F1 for verse sentiment classification. Bold: our results. Word/Char are from Dufter et al. (2018).

ISO	B	W	Δ	ISO	B	W	Δ
lzh1	.82	.04	+.78	eng1	.88	.84	+.04
jpn1	.86	.19	+.67	fra1	.85	.85	-.00
khm2	.87	.21	+.66	deu1	.84	.83	+.01
khm3	.86	.25	+.61	spa1	.85	.85	+.00
ksw0	.86	.32	+.54	por1	.84	.87	-.03

Table 3: The most improved (left) editions when using S-ID BPE (*B*) compared with S-ID Word (*W*). *B* and *W* perform similarly on segmented languages (right) like English (eng), French (fra), German (deu), Spanish (spa) and Portuguese (por). Numbers are in F1.

For the more challenging subtask POS, we find the biggest improvement of S-ID BPE over Word is for non-segmented languages like Classical Chinese (lzh), Japanese (jpn), Khmer (khm) and S'gaw Karen (ksw) as shown in Table 3 (left). For segmented languages, S-ID BPE delivers similar performance as S-ID Word as shown in Table 3 (right). This observation also meets our expectation – lots of BPEs in segmented languages are essentially valid words.

These observations show the universality of our approach. The sentiment information derived from English is successfully transferred to heterogeneous languages without language-dependent preprocessing – even for non-segmented languages.

5.2 PBC+ ZS (Zero-Shot) Lexicon Evaluation

Sample entries in the English ZS lexicon are shown in Table 4 (left) as a **qualitative** evaluation. Table 5 shows the high consistency between the PBC+ ZS lexicons and gold lexicons in six languages. These results indicate that the

positive	negative	positive	negative
magnificent	fought	#blessedbeyondbelief	shats
privilege	blamed	alhamduillah	#worstpain
enjoyed	debauchery	#365daysofgratitude	theiving
salvation	adulter	#excellence	#stuffynose
rejoices	gloomy	co-create	sorethroat

Table 4: Sample entries in English ZS lexicon (left) and DA lexicon with Twitter embeddings (right).

two SVMs trained on English BPE embeddings perform strongly in a zero-shot crosslingual setting, and the resulting PBC+ ZS lexicons in difficult (morphologically rich, e.g., Czech; non-segmented, e.g., Japanese) languages encode clear sentiment information.

5.3 Generic DA (Domain-Adapted) Lexicon Evaluation

Table 4 (right) **qualitatively** shows the most sentiment-bearing words of the DA lexicon induced with English ZS lexicon and Twitter embeddings (EN Twitter). Lots of top ranked words are strong sentiment-bearing hashtags that never occur in the ZS lexicon domain, illustrating that our approach functions well in the *domain adaptation* setup. This observation is consistent with Densifier (Rothe et al., 2016).

Intrinsic evaluation: ranking correlation. We compute ranking correlation between our generic DA lexicons and gold standard lexicons. There are overlapping words between our PBC+ ZS lexicon BPEs and the validation/test sets used by Rothe et al. (2016) – we discard these training words for a clean comparison.

Columns (i) and (ii) of Table 6 show that REG (§3.4) delivers results comparable to Densifier (ORTH) when using the same set of generic training words (GEN) in lexicon induction. However, our method is more efficient – no need to compute the expensive SVD after every batch update.

Comparing columns (ii) and (iii), we see a marginal decrease of τ between .020 and .057 when GEN is replaced by PBC+ ZS lexicons. Note that PBC+ ZS lexicons have much fewer training BPEs than GEN (e.g., 343 vs. 4298 in JA Wiki) – this may contribute to the decrease. These comparable results also reflect the correctness of PBC+ ZS lexicons.

We also use $\alpha = 0.4$ and $\lambda = 0.01$, the optimal hyperparameter values found on the trial set of EN Twitter, to induce generic DA lexicons for the other languages. This is the common setting in real applications – other languages most likely do not have validation sets available. Results are shown in column (iv). Compared with tuned results (PBC+/T), performance slightly drops as the hyperparameters are not tuned (PBC+/NT) for languages other than EN Twitter.

	JA	CZ	DE	ES	FR	EN
F1	.883	.914	.903	.963	.916	.939
∩ size	120	141	788	63	407	1145
|PBC+|	728	1793	2827	1766	2193	2563

Table 5: High consistency between PBC+ ZS lexicons and generic gold lexicons in JA and five languages used in Rothe et al. (2016). ∩ size: intersection size. |PBC+|: ZS lexicon size.

	(i) ORTH	(ii)	(iii) REG	(iv)
	GEN	GEN	PBC+/T	PBC+/NT
CZ web	.580	.576	.529	.524
DE web	.654	.654	.634	.634
ES web	.563	.568	.524	.514
FR web	.544	.540	.514	.474
EN Tw.	.654	.629	.583	.583
EN Ne.	.622	.582	.562	.557
JA Wiki	n/a	.628	.571	.558

Table 6: Correlation (τ) of generic DA lexicons with gold standard lexicons. ORTH results are from Rothe et al. (2016). The other columns use REG (§3.4). Training words for lexicon induction are from Rothe et al. (2016) (GEN) and from PBC+ ZS lexicons.

Algorithm 1 Creating tweet representation

```
procedure RepTweet(String: Tweet, Dict: Lexicon)
    words = Tweet.split(" ")
    vec = [0.0, 0.0]
    for w ∈ words do
        val = Lexicon.get(w)
        if val > 0 then
            vec[0] = vec[0] + val
        else if val < 0 then
            vec[1] = vec[1] + val
        else
            continue
    return vec
```

Figure 3: Creating the representation of a tweet in Twitter sentiment classification using ML.

Overall, the performance differences between GEN (based on generic gold standard lexicons) and PBC+ (based on PBC+ ZS lexicons) are small and τ correlations are high. The high quality of generic DA lexicons in these six diverse (morphologically rich and non-segmented) languages shows the universality of our approach again – no language-dependent preprocessing is needed.

Extrinsic evaluation: Twitter sentiment classification. Based on the subset of frequent words only,[5] we use the top 10% most positive and most negative words for this evaluation. We compare with the closest work – lexicons from Chen and Skiena (2014).

Two classification models are used – word-count-based model COUNT (Chen and Skiena,

[5] In all discussions, we consider words that are top 50% frequent in the embedding vocabulary as "frequent" words.

		sqi	bul	hrv	deu	hun	pol	por	rus	srp	slk	slv	spa	swe	$\bar{x}$
COUNT	C&S	.55	.57	.57	.61	.61	.55	.57	.54	.51	.55	.64	.54	.57	.57
	Ours	.50	.60	.60	.56	.64	.62	.53	.65	.50	.61	.57	.55	.63	**.58**
ML	C&S	**.58**	.59	.60	.62	.64	.56	.54	.56	.51	.57	**.66**	.53	.59	.58
	Ours	.54	**.65**	**.65**	**.64**	**.66**	**.66**	.54	**.67**	.51	**.64**	.59	**.57**	**.64**	**.61**

Table 7: Accuracy of Twitter sentiment classification in Albanian (sqi), Bulgarian (bul), Croatian (hrv), German (deu), Hungarian (hun), Polish (pol), Portuguese (por), Russian (rus), Serbian (srp), Slovak (slk), Slovenian (slv), Spanish (spa) and Swedish (swe). Baseline of all experiments: 0.5.

2014), and machine-learning-based model ML (Eskander and Rambow, 2015). COUNT labels a tweet with the sentiment that has more word occurrences in the tweet (positive in case of ties). COUNT does not require training and the results are from all tweets for each language. In ML, the vector representation of a tweet is created according to Figure 3. Our generic DA lexicons support computing real-valued vectors in this way. Chen and Skiena (2014)'s lexicons are discrete (1/-1); we use these discrete values when applying ML to their lexicons. Finally, for each language, an SVM is trained on the 2-dimensional vectors.

Table 7 shows results. The baseline accuracy is 0.5 for all experiments as our dataset is balanced. Rows *Ours* and *C&S* show results using our and Chen and Skiena (2014)'s lexicons respectively. As shown, the two sets of lexicons give comparable results in COUNT. But ML generally performs better than COUNT, and our lexicons give better classification results – our real-valued representation of tweets is superior to the discrete one computed with Chen and Skiena (2014)'s lexicons.

Overall, intrinsic and extrinsic evaluations on diverse languages demonstrate the high quality of our generic DA lexicons.

5.4 Evaluation of Universality

We further conduct automatic and human evaluations on 95 diverse languages to show the universality of our approach. We focus on intrinsic evaluation – verifying the correctness of PBC+ ZS lexicons with F1, and assessing the quality of generic DA lexicons using τ. The extrinsic evaluation, i.e., Twitter sentiment classification, is not feasible here due to missing human annotated Twitter datasets in low-resource languages.

Automatic evaluation. Similar to Chen and Skiena (2014); Abdaoui et al. (2017), we use Google Translate (GT) for automatic evaluation – given a non-English language $\mathcal{L}$, we translate its PBC+ ZS lexicon and generic DA lexicon into English. Translated English lexicons are then evaluated against the gold English lexicon WHM.

GT supports 102 non-English languages. We omit ten languages that (i) are not covered by PBC+ (Corsican, Galician, Pashto, Yiddish); (ii) are covered in PBC+, but not in the alphabet used by GT (Malayalam); (iii) do not have public pretrained embeddings (Filipino, Hmong, Kyrgyz, Sesotho); or (iv) are very close to another language (we keep Croatian, but do not include Bosnian). We conduct separate experiments for Bokmål and Nynorsk, which are not distinguished by GT. Thus, we evaluate on 93 languages. When translating words to English, we discard entries where GT fails (i.e., output is identical to input). As GT requires the uploaded file to be small ($\leqslant$ 1MB), we do the evaluation on uniformly sampled 600 top 1% positive and negative words that are frequent. For ten languages (Chichewa, Hausa, Hawaiian, Igbo, Lao, Maori, Samoan, Shona, Xhosa, Zulu) that have very small embedding training corpora (<5MB Wikipedia pages and articles) and vocabulary sizes (e.g., 5000 for Hausa), we sample 200 words at 10%.

Table 8 shows results. We see that PBC+ ZS lexicons show high consistency with gold labels across all 93 languages (F1 columns), including morphologically rich languages like Czech and Turkish, and non-segmented languages like Japanese and Khmer. The generic DA lexicons show high correlation with gold labels (τ columns) – with two exceptions. First, some languages have low-quality embeddings due to small embedding training corpora (e.g., Hawaiian: 998 KB; Igbo: 1014 KB) or because the training corpora apparently have low quality – e.g., the Luxembourgish embedding vocabulary contains a large amount of French and German words, suggesting that it was trained on mixed text and that the genuine Luxembourgish part is small. Second, GT does not perform well for some of the languages, again this is the case for Luxembourgish and also for Frisian. To give an example from Lux-

Language	F1	τ	Language	F1	τ	Language	F1	τ	Language	F1	τ	Language	F1	τ
Afrikaans	.909	.508	Esperanto	.933	.361	Italian	.924	.591	Mongolian	.840	.222	Sundanese	.912	.409
Albanian	.916	.570	Estonian	.889	.606	Japanese	.901	.411	Myanmar	.916	.534	Shona	.885	.223
Amharic	.870	.418	Finnish	.932	.584	Javanese	.904	.398	Nepali	.862	.491	Swedish	.936	.621
Arabic	.905	.509	French	.919	.600	Kannada	.921	.447	Nynorsk	.853	.434	Sinhala	.880	.540
Armenian	.848	.524	Frisian	.885	.065	Kazakh	.893	.421	Punjabi	.927	.506	Tajik	.876	.436
Azerbaijani	.768	.401	Georgian	.908	.540	Khmer	.906	.474	Persian	.903	.390	Tamil	.911	.513
Basque	.898	.477	German	.898	.548	Korean	.897	.481	Polish	.923	.530	Telugu	.934	.297
Belarusian	.915	.597	Greek	.912	.570	Kurdish	.925	.258	Portuguese	.913	.574	Thai	.867	.357
Bengali	.910	.389	Gujarati	.896	.479	Latin	.927	.336	Romanian	.917	.644	Turkish	.897	.607
Bokmål	.927	.625	Haitian	.891	.238	Lao	.834	.222	Russian	.910	.596	Ukrainian	.909	.612
Bulgarian	.911	.511	Hausa	.905	.184	Latvian	.919	.538	Scots	.848	.385	Urdu	.825	.258
Catalan	.937	.453	Hawaiian	.951	.078	Lithuanian	.922	.491	Serbian	.957	.559	Uzbek	.900	.361
Cebuano	.917	.390	Hebrew	.833	.522	Luxemb'gish	.834	.031	Sindhi	.845	.169	Vietnamese	.840	.403
Chichewa	.872	.061	Hindi	.878	.447	Macedonian	.918	.425	Slovak	.942	.515	Welsh	.879	.560
Chinese	.889	.486	Hungarian	.910	.502	Malagasy	.923	.417	Samoan	.857	.116	Xhosa	.892	.057
Croatian	.926	.519	Igbo	.791	.088	Malay	.892	.494	Swahili	.842	.403	Yoruba	.873	.188
Czech	.915	.545	Icelandic	.947	.417	Maori	.836	.015	Slovenian	.957	.483	Zulu	.889	.226
Danish	.936	.359	Indonesian	.898	.498	Maltese	.938	.488	Somali	.954	.335			
Dutch	.906	.553	Irish	.902	.476	Marathi	.942	.479	Spanish	.943	.428			

Table 8: Intrinsic evaluation of our PBC+ ZS and generic DA lexicons in 93 languages. We see high consistency (F1) between PBC+ ZS lexicons and gold labels across languages. The generic DA lexicons are strongly correlated (τ) with gold labels in most languages.

	Hiligaynon		Tibetan	
	τ	size	τ	size
2-way	.474	103	.542	64
3-way	.357	188	.361	148

Table 9: Human evaluation of generic DA lexicons in Hiligaynon and Tibetan. 2-way: positive, negative. 3-way: positive, neutral, negative.

embourgish for both problems: "verglоust" and its first nearest neighbor "verglousten" are translated by GT as "glowed" and "forget about it". We recommend to use the higher quality PBC+ ZS lexicon for these languages.

Apart from above exceptions, both F1 and τ are reasonably high, evidencing that our universal approach is applicable to a broad range of typologically diverse languages.

We do **human evaluation** for Hiligaynon and Tibetan, languages not supported by GT.

There are no public pretrained embeddings for Hiligaynon. We train embeddings on a concatenation of texts from project *Palito* (Dita et al., 2009) and *Jehovah's Witnesses* e-books (www.jw.org). From the generic DA Hiligaynon and Tibetan lexicons, we uniformly sample 199 from the top 10% positive and negative frequent BPEs.

Two Tibetan scholars and three Hiligaynon speakers annotated these BPEs as positive, negative, neutral, unclear where the last category refers to cases where the intended word is not apparent from the BPE. We omit entries labeled as unclear and compute τ. Table 9 shows τ averaged over annotators. We see that our lexicons have consistent positive correlation with the human annotation in both languages.

6 Conclusion

We proposed a universal approach for sentiment lexicon induction. By creating a multilingual BPE embedding space for 1500+ languages, we successfully transfer sentiment to each language without language-dependent preprocessing. We created 1593 ZS (zero-shot) sentiment lexicons and showed for a subset that they are highly consistent with gold lexicons. To address the fact that the small-size ZS lexicons are specific to PBC+'s domain, we conduct domain adaptation and induce large-size generic DA (domain-adapted) lexicons for 200 languages. Extensive intrinsic and extrinsic, automatic and human evaluations on 95 languages confirm the correctness and good quality of our lexicons. We make our code and lexicons freely available to the community.

To induce generic lexicons, our approach requires generic embeddings, which are not always available for low-resource languages. Solving this problem is non-trivial as many low-resource languages have a limited amount of written text in electronic form (and in any form). In such cases, the PBC+ ZS lexicons can be utilized because they also have high quality.

Acknowledgements. We thank Philipp Dufter and the anonymous reviewers for comments and suggestions; and Mary Ann C. Tan, Samyo Rode and Nikolai Solmsdorf for sentiment judgments for Hiligaynon and Tibetan. This work was funded by the European Research Council (ERC #740516).

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Silvio Amir, Ramón Astudillo, Wang Ling, Bruno Martins, Mario J. Silva, and Isabel Trancoso. 2015. Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173.

Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. 2017. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark. Association for Computational Linguistics.

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

T Buckwalter. 2004. Buckwalter arabic morphological analyzer (bama) version 2.0. linguistic data consortium (ldc) catalogue number ldc2004l02. Technical report, ISBN1-58563-324-0.

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: Short Papers)*, pages 383–389. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Shirley N. Dita, Rachel Edita O. Roxas, and Paul Inventado. 2009. Building online corpora of philippine languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530. Association for Computational Linguistics.

Ramy Eskander and Owen Rambow. 2015. Slsa: A sentiment lexicon for standard arabic. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2550. Association for Computational Linguistics.

Schubert Foo and Hui Li. 2004. Chinese word segmentation and its effect on information retrieval. *Information processing & management*, 40(1):161–190.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Dehong Gao, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou. 2015. Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*, 41(1):21–40.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306. Association for Computational Linguistics.

Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff. 2016. Corpus collection for under-resourced languages with more than one million speakers. *CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity*, page 67.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Learning sentiment of nouns from selectional preferences of verbs and adjectives. In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2005. Collecting evaluative expressions for opinion extraction. In *Proceedings of the First International Joint Conference on Natural Language Processing*, IJCNLP'04, pages 596–605, Berlin, Heidelberg. Springer-Verlag.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 765–774. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327. Association for Computational Linguistics.

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

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Travis E Oliphant. 2006. *A guide to NumPy*, volume 1. Trelgol Publishing.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International*

Workshop on Semantic Evaluation (SemEval 2015), pages 451–463. Association for Computational Linguistics.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777. Association for Computational Linguistics.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018. SenSALDO: Creating a Sentiment Lexicon for Swedish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Hinrich Schütze. 1993. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan-Kaufmann.

Hinrich Schütze. 2017. Nonsymbolic text representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 785–796. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565. Association for Computational Linguistics.

Kateřina Veselovská and Ondřej Bojar. 2013. Czech SubLex 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ulli Waltinger. 2010. Germanpolarityclues: A lexical resource for german sentiment analysis. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Shih-Ming Wang and Lun-Wei Ku. 2016. Antusd: A large chinese sentiment dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.