On the Effect of Word Order
on Cross-lingual Sentiment Analysis

Sobre el efecto del orden de las palabras
en el análisis de sentimiento crosslingüe

Àlex R. Atrio[1], Toni Badia[2], Jeremy Barnes[3]

[1] HEIG-VD & EPFL
[2] Universitat Pompeu Fabra
[3] University of Oslo

[1]alejandro.ramirezatrio@heig-vd.ch, [2]toni.badia@upf.edu, [3]jeremycb@ifi.uio.no

Abstract: Current state-of-the-art models for sentiment analysis make use of word order either explicitly by pre-training on a language modeling objective or implicitly by using recurrent neural networks (Rnns) or convolutional networks (Cnns). This is a problem for cross-lingual models that use bilingual embeddings as features, as the difference in word order between source and target languages is not resolved. In this work, we explore reordering as a pre-processing step for sentence-level cross-lingual sentiment classification with two language combinations (English-Spanish, English-Catalan). We find that while reordering helps both models, Cnns are more sensitive to local reorderings, while global reordering benefits Rnns.
Keywords: sentiment analysis, cross-lingual, reordering

Resumen: Los modelos de análisis de sentimiento que actualmente representan el estado del arte utilizan el orden de las palabras, ya sea explícitamente al preentrenar con un objetivo de modelización del lenguaje, ya sea implícitamente al recurrir a redes neuronales recurrentes (RNR) o convolucionales (RNC). Esto es un problema para los acercamientos crosslingües que emplean vectores bilingües para entrenar, ya que la diferencia del orden de las palabras entre la lengua de origen y la de destino no se resuelve. En este trabajo, exploramos el reordenamiento de las palabras como etapa de procesamiento previa para la clasificación de sentimiento crosslingüe a nivel de frase, con dos combinaciones de idiomas (Inglés-Castellano, Inglés-Catalán). Descubrimos que aunque el reordenamiento ayuda a los dos modelos, los RNC son más sensibles al reordenamiento local, mientras un reordenamiento global beneficia a los RNR.
Palabras clave: análisis de sentimiento, crosslingüe, reordenamiento

1 Introduction

Cross-lingual Sentiment Analysis (CLSA) exploits resources, *e. g.* labeled data of a high-resource language, to train a sentiment classifier for low-resource languages. This approach is useful when a target language lacks plentiful labeled data, particularly for specific domains. Machine Translation (MT) is often used to bridge the gap between languages (Banea et al., 2008; Balahur and Turchi, 2014), but requires abundant parallel data, which may be difficult to find for some low-resource languages. Approaches that use bilingual distributional representations, in contrast, have proven competitive while requiring less parallel data (Chen et al., 2016; Barnes, Klinger

and Schulte im Walde, 2018).

Recently, sentiment classifiers pre-trained on a language modeling task have lead to state-of-the-art results (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2018). This improvement suggests that sentiment analysis benefits from learning word order and fine-grained relationships between tokens, which can be gleaned from unlabeled data. These approaches, however, have only been applied in a monolingual setting and it is not clear how the difference in word orders would affect them in a cross-lingual setup. In this work, we perform an analysis of the effect of word order on cross-lingual sentiment classifiers that use bilingual embeddings as features. We show that these models are sensitive to

word order and benefit from pre-reordering the target-language test data so that it resembles the source-language word order.

2 Related Work

Cross-lingual Sentiment Analysis: Cross-lingual approaches to sentiment analysis attempt to leverage available sentiment annotations in a high-resource language for target languages which lack annotated data. This is especially important when the cost of annotating a high-quality sentiment dataset, such as the Stanford Sentiment Treebank (Socher et al., 2013), can be prohibitive (215,154 phrases, each annotated by 3 annotators, at 10 cents an annotation would be 64,546€!). Therefore, it is preferable to make use of those datasets that already exist. Although most approaches to cross-lingual sentiment analysis rely on Machine Translation (Banea et al., 2008; Balahur and Turchi, 2014; Klinger and Cimiano, 2015), this requires large amounts of parallel data, making it less helpful for under-resourced languages.

Bilingual word embeddings have enabled cross-lingual transfer with small amounts of parallel data (Artetxe, Labaka, and Agirre, 2017; Lample et al., 2018b) or even none at all (Lample et al., 2018a; Artetxe, Labaka, and Agirre, 2018a), and are now used as features for state-of-the-art document-level (Chen et al., 2016), sentence-level (Barnes, Klinger, and Schulte im Walde, 2018), and targeted (Hangya et al., 2018) cross-lingual sentiment analysis approaches. The objective of bilingual embeddings is to learn a shared vector space in which translation pairs have similar vector representations. This benefits under-resourced languages as a sentiment classification model trained on the source-language can be applied directly to target-language data, without the need to translate it.

Word Order in Sentiment Analysis: Pre-training sentiment classifiers with a language-modeling task represents a successful transfer learning method. Peters et al. (2018) learn to create contextualized embeddings by training a character-level convolutional network to predict the next word in a sequence. Similarly, Howard and Ruder (2018) introduce techniques that improve the fine-tuning of the base language-model. Likewise, Devlin et al. (2018) introduce a self-attention network and adjust the language modeling

		EN	ES	CA
Binary	+	1258	1216	682
	−	473	256	467
4-class	++	379	370	256
	+	879	846	426
	−	399	218	409
	− −	74	38	58
	Total	1731	1472	1149

Table 1: Statistics for the OpeNER English (EN) and Spanish (ES) as well as the Multi-Booked Catalan (CA) sentence-level datasets (Agerri et al., 2013; Barnes, Badia, and Lambert, 2018)

task to a cloze task, where they predict missing words in a sentence, rather than the next word given a sequence. They then fine-tune their models on downstream tasks. These models that explicitly learn word order have led to state-of-the-art results on monolingual sentiment tasks.

Word Reordering: Rule-based pre-reordering has a long tradition in Machine Translation (see Bisazza and Federico (2016) for a survey), where word order directly affects the quality of the final result. Reordering rules can be determined manually (Collins, Koehn, and Kucerova, 2005; Gojun and Fraser, 2012) or with data-driven approaches that either learn POS-tag based (Crego and Mariño, 2006a; Crego and Mariño, 2006b) or tree-based (Neubig, Watanabe, and Mori, 2012; Nakagawa, 2015) reordering rules. The advantage of POS-tag based rules is that they are simple to implement and do not require full parsing of the target-language sentences.

3 Methodology

3.1 Corpora and Datasets

At document-level, bag-of-words models are often expressive enough to give good results without relying on word order (Meng et al., 2012; Iyyer et al., 2015). But because we are interested in word-order effects in cross-lingual sentiment analysis, we require datasets that are annotated at a fine-grained level, *i.e.* sentence- or aspect-level.

For this reason, we use the English and Spanish OpeNER corpora of hotel reviews (Agerri et al., 2013) as well as the Catalan MultiBooked Dataset (Barnes, Badia, and

ORIGINAL	Único punto **negativo** el **ruido** que las ventanas de madera tan típicas de la zona *no consiguen* **aislar**
REORDERED	Único **negativo** punto el **ruido** que las ventanas de tan típicas madera de la *no* zona *consiguen* **aislar**
NOUN-ADJ	Único **negativo** punto el **ruido** que las ventanas de madera tan típicas de la zona *no consiguen* **aislar**
RANDOM	aislar madera *consiguen* típicas de el de zona las ventanas punto **negativo** Único la *no* **ruido** tan que
ONLY-LEXICON	UNK **negativo** UNK UNK **ruido** UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK **aislar**
NO-LEXICON	Único punto UNK el UNK que las ventanas de madera tan típicas de la zona *no consiguen* UNK
Translation	The only **negative** point the **noise** that the typical wooden windows in the area *don't manage* to **block**

Table 2: An example of a negative Spanish sentence (ORIGINAL) with the five reordering transformations applied, as well as its English translation. The **bold tokens** are words found in the sentiment lexicon, and the *italic words* are words that convey sentiment in this instance, but are not in the lexicon

Lambert, 2018). Statistics on the corpora are shown in Table 1. Each sentence is annotated for four classes of sentiment (strong positive, positive, negative, and strong negative). We use the English subset for training our classifiers and the Spanish and Catalan for testing the effects of word order on the target languages. Although these datasets are relatively small, they are all annotated similarly and are in-domain, which avoids problems with mapping labels or domain shifts.

3.2 Bilingual Word Embeddings

VecMap (Artetxe, Labaka, and Agirre, 2016; Artetxe, Labaka, and Agirre, 2017) creates bilingual embeddings by learning an orthogonal projection between two pre-computed monolingual vector spaces and requires only a small bilingual dictionary. We use the publicly available GoogleNews vectors for the English (available at `https://code.google.com/archive/p/word2vec/`), and for Spanish and Catalan we create skip-gram embeddings with 300 dimensions trained on Wikipedia data. The bilingual dictionaries are translated sentiment lexicons (Hu and Liu, 2004) with 5700 pairs for English – Spanish (5271 for English – Catalan).

3.3 Experimental Setup

In order to test whether a sentiment classifier trained on bilingual embeddings is sensitive to word order, we test classifiers on six versions of the target-language sentiment data, which we describe in the following section. An example of these six versions is shown in Table 2.

Original: We test the model on the original data with no changes in word order.

Reordered: A competing hypothesis is that a full pre-reordering of the target-side sequences will be more familiar to the sentiment classifier trained on English and therefore lead to better results. We implement POS-tag based rewrite rules (Crego and Mariño, 2006a; Crego and Mariño, 2006b), which are then applied to the target-language test data before testing.

Noun-Adj: Given that adjectives are important for sentiment analysis, we hypothesize that adjusting the order of nouns and adjectives should be beneficial if the classifier is learning source-language word order. Therefore, we implement a simple reordering which places Spanish and Catalan adjectives before, rather than after, the noun they modify.

Random: We randomly permutate the order of the target-language sentences. If the sentiment classification models take the target language word order into consideration, this should lead to poor results.

Only-Lexicon and No-Lexicon: Finally, we provide two baselines for clarification. The ONLY-LEXICON experiment removes all words which do not appear in the Hu & Liu sentiment lexicon (Hu and Liu, 2004). If our systems take word order into account, they should be affected negatively by this, as the resulting sentence does not resemble the normal word order. If, however, the models are relying on keywords, this will have little effect.

For the NO-LEXICON experiment, we remove all of the words in a phrase which appear in the sentiment lexicon. If the models are attending to sentiment keywords, this approach should lead to the worst performance.

Baselines: We perform additional experiments with monolingual and Machine Translation (MT)-based cross-lingual approaches. For the former, we use the Google API (available at `https://translate.google.com/`)

and translate the target-language data to English.

For both baseline setups, we only test the Random reordering, Only-Lexicon, and No-Lexicon approaches. Additionally, the monolingual setup is not comparable to the MT and cross-lingual versions, as we must use the target-language data for training, development, and testing (70%/10%/20%).

3.4 Models

To test our hypotheses, we compare three different classifiers: a Support Vector Machine (SVM) with Bag-of-Embeddings feature representations, a Convolutional Neural Network (Cnn) (dos Santos and Gatti, 2014; Severyn and Moschitti, 2015), and a Bidirectional Long Short-Term Memory Network (BiLstm) (Luong, Pham, and Manning, 2015). Each of these classifiers theoretically has an increasing reliance on word order. Note that we do not use the bilingual sentiment model (Barnes, Klinger, and Schulte im Walde, 2018), as it jointly learns both projection and classifier and cannot be used as input to the Cnn and BiLstm. Although the SVM does not take into account word order at all, it is a strong baseline for sentiment analysis (Kiritchenko et al., 2014). The Cnn considers only local word order, while the BiLstm relies on both local and long-distance dependencies.

For the neural models, we train five classifiers on five random seeds and report the mean and standard deviation of the macro F_1 score, while we only report the macro F_1 score of a single run for the SVM.

BiLstm We implement a single-layered BiLstm classifier with a 100-dimensional hidden layer, which passes the concatenation of the two final hidden states to a softmax layer for classification. The cross-lingual model is initialized with the pre-trained bilingual embeddings (monolingual embeddings for the monolingual and translation models), use dropout of 0.3 for regularization, and are trained for 30 epochs with a batch size of 32 using Adam as an optimizer. We choose the parameter for training epochs on the source-language development set and test this model on the target-language data.

Cnn The Cnn has a single convolutional layer with filters $\in \{3, 4, 5\}$ followed by a max-pooling layer of length 2. The pooled representation of the sentence is passed to a feed-forward layer and finally a softmax layer

of size $\mathbb{R}^{|L|}$ where L is the set of labels. The optimization is the same as the BiLstm, with dropout applied after the feed-forward layer.

SVM Finally, we implement a baseline bag-of-embeddings SVM. For each sentence in the dataset, we create an averaged embedding representation $A = \frac{1}{n}(\sum_{i=1}^{n} e(t_i))$ where $e(t_i)$ is the embedding representation of the ith token in the sentence $S \in \{t_1, t_2, \ldots, t_n\}$ of length n. For the cross-lingual approaches we use the bilingual embeddings (monolingual embeddings for the monolingual and translation approaches) and tune the c parameter on the source-language development set.

4 Results

Table 3 shows the results of all experiments. Firstly, reordering the test data improves the results on all of the eight experiments (we do not consider SVM experiments to calculate improvements as they are invariant to word order). Specifically, the Reordered approach improves the BiLstm results the most on all experiments, while the simpler Noun-Adj flip is the best performing setup with Cnns. This indicates that local word reordering has more of an effect on Cnns, while the global reordering can be more helpful to BiLstms. While the improvements from reordering are often small (0.2 - 1.6 percentage points (ppt)), they are stable.

While it is the case that in both of the target languages Noun-Adj combinations can have a different meaning if the order is switched (for example "el amigo viejo" and "el viejo amigo"), the practical relevance of these order changes is minimal: in the Spanish dataset, of 978 occurrences of Noun-Adj, only 23 (2.35%) occur as well with a Adj-Noun order; in the Catalan dataset, of 745 occurrences of Noun-Adj, only 8 (1.07%) occur as well with a Adj-Noun order.

Random has a more substantial negative effect on monolingual models (an average decrease of 1.6 ppt for BiLstm and 3.0 ppt for Cnn) and MT-based models (1.6/4.3 ppt, respectively) than bilingual embedding models (0.8/1.1). This indicates that noise from the embedding projection renders it more difficult for models to use word order in the cross-lingual setup.

Additionally, Random has a larger effect on the Cnn (an average loss of 1.1 ppt) than on the BiLstm (0.8). This is likely because the Cnn relies on specific combinations of n-

			4-class			Binary		
			BiLstm	Cnn	SVM	BiLstm	Cnn	SVM
BWE	EN-ES	Original	33.3 (1.8)	35.4 (1.1)	34.9	64.9 (0.9)	60.0 (1.4)	66.6
		Reordered	**34.0 (1.6)**	35.6 (1.4)	34.9	**65.1 (1.3)**	60.1 (1.3)	66.6
		N-ADJ	**34.0 (1.8)**	**35.8 (1.2)**	34.9	65.0 (1.2)	**60.2 (1.4)**	66.6
		Random	33.2 (1.3)	35.3 (1.1)	34.9	63.9 (2.3)	58.8 (0.9)	66.6
		Only-Lexicon	28.2 (3.8)	26.9 (2.5)	30.7	57.6 (5.5)	34.2 (5.5)	53.0
		No-Lexicon	31.9 (1.6)	33.2 (1.4)	33.3	61.1	57.1 (2.8)	63.4
	EN-CA	Original	37.0 (1.4)	37.4 (1.5)	33.2	64.0 (1.1)	61.9 (6.8)	68.2
		Reordered	**37.8 (1.2)**	37.9 (1.5)	33.2	**65.6 (1.5)**	62.6 (5.8)	68.2
		N-ADJ	37.7 (1.5)	**38.1 (1.6)**	33.2	65.5 (1.5)	**62.8 (6.3)**	68.2
		Random	35.7 (1.0)	35.6 (1.5)	33.2	63.3 (0.8)	60.8 (5.5)	68.2
		Only-Lexicon	28.2 (1.8)	25.7 (3.2)	23.8	49.9 (4.3)	40.5 (6.7)	39.1
		No-Lexicon	35.9 (1.7)	34.3 (1.8)	31.2	61.7 (1.1)	58.1 (5.3)	63.1
MT	EN-ES	Original	**46.5 (1.2)**	**41.2 (3.7)**	44.6	**71.8 (1.1)**	**64.3 (1.6)**	70.7
		Random	46.0 (1.8)	38.9 (3.9)	44.6	71.0 (1.4)	62.2 (1.5)	70.7
		Only-Lexicon	32.9 (2.5)	28.2 (4.4)	36.2	63.0 (3.8)	44.6 (3.5)	51.9
		No-Lexicon	41.8 (0.7)	37.0 (3.0)	41.6	63.0 (1.1)	54.8 (2.6)	66.2
	EN-CA	Original	**51.5 (3.1)**	**44.1 (4.3)**	46.8	**79.9 (1.5)**	**72.8 (2.0)**	74.2
		Random	49.7 (1.4)	37.7 (3.6)	46.8	76.5 (2.2)	66.4 (1.4)	74.2
		Only-Lexicon	32.0 (2.7)	32.5 (4.0)	36.1	58.4 (7.8)	57.5 (2.9)	43.7
		No-Lexicon	48.4 (2.0)	40.9 (2.9)	46.2	75.6 (1.6)	65.6 (2.8)	70.4

Table 3: Macro F_1 results for all corpora and techniques. We denote the best performing bilingual embedding method per column with a **blue box**, the best MT method with a **pink box**. We do not denote bag-of-words SVM results, as they are invariant to word order

			4-class			Binary		
Monolingual	ES	Original	**43.2 (3.3)**	**36.2 (2.2)**	32.1	**68.5 (3.4)**	**64.8 (2.3)**	52.7
		Random	42.5 (2.6)	32.7 (1.8)	32.1	67.5 (4.2)	63.1 (2.7)	52.7
		Only-Lexicon	27.0 (0.5)	21.2 (4.5)	27.0	45.2 (0.0)	47.9 (3.9)	45.2
		No-Lexicon	37.9 (1.9)	34.3 (2.0)	30.3	64.7 (2.7)	65.0 (0.9)	51.8
	CA	Original	**48.6 (1.6)**	**46.2 (0.8)**	46.8	**77.1 (1.3)**	**76.4 (1.2)**	75.0
		Random	47.4 (1.9)	43.9 (3.0)	46.8	73.6 (1.3)	71.9 (1.9)	75.0
		Only-Lexicon	20.3 (2.8)	27.4 (3.2)	16.7	40.1 (1.5)	56.4 (5.2)	39.6
		No-Lexicon	47.5 (0.6)	45.8 (1.6)	45.8	75.0 (1.6)	74.5 (1.1)	74.8

Table 4: Macro F_1 results for all monolingual models. Although these results are not comparable to BWE or MT, as they are calculated on a smaller dataset, we provide them as a general reference to what results can be expected under monolingual conditions. We denote the best monolingual method with a **green box**

grams in order to correctly classify a sentence. If these are not present, the filters are not effective at classification.

Although they are not comparable (the test datasets have fewer examples), the monolingual models generally perform better than the cross-lingual versions, except for the SVM classifiers. The machine translation approaches perform better than the cross-lingual embedding methods.

The classification models display different trends across the setups. On the monolingual and machine translations setups, the BiLstm is the strongest model, followed by the Cnn and SVM (SVM and Cnn, respectively for machine translation). With bilingual embed-dings, however, the SVM outperforms both the BiLstm and Cnn on the Spanish binary setup, while the Cnn is strongest on the multiclass. This shows that BiLstm displays a different behavior with bilingual embeddings.

The machine translation models perform well and surprisingly suffer less than monolingual models (an average decrease of 15.4 ppt for MT BiLstm and Cnn models vs. 20.6 for monolingual) from using only features from the sentiment lexicon (Only-Lexicon). This suggests that MT models rely more on these keywords while ignoring word order effects to a higher degree.

Finally, the No-Lexicon and Only-Lexicon baselines perform poorly, with

lex R. Atrio, Toni Badia, Jeremy Barnes

model	text	prediction
ORIGINAL	relación calidad precio muy buena	negative
REORDERED	relación muy buena calidad precio	positive
translation	very good quality price relationship	positive
ORIGINAL	hotel perfecto	negative
REORDERED	perfecto hotel	positive
translation	perfect hotel	positive
ORIGINAL	el desayuno muy bueno .	negative
REORDERED	el muy bueno desayuno .	positive
translation	the breakfeast (was) very good	positive
ORIGINAL	gestión nefasta .	positive
REORDERED	nefasta gestión .	negative
translation	terrible management	negative

Table 5: Examples where reordering improves results over original on binary English-Spanish setup with the BiLSTM classifier

ONLY-LEXICON often more than 20 ppt below the performance of ORIGINAL. This is due largely to the low coverage of the sentiment lexicon used in this work, as many full sentences were completely unked (38% for Spanish, 43% for Catalan). This also explains the similar performances of ORIGINAL and NO-LEXICON.

5 Analysis

Reordering tends to help both the BiLSTM and CNN models with shorter examples (less than eight tokens long) where adjective order can easily be changed to resemble English word order, such as the examples shown in Table 5. In longer instances (more than ten tokens), however, the reordering either introduces too much noise or does not affect the final prediction. The current reordering models are therefore more adequate for sentiment tasks that deal with shorter texts, such as aspect- or sentence-level, rather than document-level sentiment analysis.

6 Conclusion and Future Work

In this work, we have shown that neural networks that rely on bilingual embeddings as features are sensitive to differences in source- and target-language word order and subsequently benefit from reordering the target language test data. The gains, however, are still relatively small, which suggests that currently bilingual embeddings introduce too much noise for a classifier to generalize well to the target language.

Although our reordering approach does improve the neural models, these more expressive models are still outperformed by the linear SVM with bag-of-embeddings representations. This is likely a side effect of the noise introduced by the bilingual embeddings. At test time, the model receives as input embeddings that are *similar* but not necessarily the same as at training. In the future, it may be helpful to develop models which are more robust to this noise, or alternatively to use low-resource machine translation techniques (Artetxe et al., 2018; Lample et al., 2018a; Artetxe, Labaka, and Agirre, 2018b; Lample et al., 2018c)

Given that language modeling pre-training is beneficial for state-of-the-art results in monolingual sentiment analysis, it is important to realize that cross-lingual models based on bilingual word embeddings do not currently benefit from word order learned in the source language. In the future, we would like to pre-train bilingual language models for cross-lingual sentiment analysis.

References

Agerri, R., M. Cuadros, S. Gaines, and G. Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51(Septiembre):215–218.

Artetxe, M., G. Labaka, and E. Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving mono-

28

lingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Artetxe, M., G. Labaka, and E. 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.

Artetxe, M., G. Labaka, and E. Agirre. 2018a. 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: Long Papers)*, pages 789–798. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April.

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

Banea, C., R. Mihalcea, J. Wiebe, and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135.

Barnes, J., T. Badia, and P. Lambert. 2018. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018.

Barnes, J., R. Klinger, and S. 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: Long Papers)*, pages 2483–2493.

Bisazza, A. and M. Federico. 2016. A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena. *Computational Linguistics*, 42:163–205.

Chen, X., B. Athiwaratkun, Y. Sun, K. Q. Weinberger, and C. Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614.

Collins, M., P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540.

Crego, J. M. and J. B. Mariño. 2006a. Improving statistical mt by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Crego, J. M. and J. B. Mariño. 2006b. Integration of pos tag-based source reordering into smt decoding by an extended search graph. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 29–36. Cambridge.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

dos Santos, C. N. and M. Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland, August.

Gojun, A. and A. Fraser. 2012. Determining the placement of german verbs in english–to–german smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 726–735.

Hangya, V., F. Braune, A. Fraser, and H. Schütze. 2018. Two methods for do-

main adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820.

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

Hu, M. and B. Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daume III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China.

Kiritchenko, S., X. Zhu, C. Cherry, and S. M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 437–442.

Klinger, R. and P. Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China, July.

Lample, G., A. Conneau, L. Denoyer, and M. Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November.

Luong, T., H. Pham, and C. D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Meng, X., F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea, July.

Nakagawa, T. 2015. Efficient top-down btg parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 208–218.

Neubig, G., T. Watanabe, and S. Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Severyn, A. and A. Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.