Improving Cross-lingual Text Classification with Zero-shot Instance-Weighting

Irene Li[1]*, Prithviraj Sen[2], Huaiyu Zhu[2], Yunyao Li[2], Dragomir Radev[1]
[1]Yale University, USA
{irene.li,dragomir.radev}@yale.edu
[2]IBM Research Almaden, USA
{senp,huaiyu,yunyaoli}@us.ibm.com

Abstract

Cross-lingual text classification (CLTC) is a challenging task made even harder still due to the lack of labeled data in low-resource languages. In this paper, we propose zero-shot instance-weighting, a general model-agnostic zero-shot learning framework for improving CLTC by leveraging source instance weighting. It adds a module on top of pre-trained language models for similarity computation of instance weights, thus aligning each source instance to the target language. During training, the framework utilizes gradient descent that is weighted by instance weights to update parameters. We evaluate this framework over seven target languages on three fundamental tasks and show its effectiveness and extensibility, by improving on F1 score up to 4% in single-source transfer and 8% in multi-source transfer. To the best of our knowledge, our method is the first to apply instance weighting in zero-shot CLTC. It is simple yet effective and easily extensible into multi-source transfer.

1 Introduction

Natural language processing (NLP) has largely benefited from recent advances in deep learning and large-scale labeled data. Unfortunately, such labeled corpora are not available for all languages. Cross-lingual transfer learning is one way to spread the success from high-resource to low-resource languages. Cross-lingual text classification (CLTC) (Prettenhofer and Stein, 2010; Ni et al., 2011) can learn a classifier in a low-resource *target* language by transferring from a resource-rich *source* language (Chen et al., 2018; Esuli et al., 2019).

Previous work has learned a classifier in the target language using a very small sample of labeled target instances or external corpora of unlabeled instances (Wang et al., 2019; Xu and Wan, 2017). In addition, other resources that may be utilized to achieve the same include, but are not limited to, parallel corpora of unlabeled instances in the target language (Xu and Wan, 2017). In this work, we address the most challenging setting, zero-shot CLTC (Arnold et al., 2007; Joachims, 2003), where no resource in the target language is given. Among the many methods for transfer learning that have been successfully employed in NLP (Mogadala and Rettinger, 2016; Zhou et al., 2016; Eriguchi et al., 2018), instance (re-) weighting is perhaps one of the oldest and most well known (Wang et al., 2017, 2019). It is best illustrated when we are given access to a few target labeled instances (few-shot learning). For example, both Dai et al. (2007) and Wang et al. (2019) learn a classifier iteratively by assigning weights to each instance in the source training data. While Dai et al. (2007) assigns weights to both source and target instances, Wang et al. (2019) pre-trains a classifier on the source training data and then re-weights the target labeled instances. Crucially, the weights are set to be a function of the error between the prediction made for the instance by the current classifier and the instance's gold label.

In a few-shot case, it is easy to see the appeal of re-weighting target language instances, since an instance that incurs a higher prediction loss can be given a larger weight, so as to improve the classifier. But in a zero-shot case, it seems impossible to compute instance weights based on prediction loss. In this work, we make it possible to assign such weights on instances in zero-shot CLTC. To the best of our knowledge, this is the first attempt to apply such a method to NLP tasks.

Our contributions are two-fold: First, we introduce **zero-shot instance-weighting**, a simple but effective, and extensible framework to enable instance weighted transfer learning for zero-shot CLTC. Second, we evaluate on three cross-lingual

[*] Work done as an intern at IBM Research Almaden.

Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 1–7
Bangkok, Thailand (Online), August 6, 2021. ©2021 Association for Computational Linguistics

classification tasks in seven different languages. Results show that it improves F1 score by up to 4% in single-source transfer and 8% in multi-source transfer, identifying a promising direction for utilizing knowledge from unlabeled data.

2 Proposed Method

We illustrate the zero-shot CLTC framework in Figure 1. The source and target language inputs are x_s and x_t respectively, during training, only the source label y_s is available and the task is to predict the target label y_t. We first apply the pre-trained model as an encoder to encode the inputs, the encoded representations are denoted by h_s and h_t. The figure illustrates four instances for each language in the mini-batch. Then there is an *Instance Weighting* module to assign weights to source language instances by considering the hidden representations h_s and h_t. Note that these layers are shared. We train the task layer and fine-tune the pre-trained language model layers.

2.1 Pre-trained Models

We compare two multilingual versions of pre-trained models for the pre-trained models: multilingual BERT (mBERT)[1] (Devlin et al., 2019) and XLM-Roberta (XLMR)[2] (Conneau et al., 2020).

We evaluate on multiple tasks in Section 3, so there are different ways to utilize the pre-trained models. For the sentiment and document classification task, we train a fully-connected layer on top of the output of the `[CLS]` token, which is considered to be the representation of the input sequence. For the opinion target extraction task, we formulate it as sequence labeling task (Agerri and Rigau, 2019; Jebbara and Cimiano, 2019). To extract such opinion target tokens is to classify each token into one of the following: **B**eginning, **I**nside and **O**utside of an aspect. We follow a typical IOB scheme for the task (Toh and Wang, 2014; San Vicente et al., 2015; Álvarez-López et al., 2016). In this case, each token should have a label, so we have a fully-connected layer that is shared for each token. We note that it may be possible to improve all the results even further by employing more powerful task layers and modules such as conditional random fields (Lafferty et al., 2001), but keep things relatively simple since our main goal is to evaluate instance weighting with zero-shot CLTC.

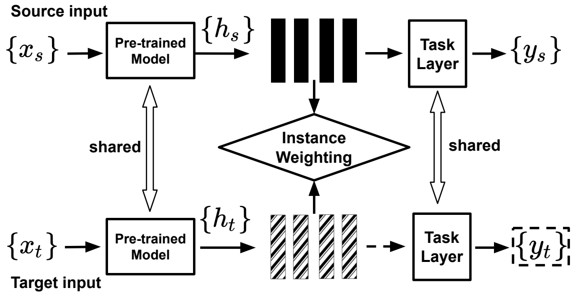

Figure 1: Framework Illustration: we illustrate 4 instances for each domain here.

2.2 Instance Weighting

The intuition behind instance weighting is the following: if the difference between a source instance and the target language is small, then it shares more common features with the target language, so it should make a larger contribution. For each instance in the source language, a large weight indicates a large contribution by the instance during training. Ideally, when deciding an instance weight, we should compare it with *all* instances from the target language. But doing so would incur prohibitively excessive computational resources. We thus approximate in small batches and calculate the weights by comparing how similar the instances are to the target ones within a small batch in each training step.

Instance Weighting-based Gradient Descent Vanilla mini-batch gradient descent is defined as:

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{k} \nabla_\theta f\left(y_i, g_\theta\left(x_i\right)\right) \qquad (1)$$

where α is the learning rate, θ is the parameter that we want to update, $g_\theta(x_i)$ is the model prediction for x_i , ∇_θ is the partial derivative, and $f(\cdot)$ is the loss function.

We modify Equation 1 to include instance weights:

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{k} w_i \cdot \nabla_\theta f\left(y_i, g_\theta\left(x_i\right)\right) \qquad (2)$$

where we assign a weight w_i to each instance within a mini-batch, and there is a weighted summation of the gradients in the mini-batch for all the instances and then update the parameter θ. It can be easily extended to multiple source languages, in this case, x_s may be training samples from more than one languages.

Unsupervised Weighting Metrics In each batch, to obtain weight w_i for each source instance i, we

[1] github.com/google-research/bert/blob/master/multilingual.md

[2] huggingface.co/XLMRoberta-base

follow a similarity-based approach. We define a scoring function to calculate a score between the current source instance representation h_i and the target instance representation h_j. Then we conduct a summation as the final score for source instance i to the *set* of target instances within this batch D_t. For $i \in D_s$:

$$w_i = score(i, D_t) = \sum_{j \in D_t} score(i, j).$$

We normalize each w_i in this batch to make sure the summation is 1, and they are plugged into Eq. 2.

Multiple ways exist to define a scoring function $score(i, j)$, and a Cosine-Similarity based scoring function is defined as:

$$score(i, j) = \frac{1}{2}(\frac{h_i \cdot h_j}{\|h_i\| \, \|h_j\|} + 1).$$

We also investigate two other ways for scoring function: Euclidean-Distance based and the CORAL Function (Sun et al., 2016). While Cosine scoring function performs the best, so we report it in our main experiments and ignoring the other two.

3 Evaluation

We test on three tasks: opinion target extraction, document classification, and sentiment classification [3]. English is the source language for all the experiments. We evaluate four settings: 1) direct adaptation with mBERT-base (mBERT), 2) mBERT with Instance Weighting (mBERT+IW), 3) direct adaption of XLMR-base (XLMR), and 4) XLMR with Instance Weighting (XLMR+IW).

Opinion Target Extraction We choose SemEval 2016 Workshop Task 5 (Pontiki et al., 2016) for opinion target extraction. It includes restaurant reviews in five languages[4]: English, Spanish (`es`), Dutch (`nl`), Russian (`ru`) and Turkish (`tr`). Given a sentence as input, one needs to classify each token into one of the three classes according to the IOB scheme. The training and testing size varies from 144 to 3,655. We compare against a list of models. Pontiki et al. (2014) and Kumar et al. (2016) are supervised and require extra corpora or resources to train. Agerri and Rigau exploits additional resources like unlabeled corpora. Jebbara and Cimiano (2019) applies multi-source (including the target) languages to train a classifier using cross-lingual embeddings and evaluates in a zero-shot manner. We summarize the results in Table 1.

Method	es	nl	ru	tr
Pontiki et al. (2014)★	0.520	0.506	0.493	0.419
Kumar et al. (2016)★	0.697	0.644	-	-
Jebbara and Cimiano (2019)	0.687	0.624	0.567	0.490
Agerri and Rigau (2019)★	0.699	0.664	0.655	0.602
mBERT	0.697	0.677	0.652	0.598
mBERT+IW	0.692	0.691	0.671	0.620
XLMR	0.690	0.700	0.664	0.674
XLMR+IW	**0.704**	**0.714**	**0.706**	**0.682**

Table 1: F1 scores on SemEval for Opinion Target Extraction. ★ indicates a supervised or semi-supervised learning method.

Cross-lingual Document Classification We conduct cross-lingual document classification task on the MLDoc dataset (Schwenk and Li, 2018). It is a set of news articles with balanced class priors in eight languages; Each language has 1,000 training documents and 4,000 test documents, and splits into four classes. We select a strong baseline (Schwenk and Li, 2018), which applies pre-trained MultiCCA word embeddings (Ammar et al., 2016) and then trained in a supervised way. Another baseline is a zero-shot method proposed by Artetxe and Schwenk (2019), which applies a single BiLSTM encoder with a shared vocabulary among all languages, and a decoder trained with parallel corpora. Artetxe and Schwenk (2019) apply mBERT as a zero-shot language transfer. Table 2 shows the results of our comparison study.

Sentiment Classification Finally, we evaluate sentiment classification task on Amazon multilingual reviews dataset (Prettenhofer and Stein, 2010). It contains positive and negative reviews from 3 domains, including `DVD`, `Music` and `Books`, in four languages: English (`en`), French (`fr`), German (`de`), and Japanese (`ja`). For each domain, there are 1,000 positive samples and 1,000 negative samples in each language for both training and testing. We choose the following baselines: translation baseline, UMM (Xu and Wan, 2017), CLDFA (Xu and Yang, 2017) and MAN-MoE (Chen et al., 2019). For the translation baseline, we translate the training and testing data for each target language into English using Watson Language Translator[5], and trained on the mBERT model, which is more

[3] We release our code in `https://github.com/IreneZihuiLi/ZSIW/`.

[4] The download script was broken and failed to obtain French data, so we do not report results for French.

[5] `https://www.ibm.com/watson/services/language-translator/`, version `2018-05-01`

Method	en	de	es	fr	it	ja	ru	zh
Schwenk and Li (2018) ★	0.9220	0.8120	0.7250	0.7238	0.6938	0.6763	0.6080	0.7473
Wu and Dredze (2019)	0.9420	0.8020	0.7260	0.7260	0.6890	0.5650	0.7370	0.7690
Artetxe and Schwenk (2019)	0.8993	0.8478	0.7733	0.7795	0.6943	0.6030	0.6778	0.7193
mBERT	0.8981	0.8680	0.7519	0.7492	0.6952	0.7222	0.6797	0.7937
mBERT+IW	-	0.8766	0.7532	0.7527	0.7122	0.7264	0.6949	0.8277
XLMR	**0.9295**	0.9245	0.8462	0.8710	0.7322	0.7824	0.6892	0.8580
XLMR+IW	-	**0.9265**	**0.8612**	**0.8797**	**0.7464**	**0.7942**	**0.7024**	**0.8712**

Table 2: F1 scores on MLDoc for Cross-lingual Document Classification. ★ indicates a supervised or semi-supervised learning method.

Method	Books	DVD	Music
Translation Baseline	0.7993	0.7789	0.7877
UMM★ (Xu and Wan, 2017)	0.7772	0.7803	0.7870
CLDFA★ (Xu and Yang, 2017)	0.8156	0.8207	0.7960
MAN-MoE (Chen et al., 2019)	0.7543	0.7738	0.7688
mBERT	0.7497	0.7378	0.7575
mBERT+IW	0.7573	0.7565	0.7553
XLMR	0.8248	0.8268	**0.8425**
XLMR+IW	**0.8452**	**0.8362**	0.8400

Table 3: F1 scores on Amazon Review for Sentiment Classification group by domains: Each cell shows the average accuracy of the three languages.★ indicates a supervised or semi-supervised learning method.

Method	es	nl	ru	tr
XLMR	0.690	0.700	0.664	0.674
Single-source	0.704	0.714	0.706	0.682
Multi-source	**0.735**	**0.738**	**0.745**	**0.688**

Table 4: Multi-source F1 scores on SemEval for Opinion Target Extraction: transfer from single-source and multi-source using XLMR+IW model.

confident in English[6]. Both UMM and CLDFA utilized more resources or tools like unlabeled corpora or machine translation. MAN-MoE is the only zero-shot baseline method. It applies MUSE (Lample et al., 2018) and VecMap (Artetxe et al., 2017) embeddings. We summarize the results in Table 3 for each domain.

Results Among the three tasks, both base models achieve competitive results for all languages thanks to the choice of pre-trained models. Instance weighting produces consistent improvements over the base models for nearly all target languages. Especially, in Table 1, the best model XLMR+IW beats the best baseline by 4.65% on average, improving from XLMR by 4% on Russian and gaining substantially on the other target languages; in Table 2, XLMR+IW outperforms the baselines, and surpassing XLMR steadily, with impressive gains on Russian, Chinese and Spanish. In Table 3, the best model shows the same trend in most cases. While our approach is model-agnostic, when the base model or the embedding improves, instance weighting will still help, as we can see the improved results obtained by switching from mBERT to XLMR. Again, the framework is simple but effective given these observations. Most importantly, it requires no additional external data and is easily adaptable into any deep models.

4 Discussion

Multi-source Expansion Studies show that multilingual transfer outperforms bilingual transfer (Guo et al., 2018). We run an experiment on the opinion extraction task to illustrate how our approach can be easily extended to enable multi-source transfer, (see Table 5). Here, we take the SemEval dataset, and for each target language, we train on the union of all other available languages. We can observe that by easily expanding into multi-source language training, we get a significant boost across the board in all target languages. Specifically, there is a 8.1% improvement on Russian. With easy adaptation, we show the extensibility and that multilingual transfer in zero-shot learning is a promising direction.

Case Study Intuitively, we should focus on the source instances with a smaller difference with target language, because they contain more common features with the target language. Thus, if we let those instances contribute more, it is possible that the model may perform better on the target language. As an example, Table 5 shows a positively-labeled French review containing adjectives with positive emotions (e.g., “exceptionnel”, “superbe”) and the instance weights for two English reviews, where the weights are generated using our best model XLMR+IW. Since English instance

[6] https://github.com/google-research/bert/blob/master/\multilingual.md explains the pre-training.

Language	Score	Content	Label
English Instance 2	0.5056	...I liked the book. Kaplan has consistently been one of my favorite authors (Atlantic Monthly) His theme is consistent: many nation states are not really nation states... Kaplan had great hope for the future of Iran as they struggle with theocracy...	Pos
English Instance 1	0.3647	One start , for some very acurate dramatic and terrorific facts about the Ebola, but very weak regarding origin of the virus, very unconvincing about possible ”theories”. sound more like that old music of desinformation, he almost blame another monkey for the Ebola...	Neg
French		**Origin**: ...ce livre est exceptionnel..La construction du livre est superbe, l'écriture magique... **Translation**: ...this book is outstanding..The construction of book is superb, magical writing ...	Pos

Table 5: A positive scenario: score comparison within the same batch.

1 contains adjectives with positive emotions (e.g. “favorite”, “great”), it has a higher score than English instance 2 containing adjectives with negative emotions (e.g., “weak”, “unconvincing”).

5 Conclusion

We proposed instance weighting for CLTC and evaluated on 3 fundamental tasks. The benefits of our approach include simplicity and effectiveness by ensuring wide applicability across NLP tasks, extensibility by involving multiple source languages and effectiveness by outperforming a variety of baselines significantly. In the future, we plan to evaluate on more tasks such as natural language inference (Conneau et al., 2018) and abstract meaning representation (Blloshmi et al., 2020).

References

Rodrigo Agerri and German Rigau. 2019. Language independent sequence labelling for opinion target extraction. *Artificial Intelligence*, 268:85–95.

Tamara Álvarez-López, Jonathan Juncal-Martínez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. 2016. GTI at SemEval-2016 task 5: SVM and CRF for aspect detection and unsupervised aspect-based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 306–311, San Diego, California. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925v2*.

Andrew Arnold, Ramesh Nallapati, and William W Cohen. 2007. A comparative study of methods for transductive transfer learning. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, pages 77–82.

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, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 193–200. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686v1*.

Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2019. Funnelling: A new ensemble method for heterogeneous transfer learning and its application to cross-lingual text classification. *ACM Transactions on Information Systems (TOIS)*, 37(3):37.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495, Minneapolis, Minnesota. Association for Computational Linguistics.

Thorsten Joachims. 2003. Transductive learning via spectral graph partitioning. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 290–297. AAAI Press.

Ayush Kumar, Sarah Kohail, Amit Kumar, Asif Ekbal, and Chris Biemann. 2016. IIT-TUDA at SemEval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1129–1135, San Diego, California. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702, San Diego, California. Association for Computational Linguistics.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 375–384. ACM.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. 2015. EliXa: A modular and flexible ABSA platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 748–752, Denver, Colorado. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2058–2065. AAAI Press.

Zhiqiang Toh and Wenting Wang. 2014. DLIREC: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240, Dublin, Ireland. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Zhi Wang, Wei Bi, Yan Wang, and Xiaojiang Liu. 2019. Better fine-tuning via instance weighting for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7241–7248.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Copenhagen, Denmark. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Vancouver, Canada. Association for Computational Linguistics.

Guangyou Zhou, Zhao Zeng, Jimmy Xiangji Huang, and Tingting He. 2016. Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 245–254.