JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages

Željko Agić
Department of Computer Science
IT University of Copenhagen, Denmark
zeag@itu.dk

Ivan Vulić
PolyAI Ltd.
London, United Kingdom
ivan@poly-ai.com

Abstract

Viable cross-lingual transfer critically depends on the availability of parallel texts. Shortage of such resources imposes a development and evaluation bottleneck in multilingual processing. We introduce JW300, a parallel corpus of over 300 languages with around 100 thousand parallel sentences per language pair on average. In this paper, we present the resource and showcase its utility in experiments with cross-lingual word embedding induction and multi-source part-of-speech projection.

1 Introduction

In natural language processing (NLP) the rule of thumb is that if we possess some parallel data for a low-resource target language, then we can yield feasible basic tools such as part-of-speech taggers for that language. Without such distant supervision, this task and many others remain unattainable, leaving the majority of languages in the world without basic language technology. Parallel data features a prominent role in building multilingual word representations (Ruder et al., 2017), annotation projection for parts-of-speech and syntactic dependencies (Das and Petrov, 2011; Tiedemann, 2014) and naturally machine translation.

The shortage of parallel data in turn creates a bottleneck in cross-lingual processing: without parallel sentences, we cannot yield usable models, nor can we robustly evaluate them, if even just approximately (cf. Agić et al. 2017). This absence has over the recent years materialized the *proxy fallacy*, whereby intended low-resource methods are tested by proxy, exclusively on resource-rich languages, because of the absence of test data or the lack of effort to produce it for approximate evaluation.

We seek to alleviate these issues by a significant new addition to the limited pool of parallel texts for low-resource languages.

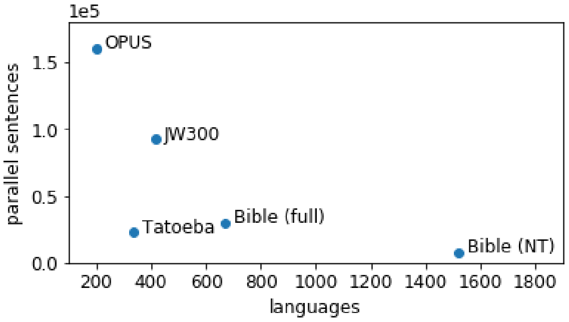

Figure 1: Our dataset JW300 in comparison to other massive parallel text collections with respect to multilingual breadth and volume of parallel sentences. The y-axis depicts the mean number of parallel sentences per language pair.

Contributions. A massive collection of parallel texts for over 300 diverse languages is our main contribution to facilitate multilingual NLP. The dataset is freely available for all non-commercial use.[1] We also show how simple techniques over our data yield competitive results in building cross-lingual word embeddings and annotation projection for part-of-speech tagger induction.

2 Dataset

JW300 spans across 343 languages, and comprises a total of 1,335,376 articles, with a bit over 109 million sentences, and 1.48 billion tokens.

Sources and structure. The data is a complete crawl of all the publications from the website `jw.org`. A vast majority of texts come from the magazines *Awake!* and *Watchtower*. While the texts do stem from a religious society, they cover an immense range of topics. The multilingual articles are mainly translations from a source in English. The dataset is organized by language and by article. Articles carry unique identifiers which

[1] http://zeljkoagic.github.io/jw300/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

span across the languages: all translations of the same article carry the same identifier number. This way we denote "parallel articles" as the base of all further processing.

Curation. All articles are converted from their HTML source into plain text format, one sentence per line, and tokenized. We also preserve the original formatting. We apply Polyglot (Al-Rfou, 2015) for sentence splitting and tokenization. For languages uncovered by Polyglot, we use its built-in language identifier to select the closest fit. Roughly 40% of all articles were split using a "neighbor language" tokenizer. Such broad strokes are necessary when dealing with massively multilingual datasets with low-resource languages where not even the basic processing is available, cf. Agić et al. (2016) who used only whitespace tokenization.

For all language pairs, and for all article pairs carrying the same identifier number, we perform sentence alignment using the aligner Yasa (Lamraoui and Langlais, 2013) with default settings. This way we align more than 50 thousand language pairs with over 90 thousand parallel sentences per language pair on average (see Table 1).

The basic statistics of JW300 in Table 1 reveal a small number of outliers with up to 2.5 million sentences like English, French, and Italian which are all rich in resources. However, the long tail of low-resource languages typically still offers between 50-100 thousand sentences.

Comparison. With its balance between multilingual breadth and monolingual depth, JW300 fills an important gap in cross-lingual resources: it comprises a multitude of low-resource languages while still offering ample sentences for each individual language, and parallel sentences for language pairs. To illustrate, for JW300 the breadth × depth ratio is 1.2x larger than for OPUS (Tiedemann, 2012), 2x larger than for the full Bible, and even 3x that of New Testament (see Figure 1).

JW300 still does come with its own caveats. The crucial one is surely bias: For example, could we indiscriminately use JW300 to train complex machine learning systems that further propagate the attitude of `jw.org` towards gender differences? From another viewpoint, however, should we rather train part-of-speech taggers through multi-source annotation projection from *Watchtower* articles on one side, or OPUS Ubuntu menu localizations or Bible psalms on the other side?

languages covered	343
language datasets	417
aligned pairs of languages	54,376

	μ	σ
articles	3,202.34	± 5,946.68
sentences	261,573.37	± 464,343.05
tokens	3,544,039.82	± 7,472,321.78
alignments	92,111.61	± 176,563.25

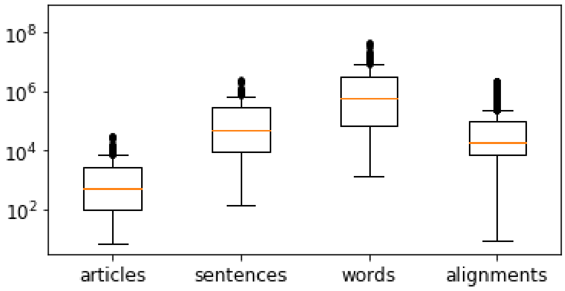

Table 1: Basic statistics for the JW300 corpus: counts of articles, sentences, words, and alignments, as well as an illustration of their distributions. Counts are reported for languages with at least one non-empty alignment to another language. Some languages have multiple datasets, e.g. different scripts, sign language.

Moreover, the ideological bias of JW300 is fairly well-defined. In that sense, while bias may invalidate the use of our corpus in some application areas, we argue that a wide-coverage collection of parallel data with *known* bias may in fact be valuable for research *on bias* in NLP (Bolukbasi et al., 2016; Caliskan et al., 2017; Dev and Phillips, 2019; Gonen and Goldberg, 2019), especially in multilingual settings (Lauscher and Glavaš, 2019).[2]

JW300 excels in low-resource language coverage. For example, OPUS offers over 100 million English-German parallel sentences, and JW300 only 2.1 million. However, in another example, for Afrikaans-Croatian the counts are 300 thousand in OPUS and 990 thousand in JW300, and moreover, the OPUS data for this language pair contains only Linux localizations.

Availability. Our dataset is freely available for all non-commercial use. The exact terms of use are provided by the copyright holder; see `https://www.jw.org/en/terms-of-use/`. For all practical purposes their custom terms of use are very closely aligned with the more well-known CC-

[2] We acknowledge the anonymous area chair who contributed this valuable argument as part of their meta-review.

	EN	ET	HR	MR	MT
EN	–	0.280	0.254	0.0	0.001
ET	**0.314**	–	0.302	0.001	0.0
HR	**0.269**	**0.334**	–	0.002	0.0
MR	**0.094**	**0.144**	**0.112**	–	0.001
MT	**0.131**	**0.206**	**0.164**	**0.141**	–

Table 2: BLI results (MRR scores) on a small subset of JW300 language pairs. The scores with the best-performing unsupervised cross-lingual word embedding model (Artetxe et al., 2018) are in gray cells over the main diagonal; the scores with a simple supervised method (Smith et al., 2017) are below the main diagonal. Better performance for each pair in bold.

BY-NC-SA license.[3]

3 Experiments

3.1 Cross-lingual word embedding induction

A recent trend in cross-lingual word embedding induction are fully unsupervised projection-based methods that learn on the basis of monolingual data only (Conneau et al., 2018; Alvarez-Melis and Jaakkola, 2018; Chen and Cardie, 2018, *inter alia*). The main idea is to construct a seed bilingual dictionary in an unsupervised fashion relying on adversarial training (Conneau et al., 2018), monolingual similarity distributions (Artetxe et al., 2018) or PCA projection similarities (Hoshen and Wolf, 2018), and then learn (gradually refined) projections of two monolingual embedding spaces into a shared cross-lingual space (by also iteratively refining the seed dictionary).

Such models hold promise to support cross-lingual representation learning for resource-poor language pairs. However, besides their problems with training divergence (Søgaard et al., 2018), a recent empirical study (Glavaš et al., 2019) has demonstrated that even most robust projection-based unsupervised models cannot match the performance of projection-based methods which require only 1K-5K seed translation pairs. The large-scale JW300 corpus offers such supervision (i.e., seed translation pairs) for a large number of language pairs. In other words, instead of resorting to fully unsupervised models for the language pairs included in JW300, we can use seed bilingual dictionaries from the parallel data to learn the projections. Based on the findings from Glavaš et al. (2019), we compare the most effective and the most robust unsupervised method of Artetxe et al. (2018) to a simple supervised method (Smith et al., 2017) in the bilingual lexicon induction task (BLI).[4]

For the demonstration purposes, we work with all pairs from the following language set: English (EN), Estonian (ET), Croatian (HR), Marathi (MR), and Maltese (MT). Our seed bilingual dictionaries are extracted from the JW300 corpora by taking the most probable target translation for each source word from IBM1-based word translation tables. Following prior work, we use the 5K most frequent translation pairs from training, while the next 2K pairs are used for testing. We use 300-dim monolingual fastText embeddings pretrained on Wikipedia for all languages (Bojanowski et al., 2017),[5] but the same trends are observed with other monolingual embeddings. The results in terms of Mean Reciprocal Rank (MRR) are summarized in Table 2. The BLI results are straightforward to interpret: for all experimental runs a simple supervised model with its supervision extracted from the JW300 corpus outperforms its unsupervised competition, further confirming the findings of Glavaš et al. (2019). The unsupervised model is even unable to converge for most language pairs, yielding extremely low MRR scores. The scores on another test set (Conneau et al., 2018) for EN-ET and EN-HR also favour the supervised model: 0.342 vs. 0.313 on EN-ET, and 0.289 vs. 0.261 on EN-HR. In sum, these preliminary experiments indicate the potential of JW300 in guiding cross-lingual representation learning.

3.2 Part-of-speech projection

Massively parallel data has proven most useful in inducing basic NLP models such as part-of-speech taggers. The formative work by Yarowsky et al. (2001) has inspired many influential works in projecting sequential labels from multiple source languages (Das and Petrov, 2011; Täckström et al., 2013), as well as projecting more complex annotations such as syntactic and semantic dependencies (Hwa et al., 2005; Padó and Lapata, 2009; Agić et al., 2016). Here we implement an experiment with projecting parts of speech from multiple sources to multiple targets following the line of work by Agić et al. (2015) and subsequently Plank et al. (2018), to showcase our corpus.

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/

[4] We expect even better performance with recently developed more sophisticated supervised methods such as RCSLS proposed by Joulin et al. (2018), see Glavaš et al. (2019).

[5] https://fasttext.cc/docs/en/english-vectors.html

Setup. We work with a large collection of multilingual sentences, where each sentence is a graph $G = (V, A)$. Its vertices V are sentence words for all involved languages, while its edges A are alignments between these words. One sentence t is declared as target sentence and indexed as $i = 0$, while the remaining n sentences are sources: Target words are then vertices $v_t \in V_0$, while the vertices $v_s \in V_i, 1 \leq i \leq n$ are the source words. The word alignments $a(v_s, v_t) \in A$ are also word aligner confidences: $a(v_s, v_t) \in (0, 1)$. The graph is thus bipartite between the target words V_0 and all the source words $V_i, i > 0$. The source sentences are tagged for parts of speech and thus each source word v_s packs a label distribution $p(l|v_s)$ of tagger confidences across parts of speech $l \in L$.

On top of this parallel dataset, we implement the best practices in annotation projection of sequential labels from multiple sources with low-resource target languages in mind:

- Word alignments are obtained from an IBM1 model Efmaral (Östling and Tiedemann, 2016) as Agić et al. (2016) show that simpler alignment models favor low-resource languages. Thus we acquire all $a(v_s, v_t) \in A$.
- Source sentences are tagged for parts of speech by a state-of-the-art neural tagger with default settings (Plank et al., 2016). That way all source words attain a tag distribution $p(l|v_s)$.
- Source tags are projected through the word alignments and accumulated at the target ends:

$$\text{BALLOT}(l|v_t) = \sum_{v_s \in V_s} p(l|v_s) a(v_s, v_t).$$

  The part-of-speech tag for each target word v_t is finally decoded through simple weighted majority voting:

$$\text{LABEL}(v_t) = \arg\max_{l} \text{BALLOT}(l|v_t).$$

- The sentences are further filtered so as to remove noisy instances. The model by Plank et al. (2018) is used, whereby for training we select only the top 10 thousand target sentences ranked by mean word alignment coverage c_t:

$$c_t = \frac{1}{n} \sum_{i=1}^{n} c_{i,t}.$$

  Mean coverage c_t is defined through individual source-target coverages, for all $i > 0$:

$$c_{i,t} = \frac{|\{v_t : \exists v_s, v_s \in V_i, a(v_s, v_t) \in A\}|}{|V_t|}.$$

	Bible	DsDs	JW300 proj
Bulgarian (BG)	77.7	83.9	82.7
Croatian (HR)	67.1	78.0	77.7
Czech (CS)	73.3	86.8	82.5
Danish (DA)	79.0	84.5	84.8
English (EN)	73.0	85.7	80.3
French (FR)	76.6	88.7	84.9
German (DE)	80.2	84.1	83.3
Greek (EL)	52.3	81.1	76.1
Hindi (HI)	67.6	63.1	73.4
Hungarian (HU)	72.0	77.3	76.3
Italian (IT)	76.9	92.1	85.2
Norwegian (NO)	76.7	86.2	83.1
Persian (FA)	59.6	43.6	66.6
Polish (PL)	75.1	84.4	83.2
Portuguese (PT)	83.8	89.4	86.9
Spanish (ES)	81.4	91.7	87.0
Swedish (SV)	75.2	83.1	79.7
μ	73.4	81.4	80.8

Table 3: Accuracy of part-of-speech taggers induced by projection from multiple sources of JW300, in comparison to projections from the Bible by Agić et al. (2015) and the DsDs system by Plank et al. (2018) which learns from multiple sources of weak supervision including annotation projection.

  We also remove all sentences under 3 and over 100 tokens. Finally, the target language taggers are trained on these 10 thousand filtered projections and evaluated on held-out test data. We use the same part-of-speech tagger by Plank et al. (2016) for the target languages as we did for the source languages.

Baselines and data. In this experiment we compare three distantly supervised systems:

- the bare-bones Bible annotation projection by Agić et al. (2015),
- a state-of-the-art system DsDs by Plank et al. (2018) which combines annotation projection, type supervision with Wiktionary and UniMorph (Kirov et al., 2018), word embeddings, and subword representations, and finally
- JW300 proj which is our own multi-source projection with JW300 data as defined above.

The training data is Universal Dependencies version 2.3 (Nivre et al., 2018). The test data amounts to 17 languages at the intersection of the three systems and comes from Plank and Agić (2018). All tags are converted to the tagset of Petrov et al. (2011) for comparability.

Results. Table 3 lists the tagging accuracy by language and system. Projections from our system JW300 PROJ are expectedly superior to those by BIBLE by +7.4 increase in mean accuracy across all 17 languages. On a more interesting note, our barebones approach to annotation projection falls only -0.6 points short of DSDS on average, which is an admirable feat since DSDS is an intricate multi-task learning system which learns from several disparate signals of distant supervision, only one of which is annotation projection.

Beyond the confines of the 17-language comparison from Table 3, we also conduct one larger experiment with 42 languages in the overlap of JW300 and Universal Dependencies v2.3. The mean accuracy for the 17 languages in Table 3 increases with this additional multi-source support by +0.8 points absolute, to 81.6 which now just surpasses the score of DSDS. Since these systems are complementary, future work could further explore the benefits of injecting the improved JW300 projections to more complex learners such as DSDS. In particular, DSDS would likely benefit from better projections, since the ones that its current instance uses are inferior to JW300.

4 Related work

Our work is a contribution to the pool of massively multilingual resources. In that pool we already singled out OPUS (Tiedemann, 2012) as the largest collection of freely available parallel sentences to date. OPUS is a collection that covers large datasets such as Europarl (Koehn, 2005), OpenSubtitles (Lison and Tiedemann, 2016), along with many others. OPUS also contains a smaller snapshot of Tatoeba, whose original collection hosts 337 languages and 22,427 (±106,815) sentences on average.[6]

Moving from OPUS and Tatoeba towards greater linguistic breadth, there are several publicly available Bible datasets, most notably those by Mayer and Cysouw (2014) and Christodouloupoulos and Steedman (2015). The Bible datasets are typically aligned by verse and not by sentence, because verse identifiers are assigned by humans, with absolute accuracy. However, a verse sometimes comprises several sentences, or alternatively just parts of one sentence, thus in effect replacing one type of alignment noise with another. Our results strongly favor JW300 for part-of-speech projection.

Prior to our work, Agić et al. (2016) have also collected a smaller dataset from `jw.org` to produce cross-lingual dependency parsers with multi-source projection. Their dataset covers 135 languages with a mean of 115,856 sentences per language (±34,898), but with sentence alignments only within a group of 27 languages.

Our contribution JW300 strikes a balance between multilingual and intra-language coverage that will greatly facilitate future research in large-scale cross-lingual processing. Our work is entirely complementary to related efforts in bringing forth massively multilingual resources.

5 Conclusion

We introduced JW300, a large collection of parallel texts that spans over more than 300 languages, and offers 54 thousand pairs of alignments, each with roughly 100 thousand parallel sentences on average. We posit that the dataset would prove immensely useful for a wide variety of research in cross-lingual processing. JW300 is freely available for all non-commercial use as per terms of the data owner.

Our two experiments show that even with simple models JW300 offers top performance in cross-lingual word embedding induction and multilingual projection for part-of-speech tagging, where we reach or even surpass more advanced models.

Acknowledgements

The authors acknowledge the NVIDIA Corporation for supporting their work. We are also thankful to the anonymous reviewers and area chairs for their incisive reviews.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of ACL*, pages 268–272.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Cross-lingual tagger evaluation without test data. In *Proceedings of EACL*, pages 248–253.

[6] https://tatoeba.org/eng/stats/sentences_by_language

Rami Al-Rfou. 2015. *Polyglot: A Massive Multilingual Natural Language Processing Pipeline*. Ph.D. thesis, Stony Brook University.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP*, pages 1881–1890.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NIPS*, pages 4356–4364.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of EMNLP*, pages 261–270.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pages 600–609.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *Proceedings of AISTATS*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *CoRR*, abs/1902.00508.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of EMNLP*, pages 469–478.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. *arXiv preprint arXiv:1810.11101*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86.

Fethi Lamraoui and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner: Time to reconsider sentence alignment? In *Proceedings of the XIV Machine Translation Summit*.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of *SEM*, pages 85–91.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: extracting large parallel corpora from movie and tv subtitles. In *Proceedings of LREC*.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of LREC*.

Joakim Nivre et al. 2018. Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of EMNLP*, pages 614–620.

Barbara Plank, Sigrid Klerke, and Željko Agić. 2018. The best of both worlds: Lexical resources to improve low-resource part-of-speech tagging. *arXiv preprint arXiv:1811.08757*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL*, pages 412–418.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*, pages 2214–2218.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING*, pages 1854–1864.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of NAACL-HLT*, pages 1–8.