Zero-Shot Cross-Lingual Dependency Parsing through Contextual Embedding Transformation

Haoran Xu and **Philipp Koehn**
Johns Hopkins University
hxu64@jhu.edu, phi@jhu.edu

Abstract

Linear embedding transformation has been shown to be effective for zero-shot cross-lingual transfer tasks and achieve surprisingly promising results. However, cross-lingual embedding space mapping is usually studied in static *word-level* embeddings, where a space transformation is derived by aligning representations of translation pairs that are referred from dictionaries. We move further from this line and investigate a contextual embedding alignment approach which is *sense-level* and dictionary-free. To enhance the quality of the mapping, we also provide a deep view of properties of contextual embeddings, i.e., the anisotropy problem and its solution. Experiments on zero-shot dependency parsing through the concept-shared space built by our embedding transformation substantially outperform state-of-the-art methods using multilingual embeddings.

1 Introduction

Cross-lingual embedding space alignment (Mikolov et al., 2013b; Artetxe et al., 2016; Xing et al., 2015; Conneau et al., 2018) recently has been attracted a lot of attention because cross-lingual model transfer is effectively facilitated by shared semantic spaces in NLP tasks, e.g., named entity recognition (Xie et al., 2018), part-of-speech tagging (Hsu et al., 2019), and dependency parsing (Schuster et al., 2019), where dependency paring is scoped out in this paper. Compared with the delexicalized parsers (McDonald et al., 2011), multilingual word embeddings have been demonstrated to significantly improve the performance of zero-shot dependency parsing by bridging the lexical feature gap (Guo et al., 2015).

With the remarkable development of monolingual contextual pre-trained models (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019), which dramatically outperform static word embeddings (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017) in broad NLP applications, increasing number of researchers have started focusing on contextual representation alignment for cross-lingual dependency parsing (Schuster et al., 2019; Wang et al., 2019). Moreover, with the appearance of multilingual pre-trained models, such as Multilingual BERT (mBERT) (Devlin et al., 2019), zero-shot dependency parsing becomes easier by utilizing the large vocabulary of the multilingual models (Kondratyuk and Straka, 2019).

Our approach is most similar to Schuster et al. (2019), which maps a target language space into a source language space through a linear transformation to realize zero-shot transfer in dependency parsing. Typically, a transformation is usually derived by word-level embedding alignment, while we explore a sense-level embedding alignment method to map bilingual spaces more precisely, where sense-level representations are split from multi-sense word-level embeddings. Furthermore, our mapping approach is dictionary-free which utilizes the silver token pairs from parallel corpora and eliminates the necessity of gold dictionaries. The experimental results of zero-shot dependency parsing demonstrate that two parser evaluation scores (UAS and LAS) of sense-level mapping are always better than of word-level one. Moreover, we also notice the anisotropy problem (Ethayarajh, 2019) (defined in Section 3.2) in contextual embeddings, which potentially deteriorate the performance of the zero-shot transfer task. We significantly mitigate this drawback by leveraging a prepossessing step, **i**terative **n**ormalization (IN) (Zhang et al., 2019), which is originally used for improving the performance of static embedding mapping on the bilingual dictionary induction task.

Zero-shot dependency parsing experiments are conducted on *Universal Dependencies treebank*

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 204–213
April 20, 2021.

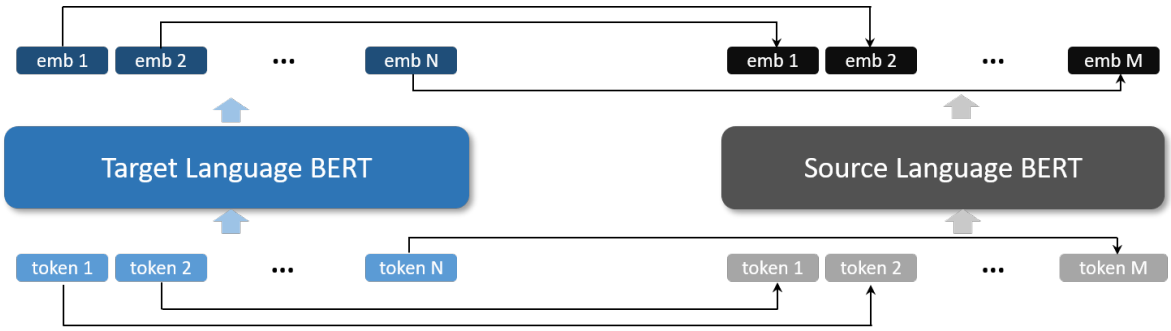


Figure 1: The target tokens (left, blue) and the source tokens (right, black) are aligned by *Fast Align*, so their contextual embeddings can be aligned as well.

v2.6 (Zeman et al., 2020), which shows that our results obtain a substantial gain compared with state-of-the-art methods using multilingual fastText and mBERT [1].

2 Linear Cross-lingual Space Alignment

Let denote $X \in \mathbb{R}^{d \times N}$ as the word embedding matrix for a target language [2], and Y as the word embedding matrix for a source language. For each column of the target embedding matrix $x_i \in \mathbb{R}^d$, it has one source embedding vector $y_i \in \mathbb{R}^d$ corresponding to a source word translated from the target word i. We aim to derive a linear transformation matrix $\hat{W}$ used for mapping from the target language space to the source language space. This can be learned by minimizing the Frobenius norm:

$$\hat{W} = \arg\min_{W \in \mathbb{R}^{d \times d}} \|WX - Y\|_F \quad (1)$$

Furthermore, Xing et al. (2015) show that the quality of space alignment is successfully improved with the orthogonal restriction, i.e, $W^T W = I$. Thus, the problem can be solved by Procrustes approach (Schönemann, 1966):

$$\begin{aligned} \hat{W} &= \arg\min_{W \in \mathbb{O}^{d \times d}} \|WX - Y\|_F = UV^T \\ &\text{s.t. } U\Sigma V^T = \text{svd}(YX^T) \end{aligned} \quad (2)$$

where $\mathbb{O}^{d \times d}$ is the set of orthogonal matrices.

3 Method

3.1 Contextual Embedding Transformation

An unsupervised bidirectional word alignment algorithm based on *IBM Model 2* (Brown et al., 1993), *Fast Align* (Dyer et al., 2013), is first applied to a parallel corpus to derive silver aligned token pairs. We then respectively feed the parallel corpus to the BERTs of the target and the source languages and extract the outputs as contextual embeddings. As shown in Figure 1, *Fast Align* bridges "links" between silver token pairs, and between the embeddings of the token pairs as well. Thus, for each target type, a collection of its contextual embeddings can be obtained, as well as a collection of contextual embeddings of its aligned source tokens. Vectors are normalized to satisfy the orthogonal condition.

Motivated by the assumption that multiple senses of a type can construct multiple distinct clusters in its collection (Schuster et al., 2019), we derive several sense-level (cluster-level) embeddings for a type by averaging vectors in each cluster. This splits the representations of multi-sense words and helps the anchor-driven space mapping in a finer resolution. To find clusters, we utilize k-means to cluster contextual embeddings in the vector collection of each type, and adaptively find the optimal k by an elbow-based method Satopaa et al. (2011). Contextual vectors are only clustered in the target side to obtain sense-level embeddings, while the aligned sense-level embeddings in the source side can also be simultaneously derived because embeddings have been already "linked" by *Fast Align*. We next build a sense embedding matrix X_s for the target language by putting the sense-level embeddings in each column, and meanwhile construct a column-wise aligned sense embedding matrix Y_s in the source side. Finally, we obtain the optimal linear mapping $\hat{W}$ from X_s to Y_s by Equation 2. Pseudo code of transformation method is in Appendix A.

[1] Code is available at: https://github.com/felixxu/ZeroShot-CrossLing-Parsing.

[2] Different from usual settings, we use x-related symbols for target data and y-related ones for source data.

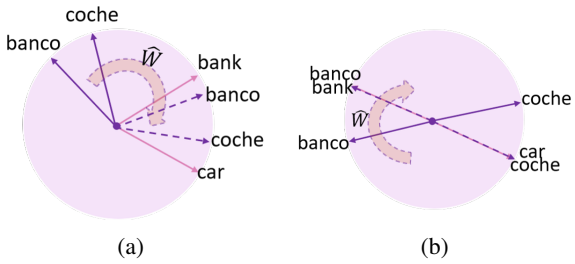


Figure 2: (a) Spanish vectors (purple arrows) cannot well fit to English vectors (pink arrows) by a linear transformation because they gather in different degrees of cones (different angles between vectors), where dash lines are mapped vectors. (b) After iterative normalization, Spanish and English vectors are uniformly distributed (same angles between vectors). They can be perfectly fit after mapping now.

3.2 Anisotropy in Embedding Spaces

Our findings show that contextual embeddings always hold anisotropic property, i.e., they are not uniformly distributed in the space and gather toward a narrow range of orientations. Importantly, degrees of anisotropy across languages are various, which undermines the quality of cross-lingual mappings. A toy example of how the anisotropy affect mappings is illustrated in Figure 2a. One metric for anisotropy is to calculate the average cosine similarity distance of randomly selected vectors. The higher the distance is, the narrower directions vectors point to. Note that the distance for an isotropic space is 0. To mitigate this problem, we introduce iterative normalization. For each token i, the embedding vector x_i is forced to be zero-mean firstly in each iteration:

$$x_i = x_i - \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3}$$

and then normalize it to a fixed length:

$$x_i = \frac{x_i}{\|x_i\|_2} \tag{4}$$

The two steps are repeated until convergence. N is the total number of embeddings. The iterative preprocessing enforces the space to be uniformally distributed, and relative angles between vectors across languages to be more similar (Figure 2b).

3.3 Zero-shot Transfer

A parser is first trained on a source language treebank, where outputs of a frozen BERT are used as embeddings. To apply the pre-trained parser to the target languages, we first replace the source BERT with the target BERT. Then, iterative normalization is operated to enforce contextual embeddings in a near-isotropic space. At last, we map the embeddings to the source language space. Specifically, for each target token i, its contextual representation x_i is mapped by $\hat{W}x_i$. The processing of zero-shot dependency parsing is visualized in Figure 3. Note that the space of pre-trained model has already fit to be near-isotropic by utilizing iterative normalization during training.

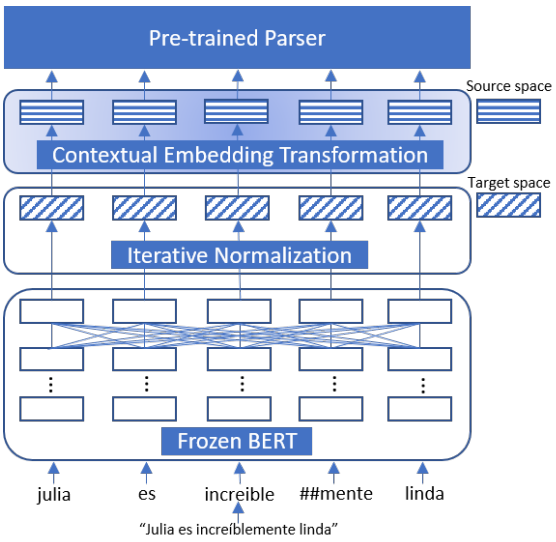


Figure 3: The workflow of how zero-shot transfer processes in our model architecture.

4 Experiment

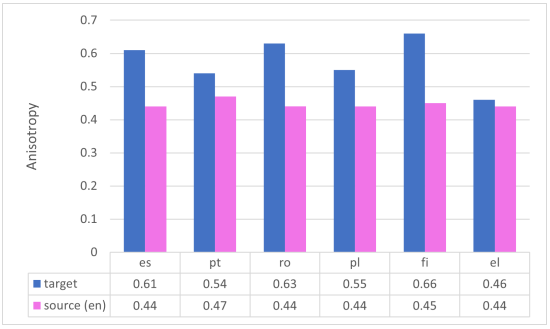
Our parser is the deep biaffine model from Dozat and Manning (2016) where hyperparameters are almost unchanged. The settings of all hyperparameters are listed in Appendix B. English is set as the source language and other languages are targets. In our experiments, we select 6 target languages from 4 language families for which we have off-the-shelf monolingual pre-trained BERT models (base-size). We train the parsing model only in the English treebank, and directly evaluate zero-shot transfer performance on the target languages.

4.1 Baseline

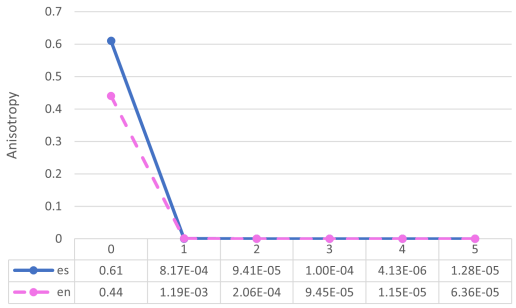
Aligned fastText: Our first baseline is multilingual fastText aligned by the RCSLS method (Joulin et al., 2018; Bojanowski et al., 2017) which is straightforwardly employed to the embedding layer for the corresponding language.

lang (*treebank*)	**en** (*ewt*)		**es** (*gsd*)		**pt** (*gsd*)		**ro** (*rrt*)		**pl** (*lfg*)		**fi** (*tdt*)		**el** (*gdt*)	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
aligned fastText	88.55	86.36	73.57	65.13	72.41	60.69	60.27	46.79	75.88	59.48	62.32	40.78	71.51	61.46
mBERT uncased	93.45	91.52	82.11	72.51	80.89	68.90	72.08	56.91	**85.27**	69.76	72.76	49.64	81.72	68.35
mBERT cased	93.32	91.34	82.83	74.08	80.80	68.68	70.76	56.04	83.77	68.01	71.82	48.84	78.24	65.92
word-level	93.70	91.78	82.43	73.86	79.77	67.35	71.13	57.28	84.58	69.53	74.65	51.06	82.29	69.88
sense-level			82.55	73.92	80.34	67.80	71.46	57.57	84.71	69.56	74.81	51.14	82.33	70.10
word-level + IN	**94.21**	**92.01**	83.70	75.14	81.48	69.04	74.65	59.68	84.32	70.45	75.07	51.75	83.76	71.11
sense-level + IN			**83.91**	**75.39**	**81.99**	**69.49**	**74.78**	**59.83**	84.57	**70.52**	**75.31**	**51.99**	**84.05**	**71.26**

Table 1: UAS and LAS of zero-shot evaluation for various languages on test files. The highest scores are bolded and the second highest scores are underlined. Language families are split by dash lines. lang = language, en = English, es = Spanish, pt = Portuguese, ro = Romanian, pl = Polish, fi = Finnish, el = Greek.



(a)



(b)

Figure 4: (a) Discrepancy of anisotropic degrees for all tested language pairs, where scores of anisotropic degree are calculated by the mean cosine similarity between 1000 randomly selected vectors in their language spaces. (b) The isotropic degrees basically decrease to 0 at the first iteration and converge afterwards.

mBERT: We compare our approach with both uncased and cased version of mBERT. Outputs of mBERT are directly used for the embedding layer.

4.2 Settings

Following the analysis that top layers of BERT contain more semantic information (Jawahar et al., 2019), our contextual representation are normalized mean vector of the last 4 layers of BERT. The parallel corpora used to extract contextual embeddings are obtained from *ParaCrawl v6.0* [3]. For each language pairs, we select 1M parallel sentences whose length is shorter than 150. Since some noisy alignments are produced during *Fast Align*, we only take one-to-one token alignment into consideration. The dataset used for cross-lingual dependency parsing is the *Universal Dependencies treebank v2.6* [4] (Zeman et al., 2020).

We store up to 10K contextual vectors extracted from BERT for non-OOV tokens [5]. Vectors in the collection of a target type are clustered to derive sense-level embeddings only if the token occurs more than 100 times. Otherwise, the representation for the token is the basic word-level embedding, i.e., the mean vector of its vector collection. Experiments of word-level embedding alignment are also conducted to compare with sense-level results.

4.3 Iterative Normalization Preprocessing

Forcing contextual embedding vectors in X_s and Y_s to be zero-mean is straightforward. Nevertheless, it is difficult to look for the universal mean vector of contextual embeddings when we train the English parser, because we do not have such an exact mean vector for all possible contextual embeddings. Thus, to successfully implement IN for pre-training the parser, we calculate the approximate universal mean vector by averaging all contextual vectors of every occurrence of tokens from the given training dataset in each iteration. IN runs for 5 iterations, which is sufficient for convergence.

5 Discussion

5.1 Why Contextual Embedding Mapping?

Compare with Previous Methods: Overall results are shown in Table 1. In the first place, our contextual-aware embedding mapping (row 4 - 7)

[3] www.paracrawl.eu

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3226

[5] We do not use the composition of subword vectors to approximately represent OOV tokens, because our preliminary results show this hurts the mapping.

exceeds the aligned fastText (row 1) by a large margin. Moreover, our sense-level mapping without IN preprocessing outperforms uncased and cased mBERT by 0.67% and 1.42% on LAS on average, and mapping with preprocessing further outperforms them by 2.07% and 2.82% on average.

Dictionary-free Mapping: Typically, aligned embeddings take a static dictionary as reference but high-quality manual dictionaries are still rare (Ruder et al., 2019). Our mapping skips the word-level alignment in dictionaries, and directly aligns the embeddings from parallel corpora which offers a large scope of token alignments.

Sense-level Mapping: Different from static embeddings whose words only have one unique representation, our contextual embeddings also take advantage of multiple representations for multi-sense words to improve the quality of anchor-driven mapping. In Table 1, the performance of sense-level mapping always surpasses word-level mapping.

5.2 Effect of Iterative Normalization

Figure 4a illustrates the various degrees of anisotropy among different language pairs. As we expect, the anisotropic degree for English (pink, right) is basically constant, but there is large discrepancy between other target languages (blue, left). After IN preprocessing, all language spaces are approximately isotropic, where their scores of anisotropy dramatically reduce near to zero. One example of how the anisotropic degree drops down in each iteration of IN for the Spanish-English pair is illustrated in Figure 4b. IN assists the aligned embeddings in building more similar relative angles across embeddings in different language spaces. As shown in Table 1, this preprocessing improves an absolute gain of 1.37% for word-level mapping and 1.40% for sense-level mapping on average.

6 Conclusion

We proposed a linear, dictionary-free and sense-level contextual mapping approach by exploiting parallel corpus which has shown promising results and substantial improvement compared with multilingual fastText and mBERT in the zero-shot dependency parsing task. We also revealed that various degrees of anisotropy hurts the performance of mapping, and introduced iterative normalization to alleviate it by enforcing contextual embeddings to be uniformly distributed, which also has indicated the benefits of isotropy.

Acknowledgments

We thank the anonymous reviewers for their valuable comments.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2289–2294, Austin, Texas. 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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311.

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

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.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. 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 1234–1244, Beijing, China. Association for Computational Linguistics.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

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 the 2018 Conference on Empirical Methods in Natural Language Processing.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke 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, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65:569–631.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In 2011 31st international conference on distributed computing systems workshops, pages 166–171. IEEE.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. Psychometrika, 31(1):1–10.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. 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 5725–5731.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In Proceedings

of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor

Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3180–3189, Florence, Italy. Association for Computational Linguistics.

A Pseudo Code of Contextual Embedding Transformation

Pseudo Code is shown in Alogirthm 1.

B Hyperparamters

Here we list all hyperparamters for our pre-trained parser in Table 2.

Hyperparameters	Value
Batch size	128
Arc representation dim	500
Tag representation dim	100
Dropout	0.3
LSTM hidden size	500
LSTM # layers	3
Pos tag embedding dim	100
Grad norm	5
# epochs	200
Patience	25
Optimizer	Dense Sparse ADAM
Learning rate	0.0008
Encoding Layer	Bi-LSTM

Table 2: Hyperparameters for deep biaffine dependency parser training.

C pre-trained Monolingual BERTs

In Table 3, we list the names of pre-trained monolingual BERTs from huggingface [6] that we used in our experiments.

Language	Model name
mbert uncased	bert-base-multilingual-uncased
mbert cased	bert-base-multilingual-cased
en	bert-base-uncased
es	dccuchile/bert-base-spanish-wwm-uncased
pt	neuralmind/bert-base-portuguese-cased
ro	dumitrescustefan/bert-base-romanian-uncased-v1
pl	dkleczek/bert-base-polish-uncased-v1
fi	bert-base-finnish-uncased-v1
el	nlpaueb/bert-base-greek-uncased-v1

Table 3: Names of Pre-trained BERT models.

[6] https://huggingface.co/models

Algorithm 1 Contextual Embedding Transformation

Require: Target Corpus $\mathcal{X}$, source Corpus $\mathcal{Y}$, target pre-trained BERT $\mathcal{B}_x$, source pre-trained BERT $\mathcal{B}_y$, where $\mathcal{X}$ is the translation corpus of $\mathcal{Y}$

function CONTEXTUAL-TRANSFORMATION($\mathcal{X}$, $\mathcal{Y}$, $\mathcal{B}_x$, $\mathcal{B}_y$)
    # Part 1: Collect embeddings
    $\mathcal{I} \leftarrow$ FAST-ALIGN($\mathcal{X}$, $\mathcal{Y}$) ▷ $\mathcal{I}$ is an index-aligned corpus, where each line is composed of index pairs of aligned tokens for each parallel sentence.
    Initialize $\mathcal{C} \leftarrow$ Empty Hash Map
    for index i in LENGTH($\mathcal{X}$) **do** ▷ number of sentences in the corpus
        $X \leftarrow \mathcal{X}[i], Y \leftarrow \mathcal{Y}[i], I \leftarrow \mathcal{I}[i]$
        $E_X \leftarrow \mathcal{B}_x(X)$ ▷ Contextual embeddings of tokens:
        $E_Y \leftarrow \mathcal{B}_y(Y)$
        for index j in LENGTH(X) **do** ▷ number of tokens in the sentence
            $x \leftarrow X[j], e_x \leftarrow E_X[j]$
            $e_y \leftarrow E_Y[I(j)]$ ▷ Find the aligned embedding by looking at I, where $I(j)$ is the index of aligned source token.
            $\mathcal{C}[x]$.append((e_x, e_y))
        end for
    end for

    # Part 2: Obtain aligned sense-level embeddings
    Initialize Empty matrix X_s, Y_s
    for target type x in $\mathcal{C}$.keys() **do**
        $c_x \leftarrow$ all target embeddings e_x in $\mathcal{C}[x]$
        $c_y \leftarrow$ all target embeddings e_y in $\mathcal{C}[x]$
        $k \leftarrow$ ELBOW-BASED(c_x) ▷ Find optimal number of clusters
        for Subcluster c_{x_i} in K-MEANS(k, c_x) **do**
            Get subcluster c_{y_i} due to aligned pair $((e_x, e_y))$ in $\mathcal{C}[x]$
            $mean_x \leftarrow$ mean vector of c_{x_i}
            $mean_y \leftarrow$ mean vector of c_{y_i}
            Put $mean_x$ in X_s as a column
            Put $mean_y$ in Y_s as a column
        end for
    end for

    # Part 3: Derive embedding transformation
    $U\Sigma V^T = \mathrm{svd}(YX^T)$
    $\hat{W} = UV^T$
    return $\hat{W}$
end function