XL-WA: a Gold Evaluation Benchmark for Word Alignment in 14 Language Pairs

Federico Martelli[11,*], Andrei Stefan Bejgu[2,11], Cesare Campagnano[11], Jaka Čibej[14], Rute Costa[10], Apolonija Gantar[14], Jelena Kallas[6], Svetla Koeva[3], Kristina Koppel[6], Simon Krek[9], Margit Langemets[6], Veronika Lipp[5], Sanni Nimb[12], Sussi Olsen[13], Bolette Sandford Pedersen[13], Valeria Quochi[8], Ana Salgado[1,10], László Simon[5], Carole Tiberius[7], Rafael-J Ureña-Ruiz[4] and Roberto Navigli[11]

[1]*Academia das Ciências de Lisboa, Portugal*
[2]*Babelscape, Italy*
[3]*Bulgarian Academy of Sciences, Bulgaria*
[4]*Centro de Estudios de la Real Academia Española, Spain*
[5]*HUN-REN Hungarian Research Centre for Linguistics, Hungary*
[6]*Institute of the Estonian Language, Estonia*
[7]*Instituut voor de Nederlandse Taal, The Netherlands*
[8]*Istituto di Linguistica Computazionale "A.Zampolli", Consiglio Nazionale delle Ricerche, Italy*
[9]*Jožef Stefan Institute, Slovenia*
[10]*NOVA CLUNL, Portugal*
[11]*Sapienza University of Rome, Italy*
[12]*Society for Danish Language and Literature, Denmark*
[13]*University of Copenhagen, Denmark*
[14]*University of Ljubljana, Slovenia*

Abstract

Word alignment plays a crucial role in several Natural Language Processing tasks, such as lexicon injection and cross-lingual label projection. The evaluation of word alignment systems relies heavily on manually-curated datasets, which are not always available, especially in mid- and low-resource languages. In order to address this limitation, we propose XL-WA, a novel entirely manually-curated evaluation benchmark for word alignment covering 14 language pairs. We illustrate the creation process of our benchmark and compare statistical and neural approaches to word alignment in both language-specific and zero-shot settings, thus investigating the ability of state-of-the-art models to generalize on unseen language pairs. We release our new benchmark at: https://github.com/SapienzaNLP/XL-WA.

Keywords

Word Alignment, Deep Learning, Natural Language Processing, Multilinguality

1. Introduction

Word alignment is the computational task of identifying translation correspondences at word and multi-word level between parallel sentences [1, 2]. Historically, word alignment played a crucial role in Statistical Machine Translation [3, 4, SMT]. However, while SMT has been replaced by end-to-end neural architectures which attain considerably higher performances, word alignment – also thanks to novel neural approaches – still plays a crucial role in many other Natural Language Processing (NLP) tasks, such as lexicon injection and, most importantly, cross-lingual annotation projection [5]. For instance, Procopio et al. [6] recently proposed a state-of-the-art approach to cross-lingual label projection based on word alignment which allows high-quality sense-tagged datasets to be produced automatically. Furthermore, word alignment has also been leveraged effectively

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
martelli@diag.uniroma1.it (F. Martelli); bejgu@babelscape.com (A. S. Bejgu); campagnano@di.uniroma1.it (C. Campagnano); Jaka.Cibej@ijs.si (J. Čibej); costamrv@gmail.com (R. Costa); apolonija.gantar@guest.arnes.si (A. Gantar); jelena.kallas@eki.ee (J. Kallas); svetla@dcl.bas.bg (S. Koeva); kristina.koppel@eki.ee (K. Koppel); simon.krek@ijs.si (S. Krek); margit@eki.ee (M. Langemets); lipp.veronika@nytud.hu (V. Lipp); sn@dsl.dk (S. Nimb); saolsen@hum.ku.dk (S. Olsen); vnb282@ku.dk (B. S. Pedersen); valeria.quochi@ilc.cnr.it (V. Quochi); anasalgado@fcsh.unl.pt (A. Salgado); simon.laszlo@nytud.hu (L. Simon); Carole.Tiberius@ivdnt.org (C. Tiberius); rafa@rae.es (R. Ureña-Ruiz); navigli@diag.uniroma1.it (R. Navigli)

to create silver datasets, not only for Word Sense Disambiguation [7, 8, WSD] but also for other semantic tasks, such as Semantic Role Labeling [9, 10, SRL], thereby addressing the knowledge acquisition bottleneck [11], especially when dealing with mid- and low-resource languages.

While, on the one hand, current architectures for word alignment are achieving increasingly better performance, on the other hand, the lack of high-quality manual data in multiple languages significantly limits their potential and scalability. With a view to addressing the aforementioned drawbacks, our contributions are as follows:

1. We propose a fully manually-annotated evaluation benchmark for word alignment with a total of 14 language pairs, each composed of English and one of the following languages: Arabic, Bulgarian, Chinese, Danish, Dutch, Estonian, Hungarian, Italian, Korean, Portuguese, Russian, Slovenian, Spanish and Swedish.

2. We experiment with statistical and neural approaches to word alignment and evaluate them against our newly created benchmark.

3. We demonstrate that the concatenation of our novel datasets can be exploited effectively to train a neural approach that generalizes on unseen languages in a zero-shot setting, thereby addressing the lack of training data in low-resource languages.

2. Related Work

Approaches Initial approaches to word alignment leveraged statistical and heuristic models [12]. Along these lines, several systems were proposed such as HMM [1], GIZA++[1] [13], PGIZA++, MGIZA++ [14] and FastAlign[2] [15]. Subsequently, statistical approaches were gradually substituted by neural counterparts and the advent of Transformer architectures [16] set a new standard in this task [17, 18, 19, 20, 21]. More recently, Procopio et al. [6] proposed a novel neural discriminative model for word alignment based on multilingual BERT [22], capable of significantly reducing the processing time.

Data Over the course of the last few decades, a number of datasets for word alignment, both manual and automatic, have been created, e.g. Czech-English[3] [23], Dutch-English [24], English-Turkish [25], English-Swedish[4] [26], Chinese-English[5] [27]. Interestingly, Graca et al. [28] proposed a collection of small datasets for word alignment in 6 language combinations; each dataset being composed of 100 sentences derived from the Europarl corpus[6] [29]. Among the currently available resources, we highlight the following contributions which we use in our experiments: the English-French and Romanian-English corpora released during the HLT-NAACL-2003 workshop on Building and Using Parallel Texts[7] [30], and the German-English dataset[8] proposed by Vilar et al. [31]. Finally, Neubig [32] presented a Japanese-English dataset[9] obtained by translating Wikipedia pages. However, despite the preceding efforts undertaken in this direction, to the best of our knowledge, no entirely manually-curated evaluation benchmark, which matches XL-WA in both size and language pairs covered, is currently available.

	# of Sentences		# of Alignments	
Lang.	**Dev**	**Test**	**Dev**	**Test**
EN-AR	90	210	1591	3597
EN-BG	105	245	1719	4179
EN-DA	105	245	1841	4136
EN-ES	105	245	1961	4722
EN-ET	105	245	1614	3722
EN-HU	105	245	1580	3781
EN-IT	103	243	1980	4765
EN-KO	90	210	1277	3007
EN-NL	105	245	1886	4490
EN-PT	105	245	1849	4578
EN-RU	90	210	1114	2582
EN-SL	105	245	1942	4537
EN-SV	90	210	1522	3530
EN-ZH	90	210	1724	4135
$\sum$	1393	3253	23600	55761

Table 1
Composition of XL-WA. We report from left to right: the available language combinations, the number of sentences and alignments divided by data split. In our experiments, we use approximately 30% of our data for development so as to obtain a more representative set.

3. XL-WA

To tackle the aforementioned gap, we introduce XL-WA, a novel entirely manually-curated evaluation benchmark

[1] https://github.com/moses-smt/giza-pp
[2] https://github.com/clab/fast_align
[3] https://ufal.mff.cuni.cz/czech-english-manual-word-alignment
[4] https://www.ida.liu.se/divisions/hcs/nlplab/resources/ges/
[5] https://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html
[6] https://www.statmt.org/europarl/
[7] https://web.eecs.umich.edu/~mihalcea/wpt/
[8] https://www-i6.informatik.rwth-aachen.de/goldAlignment/
[9] http://www.phontron.com/kftt

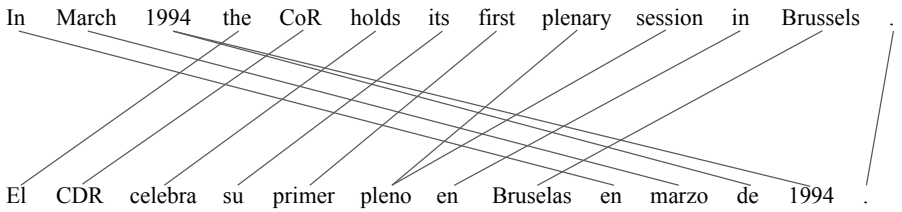

Figure 1: An example of alignment between English and Spanish, derived from the EN-ES dataset in XL-WA.

for word alignment. XL-WA is currently composed of 14 datasets, out of which 9 are parallel. The languages included in XL-WA cover 7 different language families, i.e. Afro-Asiatic, Indo-European (Germanic), Indo-European (Romance), Indo-European (Slavic), Sino-Tibetan, Uralic (Finno-Ugric) and Koreanic.

Importantly, all datasets include English as source language. This choice is motivated by the fact that enabling word alignment from English to multiple target languages is crucial for tasks such as label projection, where the majority of high-quality annotated data whose labels can be propagated is typically available in English.

We show the composition of our dataset in Table 1. Importantly, XL-WA is annotated exclusively by professional mother tongue annotators with a solid academic background and proven experience in linguistic annotation tasks. A detailed description of the format which we adopt is provided in Section 3.2.

3.1. Creation process

In this section, we detail the creation process and illustrate the guidelines adopted during the annotation phase.

The creation of XL-WA can be divided into three steps: i) *automatic extraction* of candidate sentences from a corpus, ii) *manual selection* of sentences satisfying specific linguistic criteria, and iii) *manual alignment.*

In order to obtain a balanced corpus in terms of domains and genres, similarly to the procedure adopted by Martelli et al. [33], we extract our data from WikiMatrix[10] [34], a wide-coverage collection of parallel sentences derived from the Wikipedia[11] corpus using an automatic approach based on multilingual sentence embeddings, covering 1620 different language combinations. First, we consider the WikiMatrix datasets containing English as the source language, and extract the highest number of overlapping source sentences across datasets. To this end, we compute a Boolean matrix $A \in \{0,1\}^{m \times n}$ where m is the number of English sentences in WikiMatrix and n the number of the target languages other than English covered in XL-WA.

We compute A such that A_{ij} contains: i) 1 if, for the i-th English sentence, a translation into the j-th target language is available, ii) 0 otherwise. We first extract the sentences shared in the highest number of languages. Subsequently, we manually discard sentences which are not well-formatted or contain significant grammatical errors. We then ask annotators to provide the missing translations in order to fill the gaps in our parallel dataset[12]. Finally, we ask our annotators to perform word alignment from scratch.

Guidelines All annotators are required to follow specific annotation guidelines for word alignment inspired by Lambert et al. [35], who provide detailed instructions and suggestions regarding the annotation of datasets for word alignment, including specific cases and exceptions. Importantly, annotators are asked to align source and target words also when these do not share the same part of speech. Furthermore, annotators are required to align complex lexical units such as compounds and multi-word expressions. For instance, given an open compound word c_{en}, e.g. *bus driver* in the English source sentence, translated into Dutch with the compound word c_{nl} *buschauffeur*, each component of c_{en} should be aligned to c_{nl}.

3.2. Alignment Format

We now describe our alignment format and provide an example for the language combination English-Spanish (EN-ES).

We adopt the Pharaoh alignment format [36]. Specifically, we use a Tab-Separated Values (TSV) format, where each row is formatted as follows: `source sentence<tab>target sentence<tab>alignments`. Tokens and alignments are separated by spaces; each alignment is composed of a pair of integers which identify the corresponding positions of source and target tokens, starting from zero. In order to deal with multi-word expressions in which 1:1 alignments are not possible, e.g., due to collocations or idiomatic expressions, we align all components of a given

[10] https://ai.facebook.com/blog/wikimatrix/
[11] https://www.wikipedia.org/

[12] Due to time constraints and the limited availability of professional annotators for specific language combinations, we carry out this step in 9 language combinations only.

multi-word expression in English with all components of the corresponding multi-word expression in the target language.

Below we report an example extracted from the EN-ES dataset:

- **Source:** `In March 1994 the CoR holds its first plenary session in Brussels .`
- **Target:** `El CDR celebra su primer pleno en Bruselas en marzo de 1994 .`
- **Alignments:** `3-0 4-1 5-2 6-3 7-4 8-5 9-5 10-6 11-7 0-8 1-9 2-10 2-11 12-12`

A visual representation of the above example is provided in Figure 1.

3.3. Inter-annotator agreement

Finally, in order to assess the reliability of our manual annotations, we compute the inter-annotator agreement[13]. To this end, we randomly select a sample of approximately 50 sentence pairs in two language combinations, namely EN-DA and EN-IT, and ask new annotators to align these manually. We compute the Cohen's kappa and obtain 0.94 and 0.89 in EN-DA and EN-IT, respectively. Importantly, these results indicate a remarkable level of agreement, which suggests a high degree of annotation consistency across datasets.

4. Experimental Setup

In this section, we illustrate our experimental setup and carry out a performance analysis. To this end, we put forward two different experimental settings. Specifically, we propose a comparison between statistical and neural approaches tested against our novel benchmark in a language-specific setting, i.e. we train and test on the same language pairs (Section 4.1.1). Subsequently, we investigate the behavior of models in a zero-shot setting, thus exploring the ability of state-of-the-art models to deal with languages unseen during the training phase (Section 4.1.2). Finally, we describe the evaluation metrics adopted.

4.1. Settings

We now describe our two experimental settings. Technical details regarding hyperparameters and hardware are reported in Appendix A.

4.1.1. Language-specific setting

Systems In this setting, we experiment with two statistical approaches, namely GIZA++ and FastAlign, and two state-of-the-art neural models, i.e. the SQuAD-style formulation for word alignment[14], which relies on multilingual BERT, proposed by Nagata et al. [20] and the MultiMirror neural word aligner by Procopio et al. [6].

For each language pair, the aforementioned statistical systems are trained on a randomly selected sample of 0.5M parallel sentences concatenated with our test data. Instead, for neural approaches requiring aligned data, which is not available in all our language combinations, we follow Garg et al. [17]. Specifically, we use sentences derived from the aforementioned silver training data, tagged both with GIZA++ and FastAlign, and randomly choose 1,000 sentences with the highest number of overlapping alignments.

Data For this setting, we derive training data from three well-established parallel corpora, namely, Europarl, WikiMatrix and UNPC[15] [37]. Importantly, this choice allows us to cover all language combinations considered. Instead, for validation and evaluation purposes we use the XL-WA datasets whose composition is reported in Table 1. In this case, our goal is to show and analyze the performance achieved by state-of-the-art models on each language pair.

4.1.2. Zero-shot setting

In the zero-shot setting, we experiment with MultiMirror only, since this model shows a reasonable balance between results and processing speed. Specifically, we train MultiMirror on the concatenation of our datasets and evaluate it against unseen language pairs, thus demonstrating the effectiveness of XL-WA when no aligned data is available in a given language combination. In this case, our goal is to determine the extent to which the model is able to generalize on language pairs unseen during training, i.e. EN-DE, EN-FR, EN-JA and EN-RO. The data is split as in Nagata et al. [20].

4.2. Evaluation metrics

As customary in the word alignment task, we adopt the following evaluation metrics: precision, recall and F1. In this work, we do not use the Alignment Error Rate (AER) metric, since previous works argue that AER is unlikely to be a useful metric for word alignment, due to its bias towards precision [4].

[13] Due to time constraints we compute the inter-annotator agreement in two language combinations.

[14] https://github.com/nttcslab-nlp/word_align

[15] https://opus.nlpl.eu/UNPC.php

	GIZA++ [13]			FastAlign [15]			SQuAD BERT [20]			MultiMirror [6]		
Lang.	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EN-AR	66.0	57.8	61.6	68.6	66.3	67.4	87.3	78.8	82.9	88.3	77.9	82.8
EN-BG	73.6	74.7	74.1	65.7	75.7	70.4	83.3	88.5	85.8	85.5	88.3	86.9
EN-DA	73.6	75.0	74.3	68.7	75.7	72.0	90.6	94.0	92.3	90.8	93.4	92.1
EN-ES	78.1	71.7	74.8	72.3	74.8	73.5	90.7	84.7	87.6	89.5	84.4	86.8
EN-ET	59.6	67.4	63.3	61.1	68.1	64.4	76.3	86.4	81.0	77.2	86.8	81.7
EN-HU	55.9	63.7	59.6	53.3	63.0	57.7	71.4	82.6	76.6	72.4	80.1	76.1
EN-IT	56.3	49.8	52.9	53.7	55.5	54.6	86.9	81.2	84.0	88.2	78.7	83.2
EN-KO	51.2	53.3	52.2	50.7	52.8	51.7	31.4	64.2	42.1	69.5	70.4	69.9
EN-NL	80.3	77.7	79.0	76.2	79.9	78.0	94.9	93.7	94.3	94.2	93.4	93.8
EN-PT	78.4	75.2	76.7	72.6	77.8	75.1	89.2	88.4	88.8	87.9	87.9	87.9
EN-RU	74.0	73.6	73.8	71.8	77.4	74.5	84.1	85.6	84.9	87.6	84.0	85.8
EN-SL	71.0	66.4	68.6	67.9	67.9	67.9	83.5	81.4	82.4	85.4	81.4	83.3
EN-SV	79.4	72.6	75.8	74.7	73.0	73.9	92.1	86.5	89.2	91.5	87.2	89.3
EN-ZH	47.4	41.9	44.5	49.6	48.2	48.9	78.8	70.7	74.5	79.2	71.7	75.3
Avg	67.5	65.8	66.5	64.8	68.3	66.4	81.5	83.3	81.9	84.8	83.3	83.9

Table 2
Comparison between statistical baselines (GIZA++ and FastAlign) and current state-of-the-art approaches (SQuAD BERT and MultiMirror) on our datasets. **P**, **R** and **F1** stand for **P**recision, **R**ecall and **F1**-score, respectively; all the scores are calculated using the Micro average. Note that the neural approaches are trained on silver data generated with the statistical baselines.

Lang.	**P**	**R**	**F1**
EN-DE	89.4	78.5	83.6
EN-FR	94.7	55.7	70.1
EN-JA	79.4	42.5	55.3
EN-RO	86.7	80.7	83.6
Avg	87.5	64.4	73.2

Table 3
Results of MultiMirror trained on all XL-WA datasets and evaluated on unseen data, i.e., in a zero-shot setting. To facilitate analysis and comparison, we keep English as source language.

5. Results

In this section, we discuss the results obtained. As can be seen in Table 2, in the language-specific setting, we observe a remarkable difference between statistical and neural approaches, with the latter outperforming the former by up to 17.5 points in terms of F1 score on average. In this setting, the best results are attained by Nagata et al. [20] in the English-Dutch (EN-NL) combination. Interestingly, we note that even neural models struggle to achieve good results in topic-prominent languages such as Chinese, Hungarian and Korean. In fact, in these languages, both statistical and neural approaches obtain significantly below-average results.

Instead, as far as the zero-shot scenario is concerned, we observe a good generalization capability of MultiMirror when trained on the concatenation of our novel datasets and tested against unseen language pairs, as reported in Table 3. In particular, the language combinations EN-DE and EN-RO attain a remarkable 83.6 F1 score. Importantly, this seems to suggest that the zero-shot paradigm can be employed as a viable approach to compensate effectively for the lack of annotated data in many low-resource languages.

Finally, we investigate the impact of the size of the training data generated with GIZA++ and FastAlign, as described in Section 4.1, on the overall performance achieved by MultiMirror[16]. To this end, we increase the size of the silver training data to 10,000 sentence pairs and compare the results obtained with those achieved in the previous setting where we use 1,000 sentence pairs. As can be seen in Table 4, the greater quantity of data allows us to achieve better results in terms of both precision and F1 score. However, interestingly, when training on 10,000 sentence pairs, MultiMirror reports a slightly inferior performance in terms of recall, with a decrease of 0.3 on average.

[16]As mentioned in Section 4.1.2, we use MultiMirror in this experiment due to a satisfactory trade-off between performance and processing speed.

	MultiMirror 1k sentences			MultiMirror 10k sentences		
Lang.	**P**	**R**	**F1**	**P**	**R**	**F1**
EN-AR	88.3	77.9	82.8	88.9	79.3	83.8
EN-BG	85.5	88.3	86.9	84.7	88.5	86.6
EN-DA	90.8	93.4	92.1	91.3	92.0	91.7
EN-ES	89.5	84.4	86.8	91.8	82.6	86.9
EN-ET	77.2	86.8	81.7	78.0	84.8	81.3
EN-HU	72.4	80.1	76.1	74.2	79.6	76.8
EN-IT	88.2	78.7	83.2	89.5	79.7	84.3
EN-KO	69.5	70.4	69.9	71.1	72.2	71.6
EN-NL	94.2	93.4	93.8	95.9	92.5	94.2
EN-PT	87.9	87.9	87.9	89.0	86.8	87.9
EN-RU	87.6	84.0	85.8	87.5	85.2	86.3
EN-SL	85.4	81.4	83.3	84.4	81.3	82.8
EN-SV	91.5	87.2	89.3	92.3	85.9	89.0
EN-ZH	79.2	71.7	75.3	80.5	72.3	76.2
Avg	84.8	83.3	83.9	85.7	83.0	84.2

Table 4
Comparison between MultiMirror trained on different size silver datasets. P, R and F1 stand for Precision, Recall and F1-score, respectively; all the scores are calculated using the Micro average.

6. Conclusion

In this work, we introduce XL-WA, a novel evaluation benchmark for word alignment in 14 language pairs. We detail the creation process for our novel evaluation suite, as well as our experimental setup in which we compare statistical and neural approaches to word alignment. We investigate the behavior of models in zero-shot scenarios and show that the concatenation of our datasets can be used effectively to align languages unseen during training, thus tackling the paucity or limited availability of data for word alignment in low-resource languages. We release our new benchmark at: https://github.com/SapienzaNLP/XL-WA.

As future work, we intend to investigate the impact of language-specific peculiarities on the overall performance of neural models for word alignment. Furthermore, we plan to increase the language coverage of XL-WA and, importantly, investigate the role played by additional low-resource languages in zero-shot settings. Finally, we aim to explore novel neural approaches to word alignment which can be employed in the field of cross-lingual label projection in order to create multilingual silver training datasets for several Natural Language Understanding tasks, such as WSD, SRL and Semantic Parsing.

Acknowledgments

The authors gratefully acknowledge the support of the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme. Furthermore, the authors are sincerely thankful for the support of the Estonian Research Council grant (PRG 1978). Finally, the authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR. This work has been carried out while Andrei Stefan Bejgu was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

[1] S. Vogel, H. Ney, C. Tillmann, Hmm-based word alignment in statistical translation, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-2141.

[2] J. Tiedemann, Word to word alignment strategies, in: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 212–218. URL: https://aclanthology.org/C04-1031.

[3] F. J. Och, C. Tillmann, H. Ney, Improved alignment models for statistical machine translation, in: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999. URL: https://aclanthology.org/W99-0604.

[4] A. Fraser, D. Marcu, Measuring word alignment quality for statistical machine translation, Computational Linguistics 33 (2007) 293–303. URL: https://aclanthology.org/J07-3002.

[5] D. Yarowsky, G. Ngai, Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora, in: Second Meeting of the North American Chapter of the Association for Computational Linguistics, 2001. URL: https://aclanthology.org/N01-1026.

[6] L. Procopio, E. Barba, F. Martelli, R. Navigli, Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 3915–3921. URL: https://www.ijcai.org/proceedings/2021/0539.pdf.

[7] E. Barba, L. Procopio, N. Campolungo, T. Pasini, R. Navigli, Mulan: Multilingual label propagation for word sense disambiguation, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Arti-

ficial Intelligence, 2021, pp. 3837–3844. URL: https://www.ijcai.org/Proceedings/2020/0531.pdf.

[8] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4330–4338. URL: https://doi.org/10.24963/ijcai.2021/593. doi:10.24963/ijcai.2021/593.

[9] S. Padó, M. Lapata, Cross-lingual annotation projection for semantic roles, Journal of Artificial Intelligence Research 36 (2009) 307–340. URL: https://www.jair.org/index.php/jair/article/download/10629/25416.

[10] A. Daza, A. Frank, X-SRL: A parallel cross-lingual semantic role labeling dataset, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3904–3914. URL: https://aclanthology.org/2020.emnlp-main.321. doi:10.18653/v1/2020.emnlp-main.321.

[11] W. A. Gale, K. W. Church, D. Yarowsky, A method for disambiguating word senses in a large corpus, Computers and the Humanities 26 (1992) 415–439. URL: https://www.jstor.org/stable/30204634.

[12] V. J. Della Pietra, The mathematics of statistical machine translation: Parameter estimation, Using Large Corpora (1994) 223. URL: https://aclanthology.org/J93-2003.pdf.

[13] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational linguistics 29 (2003) 19–51. URL: https://aclanthology.org/J03-1002.

[14] Q. Gao, S. Vogel, Parallel implementations of word alignment tool, in: Software engineering, testing, and quality assurance for natural language processing, 2008, pp. 49–57. URL: https://aclanthology.org/W08-0509.

[15] C. Dyer, V. Chahuneau, N. A. Smith, 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, 2013, pp. 644–648. URL: https://aclanthology.org/N13-1073.pdf.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017). URL: http://arxiv.org/abs/1706.03762.

[17] S. Garg, S. Peitz, U. Nallasamy, M. Paulik, Jointly learning to align and translate with transformer models, arXiv preprint arXiv:1909.02074 (2019). URL: https://aclanthology.org/D19-1453.

[18] E. Stengel-Eskin, T.-r. Su, M. Post, B. Van Durme, A discriminative neural model for cross-lingual word alignment, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 910–920. URL: https://aclanthology.org/D19-1084. doi:10.18653/v1/D19-1084.

[19] T. Zenkel, J. Wuebker, J. DeNero, Adding interpretable attention to neural translation models improves word alignment, arXiv preprint arXiv:1901.11359 (2019). URL: http://arxiv.org/abs/1901.11359.

[20] M. Nagata, K. Chousa, M. Nishino, A supervised word alignment method based on cross-language span prediction using multilingual BERT (2020). URL: https://aclanthology.org/2020.emnlp-main.41.

[21] M. J. Sabet, P. Dufter, F. Yvon, H. Schütze, Simalign: High quality word alignments without parallel training data using static and contextualized embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1627–1643. URL: https://aclanthology.org/2020.findings-emnlp.147.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[23] D. Mareček, Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus, Master's thesis, Charles University, MFF UK, 2008. URL: https://ufal.mff.cuni.cz/pcedt3.0/pubs/Marecek2008_diplomka.pdf.

[24] L. Macken, An annotation scheme and gold standard for dutch-english word alignment, in: 7th conference on International Language Resources and Evaluation (LREC 2010), European Language Resources Association (ELRA), 2010, pp. 3369–3374. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/100_Paper.pdf.

[25] M. T. Cakmak, S. Acar, G. Eryigit, Word alignment for english-turkish language pair., in: LREC, 2012, pp. 2177–2180. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/380_Paper.pdf.

[26] M. Holmqvist, L. Ahrenberg, A gold standard for english-swedish word alignment, in: Proceedings

of the 18th Nordic conference of computational linguistics (NODALIDA 2011), 2011, pp. 106–113. URL: https://aclanthology.org/W11-4615.pdf.

[27] Y. Liu, M. Sun, Contrastive unsupervised word alignment with non-local features, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015. URL: http://arxiv.org/abs/1410.2082.

[28] J. Graca, J. P. Pardal, L. Coheur, D. Caseiro, Building a golden collection of parallel multi-language word alignment., in: LREC, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/250_paper.pdf.

[29] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of machine translation summit x: papers, 2005, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11.

[30] R. Mihalcea, T. Pedersen, An evaluation exercise for word alignment, in: Proc. of HLT-NAACL, 2003, pp. 1–10. URL: https://aclanthology.org/W03-0301.

[31] D. Vilar, M. Popović, H. Ney, Aer: Do we need to "improve" our alignments?, in: Proc. of Workshop on Spoken Language Translation, 2006. URL: https://aclanthology.org/2006.iwslt-papers.7.pdf.

[32] G. Neubig, The Kyoto free translation task, 2011. URL: http://www.phontron.com/kftt.

[33] F. Martelli, R. Navigli, S. Krek, J. Kallas, P. Gantar, S. Koeva, S. Nimb, B. Sandford Pedersen, S. Olsen, M. Langemets, K. Koppel, T. Üksik, J. Dobrovoljc, R.-J. Ureña-Ruiz, J.-L. Sancho-Sánchez, V. Lipp, T. Váradi, A. Győrffy, S. László, V. Quochi, M. Monachini, F. Frontini, C. Tiberius, R. Tempelaars, R. Costa, A. Salgado, J. Čibej, T. Munda, Designing the ELEXIS parallel sense-annotated dataset in 10 european languages, in: Proceedings of the eLex Conference, 2021. URL: https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_22_pp377-395.pdf.

[34] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, arXiv preprint arXiv:1907.05791 (2019). URL: https://arxiv.org/pdf/1907.05791.pdf.

[35] P. Lambert, A. De Gispert, R. Banchs, J. B. Mariño, Guidelines for word alignment evaluation and manual alignment, Language Resources and Evaluation 39 (2005) 267–285. URL: https://link.springer.com/article/10.1007/s10579-005-4822-5.

[36] P. Koehn, Pharaoh: A beam search decoder for phrase-based statistical machine translation models, in: R. E. Frederking, K. B. Taylor (Eds.), Machine Translation: From Real Users to Research, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 115–124. URL: https://aclanthology.org/2004.amta-papers.13/.

[37] M. Ziemski, M. Junczys-Dowmunt, B. Pouliquen, The united nations parallel corpus v1. 0, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3530–3534. URL: https://aclanthology.org/L16-1561/.

[38] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, in: International Conference on Learning Representations, 2019. URL: https://iclr.cc/virtual_2020/poster_rkgz2aEKDr.html.

A. Hyperparameters and Hardware

In this appendix, we report the hyperparameters and hardware setup for the experiments described in the paper.

We adopt four approaches with the following hyperparameters:

- We use two statistical approaches, namely GIZA++ [13] and FastAlign [15]. We compile the code downloaded from the original repositories and we run all the experiments on CPU. Neither of the approaches requires any parameter tuning.
- SQuAD mBERT-based model [20], whose code is downloaded from the official repository. We run all the experiments using the default hyperparameters. For the sake of consistency and fairness, we do not tune any hyperparameters and use the optimal ones according to the authors, as specified in their paper. All the experiments run for 2 training epochs with a learning rate of 3×10^{-5} and a batch size of 6. Language-specific experiments run for approximately 20 minutes each. We also experiment with the whole multilingual dataset, which requires 4 hours to complete the training. Inference for the language-specific experiments takes around one minute per language on GPU.
- MultiMirror [6] is an mBERT-based model whose code is obtained from the authors for research purposes. All the experiments run with a patience of 50, using the RAdam [38] optimizer with a learning rate of $1e-05$ and a token batch size of 512. Language-specific experiments run for approximately 10 to 15 minutes each, while the multilingual experiment on the whole XL-WA dataset runs for approximately 1 hour. Inferences time is negligible: a few seconds on CPU for the language-specific data and around one minute for the whole dataset.

All the experiments are conducted on the same hardware, i.e. an Intel Core i7 7800x CPU and NVidia RTX 2080ti GPU with 11GB of VRAM.