Esposito: An English-Persian Scientific Parallel Corpus for Machine Translation

Anonymous submission

Abstract

Neural machine translation requires large number of parallel sentences along with in-domain parallel data to attain best results. Nevertheless, no scientific parallel corpus for English-Persian language pair is available. In this paper, a parallel corpus called Esposito is introduced, which contains 3.5 million parallel sentences in the scientific domain for English-Persian language pair. In addition, we present a manually validated scientific test set that might serve as a baseline for future studies. We show that a system trained using Esposito along with other publicly available data improves the baseline on average by 7.6 and 8.4 BLEU scores for En→Fa and Fa→En directions, respectively. Additionally, domain analysis using the 5-gram KenLM model revealed notable distinctions between our parallel corpus and the existing generic parallel corpus. This dataset will be available to the public upon the acceptance of the paper.

Keywords: Neural Machine Translation, Parallel Corpus, Domain Adaptation

1. Introduction

Machine translation is a natural language processing task that involves automatic translation of sentences from a source language into a target language. In recent years, neural machine translation has established itself as the most promising method to the field of machine translation. It shows superior performance on public benchmarks (Goyal et al., 2022) and rapid adoption in deployments, such as Google (Wu et al., 2016), Systran (Crego et al., 2016), and WIPO (Junczys-Dowmunt et al., 2016).

In order to train supervised neural machine translation models, millions of parallel sentences are needed. Nonetheless, this amount of data is not always available. Only for a small number of language pairs with high resource condition and particular domains, there are sufficient and openly accessible parallel corpora. There are several publicly available parallel corpora for the different language pairs, covering a wide range of topics and domains. Nevertheless, no scientific parallel corpus for English-Persian language pair is available.

Although English is the common language of scientific community, many researchers only have a basic command of the language and prefer to read scientific literature written in their native language. Machine translation can provide a solution to increase access to scientific publications. Even though there has been much work in the field of domain adaptation (Kocmi et al., 2022), the automatic translation of scientific publications has not received much attention from the community, in part because of the difficulty of collecting parallel documents.

In this paper, we introduce Esposito, which is an English-Persian scientific parallel corpus designated to improve the quality of machine translation. This corpus contains 3.5 million parallel sentence pairs for English-Persian, which is created from scientific publications' abstracts. Documents used to create this corpus are crawled from Open-Access (OA) journals registered in the Scientific Information Database (SID) portal[1]. We also apply Esposito as a training corpus for machine translation systems and show that a system trained using Esposito along with other publicly available data improves the baseline on average by 7.6 and 8.4 BLEU scores for En→Fa and Fa→En directions, respectively. We plan to publish Esposito upon the acceptance of the paper. We think this resource will be useful, especially for research related to scientific texts translation between English and Persian. This will facilitate equitable access to scientific knowledge and accelerate research in many fields.

The rest of the paper is laid out as follows. Section 2 discusses relevant studies and approaches in the field. Section 3 goes into further depth about how Esposito is built, outlining the procedures and methods used to create this corpus. We demonstrate the advantages of using Esposito in Section 4 by presenting results of our experiments. Finally, Section 5 summarizes the paper and provides potential future directions and research areas for further development.

2. Related Work

The development of parallel corpora for training machine translation systems have been an active research area in recent years. OPUS (Tiedemann, 2012a) is a collection of various corpora which covers many language pairs including English-Persian such as CCMatrix (Schwenk et al., 2021), Tanzil (Tiedemann, 2012b), TEP (Pilevar et al., 2011), along with many others. Among these corpora, CCMatrix is the largest parallel corpus obtained by mining unstructured web for parallel data,

[1] https://www.sid.ir/en/

a technique which is employed in the retrieval of unstructured web data. Majority of English-Persian parallel corpora in OPUS contain generic text; however, there are a few domain-specific parallel corpora available. Unfortunately, none of these corpora cover the scientific domain.

One of the first large parallel corpora of scientific papers was ASPEC (Asian Scientific Paper Excerpt Corpus) which consists of about 3.7 million sentence pairs in English, Japanese and Chinese (Nakazawa et al., 2016). In addition, SciELO is an English-Portuguese and Spanish corpus which is also available on OPUS and relies on the SciELO database of scientific articles (Névéol et al., 2018). This corpus is based on full article texts and contains 3.3 million aligned sentences. To the best of our knowledge, there are no public parallel corpora based on scientific publications for the English-Persian language pair.

3. Dataset Construction

In this section, we present our methodology for constructing Esposito, a collection of scientific publications' abstracts derived from the SID database. All abstracts' translations are provided by the publications' authors and are peer reviewed. We organize scientific journals to three main domains including *Human science*, *Medicine*, and *Science & engineering*. The workflow to create Esposito is illustrated in Figure 1 and each phase is described in detail below.

SID was established on August 16, 2013, by Academic Center for Education, Culture, and Research (ACECR)[2] in Iran, to advance and disseminate scientific knowledge. The SID's bank of scientific publications indexes the full text of articles in both Persian and English sections and creates a complete archive of publications from 2000 to present.

3.1. Document Retrieval and HTML Parsing

We develop a crawler for the SID website and obtained a list of scientific journals. From the list of journals, it is possible to retrieve a list of all volumes of a particular journal. The HTML page of the journal's list of volumes was further parsed to retrieve the page containing the list of articles in a given volume. Finally, the page of a particular paper was fetched as HTML. Web crawling is done using GNU Wget[3], while boilerplates, such as headers and footer, are removed using Python's BeautifulSoup[4] in order to keep the main content of each webpage in plain text format. Custom scripts are developed to extract text from web pages and create document pairs for each journal. We use SpaCy's multilingual sentenceRecognizer[5], a pre-trained pipeline component for sentence segmentation in various languages, for splitting documents into sentence levels. The reason we choose this package over others (other than its high quality) is the high inference speed due to the support of GPU.

Table 1 summarizes the number of journals and papers retrieved for each domain. Note that we only consider journals on SID which contain publications in both English and Persian both. After parsing HTML files and extracting the main content of each webpage, about 5% creates an empty text file in at least one of the languages. As a result, presented papers count only shows non-empty files.

3.2. Sentence Alignment

Sentence alignment is the task of taking parallel documents, which have been split into sentences, and finding high-quality matching translated sentences within the parallel documents. To do this, one can create all candidate sentence pairs from bilingual documents and then compute the semantic similarity between these sentence pairs using the information contained in each sentence. A good sentence alignment algorithm should be able to detect similar and dissimilar sentences and rank them according to their relevance. It should also be able to identify and discard noisy pairs of sentences.

Early sentence aligners (Brown et al., 1991; Gale and Church, 1993) use dynamic programming (Bellman, 1954) and work based on the intuition that the length of the translated sentence is likely to be similar to that of the source sentence. Recently, automatic sentence alignment methods using neural networks have gained popularity (Grégoire and Langlais, 2018; Artetxe and Schwenk, 2019a; Thompson and Koehn, 2019; Chousa et al., 2020). Such systems use a scoring function to calculate how two sentences translate each other in embedding space, and an alignment algorithm is used to generate an alignment.

In this paper, we use the Vecalign (Thompson and Koehn, 2019) algorithm which has linear complexity for time and space with respect to the number of sentences being aligned and only requires bilingual sentence embeddings. To choose a sentence embedding model for Vecalign, we manually align several documents in each domain as gold data. We experiment with different combinations of bilingual embeddings to determine their effect on the alignment accuracy. We evaluate the performance of the Vecalign algorithm with dif-

[2] http://acecr.ir/en

[3] https://www.gnu.org/software/wget/

[4] https://www.crummy.com/software/BeautifulSoup/

[5] https://spacy.io/api/sentencerecognizer

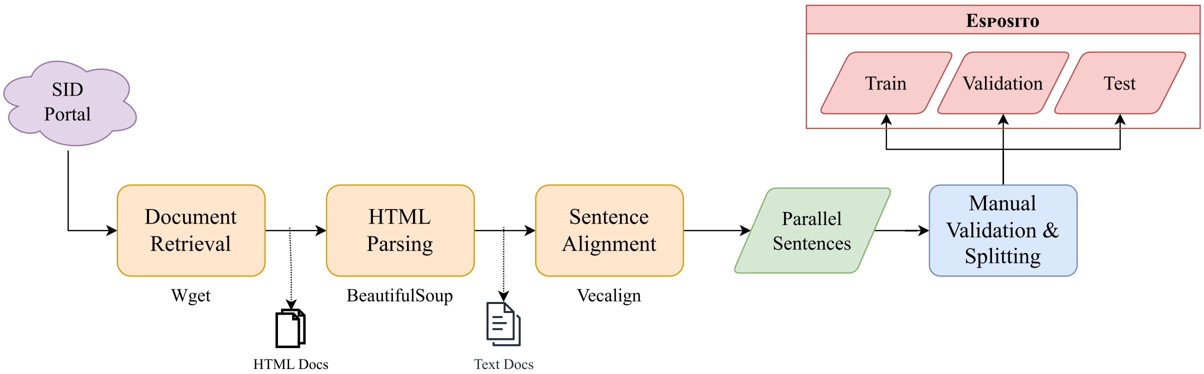

Figure 1: Workflow for the construction of Esposito.

Domain	Subject	Journals	Papers	Sentences
Human science	Human Science	775	386,994	1,261,809
	Art & Architecture	42	16,180	99,777
Medicine	Medical Science	139	101,855	975,366
	Veterinary Science	11	11,348	39,188
Science & engineering	Agriculture & Natural Resources	159	111,938	610,403
	Engineering & Technology	123	55,952	278,369
	Basic Science	81	47,499	232,188
		1330	731,766	3,497,100

Table 1: Document retrieval and HTML parsing report.

ferent sentence embeddings in term of F1 score. More concretely, we consider LASER2 (Artetxe and Schwenk, 2019b), LaBSE (Feng et al., 2022), and LASER3 (Heffernan et al., 2022) embedding models by employing their sentence embeddings in the Vecalign algorithm. According to Table 2, using Vecalign algorithm with LASER3 embeddings outperforms others in almost all domains. Consequently, we employ the Vecalign algorithm with LASER3 sentence embedding.

We provide statistics on parallel corpus including the number of bilingual sentence pairs. Our parallel corpus includes three main domains and seven subjects. The number of distinct sentences for each one is shown in Table 1.

3.3. Manual Validation and Splitting

In order to create high quality test set for Esposito as well as assess its quality, we randomly selected 1500 bilingual sentence pairs from our corpus to submit for evaluation via crowdsourcing. We recruited 45 undergraduate students majoring in computer science as annotators. The annotators are given a guideline which is summarized in Table 3. The semantic similarity of each sentence pair is assessed by three different annotators, and results are given as a value between 0 and 100. We assign the semantic similarity of sentence pairs to one of five quality levels, namely, “Very Good”, “Good”, “Needs Correction”, “Bad” and “Very Bad”. On average, each annotator efficiently annotates 100 sentence pairs, dedicating around 5 hours to thoroughly review the data. The annotation process is successfully completed within a reasonable duration of three weeks.

The results of the annotators’ quality assessment of sentence pairs are presented in Table 4. The evaluation demonstrates the high quality of our corpus in terms of human assessments across all three domains. More concretely, 82% of manually validated samples belong to “Very Good” or “Good” quality level and only 3% of samples belong to “Very Bad” quality level.

As a way to get a sense of how reliable the results of the annotators’ evaluations are, we use inter-annotator agreement scores, namely Fleiss’ kappa coefficient (Fleiss, 1971) and Spearman correlation, in order to analyze how reliable the results of the annotators’ evaluations are. According to the consensus evaluation results are presented in Table 5. Based on (Gwet, 2014) we have substantial agreement between annotators in all domains. This shows that evaluation results are reliable and can be used to draw conclusions.

We created a test set to address the issue of the English-Persian language pair lacking an official test set in the scientific domain. This test set

Domain	Subject	Docs	Vecalign		
			LaBSE	LASER2	LASER3
Human science	Human Science	8	0.86	0.88	**0.94**
	Art & Architecture	8	0.65	0.69	**0.83**
Medicine	Medical Science	7	**0.92**	0.84	0.89
	Veterinary Science	4	0.82	0.81	**0.89**
Science & engineering	Agriculture & Natural Resources	7	**0.88**	0.82	0.86
	Engineering & Technology	7	0.95	0.92	**0.98**
	Basic Science	5	**0.89**	0.86	0.85
		46	0.85	0.83	**0.89**

Table 2: Evaluation of different sentence embedding models employed in the Vecalign algorithm in term of F1 score.

Title	Scale	Description
Very Good	90-100	Two sentences are completely similar in meaning. Two sentences that refer to the same object or concept, using words that have semantic similarity or synonyms to describe them. The length of the two sentences is equivalent.
Good	70-89	Two sentences with similarities in meaning, referring to the same object or concept. The length of the two sentences may vary slightly.
Need Correction	50-69	Two sentences that are related in meaning, each referring to objects or concepts but they are related. The length of two sentences may vary slightly.
Bad	30-49	Two sentences that are different in meaning but have a slight semantic relation, may share the same topic. The length of two sentences can vary greatly.
Very Bad	0-29	The two sentences are completely different in meaning, their content is not related to each other. The length of two sentences can vary greatly.

Table 3: Annotation guidelines provided to annotators (Nguyen et al., 2022).

consists of 1200 sentences rated as “Very Good” or “Good” by annotators. We believe this test set will be useful for researchers working on English-Persian language pair. In addition, we use a random selection technique to create a validation set of 1000 sentences from each domain. The detailed statistics of Esposito is shown in Table 6.

4. Experiments

In this section, we evaluate the quality of our parallel corpus. Specifically, subsection 4.1 describes domain analysis experiments using an n-gram language model (LM). Subsection 4.2, presents evaluations using a neural network machine translation model.

4.1. Domain Analysis

In this section, we analyze differences between our parallel corpus and publicly available English-Persian corpora used for machine translation. Different datasets have different characteristics, and the domain of a parallel corpus can vary dramatically from one dataset to another. For instance, one dataset may contain more technical language, while another may contain more informal language. Understanding domains of various datasets can help improve machine translation performance. Nonetheless, defining the domain of a dataset is a challenging task to accomplish. Different strategies must be employed to determine the domain of a dataset. These can range from manual annotation to text analysis techniques. Here, we report the perplexity observed in the test set when an LM is trained using our training set.

We trained a separate LM of order five with KenLM (Heafield, 2011) on CCMatrix along with each Esposito domain to estimate the perplexity. We performed byte-pair encoding (BPE) (Sennrich et al., 2016b) on the test and train dataset to address the out-of-vocabulary issue.

Subject	Count	Very Bad	Bad	Needs Correction	Good	Very Good
Human Science	375	14	16	39	97	209
Art & Architecture	125	16	5	14	32	57
Medical Science	440	3	10	51	152	224
Veterinary Science	60	1	0	1	26	32
Agriculture & Natural Resources	125	2	6	9	42	66
Engineering & Technology	250	4	8	29	89	120
Basic Science	125	7	1	16	33	68
	1500	47	46	159	471	776

Table 4: Corpus manual validation results.

Domain	Kappa	Spearman
Human science	0.70	0.61
Medical	0.68	0.67
Science & engineering	0.64	0.64
	0.67	0.64

Table 5: Annotators consensus evaluation results.

Domain	Train	Validation	Test
Human science	1.36M	1000	400
Medical	1.01M	1000	400
Science & engineering	1.10M	1000	400
	3.49M	3000	1200

Table 6: ESPOSITO dataset statistics.

According to Figure 2, the average per-sentence perplexity decreases as training data increases. Due to the generic domain of both CCMatrix and FLORES-200 test set the decrease in perplexity is small when adding ESPOSITO. The results in Figure 2 show that adding ESPOSITO largely decreases perplexity across all domains. In other words, we can observe on average 69% and 78% reduction in perplexity for English and Persian, respectively.

4.2. Machine Translation Evaluation

Our goal is to build a high-quality parallel corpus for machine translation. To achieve this, we use neural machine translation systems for evaluations. We conduct experiments on ESPOSITO in order to determine the performance of bilingual systems which are trained using each domain of ESPOSITO separately in comparison with a baseline system. Additionally, we assess the the effectiveness of our parallel corpus when it is used as fine-tuning data in order to pre-train multilingual machine translation models. To accomplish this, we describe various experiment scenarios.

First, a baseline neural machine translation system was trained using CCMatrix parallel corpus, which is, to the best of our knowledge, the largest generic publicly available parallel corpus for English-Persian language pair. We compared generic systems against a model trained on CCMatrix and ESPOSITO. This allows the model to learn in-domain knowledge and also leverages the generic knowledge in the CCMatrix corpus. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU with 24GB of video memory.

Data Preprocessing. We used Moses' scripts (Koehn et al., 2007) for sentence tokenization in both languages. For each system, we trained a BPE with the vocabulary size of 20K using *subword BPE* (Sennrich et al., 2016b).

Systems and Training. Our models were trained using Fairseq (Ott et al., 2019)[6]. We used the Transformer architecture with an embedding size of 512, transformer hidden size of 1024, 4 attention heads, 4 transformer layers, dropout of 0.4, and attention dropout of 0.2. We trained with 0.2 label smoothing, 0.0001 weight decay, and Adam optimizer with a batch size of 4000 tokens with an update frequency of 4. Training was continued for 10 epochs and the best checkpoint was chosen based on validation perplexity.

Evaluation. Systems were evaluated using the BLEU score (Papineni et al., 2002) on the *devtest* of FLORES-200 (Goyal et al., 2022) and the test set for each domain of ESPOSITO.

In Table 7, we evaluated the performance of the systems which were trained with various combinations of the CCMatrix and different ESPOSITO domains. According to the findings, our system outperformed a generic system trained only on CCMatrix and improved on average 7.6 and 8.4 BLEU scores for En→Fa and Fa→En directions,

[6] `https://github.com/facebookresearch/fairseq`

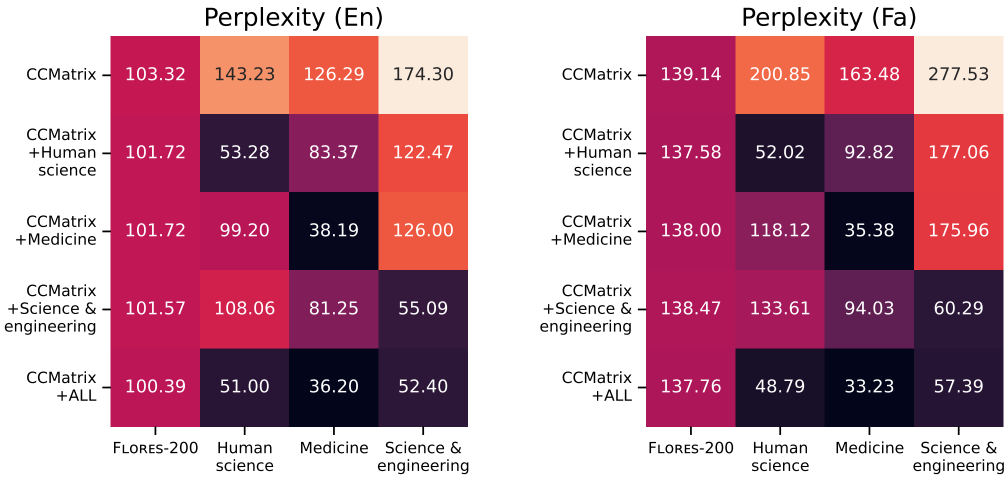

Figure 2: Perplexities of KenLM 5-gram language model trained on different domains and evaluated on FLORES-200 and ESPOSITO test sets for English (left) and Persian (right).

Test Domain	CCMatrix				
	Base	+Human science	+Medicine	+Science & engineering	+ALL
Human science	13.4 / 20.5	19.5 / 27.8	16.8 / 25.2	15.0 / 23.5	**20.7 / 28.2**
Medicine	15.7 / 20.8	19.1 / 24.9	22.3 / 29.4	16.9 / 22.5	**24.1 / 30.2**
Science & engineering	13.4 / 18.8	15.9 / 21.4	16.6 / 21.8	17.8 / 23.8	**20.6 / 26.8**
FLORES-200	**21.6** / 26.8	**21.6 / 27.0**	21.3 / 26.8	21.4 / 27.0	21.4 / 26.4

Table 7: En→Fa / Fa→En directions BLEU scores calculated for various neural machine translation systems trained on various combinations of CCMatrix and ESPOSITO domains.

respectively. As expected, all systems show no significant superiority for the FLORES-200 test set. This underlines the importance of data domain on the quality of machine translation.

Considering BLEU scores presented in Table 7 and perplexities reported in Figure 2, one can expect them to correlate negatively with each other. We verified this by calculating the Pearson's correlation coefficient and found that the correlation between BLEU scores of trained models for En→Fa direction and English LM perplexities is -0.701 and for Persian LM perplexities is -0.659. Similarly, BLEU scores of trained models for Fa→En direction and English LM perplexities is -0.807 and for Persian LM perplexities is -0.783.

Multilingual machine translation models. We further evaluated the performance of three multilingual machine translation models and Google Translation service on the test set: mBART (Liu et al., 2020), M2M100 (Fan et al., 2021), and NLLB-200 (3.3B) (Costa-jussà et al., 2022). mBART is a pre-trained, multilingual model designed specifically for machine translation tasks. It makes use of denoising objectives, which distort noisy input words before training the model to recreate the original ones. We used mBART50, which is available in *Hugging Face*[7] and supports English and Persian, to translate our test sets. Meta's M2M-100 model is capable of translating among every pair of 100 languages. Large monolingual datasets were mined using LASER (Artetxe and Schwenk, 2019b) to extract parallel sentences for M2M-100 training. We employed the pre-trained models offered by *Hugging Face*[8]. The most recent multilingual model released by Meta is called NLLB-200 which supports 200 languages. The pre-trained 3.3B parameter models provided by Fairseq is used for comparison.

Our model. We study the quality of ESPOSITO by fine-tuning DeltaLM (Ma et al., 2021), a pre-trained multilingual language model, which is among the best pre-trained models for language generation tasks such as translation and summarization. This model uses InfoXLM (Chi et al., 2021) weights as the initialization point and adopts the span corruption and translation span corruption as the pre-training task. DeltaLM takes advantage of both large-scale monolingual data and bilingual

[7] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[8] https://huggingface.co/facebook/m2m100_418M

Test Domain	Pretrained MNMT Models			Google Translate	DeltaLM (ours)	
	mBART50	M2M100	NLLB-200		CCMatrix	CCMatrix+ Esposito
Human science	10.9 / 16.6	12.2 / 19.7	12.3 / 20.3	20.5 / 29.8	17.1 / 27.0	**25.3 / 33.6**
Medicine	12.6 / 15.9	13.9 / 20.3	12.9 / 21.3	22.4 / 30.4	18.7 / 26.8	**28.5 / 36.0**
Science & engineering	11.2 / 14.5	12.3 / 18.2	11.5 / 19.7	21.7 / 28.0	16.4 / 24.1	**26.3 / 31.6**
Flores-200	14.7 / 27.0	19.9 / 28.2	18.2 / 31.7	**28.5 / 41.8**	25.4 / 36.4	24.8 / 36.8

Table 8: En→Fa / Fa→En directions BLEU scores calculated for various state-of-the-art multilingual machine translation models and the Google Translate service compared against DeltaLM model.

data. Experiments show that DeltaLM outperforms various strong baselines such as M2M and mBART on translation tasks. Microsoft released DeltaLM model in two different checkpoints, `base` and `large`. Here, we only report results of experiments on the large checkpoint.

In Table 8, we compared DeltaLM model fine-tuned using CCMatrix and Esposito datasets against multilingual machine translation models and the Google Translate system. As can be seen, DeltaLM model surpasses other multilingual machine translation models, even on the Flores-200 test set. This achievement is remarkable considering the large amount of data used for training pre-trained multilingual models. Moreover, the DeltaLM model fine-tuned using CCMatrix and Esposito outperformed Google Translate by an average of 5.1 and 4.3 BLEU scores for En→Fa and Fa→En directions, respectively. Google Translate only outperformed DeltaLM for the Flores-200 dataset. This observation leads us to speculate that Google Translate might have been trained on data resembling Flores-200, thereby contributing to its performance advantage.

5. Conclusion and Future Work

In this paper, we introduced Esposito, which is a parallel corpus containig 3.5 million sentence pairs in English-Persian language pairs. Additionally, we presented a manually validated domain-specific test set, which can be used as a baseline for future studies. We also demonstrated the usefulness of Esposito in the task of English-Persian language pair neural machine translation. Results showed that Esposito can be used to improve machine translation performance.

In the future, we plan to expand the language pair coverage of Esposito. Moreover, we want to expand our dataset using the back-translation (Sennrich et al., 2016a) technique, which leverages monolingual sentences to improve the quality of neural machine translation systems. Last but not least, we aim to improve the domain adaptation using techniques such as the one presented in (Mahdieh et al., 2020).

6. Ethics Statement

Our parallel corpus is derived from Open-Access (OA) journals indexes in SID. Open access literature is defined as “digital, online, free of charge, and free of most copyright and licensing restrictions.” The recommendations of the Budapest Open Access Declaration[9], including the use of liberal licensing (such as CC-BY), is widely recognized in the community as a means to make a work truly open access. Nevertheless, we should note that although texts of some scientific publications are copyrighted or do not allow derivative works, titles and abstracts by themselves constitute freely and publicly available metadata. Therefore, Esposito can be and will be made publicly available upon the acceptance of the paper.

7. Limitations

Due to the fact that our corpus only supports the English-Persian language pair, the applicability of our corpus to other language pairs is limited. This constraint is a result of the resources that are at our disposal as well as the concentration of our research on a particular language combination.

The process used to construct our parallel corpus is another drawback. Because the procedure is automated, there might be some instances of imprecise or erroneous translations in the corpus. Different things, such inconsistencies in alignment techniques, can cause these problems. Additionally, the quality of only a small portion of the corpus's sentences has been evaluated by annotators. The corpus' overall quality can be gauged from this sample, but it does not imply that all of the corpus's sentences will be equally accurate.

[9] https://creativecommons.org/about/program-areas/open-access/

8. Bibliographical References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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

Richard Bellman. 1954. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503 – 515.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, page 169–176, USA. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761. International Committee on Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1442–1453, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Mahdis Mahdieh, Mia Xu Chen, Yuan Cao, and Orhan Firat. 2020. Rapid domain adaptation for machine translation with monolingual data. *arXiv preprint arXiv:2010.12652*.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vinh Van Nguyen, Ha Nguyen, Huong Thanh Le, Thai Phuong Nguyen, Tan Van Bui, Luan Nghia Pham, Anh Tuan Phan, Cong Hoang-Minh Nguyen, Viet Hong Tran, and Anh Huu Tran. 2022. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5494–5502, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. TEP: Tehran english-persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 68–79. Springer Berlin Heidelberg.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348. Association for Computational Linguistics.

Jörg Tiedemann. 2012a. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Jörg Tiedemann. 2012b. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.