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

Mersad Esalati, Mohammad Javad Dousti, and Heshaam Faili
School of Electrical and Computer Engineering, College of Engineering, University of Tehran
Tehran, Iran
{mersad.esalati, mjdousti, and hfaili}@ut.ac.ir

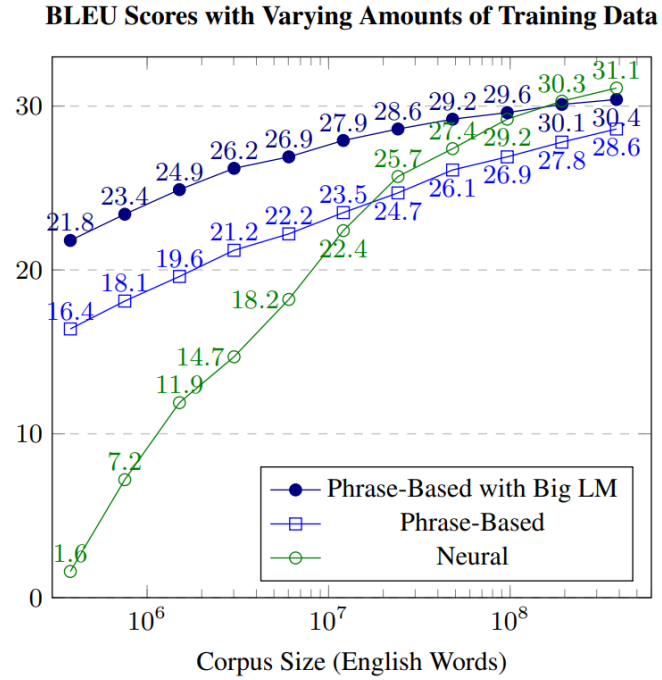
Table of Contents

Section One
Introduction

- Machine Translation
- Importance of Parallel Corpus

Machine Translation

- **Machine Translation Models**
  - Definition
  - Neural Machine Translation
- **Importance of Parallel Corpus**

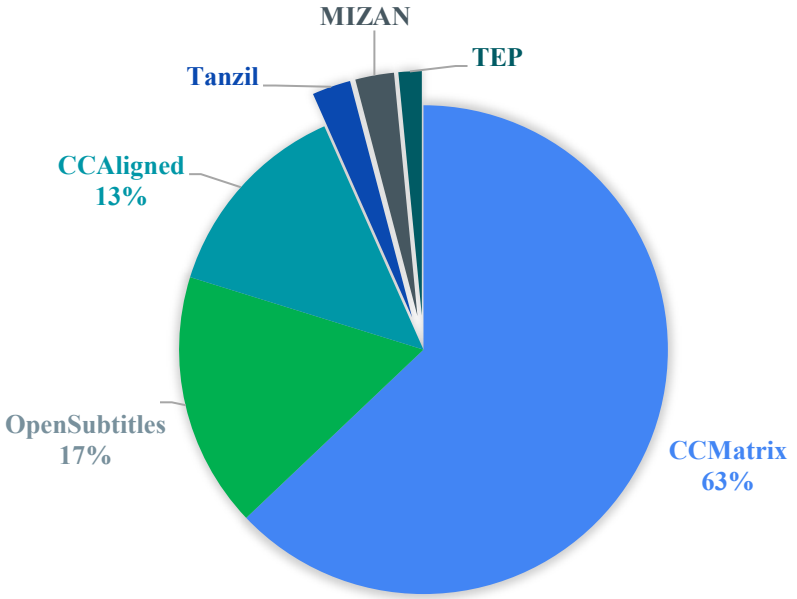


Koehn and Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the Workshop on Neural Machine Translation*.

Importance of Parallel Corpus

- **CCMatrix**
  - 24.6M Parallel Sentences
  - Very Noisy !!

- **Importance of Scientific Domain**



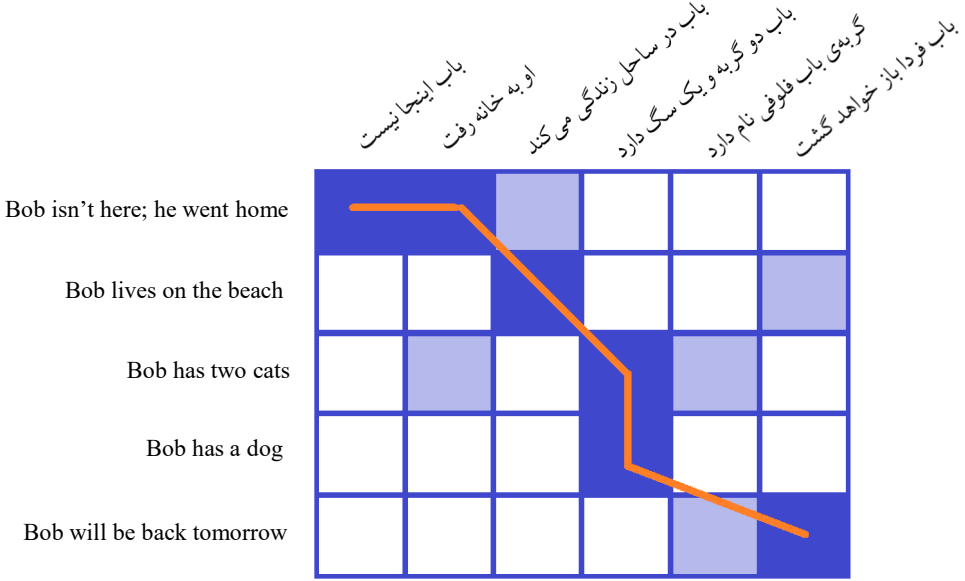
Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Section Two
Related Works

- Sentence Alignment
- Multilingual Sentence Embedding

Sentence Alignment

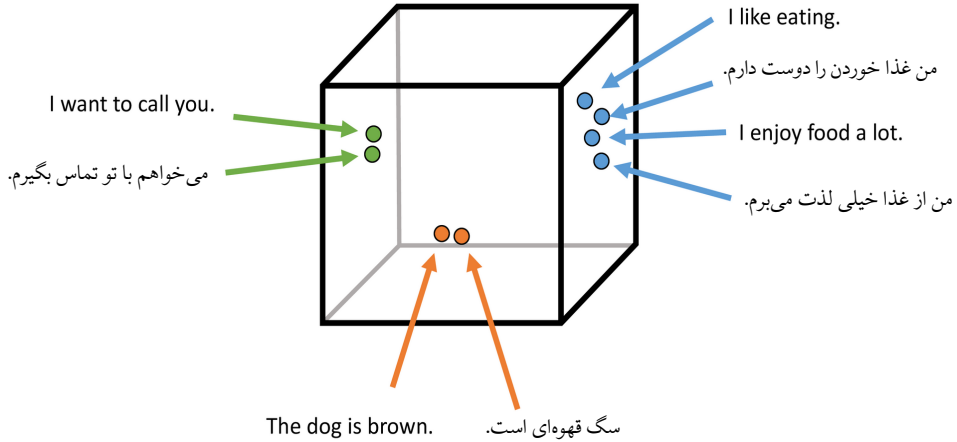
- **Problem Definition**
- **Model Structure**
  - Score Function
  - Alignment Algorithm



Neural Sentence Alignment

Vecalign *(Thompson and Koehn, 2019)*

- **Score Function**
  - Multilingual Sentence Embedding
  - Normalized Cosine Distance



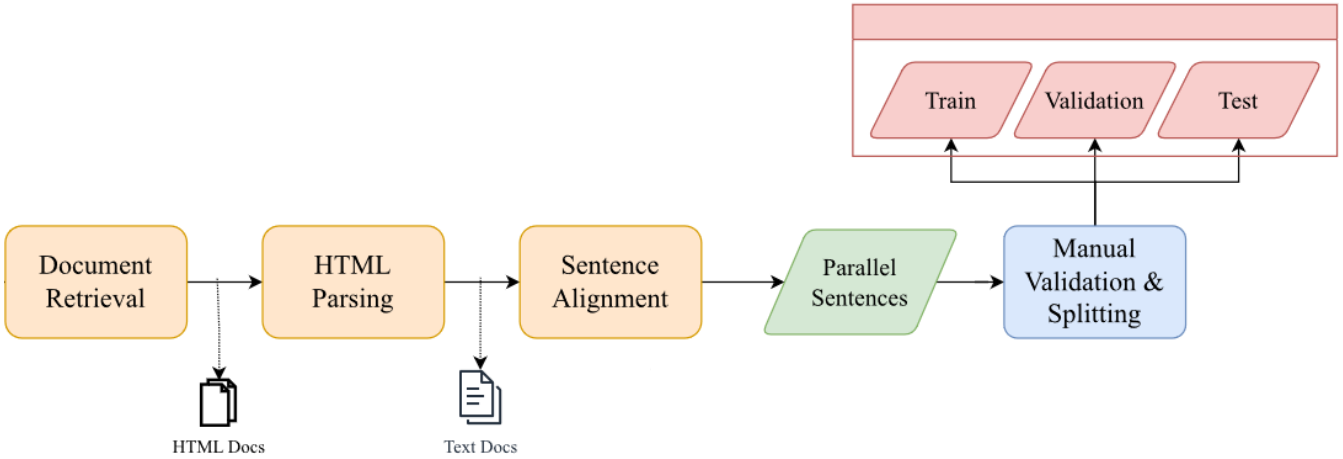
Thompson and Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*

Artetxe and Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*.

Section Three
Parallel Corpus Creation

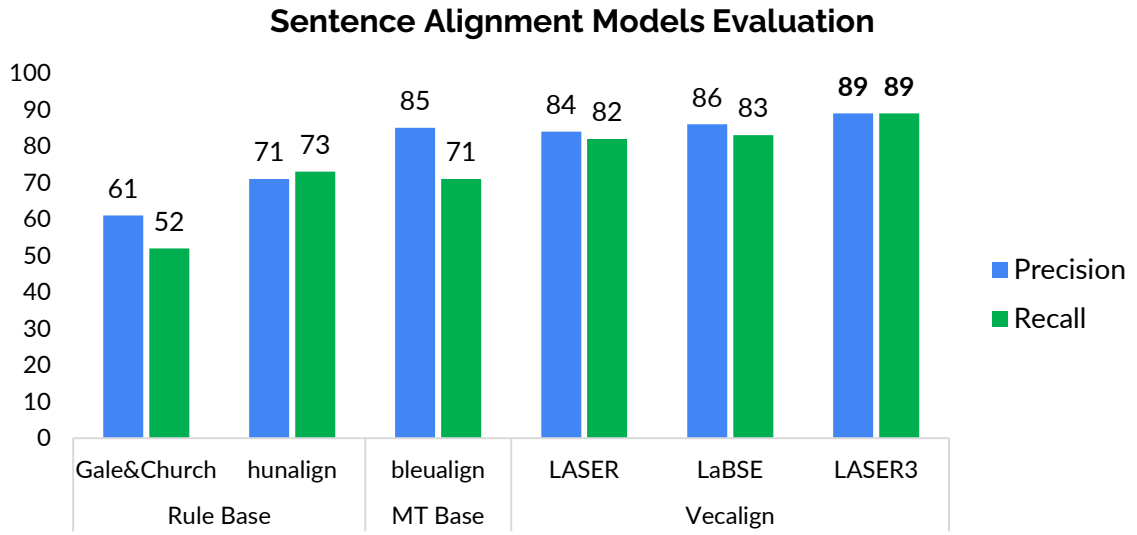
- Parallel Corpus Creation Workflow
- Parallel Corpus

Parallel Corpus Creation Workflow



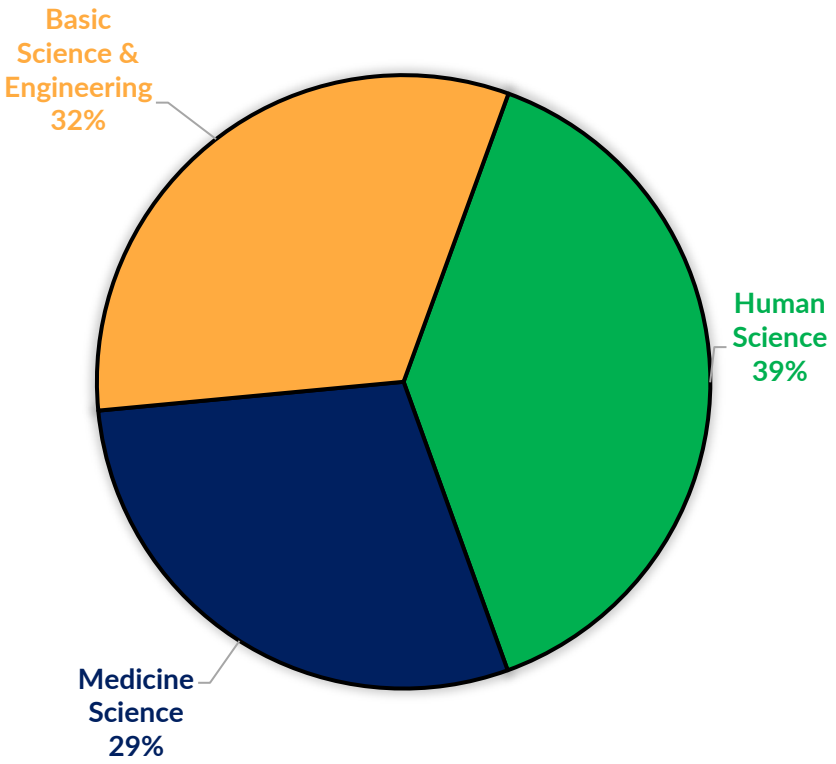
Sentence Alignment

- **Evaluation Benchmark**
  - 45 Manually Aligned English-Persian Parallel Documents



Document Retrieval & HTML Parsing

- **Scientific Information Database (SID)**
  - Open-Access
  - A Complete Archive of Scientific Journals from 2000
  - **731K** Parallel Documents
- **Sentence Alignment**
  - **3.5M** parallel sentences



Manual Validation

- **Validation Steps**
  - 1500 random bilingual sentence pairs
  - 45 undergraduate students
  - Guideline
- **Validation Results**
  - Annotators Consensus Evaluation

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

82%

3%

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.

Section Four
Experiments & Results

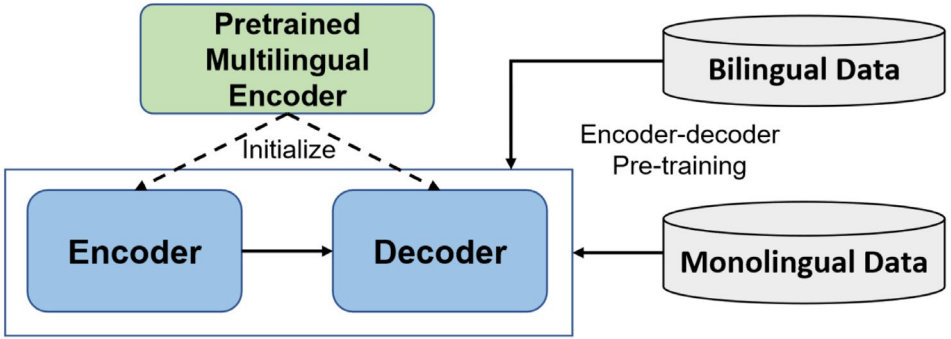
- Fine-tune Pretrained Multilingual LM

Multilingual NMT

- **Multilinguality as Transfer Learning**

- **Models**
  - mBART50 *(Liu et al., 2020)*
  - M2M-100 *(Fan et al., 2020)*
  - NLLB *(Costa-jussà et al, 2022)*

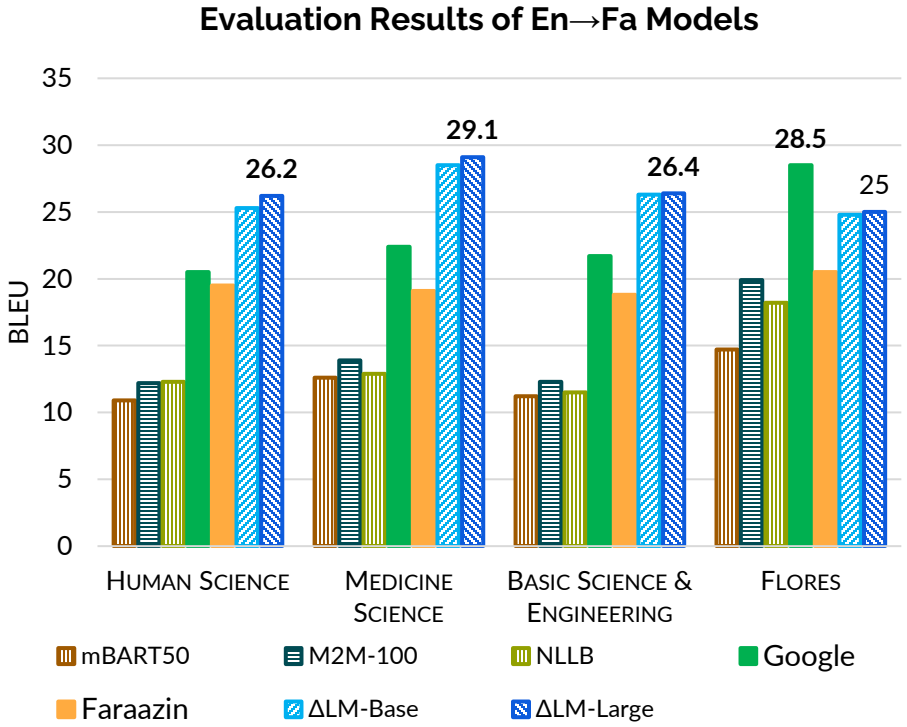
- **DeltaLM** *(Ma et al., 2021)*
  - Encoder-Decoder Architecture
  - Base: **360M** Parameters
  - Large: **830M** Parameters



Liu et al., 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*.
Fan et al., 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*.
Costa-jussà et al., 2022. No language left behind: Scaling human-centered machine translation. *arXiv*.
Ma et al., 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv*.

Fine-tune DeltaLM

- **Model Specs**
  - Base DeltaLM (ΔLM-Base)
  - Large DeltaLM (ΔLM-Large)
  - Epoch: 1
- **Training Data**
  - ESPOSITO+CCMatrix
- **Evaluation Results**



Conclusion

- Presenting the Workflow of Unsupervised Parallel Corpus Construction
- Create Standard Benchmark for Scientific Domain Machine Translation
- Future Works
  - Constructing a Multilingual Corpus for Scientific Texts
  - Utilizing Data Augmentation Techniques

Thank You