The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Azerbaijani

Mahsa Amani

Sharif University of Technology

mahsa.ama1391@gmail.com

Ph.D. Application Talk
February 20, 2024

Presentation Overview

1. Background
2. Introduction
3. Workflow
4. Datasets
5. Models
6. Results
7. References

Projects

- Eye Disease Recognition
- 3D Vision Transformers and Convolutional Models for Brain Tumor Segmentation
- Pipelines for low-resource Iranian Languages
- Linguistic Resources and Transformer-based Models for the Machine Translations between Luri and Yazdi Dialects versus Standard Persian
- **The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Azerbaijani**
- A Large Language Model for Persian
- Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection

Introduction

Azerbaijani Language:

1. Azerbaijani in Azerbaijan, written in Latin script
2. Azerbaijani in Iran, written in Perso-Arabic script

Iranian Azerbaijani:

- Spoken by more than 16% of the population in Iran
- Lack of computational resources
- Risk of extinction
- No prior work

We introduce: [Nouri et al., 2023]

1. A comprehensive linguistic resources
2. Important starter NLP models

Workflow

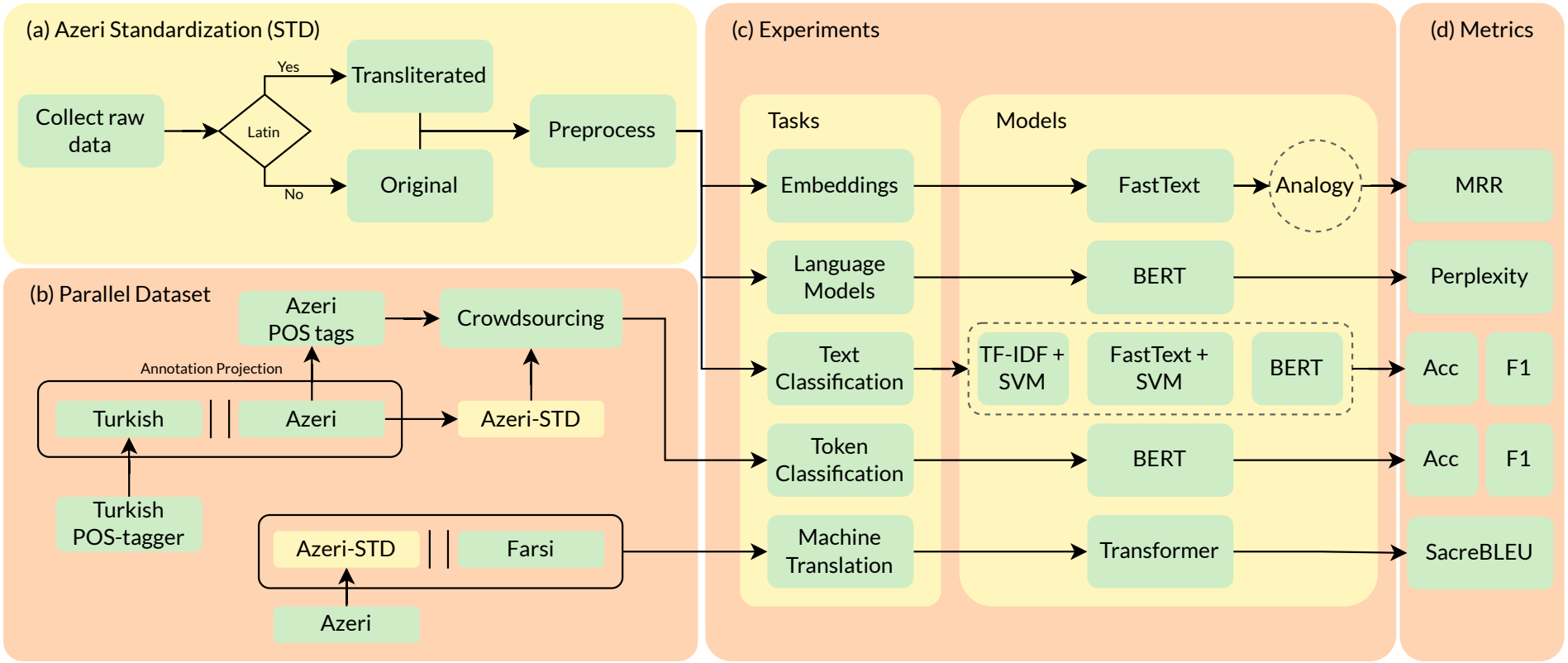


Figure: An overview of our pipeline for natural language processing of Iranian Azerbaijani, including data collection and preprocessing (block a), parallel corpus creation (block b), model development and fine-tuning (block c), and evaluation using various metrics (block d).

Datasets

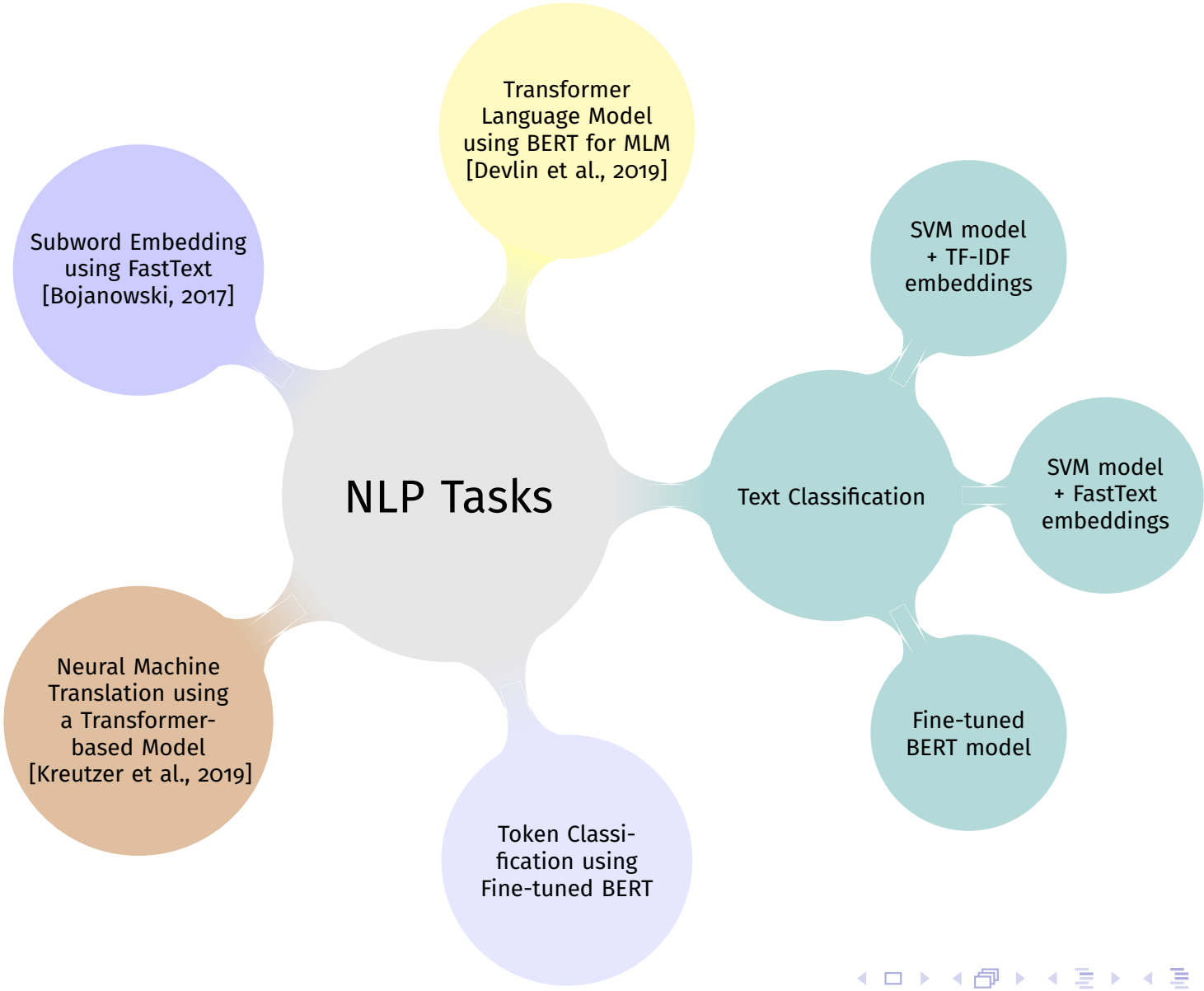
Raw text dataset	Transliterated text using a transformer-based solution [Zohrabi et al., 2023], Text originally in the Perso-Arabic script
Word analogy dataset	100 word analogies from four categories [Gladkova et al., 2016]
Text classification dataset	400 articles divided into 4 categories: `Literature`, `Sports`, `History`, and `Geography`
Token classification dataset	Annotation projection [Jalili et al., 2020, Alecakir et al., 2022], Transliteration, Crowdsourcing, 200 Sentences, 11 tags
Machine translation dataset	Parallel dataset between Persian and Iranian Azerbaijani

Data Collection

Name	Transliterated	#Sentences	#Words	#Avg. Words in Sent.
NewsCrawl	Yes	301403	210258	15.21
Books	Yes	116001	92891	6.08
Wikipedia	No	66449	88112	11.34
Ishiq	No	65321	146833	16.26
Bible (P)	Yes	42936	45693	13.36
News	Yes	19878	36875	15.68
DashQapisi	No	11071	27870	10.96
Quran (P)	Yes	8355	13176	11.3
Telegram	No	2263	10089	14.75
Varliq	No	816	5846	22.2
Stories (P)	No	676	2898	11.92
Others	Yes	699603	284642	5.98
Total	-	1323130	641861	9.55

Table: A summary of our collected datasets in Iranian Azerbaijani: (P) shows the parallel corpora.

Models


NLP Tasks
Transformer Language Model using BERT for MLM [Devlin et al., 2019]
Subword Embedding using FastText [Bojanowski, 2017]
Text Classification
SVM model + TF-IDF embeddings
SVM model + FastText embeddings
Fine-tuned BERT model
Neural Machine Translation using a Transformer-based Model [Kreutzer et al., 2019]
Token Classi-fication using Fine-tuned BERT

Results

Task	Model	Evaluation Metric	Performance
Language model-based Embedding	FastText	MRR	0.46
Language Model	BERT	Perplexity	48.05
Text Classification	TF-IDF + SVM	Accuracy	0.79
	TF-IDF + SVM	F1-score	0.78
	FastText + SVM	Accuracy	0.86
	FastText + SVM	F1-score	0.86
	BERT	Accuracy	0.89
	BERT	F1-score	0.89
Token Classification	BERT POS-tagger	Accuracy	0.86
	BERT POS-tagger	Macro F1-score	0.67
Machine Translation	Text Translation azb2fa	SacreBLEU	10.34
	Text Translation fa2azb	SacreBLEU	8.07

Table: **Summary of performance results for various NLP tasks on Iranian Azerbaijani language.** The models and evaluation metrics are detailed for each task (azb: Iranian Azerbaijani, fa: Persian).

Metrics I

Mean Reciprocal Rank (MRR)

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{1}$$

Perplexity

$$\text{PP(W)} = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}} \tag{2}$$

Accuracy

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}} \tag{3}$$

Metrics II

Macro F1-score

$$\text{MacroF1} = 2\frac{\text{Macro Precision} \cdot \text{MacroRecall}}{\text{MacroPrecision} + \text{MacroRecall}} \tag{4}$$

BLEU

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right), \text{where BP} = \begin{cases} 1 & c > r \\ e^{1-\frac{r}{c}} & c \leqslant r \end{cases} \tag{5}$$

References I

Marzia Nouri, Mahsa Amani, Reihaneh Zohrabi, and Ehsaneddin Asgari (2023)

The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Azerbaijani

Association for Computational Linguistics, 166 – 174.

Reihaneh Zohrabi, Mostafa Masumi, Omid Ghahroodi, Parham AbedAzad, Hamid Beigy, Mohammad Hossein Rohban, and Ehsaneddin Asgari (2023)

Borderless Azerbaijani Processing: Linguistic Resources and a Transformer-based Approach for Azerbaijani Transliteration

Association for Computational Linguistics, 175 – 183.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka (2016)

Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.

Association for Computational Linguistics, 8 – 15.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze (2020)

SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings

Association for Computational Linguistics, 1627 – 1643.

References II

Huseyin Alecakir, Necva Bölücü, and Burcu Can (2022)
TurkishDelightNLP: A Neural Turkish NLP Toolkit
Association for Computational Linguistics, 17 – 26.

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov (2017)
Enriching Word Vectors with Subword Information
Association for Computational Linguistics, 135 – 146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019)
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Association for Computational Linguistics, 4171 – 4186.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler (2019)
Joey NMT: A Minimalist NMT Toolkit for Novices
Association for Computational Linguistics, 109 – 114.

The End ⌛

Thank you for your attention

Comments? Questions?