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

Anonymous ACL submission

Abstract

Iranian Turkic is a dialect of the Azerbaijani language spoken by more than 16% of the population in Iran (>14 million). Unfortunately, a lack of computational resources is one of the factors that puts this language and its rich culture at risk of extinction. This work aims to create fundamental natural language processing (NLP) resources and pipelines for the processing and analysis of Iranian Turkic introducing standard datasets and starter models for various NLP tasks such as language modeling, text classification, part-of-speech (POS) tagging, and machine translation. The proposed resources have been curated and preprocessed to facilitate the development of NLP models for Iranian Turkic and provide a strong baseline for further research and development. This study is an example of bridging the gap in NLP for low-resource languages and promoting the advancement of language technologies in underrepresented languages. To the best of our knowledge, for the first time, this paper presents major infrastructures for the processing and analysis of Iranian Turkic, with the ultimate goal of improving communication and information access for millions of individuals.

1 Introduction

While a few of the world's languages are blessed with a wealth of linguistic resources, most of the world's 7,000 languages are considered low-resource and face the danger of extinction (Cieri et al., 2016). Each of these low-resource languages is crucial in preserving humanity's shared heritage, benefiting all. Developing techniques for analyzing these languages is currently a major challenge in the field of NLP, especially in different regions (Zoph et al., 2016; Duthoo and Mesnard, 2018; Bansal et al., 2021; Han et al., 2022). Despite significant advancements in deep learning for NLP in high-resource languages, some low-resource languages lack even sufficient digitized raw texts (ImaniGooghari et al., 2021).

Azerbaijani, spoken in Iran, which we refer to as Iranian Turkic in this paper, is a dialect of the Azerbaijani language spoken by a significant population in Iran written in Perso-Arabic script. This dialect, along with Azerbaijani spoken in Azerbaijan, which we denote as Azerbaijani Turkic, constitutes two distinct branches within the Azerbaijani language family. Azerbaijani Turkic with minor phonological, lexical, syntactic, and morphological variations uses the Latin script (Mokari and Werner, 2017; Rezaei et al., 2017). Despite the large number of speakers of Iranian Turkic, the digitized resources are very limited placing this language among low-resource languages and putting this language and its associated culture at risk of extinction (Kuriyozov et al., 2020; Park et al., 2021).

Related Work

The field of low-resource language research encompasses two main streams: (i) resource building through collaborative effort (e.g. Unimorph (McCarthy et al., 2020a)) and (ii) parallel projection from high resource languages (Agić et al., 2016; Eger et al., 2018; Subburathinam et al., 2019; Xia et al., 2021), particularly from the related languages (Hedderich et al., 2021). Iranian Turkic is a member of the Turkic language family, which also includes Turkish, Uzbek, Azerbaijani, Kazakh, and Uyghur (Mirzakhalov et al., 2021a).

Here we summarize the recent computational efforts on Turkic languages: **(i) High-resource Turkic NLP**: Turkish is a high-resource language among Turkic languages, with available datasets and models for various NLP tasks, such as stemming, segmentation, POS-tagging, parsing, and named entity recognition (Ehsani et al., 2012; Safaya et al., 2022). Almost the entire NLP pipeline for Turkish exists for Turkish in a toolkit, called TurkishDelightNLP (Alecakir et al., 2022). Text classification studies can also be observed for Turkish and Azerbaijani languages e.g., sentiment of social news articles in Azerbaijani (Mammadli et al., 2019), tweet topic classification (Yüksel et al., 2019) and sentiment analysis (Mutlu and Özgür, 2022) in Turkish. **(ii) Cross/multi-lingual models:** this track of research includes efforts on aligning monolingual embedding spaces of various Turkic languages, which are often affected by low-resource constraints (Kuriyozov et al., 2020). **(iii) Machine translation models:** machine translation have been developed for instances of Turkic languages (Gökırmak et al., 2019; Fatullayev et al., 2008)) as well as family-scale translations among Turkic languages (22 languages) (Mirzakhalov et al., 2021a,b). To the best of our knowledge, no prior work has developed a comprehensive NLP dataset or pipeline for Iranian Turkic, which is a language spoken by more than 14 million individuals in Iran and written in the Perso-Arabic script. In addition, the translation scenario of Iranian Turkic to Persian is significant in Iran as it can enhance communication among different generations and regions.

Contributions: our paper to the best of our knowledge, for the first time introduces: **(i)** comprehensive linguistic resources for Iranian Turkic including raw texts of various genres, a POS-tagged corpus, text classification collection, and parallel corpora (in both Turkish and Persian) as well as **(ii)** important starter NLP models for Iranian Turkic consisting of data cleanings, word embeddings, language modeling, post-tagging model, text classification models, and machine translation. Our primary focus has been to achieve a remarkable milestone by creating the first NLP pipeline and resource collection for a language spoken by at least 14 million people, while leveraging proven methodologies already established for other languages. In addition, through proposing the above-mentioned resources and models, we attempt **(iii)** to improve the language technology for the communication of millions of individuals and **(iv)** to contribute to preserving the Iranian Turkic and its rich culture.

2 Materials and Methods

Workflow: the overview of our approach for Iranian Turkic resource creation and model benchmarking is outlined in blocks of Figure 1: **(a) Azeri-standardization**: this part includes unifying the scripts of Azerbaijani Turkic and Iranian Turkic to the Perso-Arabic script and a comprehensive preprocessing spanning removal of URLs, digits, text within parentheses, elimination of non-Azerbaijani characters, and discarding sentences shorter than 10 characters. We refer to the resulting cleaned and standardized text as Azeri-STD. **(b) Parallel dataset creation**: we create two parallel corpora for two different reasons: *Parallel to Turkish:* we use a parallel corpus between Azerbaijani Turkic and Turkish (the most high-resource Turkic language) for the purpose of annotation projection (Eger et al., 2018) and run Azeri-STD to generate the parallel corpus for the Iranian Turkic, *Parallel to Persian:* we create this dataset for translation between Iranian Turkic and Persian again using our Azeri-STD on collected data from different sources. **(c) Training of the starter models**: we develop and fine-tune starter models of different NLP tasks, including word embeddings, language modeling, text and token classification, and translation. **(d) Model evaluations**: we evaluate each task using appropriate metrics and evaluation datasets.

2.1 Datasets

Raw text dataset: Our monolingual data comes from two primary sources: transliterated text using a transformer-based solution (Anonymous, 2023), and text originally written in the Perso-Arabic script. Table 2 provides information about our data (See Appendix A). The dataset includes $1.3M$ sentences spanning approximately $640K$ unique words.

Word analogy dataset: We propose a word analogy dataset for intrinsic evaluation of embedding spaces, inspired by previous literature such as (Gladkova et al., 2016). Our dataset includes 100 word analogies from four categories: inflectional morphology, derivational morphology, lexico-graphic, and encyclopedic semantics

Text classification dataset: For text classification, we use a collection of 400 articles from the Iranian Turkic Wikipedia, divided into 4 categories: Literature, Sports, History, and Geography (100 articles per category). This dataset provides a diverse set of texts for training and evaluating text classification models. We use 80% for training and dev and 20% for test purpose.

Token classification dataset: We create a token classification dataset based on the POS-tagging of our parallel Turkish corpus. We use annotation projection techniques to align (Jalili Sabet et al., 2020) the Turkish POS-tags (Alecakir et al., 2022) with

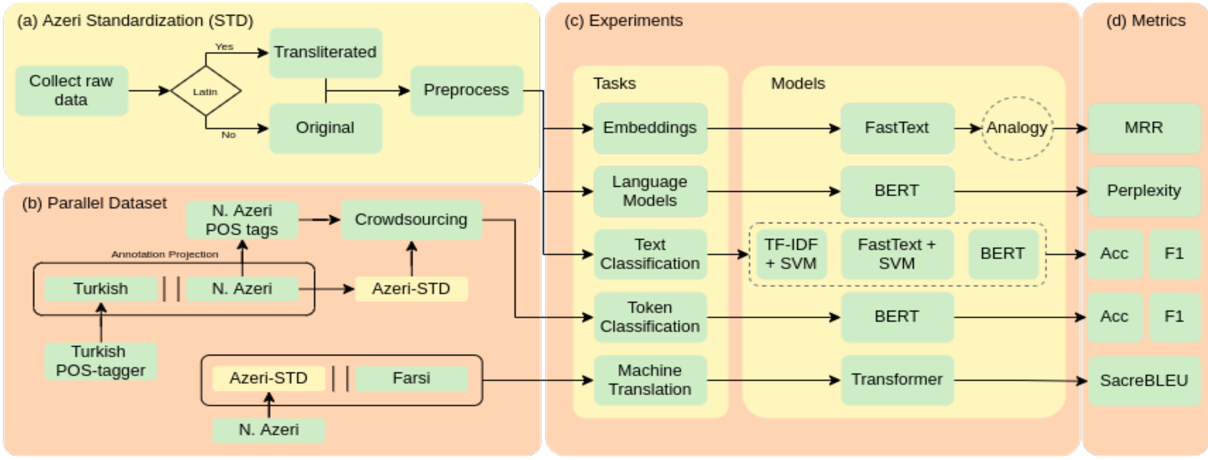

Figure 1: An overview of our pipeline for natural language processing of Iranian Turkic, including data collection and preprocessing (block a), parallel corpus creation (block b), model development and fine-tuning (block c), and evaluation using various metrics (block d).

those of Azerbaijani Turkic. To ensure script consistency across the different dialects of Azerbaijani, the results are then transliterated to the Iranian Turkic dialect. To improve the quality of the dataset, we leverage crowdsourcing to edit the tags. To summarize, we achieved a set of 200 tagged sentences. We use 90% for training and dev and 10% for test purpose. The agreement between the two annotators in the annotation task was evaluated using the kappa score, resulting in a value of 0.93, indicating substantial level of agreement. **Machine translation dataset:** we create a parallel dataset between Persian and Iranian Turkic languages. This dataset comprises a total of 14,972 aligned sentence pairs. It is composed of three main sources (marked with (p) in Table 2 in Appendix A): 7851 pairs from the Bible (Mayer and Cysouw, 2014), 6175 pairs from the Quran [1], and 946 pairs from a compilation of short stories we carefully extracted from different web forums manually. We use 90% for training and dev and 10% for test purpose.

The only available bilingual data for Iranian Turkic consists of the Quran, the Bible, and a few stories. Within the NLP community, religious texts are frequently employed as valuable resources for low-resource languages, primarily because of their inter-cultural nature, making them widely accessible across various languages (McCarthy et al., 2020b). The creation of high-quality aligned bibles in approximately 1000 languages has been a significant effort in this area (McCarthy et al., 2019).

To ensure data quality, our comprehensive preprocessing pipeline involved manual checks in some cases, successfully eliminating duplicates and noisy data from the dataset, resulting in a reduction in collection size from 2M to 1.3M sentences.

[1] https://tanzil.net/download/

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

2.2 Models

Subword embedding: A proper word representation is critical for almost all NLP tasks. Since Azerbaijani languages are agglutinative, we use fastText embeddings that can properly use the subword information in the skip-gram architecture (Bojanowski et al., 2017). We evaluate this embedding extrinsically in the text classification task and intrinsically by measuring the Mean Reciprocal Rank (MRR) in the word analogy inference task.

Transformer language model: Transformer-based language-model embeddings proved to be state-of-the-art approaches on a variety of NLP tasks benefiting from proper modeling of contextual information of tokens (Devlin et al., 2019). Therefore, we train a BERT language model with a masked language modeling objective on our standardized raw text. We evaluate this model by measuring perplexity of the language model (Chen et al., 1998).

Text Classification: We include a text classification use case in our pipeline for Iranian Turkic comparing three types of approaches: (i) an SVM model using TF-IDF embeddings, (ii) an SVM model using average fastText embeddings of a document, and (iii) supervised fine-tuning of our BERT model (Devlin et al., 2019). We evaluate the classification part by measuring accuracy and the F1 score on the test set.

Token Classification: For the example of token classification we use our POS-tagging dataset, that can benefit a range of NLP tasks. We fine-tune our BERT embedding model for the POS tagging. Since we have 11 categories, other than accuracy we evaluate the tagging on macro-F1 score as well.

Machine Translation: We train a low-resource transformer-based machine translation model between Iranian Turkic and Persian. The model's computational efficiency makes it practical for use in situations where resources are limited (Kreutzer et al., 2019). We evaluate the quality of translation using the SacreBLEU (Post, 2018) on the test set.

3 Results

The objective of this research was to establish fundamental pipelines and resources for the Iranian Turkic language. The obtained results are summarized in Table 1: **Embedding intrinsic evaluation:** Our fastText model obtained an MRR of 0.46 in word analogy intrinsic evaluation indicating that the model can guess the analogies on average in the second guess. **Language modeling perplexity:** We evaluated the model perplexity of our BERT language model, and achieved a perplexity score of 48.05. Given the constraints of a low-resource language, achieving a perplexity of 48.05 is quite commendable and suggests that despite the scarcity of training data, our model was able to produce relatively accurate predictions. **Text classification:** our fine-tuned BERT models performed better than the other two models on the text classification task. After the BERT model, the fastText-based baseline showed superior performance in comparison with the TF-IDF baseline (an extrinsic evaluation of the fastText embedding). We conducted a text classification comparison to showcase the impact of transliteration data for Iranian Turkic in BERT masked language model pretraining. Our BERT model, trained on both transliterated and original Iranian Turkic data, achieved an impressive macro-F1 of 0.89 in supervised text categorization. In contrast, the BERT model trained solely on Iranian Turkic data attained a significantly lower macro-F1 of 0.48. Moreover, training the model on transliterated data resulted in a mBert score of 0.85 macro-F1, further confirming the efficacy of utilizing transliterated data in transformer language models for downstream tasks. **Token classification:** The transformer-based tagger achieved a satisfactory performance with an accuracy of 0.86 and an F1-score of 0.67. This performance indicates that the fine-tuned BERT tagger is able to identify and classify language elements in the dataset with a moderate degree of accuracy and completeness. **Machine translation:** We assessed the model's performance using the SacreBLEU metric and obtained scores of 10.34 for Iranian Turkic to Persian translation and 8.07 for Persian to Iranian Turkic translation. Although these scores may not reach the level of high-resource settings, when compared to other low-resource languages and their respective scores, our model achieved a reasonable performance for a low-resource machine translation setting.(Mirzakhalov et al., 2021a).

4 Conclusions

In this paper, to the best of our knowledge, for the first time, we introduced computational resources and pipelines for Iranian Turkic language processing. Language technologies developed for this language can significantly contribute to the communications of >14M speakers of this endangered language. We introduced data sources and models on major NLP tasks including text cleaning, word embeddings, language modeling, text and token classifications, and machine translation. Our introduced embedding space, pos-tagger, and BERT language modeling can be used in a variety of other NLP tasks. Our translation model is the first technological effort toward closing the gap between generations that are not acquiring their grandparents' language. Our pipeline and prepared resources can play a key role in addressing the scarcity of computational resources for Iranian Turkic and preserving the language and its culture. We make the resource and models available in the HuggingFace library for the use of the public.

5 Limitations

Our study has several limitations that must be acknowledged. A major limitation is the limited resources available for Iranian Turkic, which resulted in a scarcity of data for our pipeline. This scarcity poses a significant challenge for training and evaluating our models and may impede their overall performance. Additionally, Azerbaijani is an agglutinative language, with postfixes added to words to indicate grammatical relationships and functions. However, the way postfixes are written and separated from words varies between Azerbaijani Turkic and Iranian Turkic, In Iranian Turkic, there are no clear rules for written language, leading to variations in the use of spaces and half-spaces between words and postfixes. The absence of standard and pre-defined rules also results in considerable noise in the data, making accurate analysis and understanding of the language difficult. We faced challenges in accurately tokenizing Azerbaijani because of these variations and decided to use spaces to tokenize words in our data, but this method sometimes resulted in incorrect segmentation. Furthermore, we used a significant portion of transliterated data from resources in Azerbaijani, which may be affected by phonological, lexical, syntactic, and morphological differences between the two dialects, and thus may impact the performance of our pipeline and limit the accuracy of our models.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Huseyin Alecakir, Necva Bölücü, and Burcu Can. 2022. TurkishDelightNLP: A neural Turkish NLP toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 17–26, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Anonymous. 2023. Borderless azerbaijani processing: Linguistic resources and a transformer-based approach for azerbaijani transliteration. Anonymous preprint under review.

Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Émilie Pagé-Perron, and Jacob Dahl. 2021. How low is too low? a computational perspective on extremely low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 44–59, Online. 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.

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

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.

Elie Duthoo and Olivier Mesnard. 2018. CEA LIST: Processing low-resource languages for CoNLL 2018. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 34–44, Brussels, Belgium. Association for Computational Linguistics.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2018. PD3: Better low-resource cross-lingual transfer by combining direct transfer and annotation projection. In *Proceedings of the 5th Workshop on Argument Mining*, pages 131–143, Brussels, Belgium. Association for Computational Linguistics.

Razieh Ehsani, Muzaffer Ege Alper, Gülşen Eryiğit, and Eşref Adali. 2012. Disambiguating main POS tags for Turkish. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*, pages 202–213, Chung-Li, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Rauf Fatullayev, Ali Abbasov, and Abulfat Fatullayev. 2008. Dilmanc is the 1st mt system for azerbaijani. *Proc. of SLTC-08, Stockholm, Sweden*, pages 63–64.

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. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San

Diego, California. Association for Computational Linguistics.

Memduh Gökırmak, Francis Tyers, and Jonathan Washington. 2019. A free/open-source rule-based machine translation system for crimean tatar to turkish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 24–31, Dublin, Ireland. European Association for Machine Translation.

Xu Han, Yuqi Luo, Weize Chen, Zhiyuan Liu, Maosong Sun, Zhou Botong, Hao Fei, and Suncong Zheng. 2022. Cross-lingual contrastive learning for fine-grained entity typing for low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2241–2250, Dublin, Ireland. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Masoud Jalili Sabet, Philipp Dufter, Michael Cysou, and Hinrich Schütze. 2021. ParCourE: A parallel corpus explorer for a massively multilingual corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 63–72, Online. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Elmurod Kuriyozov, Yerai Doval, and Carlos Gómez-Rodríguez. 2020. Cross-lingual word embeddings for Turkic languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4054–4062, Marseille, France. European Language Resources Association.

Sevda Mammadli, Shamsaddin Huseynov, Huseyn Alkaramov, Ulviyya Jafarli, Umid Suleymanov, and Samir Rustamov. 2019. Sentiment polarity detection in Azerbaijani social news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 703–710, Varna, Bulgaria. INCOMA Ltd.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020a. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Winston Wu, Aaron Mueller, William Watson, and David Yarowsky. 2019. Modeling color terminology across thousands of languages. 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 2241–2250, Hong Kong, China. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021a. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Bekhzodbek Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, Francis Tyers, Orhan Firat, John Licato, and Sriram Chellappan. 2021b. Evaluating multiway multilingual NMT in the Turkic languages. In *Proceedings of the Sixth Conference* *on Machine Translation*, pages 518–530, Online. Association for Computational Linguistics.

Payam Ghaffarvand Mokari and Stefan Werner. 2017. Azerbaijani. *Journal of the International Phonetic Association*, 47(2):207–212.

Mustafa Melih Mutlu and Arzucan Özgür. 2022. A dataset and BERT-based models for targeted sentiment analysis on Turkish texts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 467–472, Dublin, Ireland. Association for Computational Linguistics.

Cheonbok Park, Yunwon Tae, TaeHee Kim, Soyoung Yang, Mohammad Azam Khan, Lucy Park, and Jaegul Choo. 2021. Unsupervised neural machine translation for low-resource domains via meta-learning. 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 2888–2901, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Saeed Rezaei, Ashkan Latifi, and Arash Nematzadeh. 2017. Attitude towards azeri language in iran: a large-scale survey research. *Journal of Multilingual and Multicultural Development*, 38(10):931–941.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. 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 313–325, Hong Kong, China. Association for Computational Linguistics.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Atıf Emre Yüksel, Yaşar Alim Türkmen, Arzucan Özgür, and Berna Altınel. 2019. Turkish tweet classification with transformer encoder. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1380–1387, Varna, Bulgaria. INCOMA Ltd.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

A Data Collection

Our resources come from two sources: transliterated data from Azerbaijani Turkic and original data in Iranian Turkic, We collected the original Iranian Turkic data through various methods, including parsing Wikipedia dumps[2], gathering data from İshiq website[3], crawling Dashqapisi website archive[4], importing telegram channels[5], requesting content from Varliq quarterly journal[6], and manually extracting parallel sentences from translated short stories. For the Azerbaijani Turkic data, we collected news articles[7, 8], books[9], Quran[10] and Bible parallel corpora, and other textual resources from various sources including Github repositories[11]. Table 2 provides information about our data, including dataset name, transliteration status, number of sentences, unique words, and average words per sentence.

B Azerbaijani Turkic vs. Iranian Turkic

Azerbaijani, spoken in the Republic of Azerbaijan, commonly referred to as Azerbaijani Turkic, and Azerbaijani spoken in Iran, often denoted as Iranian Turkic, are recognized as two distinct branches within the Azerbaijani language family. The usage patterns differ between the two branches, as

[2] https://azb.wikipedia.org/
[3] https://ishiq.net/
[4] https://dashqapi.blogsky.com/
[5] https://t.me/abcmedrese
[6] http://varliq.ir/
[7] https://wortschatz.uni-leipzig.de/en/download/Azerbaijani
[8] https://wortschatz.uni-leipzig.de/en/download/Azerbaijani
[9] https://github.com/raminrahimzada/az-corpus-nlp/blob/master/sentences/books_starting_with_a.txt
[10] https://tanzil.net/
[11] https://github.com/raminrahimzada/az-corpus-nlp/blob/master/sentences/others.zip

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
New	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 2: A summary of our collected datasets in Iranian Turkic : (P) shows the parallel corpora.

Iranian Turkic is primarily used as a spoken language, whereas Azerbaijani Turkic serves as an official, scientific, and literary language. Notably, the alphabets used by these branches exhibit dissimilarities. Azerbaijani Turkic has experienced multiple changes since 1928, whereas the Iranian branch continues to employ the Perso-Arabic alphabet. Vocabulary-wise, Azerbaijani in Iran incorporates loanwords from Persian, Arabic, and English, whereas the Azerbaijani Turkic branch includes loanwords from Russian, Arabic, Persian, and English. Furthermore, grammatical disparities exist between the two branches. The Iranian branch is primarily influenced by Persian in Iran, while the Azerbaijani Turkic branch draws influence from Russian in Azerbaijan. In summary, Azerbaijani Turkic and Iranian Turkic are two distinct branches of the Azerbaijani language, differing in their usage patterns, alphabets, vocabulary, and grammatical features. These variations reflect the influence of Persian, Arabic, Russian, and English on the respective branches in their respective regions.

C POS Guideline

Introduction: This guideline provides instructions for annotating Part-of-Speech (POS) tags in the Azerbaijani language. The POS tags help identify the grammatical category of each word in a sentence. We have developed a comprehensive guideline featuring 11 tag categories. The tag categories include Noun, Punctuation, Verb, Pronoun, Adverb, Conjunction, Number, Adjective, Postposition, Interjection, and Determiner. Examples for each category have been provided to assist in the annotation process.

Instructions: Each word should be tagged with one and only one POS tag from the provided categories. The function and the grammatical properties of the word while assigning the POS tag should be considered.

POS Tag Categories: a. Noun: Tags for common and proper nouns, including names of people, places, objects, etc. Example: "کیتاب" (book), "تهران" (Tehran).

b. Punctuation: Tags for punctuation marks. Example: ".", ",", "?".

c. Verb: Tags for verbs. Example: "یازدیم" (I wrote), "گئدیرم" (I am going).

d. Pronoun: Tags for words that replace nouns. Example: "من" (I), "سنین" (yours).

e. Adverb: Tags for words that modify verbs, adjectives, or other adverbs. Example: "یاواشجا" (quietly), "همیشه" (always).

f. Conjunction: Tags for words that connect words, phrases, or clauses. Example: "و" (and), "کي" (that).

g. Number: Tags for numeric values. Example: "ایکي" (two), "۱۰۰" (hundred).

h. Adjective: Tags for words that describe or modify nouns. Example: "گؤزل" (beautiful), "یاخشي" (good).

i. Postposition: Tags for words that come after nouns and show relationships. Example: "کیمي" (like), "اۆچۆن" (for).

j. Interjection: Tags for words that express strong emotions or surprise. Example: "آي!" (oh!), "آه!" (ah!).

k. Determiner: Tags for words that introduce or specify nouns. Example: "بو" (this), "هئچ" (any).

D Hyperparameters

The **BERT language model** was trained with hyperparameters set as follows: for pre-training, the number of epochs was 10, the batch size was 128, the learning rate was 5e-5, the vocabulary size was 10,000, and the maximum size of position embeddings was set to 64. For **text classification** tasks, the maximum sequence length was set to 64, the batch size was 32, and the number of epochs was 10. The learning rate for text classification was set to 275e-7. For **token classification** tasks, the maximum sequence length was set to 64, the learning rate was set to 2e-5, the batch size was 64, and the number of training epochs was 20. In the case of **machine translation**, the early stopping metric used was "loss," and the model was trained for 500 epochs. The embedding dimension was set to 256, the vocabulary limit was 2000, the batch size was 512, the number of layers was 2, and the maximum output length was 100.