C2FDataset: A Comprehensive Parallel Dataset for Transitions between Persian Colloquial and Formal Variations

Anonymous ACL submission

Abstract

Natural language processing models have shown promising results in analyzing Persian text. However, their performance drops significantly when applied to colloquial Persian sentences, resulting in inaccurate predictions. This challenge arises from the substantial differences between colloquial and formal Persian and the lack of parallel data for colloquial to formal Persian translation. In addressing this gap, our research is dedicated to the development of the C2FDataset, a large-scale colloquial to formal Persian parallel dataset. Our proposed dataset is a critical resource for training models that can effectively bridge the linguistic variations between colloquial and formal Persian. To illustrate the utility of our dataset, we trained a GPT-2 model on it. The model exhibited remarkable proficiency in colloquial to formal text style transfer, outperforming both ChatGPT and a leading rule-based system in this task. This conclusion is supported by our extensive human evaluation, conducted using a ranking-based scoring system we designed for this purpose. The results underscore the significance of the C2FDataset in enhancing the performance of natural language processing models in the challenging domain of colloquial to formal Persian conversion.

1 Introduction

Persian is among the most spoken and widely used languages and the first ten languages used in Internet content. The Persian language has two main forms: formal (standard or written style) and colloquial (informal or conversational form).

The high daily internet and social media usage has led to increased colloquial textual content generated. These two forms differ in vocabulary, grammar, and style. Informal Persian is spoken in everyday conversations and reflects the dynamic form of the language that adapts to social and cultural adjustments. Most of this content is colloquial, meaning it does not follow standard grammar and structure. The coexistence of these two types of Persian languages in written form proposes a challenge in computational linguistics. Understanding these styles requires linguistic rules and vocabulary in both formal and informal aspects. This paper presents an approach to solving this challenge by providing a large-scale colloquial to formal dataset and further training deep learning models to bridge the gap between these two domains.

2 Related Work

Despite the availability of numerous colloquial language resources in Persian, there is a shortage of parallel datasets for both formal and colloquial language. Previous works have only covered a portion of colloquial language without providing a parallel corpus. In general, approaches can be divided into rule-based systems and learning methods.

Naeimi (Naemi et al., 2021) presents a rule-based approach that involves converting colloquial words into their formal counterparts. The approach involves extracting rules that convert colloquial words into formal ones by analyzing the changing processes in Persian words. Khojasteh (Khojasteh et al., 2020) proposes a large-scale colloquial Persian (LSCP) dataset that involves 120M colloquial sentences from Persian Twitter that translate to five different languages. These sentences resulted from 27M tweets annotated with parsing trees, part-of-speech tags, sentiment polarity, and translation. (Kabiri et al., 2022) presents an Informal Persian Universal Dependency Treebank, showing differences between informal and formal Persian language from phonological, morphological, and syntactic views.

Indeed, (Kabiri et al., 2022) and (Khojasteh et al., 2020) concentrate on the colloquial style of the Persian language and do not provide parallel corpora.

Also, (Faili et al., 2016) presents a spell-checker

that detects and corrects types of errors.

3 Persian Linguistics Background

The Persian language is primarily spoken in Iran, Afghanistan, Tajikistan, and other countries in the Middle East and Central Asia, with more than 100 million native speakers. The Persian language is complex and poses difficulties for natural language processing and understanding, such as: (Reference)

- Homomorphic letters exhibit distinct written forms based on their positional placement within a word. (Naemi et al., 2021) An exemplar of this linguistic phenomenon is the letter "غ" (gh) according to its place we can write it in four forms, "ـغ" , "ـغـ" , "غـ", and "غ".
- Homophone letters are characterized by having the same or similar pronunciations but distinct written forms. (Naemi et al., 2021) Examples include "ذ", "ض" , "ز", and "ظ". all sound like z. Another example is "ث", "ص", and "س" which sounds like s.
- Polysemy is the phenomenon of a single word having different meanings. For example, the word "شیر" (shir) can refer to both "lion", and "milk", as well as "tap", highlighting the linguistic intricacies and contextual dependence in the Persian language. (Naemi et al., 2021)
- Vowels include six types in Persian, encompassing three long vowels: "آ" (ā), "ای" (ī), "او" (ū); and three short ones: "اَ" (a), "اِ" (e), "اُ" (o), with the short vowels typically not written in the script which can result in different interpretations of the word. For instance, the word "ترک" can take on different meanings based on its vocalization: "تَرک" (Tark) means "leaving", "تَرَک" (Tarak) refers to a "crack", and "تُرک" (Turk) signifies the Turkic people. These variations underscore the importance of context in interpreting Persian texts. (Esmaeelpour et al., 2022) (Kabiri et al., 2022)
- English words written with Persian letters, e.g., "کامنت" (comment).
- Some verbs can have varying meanings depending on their position within a sentence. It is important to take note of their context to understand their intended meaning. (Naemi et al., 2021) For example, the verb "خوردن", which according to the meaning of the sentence, can either mean "to eat" or "to collide".
- Disordered sentences in Persian involve altered word order for stylistic purposes, emphasizing, conveying tone, or accommodating poetry, highlighting the language's flexibility and richness.

Colloquial Persian is known for shorter sentence lengths, recurrent use of idioms, and less compliance with the "subject-object-verb" structure; many formal expressions and even the dictation of some phrases change into their colloquial form. The colloquial style is suitable for common communication, such as casual conversations and social media. In contrast, formal Persian is characterized by rigid grammar rules, formal vocabulary, and a more structured and literary style, and most of the time follows the "subject-object-verb" structure. (Tajali et al., 2023) It is commonly used in written documents, academic publications, legal texts, formal speeches, etc. Studies show colloquial and formal language change through deletion, substitution, insertion, and interchange (Martins and Silva, 2004).

- Deletion: deletion of one or more letters, an example is "کتاب ها" (books), which changes into "کتابا".
- Substitution: change of a letter into another. For example, "باران" (rain) becomes "بارون".
- Insertion: addition of one or more letters. As an example, "آن" (that) changes into "اون".
- Interchange: the reverse ordering of two adjacent letters of a word. For example, "قفل" (lock) becomes "قلف".

Understanding linguistic complexities is crucial for successfully developing an automatic system to convert colloquial Farsi to formal Farsi. This paper presents a comprehensive technique leveraging linguistic foundations to bridge the gap between colloquial and formal Persian expressions, emphasizing practical applications.

4 Building Parallel Corpus

This section introduces the large-scale colloquial to formal Persian dataset (C2FPersian), consisting of manual and machine-generated corpora.

4.1 Manual Dataset

The manual dataset consists of 12K colloquial and formal Persian pairs. In the development process, we sampled 3K lines from the DegarBayan subtitle dataset (Group, 2022) and crawled 9K entries from Persian content on e-commerce platforms and social media. To generate the corresponding formal sentences, Native Persian speakers annotated both corpora.

4.2 Large-scale Machine Generated Dataset

To create the large-scale dataset, we used Persian movie subtitles from Opensubtitle (Tiedemann, 2012) and the DegarBayan dataset, collecting 6M colloquial Persian sentences. Recognizing the time-consuming nature of manual data annotation, we decided to use alternative methods to generate the formal component of this large-scale data. We used Google Translate, which can translate English into formal Persian proficiently, to back-translate the colloquial sentences. Initially, we translated the colloquial dataset into English and retranslated it to formal Persian. the basic statistics of the C2FPersian dataset are shown in Table 1.

	Formal	**Colloquial**
# sentences	5,995,304	5,995,304
# tokens	67,224,318	62,815,506
# unique tokens	191,058	533,773
Average length	48.37	47.18

Table 1: Statistics of C2FPersian Dataset.

5 Evaluation

A substantial colloquial-to-formal Persian dataset is crucial for training effective deep-learning models in NLP tasks. To demonstrate the effectiveness of our dataset, we implemented a scoring system to evaluate the quality of formal sentences generated from colloquial ones using various methods. Persian native speakers participated in evaluating the output. This evaluation approach is further discussed in this section.

5.1 Models

We conducted fine-tuning on both the ParsGPT2 (Team, 2021) and ParsT5 (Par, 2022) models, which had previously been fine-tuned on Persian data, to generate formal correspondences for colloquial Persian sentences. Each model was fine-tuned separately on the C2FPersian dataset during this process. Additionally, we fine-tuned the models specifically on the manual dataset.

5.2 Scoring System

To compare model results, a web app with a scoring system was developed, allowing users to rank formal sentences based on given colloquial sentences. We selected 500 colloquial sentences from our manual dataset as test data, excluding them from C2FPersian to prevent data leaks during fine-tuning. Moreover, we added 500 colloquial sentences from LSCP (Khojasteh et al., 2020); these sentences are derived from informal Persian tweets, featuring a lack of structure and writing errors. Formal correspondences for this test dataset, consisting of 1000 sentences, were generated using ParsGPT2 and ParsT5 models, which had been fine-tuned on both the C2FPersian and manual datasets. In addition, to make the scoring mechanism more challenging, we incorporated ChatGPT and the FarsiYar rule-based system (Group, 2020) to generate formal sentences. In conclusion, the scoring system presented six different formal outputs for a single colloquial sentence, with these outputs being randomly ordered and the sources concealed. One ranked sample of the scoring system is shown in Figure 1.

5.3 Human Evaluation

We asked ten native Persian speakers to rank a set of 200 sentences based on the quality of the generated formal sentences, Using the scoring system mentioned earlier. Each sentence in the test dataset underwent two separate evaluations. The evaluation process involved counting instances where a model was ranked first, second, and third.

Model	**acc@1**	**acc@2**	**acc@3**
GPT2 manual	8.321	25.610	45.441
T5 manual	4.484	12.656	23.070
T5 Full	5.431	19.183	38.465
Farsiyar	18.087	43.597	65.471
ChatGPT	20.578	37.519	53.961
GPT2 C2F	43.099	61.435	73.592

Table 2: Model performance comparison based on overall scores. The table presents accuracy metrics at the top 1, 2, and 3 predictions for each model. Higher values indicate better performance.

Subsequently, we calculated the cumulative sum of ranks and determined the percentage of occurrences for each model being ranked in the first,

Model	Short			Medium			Long		
	acc@1	acc@2	acc@3	acc@1	acc@2	acc@3	acc@1	acc@2	acc@3
GPT2 manual	10.15	30.85	47.90	7.413	24.357	45.537	7.084	20.593	42.339
T5 manual	5.548	15.426	25.440	3.933	9.985	20.121	3.789	12.191	23.394
T5 C2F	7.713	26.387	46.955	5.144	18.760	35.703	2.965	10.873	31.137
Farsiyar	13.802	36.401	61.434	17.852	43.116	64.902	23.558	52.883	71.005
ChatGPT	12.720	22.598	38.024	16.944	36.762	56.581	**34.102**	**56.507**	**70.511**
GPT2 C2F	**50.068**	**68.336**	**80.2436**	**48.714**	**67.020**	**77.156**	28.501	46.952	61.614

Table 3: Performance comparison of GPT2 C2F on evaluation instances in different settings. Accuracy at top 1, 2, and 3 is reported. Results in **bold** indicate the best-performing metrics in each category.

Figure 1: Example of a ranked sample in the scoring system. The figure illustrates six different formal outputs generated for a single colloquial sentence. The outputs are randomly ordered, and the sources are concealed to ensure an unbiased evaluation.

second, and third positions.

The overall results of the ranking system are shown in Table 2. The GPT-2 model, trained on the entire C2FDataset (GPT2-C2F), outperformed ChatGPT, a large language model (LLM), and Farsi-Yar, a rule-based system, in colloquial to formal conversion.

Moreover, when fine-tuned on the entire dataset, both T5 and GPT-2 models (T5-C2F and GPT2-C2F, respectively) demonstrated better results than those fine-tuned only on the manually annotated dataset. This highlights the effectiveness of the C2FDataset in providing a source for colloquial to formal data.

Furthermore, to evaluate the strengths and weaknesses of our model, we divided the test dataset based on the length of the sentences into three subgroups: short (less than 50 characters), medium (between 50 and 150 characters), and long (more than 150 characters) sentences. The number of sentences in each group is 739, 661, and 607, respectively. The results are shown in Table 3. Since our models were fine-tuned on relatively short sentences, our leading model, GPT2-C2F, ranked first in evaluating short and medium-length sentences.

6 Conclusion

In this paper, we introduce C2FPersian, the first large-scale Persian colloquial to formal parallel dataset. The dataset contains 6 million pairs of colloquial and formal Persian sentences, which consists of a manually annotated corpus and a large-scale machine-generated corpus. We employed the back-translation method to create the machine-generated corpus by translating the colloquial dataset into English and then retranslating it into formal Persian. We trained two models on the C2FPersian dataset and used a scoring system to evaluate their performance. To ensure accuracy, we enlisted the help of native Persian speakers to rank formal sentences generated by these models against those produced by other methods. Our evaluation results demonstrate the effectiveness of this dataset in bridging the gap between colloquial and formal Persian language, highlighting its significance as a valuable resource in this domain.

7 Limitations

Certain limitations associated with this work have not been explored within the scope of this project. A subset of potential limitations related

to C2FDataset is as follows: **(i)** Our dataset might not cover all topics and might cause generalization problems. **(ii)** The Persian language includes many different dialects, but this dataset is focused only on the most common dialect (Tehrani). **(iii)** We need to investigate whether we have covered all the types of vocabulary, idioms, and common forms of colloquial Persian.

References

2022. A t5 model for persian. https://github.com/puraminy/parsT5.

Elmira Esmaeelpour, Sarah Saneei, and Mandana Nourbakhsh. 2022. Wordpars: A tool for orthographic and phonological neighborhood and other psycholinguistic statistics in persian. *Behavior Research Methods*, 54(4):1902–1911.

Heshaam Faili, Nava Ehsan, Mortaza Montazery, and Mohammad Taher Pilehvar. 2016. Vafa spell-checker for detecting spelling, grammatical, and real-word errors of persian language. *Digital Scholarship in the Humanities*, 31(1):95–117.

FarsiYar Text Mining Group. 2020. Farsiyar text mining tools. https://github.com/Text-Mining?language=go.

MUT Deep Group. 2022. Degarbayan-sc: A colloquial paraphrase farsi subtitles dataset. https://github.com/mut-deep/Degarbayan-SC.

Roya Kabiri, Simin Karimi, and Mihai Surdeanu. 2022. Informal persian universal dependency treebank. *arXiv preprint arXiv:2201.03679*.

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. Lscp: Enhanced large scale colloquial persian language understanding. *arXiv preprint arXiv:2003.06499*.

Bruno Martins and Mário J Silva. 2004. Spelling correction for search engine queries. In *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004. Proceedings 4*, pages 372–383. Springer.

Amin Naemi, Marjan Mansourvar, Mostafa Naemi, Bahman Damirchilu, Ali Ebrahimi, and Uffe Kock Wiil. 2021. Informal-to-formal word conversion for persian language using natural language processing techniques. In *2021 2nd international conference on computing, networks and internet of things*, pages 1–7.

Vahideh Tajali, Fateme Kalantari, and Mehrnoush Shamsfard. 2023. Developing an informal-formal persian corpus. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 3027–3037. Association for Computational Linguistics.

Hooshvare Team. 2021. Parsgpt2, a persian version of gpt2. https://github.com/hooshvare/parsgpt.

Jörg Tiedemann. 2012. 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).