Journal of Innovations in Computer Science and Engineering
Shahid Beheshti Unversity

June 2024, Special Volume 1, Issue 2

Split and rephrase: Simple Syntactic Sentences for NLP applications

Mahdi Asghari ✉, Code Orkid: **0009-0009-8225-324X**
Interdisciplinary Studies of Quran, Shahid Beheshti University, Tehran, Iran, mahdi.asghari1995@gmail.com
Alireza Talebpour, Code Orkid: **0000-0003-2538-3928**
Computer Science and Engineering, hahid Beheshti University, ehran, Iran, alebpour@sbu.ac.ir
Ghasem Darzi, Code Orkid: **0000-0001-8945-5666**
Interdisciplinary Studies of Quran, hahid Beheshti University, ehran, Iran, h_darzi@sbu.ac.ir

Abstract—In today's world, simplifying compound and complex sentences into simple sentences is crucial for enhancing machine understanding in various natural language processing (NLP) tasks, such as inference, machine translation, and information extraction. This simplification process improves accuracy. Consequently, our research is inspired by a text simplification method called "split and rephrase." We introduce a new sequence-to-sequence text generation model that transforms complex sentences into simple ones based on the conjunction "and" in Persian. By utilizing linguistic models with millions or even billions of parameters, our approach facilitates a better understanding of text complexities and more accurate identification of breaking points. Our results show an output accuracy of 0.47 in the BLEU score for the generated simple sentences, which are both grammatically correct and fluent.

Keywords—Split and rephrase, Text simplification, Compound sentence, Complex sentence, Simple sentence, Text generation

I. Introduction

In recent decades, there has been a growing need to use textual data with simpler grammatical structures to enhance machine comprehension in various NLP tasks. Texts with complex syntactic structures often lead to increased ambiguity and decreased accuracy in tasks such as information extraction, machine translation, and semantic role labeling, etc [1], [2]. Previous efforts in this area typically treat the problem as a monolingual translation, wherein complex texts are translated into simpler syntactic forms. However, it serves a different purpose than translating from one language to another. In fact, the goal of syntactic simplification of a text is to produce sentences with a simpler syntactic structure along with preserving the meaning, which may even lead to the production of a longer text than the original text [3]. Text simplification can be approached at two levels: lexical simplification and structural simplification. Utilizing simplified vocabulary and text structures can greatly benefit individuals with limited language skills, such as those with lower levels of education, children, non-native speakers, and individuals with learning disabilities like autism, dyslexia, or aphasia [4]. Furthermore, text simplification can act as a preprocessing step, yielding simpler, more accurate, and higher-quality text data. This enhancement in quality and the generation of augmented data can significantly improve the performance of various natural language processing (NLP) tasks, including parsing, machine translation, semantic role labeling, text summarization, and information extraction. [1], [2]. Most of the work of text simplification is focused on the analysis of special features at the sentence level, therefore, a more specific task called sentence simplification [5] can be formed, which includes two main goals:

1. Reducing lexical complexity: simplifying words Or difficult expressions in the text such as uncommon words, special terms, foreign words, etc [6] [7].

2. Reducing syntactic complexity: It refers to reducing sentence length and reducing grammatical complexity (for example, the number of functional clauses and coordinating clauses, as well as converting functional clauses and subordinate clauses into main clauses in the process of extracting simple sentences[8]).

In the realm of syntactic complexity reduction, "split and rephrase" is a sentence simplification method introduced by Narayan and his colleagues [8]. They proposed a novel sentence simplification task that has attracted significant research interest within the field of natural language processing. The goal of this approach is to decompose a complex input sentence into shorter, more manageable sentences while preserving the original meaning. Notably, this method does not involve deletion or lexical/phrasal simplification [9].

The inventors of this method [8] believe that the transformation of "separation" should be done with the help of semantic because, in many cases, separation occurs when a semantic entity participates in two (or more) distinct events described in a single sentence. For example, in the following sentence, bricks play a role in two events: “resistance to cold” and “the possibility of building permanent buildings.”

Submit Date: 2025-04-15
Accept Date: 2025-04-21
✉ Corresponding author

Original sentence: آجر به دلیل مقاومت بیشتر در برابر سرما، ساخت ساختمان‌های دائمی را امکان‌پذیر می‌کند.
Original sentence: Bricks allow for the construction of permanent buildings due to their greater resistance to cold.

Simplified sentence: آجرها در برابر سرما مقاومت بیشتری داشتند. آجرها امکان ساخت بناهای دائمی را فراهم کردند.
Simplified sentence: Bricks were more resistant to cold. Bricks made it possible to build permanent buildings

In this paper, we introduce a novel model designed to implement the split and rephrase technique for generating simple sentences utilizing the conjunction "and". We chose the word "and" because it can connect semantic entities (groups of words) that participate in one or more (semantic) relations. Also, the presence of inflectional phrases in a sentence can cause ambiguity and errors in various tasks such as information extraction. By selecting "and" as a breakpoint, the sentence can be segmented into independent sentences in which semantic entities engage in relationships. This approach ensures that the original meaning of the sentence is maintained. In addition, this method identifies the types of events in which a semantic entity participates in the verb inflectional phrase, and thus new triple relations are also generated. This approach addresses the split and rephrase challenge within the framework of sequence-to-sequence neural models. Our primary contribution lies in the development and training of a neural network model that leverages transfer learning to effectively disaggregate complex sentences and reconstruct sequences of simple sentences built around the conjunction "and" in Persian. This model is trained on a carefully curated Persian parallel corpus comprising pairs of complex sentences and their corresponding sequences of simple sentences.

II. RELATED WORK

III. Split and Rephrase

As we said earlier, The split and rephrase method, described by Narayan and his colleagues [8], is different from other rewriting tasks used in sentence simplification (SS), such as compressing, merging, or paraphrasing sentences. Unlike traditional sentence simplification methods, split and rephrase does not require deletion of information. Its goal is to keep the original meaning intact, even when the sentence is divided [5].

IV. Transforming complex sentences into a semantic hierarchy

Among the techniques for performing transformations required for sentence splitting, Niklaus and his colleagues [1] divide them into three categories: 1. *Syntactic rule-based approaches* These methods utilize a predefined set of handwritten syntactic rules to identify potential sentence split points. In the context of our study, these split points primarily include conjunctions such as "and". 2. *Semantic parsing based approaches* that aim to decompose sentences into minimal semantic units that may be divided into individual output sentences [7]. 3. *Data-driven approaches* These strategies leverage a parallel corpus comprising complex sentence and sequence of simple sentence pairs, allowing for the automatic learning of split points and various rewrite transformations to generate simplified sentences [8], [10].

Our proposed framework adopts the third approach, which relies on data, learning breakpoints, and rewriting transformations from a parallel corpus of pairs of complex and simple sentences.

V. Learning to split and rephrase from Wikipedia edit history

In the research conducted by Narayan et al. [8] and Goldberg et al. [10], two benchmark datasets, WebSplit and WikiSplit, were introduced, each comprising approximately one million records. The WebSplit dataset consists of complex sentences alongside their corresponding RDF triples, indicating that each complex sentence is associated with a set of simple semantic triples (semantic relations). Conversely, the WikiSplit dataset is derived from the rewriting activities in the editing history of Wikipedia. Generally, each complex sentence within this dataset corresponds to a singular simplified reference that features only one split, meaning that a complex sentence is divided into two distinct sentences. Notably, the WikiSplit dataset offers greater diversity in sentence structures compared to the WebSplit dataset and does not exhibit the limitations associated with the latter.

VI. Data and methodology

In the context of existing reference datasets, including Websplit and WikiSplit, we observe a notable absence of similar datasets in the Persian language that provide aligned data comprising complex sentences paired with corresponding sequences of simple sentences. To address this gap, we developed a novel dataset by systematically collecting pairs of compound/complex sentences and their corresponding sequences of simple sentences. Our dataset encompasses a total of 4,000 sentences, representing various types of inflectional sentences in Persian. These sentences were meticulously extracted from the pn-summary dataset, which is a well-structured resource designed for summarization tasks in the Persian language, containing a total of 93,207 records. This dataset is suitable for both abstractive and extractive summarization tasks, and it can also be utilized in diverse applications such as text generation, title generation, and news category classification [11]. To enhance the dataset further, we employed data augmentation techniques, resulting in an expanded collection of 16,000 sentences. Consequently, we created a parallel dataset that features pairs of complex sentences and sequences of simple sentences, specifically including constructs with 1 to 3 instances of the conjunction "and". In the subsequent sections, we provide precise definitions of key terms relevant to our study:

Inflectional sentence: It is called a sentence that has an inflectional phrase.
Nominal inflectional phrase: It is a phrase in which two or more noun groups are inflected with the word "and".
verb inflectional phrase: It is a phrase in which two or more verb groups are inflected with the word "and".

Table I presents various types of inflectional sentences that utilize the inflectional word "and," accompanied by illustrative examples.

TABLE I. TYPES OF INFLECTIONAL SENTENCES

Type	**Complex Sentence**	**Inflectional Phrases**	
عطفی اسمی	کارگزاران و فروشندگان نیز به اطلاعات دسترسی ندارند.	کارگزاران	فروشندگان
عطفی اسمی دارای حذف	سیاست‌گذاران باید با ارزیابی سیاست‌های مالیاتی و بانکی شرایط بد اقتصادی را تا حدودی تسکین دهند.	ارزیابی سیاست های مالیاتی	ارزیابی سیاست های بانکی
عطفی فعلی	این شرکت‌ها به‌وسیله ابزارهای خود اطلاعات کاربران را جمع‌آوری می‌کنند و برای تبلیغات هدفمند می‌فروشند.	اطلاعات کاربران را جمع‌آوری می‌کنند	اطلاعات کاربران را برای تبلیغات هدفمند می‌فروشند
ترکیبی: عطفی فعلی دارای حذف+ عطفی اسمی دارای حذف	از طرفی بارها کارشناسان اقتصادی از کم‌شدن در آمد دولت در ایام کرونا و تحریم نفت گفته‌اند.	در ایام کرونا گفته‌اند	در ایام تحریم نفت گفته‌اند
عطفی کاملا مجزا	باید قدرت رقابت محصولات خود را بهبود بخشیم و با کمک‌کردن به نوآوری ارزش بالاتری برای مشتریان ایجاد کنیم.	باید قدرت رقابت محصولات خود را بهبود بخشیم	با کمک‌کردن به نوآوری ارزش بالاتری برای مشتریان ایجاد کنیم

The collected dataset consists of all types of inflectional complex sentences that are listed in Table I, so a sample of the dataset will be in Table II.

TABLE II. SAMPLE OF THE COLLECTED DATASET

Complex Sentence	**Sequence of simple sentences**
این شرکت‌ها به‌وسیله ابزارهای خود اطلاعات کاربران را جمع‌آوری می‌کنند و برای تبلیغات هدفمند می‌فروشند.	این شرکت‌ها به‌وسیله ابزارهای خود اطلاعات کاربران را جمع‌آوری می‌کنند • این شرکت‌ها به‌وسیله ابزارهای خود اطلاعات کاربران را برای تبلیغات هدفمند می‌فروشند•

In this section, we transition into the methodology of our study. We define the split-and-rephrase task as follows: given a complex sentence C, the goal is to produce a simple text T, which consists of a sequence of sentences T1, T2, . . . , Tn, $n \geq 2$, so that T preserves the meaning of C [9].
We will provide details on the implementation of our proposed framework based on this definition.
Sequence-to-sequence neural model based on transformers

Our built MT5 model is based on encoder-decoder architecture consisting of attention units [12]. We used pre-trained MT5-Base-Parsinlu-translation-en-fa weights to train the model due to the limited resources and better understanding of the hidden relationships between the sentence components. We performed the model training process in a two-stage fine-tuning. Two-stage fine-tuning trains the sequence-to-sequence generative model in a task-specific manner in the split and rephrase task.
In the first stage, similar to the pre-training phase [13], the model was trained on a larger dataset containing approximately 14500 samples. This stage aimed to provide a foundational understanding of the specific task of splitting and rephrasing sentences. In the second stage, we fine-tuned the model on a smaller dataset of 1,500 samples, allowing it to learn the patterns involved in splitting and rephrasing inflectional sentences.
For training our model, we employed the HuggingFace Transformers library [14], using a batch size of 4, a learning rate of 0.00002, and training for 3 epochs.

VII. Results and discussion

VIII. Results

Based on the works of Narayan et al. [8] and Aharoni et al. [15], we present results at the sentence level using the BLEU score [16], BiLingual Evaluation Understudy, which is a primarily known metric comes from machine translation. This metric measures how closely a machine-generated text matches one or more reference texts .BLEU metric is the ratio of the overlap of common n-grams between the candidate text and the reference texts to the total number of n-grams of the candidate text.BLEU score is between 0 and 1. The higher value in BLEU is better.
Also, based on the mentioned reference works, we report the average number of simplified output sentences per complex input sentence (#S/C) .
And the average number of tokens in each simplified output sentence is calculated (#T/S) .
Therefore, it is a good model where the BLEU score is high and can produce more simple sentences on average (#S/C).
Unlike the #S/C metric, our objective is not to produce short, simple sentences. Consequently, a high #T/S value is desirable, as token deletion does not preserve the reference syntactic structure. Therefore, we do not aim to reduce the length of the generated sentences or engage in any form of summarization.
For the evaluation stage, we extracted a parallel test set of 100 sentence samples containing the inflectional word "and" from the pn-summary [11] test set, along with their corresponding sequence of simple sentences, which includes all types of inflectional sentences.
Table III reports the results obtained from evaluating the trained model against the test set, along with the scores of the

benchmark approaches (Copy512 , DisSim) on the WikiSplit test set and SEQ2SEQ256 on the WebSplit test set. We established our model as a competitive method. Copy512 is the strongest baseline reported by Aharoni and Goldberg [15] work. It is a sequence-to-sequence neural model augmented with a copy mechanism [17] that bias the model towards copying tokens from the complex input sentences, taking into account that many of them should appear in the simplified output sentences [9].

DisSim framework, by Niklaus et al. [1], is a recursive sentence splitting approach that applies a set of 35 hand-written rules to decompose a wide range of linguistic constructs, more oriented to generate simple and regular structures to support downstream semantic applications and faster generalization in machine learning tasks [9].
```

## IX. Discussion

**To achieve a detailed analysis,** we manually checked some of the predictions of our trained model and provided examples to help explain the scores shown in Table III. These extracted examples are listed in Table IV and show general patterns with some exciting behaviors produced by our method. Some of the results are as follows:

*X.* Based on Table III, it can be seen that the higher the degree of spliting of the model (#S/C), the lower the BLEU score of the model.

*XI.* Based on Table III, it can be seen that in an almost equal condition between the proposed model and DisSim, it can be seen that the DisSim model is more interested in producing shorter (more concise) sentences (#T/S).

*XII.* As can be seen in Table III, it can be said that one of the limitations of BLEU is the low correlation with simplicity during sentence spliting, [9] but it still has a high correlation with human evaluations of grammar and meaning retention [5].

*XIII.* Based on the examples in Table IV, it can be seen that the output sentences are gramatically correct and fluent, and it is also understood from the output that the meaning of the input sentence is also preserved.

XIV. As seen in example 5, and the general problem of generative models is their desire to repeat a part of the text in the output, there is a possibility of producing repeated section of output by the model, in this example, the sentence " **دولت بخش مربوط به احکام را پنجم دی ماه تقدیم به مجلس کرد** " is repeated.we can remove duplicate generated sentences if they are identical after remove punctuations. Also, in examples 2 and 5, the model once mistakenly used the symbol of the end of the sentence as "،" instead of ".".

XV. There is a small chance that the model will not be able to generate all simple sentences on the first try, so we can repeat the generation process up to 3 times to generate all possible sentences.

TABLE III. EVALUATION OF RESULTS

| **Models** | **Dataset** | **BLEU** | **#S/C** | **#T/S** |
|---|---|---|---|---|
| OurModel | OurDataset | 0.47 | 3.86 | **19.03** |
| SEQ2SEQ256 [8] | WebSplit | **77.48** | 2.57 | 10.56 |
| Copy512[15] | WikiSplit | **76.42** | 2.08 | 16.55 |
| DisSim[1] | WikiSplit | 51.96 | **4.09** | 11.91 |
```

TABLE IV. EXAMPLES PREDICTED BY THE OUR MODEL

Ex 1_Input sentence
وی از تلاش ویژه برای تسهیل در فعالیت تجار و بازرگانان خبر داد و گفت: تلاش داریم که شاهد رونق هر چه بیشتر بنادر استان بوشهر و توسعه اقتصاد دریا محور در استان باشیم.
Ex 1_Output sentences
وی گفت: تلاش داریم که شاهد رونق هر چه بیشتر بنادر استان بوشهر باشیم.
وی از تلاش ویژه برای تسهیل در فعالیت تجار خبر داد.
وی گفت: تلاش داریم که شاهد توسعه اقتصاد دریا محور در استان باشیم.
وی از تلاش ویژه برای تسهیل در فعالیت بازرگانان خبر داد.
Ex 2_Input sentence
اقلیم کردستان متاثر از توده‌های هوای گرم و مرطوب مدیترانه‌ای به گونه ای است که بارندگی‌هایی در بهار و ریزش برف در زمستان دارد.
Ex 2_Output sentences
اقلیم کردستان متاثر از توده های هوای گرم مدیترانه ای به گونه ای است که ریزش برف در زمستان دارد.
اقلیم کردستان متاثر از توده های هوای مرطوب مدیترانه ای به گونه ای است که ریزش برف در زمستان دارد.
اقلیم کردستان متاثر از توده های هوای گرم مدیترانه ای به گونه ای است که بارندگی هایی در بهار دارد، اقلیم کردستان متاثر از توده های هوای مرطوب مدیترانه ای به گونه ای است که ریزش برف در زمستان دارد،
اقلیم کردستان متاثر از توده های هوای مرطوب مدیترانه ای به گونه ای است که بارندگی هایی در بهار دارد.
اقلیم کردستان متاثر از توده های هوای گرم مدیترانه ای به گونه ای است که بارندگی هایی در بهار دارد.
Ex 3_ Input sentence
علی حسن را دید و گرفت کتک زد.
Ex 3_Output sentences
علی حسن کتک زد.
علی حسن را گرفت کتک زد.
علی حسن را گرفت.
علی حسن را دید.
Ex 4_Input sentence
اگر مایع گوش داخلی حرکت کند و وارد کانال‌های گوش داخلی شوند، این احساس سرگیجه و چرخیدن محیط به دور سر، بیشتر می‌شود.
Ex 4_ Output sentences
این احساس چرخیدن محیط به دور سر، بیشتر می شود.
اگر مایع گوش داخلی وارد کانال های گوش داخلی شوند، این احساس سرگیجه بیشتر می شود.
این احساس سرگیجه بیشتر می شود.
اگر مایع گوش داخلی حرکت کند، این احساس سرگیجه بیشتر می شود.
اگر مایع گوش داخلی وارد کانال های گوش داخلی شوند.
اگر مایع گوش داخلی حرکت کند، این احساس سرگیجه، بیشتر می شود.
اگر مایع گوش داخلی وارد کانال های گوش داخلی شوند، این احساس چرخیدن محیط به دور سر، بیشتر می شود.
اگر مایع گوش داخلی حرکت کند.
Ex 5_Input sentence
دولت بخش مربوط به بودجه شرکت های دولتی را آبان ماه و بخش مربوط به تبصره ها و احکام را پنجم دی ماه تقدیم به مجلس کرد.
Ex 5_Output sentences
دولت بخش مربوط به احکام را پنجم دی ماه تقدیم به مجلس کرد،
دولت بخش مربوط به بودجه شرکت های دولتی را آبان ماه تقدیم به مجلس کرد.
دولت بخش مربوط به احکام را پنجم دی ماه تقدیم به مجلس کرد.
دولت بخش مربوط به تبصره ها را پنجم دی ماه تقدیم به مجلس کرد.
دولت بخش مربوط به بودجه شرکت های دولتی را آبان ماه تقدیم به مجلس کرد بخش مربوط به تبصره ها را پنجم دی ماه تقدیم به مجلس کرد بخش مربوط به احکام را پنجم دی ماه تقدیم به مجلس کرد

XVI. conclusion

The split-and-rephrase task demonstrates that simpler grammatical constructions are generally more effectively processed by both humans and downstream NLP applications. We introduced a novel model utilizing the split-and-rephrase approach, centered around the inflectional conjunction "and," which leverages the capabilities of the transformer-based sequence-to-sequence MT5 model in conjunction with transfer learning techniques. Notably, prior to our work, there was a lack of research on the split-and-

rephrase process in Persian language. We have developed a comprehensive dataset and achieved competitive results compared to existing methodologies.

As for future work, we can expand dataset to a larger and more diverse dataset for better generalization and greater robustness. Also we plan to train our model on other inflectional words in Persian language. Additionally, we will investigate the effectiveness of decoder-only large language models, such as GPT-3.5, as replacements for the sequence-to-sequence transformer models. Moreover, we aim to leverage the results of our work to enhance the research outcomes of other researchers in the fields of information extraction, logical reasoning, and semantic role labeling, by incorporating our approach as a preprocessing step in those tasks.

1.1.1.1.1 References

[1] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "Transforming complex sentences into a semantic hierarchy," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1333.

[2] S. Štajner and M. Popović, "Automated text simplification as a preprocessing step for machine translation into an under-resourced language," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2019. doi: 10.26615/978-954-452-056-4_131.

[3] A. Menta and A. Garcia-Serrano, "Controllable Sentence Simplification Using Transfer Learning," in *CEUR Workshop Proceedings*, 2022.

[4] H. Guo, R. Pasunuru, and M. Bansal, "Dynamic multi-level multi-task learning for sentence simplification," in *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 2018.

[5] F. Alva-Manchego, C. Scarton, and L. Specia, "Data-driven sentence simplification: Survey and benchmark," *Computational Linguistics*, vol. 46, no. 1, 2020, doi: 10.1162/COLI_a_00370.

[6] S. Štajner, S. Nisioi, and I. Hulpus, "CoCo: A tool for automatically assessing conceptual complexity of texts," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020.

[7] S. Narayan and C. Gardent, "Hybrid simplification using deep semantics and machine translation," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 2014. doi: 10.3115/v1/p14-1041.

[8] S. Narayan, C. Gardent, S. B. Cohen, and A. Shimorina, "Split and Rephrase," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 606–616. doi: 10.18653/v1/D17-1064.

[9] P. B. Neto and E. E. S. Ruiz, "Split-and-Rephrase in a Cross-Lingual Manner: a Complete Pipeline," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021. doi: 10.26615/978-954-452-072-4_019.

[10] J. A. Botha, M. Faruqui, J. Alex, J. Baldridge, and D. Das, "Learning to split and rephrase from Wikipedia edit history," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018. doi: 10.18653/v1/d18-1080.

[11] M. Farahani, M. Gharachorloo, and M. Manthouri, "Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization," in *26th International Computer Conference, Computer Society of Iran, CSICC 2021*, 2021. doi: 10.1109/CSICC52343.2021.9420563.

[12] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[13] R. Sun, W. Xu, and X. Wan, "Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9345–9355. doi: 10.18653/v1/2023.findings-acl.595.

[14] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

[15] R. Aharoni and Y. Goldberg, "Split and rephrase: Better evaluation and a stronger baseline," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018. doi: 10.18653/v1/p18-2114.

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[17] J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1631–1640. doi: 10.18653/v1/P16-1154.

Mahdi Asghari: I am a Master's student in Computational Linguistics at Shahid Beheshti University, Tehran. I specialize in AI, ML, and Natural Language Processing (NLP), with over 3 years of experience in both research and industry. My areas of interest include Generative AI and Reinforcement Learning. One of my recent research projects involved designing a system to identify the domain of word groups in Persian using Transformer-based models. I am eager to collaborate with researchers in these fields in the future.
mahdi asghari
0009-0009-8225-324X

Alireza talebpour: I am an Associate Professor at Shahid Beheshti University (SBU), Tehran, specializing in AI, ML, NLP, and digital humanities, with 20+ years of experience in academia, leadership, and entrepreneurship. I co-founded five tech companies, including Tebyan, a leading tech enterprise in Iran during the 2000s, which played a key role in advancing AI-driven content distribution, data analytics, and cybersecurity solutions. I also established SBU's Cybersecurity Research Center to drive cutting-edge research and industry collaboration. My expertise spans ML, healthcare AI, NLP, and AI applications in social sciences, with a strong track record in innovation, research, and startup acceleration. Having led AI commercialization and large-scale technology projects, I am committed to fostering impactful advancements in AI.
With a Ph.D. from the University of Surrey and strong ties to the UK AI and engineering ecosystem, I seek to contribute my expertise in AI-driven innovation, research, and entrepreneurship to the UK's technology landscape through the Royal Academy of Engineering's Global Talent program.
0000-0003-2538-3928

Ghasem Darzi: Dr. Ghasem Darzi is an accomplished scholar and Assistant Professor at the Interdisciplinary Quranic Studies Research Institute at Shahid Beheshti University. He specializes in integrating Quranic studies with contemporary academic disciplines. Dr. Darzi earned his M.A. and Ph.D. in Theology (Quranic Sciences and Hadith) from the University of Tehran. His seminal works include the book 'The Methodology of Quranic Interdisciplinary Studies' and a Persian translation of the Oxford Handbook of Interdisciplinary Studies.
As an educator, Dr. Darzi has designed courses such as 'Qurʾān & Human Sciences' and 'Scientific Miracle and Humanities,' emphasizing the relevance of Quranic insights to modern challenges. His research spans diverse topics including gender discourse, cognitive science, and the integration of natural sciences with Quranic teachings. He advocates for inclusive education and has supported underrepresented students through workshops and lectures.
Dr. Darzi is a key leader in the academic community, serving as the head of the First International Conference on Interdisciplinary Quranic Studies and planning the second conference for 2025. He is also an active member of prominent academic organizations such as the International Organization for the Study of the Quran (IOSA) and the Society of Biblical Literature (SBL).
0000-0001-8945-5666