Salud, Ciencia y Tecnología - Serie de Conferencias. 2023; 2:472
doi: 10.56294/sctconf2023472

Categoría: Congreso Científico de la Fundación Salud, Ciencia y Tecnología 2023

ORIGINAL

Deep Learning Applied on Arabic language for punctuation marks prediction

Deep Learning aplicado al árabe para la predicción de signos de puntuación

Abdelkarim Aboutaib[1], Imad Zeroual[1], Ahmad EL Allaoui[1]

[1]L-STI, T-IDMS, FST Errachidia, Moulay Ismail University of Meknes, Morocco.

Citar como: Aboutaib A, Zeroual I, EL Allaoui A. Deep Learning Applied on Arabic language for punctuation marks prediction. Salud, Ciencia y Tecnología - Serie de Conferencias 2023; 2:472. https://doi.org/10.56294/sctconf2023472

Recibido: 08-06-2023 **Revisado:** 07-08-2023 **Aceptado:** 09-10-2023 **Publicado:** 10-10-2023

ABSTRACT

In the absence of explicit punctuation, the Arabic language's semantic and contextual nature poses a unique challenge, necessitating the reintroduction of punctuation marks for elucidating sentence structure and meaning. We investigate the impact of sentence length on punctuation prediction in the context of Arabic language processing. Leveraging Deep Neural Networks (DNNs), specifically Bi-Directional Long Short-Term Memory (Bi-LSTM) models. Our study goes beyond restoration, aiming to accurately predict punctuation marks in unprocessed text. The investigation focuses on five primary punctuation marks (.?,: and !), contributing to a more comprehensive understanding of predicting diverse punctuation marks in Arabic texts and we have achieved 85 % in accuracy . This research not only advances our understanding of Arabic language processing but also serves as a broader exploration of the relationship between sentence length and punctuation prediction.

Keywords: Deep Learning; Bi-LSTM; NLP; Attention.

En ausencia de signos de puntuación explícitos, la naturaleza semántica y contextual de la lengua árabe plantea un reto único, que hace necesaria la reintroducción de los signos de puntuación para dilucidar la estructura y el significado de las frases. Investigamos el impacto de la longitud de la frase en la predicción de la puntuación en el contexto del procesamiento de la lengua árabe. Aprovechando las redes neuronales profundas (DNN), en concreto los modelos bidireccionales de memoria larga a corto plazo (Bi-LSTM). Nuestro estudio va más allá de la restauración, con el objetivo de predecir con precisión los signos de puntuación en texto no procesado. La investigación se centra en cinco signos de puntuación principales (.?,: y !), lo que contribuye a una comprensión más completa de la predicción de diversos signos de puntuación en textos árabes, y hemos logrado un 85 % de precisión. Esta investigación no sólo avanza en nuestra comprensión del procesamiento de la lengua árabe, sino que también sirve como una exploración más amplia de la relación entre la longitud de la frase y la predicción de la puntuación.

Palabras clave: Aprendizaje profundo; Bi-LSTM; PNL; Atención.

INTRODUCTION

In the realm of Arabic language processing, the absence of explicit punctuation presents a formidable challenge, compelling the reintroduction of punctuation marks to elucidate sentence structure and contextual meaning. The inherently semantic and contextual nature of Arabic underscores the critical role of punctuation in conveying precise meaning and distinguishing various components of speech. Punctuation, through correct placement, not only organizes written discourse but also signifies the boundaries between sentences, facilitating the initiation of new ideas and concluding preceding ones.

Motivated by the imperative to enhance automated processing of Arabic texts, While [1] conclude that the quality of text perplexity can be influenced by punctuation usage. Our research leverages Deep Neural Networks (DNNs), specifically Bi-Directional Long Short-Term Memory (Bi-LSTM) models,[2] for predicting punctuation marks. This focus goes beyond mere restoration; our goal is to accurately predict the correct punctuation marks in unprocessed text. This predictive capability holds promise in scenarios where punctuation is absent or requires augmentation for improved text coherence and significance.

The applicability of our research extends beyond punctuation prediction, with potential implications for post-processing tasks in Automatic Speech Recognition (ASR) systems, such as generating automatic subtitles for videos.[3] Punctuation errors, identified as a common challenge in Arabic linguistic annotation studies, can be effectively addressed through advanced predictive models.

Several researchers have focused on punctuation prediction for various languages, including Slovenian,[4] Chinese,[5] Portuguese,[6] Arabic,[7] and others. Additionally, some researchers have explored the development of generalized models for this purpose.[8]

Our investigation delves into the nuanced relationship between sentence length and punctuation prediction in the Arabic language, addressing a significant gap in existing literature. While prior studies have often focused on specific punctuation types, our research concentrates on five primary punctuation marks (.?,: and !), allowing for a more comprehensive examination of the complexities associated with predicting diverse punctuation marks in Arabic text.

In [9] highlights the importance of punctuation within automatic speech recognition (ASR) and broader cognitive info-communication.[10] The study shows that both text-based and prosody-based approaches can provide reliable punctuation with low latency in practical applications of ASR technology. In the area of lexical features, several methods have been proposed. These include n-gram models,[11] transition-based dependency parsing,[12,13] conditional random fields (CRFs),[14] and deep neural networks.[16,17] Some systems exploit the encoder-decoder framework with an attention mechanism,[18] a structure widely used in numerous sequence-to-sequence translation tasks.[19] This review highlights the different strategies used to tackle punctuation tasks and provides a broad understanding of the methods used in ASR and related fields.

The structure of our paper aligns with scientific inquiry conventions, commencing with a detailed exposition of our methods. In this section, we elaborate on the models employed and the intricacies of the data used in our experimentation. Bi-Directional Long Short-Term Memory (Bi-LSTM) architectures, chosen for their efficacy in capturing contextual dependencies within sequences, feature prominently. The datasets encompass a rich variety of sentences, meticulously categorized by length, providing a robust foundation for our investigation.

Moving to the results section, we present the outcomes of our rigorous experimentation and comparative analyses. The impact of sentence length on punctuation prediction is systematically explored through dedicated models, each exclusively trained on sentences of a specific length. Our approach ensures a nuanced understanding of how sentence structures influence the predictive efficacy of Bi-LSTM models across various punctuation marks.

In the subsequent discussion, we engage in a thoughtful analysis of our findings, drawing connections between sentence length and the accuracy of punctuation prediction. We scrutinize the nuances revealed by our experiments, considering the implications for the broader field of Arabic language processing. As

https://doi.org/10.56294/sctconf2023472

we navigate through our results and interpretations, our overarching goal is to shed light on the effectiveness of Bi-LSTM models in capturing the intricate interplay between sentence length and punctuation prediction in Arabic text.

In conclusion, our research contributes valuable insights to the specific domain of Arabic language processing and serves as a broader exploration of the impact of sentence length on punctuation prediction. By adhering to the outlined structure and drawing upon the foundations laid by existing literature, we aim to elevate our understanding of this intricate relationship and pave the way for further advancements in the field.

METHODS

Various deep learning models and techniques have been explored in the field of document analysis research. This paper examines the importance and shortcomings of these approaches, highlighting their effectiveness and limitations. While progress has been made in using deep learning for document analysis, there are still unexplored areas that warrant further investigation. The study highlights the need for future research to delve into these untapped areas and unlock the full potential of deep learning to improve document analysis. In addition, the paper presents the latest deep learning frameworks and provides insights into the cutting-edge tools and technologies that are shaping the landscape of document analysis research,[20,21] These approaches where we have used in our previous work[7] to compare many models such as GRU-based[22] and LSTM-based models.

Bi-Directional Long Short-Term Memory (Bi-LSTM) models

In accordance with the findings presented in [7,23], we opt for a Bi-LSTM-based model to assess the influence of sentence length on punctuation prediction. Figure 1 presents the architecture of the Bi-LSTM model. Where Xi is the input token, Yi is the output token, and A and A' are LSTM nodes. The final output of Yi is the combination of A and A' LSTM nodes.

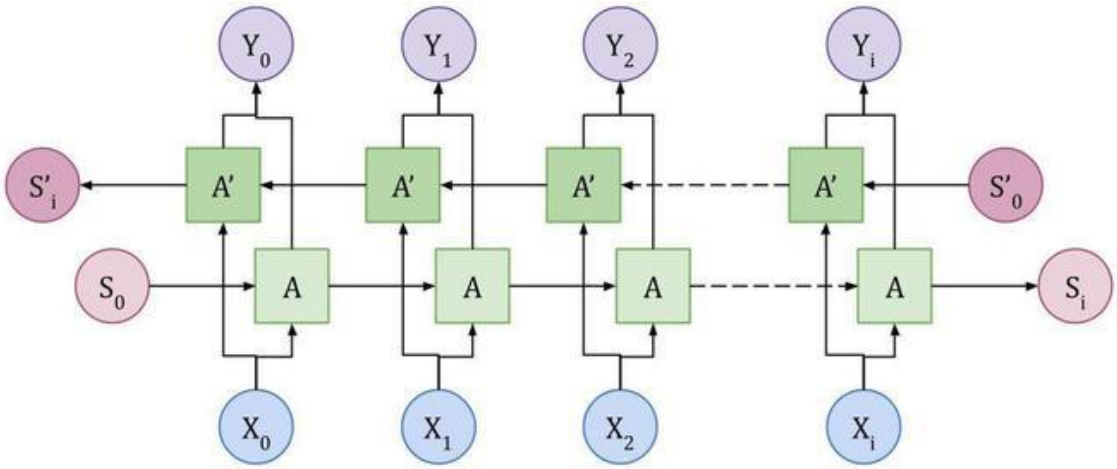

Figure 1. Architecture of the Bi-LSTM model

In the experimental setup, a Bi-LSTM model was chosen for its efficacy in sequence modeling. The architecture featured an embedding layer with an output dimension of 128, aimed at capturing the semantic nuances of the input sentences. The subsequent two LSTM layers,[24] each comprising 64 units, provided the model with the capacity to capture long-term dependencies. Dropout regularization with a rate of 0,3 was implemented after each LSTM layer to mitigate overfitting, and a learning rate of 1e-4 facilitated fine-tuning during training using Adam optimizer.[25]

https://doi.org/10.56294/sctconf2023472

Dataset

In our dataset,[26] an extensive variety of sentence lengths is captured, spanning from succinct sentences composed of merely three words to more extensive ones comprising up to twenty words. The dataset reveals nuanced patterns in punctuation frequencies, showcasing variations tied to differing sentence lengths. This variation reflects not only the diverse structures of sentences but also the distinct stylistic choices employed within the text. The dataset comprises a substantial corpus, encompassing a total of 57M sentences, with an expansive word count totaling 1,170M words. Within this linguistic landscape, a rich vocabulary is evident, consisting of 4M unique words. This comprehensive dataset is particularly well-suited for exploring the intricate interplay between sentence characteristics, punctuation usage, and linguistic styles, making it valuable for research in the realms of deep learning, natural language processing (NLP), and related fields.

The dataset derived from the ArPM corpus[26] has been carefully pre-processed, as shown in figure 2. Sentences were tokenized, and punctuation marks were assigned as labels, resulting in a comprehensive dataset of 100k sentences. Each sentence pertained to one of five punctuation marks: period, colon, exclamation, question, and comma. These sentences were organized into sequences of varying lengths (3, 5, 7, 9, 11, 13, 15, 17), forming the basis for a nuanced investigation into the impact of sentence length on punctuation prediction.

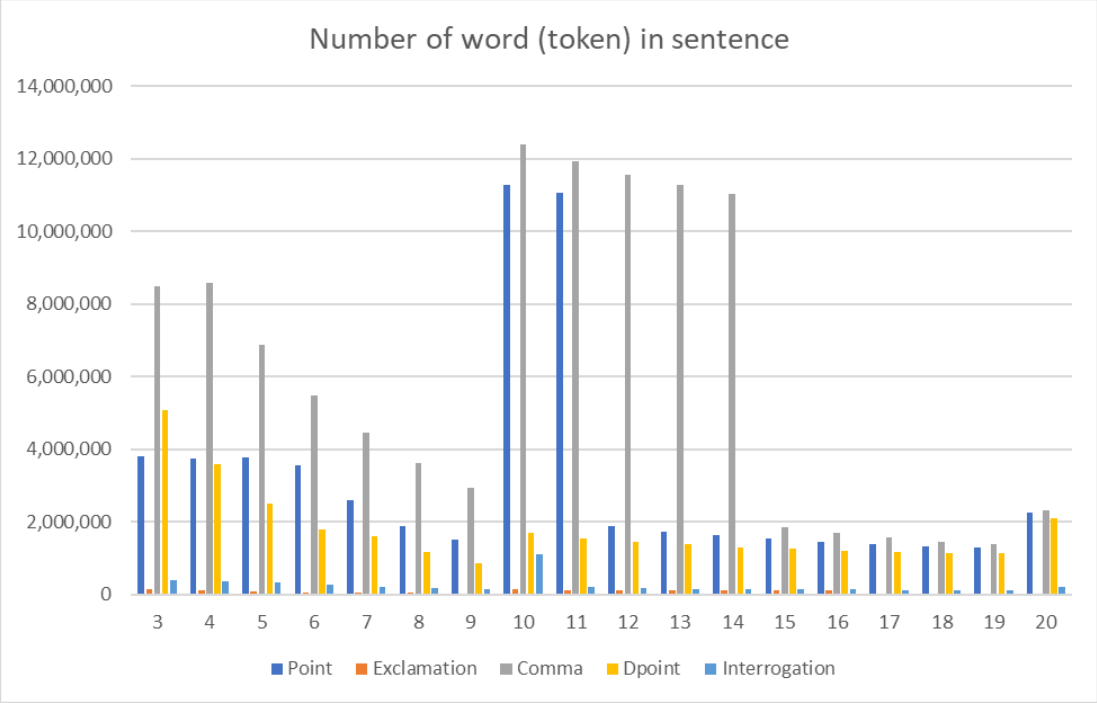

Figure 2. Number of words in a sentence ArPM corpus

Training Model

For the training phase, an innovative approach was adopted. Eight distinct models were trained, each dedicated to a specific sentence length. This deliberate strategy facilitated an in-depth exploration of the model's performance across diverse sentence structures, providing granularity in understanding the interplay between sentence length and punctuation prediction. The tables, denoted as table 1 and table 2, present the empirical results, offering valuable insights into the model's predictive efficacy for each punctuation mark across varying sentence lengths. This meticulous methodology aligns with the rigorous standards of scientific inquiry, ensuring a comprehensive exploration of the research question at hand.

https://doi.org/10.56294/sctconf2023472

RESULTS

The model's performance varies across different punctuation marks, with periods showing fluctuations, colons demonstrating stability, exclamation marks exhibiting variability, commas indicating robust performance, and question marks displaying some variability with a slight decreasing trend. The analysis provides insights into the model's strengths and potential areas for improvement in predicting different punctuation marks.

Table 1. Bi-LSTM model results (Precision, Recall, F1-score for each length sentence per Mark)

Length of sentence(words)	3	5	7	9	11	13	15	17
Metrics (%)	P R F	P R F	P R F	P R F	P R F	P R F	P R F	P R F
Period	51 47 49	58 52 55	60 52 56	47 44 45	46 43 44	39 48 43	40 50 44	41 45 43
Colon	46 51 48	44 49 46	52 55 53	53 55 54	49 49 49	49 39 43	54 40 46	51 41 45
Exclamation Mark	59 55 57	49 47 48	55 54 55	57 54 55	57 61 59	61 56 58	53 55 54	44 50 47
Comma	70 72 71	64 68 66	69 71 70	70 72 71	68 72 70	71 69 70	72 71 72	72 76 74
Question Mark	83 82 83	80 78 79	76 82 79	75 78 76	75 72 73	68 74 71	73 72 73	73 70 71

1. Period

- The predictions for the period punctuation mark show some variability, starting at 51, reaching a peak at 60, and then stabilizing in the range of 47 to 56. The overall pattern suggests fluctuations in the model's confidence for predicting periods.

2. Colon

- Predictions for the colon punctuation mark exhibit a relatively stable pattern, with values varying within the range of 44 to 55. This indicates a consistent level of confidence in predicting colons across the given data.

3. Exclamation Mark

- The model's predictions for the exclamation mark show a more varied trend, with values oscillating between 47 and 61. This suggests a range of confidence levels in predicting exclamation marks, indicating potential sensitivity to contextual variations.

4. Comma

- Predictions for the comma punctuation mark reveal a relatively consistent and high range, ranging from 64 to 76. This indicates a robust and stable performance in predicting commas, with the model consistently confident in these predictions.

5. Question Mark

- The model's predictions for the question mark display variations between 68 and 83, with a slight overall decreasing trend. This suggests some variability in the model's confidence for predicting question marks, possibly influenced by specific contexts within the data.

In examining the influence of sentence length on punctuation prediction, distinct patterns emerge for each punctuation mark. The model's predictions for periods exhibit fluctuations, prompting an investigation into whether sentence length correlates with increased confidence in predicting periods. Conversely, predictions for colons remain relatively stable across varying sentence lengths, suggesting a consistent model performance irrespective of sentence length. The varied trend in exclamation mark predictions prompts an exploration of potential connections between sentence length and the model's confidence. For commas, the consistently high prediction range implies that sentence length may have

minimal impact on accuracy. Lastly, the variations and slight decreasing trend in question mark predictions warrant a closer look into how sentence length affects the model's proficiency in predicting question marks, discerning potential challenges or advantages associated with shorter or longer sentences. This nuanced analysis provides valuable insights into the interplay between sentence length and punctuation prediction outcomes.

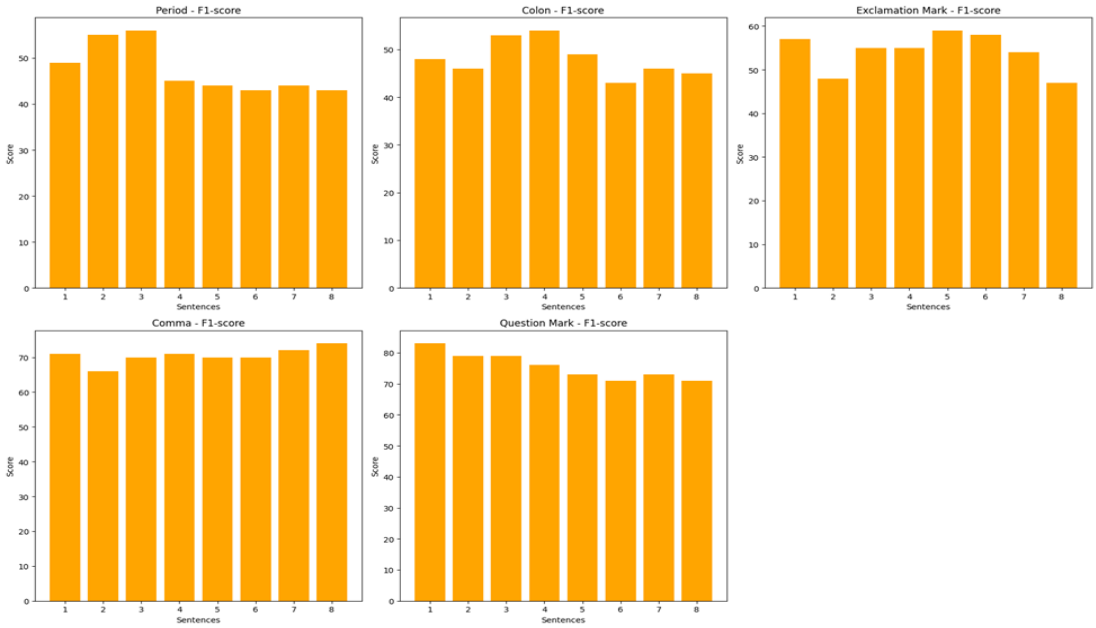

Figure 3: F1-score for Each Punctuation Mark Across Sentences with Different Lengths for Bi-LSTM

Table 2. Bi-LSTM-Att model results (Precision, Recall, F1-score for each length sentence per Mark)

Length of sentence(words)	3	5	7	9	11	13	15	17
Metrics (%)	P R F	P R F	P R F	P R F	P R F	P R F	P R F	P R F
Period	48 53 50	54 56 55	57 55 56	44 49 47	46 43 44	41 41 41	40 46 43	43 44 43
Colon	49 49 49	46 44 45	53 51 52	55 51 53	48 50 49	49 36 41	50 44 47	50 43 46
Exclamation Mark	60 55 57	49 48 48	54 57 56	58 57 57	60 59 59	59 58 59	58 56 57	46 48 47
Comma	72 71 71	65 66 65	70 69 70	70 71 70	69 73 71	66 74 70	74 70 72	72 73 73
Question Mark	84 83 83	78 79 78	76 80 78	78 74 76	74 72 73	67 76 71	70 76 73	67 75 71

For Bi-LSTM-Att model we can observe:

1. Period
   - Accuracy: Exhibits an incremental trend, rising from 48 to 57.
   - Precision: Shows an initial increase from 53 to 56, followed by a slight decrease to 55.
   - Recall: Demonstrates an overall incremental pattern, ascending from 50 to 55 and then experiencing a minor decline to 56.
2. Colon

https://doi.org/10.56294/sctconf2023472

- Accuracy: Varied, with values oscillating between 46 and 55.
- Precision: Initially decreases from 49 to 45, then rises again to 52.
- Recall: Displays fluctuations, with values ranging from 49 to 53.

3. Exclamation Mark
   - Accuracy: Varied, with values ranging from 48 to 60.
   - Precision: Shows slight variations, generally maintaining values in the mid-50s.
   - Recall: Remains relatively stable, with values ranging from 48 to 60.
4. Comma
   - Accuracy: Relatively stable, with values around 70.
   - Precision: Shows minor fluctuations, maintaining values around 70.
   - Recall: Remains consistent, with values around 71.
5. Question Mark
   - Accuracy: Varied, with values ranging from 67 to 84.
   - Precision: Displays fluctuations, with values ranging from 72 to 83.
   - Recall: Shows some variations, with values ranging from 71 to 84.

Specifically, for the period punctuation, there is a discernible incremental trend in accuracy, implying that the predictive performance improves as sentence length increases. However, precision and recall exhibit more nuanced behaviors, with precision initially increasing before experiencing a slight decrease, and recall showing an overall incremental pattern but with a minor decline in the middle.

For colon punctuation, the accuracy trend appears varied, suggesting that sentence length may not have a consistent impact on prediction accuracy. Precision, on the other hand, shows fluctuations, indicating potential sensitivity to sentence length changes. Recall also displays variability, further emphasizing the nuanced impact of sentence length on colon prediction.

The exclamation mark punctuation reveals mixed results, with accuracy and recall exhibiting varied patterns, possibly suggesting a complex relationship with sentence length. Precision remains relatively stable, implying a consistent predictive precision regardless of sentence length.

In the case of the comma punctuation, the stability in accuracy and recall suggests that sentence length might have a limited impact on predicting commas. Precision, with minor fluctuations, indicates a subtle sensitivity to sentence length variations.

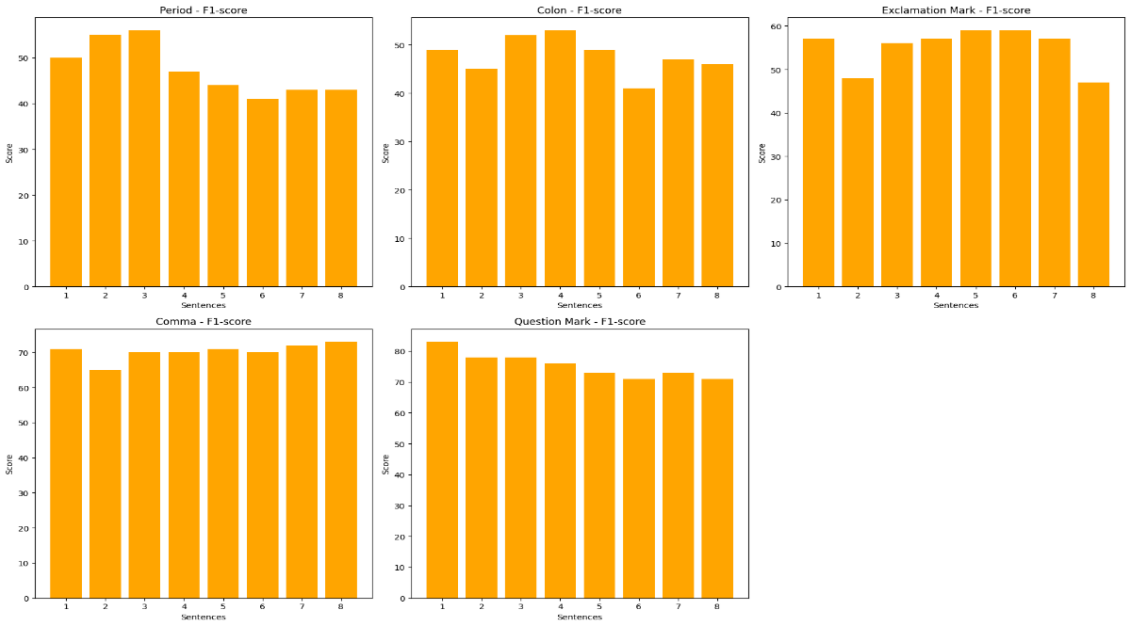

Figure 4. F1-score for Each Punctuation Mark Across Sentences with Different Lengths for Bi-LSTM-Att

https://doi.org/10.56294/sctconf2023472

Finally, for the question mark punctuation, the data suggests a diverse impact of sentence length on prediction outcomes, with accuracy and recall showing variations and precision displaying sensitivity to changes in sentence length.

Overall, the observed trends underscore the importance of considering sentence length as a factor in punctuation prediction models, with different punctuation marks exhibiting unique sensitivities to varying sentence lengths. Further investigation and fine-tuning of models may be warranted to optimize punctuation prediction across a range of sentence lengths.

DISCUSSION

In comparing the results of the Bi-LSTM model with those of the Bi-LSTM-Att model (the first model incorporating an attention mechanism), several observations can be made.

For the Bi-LSTM model as illustrate in the figure 3:

- Period predictions range from 39 to 60.
- Colon predictions vary between 39 and 55.
- Exclamation mark predictions fluctuate from 44 to 61.
- Comma predictions fall within the range of 66 to 76.
- Question mark predictions span from 70 to 83.

For the Bi-LSTM-Att model as shown in figure 4:

- Period predictions range from 40 to 57.
- Colon predictions vary between 36 and 55.
- Exclamation mark predictions fluctuate from 46 to 60.
- Comma predictions fall within the range of 65 to 74.
- Question mark predictions span from 67 to 84.

Upon comparison, it is evident that the Bi-LSTM-Att model generally yields predictions within a narrower range for each punctuation mark, suggesting a more constrained and potentially focused output. Additionally, the attention mechanism in the Bi-LSTM-Att model appears to contribute to a more refined prediction pattern. The Bi-LSTM-Att model, with its attention mechanism, may offer advantages in scenarios where nuanced attention to certain parts of the input sequence is crucial for improved punctuation prediction.

CONCLUSION

The results offer a clear insight into the influence of sentence length on prediction. Notably, as sentence length increases, there is a slight decrease in prediction accuracy. However, overall, the predictions remain relatively stable, hovering around 5 %, contingent upon the specific punctuation mark in consideration.

REFERENCES

1. Y. Wang, J. Deng, A. Sun, and X. Meng, “Perplexity from PLM Is Unreliable for Evaluating Text Quality.” arXiv, Mar. 15, 2023. Accessed: Dec. 26, 2023. [Online]. Available: http://arxiv.org/abs/2210.05892

2. M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, no. 11, Art. no. 11, 1997, doi: 10.1109/78.650093.

3. Ł. Augustyniak et al., “Punctuation Prediction in Spontaneous Conversations: Can We Mitigate ASR Errors with Retrofitted Word Embeddings?,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.05985

4. M. Bajec, M. Janković, S. Žitnik, and I. L. Bajec, “Punctuation Restoration System for Slovene Language,” in Research Challenges in Information Science, F. Dalpiaz, J. Zdravkovic, and P. Loucopoulos, Eds., Cham: Springer International Publishing, 2020, pp. 509–514.

5. International Association for Pattern Recognition, Zhongguo ke xue yuan, and Chinese Association of Automation, 2018 24th International Conference on Pattern Recognition (ICPR).

6. T. B. D. Lima et al., “Sequence Labeling Algorithms for Punctuation Restoration in Brazilian Portuguese Texts,” in Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28 – December 1, 2022, Proceedings, Part II, Berlin, Heidelberg: Springer-Verlag, 2022, pp. 616–630. doi: 10.1007/978-3-031-21689-3_43.

7. A. Aboutaib, A. El allaoui, I. Zeroual, and E. W. Dadi, “Punctuation Prediction for the Arabic Language,” in Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security, in NISS ’23. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3607720.3607734.

8. X. Li and E. Lin, “A 43 Language Multilingual Punctuation Prediction Neural Network Model.” [Online]. Available: https://github.com/pytorch/pytorch

9. M. Á. Tündik and G. Szaszák, “Joint Word- and Character-level Embedding CNN-RNN Models for Punctuation Restoration,” in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2018, pp. 000135–000140. doi: 10.1109/CogInfoCom.2018.8639876.

10. P. Baranyi, A. Csapo, and G. Sallai, Cognitive infocommunications (coginfocom). Springer, 2015.

11. A. Gravano, M. Jansche, and M. Bacchiani, “Restoring punctuation and capitalization in transcribed speech,” 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4741–4744, 2009.

12. R. Pan, J. A. García-Díaz, and R. Valencia-García, “Evaluation of Transformer-Based Models for Punctuation and Capitalization Restoration in Spanish and Portuguese,” in Natural Language Processing and Information Systems: 28th International Conference on Applications of Natural Language to Information Systems, NLDB 2023, Derby, UK, June 21–23, 2023, Proceedings, Berlin, Heidelberg: Springer-Verlag, 2023, pp. 243–256. doi: 10.1007/978-3-031-35320-8_17.

13. R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, “Exploring the Limits of Language Modeling.” 2016. [Online]. Available: http://arxiv.org/abs/1602.02410

14. W. Lu and H. T. Ng, “Better punctuation prediction with dynamic conditional random fields,” in Proceedings of the 2010 conference on empirical methods in natural language processing, 2010, pp. 177–186.

15. X. Che, C. Wang, H. Yang, and C. Meinel, “Punctuation Prediction for Unsegmented Transcript Based on Word Vector.” [Online]. Available: http://nlp.stanford.edu/projects/glove/

16. F. Wang, W. Chen, Z. Yang, and B. Xu, "Self-Attention Based Network for Punctuation Restoration," in 2018 24th International Conference on Pattern Recognition (ICPR), Beijing: IEEE, Aug. 2018, pp. 2803–2808. doi: 10.1109/ICPR.2018.8545470.

17. O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech and Communication Association, 2016, pp. 3047–3051. doi: 10.21437/Interspeech.2016-1517.

18. A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

19. O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in Interspeech 2016, ISCA, Sep. 2016, pp. 3047–3051. doi: 10.21437/Interspeech.2016-1517.

20. R. Al-Shalabi, G. Kanaan, T. Kanan, and M. ElBes, "A Review Study for Arabic Machine Learning and Deep Learning Methods," in 2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), 2022, pp. 225–232. doi: 10.1109/ICETSIS55481.2022.9888948.

21. M. K. Siddhu and S. N. Yaakob, "Deep learning applied to arabic and latin scripts: A review," International Journal of Scientific and Technology Research, vol. 8, no. 11, pp. 1510–1521, 2019.

22. R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), IEEE, 2017, pp. 1597–1600.

23. C. C. Juin, R. X. J. Wei, L. F. D'Haro, and R. E. Banchs, "Punctuation prediction using a bidirectional recurrent neural network with part-of-speech tagging," in TENCON 2017-2017 IEEE Region 10 Conference, IEEE, 2017, pp. 1806–1811.

24. O. Tilk and T. Alumäe, "LSTM for Punctuation Restoration in Speech Transcripts," 2015. [Online]. Available: http://bark.phon.ioc.ee/tsab

25. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," CoRR, vol. abs/1412.6980, 2014.

26. Abdelkarim Aboutaib, "Punctuations corpus for Arabic"," Mendeley Data, vol. V1, 2023, doi: 10.17632/jnz483dypx.1.

27. Farhaoui, Y. and All, Big Data Mining and Analytics, 2022, 5(4), pp. I IIDOI: 10.26599/BDMA.2022.9020004

28. Farhaoui, Y.and All, Big Data Mining and Analytics, 2023, 6(3), pp. I-II, DOI: 10.26599/BDMA.2022.9020045

FINANCING

The authors did not receive funding for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Data curation: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Formal analysis: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Research: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Methodology: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Project administration: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Resources: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Software: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Supervision: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Validation: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Visualization: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Writing - original draft: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.
Writing - proofreading and editing: Abdelkarim Aboutaib, Imad Zeroual, Ahmad EL Allaoui.