Enhancing Low-Resource Question-Answering Performance Through Word Seeding and Customized Refinement

Hariom Pandya, Brijesh Bhatt
Computer Engineering Department, Dharmsinh Desai University
College Road, Nadiad, 387001, Gujarat, India

***Abstract*—The state-of-the-art approaches in Question-Answering (QA) systems necessitate extensive supervised training datasets. In low-resource languages (LRL), the scarcity of data poses a bottleneck, and the manual annotation of labeled data is a rigorous process. Addressing this challenge, some recent efforts have explored cross-lingual or multilingual QA learning by leveraging training data from resource-rich languages (RRL). However, the efficiency of such approaches relies on syntactic compatibility between languages. The paper introduces the innovative method that involves seeding LRL data into RRL to create a bilingual supervised corpus while preserving the syntactical structure of RRL. The method employs the translation and transliteration of selected parts-of-speech (POS) category words. Additionally, the paper also proposes a customized approach to fine-tune the models using bilingual data. Employing the bilingual data and the proposed fine-tuning approach, the most successful model has achieved a 75.62 F1 score on the XQuAD Hindi dataset and a 68.92 F1 score on the MLQA Hindi dataset in a zero-shot architecture. In the experiments conducted using few-shot learning setup, the highest F1 scores of 79.17 on the XQuAD Hindi dataset and 70.42 on the MLQA Hindi dataset have been achieved.**

***Keywords*—Embedding learning; words seeding; bilingual dataset generation; low-resource question-answering**

I. Introduction

In recent years the pre-trained models have shown notable performance on many downstream Natural Language Processing (NLP) tasks such as Question-Answering(QA), summarization, machine translation, sentiment analysis, etc. [1], [2], [3], [4], [5], [6]. To use the pre-train models for the task other than the one on which it has been trained [7], fine-tuning on the task-specific supervised dataset is required. While the fine-tuning datasets are available in many resource-rich languages(RRLs) like English, French, and German[8], there are some languages that suffer from the bottleneck of the unavailability of supervised task-specific data.

In various fields of NLP [9], [10], [11], [12], [13], there have been efforts to tackle the situation of LRL data scarcity by annotating RRL datasets.

This paper introduces a method for integrating Hindi terms into English supervised corpora. It is noted that variations in syntactic structures between languages can detrimentally impact the effectiveness of question answering tasks. For example, English follows SVO (Subject - Verb - Object) word order whereas SOV (Subject - Object - Verb) word order is followed in the Hindi language. The proposed approach not only maintains a syntactic structure but also improves the word overlapping between question and context tokens.

It is observed that through the integration of Hindi noun category terms into English supervised data, a supervised QA dataset for LRL can be produced with minimal manual labeling required. Furthermore, it has been demonstrated that this newly generated LRL dataset can be effectively utilized alongside a tailored transfer-learning approach to attain benchmark performance levels. The methodology of transfer-learning is discussed in IV section.

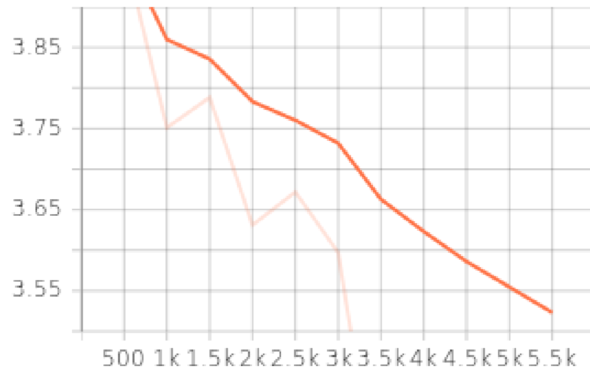
Our major contributions are as follows:

1) For the LRL, a method is presented to construct a bilingual QA supervised dataset by integrating LRL words into the RRL corpora.
2) The proposed transfer-learning mechanism leverages bilingual supervised QA dataset to enable task-specific learning and language structure learning together.
3) A method is proposed to modify the position of *answer_start* during the generation of bilingual annotated data. This method relies on n-gram matching between the answer and context tokens.
4) An analysis of the translation and transliteration of nouns from the source RRL to the destination LRL is also furnished, along with its repercussions on the QA task.

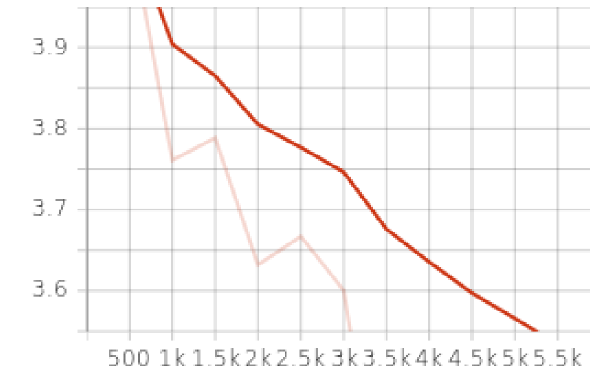
The remaining paper is organized as follows. The next section describes the existing work in the directions of LM learning and QA task. The noun seeding approaches and challenges of transliteration and translation are given in Section III. The proposed approach to QA learning is mentioned in the Section IV. In Section V, the discussion revolves around the impact on performance and the analysis of the obtained results.

II. Related Work

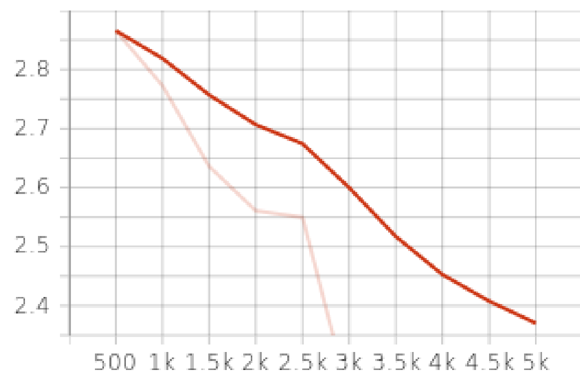
The development of the state-of-the-art QA models ([14], [15], [16], [17], [18], [19], [20], [21], [22]) is facilitated by numerous supervised large-scale question-answering datasets. Majority of QA datasets are either labelled manually by crowdworkers (e.g., SQuAD [23], HotPotQA [24], NewsQA [25]) or originated from human inputs such as conversations or search query logs (e.g., MS MARCO [26], NaturalQuestions



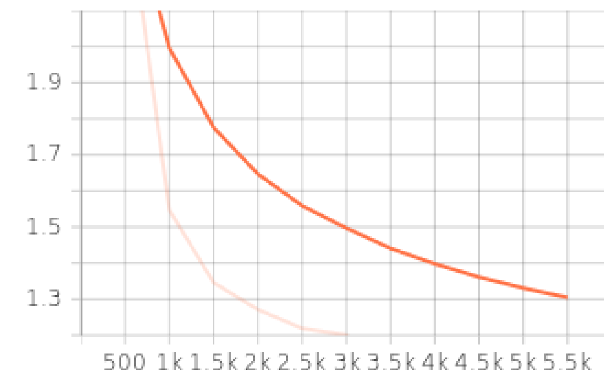
(a) mBERT-MLM



(b) mBERT-SQuAD



(c) XLM-R$_{Large}$-MLM



(d) XLM-R$_{Large}$-SQuAD

Fig. 3. Training loss of the zero-shot learning steps on mBERT and XLM-R$_{Large}$ models.

the models for the embedding training.

b) Supervised data for Question-Answering training: On the SQuAD 1.1 ([23]) English dataset, the models were trained for single epoch. To train the model further on bilingual QA data, the model were trained on task-specific bilingual corpora. The Hindi subset of MLQA dataset ([59]) and XQuAD ([56]) were used to train the models on Hindi QA task in few-shot setup. The few-shot training was executed for two epochs on XQuAD or MLQA dataset depending on the model. Our models, trained on MLQA, are evaluated on the XQuAD Hindi dataset and visa-versa.

C. Result Analysis

Table III shows an example context paragraph from SQuAD training set. The table indicates the approach of translation of common nouns and transliteration of a proper noun has more word overlapping with the Hindi translation version as compared to the transliteration of all nouns (overlapping is highlighted in blue color). However, there are few cases where the translation-transliteration approach leads to incorrect translation as the Hindi translation of a word is independent of the statement structure and neighborhood words (highlighted in red color). For example, the translation tool has converted the word *end* to समाप्त which is the correct translation. However, for the current context, it should be समाप्ति. Table III also depicts that the synonyms are also playing vital role in the translation as mentioned in III-A0b. Some examples of synonym pairs from the table are (यूनियन-संघ), (आजादी-स्वतंत्रता), and (बढ़ोतरी-वृद्धि).

Table V indicates zero-shot and few-shot learning results on the MLQA Hindi dataset. The baseline results obtained for mBERT and XLM-R$_{Large}$ models are highlighted with † sign in the table. The models trained after all noun replacement are producing the best results. In the zero-shot configuration, XLM-R$_{Large}$ model has achieved the best (68.92/52.24) (F1/EM) scores and the best score of the mBERT model is (49.45/34.55). In the few-shot configuration when the same models are trained on XQuAD, the XLM-R$_{Large}$ model has achieved (70.42/54.51) (F1/EM) scores. The best few-shot F1 score is 1.5% better than zero-shot. Additionally, for the MLQA dataset, the best performance difference between zero-shot and few-shot setup for the mBERT is 11.29% which is just 1.5% in XLM-R$_{Large}$ model. This shows for the mBERT models, the few-shot XQuAD training helps in boosting the overall performance.

Table VI shows zero-shot and few-shot learning results on the XQuAD Hindi dataset. The baseline results obtained for mBERT and XLM-R$_{Large}$ models are highlighted with † sign in the table. In the zero-shot setup, the best performance on the XQuAD Hindi dataset has been observed by the setup of the models trained on all nouns seeding dataset, followed by SQuAD training. Specifically, XLM-R$_{Large}$ model has achieved (75.62/58.65) (F1/EM) and (56.04/40.50) (F1/EM) is the score of the mBERT for the same configuration. When the same models were trained on MLQA to report a few-shot learning outcome, the same XLM-R$_{Large}$ model has achieved (79.17/62.18) (F1/EM) scores and (71.52/55.46) (F1/EM) is the mBERT result. The best few-shot F1 score is 3.55% better than zero-shot.

Results obtained in both tables suggest that common noun translation and proper noun transliteration have improved the performance of XLM-R and mBERT models for both MLQA and XQuAD datasets as it involves the replacement of 31.93% English tokens by its aligned Hindi version.

VI. Conclusion and Future work

In this paper, a novel method is introduced aimed at seeding low-resource words to establish a bilingual supervised QA dataset while ensuring the syntactic structure of the RRL is maintained. The proposed approach leverages the RRL and incorporates transliteration or translation techniques for nouns into the LRL. This method facilitates the creation of a robust bilingual dataset for question-answering tasks, addressing the challenge of limited resources in certain languages while preserving syntactic coherence and linguistic structure across languages. By utilizing this approach, the availability and quality of datasets for training and evaluating QA systems in bilingual settings has been enhanced, contributing to advancements in NLP and QA research. Moreover, the issue of aligning *answer_start* following the LRL word seeding process, has been addressed. Performance analysis of our approach and bilingual corpora on MLQA and XQuAD Hindi datasets has been conducted utilizing the mBERT and XLM$_{Large}$ architectures. In the zero-shot setup, our best-performing models have shown (75.62 / 58.65) (F1/EM) on the XQuAD Hindi dataset and (68.92/52.24) (F1/EM) scores on the MLQA Hindi dataset. In the few-shot setup, our best-performing models have shown (79.17/62.18) (F1/EM) on the XQuAD Hindi dataset and (70.42/54.51) (F1/EM) scores on the MLQA Hindi testset.

The proposed work opens avenues for future research in several areas. An intriguing direction is the analysis of POS category-based Hindi translation or transliteration and text annotation using all possible translated synonyms. However, it is important to acknowledge that in translation, synonyms might alter the sentence focus, even though they refer to the same concept, thus potentially introducing ambiguity. Another area worth exploring is the identification of the most suitable word replacement by translation or transliteration based on POS category, coupled with an in-depth analysis of the impact of all word replacements. This comprehensive approach would help address the limitations inherent in the current method and provide insights for improving accuracy and effectiveness. Additionally, examining the impact of word replacement by synonyms could be a promising avenue for further investigation, shedding light on potential limitations and challenges. Furthermore, regarding the mBERT model, while it demonstrates a notable improvement in few-shot learning compared to XLM-R$_{Large}$, further investigation into the underlying reasons for this disparity is warranted to gain a deeper understanding of model performance. By addressing these limitations and delving into these research directions, future studies can enhance the current work of multilingual QA systems.

References

[1] J. Liu, Y. Chen, and J. Xu, "Document-level event argument linking as machine reading comprehension," *Neurocomputing*, vol. 488, pp. 414–423, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222002867

[2] B. Ofoghi, M. Mahdiloo, and J. Yearwood, "Data envelopment analysis of linguistic features and passage relevance for open-domain question answering," *Knowledge-Based Systems*, vol. 244, p. 108574, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122002568

[3] D. Suleiman and A. Awajan, "Multilayer encoder and single-layer decoder for abstractive arabic text summarization," *Knowledge-Based Systems*, vol. 237, p. 107791, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121010005

[4] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Question-aware transformer models for consumer health question summarization," *Journal of Biomedical Informatics*, vol. 128, p. 104040, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422000569

[5] S. M. Jain, *Fine-Tuning Pretrained Models*. Berkeley, CA: Apress, 2022, pp. 137–151. [Online]. Available: https://doi.org/10.1007/978-1-4842-8844-3_6

[6] S. Tarek, H. M. Noaman, and M. Kayed, "Enhancing question pairs identification with ensemble learning: Integrating machine learning and deep learning models," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.01411100

[7] C. Zhang, Y. Lai, Y. Feng, and D. Zhao, "A review of deep learning in question answering over knowledge bases," *AI Open*, vol. 2, pp. 205–215, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000292

[8] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021, https://www.sciencedirect.com/science/article/pii/S1574013720304366.

[9] M. Marchal, M. Scholman, and V. Demberg, "Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin," in *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, Nov. 2021, pp. 84–94. [Online]. Available: https://aclanthology.org/2021.codi-main.8

[10] M. A. Hedderich, L. Lange, and D. Klakow, "ANEA: distant supervision for low-resource named entity recognition," *CoRR*, vol. abs/2102.13129, 2021. [Online]. Available: https://arxiv.org/abs/2102.13129

[11] W. Ali, N. Ali, Y. Dai, J. Kumar, S. Tumrani, and Z. Xu, "Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: Association for Computational Linguistics, Apr. 2021, pp. 188–194. [Online]. Available: https://aclanthology.org/2021.wassa-1.20

[12] G. Singh, Z. Sabet, J. Shawe-Taylor, and J. Thomas, *Constructing Artificial Data for Fine-Tuning for Low-Resource Biomedical Text Tagging with Applications in PICO Annotation*. Cham: Springer International Publishing, 2021, pp. 131–145. [Online]. Available: https://doi.org/10.1007/978-3-030-53352-6_12

[13] X. Li, D. R. Mortensen, F. Metze, and A. W. Black, "Multilingual phonetic dataset for low resource speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6958–6962.

[14] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191–211, 03 2023. [Online]. Available: https://doi.org/10.1162/tacl_a_00542

[15] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. [Online]. Available: https://aclanthology.org/P17-1171

[16] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang, "R^3: Reinforced reader-ranker for open-domain question answering," *CoRR*, vol. abs/1709.00023, 2017. [Online]. Available: http://arxiv.org/abs/1709.00023

[17] S. Min, V. Zhong, R. Socher, and C. Xiong, "Efficient and robust question answering from minimal context over documents," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1725–1735. [Online]. Available: https://aclanthology.org/P18-1160

[18] J. Lee, S. Yun, H. Kim, M. Ko, and J. Kang, "Ranking paragraphs for improving answer recall in open-domain question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 565–569. [Online]. Available: https://aclanthology.org/D18-1053

[19] B. Kratzwald and S. Feuerriegel, "Adaptive document retrieval for deep question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 576–581. [Online]. Available: https://aclanthology.org/D18-1055

[20] B. Kratzwald, S. Feuerriegel, and H. Sun, "Learning a Cost-Effective Annotation Policy for Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3051–3062. [Online]. Available: https://aclanthology.org/2020.emnlp-main.246

[21] Y. Xie, W. Yang, L. Tan, K. Xiong, N. J. Yuan, B. Huai, M. Li, and J. Lin, *Distant Supervision for Multi-Stage Fine-Tuning in Retrieval-Based Question Answering*. New York, NY, USA: Association for Computing Machinery, 2020, p. 2934–2940. [Online]. Available: https://doi.org/10.1145/3366423.3380060

[22] H. A. Pandya and B. S. Bhatt, "Question answering survey: Directions, challenges, datasets, evaluation matrices," *CoRR*, vol. abs/2112.03572, 2021. [Online]. Available: https://arxiv.org/abs/2112.03572

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[24] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2369–2380. [Online]. Available: https://aclanthology.org/D18-1259

[25] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "NewsQA: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 191–200. [Online]. Available: https://aclanthology.org/W17-2623

[26] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *CoRR*, vol. abs/1611.09268, 2016. [Online]. Available: http://arxiv.org/abs/1611.09268

[27] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 08 2019. [Online]. Available: https://doi.org/10.1162/tacl_a_00276

[28] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. [Online]. Available: https://aclanthology.org/Q19-1016

[29] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 874–880. [Online]. Available: https://aclanthology.org/2021.eacl-main.74

[30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[31] X. Cao, Y. Zhao, and B. Shen, "Improving and evaluating complex question answering over knowledge bases by constructing strongly supervised data," *Neural Computing and Applications*, vol. 35, no. 7, pp. 5513–5533, 2023.

[32] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "Zerogen: Efficient zero-shot learning via dataset generation," 2022.

[33] F. Ture and E. Boschee, "Learning to translate for multilingual question answering," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 573–584. [Online]. Available: https://aclanthology.org/D16-1055

[34] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 547–564. [Online]. Available: https://aclanthology.org/2021.naacl-main.46

[35] J. Zhu, Q. Wang, Y. Wang, Y. Zhou, J. Zhang, S. Wang, and C. Zong, "NCLS: Neural cross-lingual summarization," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3054–3064. [Online]. Available: https://aclanthology.org/D19-1302

[36] E. Trandafili, N. Kote, and G. Plepi, "Question classification in albanian through deep learning approaches," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140385

[37] Y.-H. Lin, C.-Y. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig, "Choosing transfer languages for cross-lingual learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3125–3135. [Online]. Available: https://aclanthology.org/P19-1301

[38] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-shot cross-lingual transfer with meta learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4547–4562. [Online]. Available: https://aclanthology.org/2020.emnlp-main.368

[39] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 87–94. [Online]. Available: https://aclanthology.org/2020.acl-demos.12

[40] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. [Online]. Available: https://aclanthology.org/2022.acl-long.62

[41] J. Ma, Q. Chai, J. Huang, J. Liu, Y. You, and Q. Zheng, "Weakly supervised learning for textbook question answering," *IEEE Transactions on Image Processing*, vol. 31, pp. 7378–7388, 2022.

[42] A. F. T. Martins, M. Junczys-Dowmunt, F. N. Kepler, R. Astudillo, C. Hokamp, and R. Grundkiewicz, "Pushing the Limits of Translation Quality Estimation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 205–218, 07 2017. [Online]. Available: https://doi.org/10.1162/tacl_a_00056

[43] J. Zhao, Y. Su, Z. Guan, and H. Sun, "An end-to-end deep framework for answer triggering with a novel group-level objective," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1276–1282. [Online]. Available: https://aclanthology.org/D17-1131

[44] A. Kamath, R. Jia, and P. Liang, "Selective question answering under domain shift," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5684–5696. [Online]. Available: https://aclanthology.org/2020.acl-main.503

[45] P. Petrushkov, S. Khadivi, and E. Matusov, "Learning from chunk-based feedback in neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 326–331. [Online]. Available: https://aclanthology.org/P18-2052

[46] K. Clark, M.-T. Luong, C. D. Manning, and Q. Le, "Semi-supervised sequence modeling with cross-view training," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1914–1925. [Online]. Available: https://aclanthology.org/D18-1217

[47] C. Liang, J. Berant, Q. Le, K. D. Forbus, and N. Lao, "Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 23–33. [Online]. Available: https://aclanthology.org/P17-1003

[48] S. Iyer, I. Konstas, A. Cheung, J. Krishnamurthy, and L. Zettlemoyer, "Learning a neural semantic parser from user feedback," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 963–973. [Online]. Available: https://aclanthology.org/P17-1089

[49] I. Gur, S. Yavuz, Y. Su, and X. Yan, "DialSQL: Dialogue based structured query generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1339–1349. [Online]. Available: https://aclanthology.org/P18-1124

[50] Z. Yao, Y. Su, H. Sun, and W.-t. Yih, "Model-based interactive semantic parsing: A unified framework and a text-to-SQL case study," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5447–5458. [Online]. Available: https://aclanthology.org/D19-1547

[51] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: https://aclanthology.org/P17-1147

[52] B. Kratzwald and S. Feuerriegel, "Learning from on-line user feedback in neural question answering on the web," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 906–916. [Online]. Available: https://doi.org/10.1145/3308558.3313661

[53] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 51–60.

[54] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 595–605. [Online]. Available: https://aclanthology.org/D17-1063

[55] A. Siddhant and Z. C. Lipton, "Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2904–2909. [Online]. Available: https://aclanthology.org/D18-1318

[56] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4623–4637. [Online]. Available: https://aclanthology.org/2020.acl-main.421

[57] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4948–4961. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.445

[58] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra, "Samanantar: The largest publicly available parallel corpora collection for 11 indic languages," 2021.

[59] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7315–7330. [Online]. Available: https://aclanthology.org/2020.acl-main.653