Cross-Lingual Fact Checking: Automated Extraction and Verification of Information from Wikipedia using References

Shivansh Subramanian*, Ankita Maity*, Aakash Jain*, Bhavyajeet Singh*, Harshit Gupta*, Lakshya Khanna* and **Vasudeva Varma**
IIIT Hyderabad
{shivansh.s, ankita.maity, bhavyajeet.singh, harshit.g, lakshya.khanna}@research.iiit.ac.in
aakash.jain@students.iiit.ac.in, vv@iiit.ac.in

Abstract

The paper presents a novel approach for automated cross-lingual fact-checking that extracts and verifies information from Wikipedia using references. The problem involves determining whether a factoid in an article is supported or needs additional citations based on the provided references, with granularity at the fact level. We introduce a cross-lingual manually annotated dataset for fact extraction and verification and an entirely automated pipeline for the task. The proposed solution operates entirely in a cross-lingual setting, where the article text and the references can be in any language. The pipeline integrates several natural language processing techniques to extract the relevant facts from the input sources. The extracted facts are then verified against the references, leveraging the semantic relationships between the facts and the reference sources. Experimental evaluation on a large-scale dataset demonstrates the effectiveness and efficiency of the proposed approach in handling cross-lingual fact-checking tasks. We make our code and data publicly available[1].

1 Introduction

Wikipedia is one of the world's most widely used sources of information, and its articles cover a vast array of topics in many different languages. Hence, the accuracy and reliability of the information on Wikipedia becomes a topic of concern and importance. To maintain the high standard of its articles, all material in Wikipedia must be attributable to a reliable, published source.

While there have been efforts at identifying if the information in a sentence is factually correct or needs a citation, most of these approaches are monolingual and only present for high-resource languages. Furthermore, these solutions work on the granularity of a sentence. Complex sentences from Wikipedia articles can be made up of multiple facts. In such cases, the correctness of each of these facts can be more helpful than the correctness of the sentence as a whole. For this, we need to have specific information about the availability of citations for each fact. Thus, it becomes necessary first to extract factual information from the sentences and then predict the label for each of those. Figure 1 shows an example of the XFactVer problem.

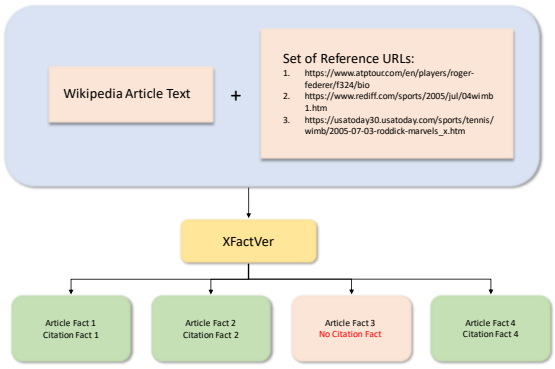

Figure 1: Example of the cross-lingual fact extraction and verification problem.

The pipeline for cross-lingually extracting factual information can also be used for multiple purposes, like automatically populating knowledge graphs such as Wikidata or utilising natural language text from multiple sources to create a common knowledge graph. Once the facts are extracted, we pass each of the facts along with semantically selected sentences from the reference through a classifier pipeline, which predicts if the citations support the fact or if the fact is in need of further citation. Such a pipeline can be used for automatically citing text on the low-resource editions of Wikipedia and reducing the manual efforts needed

*Equal contribution.
[1]https://github.com/TheAthleticCoder/Cross-Lingual-Fact-Checking

to identify sentences needing citations. This becomes particularly important for the low-resource versions of Wikipedia, which have a lower quality of articles and fewer editors.

Thus, the major contributions of this paper include:

- A cross-lingual dataset for fact extraction and verification, covering English and five Indian languages.
- A pipeline for automated cross-lingual fact extraction and verification, with the granularity at the fact level instead of the sentence level.

To the best of our knowledge, this is the first attempt at solving this task.

2 Related Work

In this section, we discuss related work on the two stages of our approach - fact extraction and verification.

2.1 Fact Extraction

Structured fact extraction from unstructured textual data is a widely studied problem. Two Indian languages - Hindi and Telugu have been covered in a prior work (Kolluru et al., 2022). We extend this to four other Indian languages while avoiding translation. Our work is most similar to Singh et al. (2022) - we experiment with other fact extraction methods and extend their work to include verification as well.

2.2 Fact Verification

Prior work on fact verification has centred around the FEVER benchmark (Pan et al., 2021; Krishna et al., 2022). Most prior work on fact verification is monolingual and works on sentence level instead of fact level (Subramanian and Lee, 2020; Huang et al., 2022). Individual facts can be added to a knowledge graph (Nadgeri et al., 2021), and then various knowledge graph comparison methods can be used for comparing and verifying facts across the two graphs (Zhu et al., 2020; Mondal et al., 2021; Chen et al., 2022).

3 Dataset

We construct the XFactVer dataset using two existing datasets, the XAlign (Abhishek et al., 2022) and the XWikiRef (Taunk et al., 2023) datasets. The XAlign dataset contains sentences from Indian language Wikipedia articles from the persons domain along with aligned facts from Wikidata. The XWikiRef dataset contains articles in Indian languages along with text from their references. We extract the intersection of these two datasets by getting the entities which are present in both XAlign and XWikiRef. We extract the top 10 sentences from the reference text for all the article sentences in the dataset. We do this by checking semantic similarity between (article text, reference text) as (question, answer) pairs [2]. In order to construct the golden test data for every sentence, we manually annotate each fact. The manual annotators provide two possible labels - either the fact is supported with respect to the reference sentences, or it isn't supported. Using this approach, we construct the XFactVer dataset. The constructed golden test dataset contains a sentence from the Indian language Wikipedia article, context from citations, manually aligned facts, and a manually annotated label. Figure 2 describes the components of the dataset.

Language	Articles	Sentences	Facts
Bengali	11,468	53,522	106,165
Odia	1,635	7,601	13,035
English	4,715	17,326	39,540
Punjabi	3,491	12,324	25,758
Tamil	6,003	21,937	38,100
Hindi	5,796	20,277	40,062
Total	**33,108**	**132,987**	**262,660**

Table 1: Dataset statistics for each of the languages.

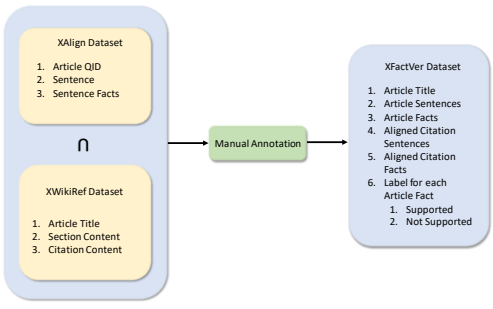

Figure 2: Components of the XFactVer dataset.

[2] https://huggingface.co/SeyedAli/Multilingual-Text-Semantic-Search-Siamese-BERT-V1

4 Methods

The automated pipeline for fact-level verification is constructed in the following phases. Figure 3 gives a diagrammatic overview of this process.

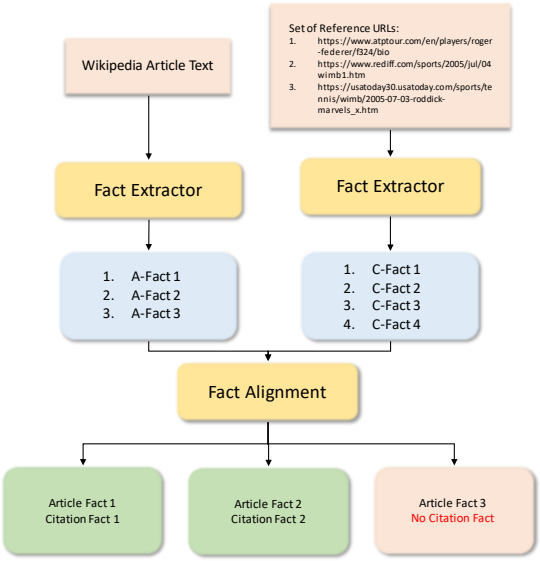

Figure 3: Pipeline for automated fact extraction and verification.

4.1 Fact Extraction

The task of cross-lingual fact extraction (CLFE) involves extracting English facts from the natural language text of multiple low-resource Indian languages. For this task, we propose two methods. The first approach formulates this problem as a text-to-text task and finetunes a pre-trained mT5 model for extracting the English facts (Singh et al., 2022). For our second method, to check the viability of LLMs for this task, we prompt GPT-4 to extract facts in English from the multilingual sentences.

4.2 Fact Verification

Our best-performing fact verification approach utilizes passing each fact along with the corresponding reference sentences from a classifier. Since the reference text can be very long, we tokenize the reference text into sentences and then use LABSE (Feng et al., 2022) to find the top 10 semantically similar sentences from the reference to the article sentence.

	mT5-small		GPT-4	
Metric	**ROUGE-L**	**BERTScore**	**ROUGE-L**	**BERTScore**
bn	0.838	0.890	**0.902**	**0.954**
or	**0.711**	**0.860**	0.600	0.822
en	**0.768**	**0.883**	0.656	0.868
pa	**0.692**	**0.865**	0.601	0.847
ta	**0.842**	**0.924**	0.766	0.902
hi	**0.854**	**0.932**	0.596	0.833
avg	**0.784**	**0.893**	0.687	0.871

Table 2: Language-wise fact extraction results.

	bn	**or**	**en**	**pa**	**ta**	**hi**	**avg**
Accuracy	66.59	**70.52**	61.90	60.39	66.43	57.76	63.93

Table 3: Language-wise fact verification results.

4.3 Implementation Details

For fact extraction using mT5, we use the mT5-small model having 8 encoder and 8 decoder layers. We use Adam optimizer with a learning rate of 2e-5 and train the model for 10 epochs with a batch size of 4. For fact verification, in particular, the threshold to determine semantic similarity was kept at 0.7.

5 Results

5.1 Fact Extraction

The results for fact extraction are shown in Table **??**. We observe that ROUGE-L and BERTScore correlate well, and thus, either metric can be used to find our best-performing model. Other than Bengali, mT5-small outperforms GPT-4 in all the languages, with the best results observed for Hindi. Thus, fine-tuning a much smaller model outperforms a SOTA model trained with few shot prompting, even for English.

5.2 Fact Verification

For fact verification, as shown in Table 3, our proposed system achieves an average accuracy of 63.93%. It can be observed that the system does not suffer from a language divide. Even extremely low-resource languages like Odia and Punjabi perform very close to or higher than the average, while higher-resource languages like English perform worse than average.

6 Conclusion and Future Work

In this work, we proposed the task of cross-lingual fact extraction and verification and contributed relevant baselines for the same. We also contribute a manually annotated golden test set for verifying our

pipeline or those devised in future. Surprisingly, we find lower-resource Indian languages to perform comparably, or in a few cases, even better than English, all without relying on translation. Further work can utilize Set Transformers (Lee et al., 2019) to augment the mT5-based generation, which is particularly useful in our case since the extracted facts are permutation invariant.

References

Tushar Abhishek, Shivprasad Sagare, Bhavyajeet Singh, Anubhav Sharma, Manish Gupta, and Vasudeva Varma. 2022. Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 171–175, New York, NY, USA. Association for Computing Machinery.

Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. 2022. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 927–937, New York, NY, USA. Association for Computing Machinery.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895, Online. Association for Computational Linguistics.

Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang', Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. 2021. KGPool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Bhavyajeet Singh, Siri Venkata Pavan Kumar Kandru, Anubhav Sharma, and Vasudeva Varma. 2022. Massively multilingual language models for cross lingual fact extraction from low resource Indian languages. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 11–18, New Delhi, India. Association for Computational Linguistics.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical Evidence Set Modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7798–7809, Online. Association for Computational Linguistics.

Dhaval Taunk, Shivprasad Sagare, Anupam Patil, Shivansh Subramanian, Manish Gupta, and Vasudeva Varma. 2023. Xwikigen: Cross-lingual summarization for encyclopedic text generation in low resource languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1703–1713, New York, NY, USA. Association for Computing Machinery.

Qi Zhu, Hao Wei, Bunyamin Sisman, Da Zheng, Christos Faloutsos, Xin Luna Dong, and Jiawei Han. 2020. Collective multi-type entity alignment between knowledge graphs. In *Proceedings of The Web Conference 2020*, WWW '20, page 2241–2252, New York, NY, USA. Association for Computing Machinery.