Bridging the Gap Between Wikipedians and Scientists with Terminology-Aware Translation: A Case Study in Turkish

Gözde Gül Şahin (PI)
Department of Computer Engineering, Koç University, İstanbul, Türkiye

Abstract

This project addresses the gap between the escalating volume of English-to-Turkish Wikipedia translations and the insufficient number of contributors, particularly in technical domains. Leveraging expertise from academics' collaborative terminology dictionary effort, we propose a pipeline system to enhance translation quality. Our focus is on bridging academic and Wikipedia communities, creating datasets, and developing NLP models for terminology identification and retrieval, and terminology-aware translation. The aim is to foster sustained contributions and improve the overall quality of Turkish Wikipedia articles.

Introduction

According to the most recent dump of *contenttranslation*[1], (editor tool for automatic translation) 418,000 short paragraphs are translated from English to Turkish, followed by 10,000 translated from German. The volume of articles is increasing significantly, but the number of active Turkish Wikipedia contributors remains insufficient to keep pace. This poses a particular concern for articles demanding specialized domain knowledge, especially those featuring technical and scientific content laden with rigorous terminology.

On the other hand, Turkish Academy of Sciences (TÜBA) has been supporting a collaborative effort among 135 Turkish academics (list is still growing) that provide expert translations for scientific terms in a wide range of topics including engineering, biology and chemistry. This dictionary, terimlor.org, has been maintained for an impressive 49 years now. We hypothesize that bridging these two communities will significantly enhance the quality of Turkish Wikipedia articles, fostering sustained contributions from academics to expand and maintain the dictionary, as demonstrated in Figure 1.

Here, we aim to create a pipeline system that: i) automatically identifies scientific and technical terms, ii) consults an expert dictionary for accurate translations, and iii) suggests automatic content rewriting with the translations. Additionally, the system will help identify terms lacking translations, informing the expansion of the dictionary.

We aim to address three key research questions:
(RQ1) Community: Strategies for integrating domain experts with Wikipedians, aiming to recruit domain experts as contributors.
(RQ2) Data: Development of datasets for training and evaluating NLP models targeted at i) term

[1] https://dumps.wikimedia.org/other/contenttranslation/20230908/

identification, ii) term sense detection, and iii) terminology-aware translation.
(RQ3) Model: Designing and implementing Turkish language-capable NLP models for the specified tasks.

Date: June 1, 2024 - May 31, 2025.

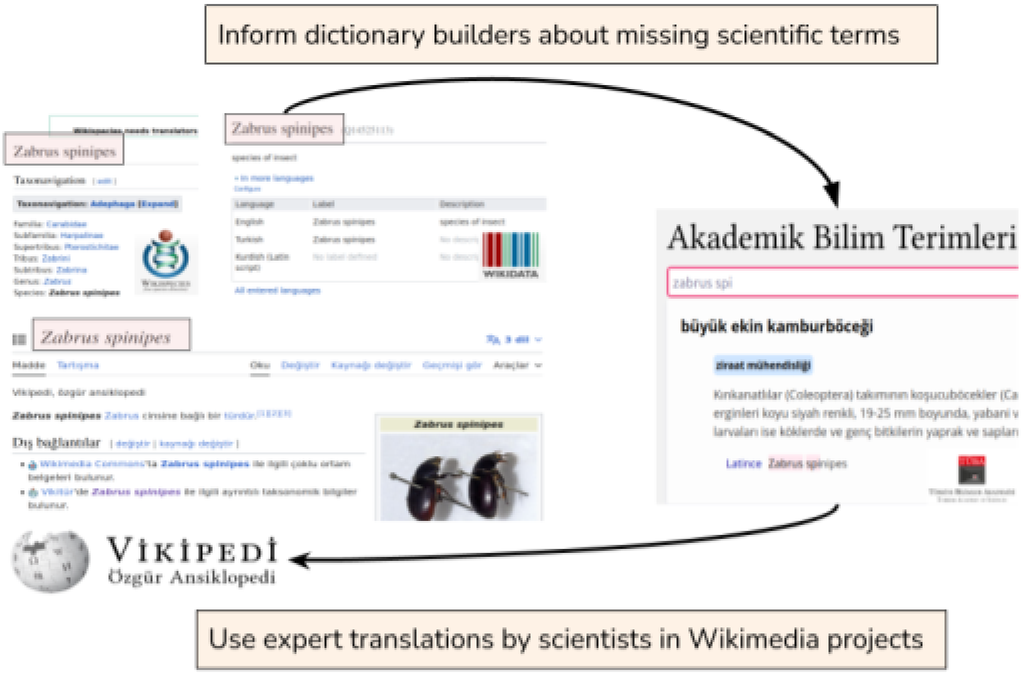

Figure 1: Left: Vikipedi (Turkish Wikipedia), Wikidata and Wikispecies pages of insect '*Zabrus Spinipes*'. Right: terimler.org page that contains the correct Turkish translation '*büyük ekin kamburböceği*'. Note that the current draft only targets Wikipedia articles, however, can later be extended to other Wikimedia projects

Related work

This work aligns with the emerging field[2] of terminology-aware translation, highlighted by a recent WMT23 shared task [1]. While prior efforts focus on Chinese, English, Czech, and German, and assume access to MT model weights, our approach differs by i) concentrating on English to Turkish, introducing additional challenges with complex morphology, and ii) treating the MT model as a black box for a more realistic real-world scenario. Unlike the 21 proposed approaches, we decompose the terminology-aware translation problem into distinct stages: *term identification, term sense detection, and rewriting with lexical constraints*, instead of pursuing end-to-end translation. Leveraging insights from our previous work on grammar rule-aware text correction[2], we posit that this modular approach will yield superior results and provide a reusable term identification model.

Methods

Task 1: Term identification can be formulated as named entity recognition (NER). We can easily build a synthetic, NER dataset, automatically annotating existing terms (e.g., *Zabrus Spinipes*) in the latest *contenttranslation*[3] dump. We can then fine-tune a small pretrained

[2] The shared task received 21 submissions

[3] https://dumps.wikimedia.org/other/contenttranslation/20230908/

LM (e.g., BERTurk, mGPT) for the span detection task.

Task 2: Term sense detection can be approached as a retrieval task, utilizing efficient tools like FAISS[4] to index the dictionary and the contextual term. We propose synthesizing a retrieval dataset using this method, with subsequent human annotation for quality assurance. If agreement between human annotators and the proposed approach is low, manual annotation will be exclusively employed.

Task 3: Rewrite with lexical constraint
The task involves replacing detected terms in a short text with their correct translation. For morphologically rich languages, like Turkish, simple Find/Replace is inadequate; it requires preserving morphology. For example, translating *'Zabrus Spinipes*lerin hayatı' (the life of Zabrus Spinipes), into *'büyük ekin kambur böceklerinin hayatı'* entails analyzing and inflecting the scientific term's lemma with appropriate morphological features, such as Plural+Possessive. This process, termed **reinflection**, can be approached through various methods, ranging from rule-based, such as using external morphological tools, to data-driven, such as creating synthetic data and training Transformer-based models.

Task 4: Build a communication channel between the communities Deploy and offer API access for the developed models in the Wikipedia content editor. Host a collaborative Wiki event inviting academic contributors from terimler.org and Wikipedians. Conduct an editing marathon for both groups, assessing the System Usability Score (SUS) post-event.

Expected output

- Editor tool: Enables **Wikipedians** to auto-edit text with accurate Turkish scientific terminology. Highlights terms not in the dictionary, providing automatic feedback suggestions to the **scientific community**.
- Public datasets and models: **NLP researchers** can use them to train/evaluate/compare their own models.
- Wiki Event: Engage Turkish scientists and Wikipedians in a public presentation to introduce the tool and gather feedback.
- Scientific publication at a top-tier NLP venue (e.g., *CL, EMNLP or CL, TACL journal)

Risks

We anticipate that *contenttranslation* will not translate terms (e.g., Zabrus Spinipes), leaving them unchanged by the editor. We will expand the proposed methods to include source text if necessary. Additionally, errors in the initial stages may be irrecoverable.

Community impact plan

PI plans to work with Başak Tosun and Zafer Batık (Wikimedia volunteer editors and organizers) and Bülent Sankur (main contact for terimler.org) to organize WikiEvent.

Evaluation

Task 1: Accuracy, Task 2: R@n (percentage of the ground-truth term being in the top-n), Task 3: Exact Match, Task 4: SUS.

Budget

Conference participation, 2K$
Event organization, 2K$

[4] https://github.com/facebookresearch/faiss

Cloud services 2K$
PhD Salary 18K$
PI Salary (Part-time) 12K$
Community Staff (Event organization, Feedback Collection) 2K$
Engineer for tool development (3 months) - 6K$
Organizational overhead - 4K$
Annotation - 3K$
Total: ~40K-50K

Prior contributions

PI has co-organized the Multilingual Representation Learning (MRL) Workshop at EMNLP for the last 3 years in a row.
PI has significant experience in all tasks: multilingual NER[3] and similar span detection tasks (e.g., extractive QA[4,5]), retrieval[4,6], reinflection [7] and tool building [6, 8, GECTurk-WEB[5]].

References

[1] Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In Proceedings of the Eighth Conference on Machine Translation, pages 663–671, Singapore. Association for Computational Linguistics.

[2] GECTurk: Grammatical Error Correction and Detection Dataset for Turkish
Atakan Kara, Farrin Marouf Sofian, Andrew Bond, and Gözde Şahin
In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Nov 2023

[3] Şahin, Gözde Gül, et al. "LINSPECTOR: Multilingual probing tasks for word representations." *Computational Linguistics* 46.2 (2020): 335-385.

[4] Uzunoğlu, Arda, and Gözde Gül Şahin. "Benchmarking Procedural Language Understanding for Low-Resource Languages: A Case Study on Turkish." In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Nov 2023

[5] Puerto, Haritz, Gözde Şahin, and Iryna Gurevych. "MetaQA: Combining Expert Agents for Multi-Skill Question Answering." *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023.

[6] Baumgärtner, Tim, et al. "UKP-SQUARE: An Online Platform for Question Answering Research." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2022.

[7] Acikgoz, Emre Can, et al. "Transformers on Multilingual Clause-Level Morphology." *MRL 2022* (2022): 100.

[8] Eichler, Max, Gözde Gül Sahin, and Iryna Gurevych. "LINSPECTOR WEB: A Multilingual Probing Suite for Word Representations." *EMNLP-IJCNLP 2019* (2019): 127.

[5] https://www.gecturk.net/