Wiki Workshop (11th edition) – June 20, 2024

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

Ali Gebeşçe* **Gürkan Soykan*** **Gözde Gül Şahin**
Department of Computer Engineering, Koç University, İstanbul, Türkiye
KUIS AI, Koç University, İstanbul, Türkiye
https://gglab-ku.github.io/

Keywords: Natural Language Processing (NLP), Terminology-Aware Translation, Terminology Extraction and Linking, Community-Driven Content Development, English-Turkish Machine Translation

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, terimler.org terminology dictionary, 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. 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.

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 automated terminology-aware translation. Additionally, the system will help identify terms lacking translations, informing the expansion of the dictionary.

We aim to address three key research objectives:

Objective 1 Community: Strategies for integrating domain experts with Wikipedians, aiming to recruit domain experts as contributors and train existing/new ones to translate technical content more accurately.

Objective 2 Data: Development of datasets for training and evaluating NLP models targeted at (i) term identification, (ii) term linking, and (iii) terminology-aware translation.

Objective 3 Model: Designing and implementing Turkish language-capable NLP models for the specified tasks.

Methods

Preliminary Work

In our feasibility study, we addressed the significance of terminology-aware translation for Wikipedia and the proficiency of current tools like automatic machine translation and ChatGPT in this context. Analyzing 53,162 terms from *terimler.org*, we found 67.5% lacked dedicated Wikipedia pages in English or Turkish, and 20% existed only in English. Our evaluation of bilingual pages for the remaining 12,5% terms revealed a mismatch in 1,063 out of 2,927 terms with expert-curated definitions, indicating a need for extensive translation and alignment.

Our examination of the *contenttranslation* dump highlighted the limitations of existing tools in ensuring terminological accuracy. Despite reviewing only 217 entries containing terminological phrases, we identified numerous inaccuracies in both machine-generated and human-edited translations, particularly in the translation of scientific content, see Table 1 for a sample entry. While human translations aligned most closely with our terminological database, all methods struggled with scientific terminology, underscoring the need for improved translation practices in this domain (see Figure 2).

Proposed Research

We will first create a dataset that can be used for all three tasks: term identification, term linking, and terminology-aware translation, then iteratively build models and seek feedback from the community as explained below (see Figure 1).

Data Curation and Annotation Protocol

We aim to create a corpus of aligned English-Turkish sentences using the *contenttranslation* dump and abstracts of theses from the Council of Higher Education. These resources provide a vast amount of data, which will be refined using existing alignment tools (Steingr'imsson et al., 2023). Our goal is to generate 3,000 parallel sentences annotated with technical terms, linked to the correct database entries, and featuring accurately translated terms in Turkish.

*These authors contributed equally to this work.

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

We plan to predetermine the scientific keyphrases in the source text with existing tools (Ferragina and Scaiella, 2010), identify the URLs for the terms, and insert hyperlinks, if any. Given the identified and linked terms, we will first ask the annotators to check whether the links and terms are correct. Next, they will be asked to post-edit the human translations following the term URL. Three annotators will check each parallel entry, and the final results will be aggregated.

Building NLP Models

We plan to build separate models for each task, with the potential of exploring joint models if time permits. (i) **Term Identification:** We aim to create a system compatible with any terminology database, initially testing state-of-the-art methods for term identification. We may finetune a smaller pretrained model focusing on span detection in both English and Turkish, depending on performance. Our approach will include the development of both multilingual and monolingual models. (ii) **Term Linking:** We will treat term linking as a retrieval task, employing tools like FAISS [2] to index terms and their contextual embeddings. The process includes querying the database beforehand to identify unlinkable terms, aiding domain experts in recognizing gaps. (iii) **Terminology-aware Translation** As shown in Table 1, there were cases where the terms were left unchanged (e.g., hypernucleus). However, MT-engine mostly caused mistranslations, drastically altering the term. Therefore, we realized the need for incorporating the source text into the models and designed the annotation protocol accordingly. We plan to do that by replacing technical terms with placeholders and translating the text with an existing MT model. Then we will build a contextualized reinflection model. The model will need to perform reinflection with missing morphological tags and missing word order information. Alternatively, we plan to experiment with concatenating the terminological constraints to the input in various ways (e.g., lemmatized, surface, etc.).

The scalability of our research to other languages depends on access to terminology databases and the ability to extend NLP models (like mGPT, mT5, mBART) to additional languages with little extra effort, despite the challenge of creating large, quality parallel corpora. We aim to leverage existing resources and community events like Wikimania 2024 to facilitate dataset curation and annotation for diverse languages.

Building a Communication Channel between the Communities

We aim to foster collaboration between volunteers at the *terimler.org* terminology database and the Wikimedia Community User Group Turkey, targeting a synergy between academic and Wikipedia contributors. An introductory online seminar will set the stage for this partnership, followed by a series of user studies involving active members from both groups to evaluate and refine our NLP models through practical translation tasks. Feedback from these sessions will guide the development of training materials designed to enhance the accuracy of technical term translations on Wikipedia. This collaborative approach will be continuously assessed through surveys, e.g., system usability score (SUS), with the ultimate goal of integrating the developed models into the Wikipedia editing process, thus enriching the content quality and expanding the editor's expertise.

Expected output

With this project, we aim to produce public datasets and models, facilitating training, evaluation, and comparison by the NLP research community. Two public seminars aim to connect Turkish scientists with Wikipedians, showcasing the developed models and integrating feedback. Regular office hours will further this dialogue, discussing findings from interviews and user studies. Additionally, guidelines will be created to aid Wikipedia editors in accurate technical term translation using the developed models. The culmination of these efforts will be a scientific publication in a top-tier NLP venue, such as ACL, EMNLP, or the TACL journal.

References

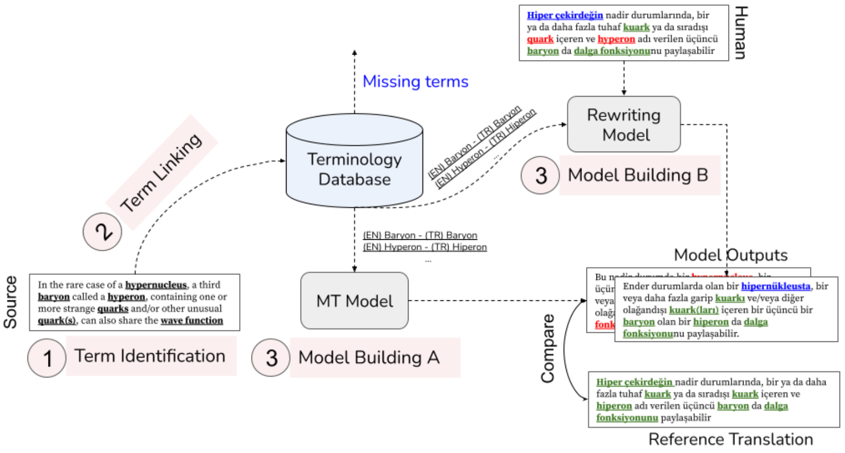
[Ferragina and Scaiella2010] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). *Proceedings of the 19th ACM international conference on Information and knowledge management*.

[Steingr'imsson et al.2023] Steinth'or Steingr'imsson, Hrafn Loftsson, and Andy Way. 2023. Sentalign: Accurate and scalable sentence alignment. *ArXiv*, abs/2311.08982.

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

Source Text (Trans. ID:154396)	In the rare case of a **hypernucleus**, a third **baryon** called a **hyperon**, containing one or more strange **quarks** and/or other unusual **quark(s)**, can also share the **wave function**.
MT (Yandex)	Bu nadir durumda bir **hypernucleus**, bir üçüncü **baryon** denilen bir **hyperoniçeren** bir veya daha fazla garip **kuarklar** ve/veya diğer olağandışı **kuark(s)**, ayrıca **paylaşım dalga fonksiyonu**
ChatGPT	Ender durumlarda olan bir **hipernükleusta**, bir veya daha fazla garip **kuarkı** ve/veya diğer olağandışı **kuark(ları)** içeren bir üçüncü bir **baryon** olan bir **hiperon** da **dalga fonksiyonunu** paylaşabilir.
Post-Edit	**Hiper çekirdeğin** nadir durumlarında, bir ya da daha fazla tuhaf **kuark** ya da sıradışı **quark** içeren ve **hyperon** adı verilen üçüncü **baryon** da **dalga fonksiyonunu** paylaşabilir
Terminology Database	**(EN) Hypernucleus - (TR) Missing Term** (EN) Baryon - (TR) Baryon (EN) Hyperon - (TR) Hiperon (EN) Quark - (TR) Kuark (EN) Wave Function - (TR) Dalga Fonksiyonu

Table 1: Annotated samples. **Blue:** Missing, **Red:** Wrong, **Green:** Correct



Hiper çekirdeğin nadir durumlarında, bir ya da daha fazla tuhaf **kuark** ya da sıradışı **quark** içeren ve **hyperon** adı verilen üçüncü **baryon** da **dalga fonksiyonu**nu paylaşabilir

In the rare case of a **hypernucleus**, a third **baryon** called a **hyperon**, containing one or more strange **quarks** and/or other unusual **quark(s)**, can also share the **wave function**

Ender durumlarda olan bir **hipernükleusta**, bir veya daha fazla garip **kuarkı** ve/veya diğer olağandışı **kuark(ları)** içeren bir üçüncü bir **baryon** olan bir **hiperon** da **dalga fonksiyonu**nu paylaşabilir.

Hiper çekirdeğin nadir durumlarında, bir ya da daha fazla tuhaf **kuark** ya da sıradışı **kuark** içeren ve **hiperon** adı verilen üçüncü **baryon** da **dalga fonksiyonunu** paylaşabilir

Figure 1: Technical overview of our proposed research

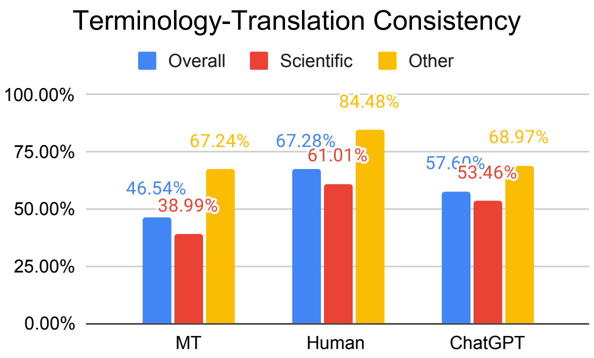


Figure 2: Preliminary investigation of terminology translation success rates