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 linking, 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, [terimler.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.

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 questions:
(RQ1) 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.
(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 linking, 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.

Related work

Entity Linking

Most similar work to term identification and linking is the task of *entity linking,* which aims to identify phrases that mention an entity, such as *"Barack Obama",* and link them to a unique entry in the target knowledge base. Majority of previous datasets/models consider broad coverage knowledge bases such as Wikipedia (referred to as wikification), YAGO, Freebase, BabelNet[1] (referred to as Babelfy[2][2]); and several other works [3,4] provide benchmarks/shared tasks to bring those together.

Unfortunately, entity linking research for the science domain is rather limited. There are two main categories: linking entities in scientific text to i) Wikipedia URLs (or a similar broad-coverage KB) or ii) to domain-specific KB. In the first category, [5] introduce a technique that leverage author-citation networks to link entities in scholarly abstracts to their associated Wikipedia pages, whereas [6] leverage wikilinks to populate the references section of scientific Wikipedia pages. More similar to ours, [7] focus on scientific (i.e., STEM) entity extraction. However they consider text only from scholarly abstracts, and use Wikipedia and Wiktionary as the KB, which might lack domain expertise as shown in our preliminary analysis.

In the second category, researchers [8,9,10] create a large-scale annotated corpora of domain-specific entities (biomedical, chemical) on domain-specific text such radiology reports and perform linking on domain-specific ontologies such as UMLS. Although related, our focus is on a broader definition of science, and more accessible content (e.g., Wikipedia articles) rather than domain-expert accessible text. There are also studies [11,12,13] that tackle scientific keyphrase extraction, however they focus on scholarly text and ignore the linking process. It is also worth noting that the majority of aforementioned works focus on English—with some exceptions such as BabelNet.

Terminology-aware Translation

The proposed work also aligns with the emerging field[3] of terminology-aware translation, highlighted by two recent WMT shared tasks [14,15]. In the 2021 shared task, organizers focus on COVID-19 domain and annotate 5 language pairs (En->Fr, En->Chinese, En->Ru, En->Korean and Czech->German) using the TICO-19 terminology database[16]-limited to 600 terms per language. For the Czech->German pair, they automatically generate the terminology from Wikipedia, which might again be ill-defined. Final development and test data splits per language is 3,5K sentences for development and 1,1K for evaluation. The 2023 Shared Task redefine the task to better distinguish the terminology-injection ability by introducing one real "hint" for the actual terminology, and another random "hint". They also change the focus to more general domains (e.g., web novels) and scholarly abstracts, reducing the number of language pairs to three (German->English, English->Czech and Chinese->English), and dataset size to 3K sentences per language pair. Unlike 2021, they don't use predefined terminology translations;

[2] http://babelfy.org/

[3] The 2021 and 2023 shared tasks received 43 and 21 submissions accordingly.

id	Turkish Term	Category	Turkish Definition	German	English	French
5872	beton karışımı	Civil Eng.	*"Cement, sand, gravel or crushed stone, water and additives mixed together in certain proportions to form concrete."*	Betonmischung	concrete mixture	mélange du béton

Table 1: A sample terminology entry for "concrete mixture" from terimler.org. The English definition is provided for convenience only (original is Turkish).

but semi-automatically *extract* it from aligned scientific abstracts.

There are three categories of models that inject constraints to the translation process: i) post-processing ii) constrained decoding and iii) weakly supervised learning of the constraints. First set of techniques [17,18] use placeholders for named entities, numbers and markups and perform post-processing on the placeholders. Second category[19,20] mostly manipulate the beam search decoding process leveraging attention weights and specialized FSM, however, with the cost of slow inference and less fluent output. Final category[21,22] is the most straightforward. They propose ways to provide the constraints (e.g., surface, lemmatized) along with the input (e.g., by appending to the beginning/end) and further train the model; or augment the training data with the constraints[23]. Most recent techniques [24,25] follow the same approaches however also utilize LLMs for various steps e.g., to *post-process the output, generate synthetic data given with constraints* and to *perform constrained decoding*. This will be the first work to tackle terminology-aware translation together with entity-linking; and in the English->Turkish direction.

Methods

Preliminary Work

We first perform a feasibility study to answer two questions:

Q1) How important is the problem of terminology-aware translation for Wikipedia?
Q2) How good are existing tools (e.g., automatic machine translation, ChatGPT etc...) in terminology-aware translation for the given problem?

How severe is the problem?
To answer this question, we use the proposed terminology database, terimler.org, and developed a script to query the available terms for all available metadata. One example dictionary entry is given in Table 1.

Here, one can easily access the expert-curated translations for Turkish, German and French[4], along with additional information such as the scientific category and Turkish definition of the term. Using the developed scripts, we have successfully retrieved 53,162 entries with 47,823 being unique. We have then asked the following question: "***What is the coverage rate of English and Turkish Wikipedia for the retrieved terms?***"

Since this is a preliminary study, we have done several simplifying assumptions. We only considered the terms with a one-to-one English-Turkish correspondence (i.e., ignored the synonyms, terms with multiple senses, and other languages). We considered the term as covered if there is an exact match between the term and Wikipedia article title. As given in Figure 2, a striking number of terms (67,5% of all terms) do not have a dedicated English or

[4] For a large number of terms, Latin names are also provided.

Turkish Wikipedia page. Furthermore, 20% (around 5,000) of the terms have only an English Wikipedia page. We have then analyzed the bilingual pages (12,5% of all terms), to check if the terms in translated pages match with their expert-curated counterparts, and found that **1,063** out of **2,927** do not match. Our results suggest that, i) there is a large number of potential English pages to translate, and ii) considerably large amount of translations to align with expert-curated terminology.

During our explanatory analysis, we have also performed manual checks for mismatched terms, and encountered several cases where the terminology database lacks one of the correct translations. This supports our initial hypothesis that this research has potential to provide benefit to the scientists, i.e., the database creators.

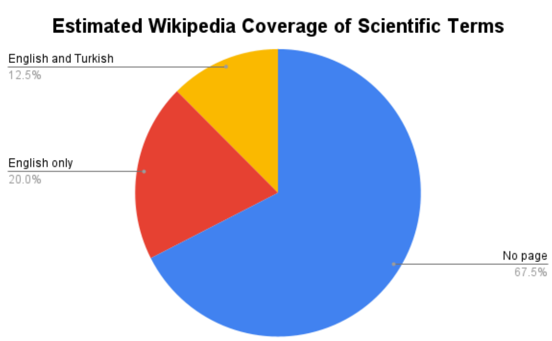


Figure 2: Percentage of Wikipedia pages that cover the retrieved terms

"How good are existing tools?"
To answer this question, we parsed the recent *contenttranslation* dump for English to Turkish, that contains the original English source, machine translation output, and human post-edits. We have then randomly picked around 1000 documents and manually annotated them as *scientific* or *non-scientific* for preliminary analysis purposes. We categorized them as scientific if the topic is related to any of the STEM fields (e.g., biology, physics, chemistry, computer science etc…); and non-scientific otherwise. We filtered out the documents with empty machine translation or post-edit fields, which eliminated around half. We have then manually searched the English source content for terminological phrases and created a small parallel corpora with 217 entries. To answer the initial question, we have prompted *ChatGPT* to generate a translation *like a Wikipedia expert*. As given in Table 2, we have observed erroneous cases both in automatically generated content and the human post-edits. We additionally detected several missing terms in our database. For the sake of simplicity, we have only checked for terminological consistency, however, we have also encountered many low-quality automatic generations both from MT and ChatGPT (wrong semantics, ungrammatical etc…) even when the term was correctly translated[5]. Our preliminary results are given in Figure 3. Even though humans have the highest consistency with our terminology database, all models suffered in scientific documents more.

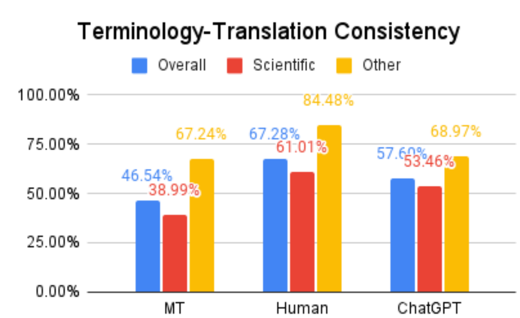


Figure 3: Preliminary investigation of terminology translation success rates

[5] Since the search, annotation, evaluation and corpora creation processes will be refined later, we avoid giving the details at this stage.

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 fonksiyonu**nu 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 fonksiyonu**nu 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 2: Annotated samples. **Blue:** Missing, **Red:** Wrong, **Green:** Correct

Proposed Research

Technical overview of our proposed research is given in the Figure (next page). 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.

Data Curation

We will rely on two resources: *contenttranslation* and abstracts of theses from the Council of Higher Education (CoHE) website. *Contenttranslation* is dumped bi-weekly and contains around 400,000 entries; whereas CoHE contains all abstracts of theses submitted to any Turkish university from 2006 until today—containing more than one million sentences. Since the translations are mostly provided as long paragraphs in both resources, we will first align the sentences with highly-accurate existing tools such as SentAlign[26], and VecAlign[27]. Then we will perform simple preprocessing, e.g., filtering URLs, special characters, short/long sentences, and content with mostly numbers. Next, following [28], we aim to generate 3,000 parallel sentences in English-Turkish containing the following: i) English text annotated with the technical terms, ii) links to correct terminology entries in the database, and iii) edited translations using the correct terminology with Turkish terms, similar to:

{
  source_text: ...[EN-Term$_1$].. [EN-Term$_2$],
  links: [None],[TermURL$_2$],
  target_text:...[TR-Term$_1$]..[TR-Term$_2$]
}

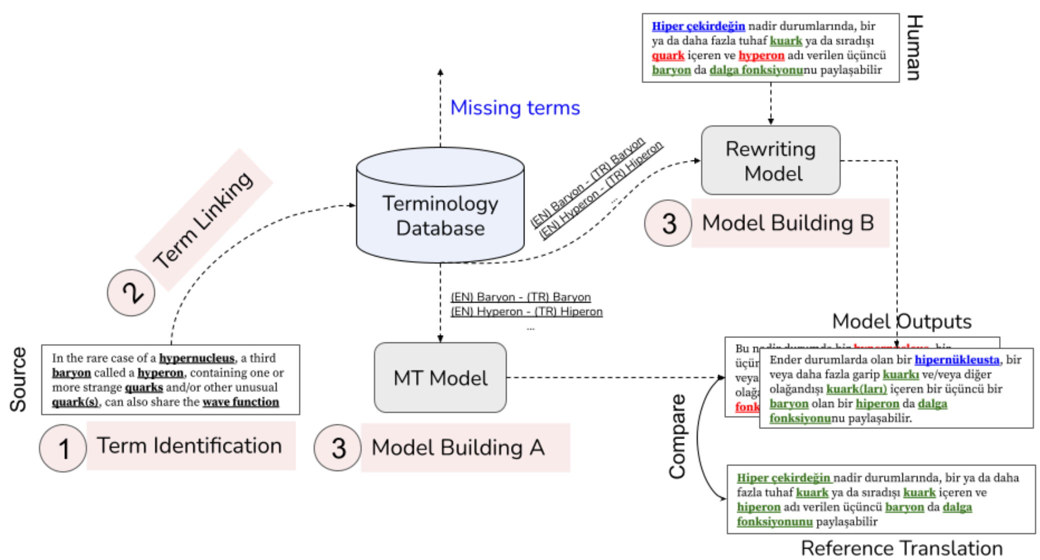
Tentative Annotation Protocol

To reduce the annotation costs, we plan to predetermine the scientific keyphrases in the source text with existing tools such as TagMe[29], AIDA[30], or babelfy[6]. Similarly, we plan to use a simple search code to identify the URLs for the terms, and insert hyperlinks if any. At this stage, we can estimate whether the identified terms exist in the DB, or have more than one entry. To ensure that we include enough terms with multiple senses or without URLs, we will sample sentences accordingly. Given the identified and linked terms, we will

[6] Our preliminary investigation suggests that they provide too many candidates, however, candidates are more accurate for English text

first ask the annotators to check whether links and terms are correct—correct them if not. Next, they will be asked to post-edit the human translations following the term URL. Each parallel entry will be checked by 3 annotators and final results will be aggregated by graduate students. We will calculate the inter-annotator agreement among the participants and against the experts. We will employ additional quality checks (e.g., quiz before entry, test at regular intervals, remove underperformers etc...).

We will advertise the annotation task among college students in Turkiye, who are fluent in English. Note that domain-expertise is desired but not crucial for this task, since terminology will already be provided to annotators. All participants will be compensated fairly (see the budget). The expected outcome is a high-quality dataset containing 3,000 English-Turkish parallel sentences, that can be used for terminology identification, terminology linking and terminology-aware translation.



Building NLP Models

Next, we plan to build models for each of the tasks separately. If time allows, we will also investigate building joint models for comparison.

Term Identification

To develop a flexible system, we consider this as a separate task that can later be integrated with any terminology database. We first plan to split the curated dataset into development and test; and explore the capacity of existing state-of-the-art methods[31]on evaluation split. If the performance is below desired threshold, we will use the development split to further finetune a small pretrained model for the span detection task. Since the terminology spans will also be detected for the Turkish text, we will explore building both multilingual and monolingual identification models drawing upon our previous works[32,33,34].

Evaluation: We will calculate the F1 score considering only the exact matches on the *lemmatized* phrases (e.g., cats-cat).

Term Linking

Using the development split of the curated dataset, we will approach the problem as a retrieval task utilizing efficient tools like FAISS[7] to index the dictionary and the contextualized term following our previous works[33,35]. In order to detect the missing terms to inform the domain-experts, our model will also handle the cases where the *term is not linkable*, simply by querying the DB beforehand.

Evaluation: We will use R@n (percentage of the ground-truth term being in the top n) to evaluate the performance of the models.

Terminology-aware Translation

We will explore two common approaches: i) post-processing the output, i.e., rewriting the content with lexical constraints (as given in Stage I) and ii) incorporating constraints into the training of the machine translation model. In Stage I, we only mentioned the first approach, assuming that MT engine would mostly leave the technical terms unchanged, however our preliminary analysis showed otherwise. As shown in Table 2, there were cases where the terms left unchanged (e.g., hypernucleus), however, MT-engine mostly caused mistranslations altering the term drastically. For instance, it translates "the ground-state" configuration of the atom as "zemin-devlet", which literally means "basement-government". Therefore, we realized the need for *incorporating the source text* into the models and designed the annotation protocol accordingly.

As for the first approach, we plan to replace technical terms with the placeholders, and translate the text with an existing MT model. Then we will build upon our previous work[34] and redesign a contextualized reinflection model. Different from our previous work, the model will need to perform reinflection with *missing morphological tags*[8] and *missing word order information (in case of multiple placeholders, it needs to figure out which term comes where)*. Despite being challenging, we believe word-alignment techniques based on pretrained multilingual language models such as AwesomeAlign[36] can provide acceptable alignments for the placeholders which can be used to handle both missing tags and missing order problems. As for the second approach, we plan to experiment with concatenating the terminological constraints to the input in various ways (e.g., lemmatized, surface etc…) as it provided promising results for Czech[22].

Evaluation: It is common to perform two step evaluation: i) general translation accuracy ii) terminology consistency. Although human evaluation is also necessary for NLG evaluation, commonly used metrics for MT are BLEU, chrF, BERTScore and COMET[37]. Terminology consistency evaluation is yet an active field of research[38,39], however the most common metric is the exact-match term accuracy[40]. We plan to use all possible metrics and conduct a small user study to rank the generation outputs from different models (see the next task).

Building a Communication Channel between the Communities

We will work together with two distinct communities that do not normally engage: academics, researchers and students that voluntarily contribute to terminology database[9] and Wikimedia Community User Group Turkey[10]. We first plan to connect with the Wikimedia group with a particular focus on editors that generate STEM content. We will do a

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

[8] Source text doesn't have morphological tags, and MT output is unreliable.

[9] https://terimler.org/katkida-bulunanlar

[10] https://meta.wikimedia.org/wiki/Wikimedia_Community_User_Group_Turkey

semi-structured interview with them to gain insights on their profiles and editing habits (e.g., Do they consider themselves domain-experts? How do they translate technical terms? Are they aware of any terminological databases? Would they contact a domain-expert for guidance/help? etc…) Similarly we will interview the academic community for their desire to contribute (e.g., would they consider reviewing a wiki article on their domain of expertise? would they answer requests for missing terms? etc…). In this initial phase, we will reach as many people as possible through maintainers' immediate networks, emailing lists, our collaborators from both communities, and student/university organizations that we have access to. Expected outcome is to identify potential new Wikipedia contributors with domain-expertise and existing contributors who are open to feedback/or have domain-expertise.

Seminar I. We will organize an online seminar within the first three months to introduce the project and present the interview results. During the seminar, we will hold a panel discussion on best strategies to bridge two communities (e.g., how to give feedback/how to use feedback from one another) to make informed decisions for our research. Expected outcome is an established communication channel (online e.g., discord,slack or offline e.g., feedback button, email etc…).

User study I. Using the previously established communication channel, we will recruit around 10 active participants from different groups (both old and newly recruited contributors) to perform a small scale user study. They will be shown a source text with highlighted and hyperlinked terms using the developed models. We will ask them to translate the content in a contrastive setting (one showing the links, other without), and evaluate the output against the gold standard(s).

User study II. We will do another user study with the same participants to gain insights on the generation models. We will show them the output of our rewriting and weakly supervised models along with modern tools (e.g., Google Translate, ChatGPT) and ask them to rank those outputs. Then we'll evaluate the rank of our proposed models against others. Expected outcome for user studies is a quantitative measure for potential usability of the developed models.

Seminar II. We will organize another online seminar to discuss the project progress and user study results. We will hold a discussion panel on how to design training materials for editors to more accurately translate the technical terms (using developed models).

Evaluation. After each user study, we will conduct a survey to measure system usability score (SUS). We will also reveal the results to the participants, and ask them whether they would incorporate any of these models into their editing process.

Expected Outcome(s) Together with the community staff, we will design training materials for editors to more accurately translate the technical terms (using developed models) drawing conclusions from the user studies. Our community staff will include them in the next editing marathons (usually conducted in universities, NGOs or high-schools).

Extensions to other languages

Terminology DB: The most specific component of our proposed research, hence the bottleneck for expansion, is the domain-curated terminology database, terimler.org. As given in Table 1, the DB already contains German and French translations; hence the proposed research can be transferred to those languages with minimal effort. We have also searched for the availability of such a terminological dictionary and found many (A screenshot is given in Appendix). For instance IATE[11] contains 7 million entries for all EU languages. Termium[12] is maintained by the Government of Canada, and contains millions of terms in English, French, Spanish and Portuguese. Terms for specific categories can be downloaded as a *.csv file. dict.cc is a more general dictionary, however, contains many scientific terms for many languages as well.

Parallel corpus. *contenttranslation* dumps are available for a wide range of language pairs. Therefore no effort would be needed.

Annotation effort. To detect only 1 BLEU point with 75% power, one needs around 2,000 evaluation sentences[28]. To increase the power, the common practice is to create 3,000 high-quality parallel sentences. The annotation effort for identifying and linking the terms, and correcting the translations require a substantial amount of resources. It should also be noted that annotation should be performed by 3 different annotators to ensure the quality. Therefore, given the limited time, budget and staff, repeating this effort for other languages is not currently feasible. However, it should be possible to add an extra annotation layer with terminology links to existing parallel corpora from WMT shared tasks using the aforementioned terminology DBs, with minimal effort. We plan to reach out to the WMT organizers to discuss this further. We also plan to attend Wikimania 2024 to discuss the roadmap for curating such a dataset for other Wikipedia editions.

[11]https://iate.europa.eu/home
[12]https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng

NLP Models.
Since we plan to develop models based on pre-trained multilingual LMs for all NLP tasks (e.g., mGPT, mT5, mBART etc...), our models can directly be extended to other languages with a small development set—which can be created with less effort.

Expected output

Expected intermediate outcomes from each task are given separately under related sections.

Public datasets and models: NLP researchers can use them to train/evaluate/compare their own models.

Online Seminars & Office Hours: Bridge Turkish scientists and Wikipedians in two public seminars to introduce the developed models and gather feedback. Additionally, communicate the results of related interviews/user studies with Wikipedia communities by organizing office hours for discussion.

Guidelines for Wikipedia editors: Training materials for editors to more accurately translate the technical terms (using developed models).

Scientific publication at a top-tier NLP venue (e.g., *CL, EMNLP or CL, TACL journal)

Risks

Low participation from the scientific community: There are currently around 150 academics listed on the website as contributors, with all of them having a large pool of alumni, advisees and undergraduate students, enlarging the pool to thousands. As previously discussed, we will reach as many people as possible in the initial phases, using maintainers' immediate networks, university and student organization emailing lists we have access to, and our collaborators from both communities. To mitigate the risk further, we also budgeted for a "Human Subject Compensation" for user study participants.

Low participation for data curation: According to our prior experiences[41][13], with fair compensation for the work, this is a very low risk. Furthermore, the terminologies will be provided externally, so college-level education would be enough for labeling. In case of low participation, we can hire professional translators.

Turkish Wikipedia Community is an active, vibrant community with an adequate size of contributors and Wikipedia articles. The only risk is that it is hard to predict how the Turkish Wikipedia editors will perceive the developed models. Hence, we've budgeted for a community staff to gather continuous editor feedback to enhance future acceptance and adoption of the models.

Community impact plan

We will work closely with Başak Tosun and Zafer Batık, leaders of Wikimedia Community User Group Turkey, to conduct community related tasks outlined in section "Building a Communication Channel between the Communities", as well as Bülent Sankur, the lead of terimler.org. We will present our findings at Wiki Workshop, and Wikimania. We will maintain an up-to-date project page on MetaWiki:Research and consistently update it with the findings from each step.

Evaluation

Task specific evaluation metrics are given under each task. Other than those, we will consider the level of community engagement as another metric, which will be approximated with the number of participants in the seminars, number and frequency of messages in the communication channel, number of stars in our github repo, number of edited articles following the developed guidelines, and number of reported missing terms.

[13] https://turkishpropbank.github.io/

Budget

GGLAB - Conference Fund Budget Template

Response to reviewers and meta-reviewers

Thank you all for your insightful feedback on our proposal.

Regarding tool development:
We wholeheartedly agree with your perspective. We have revised our proposal to ensure that it strictly adheres to this principle. Here are the changes:

1. In the "Building a Communication Channel between the Communities" section, we have redesigned the User Study to have two-stages on the developed **model outputs**, rather than live tool usage.
2. We have removed the tool from the expected outputs.
3. We have removed the engineering cost.

Regarding expanding the scope to other languages & generalizability:

Please see the Section "Extensions to other languages".

Regarding risk planning:
Please read our detailed plan in "Building a Communication Channel between the Communities" regarding our effort to **minimize**

the risks associated with **low** community participation. Please also see the Risks section for the contingency plans.

Regarding quantitative comparison with current translation models and LLMs:

Thanks for this suggestion. We have conducted a comprehensive preliminary study to answer this question, given in Section "Preliminary Work". In short, we find that terminology-aware translation is a challenging task for all automated tools. We also find that, even though human performance surpasses tools by a large margin, editors also suffer translating the STEM terms. In addition, we find that both English and Turkish Wikipedia have a large room for improvement for STEM content. All our findings suggest that this project is indeed necessary.

Specific Response to reviewer uKXk: Thank you so much for the suggested papers, the proposal benefited a lot from them to position the topic better. We have also added them to the "Related Work" section.

References

[1] Navigli, R., & Ponzetto, S. P. (2010, July). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216-225).

[2]Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics, 2*, 231-244.

[3] Moro, A., & Navigli, R. (2015, June). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 288-297).

[4] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., ... & Riedel, S. (2021, June). KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2523-2544).

[5] Abhik Jana, Sruthi Mooriyath, Animesh Mukherjee, and Pawan Goyal. 2016. WikiM: Metapaths based Wikification of Scientific Abstracts. In *Proceedings of ACM Conference*, Washington, DC, USA, July 2017

[6] Jana, A., Kanojiya, P., Goyal, P., & Mukherjee, A. (2018, August). WikiRef: Wikilinks as a route to recommending appropriate references for scientific Wikipedia pages. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 379-389).

[7] D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M. Y., Auer, S., & Ewerth, R. (2020, May). The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2192-2203).

[8] Mohan, S., & Li, D. (2018, November). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*.

[9] Dong, H., Suárez-Paniagua, V., Zhang, H., Wang, M., Whitfield, E., & Wu, H. (2021, November). Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2294-2298). IEEE.

[10] Islamaj, R., Leaman, R., Kim, S., Kwon, D., Wei, C. H., Comeau, D. C., ... & Lu, Z. (2021). NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific data*, *8*(1), 91.

[11] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden. Association for Computational Linguistics.

[12] Augenstein, I., Das, M., Riedel, S., Vikraman, L., & Mccallum, A. (2017, August). SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 546-555).

[13] Kontoulis, C. G., Papagiannopoulou, E., & Tsoumakas, G. (2021, June). Keyphrase extraction from scientific articles via extractive summarization. In *Proceedings of the Second Workshop on Scholarly Document Processing* (pp. 49-55).

[14] Alam, Md Mahfuz Ibn, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. "Findings of the WMT shared task on machine translation using terminologies." In *Proceedings of the Sixth Conference on Machine Translation*, pp. 652-663. 2021.

[15] Semenov, K., Zouhar, V., Kocmi, T., Zhang, D., Zhou, W., & Jiang, Y. E. (2023, December). Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 663-671).

[16] Anastasopoulos, A., Cattelan, A., Dou, Z. Y., Federico, M., Federmann, C., Genzel, D., ... & Tur, S. (2020, December). TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

[17] Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., ... & Zoldan, P. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

[18] Hanneman, G., & Dinu, G. (2020, November). How should markup tags be translated?. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 1160-1173).

[19] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2017, September). Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 936-945).

[20] Hu, J. E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., & Van Durme, B. (2019, June). Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 839-850).

[21] Jon, J., Aires, J. P., Varis, D., & Bojar, O. (2021, August). End-to-End Lexically Constrained Machine Translation for Morphologically Rich Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers) (pp. 4019-4033).

[22] Bergmanis, T., & Pinnis, M. (2021). Facilitating terminology translation with target lemma annotations. *arXiv preprint arXiv:2101.10035*.

[23] Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., & Zhang, M. (2019). Code-switching for enhancing NMT with pre-specified translation. *arXiv preprint arXiv:1904.09107*.

[24] Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive Machine Translation with Large Language Models. In *24th Annual Conference of the European Association for Machine Translation* (p. 227).

[25] Bogoychev, N., & Chen, P. (2023, December). Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 890-896).

[26] Steingrímsson, S., Loftsson, H., & Way, A. (2023, December). SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 256-263).

[27] Thompson, B., & Koehn, P. (2019, November). Vecalign: Improved sentence alignment in linear time and space. 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)* (pp. 1342-1348).

[28] Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020, November). With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9263-9274).

[29] Ferragina, P., & Scaiella, U. (2010, October). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1625-1628).

[30] Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., & Weikum, G. (2011). Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment, 4*(12), 1450-1453.

[31] Shavarani, H. S., & Sarkar, A. (2023, December). SpEL: Structured Prediction for Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 11123-11137).

[32] 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

[33] 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

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

[35] 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.

[36] Dou, Z. Y., & Neubig, G. (2021, April). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2112-2128).

[37] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, November). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685-2702).

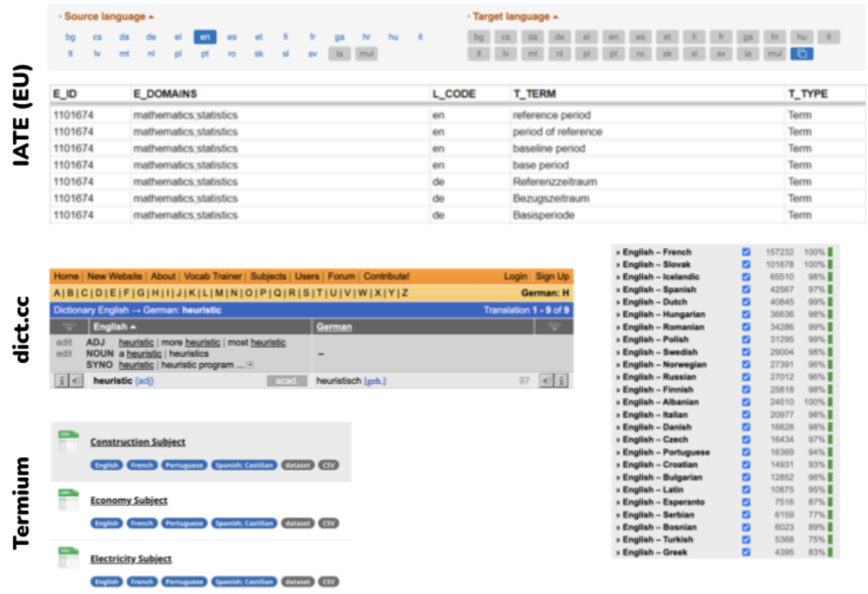
[38] Semenov, K., & Bojar, O. (2022, December). Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 450-457).

[39] Haque, R., Hasanuzzaman, M., & Way, A. (2020). Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation, 34*(2/3), 149-195.

[40] Anastasopoulos, A., Besacier, L., Cross, J., Gallé, M., Koehn, P., & Nikoulina, V. (2021). On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

[41] Gözde Gül Şahin. Building of Turkish PropBank and Semantic Role Labeling of Turkish PhD thesis, Istanbul Technical University, January 2018

Appendix


IATE (EU)
Source language
Target language
E_ID
E_DOMAINS
L_CODE
T_TERM
T_TYPE
1101674
mathematics,statistics
reference period
period of reference
baseline period
base period
Referenzzeitraum
Bezugszeitraum
Basisperiode
Term
dict.cc
Dictionary English → German: heuristic
heuristic
heuristisch
Termium
Construction Subject
Economy Subject
Electricity Subject
English – French
English – Slovak
English – Icelandic
English – Spanish
English – Dutch
English – Hungarian
English – Romanian
English – Polish
English – Swedish
English – Norwegian
English – Russian
English – Finnish
English – Albanian
English – Italian
English – Danish
English – Czech
English – Portuguese
English – Croatian
English – Bulgarian
English – Latin
English – Esperanto
English – Serbian
English – Bosnian
English – Turkish
English – Greek