HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

The Quality of Translation of Turkic Languages by AI Translation Tools

Aikhan Jumashukurov

International School of Kyrgyztan

ABSTRACT

AI translation tools came into use in 2006 when Google introduced Google Translate. AI translation tools are not effective in translating Turkic languages. Turkic languages are a language family including Kyrgyz, Turkish, Kazakh, and Uzbek, and are agglutinative in nature. This paper explores the challenges and advancements in AI translation for Turkic languages, focusing on their complex morphology and syntax. This literature review, including 9 articles, highlights the potential of deep learning and hybrid approaches to improve translation accuracy while emphasizing the need for expanded linguistic resources and specialized tools. The review ends with suggestions for improving AI translation of Turkic languages.

Introduction

AI Translation

AI translation tools came into use in 2006 when Google introduced Google Translate, which utilized Statistical Machine Translation (SMT). However, 10 years later, in 2016, Google changed to Neural Machine Translation (NMT). Unlike SMT, NMT interprets full sentences at once instead of individual phrases, which makes the translations more fluent. This led to a more widespread use of Google Translate and other AI tools that were either based on Google Translate and fine-tuned or made from scratch. Popular examples include Microsoft Translator, DeepL, and SYSTRAN (which was created earlier than Google Translate in 1968) (Talaván & Noa, 2005).

Turkic Languages

The Turkic language family is a diverse group of languages, including Kyrgyz, Turkish, Kazakh, and Uzbek. Languages in this family are spoken by more than 200 million people (Rybatzki, 2020) and are known for their agglutinative nature. This means that words are formed by adding multiple affixes to a root, with each affix representing a different grammatical function. Consequently, the meanings of words in these languages can be quite complex and nuanced. Historically, the Turkic languages have evolved and spread across a wide geographical area due to migration, trade, and cultural exchange, especially along the Silk Road. Despite their shared characteristics, Turkic languages exhibit considerable diversity in vocabulary, pronunciation, and syntax.

Foundational NLP Techniques

The transformer architecture plays a significant role in modern AI translation tools (Vaswani et al., 2017). It uses a self-attention mechanism to understand the relationships between words in a sentence, regardless of their position. In translation, this process begins with encoding, where each word is converted into a vector through embeddings. These vectors are processed through layers that capture contextual relationships (Bird et al., 2009).

The decoding phase then generates the translated sentence by predicting each word based on the encoded input and previously generated words. This approach allows transformers to handle complex linguistic structures, making them effective for translation tasks (Hirschberg & Manning, 2015). However, challenges remain in applying this model to Turkic languages, which have unique morphological and syntactic features (Balabekova et al., 2023).

AI translation tools for Turkic languages leverage several key NLP techniques. Word embeddings like BERT and GloVe are used to capture semantic meanings and context, enabling models to understand the nuances of language. Sequence-to-sequence models are employed to manage variable-length input and output, which is crucial for handling the complex syntax found in Turkic languages. Attention mechanisms within transformers allow these models to focus on relevant parts of sentences, preserving the correct context and meaning during translation. Additionally, morphological analysis tools are integrated to break down complex word forms, which is essential for accurately processing the agglutinative nature of Turkic languages. These techniques work together to enhance the translation of complex linguistic structures, though challenges remain in achieving full accuracy.

Methods

The literature review conducted for this study contains academic publications and book chapters about language usage patterns and the methods used to find these patterns and gather information. Dates range from 1994 (Biber et al., 1994) to the most recent in 2023 (Balabekova et al., 2023). Search terms consist of “Linguistics,” “NLP Python,” “Turkic Languages Morphology,” “Turkic Sentiment Analysis,” and “Discourse Connectives Turkic.” Literature databases utilized include Google Scholar and JSTOR. Additionally, publications from reputable institutions such as the Association for Computational Linguistics were included. Inclusion was based on whether articles discussed computational models for Turkic languages, challenges in language processing, deep learning methods, hybrid and advanced techniques, or foundational NLP techniques. Overall, 20 Articles were screened, and 9 were included.

Results

Challenges with Turkic Languages for These NLP Techniques

Current AI translation tools, like Google Translate, often struggle with Turkic languages due to their complex morphology and flexible syntax. These tools frequently produce inaccuracies, particularly with the numerous word forms and varied sentence structures typical of Turkic languages. While they offer basic translations, these often require significant post-editing to correct errors. This underscores the need for more specialized tools or further enhancements to existing models to effectively handle the unique challenges of Turkic languages (Hirschberg & Manning, 2015).

Turkic languages, being agglutinative, present specific challenges for NLP models like transformers. The vast number of word forms generated by affixes, encoding multiple grammatical meanings within a single word, can overwhelm standard models and complicate accurate processing and translation. Flexible word order further increases syntactic ambiguity, making it difficult for models to maintain context and meaning during translation. Additionally, Turkic languages often rely heavily on contextual cues for disambiguation, complicating the translation process (Tantuğ et al., 2006). These challenges are compounded by the scarcity of high-quality, annotated resources, particularly for less commonly studied Turkic languages such as Kyrgyz and Uzbek, which hinders the training and fine-tuning of NLP models.

In Kyrgyz, the word "үйдө" (üýdö) means "at home," where "үй" (üý) is the root word for "home," and "-дө" (-dö) is the locative suffix meaning "at." When adding possessive and plural suffixes, the word can become "үйлөрүңөрдө" (üýlörüñördö), meaning "at your homes." Here, "үй" (üý) is still the root, "-лөр" (-lör) is a plural suffix, "-үңөр" (-üñör) is a possessive suffix indicating "your," and "-дө" (-dö) is the locative suffix. The layering of affixes within a single word poses significant challenges for NLP models, which must accurately identify and process

each grammatical element. Other language families, such as Uralic and Dravidian languages, also face similar challenges due to their agglutinative nature and flexible syntax.

Deep Learning Approaches to Combat These Challenges

Deep learning approaches have been increasingly employed to address the unique challenges of translating Turkic languages. Techniques such as sequence-to-sequence models with attention mechanisms and more advanced neural architectures like transformers have shown promise in handling the complex morphology and syntax of Turkic languages (Boz, 2018).

One effective approach is contextual embeddings, such as those provided by models like BERT and GPT. These models generate word representations that consider the surrounding context, which helps better understand and translate the complex word forms typical of agglutinative languages (Yildiz & Tantuğ, 2012). This context-aware processing allows the model to more accurately disambiguate meaning based on the specific linguistic environment, improving translation quality.

Hybrid and Advanced Techniques

Hybrid techniques, combining rule-based methods with deep learning, offer improved translation quality for Turkic languages. Rule-based approaches can pre-process text by applying linguistic rules specific to Turkic languages, such as managing affixation and disambiguating word forms. This preparation enhances the efficiency and accuracy of deep learning models (Biber et al., 1994).

Advanced techniques like dependency parsing help maintain the correct syntactic structure during translation, ensuring that the intended meaning and grammar are preserved (Tantuğ et al., 2006). By blending these methods with deep learning, hybrid approaches create a more robust solution, better handling the complex linguistic features of Turkic languages and delivering more accurate translations.

Discussion

The task of translating Turkic languages using NLP models presents substantial challenges due to the unique linguistic characteristics of these languages. Their agglutinative nature, which involves forming words through the addition of multiple affixes, creates a vast array of word forms, each encapsulating several grammatical meanings within a single term. This complexity frequently overwhelms standard NLP models, leading to significant errors in both processing and translation tasks. Moreover, the flexible word order characteristic of Turkic languages introduces a high degree of syntactic ambiguity, complicating the preservation of context and meaning. These difficulties are exacerbated by the limited availability of high-quality, annotated linguistic resources, particularly for less commonly studied Turkic languages such as Kyrgyz.

These issues have direct implications for the effectiveness of current NLP models. The challenges associated with processing complex word structures and maintaining syntactic coherence often result in translations that are either inaccurate or lose crucial contextual information. Furthermore, the shortage of sufficient linguistic resources, as mentioned previously, further aggravates these problems, restricting the capacity to train and refine models tailored to Turkic languages effectively.

The obstacles encountered with Turkic languages are not unique. Similar challenges arise in other language families, including Uralic languages like Finnish and Dravidian languages like Tamil, which also feature agglutinative structures and flexible syntax. By studying the strategies developed to address these challenges in other linguistic contexts, it may be possible to identify solutions that could be adapted to Turkic languages.

Looking ahead, the development of more sophisticated morphological analysis tools is essential. Such tools would significantly improve the management of the complex word forms that typify Turkic languages. For example,

the Zemberek project which provides open-source NLP tools for the Turkish language. Moreover, there is the Apertium project, which translates Kyrgyz and Kazakh very well, but it is still not completely accurate. Consequently, there is a critical need to expand annotated corpora – which are collections of text that have been labeled with additional information for linguistic analysis and use in training NLP models – especially for underrepresented Turkic languages such as Kyrgyz and Uzbek. Turkish has a relatively extensive corpus. However, Kyrgyz, Uzbek, and Turkmen, for example, do not have much. Collaboration between computational linguists, language experts, and native speakers will be vital in creating these resources. The exploration of hybrid models that combine rule-based approaches with deep learning techniques also holds promise for addressing the unique challenges posed by these languages.

Advancing NLP capabilities for Turkic languages could have broader implications. The techniques and methodologies developed through this work could be adapted to other underrepresented or linguistically complex languages, thereby enhancing the inclusivity and accuracy of AI technologies across a broader spectrum of linguistic contexts. While these proposed strategies require further refinement and validation, focusing on these areas could lead to significant advancements.

In summary, addressing these challenges is crucial for enhancing the accuracy and effectiveness of NLP models. By developing advanced tools, expanding linguistic resources, and drawing lessons from languages facing similar challenges, it is possible to make substantial progress in this field. This would not only benefit Turkic languages but also contribute to the overarching goal of improving AI's ability to process and translate a diverse range of languages.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Balabekova, T., Kairatuly, B., & Tukeyev, U. (2023, September). Kazakh-Uzbek Speech Cascade Machine Translation on Complete Set of Endings. In *International Conference on Computational Collective Intelligence* (pp. 430-442). Cham: Springer Nature Switzerland.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based Approaches to Issues in Applied Linguistics. Applied Linguistics, 15(2), 169–189. https://doi.org/10.1093/applin/15.2.169

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Boz, E. (2018). SÖZLÜKLER İÇİN YENİ BİR DİLBİLGİSEL BİLGİ ÖNERİSİ: İLGEÇLERİN ATADIKLARI BİÇİMBİRİMLER. *Uluslararası Türkçe Edebiyat Kültür Eğitim (TEKE) Dergisi*, *7*(2), 749-758.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Rybatzki, V. (2020). The altaic languages: Tungusic, mongolic, turkic. In *The Oxford Guide to the Transeurasian Languages* (pp. 22-28). Oxford University Press.

Tantuğ, A. C., Adalı, E., & Oflazer, K. (2006, August). Computer analysis of the Turkmen language morphology. In *International Conference on Natural Language Processing (in Finland)* (pp. 186-193). Berlin, Heidelberg: Springer Berlin Heidelberg.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. https://user.phil.hhu.de/~cwurm/wp-content/uploads/2020/01/7181-attention-is-all-you-need.pdf

Yildiz, E., & Tantuğ, A. C. (2012). Evaluation of sentence alignment methods for English-Turkish parallel texts. In First Workshop on Language Resources and Technologies for Turkic Languages (p. 64).

Talaván Zanón, Noa. "Evaluating the output quality of machine translation systems: systran 4.0." (2005).