CHALLENGES IN CORPUS LINGUISTICS AND MACHINE TRANSLATION

Nazirova Elmira
DSc of the Tashkent university of information technologies.
E-mail: elmira_nazirova@mail.ru

Abdurakhmonova Nilufar
Doctor of Philology,
Professor of the National university of Uzbekistan.
E-mail: n.abdurakhmonova@nuu.uz

Usmonova Kamola
PhD of the Tashkent university of information technologies.
E-mail: kamolausmonova93@gmail.com

Abstract: This article addresses systematic cases of progress problems on corpora. The process of corpus generation creates linguistic and statistical issues, which eventually govern the entire process. During corpus generation, it needs careful attention to different factors such as corpus size, methods of data collection, organization of textual materials and others. These issues are significant not only for widely spoken languages like English and Turkish but also hold considerable vital Importance for less-resourced languages used in developing countries. We decided to explore these issues in detail in this article.

Keywords: corpus linguistics, language patterns, machine translation, morphological complexity lexical gaps, cultural nuances, linguistic diversity, data collection.

Аннотация: Данный статья рассматривает систематические случаи проблем прогресса в корпусной лингвистике. Процесс создания корпуса порождает лингвистические и статистические проблемы, которые в конечном итоге определяют весь процесс. Во время создания корпуса необходимо уделить внимание таким различным факторам, как размер корпуса, методы сбора данных, организация текстовых материалов и другие. Эти проблемы имеют важное значение не только для широко используемых языков, таких как английский и турецкий, но и имеют значительное значение для менее ресурсных языков, используемых в развивающихся странах. Мы решили подробно изучить эти вопросы в этой статье.

Ключевые слова: корпусная лингвистика, языковые шаблоны, машинный перевод, морфологическая сложность, лексические пробелы, культурные нюансы, лингвистическое разнообразие, сбор данных.

Introduction

In the linguistic field, the implementation of the corpus has increased exponentially for the last 10 years. As Cheng (2011) mentioned that text compilation which gathered for a specific purpose. Meanwhile, the definition of the corpus as a text collection, which is clear, among linguists there are arguments about corpus linguistics as a theory or methodology. According to Tognini-Bonelli (2001) it is necessary a methodological framework for studying language rather than corpus linguistics distinct field of research. There is serious attention involved in the process of generating a corpus in digital form from written sources. It is necessary to take into consideration of various linguistic, linguistic and statistical considerations. As McEnery and Haedie (2011) and Crawford and Csomay (2015) stated the determination of target and careful execution by corpus designers to ensure the successful work of corpus. Likewise, as Weisser (2016) mentioned that maintains that corpus linguistics serves to enhance our understanding of language by analyzing real-life data, emphasizing the practical application of methodology. Therefore, conducting corpus linguistics transcends mere methodology. The advancement of Corpus mechanisms and systems makes easier for people around the world who want to do research in their areas. Over time these mechanisms have gone from being mainly used for linguistic studies and education to being used for research in other specific areas. More and more researchers are using corpus technology in their studies and making their own observations and finding out some challenges and this paper also looks at the progress in research findings in various studies and several challenges in machine translation systems from Uzbek into English within the realm of corpus linguistics.

Discussion

It is essential to understand what corpus linguistics is and what is not corpus linguistics if people can get enough information, then it is easy to enhance after deep observation. Corpus linguistics cannot provide negative evidence and all possible language at a one time as well as explain why. This means corpus tells us what is or is not have in the corpus. It cannot give information what is possible or incorrect in language. That is why many users believe erroneously that whether it cannot show all manners to convey the concept then they may consider the corpus is entirely flawed. Corpus linguistics helps spot patterns in language use, but it cannot always explain why those patterns exist. Figuring out the reasons behind language use often depends on the instincts of language speakers. Even though corpora are carefully put together and cover a wide range of texts, they cannot capture every possible way language is used [Abdurakhmonova, N. 2019,2021,2022]. Hence, while corpora are organized collections of language data, they have limits when it comes to representing the full diversity of language.

Challenges corpora linguistics from Uzbek into English

There are several issues occurring during the creation of machine translation. The issues are Morphological complexity, lexical Gaps, word order differences, and cultural nuances. The reason is there is a rich morphology with complex inflection and derivational patterns. While translating it can be challenging to do more accurately into English and make it simpler morphology. To be more precise, Uzbek has agglutination and extensive suffixation. To illustrate, “daftarlarimizdan” is Uzbek word, which consists of “daftar” (note-book) with suffix – “lar” (plural) and “imizdan” (from our). When we translate it, it is from our note-books that requires not only identifying the root but also correctly handling the plural and possessive markers [Abdurakhmonova, N. 2019,2021,2022].

Word order differences: Uzbek and English word order patterns is not similar. Uzbek word order follows S-O-V, while English has S-V-O order. It can be grammatical errors if not handling properly while translating from one language to another. To cite a relevant example is that “U o’qidi maqolani” is Uzbek sentence when we translate it, we have to rearrange it to She reads the article. In order to translate accurately the correct placement of subject, verb, and object is vital.

Lexical gaps: there are some words in Uzbek, which concepts do not have direct equivalents in English or vice versa. In this case, we need creative solutions sometimes. In Uzbek, the word "dustlik" means more than just friendship. It's about having close friends who support each other like family, showing loyalty and sticking together through thick and thin. While "friendship" is close, it does not fully capture the strong emotional bond and mutual care that "dustlik" represents in Uzbek culture. A useful example to mention here is that in Uzbek, the word "dustlik" means more than just friendship. It's about having close friends who support each other like family, showing loyalty and sticking together through thick and thin. While "friendship" is close, it does not fully capture the strong emotional bond and mutual care that "dustlik" represents in Uzbek culture.

Cultural Nuances: Languages often have words and phrases that carry unique cultural meanings, making direct translation difficult. Ensuring that translations capture both the intended meaning and cultural context is important but can be tough. In Uzbek, when someone says “Assalom alaykum, yaxshimisiz”(Hello. How are you), it is not just a greeting- it also shows politeness and hospitality. Translating it directly as "Hello, how are you?" might not capture all the cultural meanings behind the greeting [Abdurakhmonova, N. 2019,2021,2022].

Furthermore, researchers frequently encountered some issues in Data collection, annotation and tagging, domain adaptation, evaluation metrics and resource availability. In Data collection, it is clear that Uzbek and English cover various genres and domains. Thus, the issue can be collecting a corpus that contains various texts, new articles, and conversational speech that ensures machine translation models are trained on diverse linguistic data. In annotation and tagging, linguistic data is very important in order to train accurate machine translation models. Syntactic structures and entities in the corpus help machine systems understand the relationship within sentences. In domain adaptation, it is necessary to differ domain, such as legal or medical texts so demand training on domain-specific corpora. Thus, here machine translation tool should accurately translate those medical terminology and specific language patterns. In evaluation metrics and resource availability, it is very important to focus on measuring system performance in order to develop reliable evaluations that obtain the quality and accuracy. Moreover, if languages have limited resources, it also creates problems such as size and composition balancing of the corpus with the range of linguistic variations. Thus, it is necessary to take into consideration the coverage of diverse linguistic phenomena and sufficient data for robust analysis.

Conclusion

This article gives information on the systematic challenges encountered in corpus linguistics. Generation and application of corpora are emphasized, as well as it highlights the importance of considerations during corpus generation such as size, data collection methods and textual organization, which hold significant important not only for widely spoken languages but also for less-resourced languages, especially in developing countries. Moreover, certainly, the article also sheds light on the boundaries of corpus linguistics. It highlights its inability to provide negative evidence, cover all conceivable linguistic variations, or fully elucidate the underlying reasons for language patterns. Nevertheless, despite these limitations, corpus linguistics remains an indispensable tool for uncovering recurrent language patterns. However, the challenges extend further when considering machine translation from Uzbek to English within the corpus linguistics framework. These hurdles encompass the intricate morphological structure, lexical disparities, divergent word order, and nuanced cultural subtleties present in Uzbek. These complexities demand innovative strategies and meticulous precision to achieve accurate translation into English.

References:

1. C. Chen, K.Chan, P.Wong, E.Chee, L.Wang, Q.Wang. A corpus-based online pronunciation learning system: The Pedagogical applications of a spoken corpus for improving Hong Kong/Mainland university students' English pronunciation. The Second Asia Pacific Corpus Linguistics Conference, 2014.
2. E.Tognini-Bonelli.Corpus linguistics at work. Amsterdam: J. Benjamins, 2001.
3. M. Weisser. Practical corpus linguistics: An introduction to corpus-based language analysis. John Wiley- Sons. 2016
4. Abdurakhmonova, N., Tuliyev, U., Ismailov, A., & Abduvahobo, G. (2022). Uzbek electronic corpus as a tool for linguistic analysis. In Компьютерная обработка тюркских языков. TURKLANG 2022 (pp. 231-240).
5. Abduraxmonova, N. Z. Q., & Urazaliyeva, M. Y. (2022). O 'zbek tili elektron korpusida (http://uzbekcorpus. uz/) og 'zaki matnlar korpusini yaratishning nazariy va amaliy masalalari. Academic research in educational sciences, 3(3), 644-650.
6. Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. Applied Sciences, 11(19), 9117.
7. Abdurakhmonova, N. (2019). Dependency parsing based on Uzbek Corpus. In of the International Conference on Language Technologies for All (LT4All).
8. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In Proceedings of the 11th Global Wordnet conference (pp. 8-19).
9. N.Abdurakhmonova, U.Tuliyev, A.Gatiatullin. Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus. uz.International Conference on Information Science and Communications Technologies (ICISCT), 2021. pp. 1-4.
10. T.McEnery, A.Hardie. Corpus Linguistics: Method, Theory, and Practice. Cambridge: Cambridge University Press, 2011.
11. W.Crawford, E.Csomay. Doing Corpus Linguistics. London: Routledge, 2015.