LEXICO-SEMANTIC AMBIGUITY IN THE ENGLISH-KARAKALPAK PARALLEL CORPUS AND ITS IMPACT ON THE QUALITY OF MACHINE TRANSLATION

Allanyazov Rustem Baxavedinovich,
Lecturer,
rustemallanyazov@gmail.com
Tashkent State University of Oriental Studies

Abstract. The article examines the problem of lexical-semantic ambiguity in parallel corpora and its impact on the quality of machine translation using the example of the English-Karakalpak language pair. Karakalpak is a low-resource Turkic language with agglutinative morphology, which increases the complexity of translation. It is shown that polysemy of lexical units is a key cause of semantic distortions, especially in conditions of limited linguistic resources. The necessity of integrating context-oriented and semantic methods into machine translation systems is substantiated.

Keywords: *machine translation, lexical-semantic ambiguity, parallel corpora, semantic disambiguation, translation quality.*

Annotatsiya. Maqolada parallel korpuslardagi leksik-semantik noaniqlik muammosi va uning mashina tarjimasi sifatiga ta’siri ingliz-qoraqalpoq tili juftligi misolida ko‘rib chiqiladi. Qoraqalpoq tili agglyutinativ morfologiyaga ega past resursli turkiy tillardan biri bo‘lib, bu tarjima jarayonining murakkabligini oshiradi. Leksik birliklarning ko‘p ma’noliligi, ayniqsa cheklangan til resurslari sharoitida, semantik buzilishlarning asosiy sababi ekanligi ko‘rsatilgan. Kontekstga yo‘naltirilgan va semantik usullarni mashina tarjimasi tizimlariga integratsiya qilish zarurati asoslanadi.

Kalit so‘zlar: *mashina tarjimasi, leksik-semantik noaniqlik, parallel korpuslar, semantik dizambiguatsiya, tarjima sifati.*

Introduction

Machine translation systems demonstrate real progress through the development of neural architectures and the use of large parallel corpora. Nevertheless, even with the application of transformational models, the task of correctly conveying text meaning remains not fully resolved, which is linked to the fundamental properties of natural language, primarily lexico-semantic ambiguity [3:47–59].

Polysemy of words, where a single word can mean different things depending on the context, is a serious problem for machine translation. In machine translation, this often leads to the program selecting the wrong translation, and because of this, the meaning of the entire sentence is distorted.[7:108].

A characteristic difficulty is the problem of ambiguity in translation under conditions of limited parallel data, where statistical information is insufficient for the reliable contextual selection of a word’s meaning[8:1-27]. For example, for the English-Karakalpak language pair, this problem is particularly acute: the Karakalpak language, which is distributed in the north of Uzbekistan in the Republic of Karakalpakstan, has an extremely small volume of digital parallel data and is structurally strongly distinguished from English with an analytical structure in English vs agglutination in Karakalpak, i.e., a different system of tenses and cases between languages. Under such circumstances, the significance of linguistically motivated and semantically oriented approaches to text processing increases.

Lexico-semantic ambiguity as a linguistic problem

In linguistics, ambiguity is considered the ability of a linguistic unit to permit more than one interpretation. There are lexical, syntactic, and pragmatic ambiguities, but it is precisely lexical ambiguity that has the most significant impact on automatic translation. [1:1-5].

From the perspective of cognitive linguistics and psycholinguistics, translation ambiguity arises when a single word in the source language corresponds

to several possible equivalents in the target language. Such cases are particularly characteristic of abstract vocabulary, functional verbs, and polysemous nouns.

Research shows that even for languages with a high level of resource provision, a significant portion of words have more than one correct translation, the choice of which is determined by the context. [5:170-180].

Lexico-semantic ambiguity in machine translation

In machine translation systems, the problem of ambiguity is closely linked to the task of automatically resolving word meanings (Word Sense Disambiguation, WSD), which is an important stage of semantic text analysis [5:173-183].

Modern neural models are capable of partially considering sentence context, but research shows that attention mechanisms do not always effectively utilize contextual information when selecting the meaning of polysemous words.[10:154–166]. This leads to the model reproducing a statistically probable but semantically incorrect translation variant. When translating polysemous lexical units, the model resolves semantic errors, confirming the need for further development of context-oriented methods for semantic disambiguation.

One of the main and most complex problems is the task of eliminating lexical and semantic polysemy in machine translation. In the case of correct morphological analysis and synthesis of word forms, selecting the wrong meaning for a polysemous lexical unit can lead to the distortion of the meaning of the translated statement. In the English-Karakalpak machine translation system, this problem is complicated by typological differences between the languages, as analytical ways of expressing meanings in English must be transformed into morphologically expressed forms of the Karakalpak language.

Qualitative analysis of translations shows that errors related to the incorrect choice of word meaning constitute a significant share of all semantic errors in

machine translation and have a stronger impact on text perception than syntactic inaccuracies. [4:1-8].

Parallel corpuses and the problem of limited resources

Parallel corpora are the primary source of knowledge for training and evaluating machine translation systems. However, for many language pairs, the volume of available parallel data remains insufficient for training universal neural models. The English-Karakalpak parallel corpus used in this study serves as a vivid example.The volume of available data for this pair is significantly smaller than for Russian-English or Chinese-English, which leads to a sharp increase in the effect of lexical-semantic ambiguity and requires the development of special methods for resolving polysemy, taking into account the characteristics of Karakalpak vocabulary and grammar.

As noted in the studies, hybrid approaches combining statistical methods with linguistic and semantic knowledge prove more effective for such conditions.

As noted in the studies, hybrid approaches combining statistical methods with linguistic and semantic knowledge prove more effective for such conditions. [6:27-34].

The impact of lexico-semantic ambiguity on translation quality

The quality assessment of machine translation is traditionally carried out using automatic metrics such as BLEU and chrF. However, these metrics poorly reflect semantic distortions caused by the incorrect choice of word meaning.[2:1-11].

Experimental studies show that translations with high BLEU meanings may contain serious semantic errors related to the translation of polysemous lexical units. This indicates the need to involve linguistic analysis and high-quality expert evaluation.

To quantitatively assess the impact of lexical-semantic ambiguity on machine translation quality within the English-Karakalpak parallel corpus, a neural network

model based on the NLLB (No Language Left Behind) architecture was trained.The parallel corpus used for training and assessment consisted of 5,000 English-Karakalpak sentences. The learning dynamics and key metrics are presented in Table 1.

Table 1 - Learning dynamics and evaluation metrics of the NLLB-based model

Step	Training Loss	Validation Loss	BLEU	chrF	METEOR
500	26.51	6.43	5.86	34.93	0.317
1000	25.66	6.34	9.42	39.41	0.370
1500	25.75	6.30	11.55	42.01	0.400
2000	25.38	6.27	13.67	43.53	0.419
2500	25.43	6.26	14.04	44.05	0.420
3000	25.28	6.25	**14.86**	**44.50**	**0.428**

A. Model behavior during the learning process.

The results demonstrate a steady improvement in all assessment metrics as learning progresses. Specifically:

- the BLEU indicator increases from 5.86 to 14.86, indicating an improvement in lexical and structural compatibility;

- the chrF indicator increases from 34.93 to 44.50, reflecting an improvement in accuracy at the character level;

- METEOR improves from 0.317 to 0.428, indicating better semantic adequacy.

At the same time, both training and validation losses are gradually decreasingfrom 26.51 to 25.28* and from 6.43 to 6.25 respectively, indicating stable convergence of the model without significant retraining.

Note: training loss decreased from 26.51 to 25.28.

B. Interpretation of results.

These indicators BLEU (14.86) and METEOR (0.428) are for low-resource language pairs and indicate that the model has the ability to convey individual fragments of meaning, but does not provide high translation accuracy at the level of whole sentences.This confirms the article's thesis that lexico-semantic ambiguity is a critical factor, especially under conditions of limited parallel data.

Therefore, lexico-semantic ambiguity is a critical factor that determines the real quality of machine translation, especially under conditions of limited parallel data.

C. Assessment on a controlled test set.

For additional qualitative assessment, the final model was tested on a separate test set consisting of 585 English-Karakalpak sentences.This set encompassed various syntactic constructions and lexical phenomena, including cases of potential polysemy.

The automatic metrics in this set, as expected, showed values close to those obtained in the validation sample: BLEU = 14.86; chrF = 44.50; METEOR = 0.428. This confirms the stability of the model and the absence of significant discrepancies between validation and test data.

Table 2 provides examples of correct translations demonstrating the model's capabilities in contexts with low lexical ambiguity, where choosing an equivalent does not cause difficulties.

Table 2 - Translation Examples

English Input	Karakalpak Output
Tom is a computer programmer	Tom kompyuter programmist
I'm going to grow wheat there	Men ol jerde biyday egemen
Tom isn't accustomed to walking barefooted	Tom jalan ayaq jurip uyrenbegen

At the same time, as shown in sections A and B, the model makes semantic errors when translating polysemous lexical units, which confirms the need for further development of context-oriented methods of semantic disambiguation for the English-Karakalpak language pair.

Discussion and prospects

The conducted analysis confirms that the problem of lexico-semantic ambiguity cannot be fully resolved solely through increasing data volumes or complicating neural architectures. A promising direction is the integration of context-oriented semantic methods, translation memory, and linguistic rules into the architecture of machine translation systems. An additional area of research is the automatic formation of semantically labeled corpora based on parallel data, which partially overcomes the shortage of annotated resources.

Conclusion

The article demonstrates that lexico-semantic ambiguity is a key reason for the decline in machine translation quality under conditions of limited parallel data. This data is particularly important for low-resource languages such as Karakalpak. Incorrect selection of a word's meaning can lead to semantic distortions that are not always detected by automatic metrics. The feasibility of applying semantically oriented and hybrid approaches in developing machine translation systems has been tested.

References

1. Aliboeva N. The Problem of Ambiguity in Machine Translation // Foreign Linguistics and Linguodidactics. 2024. pp 1-4.

2. Antonova N.A., Kuzmich I.V. Comparative Analysis: Machine vs Human Translation. 2024. pp 1-11.

3. Bakumenko Ya.D., Tadzhibova A.N. Current Challenges and Approaches in Machine Translation // Russian Journal of Cybernetics. 2025. pp 47–59.

4. Barreiro A. et al. When Multiwords Go Bad in Machine Translation. 2013. p 1-8.

5. Bolshina A.S. On the Methods of Automatic Creation of Semantically Annotated Collections // MSU Bulletin. Philology. 2022. pp 173-183.

6. Eisele A. et al. Hybrid Machine Translation Architectures. 2008. pp 27-34.

7. Rodina S.V., Lakiza E.V. Linguistic Problems of Machine Translation // Universum: Philology. 2023. p 108.

8. Singh P. et al. Leveraging Cross-Domain Parallel Corpora for Low-Resource NMT. 2025. pp 1-27.

9. Tokowicz N. Translation Ambiguity Affects Language Processing. 2014.pp 170-180.

10. Hatami A. et al. Enhancing Translation Quality by Leveraging Semantic Diversity. 2024. pp 154–166