

**OPEN ACCESS**

SUBMITTED 30 October 2025

ACCEPTED 21 December 2025

PUBLISHED 31 December 2025

VOLUME Vol.05 Issue 12 2025

COPYRIGHT

© 2025 Original content from this work may be used under the terms of the creative commons attributes 4.0 License.

Statistical Approach-Based Machine Translation System

Utayeva Iroda Baxodirovna

Doctor of Philosophy (PhD) in Pedagogical Sciences, Acting Professor,
Dean of the Faculty of Languages at Samarkand State Pedagogical
Institute, Uzbekistan

Abstract: This article analyses the principal theoretical and technological approaches to machine translation, in particular the linguistic and computational characteristics of rule-based, statistical, and example-based machine translation systems. Structural differences between the source and target languages, the transformation of syntactic representations, the use of bilingual parallel corpora, and the alignment of translation units are examined scientifically. Methodological problems related to developing translation models for agglutinative and low-resource languages, formulating linguistic rules, segmenting corpus units, and aligning them are also considered.

Keywords: Machine translation, rule-based translation, statistical machine translation, example-based machine translation, parallel corpus, bilingual corpus, syntactic representation.

Introduction: Statistical approach-based machine translation emerged as an alternative methodological direction to traditional rule-based models and provides higher results in terms of translation efficiency and the degree of automation in certain tasks. Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) are distinguished as the main directions of machine translation within the statistical paradigm.

In these approaches, the translation output is generated on the basis of probabilistic-statistical models. The parameters of these models are determined on the basis of aligned parallel bilingual text corpora in the source and target languages. Parallel corpora serve as the principal empirical database for identifying lexical, syntactic, and statistical correspondences between translation units and for training the translation model. However, the effectiveness of such an approach directly

depends on the availability of large volumes of high-quality parallel data. Therefore, developing and implementing statistical machine translation models for languages with no resources or limited resources creates certain methodological and technological difficulties.

LITERATURE REVIEW

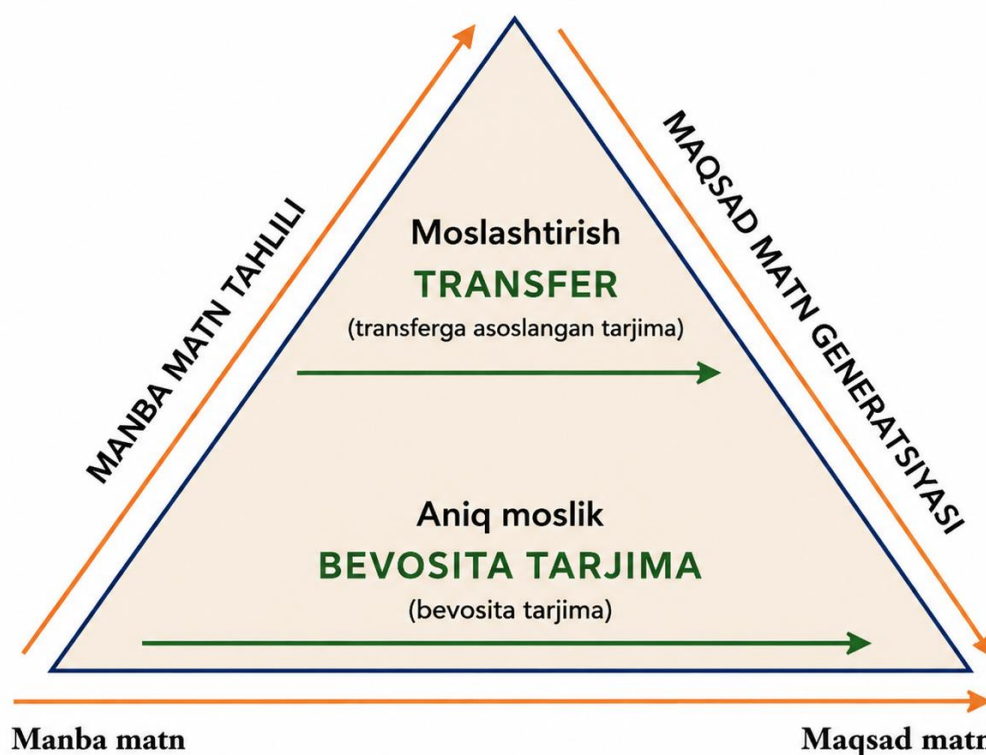
Example-Based Machine Translation (EBMT) is a corpus-based, that is, data-driven machine translation approach and is methodologically regarded as an intermediate model between Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). The initial scientific foundations of this concept were proposed by Nagao (1984), and its central principle is based on using previously translated examples as analogical models for generating new translations [20].

The Vauquois pyramid used in the classical machine translation architecture has also been adapted to the EBMT model; however, additional operations such as alignment of interlingual units and exact matching are of particular importance within it. Such a model makes it possible to identify relevant linguistic fragments from a pre-existing database of bilingual examples, find their target-language equivalents, and recombine them within a new syntactic structure during the

translation process.

According to the concept proposed by Makoto Nagao, a human translator does not limit the translation process to the formal linguistic analysis of a sentence. On the contrary, the sentence is first segmented into smaller phraseological or syntactic fragments, the corresponding units of these fragments in the target language are identified, and they are subsequently reorganized into a complete sentence in accordance with the grammatical and syntactic regularities of the target language. The translation of each fragment is carried out on the basis of the principle of analogy by using previously encountered corresponding examples as references [13].

On this basis, the functional architecture of the EBMT system is characterized by three main components. The first component covers the process of matching fragments extracted from the input text with a database containing actual translation examples. The second component performs the task of identifying translation equivalents corresponding to the identified source-language fragments. The third component ensures the generation of the final target text by recombining the identified translation fragments in accordance with the grammatical, syntactic, and semantic regularities of the target language:



4-rasm. EBMT uchun moslashtirilgan “Vokua piramidasi”.

Izoh: Piramida pastidan yuqoriga qarab tarjimada tahlil chuqurlashadi va moslashtirish darajasi ortadi.

From a research perspective, in explaining the operating principles of the Example-Based Machine Translation (EBMT) system, it is methodologically appropriate to use the traditional pyramid model proposed by Vauquois and to interpret EBMT mechanisms within this model. In the EBMT system, the sentence is regarded as the main and optimal translation unit, and this feature makes it functionally close to Translation Memory (TM) technology. If the database does not contain a ready translation variant of exactly the same sentence, the EBMT system searches for translation examples that correspond most closely to the input unit on the basis of similarity identification criteria. Therefore, the stage of linguistic analysis of the source-language sentence in traditional machine translation is replaced in EBMT by the process of comparing and matching the input text with the set of available examples.

DISCUSSION

After the system identifies the relevant translation example, it extracts the corresponding fragments in the target language. This process is expressed in the scientific literature through the concepts of “alignment” or “adaptation.” Once the required fragments have been selected, they are recombined in accordance with the structural and semantic characteristics of the target text. In some practical EBMT models, linguistic and algorithmic methods similar to those characteristics of the analysis and generation stages of traditional machine translation systems are used in the processes of identifying matches and recombining fragments. However, one of the aspects of the Vauquois pyramid that does not fully represent the characteristics of EBMT is the relationship between the concept of “exact match” and “direct translation.” These two processes differ essentially and require a minimal level of analysis. From another perspective, an exact match can be interpreted as a “perfect representation,” since no additional adaptation process is required in this case [15]; [16]. On this basis, placing the exact match phenomenon at the upper level of the pyramid is theoretically justified.

Since EBMT is a corpus-based machine translation approach, a parallel and aligned corpus constitutes its main linguistic and information base. At this point, the scientific and methodological question arises as to what the optimal size of the corpus should be. Although the volume and quality of data are among the decisive factors in the translation process, scientific results that clearly determine a universal optimal corpus size have not yet been sufficiently established. Nevertheless, experimental studies confirm that the

number of translation examples in the database has a significant effect on system efficiency.

In particular, in an experiment conducted by Mima and co-authors, an improvement in translation quality was observed as new translation examples were gradually added to the database. In an experiment aimed at translating Japanese abdominal particle constructions, a total of 774 examples were placed in the database, with 100 examples added at each stage. The experimental results showed that translation accuracy improved significantly as the number of examples increased: while the accuracy rate was 30 percent with 100 examples, this indicator reached 65 percent as the amount of data increased [6]; [7]; [8]. Initially, the increase in translation efficiency was relatively linear; however, after a certain threshold, adding further examples to the database did not result in further improvement in translation quality. This indicates that, in EBMT systems, along with corpus size, the linguistic representativeness, structural diversity, and quality indicators of the examples contained in the corpus are also of significant importance.

The main knowledge base of an EBMT system is formed from translation examples. This feature makes it possible to avoid, to a certain extent, the difficulties associated with knowledge formation and the development of formal rules characteristic of Rule-Based Machine Translation (RBMT) systems. By generalizing existing examples, EBMT identifies linguistic correspondences that can be applied to new translation cases. One of the important criteria in this process is whether or not existing examples are used directly during the translation process; it is precisely this feature that makes it possible to distinguish EBMT from purely statistical translation methods. In the traditional approach, it is not always possible to derive formal translation rules directly from examples, while in some systems there may be no mechanism for directly accessing the example database at runtime.

In general, although the Example-Based Machine Translation approach is capable of partially overcoming the problems of “structure-preserving translation,” a number of theoretical and practical issues remain within this technology. A brief scientific review of research on EBMT is summarized in Table 1. This review covers scientific studies that have made a significant contribution to the formation and evolution of EBMT technology and describes the methodological approaches proposed in them, the linguistic and computational techniques applied, the database used for model training, and the criteria for evaluating system efficiency. In particular, metrics such as the BLEU score and translation accuracy are used as the main experimental criteria for the quantitative evaluation of

translation quality.

The EBMT (Example-Based Machine Translation) system is particularly effective in translating sublanguages limited to a specific subject domain. Once an appropriate and representative corpus has been identified, the main methodological problem is related to aligning corpus units, that is, determining the correspondence between semantically equivalent sentences in the source and target languages at the level of more precise and smaller linguistic units [19]. In the EBMT approach, the translation process is not

based on reconstructing a rigid syntactic structure, because the structural model of the source language is not directly applied to the grammatical and syntactic structure of the target language.

Another important problem of this approach is explained by the necessity of storing translation examples in a complex structural form. Creating, indexing, searching, and processing a large-scale example database require significant computational resources, thereby increasing the computational complexity and resource consumption of the system.

Table 1

Review of Research on Example-Based Machine Translation (EBMT)

Year	Objective	Techniques / Dataset	Results
1994 [37]	To propose a method for analysing constituent boundaries for EBMT that supports variable and boundary-related linguistic patterns in English and Japanese.	Translation memory (TDMT) system for English–Japanese and Japanese–English.	The proposed method was stable and efficient, with a processing time of $me = 0.25$ seconds.
1995 [38]	To address the problem of how to measure similarity between sentence fragments and stored examples in EBMT.	A set of 8,000 Greek–English paired entries obtained from the CELEX database.	According to Fws (I) and the retrieved examples (G), the system was evaluated as 62% useful.
1999 [39]	To demonstrate an example-based approach using the WWW (World Wide Web) for lexical selection, to use a sufficiently large corpus, and to provide a noise-	Two bilingual dictionaries were tested: German–English and Spanish–English.	Accuracy = 86% / 87%.

	resistant (noise-free) signal.		
1999 [40]	To demonstrate that the Pangloss Example-Based Machine Translation system (PanEBMT) performs input segments more effectively in parallel translation involving multiple translation engines.	726,406 Spanish–English sentence pairs; data obtained from the UN Multilingual Corpus.	70% coverage of unrestricted Spanish news agency text was achieved.
2001 [41]	To propose an approach for extracting similar examples using DP matching (determiner phrase matching) between input and example sentences and to measure the semantic distance between words.	A bilingual corpus consisting of Japanese sentences and their English translations.	The test translation was evaluated as one of four tasks (A, B, C, D, E). The proposed approach achieved 80% accuracy.
2002 [42]	To provide an abstract linguistic representation layer for knowledge transfer between languages and for extracting and generating translation examples; to perform	English–Japanese MSR-MT system. Translation knowledge was acquired through constructed phrase structures, grammatical rules, and statistical and heuristic methods.	Evaluation metric: +1 = good; MSR-MT output sentences: –1 = worst. The MSR-MT score was slightly lower, –0.015, with a confidence measure of ± 0.16 relative to the comparison system.

	grammatical analysis and generate output strings.		
2003 [43]	To propose a method for SMT based on a greedy approach that exploits the advantages of the EBMT framework and outperforms the beam-search algorithm used for lexicon generation.	Data obtained from the Basic Travel Expression Corpus (BTEC).	The proposed method outperformed the word-by-word generation beam-search algorithm.
2004 [44]	To improve EBMT quality by applying a syntactic transfer approach that selects the best translation together with an empty translation model and an SMT language model.	Basic Travel Expression Corpus — a set of English translation and Japanese sentences.	The proposed methodology (All Search) produced the best result among the four tested methods: BLEU score = 0.498.
2005 [45]	To propose a probabilistic model formalizing EBMT and to search for the translation example with the highest probability by integrating contextually similar examples.	IWSLT04 English–Japanese sentence pairs.	The proposed model achieved a BLEU score of 0.41, slightly higher than the state-of-the-art result: Basic = 0.39. Baseline result = 0.31, C1 = 0.13, C2 = 0.27.

2006 [46]	To demonstrate that EBMT produces better results when trained on heterogeneous data, provided that subtitles are not incorrectly translated.	English–German parallel alignment data: both homogeneous and heterogeneous data from the Europarl corpus.	When trained on heterogeneous data, the BLEU score improved by 86%.
------------------	--	---	---

CONCLUSION

In general, it was determined that the functional capabilities of neural machine translation and Transformer models are extending beyond translation studies and are becoming a universal methodological tool for modelling complex sequences in various subject domains. These models are capable of identifying hidden patterns from large volumes of data, learning multifactorial and nonlinear relationships, automatically establishing relationships between input and output representations, and extracting contextually relevant information through attention mechanisms. This makes it possible to assess the scientific significance of NMT technologies not only in terms of performing automatic translation between natural languages, but also as an interdisciplinary paradigm of intelligent modelling that serves to process structurally and semantically complex data, predict transformational processes, and generate optimal results.

REFERENCES

1. Furuse O., Iida H. Constituent boundary parsing for example-based machine translation // Proceedings of COLING 1994: The 15th International Conference on Computational Linguistics. Vol. 1. Kyoto, Japan, 1994.
2. Craniias L., Papageorgiou H., Piperidis S. A matching technique in example-based machine translation // arXiv. 1995. arXiv:cmp-lg/9508005.
3. Grefenstette G. The World Wide Web as a resource for example-based machine translation tasks // Proceedings of the ASLIB Conference on Translating and the Computer. London, UK, 1999. Vol. 21. P. 517–520.
4. Brown R. D. Example-based machine translation in the Pangloss system // Proceedings of COLING 1996: The 16th International Conference on Computational Linguistics. Vol. 1. Copenhagen, Denmark, 1996.
5. Sumita E. Example-based machine translation using DP-matching between word sequences // Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation. Toulouse, France, 2001.
6. Brockett C., Aikawa T., Aue A., Menezes A., Quirk C., Suzuki H. English-Japanese example-based machine translation using abstract linguistic representations // Proceedings of COLING-02: Machine Translation in Asia. Stroudsburg, PA, USA, 2002.
7. Watanabe T., Sumita E. Example-based decoding for statistical machine translation // Proceedings of the Machine Translation Summit IX. New Orleans, LA, USA, 2003. P. 410–417.
8. Imamura K., Okuma H., Watanabe T., Sumita E. Example-based machine translation based on syntactic transfer with statistical models // Proceedings of COLING 2004: 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004. P. 99–105.
9. Aramaki E., Kurohashi S., Kashioka H., Kato N. Probabilistic model for example-based machine translation // Proceedings of the MT Summit X. Phuket, Thailand, 2005. P. 219–226.
10. Armstrong S., Way A., Caffrey C., Flanagan M., Kenny D. Improving the quality of automated DVD subtitles via example-based machine translation. 2006.
11. Och F. J., Tillmann C., Ney H. Improved alignment models for statistical machine translation // Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, China, 1999.
12. Marcu D., Wong D. A phrase-based, joint probability model for statistical machine translation // Proceedings of the 2002 Conference on Empirical

- Methods in Natural Language Processing (EMNLP 2002). Philadelphia, PA, USA, 2002. P. 133–139.
13. Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle // *Artificial and Human Intelligence*. 1984. Vol. 1. P. 351–354.
 14. Koehn P., Och F. J., Marcu D. Statistical phrase-based translation: Technical report. Marina del Rey, CA: University of Southern California, Information Sciences Institute, 2003.
 15. Zens R., Och F. J., Ney H. Phrase-based statistical machine translation // *KI 2002: Advances in Artificial Intelligence: Proceedings of the 25th Annual German Conference on AI*. Aachen, Germany. Berlin; Heidelberg: Springer, 2002. P. 18–32.
 16. Zens R., Ney H. Improvements in phrase-based statistical machine translation // *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, MA, USA, 2004. P. 257–264.
 17. Chiang D. A hierarchical phrase-based model for statistical machine translation // *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, MI, USA, 2005. P. 263–270.
 18. Xiong D., Liu Q., Lin S. Maximum entropy based phrase reordering model for statistical machine translation // *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, 2006. P. 521–528.
 19. Marcu D., Wang W., Echihiabi A., Knight K. SPMT: Statistical machine translation with syntactified target language phrases // *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia, 2006. P. 44–52.
 20. Sharma S. et al. Machine translation systems based on classical-statistical-deep-learning approaches // *Electronics*. – 2023. – Т. 12. – №. 7. – С. 1716.
 21. Wu H., Wang H. Pivot language approach for phrase-based statistical machine translation // *Machine Translation*. 2007. Vol. 21. P. 165–181.
 22. Cherry C. Cohesive phrase-based decoding for statistical machine translation // *Proceedings of ACL-08: HLT*. Columbus, OH, USA, 2008. P. 72–80.
 23. Yamada K., Knight K. A syntax-based statistical translation model // *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 2001. P. 523–530.
 24. Сарсенбаева З. Analysis of images and symbols in english non-realistic works // *Ренессанс в парадигме новаций образования и технологий в XXI веке*. – 2023. – Т. 1. – №. 1. – С. 229-232.
 25. Сарсенбаева З. Модернизм в узбекской литературе и интерпретация образов // *Зарубежная лингвистика и лингводидактика*. – 2024. – Т. 2. – №. 1. – С. 193-199.
 26. Zoya S. LITERARY ANALYSIS OF DAVID MITCHELL'S WORKS IN ENGLISH LITERATURE // *Talqin va tadqiqotlar ilmiy-uslubiy jurnali*. – 2024. – Т. 2. – №. 41. – С. 11-19.
 27. Zoya S. LITERARY ANALYSIS OF ULUGBEK HAMDAMOV'S WORKS IN UZBEK LITERATURE // *Talqin va tadqiqotlar ilmiy-uslubiy jurnali*. – 2024. – Т. 2. – №. 41. – С. 20-25.
 28. Sarsenbaeva Z. MODERN APPROACHES TO TEACHING LITERARY TERMS IN COMPARATIVE LITERATURE // *Mental Enlightenment Scientific-Methodological Journal*. – 2026. – Т. 7. – №. 02. – С. 198-209.
 29. Belgibayeva G., Sarsenbaeva Z., Zhumagulova K. MAG 'JAN JUMABAYEV IJODIDA OZODLIK G 'OYASI // *Ijtimoiy-gumanitar sohada ilmiy-innovatsion tadqiqotlar*. – 2026. – Т. 3. – №. 1. – С. 159-165.
 30. Sarsenbayeva Z., Zillolova G. COMPARATIVE ANALYSIS OF TWO TRANSLATED VERSIONS OF A LITERARY TEXT // *Ijtimoiy-gumanitar sohada ilmiy-innovatsion tadqiqotlar*. – 2025. – Т. 2. – №. 4. – С. 183-188.
 31. Sarsenbayeva Z., Zillolova G. СПАВНИТЕЛЬНЫЙ АНАЛИЗ ДВУХ ПЕРЕВОДНЫХ ВЕРСИЙ ХУДОЖЕСТВЕННОГО ТЕКСТА // *Scientific and innovative research in the social and humanitarian sphere*. – 2025. – Т. 2. – №. 4. – С. 183-188.
 32. Yamada K., Knight K. A decoder for syntax-based statistical MT // *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA, 2002. P. 303–310.