IInTÆ
Universidad Nacional de Catamarca - Facultad de Humanidades
Instituto de Investigación en Teorías del Arte y Estética
Revista Rigel – ISSN 2525-1945.

Machine Translation Between Text and Structure: A Computational Study in Transferring Arabic Structures

Dr. Imane Arioua

University of Mohamed Boudiaf, M'sila

Email: imane.arioua@univ-msila.dz

Received: 15/03/2026 Accepted: 11/05/2026 Published:25/08/2026

Abstract

This research provides an extensive and in-depth examination of the complex and interconnected challenges facing Machine Translation (MT) systems when processing Arabic texts, with a specific focus on structural and syntactic transfer levels. Given the exceptional and unique characteristics of the Arabic language—such as its immense morphological richness, complex derivational systems, word order flexibility in both verbal and nominal sentences, as well as phenomena involving null subjects and syntactic ellipsis—computational models and artificial intelligence face immense difficulties in maintaining semantic accuracy and structural coherence. These challenges are especially pronounced when transferring meaning to and from languages with fundamentally divergent structural paradigms, such as English. The present study reviews the historical and technical trajectory of computational approaches in this domain, beginning with Rule-Based Machine Translation (RBMT), transitioning through Statistical Machine Translation (SMT), and culminating in the current paradigm shift represented by Neural Machine Translation (NMT) and Large Language Models (LLMs). Furthermore, the research sheds light on how these advanced models process divergent Arabic syntactic structures and proposes future trajectories based on neuro-symbolic artificial intelligence, which aims to integrate explicit linguistic rules with neural networks to enhance overall translation fidelity and structural coherence.

Keywords: Natural Language Processing, Machine Translation, Neural Translation, Syntactic Structures, Computational Linguistics, Large Language Models, Neuro-Symbolic AI.

1. Introduction

The field of Natural Language Processing (NLP) has witnessed monumental and qualitative leaps over the past decade, driven primarily by the emergence and rapid evolution of Deep Learning technologies and Large Language Models (LLMs). Through these advancements, the computational machine has transcended the primitive stage of literal translation—which relied heavily on simple, bilingual dictionaries and rigid, word-for-word mapping—and has entered an era characterized by the simulation of contextual understanding and the generation of highly complex syntactic structures. The integration of advanced algorithms has enabled systems to interpret meaning beyond the surface level, capturing the nuances of human expression to a degree previously considered unattainable.

Despite these remarkable technological strides, the process of machine translation to and from the Arabic language remains a distinctly formidable research and technical challenge. This difficulty is largely rooted in the profound structural gap between Arabic's syntactically strict yet highly flexible grammatical system and the inherently different architectures of Indo-European languages, upon which the majority of foundational NLP algorithms and computational theories were originally built. Arabic, as a Semitic language, operates on principles of non-concatenative morphology and extensive affixation, which stand in stark contrast to the analytic nature of languages like English.

Successful translational transfer is not merely a matter of identifying lexical equivalents or substituting one word for another; rather, it extends deeply into the complete reconstruction of the syntactic architecture to align with the typological nature of the target language. This reconstructive process requires the meticulous deconstruction of the original Arabic text into its most fundamental morphological and syntactic units. The system must understand the complex network of relations binding these units together, and subsequently generate these relationships within the linguistic templates of the target language. This imposes exponentially higher computational burdens when dealing with a language like Arabic, which is characterized by heavy agglutination. In Arabic, a single orthographic word can encapsulate multiple prefixes and suffixes representing prepositions, attached pronouns, and markers for gender and plurality, functioning effectively as a complete sentence in English.

This research aims to dissect these challenges with precision and to review how successive generations of machine translation systems have attempted to overcome them. By tracing the evolutionary arc of computational linguistics—from early rule-based systems to the latest state-of-the-art technologies employing Attention Mechanisms and Transformer architectures—we seek to understand the mechanisms of structural transfer. Furthermore, this study aims to establish a forward-looking vision for developing more accurate, context-aware systems that respect and accommodate the unique linguistic idiosyncrasies of the Arabic language.

2. Theoretical Framework: Computational Linguistics and Translation

Computational linguistics is an interdisciplinary field of study that bridges computer science, artificial intelligence, and theoretical linguistics. Its primary objective is the construction of mathematical and algorithmic models capable of processing, understanding, and producing human languages in a computationally efficient manner. Within the specific context of machine translation, achieving fluency and accuracy requires a precise mathematical representation of the complex rules governing the language in question. Arabic, classified as a Semitic language, presents a unique structural model that differs radically from the predominantly sequential, concatenative languages that shaped early computational theories.

The theoretical analysis of text within machine translation is conventionally divided into several hierarchical levels: the morphological level (dealing with word formation and structure), the syntactic level (focusing on sentence structure and grammatical rules), the semantic level (concerned with meaning and interpretation), and the pragmatic level (analyzing context and intent). In the case of the Arabic language, there is a dense, almost inextricable overlap between the morphological and syntactic levels.

One of the defining features of this overlap is the Arabic diacritical system (Tashkeel). In Arabic, diacritical marks (short vowels) serve an essential syntactic function by indicating the grammatical case or syntactic position of a word within a sentence—determining whether a noun is functioning as a subject (nominative), an object (accusative), or a possessive modifier (genitive). Because these diacritics are largely absent in the vast majority of modern digital texts and written corpora, machine translation systems are forced to rely entirely on contextual cues to infer the correct grammatical relationships. This absence significantly elevates the complexity of mathematical modeling for probability distributions, as the system must resolve high levels of ambiguity before structural transfer can even begin.

Traditional theoretical frameworks often struggle to accommodate the non-linear, root-and-pattern system of Arabic morphology. While English computational models can easily segment words into stems and affixes (e.g., 'un-break-able'), Arabic words are formed by interdigitating root consonants with specific vowel patterns (e.g., root k-t-b forming 'kataba' (he wrote), 'kutiba' (it was written), or 'maktab' (office)). Consequently, translating these structures requires algorithms capable of sophisticated morphological disambiguation and deep syntactic parsing to establish equivalent constructs in the target language.

3. Structural Characteristics of Arabic and Computational Challenges

The Arabic language is distinguished by a set of intricate syntactic and morphological characteristics that make the formulation of robust machine translation algorithms an exceedingly complex endeavor. These characteristics are not limited to lexical richness or vocabulary size; they extend fundamentally to the core architecture and engineering of the sentence itself. The most prominent of these computational challenges can be categorized into four primary domains, each presenting unique obstacles for artificial intelligence.

3.1. The Case System and the Absence of Diacritics

Meaning in the Arabic language relies fundamentally on grammatical case endings indicated by short vowel diacritics (such as the Damma for nominative, Fatha for accusative, and Kasra for genitive). These marks delineate the specific syntactic function of every word within a given context. However, a significant challenge arises from the fact that the overwhelming majority of digital texts available for training machine learning models (corpora) are unvowelized (i.e., they lack diacritics).

This absence leads directly to the widespread phenomenon of syntactic and semantic ambiguity. A single undiacritized orthographic word can often be read and interpreted in multiple ways, yielding entirely different meanings and grammatical roles based on the assumed vowelization. For instance, the consonant string 'علم' can be read as 'alima' (he knew), 'allama' (he taught), 'ulima' (it was known), 'ilm' (science), or 'alam' (flag). Resolving this requires computational systems to employ highly advanced morphological and syntactic analyzers capable of decoding this textual cipher. The machine must deduce the correct diacritization based on statistical probabilities of word co-occurrence and deep contextual analysis, a process that demands massive computational resources and extensive, highly curated training data.

3.2. Word Order Flexibility

Unlike the English language, which generally adheres to a strict, rigid word order—typically Subject-Verb-Object (SVO)—the Arabic sentence is characterized by a high degree of structural flexibility. While the Verb-Subject-Object (VSO) sequence is frequently employed as the default or foundational pattern, Arabic grammar also permits the Subject-Verb-Object (SVO) sequence, as well as the deliberate fronting of the object before the subject (VOS or OVS) for specific rhetorical purposes, such as emphasis, specification, or poetic meter.

This structural fluidity deeply confounds traditional statistical translation models, which rely heavily on fixed distances and linear proximity between words to calculate translation probabilities. When transferring an Arabic sentence to English, the translation system must possess the capability to identify the main verb and its subject regardless of their spatial location within the sentence. It must then forcibly restructure these elements to fit the strict English SVO template. Accomplishing this requires a profound, programmatic understanding of the sentence's underlying parse tree, ensuring that syntactic dependencies are not lost during the reordering phase.

3.3. The Phenomenon of Pro-drop and Attached Pronouns

Arabic is classified typologically as a 'pro-drop' (pronoun-dropping) language. It frequently omits the independent subject pronoun if the subject is contextually understood, or if it is inherently encoded as a bound morpheme (an attached pronoun) at the beginning or end of the verb. For example, the word 'ذهبوا' (dhahabu) translates to 'they went', and 'أكتب'

(aktubu) translates to 'I write', without the need for explicit standalone pronouns for 'they' or 'I'.

This phenomenon necessitates that a machine translation system possesses the cognitive capacity to detect the absence of an explicit, independent subject. The system must then retrieve this latent or hidden subject by analyzing the verb's morphological conjugation or by scanning the preceding linguistic context. Once identified, the system must explicitly generate this subject as an independent word (pronoun generation) when translating into non-pro-drop languages like English or French, which strictly require an overt syntactic subject. Failure to accurately identify and reconstruct the dropped pronoun inevitably results in the generation of ungrammatical, fragmented, or entirely incomprehensible sentences in the target language.

3.4. Agglutination and High Morphological Derivation

Arabic is a highly derivational and heavily agglutinative language. A single word in Arabic can frequently equate to an entire, multi-word sentence in English. A prime example is the complex word 'وسيكتبونها' (wa-sa-yaktubunaha). This single orthographic unit is constructed from five distinct morphemes: the conjunction 'wa' (and) + the future marker 'sa' (will) + the present tense verb stem 'yaktubu' (write) + the plural subject pronoun 'na' (they) + the object pronoun 'ha' (it).

Translating such a complex morphological unit requires the system to perform highly accurate tokenization—segmenting the word into its precise constituent morphemes—before attempting any translation. If the system treats this agglutinated construct as a single, indivisible word, it will likely be classified as 'Out of Vocabulary' (OOV) by the machine's lexicon. This failure to segment leads to a catastrophic loss of the rich syntactic and semantic information encoded within the word's affixes, severely degrading the quality of the resulting translation.

4. Computational Approaches to Structural Transfer: Historical Evolution

4.1. Rule-Based Machine Translation (RBMT)

In its nascent stages, the computational approach to machine translation relied predominantly on Rule-Based Machine Translation (RBMT). This paradigm was built upon massive electronic dictionaries and explicit, rigid grammatical rules painstakingly programmed by teams of linguists and computer scientists. These systems functioned by conducting deep morphological and syntactic analyses of the source text, mapping the underlying structural framework to the target language via explicit transfer rules, and finally generating the translated text.

RBMT systems achieved a relative degree of accuracy in transferring Arabic structures within highly specialized, constrained domains where transformation rules (such as mapping VSO to SVO) could be programmed with precision. However, the fundamental flaw of this approach was its severe lack of flexibility. RBMT systems failed dramatically when confronted

with grammatical exceptions, colloquialisms, or complex rhetorical structures that had not been explicitly pre-programmed into their rule base. Furthermore, the continuous maintenance, updating, and expansion of these rule sets required grueling, endless human effort and exorbitant financial costs.

4.2. Statistical Machine Translation (SMT)

With the dawn of the third millennium, the availability of vastly greater computational power and massive bilingual text corpora (such as the United Nations document repositories) paved the way for Statistical Machine Translation (SMT). These models eschewed explicit grammatical rules in favor of calculating the probability of word sequences based on their frequency of occurrence in parallel texts. Algorithms such as the IBM Models and phrase-based SMT revolutionized the field.

While this approach excelled at achieving a reasonable degree of fluency in many language pairs, it struggled profoundly with the Arabic language. The phenomenon of long-distance dependencies—where syntactically related words are separated by large distances within a sentence—severely undermined the ability of statistical models (which primarily relied on limited N-gram windows) to capture the necessary grammatical context. Additionally, Arabic's immense morphological richness led to severe data sparsity. Because a single Arabic root can generate dozens of distinct word forms, statistical algorithms found it exceedingly difficult to calculate accurate translation probabilities without first executing complex, pre-processing morphological tokenization routines to reduce the vocabulary size.

5. The Golden Age: Neural Machine Translation (NMT)

The advent of Artificial Neural Networks represented a genuine revolution in the field of machine translation, initiating what is now considered the golden age of the discipline. Unlike statistical systems that fragmented sentences into isolated phrases, Neural Machine Translation (NMT) systems utilize an Encoder-Decoder architecture to process and translate the entire sentence as a single, cohesive unit, preserving global context.

5.1. Attention Mechanisms and Transformers

Initially, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were employed for text processing. However, they suffered from an inability to retain context over long sequences, often 'forgetting' the beginning of a long sentence by the time they reached the end. A monumental paradigm shift occurred with the invention of the Transformer architecture and the Self-Attention mechanism (Vaswani et al., 2017).

Thanks to these technologies, the machine gained the ability to evaluate the contextual importance and relevance of every single word in a sentence relative to every other word, entirely independently of the physical distance separating them. This advancement was critically transformative for the Arabic language. The model could now successfully link a verb at the beginning of a sentence to its subject, even if they were separated by lengthy appositive

clauses, multiple adjectives, or complex conjunctions. This technology revolutionized the accuracy of transferring complex syntactic structures, allowing for the generation of target language texts that exhibit an unprecedented level of fluency, increasingly approaching the natural cadence of a human translator.

5.2. Large Language Models (LLMs) and Arabic

As Transformer architectures evolved, giant Large Language Models (LLMs) emerged, including models like BERT and its specialized Arabic variants (such as AraBERT and CAMeLBERT). These architectures rely on massive pre-training regimens utilizing billions of Arabic words. Consequently, these models no longer merely 'translate' text mechanically; rather, they are capable of 'understanding' the deeper cultural, semantic, and structural context of the input.

LLMs are adept at resolving numerous issues related to structural ambiguity. They can implicitly infer the missing diacritical marks by leveraging the deep contextual relationships captured within their multidimensional word embeddings. Despite this spectacular success, however, neural models still occasionally encounter barriers when processing highly exceptional structural patterns. This can lead to the phenomenon known as 'linguistic hallucination,' where the system generates a translation that appears grammatically sound in the target language but is semantically erroneous or entirely disconnected from the source text's intended meaning.

6. Transferring Complex Structures: Applied Case Studies

To fully appreciate the capabilities and limitations of modern neural translation systems, it is necessary to examine specific case studies involving complex Arabic syntactic structures. These cases highlight the nuanced ways in which computational models must adapt to deep linguistic discrepancies.

6.1. Translating the Passive Voice

In the Arabic language, a verb is transformed into the passive voice not through the addition of auxiliary verbs (as in English), but primarily through internal vowel changes (apophony) applied to the root word. For example, the active verb 'كَتَبَ' (kataba - he wrote) becomes the passive 'كُتِبَ' (kutiba - it was written). Given that modern digital texts almost universally lack these critical vowel diacritics, the machine must rely entirely on its capacity to analyze the broader sentence context to determine whether a verb is in the active or passive state.

Modern neural systems have demonstrated a remarkable ability to discern this pattern. They achieve this by analyzing the sentence's arguments, recognizing when a direct object has been promoted to function as the subject placeholder (Na'ib Fa'il), and successfully transferring this structure into the corresponding passive voice syntax in English. However, these systems

still exhibit vulnerability and frequently err when dealing with very short sentences or isolated phrases that lack sufficient contextual padding to disambiguate the voice.

6.2. The Genitive Construct (Idhafa) and Adjective Agreement

The Arabic genitive construct, known as 'Idhafa' (e.g., 'جامعة الدول العربية' - literally 'University of the States the-Arab', translating to 'League of Arab States'), operates on structural principles quite different from English possession or noun-adjunct structures. In this construct, definiteness is inherited by the first noun from the second.

Furthermore, Arabic adjectives must strictly agree with the nouns they modify across multiple dimensions: gender (masculine/feminine), number (singular/dual/plural), case, and definiteness. Older statistical models frequently failed to correctly link an adjective to its true noun if another noun (such as the second term of an Idhafa) intervened. Transformer models, however, leveraging their multi-head attention mechanisms, have largely overcome this issue. By training on millions of grammatically correct sentences, the network implicitly learns the complex rules of structural agreement, tracking dependencies across intervening words with high accuracy.

7. Current Challenges in Neural Machine Translation

Despite the spectacular, field-defining successes of Neural Machine Translation (NMT), the academic and research community continues to observe persistent limitations and shortcomings in several areas when processing Arabic texts. These challenges highlight the remaining gap between machine processing and true linguistic comprehension.

First, the issue of 'Rare Words' (Low-Resource Vocabulary) remains significant. This includes archaic terminology, highly specialized historical jargon, or dialectal and colloquial expressions that frequently seep into Modern Standard Arabic texts. Neural systems, biased toward high-frequency patterns, tend to either marginalize these words entirely, substitute them with inaccurate high-frequency terms, or translate them overly literally, thereby corrupting the structural integrity and meaning of the sentence.

Second, the translation of complex Conditional Sentences poses a profound challenge. The Arabic conditional structure typically contains a conditional particle, a condition clause (protasis), and a result clause (apodosis), which is frequently introduced by the connective particle 'Fa' (ف). This structure represents a substantial hurdle for automatic transfer, particularly when the result clause contains lengthy parenthetical statements or complex nested clauses. Neural systems often lose track of the crucial dependency link between the initial conditional particle and its corresponding result clause, leading to fragmented or logically disconnected translations in English.

8. Conclusion and Future Prospects towards Hybrid AI

Overcoming the persistent obstacles of structural transfer in Arabic machine translation can no longer rely solely on the brute-force approach of scaling up training data or adding

more layers to deep neural networks. Recent empirical studies and linguistic analyses have demonstrated that an exclusive reliance on purely connectionist, deep learning models may soon reach a saturation point regarding the deep, deterministic understanding of strict grammatical rules.

Consequently, the focus of researchers in computational linguistics is increasingly pivoting toward a hybrid architectural approach known as Neuro-Symbolic Artificial Intelligence. This innovative paradigm seeks to fuse explicit, rule-based linguistic knowledge—such as deterministic parse trees, advanced morphological analyzers, and rigid grammatical constraints—directly into the architecture and loss functions of neural models. By guiding deep neural networks with strict linguistic rules, developers can prevent the system from making fundamental, rudimentary errors in grammatical agreement and syntax, simultaneously reducing its absolute dependence on unfathomably massive datasets.

To realize this ambitious objective, there is a pressing, critical need to develop and expand high-quality, fully diacritized, and syntactically annotated Arabic corpora (Treebanks) that have been rigorously vetted by human linguists. Treebanks such as the Prague Arabic Dependency Treebank (PADT) will remain the indispensable cornerstones for building and training the next generation of machine translation systems—systems capable of truly discerning the delicate, intricate differences between superficial text strings and deep syntactic structures. Ultimately, the Arabic language, with its immense richness, historical depth, and structural complexity, continues to serve as the ideal laboratory for pushing the boundaries of Artificial Intelligence and Natural Language Processing into uncharted territories.

References

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, 9-15.

Al-Batal, A. (2021). Automated processing of the Arabic language: Challenges and prospects. Journal of Computational Linguistic Studies, 15(2), 45-67. https://doi.org/10.1016/j.cals.2021.05.002

Guessoum, A., & Zantout, R. (2020). Machine translation of Arabic: Issues and solutions. Journal of King Saud University - Computer and Information Sciences, 32(5), 512-524.

Habash, N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies, 3(1), 1-187.

Hijazi, M., & Abdul Rahman, S. (2023). Neural translation models in processing complex syntactic structures. Annals of Arts and Languages, 8(4), 112-130.

Al-Khatib, Y., & Salem, M. (2022). Evaluating the quality of neural machine translation for academic texts to and from Arabic. Arab Journal of Automated Processing, 4(1), 88-105.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., ... & Habash, N. (2020). CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. Proceedings of the 12th Language Resources and Evaluation Conference, 7022-7032.

Al-Zantouti, S. (2019). The role of artificial intelligence in analyzing syntactic structures: A comparative study. Journal of Linguistics and Translation, 11(3), 210-235.