MTML 1.0: A Novel Interlingua Knowledge Representation Model for Machine Translation

M.A.S.T Goonatilleke[1], B Hettige[2], A.M.R.R Bandara[3]
Faculty of Applied Sciences, University of Sri Jayewardenepura, Nugegoda, Sri Lanka[1,3]
Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka[1]
Faculty of Computing, University of Sri Jayewardenepura, Nugegoda, Sri Lanka[2]

***Abstract*—Machine translation is one of the major areas of both computational linguistics and artificial intelligence that employs computer algorithms to automatically translate text between different natural languages. At present, the advent of Large Language Models (LLMs) has revolutionized this field, marking a significant turning point in its evolution. Despite their impressive capabilities, LLMs still fall short of achieving human-like translation due to key limitations, namely lack of transparency, explainability, and interpretability, the production of non-deterministic outputs, and insufficient support for low-resource languages. To address these challenges, incorporating human-aided translation mechanisms that reflect how the human brain performs translation is effective. Therefore, from a computer science perspective, this motivates the development of a novel hybrid machine translation approach that integrates a rule-based approach with LLM-based methods. This study presents a novel rule-based interlingual knowledge representation model named MTML 1.0 that has been designed and implemented to accurately analyze source language input and systematically structure the resulting linguistic information to facilitate applications, including target language generation and question-answering systems. The MTML 1.0 system consists of four key modules, namely the preprocessing module, morphological analyzer module, syntax analyzer module, and semantic analyzer module. Furthermore, the system has been fully implemented as a web-based application using the Python programming language, with spaCy serving as the foundation for natural language processing tasks. Finally, the functionality of the system has been demonstrated through the development of a prototype question-answering system.**

Keywords*—Machine translation; knowledge representation; LLMs; rule-based approach; hybrid approach

I. Introduction

Machine translation is one of the major branches of natural language processing, and a subfield of computational linguistics as well as artificial intelligence. It refers to machine-aided systems that automatically translate text or speech from one natural language to another natural language without human intervention [1]. The history of machine translation dates to the ninth century when Arab cryptologists introduced methods such as frequency analysis, probability, statistical information, and cryptanalysis which later influenced modern translation systems. Subsequently, the idea of machine translation resurfaced in the 1920s, whereas in 1956, the first machine translation conference marked a significant milestone in this field of research. Since then, machine translation has evolved significantly with various approaches, including dictionary-based approach, rule-based approach, corpus-based approach, statistical approach, hybrid approach, and sophisticated neural network-based approaches [2].

Presently, the field of machine translation has undergone a profound transformation with the emergence of Large Language Models (LLMs) that have revolutionized the way machines understand, generate, and translate human languages. In contrast to neural machine translation (NMT) systems that are trained specifically for translation tasks, Large Language Models (LLMs) like GPT (OpenAI), PaLM (Google), LLaMA (Meta), and mBART are general-purpose deep neural network models which are trained on vast corpora of multilingual data. Moreover, these models can perform a wide range of natural language processing tasks, including machine translation. Therefore, unlike traditional NMT systems, LLMs can translate across multiple language pairs without the need for separate models for each language pair, with great scalability and versatility [3]. Also, LLM-based machine translation provides some promising advantages, such as multilingual capability, contextual understanding, and the ability to perform zero-shot and few-shot learning.

Large Language Models (LLMs) have demonstrated remarkable capabilities and advancements in machine translation, but they still fall short of achieving truly human-like translation due to several major limitations, such as a lack of transparency, explainability, and interpretability, a non-deterministic nature, and their limited support for low-resource languages. Accordingly, to address and potentially overcome these issues, a solution is proposed by considering the human-aided translation processes inspired by the functioning of the human brain [4]. Therefore, the solution is a novel hybrid machine translation approach that combines the strengths of both rule-based methods and large language model (LLM)-based translation systems paradigms to overcome the existing drawbacks of any single methodology. The core premise of this approach is that by judiciously integrating these two paradigms, it is possible to develop a translation system that achieves greater accuracy, fluency, robustness, and human-like performance across a broader range of language pairs and domains [5][6]. However, at present, significant research progress has been made in LLM-based translation; still, there is no proper rule-based machine translation model or a comprehensive interlingua representation model to successfully capture, structure, and represent the linguistic knowledge encoded within source language input.

The aim of this study is to present the design and implementation of a novel system named the MTML 1.0 model (Machine Translation Markup Language 1.0 model), which is a novel rule-based interlingua knowledge representation model specifically designed for machine translation using markup language techniques. The novelty of this system is that it is a universal model that can be applied to any human language. Furthermore, in the implementation of this system, English has been adopted as the source language, and the main advantage of this model is that it can be further applied in a variety of domains, including the generation of target language outputs, question-answering systems, and other natural language processing tasks. Besides, this model has the potential to address several key limitations inherent in modern LLM-based machine translation approaches.

The rest of the study is organized as follows: Section II briefly describes the proposed approach to knowledge representation in machine translation, while Section III and Section IV deeply review the English language and the existing knowledge representation models, respectively. Section V discusses one of the data structure models that is used to represent knowledge, namely XML. Moreover, Section VI presents the design of the MTML 1.0 system, and Section VII explains the working procedure of the developed system. Finally, Section VIII is reserved to conclude the study by presenting future works.

II. Proposed Approach to Knowledge Representation in Machine Translation

The proposed approach of this research is based on how human beings perform translation from one language to another language which is tightly connected with the human brain.

The human brain is the most complex organ in the human body that controls memory, thoughts, emotions, vision, touch, breathing, temperature, motor skills, and many other processes effectively by sending and receiving chemical and electrical signals throughout the human body [7]. Structurally, the brain is divided into three main regions, namely the cerebrum, cerebellum, and brainstem. Fig. 1 illustrates the major parts of the human brain.

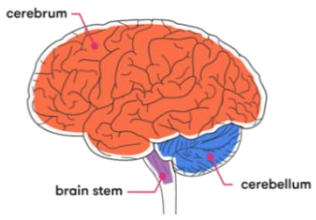

Fig. 1. Major parts of the human brain.

The cerebellum, or little brain, is located at the back of the head and performs maintaining posture, balance, and equilibrium, thinking, emotions, and social behavior. Secondly, the brainstem serves as a connection pathway between the cerebrum and the spinal cord. Thirdly, the cerebrum is the largest part of the brain that is divided into left and right hemispheres symmetrically by the longitudinal fissure. The outer part of the cerebrum is named as cerebral cortex, which is arranged as a collection of layers. This cortex is divided into fifty different functional areas named Brodmann's areas. Fig. 2 illustrates the Brodmann areas.

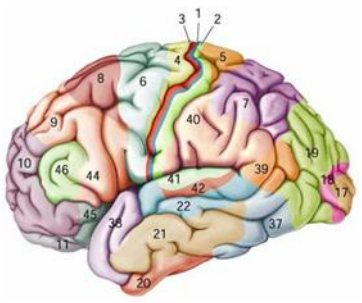

Fig. 2. Brodmann areas.

Brodmann areas were first introduced and systematically numbered in the early 20th century by the German anatomist Korbinian Brodmann based on the cytoarchitectural organization of neurons. Accordingly, Brodmann mapped the human brain and identified 52 distinct regions that were numbered from 1 to 52. These regions are associated with diverse functions such as motor control, cognition, and sensation. Notably, Brodmann area 22 is included in Wernicke's area, while areas 44 and 45 are located within Broca's area. Furthermore, both Wernicke's and Broca's areas are critical for language processing, although they perform entirely distinct functions [8]. From a translation perspective, humans must successfully comprehend and store linguistic information related to morphology, syntax, semantics, thematic roles, grammar rules, and language fluency. Achieving accurate and precise translation requires not only knowledge of grammar and linguistic structures but also language fluency, which is developed through frequent interaction and practice with the language. Within the human brain, Wernicke's area plays a key role in language fluency, while Broca's area is primarily responsible for processing grammatical rules. By simultaneously engaging both of these regions, humans are able to generate highly effective, correct, and contextually accurate translations [9].

Computationally, this scenario can be substituted with the aid of a hybrid machine translation approach that is a combination of a rule-based approach and an LLM-based translation approach. More specifically, the rule-based approach can be used to demonstrate the behavior of Broca's area, while the LLM-based translation approach can be used to demonstrate the behavior of Wernicke's area.

III. Overview of the English Language

English is the main global lingua franca that is widely used across the world in diverse domains such as communication, education, business, science and technology, tourism, trade, and the Internet. English originated in England and is currently recognized as the official or co-official language in more than fifty-nine (59) countries. Linguistically, it is classified as a West Germanic language within the Indo-European language family. The history of the English language dates back to the 5th century and is divided into four major eras, namely Old

English, Middle English, Early Modern English, and Modern English. Today, the form predominantly used worldwide is Modern English [10].

The English language comprises twenty-six (26) letters, including five (5) vowels. From a computational grammar perspective, English can be analyzed across four main areas such as grammatical units, word classes, phrases, and sentence elements. The grammatical units of English comprise words, phrases, clauses, and sentences, while the word classes (or parts of speech) are traditionally categorized into eight (8) groups, namely nouns, verbs, adjectives, adverbs, pronouns, prepositions, determiners, and conjunctions. Similarly, English phrases can be classified into five (5) types, namely noun phrases, verb phrases, adjective phrases, adverb phrases, and prepositional phrases, and finally, the sentence elements of English include the subject, verb, object, complement, and adverbial [11]. From a machine translation perspective, the English language can be systematically analyzed by considering its core linguistic elements, namely morphology, syntax, and semantics. A brief description of each of these elements is provided below.

A. English Language Morphology

Morphology is one of the major areas in linguistics that refers to the scientific study of the internal structure of words with respect to stems, root words, prefixes, and suffixes. English language morphology can be classified into four main groups such as noun morphology, adjective morphology, verb morphology, and adverb morphology. In the English language, nouns play a significant role as subjects, direct objects, indirect objects, and in various other syntactic positions. Morphologically, English nouns participate in both inflection and derivation. Inflection generates different grammatical forms of the same word, whereas derivation produces entirely new words [12]. Furthermore, based on inflectional patterns, English nouns can be classified into regular nouns and irregular nouns. Overall, English nouns conform to ten key morphological rules, which are summarized in Table I [13].

TABLE I. MORPHOLOGICAL RULES FOR ENGLISH NOUNS

Grammar	Morphology		
	Base form	***Add***	***Remove***
Singular	Noun	-	-
Plural	Noun	s	-
Plural	Noun	es	-
Plural	Noun	ies	y
Plural	Noun	ves	fe
Singular Possessive	Noun	‘s	-
Plural Possessive	Noun	s’	-
Singular	Verb	er	-
Plural	Verb	ers	-
Singular	Verb	ment	-
Plural	Verb	ments	-

Secondly, verbs in the English language exhibit five (5) primary morphological forms, such as the simple present tense, the third-person singular form, the simple past tense, the present participle, and the past participle. In addition, five major morphological rules govern the formation and transformation of English verbs, and these rules are summarized in Table II.

TABLE II. MORPHOLOGICAL RULES FOR ENGLISH VERBS

Grammar	Morphology		
	Base form	***Add***	***Remove***
Infinitive	Verb	-	-
Simple present	Verb	s	-
Present Participle	Verb	ing	-
Past	Verb	ed	-
Past Participle	Verb	ed	-

As the next morphological categories of English language, English adjective morphology shares 13 rules that are shown in Table III.

TABLE III. MORPHOLOGICAL RULES FOR ENGLISH ADJECTIVES

Grammar	Morphology		
	Base form	***Add***	***Remove***
(Positive) Adjective	Adjective	-	-
(Positive) Adjective	Noun	ish	-
(Positive) Adjective	Verb	ful	-
(Positive) Adjective	Verb	less	-
(Positive) Adjective	Noun	en	-
(Positive) Adjective	Verb	active	-
(Positive) Adjective	Verb	able	-
(Comparative) Adjective	Adjective	er	-
(Comparative) Adjective	Adjective	r	-
(Comparative) Adjective	Adjective	ier	y
(Superlative) Adjective	Adjective	est	-
(Superlative) Adjective	Adjective	st	-
(Superlative) Adjective	Adjective	iest	y

Finally, adverbs are used to provide additional information about a verb, an adjective, or another adverb. In English, many adverbs are formed by adding the suffix -ly to adjectives. Beyond this common pattern, adverbs also serve to explain aspects such as how something happens, as well as when, where, in what way, and to what extent an action occurs. Examples of these relationships are summarized in Table IV, which illustrates the connections between verbs and their corresponding adverbs.

TABLE IV. CONNECTIONS BETWEEN VERBS AND ADVERBS

Verb	Adverb	Example
When?	early	She arrives early in the morning
How?	carefully	She drives carefully
Where?	everywhere	They go everywhere together
In what way?	slowly	She walks slowly
To what extent?	slowly	It is slowly hot

B. Syntax of the English Language

The architecture of English syntax is a multifaceted domain that integrates principles of hierarchical structure, constituent arrangement, and transformational operations to govern sentence formation. Fundamentally, syntax examines how words are systematically organized to form phrases and sentences. Consequently, it encompasses key grammatical aspects such as word order, sentence structure, sentence complexity, and grammar rules. In the case of English, five basic syntactic rules can be identified, along with seven major syntactic patterns that represent the standard word order of sentences and clauses [14][15]. These are summarized in Table V and Table VI, respectively.

TABLE V. SYNTACTIC RULES IN THE ENGLISH LANGUAGE

Syntactic Rule	Description
Rule 1	All sentences need a subject and a verb, except imperative sentences.
Rule 2	In a sentence, the subject comes first, and the verb comes second. When the sentence has an object, it comes third after the verb.
Rule 3	A single sentence must contain one main idea. When a sentence contains two or more ideas, it needs to be broken up into multiple sentences.
Rule 4	Dependent clauses need a subject and a verb.
Rule 5	Adjectives and adverbs are placed in front of the appropriate words. When there are multiple adjectives describing the same noun, one needs to use the proper adjective order, known as royal order.

TABLE VI. MAJOR SYNTACTIC PATTERNS IN THE ENGLISH LANGUAGE

Pattern Type	Pattern
Pattern 1	Subject → Verb
Pattern 2	Subject → Verb → Direct Object
Pattern 3	Subject → Verb → Subject Complement
Pattern 4	Subject → Verb → Adverbial Complement
Pattern 5	Subject → Verb → Indirect Object → Direct Object
Pattern 6	Subject → Verb → Direct Object → Object Complement
Pattern 7	Subject → Verb → Direct Object → Adverbial Complement

C. Semantics of the English Language

Semantics is another major branch of linguistics that focuses on the study of meaning in language. It primarily examines how expressions are constructed across different layers of language, such as words, clauses, sentences, and texts, and how the meanings of these elements interact with and influence one another. Semantics can be viewed from two main perspectives, namely the internal side and the external side. The internal side concerns the relationship between words and mental phenomena, such as conceptual representations and ideas, while the external side investigates how words correspond to objects in the real world and the conditions under which a sentence can be judged as true or false. Furthermore, semantics can be divided into several subfields, including lexical semantics, phrasal semantics, sentence-level semantics, and paragraph-level semantics [16].

Lexical semantics is the study of lexical relations between words. The aim of lexical semantics is to analyze words in depth, as words may have multiple meanings. Lexical semantics can be divided into two approaches such as semasiology and onomasiology. The semasiology approach starts with words and examines their meanings, while onomasiology goes from meaning to word. Secondly, phrase-level semantics are used to bridge the gaps between lexical semantics and sentence-level semantics. Thirdly, sentence-level semantics are used to study the meaning of a sentence by analyzing how words and phrases combine to form meaningful statements. Moreover, thematic relationships or semantic roles are used to build the meaning of a sentence by considering the relationship between phrases presented in a sentence. Table VII shows the semantic roles of a sentence. Finally, paragraph-level semantics takes the context to contribute to meaning by analyzing the complete knowledge of each sentence in a paragraph.

TABLE VII. MAJOR SEMANTIC ROLES OF A SENTENCE

Relationship	Description
Agent	The entity that acts intentionally
Experiencer	The entity that experiences a state or sensation
Patient	The entity that is affected by the action or undergoes a change of state
Theme	The entity that will not be affected because of an action
Instrument	The means by which an action is performed, often with the help of a tool or object
Force	Mindlessly performs the action.
Location	Where the action occurs
Goal	The endpoint or destination of an action or movement
Recipient	The entity that receives something as a result of an action
Source	Where the action originated
Time	The time at which the action occurs
Beneficiary	The entity for whose benefit the action occurs
Manner	The way or method by which an action is performed
Purpose	The reason for which an action is performed
Cause	The reason why an action or state occurs

IV. KNOWLEDGE REPRESENTATION TECHNIQUES IN MACHINE TRANSLATION

Knowledge representation in machine translation is a pivotal area that focuses on developing formalisms to represent information in ways that machines can effectively process and thereby bridge the gap between different languages. Knowledge representation involves encoding human knowledge into machine-readable formats that capture meaning, context, and linguistic nuances. Furthermore, these representations enable machine translation systems to perform complex tasks requiring reasoning, inference, and understanding [17] [18]. From a machine translation point of view, knowledge representation not only maps words from one language to another, but also models morphology, syntax,

semantics, and discourse-level information and provides a foundation for reasoning, inference, and disambiguation capabilities that go beyond simple word-to-word translation. There are several major knowledge representation techniques related to machine translation, such as lexicons or bilingual dictionaries, frames, semantic networks, ontologies, interlingua representations, feature structures, and logical representations. A brief description of each of these techniques is presented below [19].

A. Lexicon

A lexicon or bilingual dictionary serves as a fundamental knowledge representation tool in machine translation, which is a structured list of word-level entries for one or more languages. It is widely employed across different MT paradigms such as the rule-based approach, hybrid approach, and neural approach. A lexicon typically consists of source-language words, their target-language equivalents, and associated linguistic information such as part-of-speech categories, morphological features, and semantic roles. It supports machine translation by enabling word lookup to find equivalent terms in the target language, facilitating disambiguation by selecting the correct sense of a word based on part of speech or context, and handling morphology effectively by adjusting word forms to match tense, number, gender, and other grammatical features. Furthermore, it supports different machine translation approaches, such as a rule-based approach as the core component that is deeply integrated with grammar rules, a hybrid approach as statistical or neural models to improve accuracy, and a neural approach as embeddings. Therefore, lexicons offer several advantages, such as providing clear word mappings, supporting low-resource languages, and enabling fine-grained control over translation. However, they have some limitations, including the time-consuming effort required for their development and maintenance, as well as difficulties in handling idiomatic expressions and context-dependent translations [20].

B. Frames

A frame is a structured knowledge representation technique introduced by Marvin Minsky in 1975. In the context of machine translation, frame representation is employed to capture semantic structures as it organizes knowledge in a way that mirrors human cognitive processes. The structure of the frame contains three components, namely frame name (the type of the concept), slots/roles (key components of the concept), and fillers (actual values in a sentence). It supports machine translation in different ways, such as facilitating semantic role labeling to identify semantic roles in a sentence, disambiguation by resolving meanings, generating target language sentences, and cross-lingual mapping. Furthermore, it supports different machine translation approaches, such as an interlingua-based approach by representing interlingual meaning using frames, a hybrid approach by combining frame semantics with statistical or neural components, and a neural approach by improving translation accuracy. Therefore, frame representation offers several advantages, such as handling variations in word order, providing better translation of idiomatic and context-rich expressions, and supporting rich semantic reasoning. However, it also has some limitations, including structural complexity, susceptibility to errors during frame extraction, and the requirement for large annotated corpora to function effectively [21].

C. Semantic Networks

Semantic networks are graph-based knowledge representation models in which concepts or words are represented as nodes and the semantic relationships between them are represented as edges. In machine translation systems, semantic networks are commonly used to support tasks such as lexical disambiguation, word sense disambiguation to identify the correct sense of a polysemous word, and synonym replacement to improve fluency by selecting contextually appropriate synonyms during target language generation. Besides, it supports different machine translation approaches, such as rule-based approaches by supporting semantic level transfer and disambiguation, a hybrid approach by improving translation fluency and accuracy via contextual knowledge, and a neural approach as an external knowledge base to enhance model understanding. Therefore, semantic networks offer several benefits, including maintaining semantic coherence, supporting meaning preservation across structurally different languages, and handling synonyms, antonyms, and context-sensitive words. Nevertheless, they also have some drawbacks, such as being labor-intensive to construct, having limited coverage across different domains, and struggling with idiomatic expressions [22].

D. Ontologies

Ontologies are formal and structured representations of knowledge that facilitate understanding the meaning of a language by accurately modeling real-world knowledge and the semantic roles associated with words. The structure of an ontology consists of five key elements such as class/concept (a category or type of thing), instance (a specific object of a class), property or attribute (describes a feature or relationship), relation (links between concepts), and axioms/ rules (constraints or logic). Ontologies support various machine translation tasks such as semantic understanding, word sense disambiguation, and cross-lingual mapping that aligns equivalent concepts across different languages to achieve more accurate translations. Moreover, it supports different machine translation approaches, such as rule-based approaches to enhance rule creation and semantic interpretation, an interlingua-based approach to define interlingual concepts to map source language input to target language output, a hybrid approach to improve context handling, and a neural approach to guide the translation via fine-tuning using external ontologies. Accordingly, ontologies offer some benefits such as improving semantic and contextual accuracy, enabling domain-specific translations, facilitating concept alignment across languages, and supporting intelligent reasoning and inference. However, it has some drawbacks, such as a labor-intensive process of creation and maintenance, limited coverage in low-resource languages, and the difficulty of integrating them with end-to-end neural models without specialized techniques [23].

E. Interlingua Representation

Interlingua serves as a universal pivot language in interlingual-based machine translation systems that enables translation across multiple source and target languages without the need for direct pairwise mappings. Interlingua captures the

deep semantic content of the source text and performs more flexible and scalable multilingual translation by abstracting away from the grammatical and lexical structures of individual languages. Primarily, an interlingua is composed of three key elements such as concepts, which represent language-independent notions; semantic roles, which define relationships between concepts; and attributes, which describe their properties. Also, it supports a rule-based approach by enabling semantic abstraction to capture meanings of source or target language grammar, context handling to represent roles and relationships, disambiguation to resolve ambiguities, and modularity to decouple language analysis from language generation. Furthermore, the interlingua approach offers several advantages, including language independence, preservation of meaning across languages, scalability through the reduction of required translation modules, and reusability as a single interlingua can support multiple target languages. However, it also presents certain drawbacks such as high complexity and domain sensitivity, high computational cost and resource intensity, and limited flexibility when compared to modern data-driven models [24].

F. Feature Structure

Feature structure is a formal framework that provides a robust mechanism for representing linguistic information through attribute–value pairs that are commonly referred to as features. This approach enables a structured and hierarchical organization of linguistic features by representing the grammatical, syntactic, and semantic properties of sentences, which are essential for effective machine translation. Fundamentally, this knowledge representation approach is employed to maintain morphological and syntactic consistency, enforce agreement rules, and support transfer rules in structural mapping between source and target languages. Furthermore, it supports different machine translation approaches, such as rule-based approaches by acting as a core component for grammar-based parsing and generation, and a hybrid approach by combining feature-based rules with statistical or neural techniques. Moreover, feature structures offer several advantages, including high expressiveness, suitability for morphologically rich languages, support for linguistic constraints and agreement checking, and facilitation of language-neutral parsing and transfer. However, they also present notable drawbacks such as structural complexity, high computational cost, limited scalability, and the need for extensive manual rule creation [25].

G. Logical Representation

Logical representation is a proper method for expressing the meaning of sentences through logic-based expressions, including propositional logic, predicate logic, and temporal logic. In machine translation, logical representation is primarily used to capture the underlying meaning of sentences at a semantic level. The structure of logical representation comprises five key elements, such as a predicate that represents properties or relationships, arguments that denote the entities involved in an action, quantifiers to express universal or existential statements, connectives that serve as logical operators, and tense, which is represented using formal logical expressions. Moreover, logical representation serves different roles in machine translation, including facilitating semantic analysis, supporting inference and reasoning, enabling disambiguation, and assisting in cross-linguistic mapping. In addition, logical representation supports various machine translation approaches, such as rule-based approaches to perform pattern matching and structure inference, an interlingua-based approach to serve as a language-neutral representation, and in hybrid or neural approaches to perform semantic annotation, evaluation, or control. Consequently, it offers some advantages, including high precision, universality, and support for reasoning. However, it also presents drawbacks such as structural complexity, high resource requirements, data scarcity, and limited suitability for idiomatic or metaphorical expressions.

V. Extensible Markup Language (XML)

Extensible Markup Language, commonly abbreviated as XML, is a markup language that has been designed to store, transport, and structure data in a platform-independent and human-readable format. Therefore, XML contains a set of rules to encode documents in a proper format that is both human-readable and machine-readable. The main purpose of this language is serialization, which involves storing, transmitting, and reconstructing arbitrary data. Generally, XML labels, categorizes, and structurally organizes information using tags. Furthermore, these tags define the data structure and contain metadata. Fundamentally, XML is used to facilitate various areas such as web services, including SOAP and RESTful services, linguistics and NLP to annotate corpora and feature structures, knowledge representation to encode ontologies, frames, and dictionaries, and different machine translation systems. There are some major features in XML, such as extensible, hierarchical, human-readable, machine-readable, and platform-independent. Moreover, XML has four key components, such as elements that can be described as a data container with opening and closing tags, attributes that store metadata inside the start tag, the root element, which is the main parent element of the XML tree, and nested elements that are elements inside other elements [26].

XML offers several benefits, such as being easy to understand and edit, highly portable and interoperable, widely supported by programming languages and tools, including Java, Python, and C#, and standardized across many domains. However, it also has notable drawbacks, including verbosity, slower parsing compared to binary formats, and inefficiency for real-time data transmission over networks.

VI. Design of the MTML Model

The Machine Translation Markup Language (MTML) 1.0 model is a novel universal rule-based interlingual knowledge representation model specifically designed for machine translation tasks using markup language techniques. The primary objective of this model is to represent and store all the linguistic knowledge extracted from the source language input, which can then be utilized to generate output in any target language. To achieve this, the system deeply analyzes the source language input by following the machine translation pyramid that is shown in Fig. 3 with respect to key linguistic levels such as morphology, syntax, and semantics.

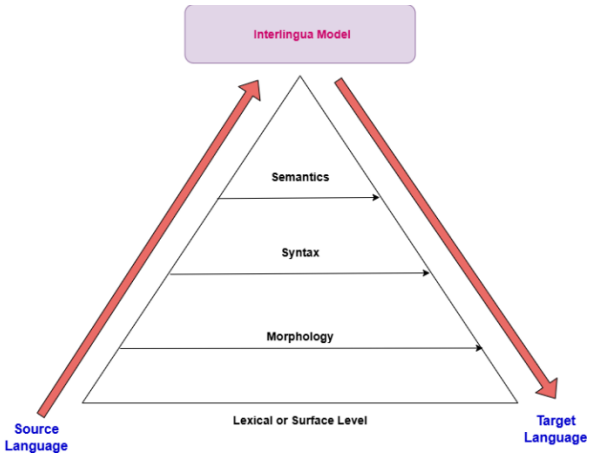

Fig. 3. Machine translation pyramid.

The extracted knowledge is subsequently represented and stored in a well-structured format, which is named as an MTML script that is a combination of XML and frame-based representation. The main advantage of this model lies in its versatility; it can be employed not only for generating target language output but also for supporting applications such as question-answering systems and other natural language processing tasks. Moreover, by following a rule-based approach, MTML 1.0 has the potential to overcome several limitations inherent in modern LLM-based machine translation approaches. Fig. 4 shows the pipeline of the MTML 1.0 model. Fig. 5 illustrates the design of the MTML 1.0 model, and the brief description of the design is provided below:

Fig. 4. Pipeline of the MTML 1.0 model.

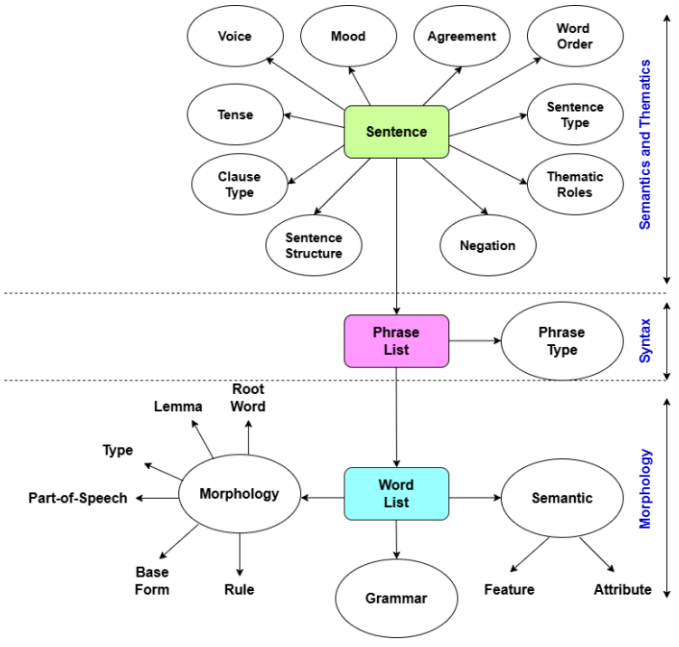

Fig. 5. Design of the MTML 1.0 model.

The MTML 1.0 model has been designed considering the common grammatical structure that is applied for any language. Furthermore, this system follows a top–down approach in its analysis. First of all, it begins with an English sentence as input and sequentially processes it through ten key linguistic properties that characterize English sentences, namely sentence structure, clause type, tense, voice, mood, agreement, word order, sentence type, negation, and thematic/semantic roles as follows:

1) Sentence structure – The model first identifies the four main structural components, namely subject, predicate, object, and modifiers.

2) Clause type – It then determines whether the sentence contains an independent clause, a dependent clause, or a relative clause.

3) Tense – The model detects the major tense (present, past, future) and further specifies the appropriate sub-tense.

4) Voice – It classifies the sentence as being in the active voice or passive voice.

5) Mood – The model identifies the speaker's intention by detecting moods such as indicative, imperative, or subjunctive.

6) Agreement – It verifies grammatical subject–verb agreement and noun–pronoun agreement.

7) Word order – The system checks if the sentence follows the standard Subject–Verb–Object (SVO) order or an inverted order (commonly used in questions).

8) Sentence type – It categorizes the sentence as declarative, interrogative, imperative, or exclamatory.

9) Negation – The model detects whether the sentence contains a negation property.

10) Thematic/semantic roles – Finally, it assigns semantic roles to constituents, including agent, experiencer, patient, theme, instrument, force, location, direction/goal, recipient, source, time, beneficiary, manner, purpose, and cause.

After completing this first stage of analysis, the model proceeds to the second stage, where it identifies and segments the sentence into five major phrase types, such as noun phrases, verb phrases, adjective phrases, adverb phrases, and prepositional phrases. These are further analyzed syntactically using a parse tree. Furthermore, in the third stage, the detected phrases are tokenized into words, and each word undergoes morphological analysis to examine its internal structure and grammatical features.

The architecture of the MTML 1.0 system is structured into four key modules, namely the Preprocessing Module, Morphological Analyzer Module, Syntactic Analyzer Module, and Semantic Analyzer Module. The overall system architecture is illustrated in Fig. 6, and a brief description of each module is provided below.

A. Preprocessing Module

This module serves as the foundational step in the MTML 1.0 system, where it accepts an English sentence as input and converts it into a structured and standardized form by functioning as a filter. Specifically, it eliminates unnecessary elements from the input sentence that are not relevant to the

analysis process, such as special characters, extra white spaces, unreadable symbols, and abbreviations. Once refined, the module produces a processed output, which is then passed to the second key component of the system, named the morphological analyzer module.

B. Morphological Analyzer Module

The morphological analyzer module is the second major module of the MTML 1.0 system, which takes the output generated by the preprocessing module as its input. The primary function of this module is to perform morphological analysis of the source sentence and identify the morphological features of each word. To achieve this, the module consists of five sub-modules, namely the tokenizer, part-of-speech tagger, morphological parser, feature extractor, and morphological output generator. At first, the tokenizer splits the sentence into individual words, which are then passed to the part-of-speech tagger to map them to their base forms with corresponding tags. Secondly, the morphological parser applies morphological rules to decompose inflected words into morphemes, and in the third step, the feature extractor identifies additional linguistic attributes such as root word, number, and word type. Finally, the morphological output generator produces the analyzed output that is enriched with structured morphological tags.

C. Syntax Analyzer Module

This module is the third major component of the MTML 1.0 system, consisting of two sub-modules, namely the syntactic parser and the syntax output generator. The primary function of the syntax analyzer module is to analyze the structure of the input sentence by searching all the phrases available in the input. In syntactic parsing, two main approaches are commonly used, namely dependency parsing and constituency parsing. For this system, a constituency parser is employed to construct the syntactic structure of the sentence. This parser detects various phrase types present in the input, such as noun phrases, verb phrases, adjective phrases, adverb phrases, and prepositional phrases. Once the parsing process is completed, the syntax output generator produces a structured representation of the sentence by visualizing its syntactic organization through a predefined tag set.

D. Semantic Analyzer Module

The semantic analyzer module is the final module of the MTML 1.0 system. Its primary responsibility is to interpret the meaning of the source language sentence and ensure that the extracted meaning is accurately preserved for subsequent applications, such as generating target language output or answering questions in a Q&A system. This module consists of three sub-modules, namely the semantic role labeler module, grammar identification module, and semantic output generator. The semantic role labeler identifies the thematic relationships present in the input sentence, while the grammar identification module handles complex grammatical tasks such as sentence structure, clause type, tense, voice, mood, agreement, word order, and sentence type. Finally, the semantic output generator produces a structured representation of the sentence's semantic features organized through a predefined tag set.

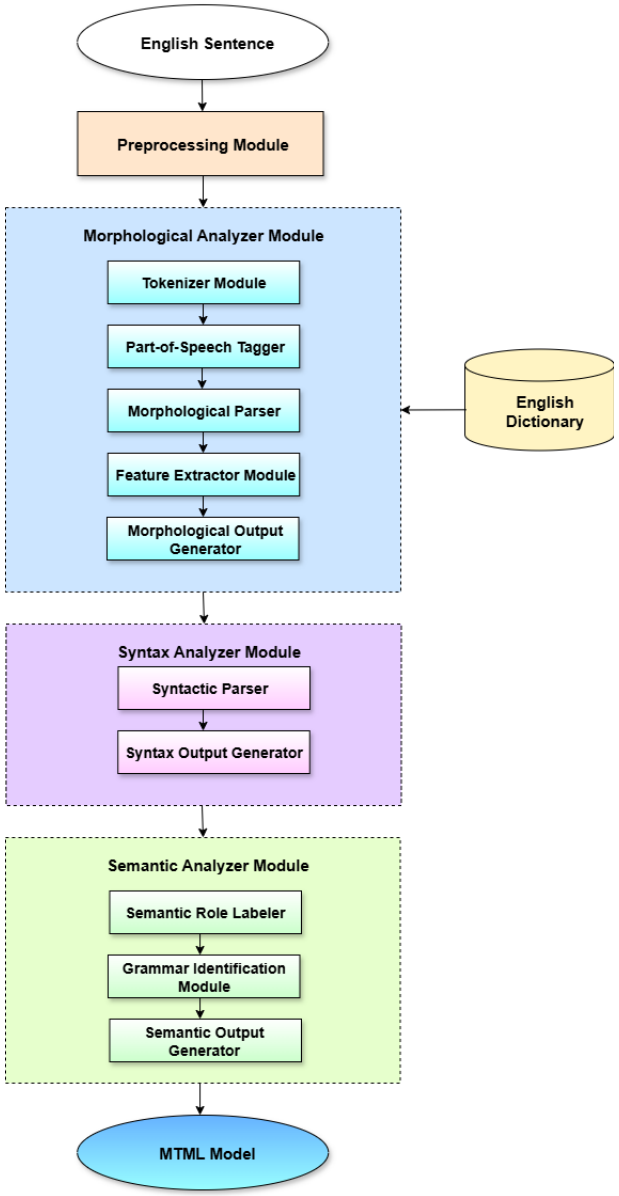

Fig. 6. Architecture of the MTML 1.0 system.

E. MTML 1.0 Model

MTML (Machine Translation Markup Language) 1.0 model represents the final output of the system that is produced after a deep analysis of the source language sentence. It can be introduced as a novel frame-based knowledge representation model specifically designed for machine translation tasks through markup language techniques. Essentially, this model represents and stores all linguistic information and knowledge contained in the source text using a structured set of tags that are automatically filled during the analysis process. The significance of this model lies in its role as a successful interlingua representation, which can be applied to various tasks such as the generation phase of machine translation systems, as well as the design and development of Q&A systems.

VII. MTML 1.0 System in Action

This section provides a brief overview of the working procedure of the MTML 1.0 system. The system has been implemented as a web-based application using the Python programming language. Furthermore, all the natural language processing (NLP) tasks within the system are modeled using spaCy, which is an advanced open-source NLP library. Unlike the Natural Language Toolkit (NLTK), which is widely adopted for educational and research purposes, spaCy is primarily designed to support NLP tasks within deep learning workflows and is well-suited for both research and industrial applications. Built on Thinc as its backend, spaCy offers a range of robust features including tokenization, part-of-speech (POS) tagging, named entity recognition (NER), dependency parsing, lemmatization, sentence segmentation, word vectors, language models, and rule-based matching, etc. Fundamentally, the MTML 1.0 system consists of two Graphical User Interfaces (GUIs) that are shown in Fig. 7 and Fig. 8, respectively.

The first graphical user interface (GUI) of the system, shown in Fig. 7, allows the user to input an English sentence. This interface includes three buttons such as "Generate", "Clear", and "Next". The "Generate" button produces the MTML script representing the linguistic knowledge extracted from the given input sentence. This script is then downloaded as an XML file, which can be utilized for further applications. The "Clear" button erases the entered input, while the "Next" button navigates the user to the second GUI, as shown in Fig. 8.

The second GUI has been designed to demonstrate a Q&A system that uses the MTML script as input to answer common questions about English sentences. In this interface, users can upload the previously generated MTML script and query the system to test its performance. Specifically, the system can respond to ten predefined questions related to English sentence analysis as follows:

1) Who did this?
2) What did you do?
3) To whom was this done?
4) When did it happen?
5) Where did it happen?
6) Why did it happen?
7) How did it happen?
8) What is the subject of the sentence?
9) What is the verb of the sentence?
10) What is the object of the sentence?

Fig. 7. GUI No. 1 of the MTML 1.0 system.

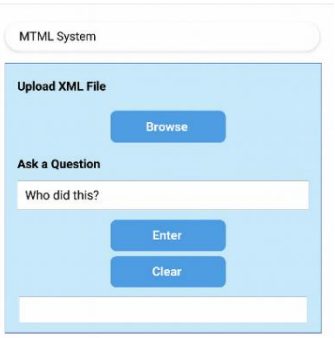

Fig. 8. GUI No. 2 of the MTML 1.0 system.

VIII. Conclusion and Future Works

This study presented the design and implementation of a novel rule-based knowledge representation model for machine translation, namely the MTML 1.0 system. The primary objective of this model is to integrate with existing LLM-based machine translation approaches in order to develop a hybrid system that more closely simulates human cognitive translation processes. The novelty of this model is that it is a universal interlingua model that can be used to analyze any natural language. In the implementation process, the model has been developed for the English language. The architecture of the MTML 1.0 system consists of four core modules, namely the preprocessing module, the morphological analyzer module, the syntax analyzer module, and the semantic analyzer module. The system has been successfully implemented as a web-based application using the Python programming language with all NLP-related tasks modeled through the spaCy library. Furthermore, the functionality of the system has been demonstrated through a prototype Q&A application, which highlights its potential for broader applications in natural language understanding and generation.

As the next stage of this research, the existing MTML 1.0 system will be extended into MTML 2.0, with the capability to address tasks that are challenging for LLM-based approaches, such as the effective handling of idiomatic expressions in the English language. In addition, the performance of the upgraded system will be systematically tested and evaluated within a controlled laboratory environment using human evaluators to ensure reliability and practical applicability.

References

[1] H. Mercan, Y. Akgün, and M. C. Odacioğlu, "The Evolution of Machine Translation: A Review Study," Int. J. Lang. Transl. Stud., vol. 4, no. 1, pp. 104–116, 2024, [Online]. Available: https://dergipark.org.tr/tr/pub/lotus.

[2] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic Machine Translation: A survey of the latest trends and challenges," Comput. Sci. Rev., vol. 38, p. 100305, Nov. 2020, doi: 10.1016/J.COSREV.2020.100305.

[3] S. Han, M. Wang, J. Zhang, D. Li, and J. Duan, "A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges," Electron. 2024, Vol. 13, Page 5040, vol. 13, no. 24, p. 5040, Dec. 2024, doi: 10.3390/ELECTRONICS13245040.

[4] W. Li et al., "TACTIC: Translation Agents with Cognitive-Theoretic Interactive Collaboration," 2025, [Online]. Available: http://arxiv.org/abs/2506.08403.

[5] H. W. Xuan, W. Li, and G. Y. Tang, "An advanced review of hybrid machine translation (HMT)," Procedia Eng., vol. 29, pp. 3017–3022, 2012, doi: 10.1016/j.proeng.2012.01.432.

[6] A. Anugu and G. Ramesh, "A Survey on Hybrid Machine Translation," E3S Web Conf., vol. 184, pp. 1–6, 2020, doi: 10.1051/e3sconf/202018401061.

[7] K. A. Maldonado and K. Alsayouri, "Physiology, Brain," StatPearls, Mar. 2023, Accessed: Sep. 02, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK551718/.

[8] K. Zilles, "Brodmann: a pioneer of human brain mapping — his impact on concepts of cortical organization," pp. 3262–3278, 2018, doi: 10.1093/brain/awy273.

[9] O. Articles, "Bidirectional Connectivity Between Broca ' s Area and Wernicke ' s Area During Interactive Verbal Communication," vol. 12, no. 3, pp. 210–222, 2022, doi: 10.1089/brain.2020.0790.

[10] F. Sharifian, "English as an international language: An overview," English as an Int. Lang. Perspect. Pedagog. Issues, pp. 1–18, 2009, doi: 10.21832/9781847691231-004.

[11] R. Hogg and D. Denison, A history of the english language. 2006.

[12] Z. Altan, "the Role of Morphological Analysis in Natural Language Processing," Anadolu Univ. J. Sci. Technol., vol. 3, no. January 2002, pp. 59–76, 2002.

[13] "Study Morphological Complexity of Non-Related Languages to Build a Universal Morphological Model for Machine Translation," Int. J. Comput. Appl. Technol. Res., Nov. 2024, doi: 10.7753/IJCATR1311.1010.

[14] "What Is Syntax? Definition, Rules, and Examples | Grammarly." https://www.grammarly.com/blog/grammar/syntax/ (accessed Sep. 02, 2025).

[15] A. Kulmizev and J. Nivre, "Schrödinger ' s tree — On syntax and neural language models."

[16] L. Rissman and A. Majid, "Thematic roles: Core knowledge or linguistic construct?," Psychon. Bull. Rev., vol. 26, no. 6, pp. 1850–1869, 2019, doi: 10.3758/s13423-019-01634-5.

[17] R. Zou, W. Chang, S. Gao, and S. Qin, "Strategies for improving the quality of computer-assisted translation based on internet," Int. J. Adv. Acad. Stud., vol. 4, no. 4, pp. 156–158, Oct. 2022, doi: 10.33545/27068919.2022.V4.I4C.890.

[18] C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating Machine Learning with Human Knowledge," iScience, vol. 23, no. 11, p. 101656, Nov. 2020, doi: 10.1016/J.ISCI.2020.101656.

[19] G. Mast, B. Hettige, and B. Amrr, "Understanding Knowledge Representation Concepts in Machine Translation: A Review," Int. J. Comput. Appl., vol. 186, no. 58, pp. 35–44, Dec. 2024, doi: 10.5120/IJCA2024924343.

[20] B. T. S. Atkins, "Bilingual dictionaries: Past, present and future," Euralex 96 Proc., pp. 515–545, 1996, [Online]. Available: http://www.euralex.org/elx_proceedings/Lexicography and Natural Language Processing/B.T.S. Atkins - Bilingual Dictionaries Past, Present and Future.pdf.

[21] B. J. Kuipers, "A Frame for Frames: Representing Knowledge for Recognition," Representation and Understanding. 1983.

[22] S. Peters and H. E. Shrobe, "Using semantic networks for knowledge representation in an intelligent environment," Proc. 1st IEEE Int. Conf. Pervasive Comput. Commun. PerCom 2003, no. November, pp. 323–329, 2003, doi: 10.1109/percom.2003.1192756.

[23] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics.," Brief. Bioinform., vol. 1, no. 4, pp. 398–414, 2000, doi: 10.1093/bib/1.4.398.

[24] H. M. Elsherif and T. R. Soomro, "Perspectives of arabic machine translation," J. Eng. Sci. Technol., vol. 12, no. 9, pp. 2315–2332, 2017.

[25] G. Penn and K. Shi, "Feature Structures in the Wild: A Case Study in Mixing Traditional Linguistic Knowledge Representation with Neural Language Models," Proc. ESSLLI 2021 Work. Comput. Semant. with Types, Fram. Relat. Struct., pp. 53–59, 2021, [Online]. Available: https://aclanthology.org/2021.cstfrs-1.6.

[26] L. Papaleo, Introduction to XML and its applications, no. September. 2013.