Hybrid-Based Machine Translation Systems

Mouiad Fadeil Alawneh[1], Tengku Mohd[2]

[1] Ajloun National University, Faculty of Information Technology, Ajloun, Jordan.
[2] National University of Malaysia, Faculty of Information Science and Technology, Bangi, Malaysia.

***Abstract*— Machine Translation (MT) is described as the method by which computer software is employed to convert text from one natural language to the other. This process includes taking into consideration each language's grammatical framework and applying examples, rules, as well as grammatical principles to adapt the grammatical structure from the source language (SL) to the target language (TL). In this paper, a method for translating well-formed English sentences into coherent Arabic sentences is introduced, utilizing grammar-based as well as example-based translation techniques to address issues related to word order and grammatical agreement. The methodology suggested is both adaptable and capable of being expanded. The primary benefits include: firstly, a hybrid approach merges the strengths of rule-based (RBMT) as well as example-based (EBMT) methodologies. Secondly, it offers the flexibility to adapt to various languages with only slight adjustments. The OAK Parser analyzes incoming English text to identify the part of speech (POS) for each word, serving as an initial step in translation, utilizing the C# programming language. To maintain data integrity, validation rules are implemented in both the database architecture as well as the programming. A key objective for this system is its capability to function independently, including its seamless integration with broader MT systems for English sentences.**

***Keywords*— MT, Agreement, Word reorder, Hybrid-based, POS.**

1. INTRODUCTION

The present Machine Translation (MT) system aids the end user in comprehending English text sentences by producing accurate equivalents in the Arabic language. Agreement is a fundamental characteristic of language. Fundamentally, agreement happens when two elements in the correct arrangement display morphology that matches their joint appearance. A clear example of this linguistic process is the number agreement between a subject and a verb: a singular noun as the subject typically appears alongside a singular verb (for example, "the cat runs"), and a plural subject noun is usually paired with a plural verb (for example, "the cats run"). If the language includes number marking on other components like adjectives or determiners, then these elements must display morphological consistency in their connection with the subject head noun. Additionally, this co-occurrence relationship applies to both gender as well as person agreement.

Contemporary Arabic dialects are recognized for displaying agreement discrepancies influenced by the order of words. These discrepancies are thought to arise from several factors: initially, from issues related to analyzing the SL, and subsequently, from challenges in generating the TLs. Yet, Arabic is not unique in exhibiting asymmetries in agreement due to word order. Asymmetries have been observed in languages such as Russian, Slovene, Hindi, French as well as Italian, as noted by Hutchins and Somers (1992). The agreement requirements across languages differ; for instance, Arabic demands agreement in number, person, gender, as well as case. Meanwhile, other languages may require only some of these forms of agreement. The development of MT systems is guided by four approaches, which vary based on their level of complexity as well as difficulty.

This paper explores various methodologies for MT, including rule-based, corpus-based, knowledge-based as well as hybrid MT systems. According to Mohammd and Sembok (2007), rule-based MT systems are further divided into direct MT, interlingua MT as well as transfer-based MT categories. The aim of this study is to develop a framework based on a hybrid approach that combines rule-based as well as example-based strategies. This approach seeks to achieve a balance between these methods for text translation and to address issues related to word agreement and sentence ordering in the translation process from English to Arabic.

2. HYBRID-BASED MT SYSTEMS

Over the last ten years, the development of new methods and the launch of innovative applications for automated translation have underscored the drawbacks of relying solely on a single strategy for translation challenges. Previously, numerous MT initiatives were initiated by researchers viewing MT as a proving ground for specific theories or methods, often leading to results that were either ambiguous or had narrow applicability. It has become increasingly acknowledged that achieving high-quality automated translation necessitates more than just one approach and that the forthcoming model will likely be 'hybrids' that integrate the strengths of statistical-based, rule-based, or example-based methodologies.
Figure 1 illustrates our methodologies in RBMT, where the translation process relies on employing bilingual

DOI:http://doi.org/10.24086/cocos2024/paper.1517

dictionaries and rules to transform SL structures into target language (TL) structures. This involves using these dictionaries and rules to create intermediate representations from which the final output is generated. Initially, the SL input strings are processed to identify suitable translation units and relationships. Subsequently, during the synthesis phase, TL texts are generated based on the TL structures or representations formulated during the stage of conversion. The functioning of the Example-Based Machine Translation (EBMT) system relies on identifying or retrieving examples of sentences in the TL that resemble the sentences in the SL being inputted. Before finding suitable translated sentences, there is a preliminary stage where input sentences are broken down into relevant segments. The stages of analysis (breaking down) and synthesis (putting back together) are specifically structured to read the input text for comparison with the database and to generate the translated output text. In HYBRID MT, rules similar to those in RBMT may be employed when a specific instance of the SL is not present in the machine's database for translation into the TL. This method has positioned HYBRID MT as an improved translation technique, effectively combining elements of EBMT with RBMT. By utilizing a matching examples strategy for translation, it has proven to be more effective across a broader range of languages.

A hybrid approach can be adapted for additional languages with slight adjustments. The database is built to be adaptable, with the majority of rules set out in tables, including those for grammar, lexicon, morphology, irregularities, derivation as well as examples of translations. Concurrently, the Parser is utilized during the morphological analysis stage to identify the parts of speech of words in the SL. Ambati and Rohini (2007) introduced a proficient English-Indian MT system designed to address challenges associated with Indian languages.

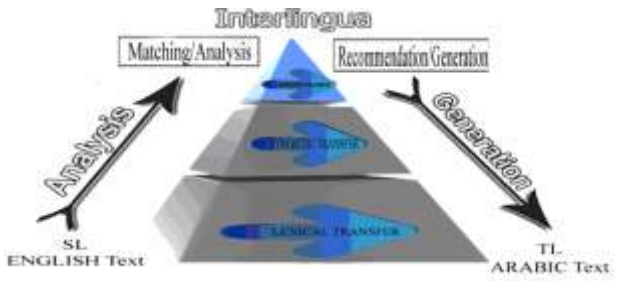

Fig 1: Hybrid based MT

3. GRAMMARS

The grammar includes patterns in English and their corresponding Arabic patterns, along with additional details necessary for implementing the rules of agreement and rearranging words in the resulting Arabic translation. An example is provided, followed by an explanation of the table's contents and their intended uses.

Example: That mad woman loved that crazy driver
Translation: احبت تلك المرأة الحمقاء ذلك السائق المجنون
[*ahbat telka almr`ah alhmka dalek alsae`g almjnoon*]

English Pattern	DT/1;JJ/2;NNX/3;VBX/4;DT/5;JJ/6;NNX/7	In this structure, a determinant and an adjective come before the subject, and the verb comes before an object, which is also preceded by a determinant and an adjective.
Subject	3	The emphasis here is on the third word being the subject.
Main verb	4	The principal action is performed by the fourth word, the verb.
Object	7	The seventh word takes the role of the object.
Verb agr.	¾	It is crucial to ensure the subject, denoted as the third word (NNX), harmonizes with the main verb, marked as the fourth word (VBX).
Adj. Agr.	2/3;6/7	Additionally, the agreement must be managed between the adjective, the second word (JJ), and the subject (NNX), as well as between the adjective, the sixth word (JJ), and the object (NNX).
Arabic Pattern	VBX/4;DT/1;XAL/1;NNX/3;XAL/1;JJ/2;DT/5;XAL/5;NNX/7;XAL/5;JJ/6	This pattern reflects the structure observed in the Arabic language, highlighting that the sequence is not merely an inverted version of the English arrangement. In this context, "XAL" symbolizes the prefix "AL alta'reef ال" attached to nouns and adjectives based on the category of the associated determinant. The numbers mentioned are consistent with those assigned to each part of speech in the English configuration, although they are not sequential.

DOI:http://doi.org/10.24086/cocos2024/paper.1517

4. RULES

The guidelines are encompassed within the "English Intelligent Rules" and "Arabic Intelligent Rules" steps, and they are categorized as follows:
- Grammar Rules
- English-Arabic Rules
- Linguistic Rules
- Translation Rules

Rules contribute to the automated translation process by capturing the broader meaning of the text, preventing our automated translator from performing literal, word-for-word translations. We posit that the incorporation of thousands of specific rules for both English and Arabic enhances the likelihood of our automated translator delivering precise translations. The rules mentioned below have been formulated through extensive text analysis. Often, we opt for implementing broad rules through coding, which enhances efficiency by minimizing search times, reducing the number of words in the database, and ensuring consistent translation outcomes.

1)

a) Ex1: He is beautiful.

b) Example: He speaks English.

Rule example one:

a) When translating into Arabic, "he" should be translated when it precedes an auxiliary verb.

b) However, if "he" does not precede an auxiliary verb, it should not be translated.

2)

a) Ex2: The book is the best friend.

b) Example: This car is the best

Rule example two: The best is to translate it as الافضل provided that it occurs at the end of a sentence, like in example b.

3)

Example: I heard the two boys speak in a low voice.

Rule example three:

=> (In the present tense for "she," the verb conjugation used for "he" is employed, with the initial ي replaced by ت.)

(in the past, for the pronoun "he," all verbs are used in their present form with the exception of altering the initial letter.)

He spoke تكلم

=> (In forming the past tense for "she," the procedure involves taking the "he" form of the verb in the past tense and appending a ت at the verb's end)

She spoke تكلمت

=> (In forming the past tense for the pronoun "they" (plural masculine), the method involves using the masculine "he" past tense form of the verb and appending _M at its conclusion. For the feminine plural, the past tense construction similarly starts with the "he" past tense verb form, followed by the addition of 2 at the verb's end).

they spoke (masculine) تكلموا

they spoke (feminine) تكلمن

4)

Example: The teacher praised both students who answered correctly.

Rule example four:

=> (The provided text translates "praise" in its noun form, which is used when the term follows a preposition, for instance, in "a praise." Conversely, when "praise" appears after a noun, it should be interpreted as a verb, and its translation is as depicted below:)

praise (to praise) مدح

both معا كلتا كلا

=> (When referring to the subject, 'both' is translated as معا.

=> "Correctly" functions as an adverb, identifiable by its "tly" suffix. In Arabic, this is conveyed by prefixing the word with بشكل or ب.

5. ADDRESSING THE ISSUES OF AGREEMENT AND WORD SEQUENCE ALTERATIONS IN MT

5.1. Reordering and Agreement with Arabic System

In this segment, we aim to investigate various aspects anticipated to lead to issues of agreement and reordering when translating from English to Arabic. The example for testing will be entered into the Arabic MT system.

Example:

Can I book tables for two good boys today?

(GOOGLE) ويمكن حجز الطاولة الاول للصبيان جيدة اليوم؟

(SYSTRAN) ? يستطيع انا كتاب الطاولة ل اثنان فتى جيد اليوم

5.2 Proposed Solution with Hybrid MT

In a hybrid-based MT approach, rules are applied similarly to RBMT but only in instances where a corresponding example from the SL intended for translation into the TL is absent in the machine's database. Following this operational method, hybrid-based MT is considered an improved translation technique over RBMT, offering advantages similar to EBMT. Fig 2 represents the whole Hybrid based MT Process.

DOI:http://doi.org/10.24086/cocos2024/paper.1517

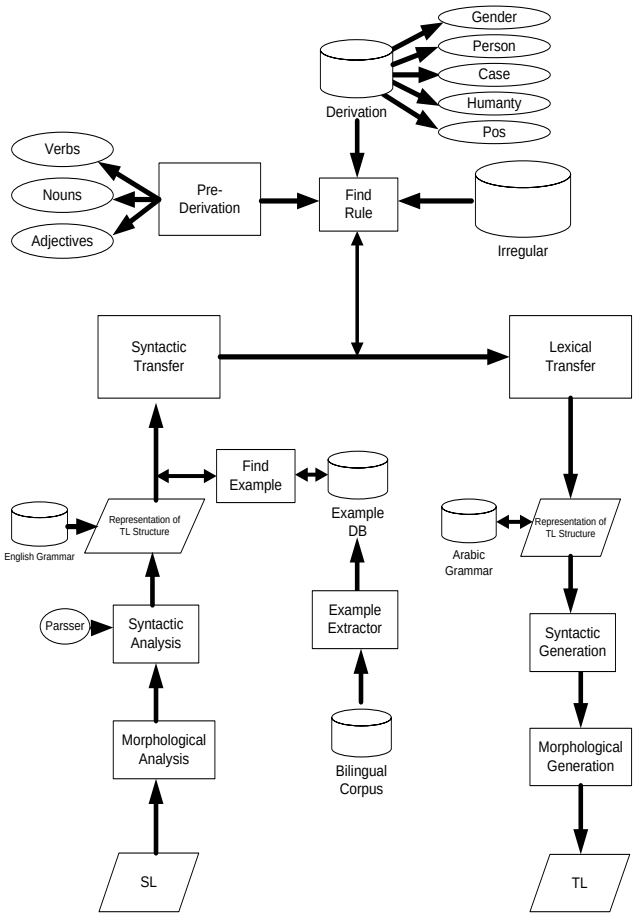

Fig 2: Hybrid based MT Process

5.3 Algorithm Steps

Let's explore how the Arabic MT system manages agreement and word-ordering by examining its translation capabilities, employing a hybrid-based MT approach in accordance with the prescribed method steps:

STEP 1: The source text in the English language is inserted

STEP 2: Submit the input text to the OAK Parser and obtain the result in the form of tagged POS.

STEP 3: Construct the English pattern based on the output provided in output 2, following the format of the grammar table.

STEP 4: Verify the protocol upon which EBMT operates as follows:
4.1 The text's alignment.

4.2 Matching input sentences with phrases (examples) stored in the database.

4.3 Selecting and extracting equivalent TL or translated phrases.

4.4 Combining translated phrases and acceptable output sentences through adaptation.

4.5 If there is no instance of the SL available in the machine's database for translation into the TL, proceed to step 5.

STEP 5: Retrieve the entry of this pattern from the grammar table to ascertain the verb, subject, agreement specifications, object as well as the corresponding pattern in the Arabic language.

POS for are, is, and am ➔ AUX
POS for were and was ➔ AUXD

STEP 6: Retrieve the characteristics and Arabic translations of all words in the sentences from the lexicon.

STEP 7: Verify any irregular word(s).

STEP 8: Enforce agreement rules for verbs and their corresponding subjects.

STEP 9: Enforce agreement rules for adjectives and the nouns they modify.

STEP 10: Apply modification guidelines to the object words.

STEP 11: Formulate the Arabic text following the patterns provided in the grammar table.

STEP 12: Steps 1 to 11 are repeated in the subsequent sentence.

هل يمكنني حجز طاواتيين لولدين جيدين اليوم

6. EVALUATION

The aim of this study is to explore whether the specified MT platforms effectively manage agreement as well as word order when translating from English to Arabic. The findings of the experiment, outlined in Table II, present the outcomes of the evaluation method applied to 100 distinct test instances.

Table II illustrates the findings of the experiment.

Grammatical Structure Form	ALKAFI	TARJIM	GOOGLE	SYSTRAN	Hybrid-based
Verb forms	75%	%89	58%	40%	95%
simple tenses	88%	91%	70%	65%	98%
Various forms of conjunction	79%	87%	71%	47%	92%
Noun phrase and Verb phrase combinations	87%	88%	69%	55%	94%
Progressive tenses	86%	90%	72%	53%	88%

7. CONCLUSION

This paper demonstrates various deficiencies in MT output resulting from errors in analyzing the SL text or generating the TL text. Improving the output requires formalizing linguistic knowledge and enhancing the computer with sufficient rules to address linguistic phenomena. Although

DOI:http://doi.org/10.24086/cocos2024/paper.1517

fully automated, high-quality MT (FAHQMT) remains elusive, there are ample opportunities to enhance MT output quality and utility.

This paper highlights the importance of addressing both agreement and word reordering in English to Arabic MT. We introduced a hybrid-based method to address these issues. The paper examines two critical factors impacting MT output: agreement and order issues. These challenges stem from the varying text orientations of different languages, some left-to-right and others right-to-left, as well as differences in word order between languages.

REFERENCES

[1] **Michelle Wendy Tan** ,2008.` acooperating hybrid mt enviroment using RULE-BASED and EXAMPLE-BASED paradigrams, manila ,philipines..

[2] **Attia, M**. 2002. 'Implications of the Agreement Features in Machine Translation'. AL-AZHAR UNIVERSITY

[3] **mohammd, and Sembok**, T. 2007a. 'TOWARD FULLY AUTOMATED ARABIC MACHINE TRANSLATION SYSTEM', IJCSNS International Journal of Computer Science and Network Security, 7 (5): 1-10.

[4] **Franck, J. Lassi, G Frauenfelder, U. & Rizzi, L**. 2006. 'Agreement and movement: A syntactic analysis of attraction'. Cognition, (101): 173-216.

[5] **Hutchins, W. and Somers. L.** 1992. 'An Introduction to Machine Translation'. London: Academic Press. Love P.E.D and Irani Z. 2003. 'A project management quality cost information system for the construction industry'. Information and Management, 40(7): 649-661.

[6] **Mohammad, M.** 1990. 'The problem of subject-verb agreement in Arabic: Towards a solution', Amsterdam, Benjamins, Publishing Company: 95-125.

[7] **mohammd, and Sembok**, T. 2007b. 'HANDLING AGREEMENT IN MACHINE TRANSLATION FROM ENGLISH TO ARABIC'. The 1st International Conference on Digital Communications and Computer Applications (DCCA2007). JUST: 385 – 379.

[8] **Trujillo, A.** 1999. 'Translation Engines Techniques for Machine Translation', Springer – Verlag Berlin Heidelberg, New Work.

[9] **Satoshi, S.** 2008, 'The manual of Apple Pie Parser v7.0' Computer science department, New York university.

[10] **H.Yamout, K.Kanso**. 2006, 'online English-Arabic Translater Electrical & Computer Engineering department, AUB university.

.

DOI:http://doi.org/10.24086/cocos2024/paper.1517