Machine Translation of Arabic Language: Challenges and Keys

Eltayeb Abuelyaman
Department of Computer Science
University of Nizwa
Nizwa, Oman
Abuelyaman@UNizwa.edu.OM

Wafaa Mukhtar
Department of Computer Science
Sudan University of Science & Technology
Khartoum, Sudan
wafaa.faisal@gmail.com

Limia Rahmatallah
Department of Computer Science
Sudan University of Science & Technology
Khartoum, Sudan
limiarahamatalla@yahoo.com

Mona Elagabani
Department of Computer Science
Sudan University of Science & Technology
Khartoum, Sudan
Minaena@gmail.com

Abstract—The morphologically rich Arabic language continues to present challenges to Arabic Word Sense Disambiguation researchers. The challenges are quite clear when one compares the gradually emerging Arabic to English Machine Translation with the rapidly developing Machine Translations available today. The lagging of the former may be attributed to two reasons. The first is the impenetrability of some Arabic words, which puzzles their decomposition into morphemes. The second is the incompatibility of existing machine-translation techniques with the Arabic language. The degree of difficulty can easily be observed by checking the results of any artless round-trip machine translation of arbitrary Arabic texts. Even the most advanced technologies that are available today are challenged. Regardless of whether the round trip is Arabic-to-English-to-Arabic or English-to-Arabic-to-English, the inevitable outcome will be disenchanting. This leads to an even more disappointing conclusion. That is, as long as no drastic improvements are made, one must not take output of any machine translation of Arabic as orthodox. Yet another disappointing fact is that most of the intelligent machines are built to produce woodenly literal or paraphrased translations. This paper proposes a framework for a biphasic Arabic translation process. Besides, the paper recommends the all-inclusive Holey Quran’s English translation as the most reliable benchmark for Arabic to English Translations.

Keywords—Natural Language processing; word sense disambiguation; machine translation; analysis; synthesis; permutations; mapping.

I. INTRODUCTION

Natural languages (NLs) are integral to our lives as means by which people communicate and document information. The power of NLs is a reality that should not be taken for granted. To learn more about, and take further advantage of such power, researchers instituted the vivid science of computational linguistics [1]. Such science mimics human processing and analysis by means of machines for the purpose of more words-power discoveries from NLs. This, in essence, leads to yet a brand new science this paper calls “the Natural Languages Mining”.

Among challenges in the study of NLs is the automatic disambiguation of word senses. According to Wilks and Stevenson [2], the task of Word Sense Disambiguation (WSD) is not an end one. Persuasively, WSD is the driving gear for Machine Translation (MT). In his technical report on MT, Sergei Nirenburg in [3] reemphasized Bar-Hillel’s position which takes the inherent difficulty of sense disambiguation as a central point. Bar Hilal, as a pioneer of MT research, considers reaching quality MT an infeasible task [4]. For more, interested readers are referred to information in [5-9]. For a quick reference thereof, the article by John Hutchins interested readers are referred to information in [5-9]. For a quick reference thereof, the article by John Hutchins

This paper focuses on challenges facing A2E-MT due to the bewildement the Arabic Word Sense Disambiguation (AWSD) brings. Nevertheless, the conclusions drawn in this paper will certainly apply to translation of Arabic to and from any other languages using parsers.

None of the available databases, dictionaries, lexicons, corpuses, statistical parsing, part of speech tagging or even techniques the likes of example-based translation will be conferred with at this stage. Interested readers are referred
to [12] for the latest work in these areas. To contribute to A2E-MT, this paper proposes two recommendations:

- The first is to develop a two-phase A2E-MT which analyzes an Arabic text during a top-down phase and then translates it into English during the bottom-up phase.
- The second is to adopt a standard benchmark for assessing performance of the first.

Each of these recommendations is introduced next.

A. The Top-Down Phase

The top-down phase is carried out by a process that analyzes Arabic texts. It basically decomposes a text into its atomic components. Analysis is the only top-down stage.

B. The Bottom-Up Phase

The bottom-up phase includes the following stages:

- **Mapping**: Maps every monosemous Arabic word as well as each sense of polysemeous ones to their appropriate English translation.
- **Permutation**: Generates a class of one or more sentence(s) constructed from translated words and their respective synonym(s). One of the sentences is expected to befit the intended sense of its counterpart Arabic sentence.
- **Synthesis**: Complements the permutation stage by selecting the most suitable translation for the Arabic sentence.

The permutation stage is not directly emphasized in the well-known pyramid of MT translation [13].

C. Adoption of a Standard Benchmark

The second recommendation is to make the English Translation of the meanings of the Quran (ETQ) a standard benchmark for A2E-MT at the part of speech level. This is due to the fact that no Arabic text has nearly as many translations as the Quran. On the other hand, no translated Arabic text had undergone as much scrutiny, or as rigorous assessment as did the ETQ.

The choice of an ETQ translation for benchmarking was difficult. Such decision was made after inspection of tens of candidates. The Noble Quran Translation, which is published by King Fahad Complex for Printing Quran [14], emerged as the most favorite choice. It is a paraphrased copy of Abdullah Yusuf Ali’s most widely used ETQ [15]. A list of the candidate ETQs and the selection criteria are not included due to page limitation. Such information will soon appear in a follow up paper.

More on the characteristics of the Arabic language is given in section II. The complexity of the Arabic language is argued in section III. Section IV introduces a framework for an A2E-MT. The scalability control is in section V. A comparison between Google Translate and the proposed MT is presented in section VI. The conclusion is drawn in section VII.

II. THE ARABIC LANGUAGE

Like Hebrew, Arabic is a highly inflectional Semitic language. Morphologically, it is characterized by the following features:

- The total number of Arabic letters is 28;
- Most of the Arabic words are derivable from, and transformable back into common roots;
- Approximately 85% of Arabic words are derived from tri-lateral roots;
- Arabic’s common nouns and particles are not transformable;
- Arabic nouns and verbs are derived from a closed set of around 10,000 roots;
- Arabic nouns and verbs have meanings and can be used as adjectives [underlined] (الرجل الطاهر The pure man) or adverbs [underlined] (Meshi beskieh He tranquilly walked);
- Arabic recognizes three genders: feminine, masculine and neuter (words borrowed from other languages for example Doctor “دكتور”);
- Arabic provides three classes for addressing the number of objects: singular, dual, and plural;
- Some of the Arabic words are made up of roots and affixes.

The composition of words from roots and affixes is discussed next.

A. Compound Words

In Arabic, a word can be coined from a morpheme and affixation as shown by the example in the following table. Since there is hardly any difference between complex and compound words in Arabic this paper uses compound words for both.

Cells in the first column are headers of their respective rows. The first row shows the example compound-word. The second breaks down the compound-ward into its four morphemes.

<table>
<thead>
<tr>
<th>word</th>
<th>هم للههم</th>
</tr>
</thead>
<tbody>
<tr>
<td>components</td>
<td>هم</td>
</tr>
<tr>
<td>transliteration</td>
<td>himm</td>
</tr>
</tbody>
</table>

The Noble Quran Translation, which is published by King Fahad Complex for Printing Quran [14], emerged as the most favorite choice. It is a paraphrased copy of Abdullah Yusuf Ali’s most widely used ETQ [15]. A list of the candidate ETQs and the selection criteria are not included due to page limitation. Such information will soon appear in a follow up paper.

More on the characteristics of the Arabic language is given in section II. The complexity of the Arabic language is argued in section III. Section IV introduces a framework for an A2E-MT. The scalability control is in section V. A comparison between Google Translate and the proposed MT is presented in section VI. The conclusion is drawn in section VII.
The third and fourth rows are the transliteration and translation of each morpheme, respectively. For the translation to be tangible, it must be rearranged (permuted), as shown by the arrows, into the phrase:

"and by their say."

The arrows show the necessary permutation that produces a palpable phrase.

Compound words processing is one of the many challenges Arabic processing researchers face. Such challenges are discussed in the next section.

B. The Differences Between Arabic and English

Arabic and English are different from the morphological and syntactic perspectives [12]. This poses a challenge to the Arabic language researchers who wish to take advantage of existing English language processing technologies. Moreover, Arabic verbs are marked explicitly for multiple forms indicating time of the action, voice and person. They are also marked with mood (indicative, imperative and interrogative). For nominal (nouns, adjectives, proper names), Arabic marks case (accusative, genitive and nominative), number, gender and definiteness features. Arabic writing is also known for its incredible reuse of names of the human body parts.

One interesting observation about the Arabic language is that it is loaded with polysemous words. For example, imagining the word 'head', which makes Arabic quite distinctive compared to English:

- Subject embedding;
- Relaxed sentence order;
- Subject object conflation;
- Word ambiguity;
- Letter ambiguity;
- Digrams & trigrams.

The rest of the section elaborates on each of these challenges.

C. Subject Embedding

Unlike English, Arabic is considered, from the syntactic standpoint, a pro-drop language where the subject of a verb may be implicitly encoded in its morphology. For example, the statement "He ate a chicken" can be expressed in Arabic as "أكل دجاجة" (أكل مرآة). The subject "He" and the verb "ate" are represented in Arabic language by the single verb-form "أكل". That is; "He ate" is translated as "أكل" and "a chicken" is translated as "دجاجة".

D. Relaxed Sentence Order

Arabic exhibits a larger degree of freedom in the order of words within a sentence. It allows permutation of the standard order of components of a sentence—the Subject-Verb-Object (SVO). As an example, the sentence "The men ate a bull" can be translated, word-by-word, to the Arabic SVO phrase → "كلم الرجال نورا". The latter may be permuted to the standard Arabic order of a sentence—the VSO form → "كلم الرجال نورا". Both forms preserve the objective of the sentence. Unfortunately, the word by word English translation of the same VSO form is → "Ate the men a bull." Ironically, most of the online A2E-MTs produce meaningless word by word translations the likes of this one.

E. Subject and Object Conflation

In some situations, the distinction between the subject and object in Arabic can be challenging. For example, the English Arabic sentence "كلم الله موسى" is in the VSO form. Its translation should be: 

God  ﷲ, spoke to  موسى, Moses  موسى. If the noun "الله" is changed to "الله" by changing only the mark on its leftmost letter, the form still remains a VSO one. However, the intended subject becomes the object and vice versa since the translation will then be: 

Moses  موسى, spoke to  ﷲ, God  ﷲ. Such a simple example demonstrates that ambiguity is highly probable if no clarifiers are provided to help with the distinction between subjects and objects.

For A2E-MT, the situation is even worse if one considers the Arabic VOS form "كلم مها المها". Its word by word translation is: 

كلم  مھا, Maha  مھا (a female); مھا  مھا, spoke to  ﷲ, the deer. The manual translation for this sentence is: ﷲ Maha ate the deer. Strangely enough, the one-way A2E translation for this sentence using Google is: ﷲ "Ate Maha Maha" and its Round Trip Translation (RTT) is: ﷲ "Maha eats Maha". The RTT sends the catastrophic message which states "The deer ate Maha."

E. Word Ambiguity

An example of an ambiguous Arabic word is "کلل", which can be translated to any of the following three words: "empty", "imagined" or "battalion." Due to the undiacritized, unvowelized writing system, the three meanings are conflated. Table 2 below shows the degree of polysemy for a selected set of Arabic words. In general, Arabic is loaded with polysemous words.

One interesting observation about the Arabic language is its incredible reuse of names of the human body parts. For example, imagining the word 'head' one would think of the neck, nose, eyes, ears, tongue and so on. Table 3 below shows the number of senses for the word "head" and the aforesaid related parts. The first (second) row shows the names in English (Arabic).
TABLE 2. EXAMPLES OF POLYSEMOUS ARABIC WORDS

<table>
<thead>
<tr>
<th>Word</th>
<th>At least three different meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>قدر</td>
<td>quantity decree abled</td>
</tr>
<tr>
<td>العين</td>
<td>eye water gold</td>
</tr>
<tr>
<td>السائل</td>
<td>liquid beggar questioner</td>
</tr>
<tr>
<td>المثال</td>
<td>quantity measure example</td>
</tr>
<tr>
<td>الشكل</td>
<td>equal similar shape</td>
</tr>
</tbody>
</table>

The degree of polysemy for each is shown on the third row.

TABLE 3. POLYSEMY OF NAMES OF A HUMAN PARTS

<table>
<thead>
<tr>
<th>head</th>
<th>neck</th>
<th>nose</th>
<th>eye</th>
<th>ear</th>
<th>tongue</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>8</td>
<td>15</td>
<td>20</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

G. Letter ambiguity

Ambiguity is not limited to Arabic words only. Some Arabic letters when affixed to morphemes lead to ambiguous compound words.

Table 3 below shows how affixing the letter “ب”, which corresponds to b in English to an atomic-word will turn it to a compound one. Such is the case because, as a prefix, the letter “ب” takes on any of the following senses: through, in, by, for and at. Table 4 below shows only five of the ten possible roles the letter “ب” plays when prefixed to different words.

TABLE 4. AFFIXED LETTER AMBIGUITY

| Word | Translation | Word || Translation of word|| |
|------|-------------|------|---------------------|
| بركة | blessing    | بسربطكة | Through blessing   |
| البيت | The house   | بسبيت | In The house       |
| المال | The money   | بسمال | By The money       |
| اي  | What        | بساي | For What           |
| الباب | The door   | بساباب | At The door       |
| الفم  | The pen    | بسلم | Using The pen      |

The letter “ب” is not the only one that is ambiguous. Table 5 shows Arabic letters and their degrees of polysemy (count) when affixed to different words.

TABLE 5. COUNTS OF SENSES OF AMBIGUOUS WORDS

<table>
<thead>
<tr>
<th>letter</th>
<th>count</th>
<th>letter</th>
<th>count</th>
<th>letter</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ب</td>
<td>10</td>
<td>ل</td>
<td>4</td>
<td>ك</td>
<td>3</td>
</tr>
</tbody>
</table>

H. Digrams and trigrams ambiguity

In the Arabic language, some of the digrams and trigrams are not considered words. As such, they do not fall in word categories. Rather, they are considered tools that require embedding contexts before the roles they play could be determined. The following table shows the count of the roles some of the digrams and trigrams can play.

TABLE 6. COUNT OF ROLES OF DIGRAMS AND TRIGRAMS

<table>
<thead>
<tr>
<th>من</th>
<th>إلى</th>
<th>عن</th>
<th>على</th>
<th>في</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

The aforementioned are examples that show how difficult it is to construct compound Arabic words from morphemes, tools and letters. One can now imagine how complex the case would be for constructing Arabic sentences. Arabic sentences are discussed next.

III. COMPLEXITY OF ARABIC EXPRESSIONS

In this section the complexity of Arabic expressions is argued. Without loss of generality, the benchmark for this task is the sentence: “على المسؤول: يعون من المسؤول؟” A possible manual translation for the sentence is: “Is it a must for the questioned to be more knowledgeable than the questioner?”

To show the difficulty an A2E-MT faces, the number of possible senses for the benchmark will be determined next.

A. Sences of The Benchmark

The benchmark uses the letter “ت” twice: as the prefix for the compound word “على” and as the prefix for the compound word “على”. The letter “ت” has two different senses/uses. Based on Arabic dictionaries [17], the words “على” and “على” have three senses each. The table below breaks the benchmark into morphemes the count of possible senses for each is given in the second row. Since the morpheme “على” can be mistaken for a proper noun that is spelled the same way, the table shows 4 as its number of senses. Consequently, any of the two senses of the letter “ت” can be used with any of the four senses of the word “على”. That results in 8 possible meanings for the
compound word “"اعلى"”. Fortunately, the actual number of
tangible meanings for the benchmark will be less than 8.
In the same manner, one can verify, using table 6
above, that there are up to 20160 possible interpretations
for the benchmark.

Table 7. COMPONENTS OF AN AMBIGUOUS SENTENCE

| اسم | من | اسم مؤنث | اسم مذكور | اسم علم | اسم
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

In some cases, reaching the intended interpretation may
not be possible when the sentence is part of a paragraph.
That is, the context that determines the intended sense may
extend beyond a single sentence. If only two such
sentences are involved and each has up to 20160 possible
interpretations, then there would be up to 406,425,600
possible interpretations. One can imagine how many
possible interpretations would there be if the context of a
sentence goes beyond a paragraph! This is a serious
scalability issue for MTs involving Arabic language.
The next section presents a translator that suggests a
solution to the scalability issue demonstrated above.

IV. THE INTELLIGENT TRANSLATOR

The Intelligent Translator, henceforth will be called
IntelTran, analyzes the text to be translated during a top-
down phase. It then performs the mapping, permutation
and synthesis stages during a bottom-up phase.

A. The Top Down Phase

The Arabic Treebank approach in [10] is similar in
many ways to IntelTran. However, as stated in [16], the
Treebank has issues that are yet to be resolved.

IntelTran decomposes objects of texts into their atomic
components using a binary tree where an object represents
a block, a paragraph or a sentence. Starting from the object
at the apex—level zero of the tree, IntelTran decomposes
the object into two halves. The halves are considered first
generation children of the apex and are placed at level one.
Each half is then broken into two second generation
children. These children are placed at level two and so on.
The decomposition is a recursive one that starts at the
topmost object— the block, and proceeds inwards to
sentences objects after passing through the paragraphs
objects. Each object type has atomic components from
which it is coined. The atomic components for a sentence
are morphemes. The atomic components for a paragraph
are sentences and those for a block are paragraphs.

For the recursive decomposition, IntelTran uses four
stacks: the first is for paragraphs and is denoted by
paragStack; the second is for sentences and is denoted by
sentStack; the third is for morphemes and is denoted by
morphStack; the fourth is for translations and is denoted by
tranStack. The analysis stage proceeds as follows:

1. Push the block of Arabic paragraphs into the
   paragStack.
2. Pop the topmost paragraph from the paragStack.
   and decompose it into atomic sentences.
3. Push all the atomic sentences into the sentStack.
4. Pop the top most sentence from the sentStack.
   and decompose it into morphemes.
5. Push every morpheme into the morphStack.
6. If the sentStack is not empty go to step 4.
7. If the paragStack is not empty go to step 2.

IntelTran handles the scalability issue during the
bottom-up phase, which is presented next.

B. The Bottom-up Phase

In addition to translation, the bottom-up phase performs
the mapping of Arabic morphemes to their English
counterparts, synthesis of all possible sentences and
permutations to mine tangible sentences. The bottom-up
phase proceeds as follows:

1. Pop the morphemes of the next sentence from the
   morphStack.
2. Translate every morpheme and its synonyms.
3. Construct all tangible sentences using the output
   of step 2 and push them in the tranStack
4. If the morphStack is not empty go to step 1.
5. Construct all tangible paragraphs from related
   sentences in the tranStack and push them back into
   the tranStack.
6. Construct all tangible blocks from related
   paragraphs and push them back into the
   TranStack.

Step 2 is for the mapping; the rest perform synthesis &
permutations; The next section handles scalability.

V. HANDLING THE SCALABILITY PROBLEM

As an example, table 8 below shows the compound
word ""اعلى"" having only 5 tangible meanings.

The benchmark itself was shown to have 20,160
possible translations. As incredible as it may seem, a
manual application of the IntelTran steps on the
benchmark resulted in a single tangible translation.

Table 8. AN EXAMPLE OF A POLYSEMOUS WORD

<table>
<thead>
<tr>
<th>&quot;&quot;اعلى&quot;&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigating questioning</td>
</tr>
<tr>
<td>A call to Ali</td>
</tr>
<tr>
<td>Is it Ali?</td>
</tr>
<tr>
<td>The top most</td>
</tr>
<tr>
<td>More superior</td>
</tr>
</tbody>
</table>
VI. GOOGLE VS INTELTRAN

This section highlights some of the features IntelTran will enjoy over the leading online translator Google. Presumably, dealing with morphemes is what gives an edge to IntelTran. Though the degree of difficulty of doing so is unknown, such step would give A2E-MT a push toward better translation. For example, Google translates the compound word "بـقولهم" of table 1 to the phrase “and Saying.” The reason being that Google decomposed the compound word into two parts: the prefix ‘بـ', which is translated as ‘and’ and “قولهم” which is translated as ‘saying’. What Google did not translate is the sense of the infixed letter ‘ال’. However, through the pop-up menu of synonyms Google associates with each word, the translation of “بـقولهم” could be changed from “saying” to “by saying”. For better accuracy, Google needs to further decompose the compound word into its morpheme “بـقول” and the suffixed subject pronoun “هم”.

The same conclusion applies to translation of other compound words. Consider the example of a seemingly simple word “فرخوزه”. This word is loaded with a standalone VSO. Its prefix ‘فرخ’ is the verb, its infix ‘وزه’ is the subject and its suffix ‘ـه’ is the object. This word is translated by Google as ‘And Karoh’. The reason for the failure in this case is that Google decomposed the compound word to ‘فرخ’ and ‘وزه’. The prefix ‘فرخ’ Google translated as ‘and’. However, the remaining part ‘وزه’ Google failed to translate. Consequently, Google ended up giving only a transliteration for it. The correct translation for “وزه” should be ‘they respected him’, which is the translation IntelTrab is expected to produce.

Google, the leading online translation service has contributed a great deal through the translation it offers free of charge to millions. However, some unsettled issues are still prevalent and will continue to irk A2E-MT.

VII. CONCLUSION

To date, little work has been published on Arabic parsing. Arabic language has many features that are inherently challenging for WSD researchers. This complicates the task of A2E-MT because researchers must start from ground zero to handle three root causes of the aforesaid challenges. Firstly is the difficulty associated with recognizing the need for full-verbs—the likes of “is”, and adverbs-of-places—the likes of “there”. Secondly is the difficulty associated with recognizing appropriate senses of undiacritized words. Thirdly is the practice of performing translation at the compound word level. The last cause annuls the effects of affixations some of the compound words are coined from. IntelTran is still under development and no conclusions can be made regarding its accuracy. However, its strategy of starting translation from morphemes is a step on the right direction in a long lasting tunnel. Such a tunnel contains all the unsolved problems the father of MT—Bar Hilal identified as “very hard problems.” In conclusion, reaching quality A2E-MT is indeed improbable as reported by Bar Hilal. The future work of the authors includes proving that A2E-MT is an NP-Complete problem. An even more challenging future project is to involve IntelTran in mining Arabic morphemes on the road to advancing A2E-MT, thereby, solving yet another simple problem.

REFERENCES

[3] Sergei Nirenburg “Bar Hillel and Machine Translation: Then and Now” Technical Report Computing Research Laboratory, New Mexico State University Las Cruces, NM 88003, USA https://www.google.com/#q=Sergei+Nirenburg%5B%5BE%5D%5BBar +Hillel%5D%5D%5B+Machine%5D%5B+Translation%5D%5B%5A%5B+Then%5D%5D%5B++ (Visited 1/2/2013)
[11] Mohamed Maamouri, Ann Bies, and Seth Kulick, “Enhanced Annotation and Parsing of the Arabic Treebank” http://www.google.com/#q=Mohamed+Maamouri%5B%5C%2C%5D%5B+Ann+Bi +es%5D%5B+and+Seth+Kulick%5D%5B+%5C%E2%80%9C%5D%5B+Enhanced%5D%5B+ (Visited 31/12/2013)
Suk+Lee+%5B+%5C%E2%80%9C+Morphological+Analysis+for+Statistical +Machine+Translation%5D%5B+IBM+Research%5D%5B+ (Visited 2/34/2013)