The Key Challenges for Arabic Machine Translation

Manar Alkhatib and Khaled Shaalan

Abstract Translating the Arabic Language into other languages engenders multiple linguistic problems, as no two languages can match, either in the meaning given to the conforming symbols or in the ways in which such symbols are arranged in phrases and sentences. Lexical, syntactic and semantic problems arise when translating the meaning of Arabic words into English. Machine translation (MT) into morphologically rich languages (MRL) poses many challenges, from handling a complex and rich vocabulary, to designing adequate MT metrics that take morphology into consideration. We present and highlight the key challenges for Arabic language translation into English.

1 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 to take further advantage of such power, researchers instituted the intense science of computational linguistics. Such science mimics human processing and analysis by means of machines for the purpose of more word-power discoveries from NLs. This led to a still brand-new science called here "Natural Languages Mining" (Abuelyaman 2014).

Machine translation has many challenges, and can be divided into linguistic and cultural categories. Linguistic problems include lexicon, syntax, morphology, text differences, rhetorical differences, and pragmatic factors.

M. Alkhatib (✉) · K. Shaalan
The British University in Dubai, Dubai, UAE
e-mail: Manaralkhatib09@gmail.com

K. Shaalan
e-mail: Khaled.shaalan@buid.ac.ae

K. Shaalan
School of Informatics, Edinburgh, UK

© Springer International Publishing AG 2018

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_8

Cultural challenges arise for the Arab translator who may find certain phrases in Arabic have no equivalents in English. For example, the term تيمم tayammum, meaning “the Islamic act of dry ablution using a purified sand or dust, which may be performed in place of ritual washing if no clean water is readily available”, doesn’t have a synonym concept in English.

Arabic has a complex morphology compared to English. Preprocessing the Arabic source by morphological segmentation has been shown to improve the performance of Arabic Machine Translation (Lee 2004; Sadat 2006; Habash 2010) by reducing the size of the source vocabulary and improving the quality of word alignments. The morphological analyzers that cause most segmentors were developed for Modern Standard Arabic (MSA), but the different dialects of Arabic share many of the morphological affixes of MSA, and so it is not unreasonable to expect MSA segmentation to also improve Dialect Arabic to English MT (Zbib et al. 2012).

Quran is a Holy book that teaching Islam, in which, it contains the main principles of Islam and how these principles should be conducted are written. The availability of digitalized translated Quran making the work of finding written knowledge in Quran becomes less complicated, and faster, especially for non-Arabic language familiar or speaker. Machine translations for Quran are available in Internet such as the websites of Islamicity.com and Tafsir.com, and there are more than 100 websites giving access to machine translation for Quran.

Much work has been done on Modern Standard Arabic (MSA) natural language processing (NLP) and machine translation (MT). MSA offers a wealth of resources in terms of morphological analyzers, disambiguation systems, annotated data, and parallel corpora. In contrast, research on dialectal Arabic (DA), the unstandardized spoken varieties of Arabic, is still lacking in NLP in general and in MT in particular (Alkhatib 2016).

The current work on natural language processing of Dialectal Arabic text is somewhat limited, especially machine translation. Earlier studies on Dialectal Arabic MT have focused on normalizing dialectal input words into MSA equivalents before translating to English, and they deal with inputs that contain a limited fraction of dialectal words. (Sawaf 2010) presented a new MT system that is adjusted to handle dialect, spontaneous and noisy text from broadcast transmissions and internet web content. The Author described a novel approach on how to deal with Arabic dialectal data by normalizing the input text to a common form, and then processing that normalized format. He successfully processed normalized source into English using a hybrid MT. By processing the training and the test corpora, his method was able to improve the translation quality.

The Word Sense Disambiguation (WSD) concept is an integral and complex part of natural language processing. The complexity has to be resolved by methods other than human clarification. In the Quran, the verses are written in a particular style, posing a challenge for humanity to dispel any confusion and grasp the intended meaning, as some words and phrases are ambiguous because the component words convey various senses or are polysemous. Problems arise in word sense disambiguation in relation to words that do not have a well-defined meaning and when the sense requires interpretation (Mussa and Tiun 2015).

2 Challenges for Arabic Translation

2.1 Classical Arabic

The Holy Quran text has remained identical and unchanged, since its revelation, over the past 1400 years. The millions of copies of the Quran circulating in the world today match completely, to the level of a single letter. God says in the Holy Quran that he will guard the Quran book: “Surely it is we who have revealed the Exposition, and surely it is we who are its guardians”. Translating the Quran has always been problematic and difficult. Many argue that the holy Quran text cannot be mimicked in another language or form. Furthermore, the Quran’s words have shades of meanings depending on the context, making an accurate translation even more difficult. Translating the holy Quran requires more wordiness to get the meaning across, which diminishes the beautiful simplicity of the Quranic message.

The various differences between Arabic and English cause many syntactic problems when translating the Holy Quran into English. Verb tense is an obvious syntactic problem that translators usually encounter in translating the Holy Quran. Verb tense means the ‘grammatical realization of location in time’ and how location in time can be expressed in language (Sadiq 2010). In translating the Holy Quran, the verb tense form should be guided by the overall context as well as by stylistic considerations. In the Holy Quran, there is a transformation from the past tense verb to the imperfect tense verb to achieve an effect, which can pose some problems and challenges in translation. For example

إِذْ جَاءُوكُم مِّن فَوْقِكُمْ وَمِنْ أَسْفَلَ مِنكُمْ وَإِذْ زَاغَتِ الْأَبْصَارُ وَبَلَغَتِ الْقُلُوبُ الْحَنَاجِرَ وَتَظُنُّونَ بِاللَّهِ الظُّنُونَا

(Behold! they came on you from above you and from below you, and behold, the eyes became dim and the hearts gaped up to the throats, and ye imagined various (vain) thoughts about Allah! (Yusuf Ali’s Translation 2000) [Surat Al-Aḥzāb 33, verse 10].

The verbs جَاءُوكُم (Ja’ukum, comes against you’), زاغت (zaghat, grew wild) and وبلغت’ (wabalaghat’, reached) are in the past tense, but the verb وتظنون (think) moves to the present tense. This move is for the purpose of conjuring an important action in the mind as if it were happening in the present. Tenses, in Classical Arabic or in the Holy Quran, cannot be transferred literally. In some cases, they need to move to convey the intended meaning to the target audience (Ali et al. 2012).

The Holy Quran has been interpreted and translated into many languages, including African, Asian, and European languages. The first translation of the Holy Quran was for Surat Al-Fatiha into Persian during the seventh century, by Salman the Persian. Another translation of the Holy Quran was completed in 884 in Alwar (Sindh, India, now Pakistan) under the orders of Abdullah bin Umar bin Abdul Aziz.

2.2 Modern Standard Arabic

A word in Arabic is comprised of morpheme, clitics and affixation, as in the example in Table 1 “وبجلوسهم” (wabajulusihim, and by their sitting). 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 the headers of their respective rows. The first row shows the example of a compound Arabic word. The second breaks down the compound word into its four morphemes.

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 in Fig. 1, into the phrase: “and by their sitting.” و بـــ جـلـوسـهـــم The arrows show the necessary permutation that produces a palpable phrase.

Arabic has different morphological and syntactic perspectives than other languages, which creates a real challenge for Arabic language researchers who wish to take advantage of current language processing technologies, especially to and from English. Moreover, Arabic verbs are indicated explicitly for multiple forms, representing the voice, the time of the action, and the person. These are also deployed with mood (indicative, imperative and interrogative). For nominal forms (nouns, adjectives, proper names), Arabic indicates case (accusative, genitive and nominative), number, gender and definiteness features. Arabic writing is also known for being underspecified for short vowels. When the genus is spiritual or educational, the Arabic text should be fully specified to avoid ambiguity.

From the syntactic standpoint, Arabic is considered as a pro-drop language where the subject of a verb can be implicitly determined in its morphology; the subject is embedded in the verb, unlike in English. For example, the sentence: *She went to the park* can be expressed in Arabic as “ذهبت الى الحديقة” (Dhahabt ‘iilaa Alhadiqa, She went to the Park). The subject *She* and the verb *went* are represented in Arabic by the single verb-form “ذهبت” (Thahabat, went) That is, the translated phrase is *She went* to *the park*, with the last part translated as “الحديقة” (Alhadiqa, The Park).

Arabic demonstrates a larger 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), and Verb Subject Object. As an example, the sentence “"الطفل أكل الطعام"” (Alttifl ‘Akl Alttaeam, The child ate the food) can be translated,

Table 1 Compound word

Word	و بـــ جـلـوسـهـــم			
Compound	هـــم	جـلـوس	بـــ	و
Transliteration	Himm	Juloos	Bi	Wa
Translation	Their	sitting	By	And

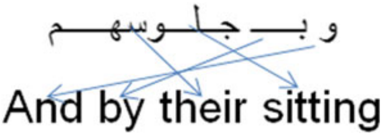

Fig. 1 Translation

word-by-word, to the English SVO phrase "the child ate the food". The latter may be permuted to the standard Arabic order of a sentence—the VSO form "أكل الطفل الطعام" "('Akl Alttifl Alttaeam, ate the child the food). Both forms preserve the objective of the sentence. Unfortunately, the word by word English translation of the same VSO form is "Ate the child the food." Ironically, most of the online translation programs produce meaningless word by word translations along the lines this one.

2.3 Dialect Arabic

Dialect is the regional, temporal or social variety of a language, distinguished by pronunciation, grammar, or vocabulary; especially a variety of speech differing from the modern standard language or classical language. A dialect is thus related to the culture of its speakers, which varies within a specific community or group of people.

Arabic Dialect poses many challenges for machine translation, especially with the lack of data resources. Since Arabic dialects are much less common in written form than in spoken form, the first challenge is to basically find instances of written Arabic dialects. The regional dialects have been classified into five main groups; Egyptian, Levantine, Gulf, Iraqi and Maghrebin.

3 Machine Translation in Natural Language Processing

3.1 Metaphor Translation

Metaphor is an expression used in everyday life communication to compare between two dissimilar things. It signifies a situation in which the unfamiliar is expressed in terms of the familiar. It is a central concept in literary studies.

Images tend to be universal in languages, as they are basically used to enhance understanding in interaction. Images, especially in speech, economize on time and effort in passing a message to its recipient. Metaphoric expressions are represented by metaphor, simile, and idioms in different languages and contexts.

Metaphor is the key figure of rhetoric, and usually implies a reference to figurative language in general. Therefore, it has always been attended to carefully by linguists, critics and writers. Traditionally, being originally a major aesthetic and rhetorical formulation, it has been analyzed and approached in terms of its constituent components (i.e. image, object, sense, etc.) and types (such as cliché, dead, anthropomorphic, recent, extended, compound, etc. metaphors). However, recently, and in the light of the latest developments of cognitive stylistics, metaphor has received even more attention from a completely different perspective, that of conceptualization and ideologization. Consequently, this change of perspective has its immediate effect on translation theory and practice, which now has to be approached differently with respect to translating metaphor. This paper is an attempt to consider the translation of metaphor from a cognitive stylistic perspective, viewing it primarily as a matter of the conceptualization of topics, objects and people (Alkhatib 2016).

Metaphor is an expression used in everyday life in languages to compare between two dissimilar things. It signifies a situation in which the unfamiliar is expressed in terms of the familiar. In addition, it is a central concept in literary studies. A metaphor is sometimes confused with a simile, especially for translators who may translate metaphor into simile or vice versa. However, it is not too difficult to decide the case of simile because of the correlative existence of simile markers like "as, similar to and like" which are not found in the metaphor (Ahmad Abdel 2014).

Simile refers to something or someone sharing a feature of something or someone else in which a significant commonality is established through one of the simile particles or through the relevant context. The rhetorical analysis of a simile requires the investigation of the two simile ends (طرفي التشبيه). These are the likened-to (المشبه) and the likened (المشبه به) entities. Simile has four components and is divided into four categories. In any simile construction, the likened should be of a higher status, as the characteristic feature is greater than that found in the likened-to. For instance, when we say كلمات كالعسل (words like honey) or وجه كالقمر (her face like the moon), we are comparing (كلمات—Kalemat, words) to (عسل—Asal, honey) in terms of sweetness and (وجه—Wajh, face) to (قمر, Qamar moon) in terms of beauty and brightness. Thus, rhetorically, the likened-to elements are represented by كلمات and وجه and the likened elements are عسل (Asal, honey) and قمر (Qamar, moon). However, the sweetness of honey and the brightness and beauty of the moon cannot be matched and are stronger than the features of the other entities.

Abdul-Raof (2006) stated that simile is realized through the following four components:

a. The likened-to (المشبه): The entity, i.e. a person or thing that is likened to another entity, which is the likened;
b. The likened (المشتبه به): The original entity to which another entity, i.e. the likened-to, is attached;
c. The simile feature: A feature that is common to both the simile ends; and
d. The simile element: The simile particles.

For example: أحمد كالأسد Ahmad Kalasad, Ahmad is like a lion, where:

- The likened to is represented by the noun) أحمد Ahmad);
- The likened is represented by the noun الأسد (Alasad, the lion);
- The simile element is represented by the particle كـ (Ka, like); and
- The simile feature is represented by the implicit notion الشجاعة (AlShaja'ah, courage), which is a semantic link that is common to and shared by both nouns الأسد and أحمد.

In Arabic rhetoric, metaphor is referred to as “الاستعارة”, which is a form of linguistic allegory and is regarded as the peak of figurative skills in spoken or written discourse. Metaphor is the master figure of speech and is a compressed analogy. Through metaphor, the communicator can turn the cognitive or abstract into a concrete phrase that can be felt, seen, or smelt. Linguistically, الاستعارة is derived from the verb اعار (A’ar, to borrow), i.e. borrowing features from someone or something and applying them to someone or something else.

Rhetorically, however, metaphor is an effective simile whose one end of the two ends, i.e. the likened-to (المشتبه) and the likened (المشتبه به), has been deleted. Metaphor represents a highly elevated effective status in Arabic rhetoric that cannot be attained by effective simile. In metaphor, the relationship between the intrinsic and non-intrinsic signification is established on the similarity between the two significations, i.e. there is a semantic link between the two meanings.

The metaphorical meaning, however, is discernible to the addressee through the lexical clue القرينة available in the speech act. In Arabic, metaphor consists of three major components. As there are different kinds of metaphor, these three components may not all be available in a single metaphor. Abdul-Raof (2006) stated that the three metaphor components are:

1. The borrowed-from: equivalent to the likened element in simile;
2. The lent-to: equivalent to the likened-to in simile; and
3. The borrowed: the borrowed lexical item taken from the borrowed-from and given to the lent-to

For example:

a. زيد أسد (Zaid Asad, Zaid is a lion). (effective simile)

b. رأيت أسدا في المدرسة (Ra’ayt ‘asadaan fi Almadrasa, I saw a lion at school). (lion refers to a brave man)

- The lent-to is represented by the noun زيد (Zaid);
- The borrowed-from is represented by the noun أسد (Asad, lion); and
- The semantic feature الشجاعة (Alshaja’a, courage) is shared by and establishes the link between زيد (Zaid) and أسد (Asad) (is the borrowed).

  In example (b), في المدرسة (Fi Almadrasa, at school) is the lexical clue to represent the metaphorical meaning of أسد lion” in this sentence, where lion refers to a brave man.

Although metaphor makes the text more beautiful and charming in the source language text (SLT) through its use of stereotyped words and new images, it can confuse the reader in the target language text (TLT) due to the linguistic and cultural differences between the two languages.

Kuiper and Allan (year) provide a definition about metaphor, as “an easy way to look at metaphor is to see the breaking down of the normal literal selection restrictions that the semantic components of words have in a sentence”. When for example, we talk about “نافذة المستقبل”, (Nafethat Almustaqbal, a window on the future), we have to ignore some of the semantic components of the word window;

for example, that it is a concrete object, and just take the fact that windows are things that allow us to look outwards from an enclosed space. The metaphor could also be seen out of a window. The metaphor lies in the suppression of some of each word's semantic features.

Metaphor can function as a means of formatting language in order to describe a certain concept, action or object to make it more comprehensive and accurate.

Hashemi (2002) classifies metaphors, i.e. isti'ara (الاستعارة), into three groups:

1. Declarative metaphors (تصريحية, Tasrihiyya): in which only the vehicle is mentioned and the tenor is deleted. In this type of isti'ara, the vehicle is explicitly stated and used to make a comparison between two different concepts that share a feature or a property in order to reveal the senses. A Declarative Metaphor is also considered as a decorative addition to ordinary plain speech. It is also used to achieve aesthetic effects (ibid). For example, in Arabic one might say (وردة, zahra, a rose) "رأيت وردة" *I saw a rose* instead of saying (a beautiful woman) امرأة جميلة, which is the vehicle in a metaphor based on the similarity between a rose and the person in terms of beauty.
2. Cognitive Metaphor (مكنية, Makniya): in which only the tenor is mentioned and the vehicle is deleted. In this type of isti'ara, the vehicle is only implied by mentioning a verb or a noun that always accompanies it. A Cognitive Metaphor is used as a means of formatting language in order to describe a certain concept, action or object to make it more comprehensive and accurate. In this case, it focuses on the denotation rather than the connotation of the metaphor that addresses the receptor in order to highlight its cognitive function.
3. Assimilative Metaphor (تمثيلي , Tamthele): which uses one of the characteristics of a vehicle for tenor. For instance "إِذا رأيتَ نُيوبَ اللَيثِ بارِزَةً فَلا تَظُنَّنَّ أَنَّ اللَيثَ يبتَسِمُ" when you see a lion baring his canines, ‹never think he is smiling.

Newmark (1988:105–113) provides another classification of metaphor, divided into six types: dead, cliché, stock, adapted, recent and original.

a. Dead metaphors

Dead metaphors are "metaphors where one is hardly conscious of the image, which relate to universal terms of space and time, the main parts of the body, general ecological features and the main human activities." Here the sense of transferred imageno longer exists. Through overuse, the metaphor has lost its figurative value. For example "خلص الوقت" (run out of time).

English words that represent dead metaphors include: "space, field, line, top, bottom, foot, mouth, arm, circle, drop, fall, and rise are particularly used graphically for the language of science to clarify or define.", some other examples are, *I didn't catch his name*, *foot filed*, *top*…etc., and an example in Arabic "عقارب الساعة" which means (hands of the clock). Dead metaphors are not difficult to translate literally; even though they could lose their figurative meaning through extensive popular use. Another example is حقل المعرفة البصرية (field of human knowledge).

b **Cliché metaphors**

Cliché Metaphors are "metaphors that have perhaps temporarily outlived their usefulness, that are used as a substitute for clear thought, often emotively, but without corresponding to the facts of the matter." One example in English would be *at the end of the day*, and an example in Arabic is في نهاية المطاف (Fi nehayat almataf).

c. **Stock or standard metaphors**

Newmark describes this kind of metaphor as "An established metaphor, in an informal context, is an efficient and concise method of covering a physical and/or mental situation both referentially and pragmatically". It has certain emotional warmth, which does not lose its brightness by overuse. These are sometimes difficult to translate since their apparent equivalents may be out of date or now used by a different social class or age group. According to Newmark, a stock metaphor that does not come naturally to you should not be used, which means, if these metaphors are unnatural or senseless in the target language, they should not be used.

d. **Recent metaphors**

Recent metaphors, where an anonymous metaphorical neologism has become something generally used in the source language. It may be a metaphor designating one of a number of 'prototypical' qualities that constantly 'renew' themselves in language. For example: تصفية الخصوم السياسية, (Tasfiyat Alkhosoom Alseyaseyah, head hunting).

e. **Adapted metaphor**

An adapted metaphor is an adaptation of an existing (stock) metaphor. This type of metaphor should be translated by an equivalent adapted metaphor; it may be incomprehensible if it is translated literally, as in "الكرة في ملعبه", (Alkora fi mal'aboh, the ball is in his court).

f. **Original metaphors**

Original metaphors refer to those created or quoted by the Source Language writers in authoritative and expressive texts. These metaphors should be translated literally, whether they are universal, cultural, or obscurely subjective.

3.2 Metaphor in Holy Quran

There are many metaphors in some verses of the Holy Qur'an. These metaphors, that the All-Wise Allah makes, are very effective and advance the understanding of those who read them. Every one of these metaphors and descriptions illustrates the subject in the most effective and the clearest way. Throughout history, critics have rarely defined this word alike (تشبيه tashbeh). The first who is known to have used the term Al-majaz is Abu Ubayda in his book, "Majazal-Quran". However, he did not mean by that Al-majaz is the counterpart of haqiqa and figurative language. He mostly uses the word in the formula: “A, its majaz is B”, where A denotes the classical word or phrase and B its “natural” equivalent. In fact, Ubayda was concerned with the first meaning of the term “majaz” which means ‘explanatory re-writing’ in ‘natural’ language of idiomatic passages in the Scripture, while the second sense of “majaz” is figurative language, which was developed later. In his Majaz Quran, Ubayda does not define *majaz*, but at the beginning of his work he does give a list of thirty nine cases of deviation from the ‘natural’ language that can be found in the Qur’an (Alshehab 2015). The following is an instance of the word “آية” Ayah from the Qur’an which is interpreted as a metaphor; a device for presenting a concept. One of the most beautiful metaphors in the Holy Qur’an is the verse:

(اللَّهُ نُورُ السَّمَاوَاتِ وَالْأَرْضِ مَثَلُ نُورِهِ كَمِشْكَاةٍ فِيهَا مِصْبَاحٌ الْمِصْبَاحُ فِي زُجَاجَةٍ الزُّجَاجَةُ كَأَنَّهَا كَوْكَبٌ دُرِّيٌّ يُوقَدُ مِنْ شَجَرَةٍ
مُبَارَكَةٍ زَيْتُونَةٍ لَا شَرْقِيَّةٍ وَلَا غَرْبِيَّةٍ يَكَادُ زَيْتُهَا يُضِيءُ وَلَوْ لَمْ تَمْسَسْهُ نَارٌ نُورٌ عَلَى نُورٍ يَهْدِي اللَّهُ لِنُورِهِ مَنْ يَشَاءُ وَيَضْرِبُ اللَّهُ
الْأَمْثَالَ لِلنَّاسِ وَاللَّهُ بِكُلِّ شَيْءٍ عَلِيمٌ)

Allah is the Light of the heavens and the earth. The example of His light is like a niche within which is a lamp, the lamp is within glass, the glass as if it were a pearly [white] star lit from [the oil of] a blessed olive tree, neither of the east nor of the west, whose oil would almost glow even if untouched by fire. Light upon light. Allah guides to His light whom He wills. And Allah presents examples for the people, and Allah is Knowing of all things. (Quran: Surah: Al-Noor, Verse 35).

When we use metaphors, it does not mean that we lie; we use metaphor to make the concepts and thoughts sharper and clearer. For instance, in the Holy Quran:

(أَلَمْ تَرَ كَيْفَ ضَرَبَ اللَّهُ مَثَلًا كَلِمَةً طَيِّبَةً كَشَجَرَةٍ طَيِّبَةٍ أَصْلُهَا ثَابِتٌ وَفَرْعُهَا فِي السَّمَاءِ (24) تُؤْتِي أُكُلَهَا كُلَّ حِينٍ بِإِذْنِ رَبِّهَا
وَيَضْرِبُ اللَّهُ الْأَمْثَالَ لِلنَّاسِ لَعَلَّهُمْ يَتَذَكَّرُونَ (25) وَمَثَلُ كَلِمَةٍ خَبِيثَةٍ كَشَجَرَةٍ خَبِيثَةٍ اجْتُثَّتْ مِنْ فَوْقِ الْأَرْضِ مَا لَهَا مِنْ قَرَارٍ
(26)

“Have you not considered how Allah presents an example: a good word is like a good tree, whose root is firmly fixed and its branches in the sky? (24) It produces its fruit all the time, by permission of its Lord. And Allah presents examples for the people that perhaps they will be reminded. (25) And the example of a bad word is like a bad tree, uprooted from the surface of the earth, not having any stability”. [Quran: Surah Ibrahim, Verse 24–26].

The metaphor here; the good word (الكلمة الطيبة) is set in similitude to a good tree (الطيبة الشجرة) that has a firm root and its branches in Heaven (sky) and gives its fruits every now and then by the will of its Lord. On the other hand, the bad word (الكلمة الخبيثة) is likened to a bad tree (الشجرة الخبيثة) which is uprooted from the earth and has no base.

The classical text is a linguistic miracle and was intended to challenge Arabs who are fluent in classic Arabic analogy, and what makes the Qur’an a miracle, is that it is impossible for a human being to compose something like it, as it lies outside the productive capacity of the nature of the Arabic language. The productive capacity of nature, concerning the Arabic language, is that any grammatically sound expression of the Arabic language will always fall within the known Arabic literary forms of prose and poetry. All of the possible combinations of Arabic words, letters and grammatical rules have been exhausted and yet their literary forms with metaphors have not been matched linguistically. The Arabs, who were known to have been Arabic linguists par excellence, failed to successfully challenge the Quran (Mohaghegh 2013).

3.3 Metaphor in Modern Standard Arabic

Metaphor is the process of ‘transporting’ qualities from one object to another, one person to another, from a thing to a person or animal, etc. When translating a metaphor, it is necessary to start by investigating the concept of metaphor, with the focus on contemporary conceptual approaches of metaphor. There have been rapid and revolutionary changes in communications, computers, and Internet technologies in recent years, along with huge changes in the conceptual studies of metaphor.

A metaphor is a figure of speech that involves a comparison, and a simile is also a figure of speech which involves a comparison. The only difference between them is that in a simile the comparison is explicitly stated, usually by a word such as “like” or “as”, while in a metaphor the comparison is implied. Machine translation is much more likely to function correctly for simile than it can for metaphor. For instance, using Google translator:

a. “اشتعل الرأس شيباً” (Eshta’al Alra’s Shayban, Flared head Chiba); and
b. “شعره كالثلج “ (Sha’aroh Kalthalj, his hair such as snow).

In the second example would help in translation, as it represents a simile, but in the first example the metaphor is implicit and so its translation is much more difficult. Another example is رأيت أسداً في المدرسة (Ra’ayt Asadan fi Almadrasa, I saw a lion in the school), it does not mean that “I saw the lion (the animal), but rather that “I saw a man like a lion in his brave demeanor”, here describing the bravery of the man like that of a lion, the king of the forest and the strongest among others.

3.4 Metaphor in Dialect Arabic

Arabic dialects, collectively referred to here as Dialectal Arabic (DA), are the day to day vernaculars spoken in the Arab world. Metaphorical expressions are pervasive in day-to-day speech. The Arabic language is a collection of historically related variants that live side by side with Modern Standard Arabic (MSA). As spoken varieties of Arabic, they differ from MSA on all levels of linguistic representation, from phonology, morphology and lexicon to syntax, semantics, and pragmatic language use. The most extreme differences are at the phonological and morphological levels. We can see the difference in meaning with the use of the word white in metaphorical expressions. For example, the expression in the dialect Arabic

سارة قلبها زي التلج (Sarah Qalbaha zay Althalj, Sara's heart [is] like snow) expresses that Sara is a good person, whereas the expression كدبة بيضة (Kedba Bedhah, a white lie) means a lie that is "honest and harmless". Another example is praise with the word "donkey" in the expression سارة حمارة شغل (Sarah Hemart Shoghol, Sara is a donkey at work) which means "She is a very patient and hard worker". However, describing a person as a donkey in the dialect Arabic is very offensive and has connotations such as foolish or stupid. In dialect metaphors, we usually use the bad words (bad expressions) to express a good adjective and the vice versa.

Dialect Metaphors expressions are day-to-day speech that people use all the time (Biadsy 2009):

- In arguments like "مافيك تدافع عن موقفك" (Mafeek Tedafe'a a;n mauqifak, you cannot defend your position) contain the word "تدافع" (defend); it must be for something like country or building. We consider the person in the argument with us as an opponent and we attack his position. Another example "حكيو ضرب علي الراس", (Haku Dhareb ala Alras, his speech is hitting it on the head), means that he is getting to the heart of the matter.
- Utilizing ideas and peech as food and commodities: "أفكاره مهضومة", (Afkaroh Mahdomeh, his ideas [are] tasty and sweet), means that his ideas are nice and appropriate, while "أفكاره بلا طعمه" (Bla Ta'meh, his ideas [are] without taste) means that they are not useful, or even harmful. Two other examples are "حط بيطنك بطيخة صيفي" (Hoot Bebatnak Batekha Saifi, Eat watermelon), which means 'relax and don't worry', and "طحن الكتب طحن" (Tahan, Elkutob Tahen, he smashed the books), which means that he studied the books thoroughly.
- To express time: "إجا وقت الجد" (Eja Waqt aljad, the time of seriousness has come) means that it is time to work hard and be serious. Other examples of time metaphors are "راح آذار'' (March went away), meaning March has ended, and "{الشتا صار على الابواب}," (Alshita sar ala alabwab, winter has reached our doorsteps), which means winter will start soon.
- Times are used as location: "نط التسعين" (Nat Altes'en, He jumped over ninety) means he is over ninety years old, and "العام الي مرق" (Alam Eli Maraq, the year that passed) means the last year, and here describes the year as a person that has walked away.

Dialect metaphors are difficult to understand correctly, unless we are familiar with them and we are from the same culture with the same dialect, as each country (and even each region) has its own metaphor dialect.

4 Named Entity Recognition Translation

The Named Entity Recognition (NER) task consists of determining and classifying proper names within an open-domain text. This Natural Language Processing task is acknowledged to be more difficult for the Arabic language, as it has such a complex morphology. NER has also been confirmed to help in Natural Language

Processing tasks such as Machine Translation, Information Retrieval and Question Answering to obtain a higher performance. NER can also be defined as a task that attempts to determine, extract, and automatically classify proper name entities into predefined classes or types in open-domain text. The importance of named entities is their pervasiveness, which is proven by the high frequency, including occurrence and co-occurrence, of named entities in corpora. Arabic is a language of rich morphology and syntax. The peculiarities and characteristics of the Arabic language pose particular challenges for NER. There has been a growing interest in addressing these challenges to encourage the development of a productive and robust Arabic Named Entity Recognition system (Shaalan 2014). End of editing 18 March (EST) —new editing from here!

The NER task was defined so that it can determine the appropriate names within an open domain text and categorize them as one of the following four classes:

1. Person: person name or family name;
2. Location: name of geographically, and defined location;
3. Organization: corporate, institute, governmental, or other organizational entity; and
4. Miscellaneous: the rest of proper names (vehicles, brand, weapons, etc.).

In the English language the determination of the named entities (NEs) in a text is a quite easy sub-task if we can use capital letters as indicators of where the NEs start and where they end. However, this is only possible when capital letters are also supported in the target language, which is not the case for the Arabic language. The absence of capital letters in the Arabic language is the main difficulty to achieving high performance in NER (Benajiba 2008; Benajiba and Rosso 2007; Shaalan 2014).

To reduce data sparseness in Arabic texts two solutions are possible: (i) Stemming: omitting all of the clitics, prefixes and suffixes that have been added to a lemma to find the needed meaning. This solution is appropriate for tasks such as Information Retrieval and Question Answering because the prepositions, articles and conjunctions are considered as stop words and are not taken into consideration when deciding whether or not a document is relevant for a query. An implementation of this solution is available in Darwish and Magdy (2014); (ii) Word segmentation: separating the different components of a word by a space (blank) character. This solution is more appropriate for NLP tasks that require maintaining the different word morphemes such as Word Sense Disambiguation, Named Entity Recognition, etc.

NER in Dialect Arabic is completely different than it is in MSA. For example, a person name in either DA or MSA could be expressed in DA by more than one form; for example, the name “قمر طارق” (Qamar Tareq) in MSD, can be “امر طارىء” (Amar Tare’a) and “كمر طارك” (Kamar Tarek); the main complication is that the first name is a girl’s name, when translated it can be ‘moon’ and not appear as a Name Entity for a person.

Another issue in NER is the ambiguity between two or more NEs. For example consider the following text: (عيد سعيد عيد مبارك). In this example, the (Eid) is both a person's name and a greeting for Al Eid, thereby giving rise to a conflict situation, where the same NE is tagged as two different NE types. The same in the following names {حصة ، موزة ، شمس ، هند، جمعة} for example "حصة مرحة" (Hesa is a funny girl) and "موزة مهضومة", which means "Mouza is cute", another example is the name "أحمد الفهد الصباح"; these are all person-names and do not refer to an animal or a timing period.

In Machine Translation (MT), NEs require different translation techniques than the rest of the words of a text. The post-editing step is also more expensive when the errors of an MT system are mainly in the translation of NEs. This situation inspired (Babych and Hartley 2003) to conduct a research study in which he tagged a text with an NER system as a pre-processing step of MT. He found achieved a higher accuracy with this new approach which helps the MT system to switch to a different translation technique when a Named Entity (NE) is detected (Othman 2009).

5 Word Sense Disambiguation Translation

The Arabic Language contains several kinds of ambiguity; many words can be in various characteristics based on certain contexts. For example, the word دين has two meaning; the first refers to religion and the second refers to rent money. Such ambiguity can be easily distinguished by a human using common sense, while machine translation cannot distinguish the difference. Instead, MT requires more complex analysis and computation in order to correctly identify the meaning; this process is called Word Sense Disambiguation (WSD) (Mussa 2014); (Hadni 2016).

Word Sense Disambiguation (WSD) is the problem of identifying the sense (meaning) of a word within a specific context. In Natural Language Processing (NLP), WSD is the task of automatically determining the meaning of a word by considering the associated context (Navigli 2009). It is a complicated but crucial task in many areas, such as Topic Detection and Indexing, Information Retrieval, Information Extraction, Machine Translation, Semantic Annotation, Cross-Document Co-Referencing and Web People Search. Given the current explosive growth of online information and content, an efficient and high-quality disambiguation method with high scalability is of vital importance to allow for a better understanding, and consequently, improved exploitation of processed linguistic material (Hadni 2016).

One example of an ambiguous Arabic word is خال (Khal), which can be translated to any of the following three words: "empty", "imagined" or "battalion." Due to the undiacritized and unvowelized Arabic writing system, the three meanings are conflated. Generally, 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 رأس 'head' one could think of the neck, nose, eyes, ears, tongue and so on (Abuelyaman et al. 2014).

Apparently, when many researchers translating Quran to English language, several semantic issues have been appear. Such issues poses the ambiguity of words for example ليلاًونهاراً (laylan wanaharan) and يوم الحساب (Yaum Alhesab), which are translated into “day and night” and “judgment day”. Such ambiguity has to be omitted by determining the correct sense of the translated word.

In MSA, synonyms are very common, for example the word year has two different synonyms in Arabic for example (سنة sanah, and عام Aam) and both of them are widely used in everyday communication. Despite the issues and complexity of Arabic morphology, this impedes the matching of the Arabic word. The word “year” is written also in two different ways in the Quran سنة sanat, and عام Aam. Both are simple singular forms occur 7 times in the entire Quran, providing one of many examples of word symmetries in the Quran. The words سنة (Sanat) and عام (Aam) are perfect synonyms. This cannot be further away from the miracle why God chose very specific words to be written in His book.

Ambiguity is not limited to Arabic words only, but also to Arabic letters when they affixed to morphemes, lead to ambiguous compound words. Table 2 shows how affixing the letter ‘ب’ which corresponds to ‘b’ in English, to an atomic word will turn it into a compound one. This is because, as a prefix, the letter ‘ب’ takes on any of the following senses: through, in, by, for and at. Table 2 shows only five of the ten possible roles the letter ‘ب’ plays when prefixed to different words (Abuelyaman et al. 2014).

The ambiguity of letters also appears in the Holy Quran. Twenty-ning surahs of Al-Quran begin with letters, such as Surat Maryam, verse 1 كَهَيَعَصَ “Kaf-Ha-Ya-’Ain-Sad”. These letters are one of the miracles of the Qur’an, and none but Allah alone knows their meanings.

Arabic texts without diacritics pose the greatest challenge for WSD, as they increase the number of a word’s possible senses and consequently make the disambiguation task much more difficult. For example, the word صوت Sawt without diacritics has 11 senses according to the Arabic WordNet (AWN) (Bouhriz and Benabbou 2016), while the use of diacritics for the same word صوّتَ Sawata cuts down the number of senses to two. Another example the word مال, which has seven senses in) (Bouhriz and Benabbou 2016):

- Sense 1{فلوس,ثَروة,دَراهم,مَال},
- Sense 2{نقود,مَال},
- Sense 3{تمايل,تَرنح,مَال},
- Sense 4 {انحدر,مَال},

Table 2 Letter ambiguity

Word	Translation	Word|| بـ	Translation of word|| بـ
بركة	Blessing	بـــبركة	Through blessing
المدرسة	The school	بالمدرسة	In the school
المال	The money	بـــالمال	By the money
أي	What	بـــأي	For what
الباب	The door	بـــالباب	At the door
القلم	The pen	بـــالقلم	Using the pen

- Sense 5 {مال، نزعَ إلى},
- Sense 6 {أقنع,أَمال,مَال},
- Sense 7 {انحرف,انحنى,مال}.

The WSD approach has shown that two words before and after an ambiguous word are sufficient for its disambiguation in almost all languages (Mohamed and Tiun 2015). For the Arabic language, the information extracted from this local context is not always sufficient. To solve this problem, an Arabic WSD system has been proposed that is not only based on the local context, but also on the global context extracted from the full text (Bouhriz and Benabbou 2016). The objective of their approach is to combine the local contextual information with the global one for a better disambiguation using the resource Arabic WordNet (AWN) to select word senses.

All of the WSD approaches make use of words in a sentence to mutually disambiguate each other (Chen et al. 2009; Agire et al. 2009; Ponzetto et al. 2010). The distinction between various approaches lies in the source and type of knowledge made by the lexical units in a sentence. Thus, all of these approaches can be classified into either corpus-based or knowledge-based methods. Corpus-based methods use machine-learning techniques to induce models of word usages from large collections of text examples. Statistical information that may be monolingual or bilingual, raw or sense-tagged is extracted from corpora. Knowledge-based methods instead use external knowledge resources that define explicit sense distinctions for assigning the correct sense of a word in context. (Dagan and Itai 1994); (Gale et al. 1992) used Machine-Readable Dictionaries (MRDs), thesauri, and computational lexicons, such as WordNet (WN). (Dagan and Itai 1994) was the first to resolve lexical ambiguities in one language using statistical data from the monolingual corpus of another language. That approach exploits the differences between the mappings of words to senses in different languages.

6 Conclusion

The paper presents the key the challenges of translating the Arabic language into the English language according to the classical Arabic, Modern Standard Arabic and Dialect Arabic. It also has suggested a line of argument in favors of the conceptualization of Word Sense Disambiguation, Metaphor, and Named Entity Recognition. Up to date, little work has been published on Arabic language translation. Arabic sentences are usually long, the punctuation are not effecting on the text interpretation. Contextual analysis is very important in the Arabic text translation, in order to understand the exact meaning of the word. The absence of diacritization in most of the MSD and completely in Dialect Arabic pose a real challenge in Arabic Natural Language Processing, especially in Machine translation. The Arabic language has many features that are inherently challenging for NLP researchers. The difficulties associated with recognizing the need for full-verbs the likes of “is”, and adverbs-of-places—the likes of “there”, recognizing the appropriate senses of undiacritized words, and the practice of performing translation at the compound word level are some of the main issues. Classical Arabic is regarded as rhetorical and eloquent because of its stylistic and linguistic manifestations. Translators who are not well-acquainted with this religious discourse cannot succeed in relaying the linguistic, stylistic and cultural aspects in the translated language. Unlike an ordinary text, the classical discourse is featured is noted to be sensitive; its language is euphemistic, indirect, and solicitous of people’s feelings. While Dialect can be a crucial element in the process of describing and individualizing characters in literature and therefore should be handled with great care. Dialect phonetic, grammatical and syntactic effect should directly or indirectly be preserved in the target language.

References

Abuelyaman, E., Rahmatallah, L., Mukhtar, W., Elagabani, M.: Machine translation of Arabic language: challenges and keys. In: 2014 5th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 111–116. IEEE Januray 2004

Abdul-Raof, H.: Arabic rhetoric: A pragmatic analysis. Routledge (2006)

Agire, E., Lacalle, O. L. d., Soroa A.: Knowledge-based WSD and specific domains: performing over supervised WSD. In: Proceedings of the International Joint Conference on Artificial Intelligence 2009, pp. 1501–1506, AAAI Press (2009)

Ahmad Abdel Tawwab Sharaf Eldin.: A Cognitive Metaphorical Analysis of Selected Verses in the Holy Quran. International J. Engl. Linguist **4**(6) (2014)

Ali, A., Brakhw, M.A., Nordin, M.Z.F.B., ShaikIsmail, S.F.: Some linguistic difficulties in translating the holy Quran from Arabic into English. Int. J. Social Sci. Humanity **2**(6), 588 (2012)

Alkhatib, M., Shaalan, K.: Natural language processing for Arabic metaphors: a conceptual approach. In: International Conference on Advanced Intelligent Systems and Informatics, pp. 170–181, Springer International Publishing October (2016)

Alshehab, M.: Two english translations of arabic metaphors in the Holy Qura’n. Browser Download This Paper (2015)

Babych, B., Hartley, A.: Improving machine translation quality with automatic named entity recognition. In: Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: resources and Tools for Building MT, Association for Computational Linguistics, pp. 1–8 April 2003

Benajiba, Y., Rosso, P., Benedíruiz, J.M.: Anersys: an Arabic named entity recognition system based on maximum entropy. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 143−153. Springer, Berlin, Heidelberg February 2007

Benajiba, Y., Rosso, P.: Arabic named entity recognition using conditional random fields. In: Proceedings. of Workshop on HLT & NLP within the Arabic World, LREC, vol. **8**, pp. 143–153 May 2008

Biadsy, F., Hirschberg, J., Habash, N.: March. Spoken Arabic dialect identification using phonotactic modeling. In Proceedings of the eacl 2009 workshop on computational approaches to semitic languages (pp. 53–61). Association for Computational Linguistics (2009)

Bouhriz, N., Benabbou, F.: Word sense disambiguation approach for Arabic text. Int. J. Adv. Compt. Sci. Appl. **1**(7), 381–385 (2016)

Chen, P., Ding, W., Bowes, C., Brown, D.: A fully unsupervised word sense disambiguation method using dependency knowledge. In: Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the ACL.pp. 28–36. ACL, Rappel Precision F1-mesure SVM 0,746 0,718 0,732 Naive Bayesien 0,747 0,71 0,782 8 (2009)

Dagan, I., Itai, A.: Word sense disambiguation using a second language monolingual corpus. Comput. linguist. **20**(4), 563–596 (1994)

Darwish, K., Magdy, W.: Arabic information retrieval. Foundations and Trends®. Inf. Retrieval **7**(4), 239–342 (2014)

El Kholy, A., Habash, N.: Techniques for Arabic morphological detokenization and orthographic denormalization. In: LREC 2010 Workshop on Language Resources and Human Language Technology for Semitic Languages, pp. 45−51 (2010)

Gale, W., Church, K., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Comput. Humanit. **26**, 415–439 (1992)

Hadni, M., Ouatik, S.E.A., Lachkar, A.: Word sense disambiguation for Arabic text categorization. Int. Arab. J. Inf. Technol. **13**(1A), 215–222 (2016)

Hashemi, A.: Javaher al-balagha. (H. Erfan, Trans.) Belaghat Publication (2002)

Lee, Y.S.: Morphological analysis for statistical machine translation. In: Proceedings of HLT-NAACL 2004: short Papers, pp. 57−60. Association for Computational Linguistics, May 2004

Mohamed, O.J., Tiun, S.: Word sense disambiguation based on yarowsky approach in english quranic information retrieval system. J. Theor. Appl. Inf. Technol. **82**(1), 163 (2015)

Mohaghegh, A., Dabaghi, A.: A comparative study of figurative language and metaphor in English, Arabic, and Persian with a focus on the role of context in translation of Qur'anic metaphors. TextRoad Publication, **3**(4), pp. 275–282 (2013)

Mussa, S.A.A., Tiun, S.: A novel method of semantic relatedness measurements for word sense disambiguation on english AlQuran. Bulletin Elect. Eng. Inform. vol. 4 (2014)

Mussa, S.A.A., Tiun, S.: Word sense disambiguation on english translation of holy quran. Bulletin Elect. Eng. Inform. **4**(3), 241–247 (2015)

Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. (CSUR) **41**(2), 10 (2009)

Newmark, P.: A textbook of translation (Vol. 66). New York: Prentice hall (1988)

Othman, R.: Trends In Information Retrieval System. IIUM Press (2009)

Ponzetto, S.P., Navigli, R.: Knowledge-rich word sense disambiguation rivaling supervised system. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 11−16, pp. 1522−1531 July 2010

Sadiq, S.: A Comparative Study of Four English Translations of Surat Ad-Dukhan on the Semantic Level. Cambridge Scholars Publishing (2010)

Sadat, F., Habash, N.: Combination of Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 1–8. Association for Computational Linguistics July 2006

Sawaf, H.: Arabic dialect handling in hybrid machine translation. In: Proceedings of the Conference of the Association for machine translation in the Americas (AMTA), Denver, CO November (2010)

Shaalan, K.: A survey of Arabic named entity recognition and classification. Computat. Linguist. **40**(2), 469–510 (2014)

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., Callison-Burch, C.: Machine translation of Arabic dialects. In: Proceedings of the 2012 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies, pp. 49−59. Association for Computational Linguistics June 2012

Zribi, I., Khemakhem, M.E., Belguith, L.H.: Morphological analysis of tunisian dialect. In: IJCNLP pp. 992−996 October 2013