Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Artificial intelligence and natural language processing: the Arabic corpora in online translation software

Mohammed Abdulmalik Ali *

Department of English, Prince Sattam Bin Abdulaziz University, AlKharj, Saudi Arabia

ARTICLE INFO

Article history:
Received 25 May 2016
Received in revised form
28 August 2016
Accepted 20 September 2016

Keywords:
Arabic corpora
Online content
Translation
Software

ABSTRACT

It is ironical to note that worldwide the Internet content in the Arabic language is mere 1%, whereas 5% of the world population speaks Arabic. This speaks of the disproportionate presence of on-line content of Arabic language as compared to other languages which may be due to many reasons including a lack of experts in the field of the Arabic language. This research study will investigate the impact of such Machine Translation (MT) software and TM tools that are widely used by the Arab community for their academic and business purposes. The study aims at finding whether it is possible to bring a paradigm shift from Arabic Localization to Arabic Globalization; hence, facilitating the usage of NLP techniques in the human interface with the computer. For this study; a few machine translation software (e.g. SYSTRAN, IBM Watson) shall be studied for their content and applications, to determine their usage without human intervention and retaining the meaning of the original text.

1. Introduction

Researchers have known Natural Language Processing (NLP) as that branch of Artificial Intelligence (AI) that deals with analyzing a language that is used by a human being to interface with a computer. A great challenge that man has faced in such an interface is to teach a computer the language that a man can learn, understand and interact in, which in the current context, is the Arabic language. Being the largest living Semitic language, official language of 23 countries, spoken by over 360 million people worldwide (The Arab world population is estimated to 369.8 million people (2013). The Arab region maps from Morocco in North Africa to Dubai in the Persian Gulf), Arabic language has ironically less than 1% of worldwide Internet content when 5% of the world population speaks Arabic. This speaks of the disproportionate presence of on-line content of Arabic language as compared to other languages. The reason given by NLP experts (Ali and Khaled, 2009; Habash, 2010; Hijjawi and Elsheikh, 2015**;** Huang, 2015) with regard to analyzing the use of the Arabic Diglossia is that Arabic has two forms existing concurrently. The first is the Modern Standard Arabic (MSA) which is widely used in formal situations like formal speeches, government and official operations, product manuals, and news media and it is perceived as the "language of the mind" in contrast with the second form known as Dialectal Arabic (DA). It is the informal private language, predominantly found as spoken vernaculars with no written standards, and perceived as "language of the heart"' although the Arab speakers perceive the use of dialects as a "deteriorated" form of Classical Arabic (Huang, 2015), a much debatable issue and outside the scope of this study.

This research paper will be citing a few studies that have principally addressed to these challenges and shall also investigate the impact of such Machine Translation (MT) software that are widely used by the Arab community for their academic and business purposes. This study will also cite examples of discrepancies found in these software and search engines. The main objective of this study is to find whether it is possible to bring a paradigm shift from Arabic Localization to Arabic Globalization in order to facilitate the use of NLP techniques or even formulate and modify the existing Arabic corpora for better understanding of the language. The article shall also discuss frequent colloquialism (e.g. Arab chat alphabet known as *Moaarab* or *Arabizi*) as found on social media platforms like Facebook and Twitter not withstanding spelling errors,

* Corresponding Author.
Email Address: mh.ali@psau.edu.sa (M. A. Ali)
https://doi.org/10.21833/ijaas.2016.09.010
2313-626X/

ungrammatical utterances, internet slangs, improper casing and like.

2. On-line content of Arabic language

The Arabic language has 5 major dialects: Iraqi, Gulf, Egyptian, Levantine and Maghrebi (Zaidan and Callison-Burch, 2011; 2014) with wide and distinct differences in their morphologies, grammatical cases, vocabularies and verb conjugations. These differences are reflected whenever Arabic utterances are transcribed or processed in any machine translation (MT) systems or automatic speech recognition (ASR) systems. The only solution, according to the authors, lies in "dialect-specific processing and modeling" and "identification and classification" of the Arabic text before making it a part of any NLP system. Moreover, there are several factors responsible for the disproportionate presence of on-line content of Arabic language as compared to other languages (Huang, 2015). However, the most significant is the penetration of the Dialectical Arabic (DA) threatening to replace the Modern Standard Arabic (MSA). Much of the Internet content and on the social media is composed in DA, which can be called the spoken variety of the Arabic language and markedly different from the MSA in terms of phonology, morphology, lexicon and even syntax (Embarki and Ennaji, 2011). Since Dialectical Arabic is traditionally confined to informal situations while writing requires the use of MSA, a need has been felt to switch the focus of the Arabic NLP on the informal communication. However, there are research studies (Zbib et al., 2012; Salloum and Habash, 2011) that advocate developing parallel corpora that would be mixtures of Dialectal Arabic and MSA e.g. Levantine-English and Egyptian-English corpora. According to these studies, a parallel corpus would act as a dialect classifier, enabling to accomplish a bias towards MSA or the Dialectical portion of the data inserted in the translation model (TM). The research carried out by Zbib et al. (2012), for instance, recommend the need of a cross-dialect data amalgamation to see the effects of translating from a particular dialect (e.g. Levantine) to MSA and then to English and vice versa. Zbib et al. (2012) however lament on the limited presence of MSA parallel data and therefore feel the need for an increase in MSA morphological segmentation through dialect-specific segmenters. The researcher in this paper will however be focusing only on the authenticity and reliability of translations and will be citing examples from leading MT software to prove his point.

3. Prior studies

There are two most recent research studies (Huang, 2015; Meftouh et al., 2015) that discuss the issues and challenges of machine translation of the Arabic language. In the first study, Huang suggests several solutions to handle challenges being faced in bringing the quality of machine translation. He suggests using the dialect classifier output to build a compatible, dialect-specific Arabic-English MT system. The dialect classifier identifies the type of dialect before translating and sending it to the corresponding MT system of English-to-Arabic (MSA) translation system. In this method, the target Language Model (LM) derived from the SL helps to improve the translation quality. Truly speaking, the researcher feels that since the in-domain data of the Arabic language contains lots of dialects, an issue taken up later in this study, it will be a good idea to maintain a cleaner LM with the assistance of an effective dialect classifier to filter out all DA elements and only retain the MSA elements. However, how effective it will prove when it comes to practical application needs to be tested holistically. In another study, Meftouh et al. (2015) devises a Parallel Arabic DIalect Corpus (PADIC) that focuses on the statistical MT experiments from MSA to DA and vice versa and conducts experiments on cross dialect Arabic machine translation. Meftouh et al. (2015) claimed that PADIC comprises major dialects of the Arab region and a special attention has been paid to align each dialect with MSA. This study uses at least eight dialects (two dialects from Algeria, three from Maghreb, two from the Middle-East (Syria and Palestine) and one from Tunisia). Meftouh et al. (2015) also felt a few constraints in devising this corpus namely a lack of dialectal Arabic parallel corpora, colloquial nature of Arabic dialects as most of them are used only for conversations and not for writing, and above all, the small size of the available corpora. However, for this purpose of analyzing the effect of LM on MT process, they vary the smoothing techniques and interpolate it with PADIC, which they claim to be the largest corpus working on dialects.

Zbib et al. (2012), too, talk about Levantine-English and Egyptian-English parallel corpora. In their study, the authors perform a few machine translation experiments in order to show a variety of attributes like a limited MSA data, the utility of inter-dialect learning, the impact of both morphological analysis and translating from a particular dialect to MSA and then to English. Zbib et al. (2012) claimed to have discovered a process of developing a Dialectal-Arabic-English parallel corpus, for which they need to select passages containing non-MSA words from any available corpus on the Arabic web text, and use crowdsourcing method for classifying the text dialect wise and segmenting them into individual sentences before translating into English. This study seems to be a canonical one because several other works in this domain have proposed a similar experimentation of processing Arabic dialects and recommending the part-of-speech (POS) tagging, diacritization, building of lexicon and analyzing text morphology (Zbib et al., 2012).

There are also a few studies on Arabic Dialect (AD) hybridization and adaptation. For instance, a few authors (Salloum and Habash, 2011; Zbib et al., 2012; Darwish et al., 2014), in order to devise a translation system from MSA to Moroccan, designed such tools that could be adapted to the Moroccan

dialect; Sawaf (2010) devises a MSA pivot approach to build a hybrid AD-English MT system in which AD is transferred into MSA. Similarly, Salloum and Habash (2011) recommend Elissa, a system that uses a rule based approach in translations from AD to MSA and also depends upon paraphrasing the dialectal sentences before transferring to MSA. Elissa has been designed to meet the requirements of all major dialects of the region, particularly Iraqi, Levantine, Egyptian, and to a little extent the Gulf Arabic. Farghaly (2010), too, discuss the use of state of the art technology in SMT and devise a sort of Phrase-based Statistical Machine Translation (PBSMT) to be used in longer translation units than the initial word-based models. By these means, according to Zbib et al. (2012), more contextual information can be captured and so lead to improving the translation quality as well. Their device also uses the log linear model which allows the integration of additional features into the model with different weights. The weights are optimized using optimization algorithms.

Despite all these experimentations, the transfer approach to machine translation has also gained prominence. Farghaly (2010), for instance, took a few examples from the SYSTRAN Arabic to English machine translation system. Alqudsi et al. (2012) too recommend to develop a transfer based MT system and a no-machine learning technique that fully meets human requirements. It is emphasized upon transfer-based technique to ensure that the meaning of the original sentence is captured before generating the correct translation. In order to meet the challenges faced by any set of two languages, the authors emphasize upon building a strong lexical MT system that will not only accept source sentences (SL) of Arabic but also will generate sentences in English as a target language (TL).

Moreover, the transfer approach is often contrasted with the Interlingua approach to machine translation (Shaalan et al., 2004). The Interlingua approach is based on the assumption that it is possible to convert the source language texts into a universal representation that is language independent. This universal representation can, in turn, be converted into the surface representation of the target language. The interlingua approach actually suits multilingual environment as it was in the case of European Commission funded Eurotra machine translation project (1974-1994) which mandated that all citizens of the European Union have the right to access and to read all the documents of the commission in their own official languages. With more countries joining the European Union (EU), this resulted in a combinatorial explosion in the number of language pairs involved and very quickly translation placed a heavy burden on the administrative budget of the EU. Moreover, the Interlingua has the advantage of making the addition of a new language to the MT system less costly and much faster. Thus the transfer approach to machine translation is particularly helpful for only a specific pair of languages.

All these studies are although dynamical but yet unclear whether machine translation can be relied upon for all the requirements in terms of retrieval time and translation quality. Thus most of these studies are inconclusive as they only talk about constraints and challenges, and solutions suggested and instruments devised too have very limited application and use.

4. Multidialectal Arabic parallel corpora

There is no such concept like Multidialectal Arabic parallel corpora, however, Bouamor et al. (2014) pioneered such a corpus recently which comprises over 2000 sentences in multiples dialects like Egyptian, Jordanian, Palestinian, Tunisian, Syrian and even MSA and English. These sentences were based upon the corpus built by Zbib et al. (2012). However, there are various software and online dictionaries that are serving the translation needs of the people across 23 countries where Arabic is the national language and several other nations where there are Arab migrants residing for education and business purposes. For instance, MTs like Al-Mounged English-Arabic-English dictionary, IBM Watson Language Translation and Amazon's Mechanical Turk (MTurk) are becoming essential tools for creating annotated resources for computational linguistics. However, all these tools have limitations that result in a comparison of two translators, Google and IBM Watson, suggestive of a need of a more complete and stronger corpus to meet the translation requirements. It is evident from the examples (Table 1) why corpus-based approach is incapable of helping translation teachers as well as students to acquire a correct translation. Besides the problems discovered in these examples, there are several other problems like synonymy issues in translation in the target language (propositional meanings vs. expressive meanings); the choice of the appropriate equivalent in the target language; semantic prosody and translation as well as differences in collocation and collocation patterns between the source and target language.

The SYSTRAN Arabic to English Transfer Machine Translation System SYSTRAN Inc. is also a pioneer in machine translation for over thirty years focusing on developing machine translation systems in more than thirty languages using the transfer approach, (http://www.systransoft.com/lp/english-arabic-translation/). Its approach is however different from other MT software. It understands that MT systems are usually designed for specific language pairs; therefore, they can capitalize on the similarities between the source and target languages. SYSTRAN therefore makes extensive use of dictionaries that annotate lexical items with morphological, syntactic and semantic features.

SYSTRAN has developed a monolingual Arabic stem-based lexicon and a bilingual Arabic to English dictionary. In order to differentiate and expedite the process, SYSTRAN has used Arabic stems rather than roots, thus eliminating the step of generating stems

from roots. The goal of the SYSTRAN Arabic to English MT system is to improve translation quality by introducing analysis, transfer and disambiguation rules in at least three existing Arabic MT systems, like ALKAFI, GOOGLE, and TARJIM SAKHR.

Table 1: A Comparative analysis of Google and IBM Watson MT software

	Original Sentences	Google Translation	IBM Watson Language Translation	Author's Comments
1	We teach at a school.	*Naħnu tadrīs fi al-madrasah.*	*ʔallamūnā fi al-madrasah.*	Both systems mistranslate the sentence: instead of saying (nuʔallem / nadrus) the verb *teach* is translated wrongly in both cases; definite article in Arabic (*al*) is added to the noun school (madrasah)
2	Salma speaks English.	*Al.inglīziyyah yataħaddaθ Salmā.*	*Al.inglīziyyah yataħaddaθ Salmā.*	Both translators could not recognize the gender of the subject. The subject is feminine, but dealt with as a masculine.
3	Switch off lights before leaving the room	*Al-ġurfahatu maġādaratu qubla al-anwar itfāa.*	*Al-ġurfahatu maġādaratu qubla al-anwar iġlaq muftāħ.*	Google is very much closer to correct translation, but couldn't recognize the imperative structure of the sentence. IBM gives a wrong translation.
4	Please, don't park in front of the loading bay	*Fatħa amam ħadīqat la faḍlak, min al-taħmīl.*	*Al-ŝaħn xalīj amam bark la arjūk.*	Neither translation system presents a correct translation of the sentence. Meaning has been lost completely. Only one meaning of the word *park* could be recognized.
5	Please, make the office look clean and tidy. Books on the shelf, Computer stuff on the desk, and pencils in the desk drawers.	*naẓīfa tabdu al-maktab jaʿala faḍlik,min al-ʿŝyaʔ al-rraf al-kutub wamurattabah. Al-maktab, waaqlām al-raṣaṣ fī adrāj maktab.*	*maktab watartīb bitanẓīf al-qiyām rajāʔ al-ʿŝyaʔ ʿala al-ħāsib al-rraf al-baħθ. Kutub al-adrāj maktab waaqlām maktab.*	Google seems closer to correct translation but with a few mistakes: - Gender problem – *Office* is masculine but not feminine in Arabic - Compound noun *computer stuff* is translated as two separate words with unmatched meanings - Deletion of the definite article in *the desk drawers* makes the translation awkward.

Google and Microsoft, too, have taken the initiative and included Arabic in their priority languages. Microsoft has announced *Maren*, a Windows extended and customized version, which can translate Arabic written in Roman characters into Arabic script. It has gained huge popularity since its release. Thus, Arabic just became the eighth language supported by Microsoft's real-time translation tool. Similarly, Skype was recently (2016) upgraded to support Arabic by introducing a new version of Skype Translator. As with the other Skype Translator languages, the translation in Arabic will happen in real time as the other person is speaking. That's because for each word of a sentence that is spoken, the translation software becomes more confident in understanding the meaning of a sentence and is able to make adjustments in real time. Arabic thus joins English, Spanish, French, German, Mandarin, Italian and Portuguese (Brazilian) in Skype's growing lineup of languages. However, the Skype translator too has limitations; it appears to handle basic conversations well, but sometimes stumbles with more complex sentences or misheard words. (For instance, it mistakes the word "bye" for "vine," while in another conversation it appears to substitute "Crown Princess" for "England.") Nevertheless, Google claims that their mission is to provide authentic tools to all Arabic users in order to enrich Arabic content and offer more opportunities of e-commerce in the region.

5. The linguistic theory

There was a paradigm shift in linguistics when Chomsky (1957) challenged the well-established theory of structural linguistics and redefined the goals of linguistic theory to account for native speakers intuitions about their language rather than simply investigating a corpus and finding regularities in that corpus. He also challenged the view held by structuralists that a child is born with a "tabula rasa" i.e. with no knowledge of language at all. Structural linguists believe that it is through listening, imitating, and repetition that a child acquires the language of his people. Chomsky proves that while comparing any two languages a child internalizes with the fragments he is exposed to in his early linguistic experiences. He also points out the gaps that need to be accounted for in language corpus and challenged the structuralists' position that in order to write a description of a language, they must obtain a corpus of the language and perform a "discovery procedure" to deduce the generalizations underlying the language. He argues that a corpus of native speakers' utterances represents only the performance of the speakers of the language. Performance is usually affected by lapses of mind, change of plans, fatigue, and distractions etc. It is not always a true reflection of native speakers' knowledge of their language. Chomsky argues that a linguist should aim at

describing the speaker's mental grammar by eliciting his intuitions. He makes a fundamental distinction between competence and performance. For him, competence is the linguistic knowledge a speaker has of his language while performance is what he actually says which a true reflection of his linguistic knowledge. Later, Chomsky (1965) himself states clearly that linguistic theory must be concerned with characterizing native speakers' competence rather than performance. The theorists of NLP and the Arabic Corpora need to pay attention to Chomsky's advice while resolving the translation issues.

6. Decline of Arabic language

Experts have expressed concern over a gradual decline of the Arabic language in all regions of the world due to many factors like globalization, use of the English language in the Arab world, especially on social media, use of the Arab chat alphabet known as Moaarab or Arabizi (created with Roman characters and English numbers mainly to communicate over the internet and cellular phones) in speech and text, foreign, non-Arabic speaking workers out-numbering native Arabic speakers as in the Emirates and like. There are also an inadequate number of Arabic language teachers who can be trusted to preserve this language; as a result, Classical Arabic is being replaced by local dialects. Not only that, but some people like the Egyptian Philosopher Safouan (2007) argue that classical Arabic is a dead language like Latin and Greek although others see it as a tool for unifying the Arab world.

Truly speaking, the growth of universal media and globalization has challenged many native languages including the Classical Arabic. Several voices are aired to draw attention of the impending decline of the Classic Arabic. Globalization Partners International (2015) finds in their study that since the educational system delivers most of the curriculum in a foreign language, the classical Arabic has failed to modernize and therefore is at risk; On the contrary, in interviews with major news networks of the region experts discuss the importance of the Arabic language as true of any Arab's identity. Speaking to *Saudi Gazette*, (Fatma, 2016) Abdelsalam Al-Masaddi, a famous Tunisian professor of linguistics opines that Arabic should not be a dying language nor the Arab world should be represented as "fragmented". Duha Akkad, who teaches at King Abdulaziz University, also comments, that the principal cause of the weakening of the Arabic language is the penetration of colonial powers into the Arab world. Athoob Al-Shuaibi, in an interview given to *Kuwait Times* (Fatma, 2016) also raises great concern on the influence of English on the Arabic vocabulary and criticizes people who have abandoned Arabic, their mother tongue, and adopted English.

However, this abandonment of one's mother language in favour of English is a global phenomenon and not limited to Arabic and may be termed as Anglicization of global culture (Hjarvard, 2004). The media has further contributed to a sort of Anglo-American culture and institutionalization of English due to its excessive usage in engineering, medicine and software industry. Thus human communication has become mediatized, as media-bound varieties of language have arisen (Hjarvard, 2004).

7. A paradigm shift

Fatma (2016) however vehemently turns down the notion that Arabic is an "endangered" language or this language is declining. Instead, she finds a paradigm shift happening from Arabic Localization to Arabic Globalization, as many studies have been carried out to apply Arabic NLP to building up a scientific corpus. She however accepts the difficulty in handling Arabic (and Hebrew) in NLP as they are heavily inflected languages. Therefore, before being able to parse an Arabic text or manipulate it to build scientific corpora, she claims that an expert linguist needs to make morphological analyses to get to the word lemmas. Having word lemmas in Arabic gives a similar ground level text to what one already has in English. This would bring a new advancement in NLP including annotated corpora in Arabic, she adds. Fatma's contention can be supported by the introduction of a new invention, SyntaxNet, which is based on an open-source neural network and is embedded in the TensorFlow software constructed on the principles of Natural Language Understanding (NLU) systems. The new invention enables procurement of all the codes needed to analyze the English text while translation. The software uses the Parsey McParseface, a small parsing unit, which is the first step in writing a parser for Arabic since there is a lack of reliable text corpora in Arabic. Parsey McParseface employs machine learning algorithms that have been developed to analyze the linguistic structure of a language, including its lexicon, syntax and morphology. It is however yet to prove whether Parsey McParseface can automatically extract information and is adapted to translation systems and other core applications of NLU.

One of the main problems that make parsing so challenging for a MT or computer software is the ambiguity that every human language shows in terms of length of sentences or their syntactic structures which may vary tremendously. The big challenge before a natural language parser is to resolve this ambiguity and search the right structure from all available alternatives in a given context. For studying the effectiveness of parsing done in this MT software, the Google translations of the following 4 examples were tested in online parsing program, (Table 2) *ZZCad Sentence Parsing Program* (zzcad.com/cgi-bin/webparse.exe) and later parsing done. The parsing found in the online software exhibits various possibilities of dependency parsing (Table 3).

Table 3 clearly shows that when the TM (Google Translate, here) does correct or logical translation as in Example 1, a remarkable reason for this accuracy

of translation can be attributed to system being able to recognize the syntactic relationships between the words and the meaning that each word has according to its structure (morphology). Hence, having the TM system trained (programed) well to recognize the structure of the Arabic sentence does not guarantee reasonable translation. This is shown in Example 4 which tells that parsing of the English sentence and its Arabic translation is the same, but the meaning cannot look accurate without human's interaction. Although the sentence has been parsed correctly and the syntactical relations are kept the same, even the nominal form of the sentence is kept in both cases; some words are wrongly translated (ħalaqa = circle instead of ħabah = ring). Also the translator dealt with Alice as a masculine so the verbs "reading" and "saw" are presented in the masculine form although they should be put in the feminine form (yarā & yaqra? instead of ra?at & taqra?, respectively).

With reference to example 3, it is noticed that Google Translate recognizes the correct parsing of the Arabic equivalent sentence except for the last part of the sentence of the prepositional phrase; the position of the adjective in this phrase is kept the same and so it makes a difference in the meaning. The meaning in the output sentence (Arabic sentence) is different from that conveyed in the original English one. The English sentence means that we teach the different modules in a number of colleges, whereas the prepositional phrase (in different colleges) is mistranslated into Arabic showing that the modules we teach are in all the colleges available in that context (fī muxtalaf al-kulliyyat) instead of (fī kulliyyat muxtalafah).

In the same sequence, the meaning implied in the original English sentence is conveyed in the Arabic translation, but the sentence looks a bit awkward as the output is a nominal sentence which not preferable in this case, not to forget the choice of the word (al-ifrāj) instead of (ṭarħ) which is not appropriate for (release).

An important point noticed in all the above examples is that the Translation of English sentences into Arabic results in Nominal sentences although occurrence of Arabic verbal sentences in standard and even colloquial Arabic is very much greater and more common. These are no exceptions nor can SyntaxNet claim to resolve these issues since it applies neural networks to the ambiguity problem. In all these examples of translations, the software processes an input sentence from left to right, while it cannot stop the dependencies between words which increase as parsing progresses. Moreover, in order to achieve highest prediction accuracy (Hijjawi and Elsheikh (2015), it is essential to integrate all searches before making any left-to-right sequence of decisions during parsing. Any software, including the Parsey McParseface, cannot achieve parse accuracy and therefore it may not be useful to implement in many applications.

At this stage, one may like to try to understand why Arabic processing is hard. Look at the problem areas below (Table 4).

Cohen (2015) opines that English has relatively impoverished morphology because of its simplicity. Languages like Turkish, Arabic, Hungarian, Korean, and many more have rich and complex morphology in comparison to English which have relatively simple morphology. For example, Turkish is an agglutinative language, and words are constructed by concatenating morphemes together without changing them much, according to Cohen.

English on the other hand has concatenative morphology in which words can be made up of a main stem (carrying the basic dictionary meaning) plus one or more affixes carrying grammatical information. E.g.: Surface form: cats walking smoothest Lexical form: cat+N+PL walk+V+PresPart smooth+Adj+Sup. In effect, morphological parsing is the problem of extracting the lexical form from the surface form as speech processing, too, like if we see irregular verb forms (e.g. tooth → teeth) systematic rules (e.g. 'e' inserted before suffix 's' after s,x,z,ch,sh: brush → brushes, box→ boxes) and so on (Alhihi, 2015).

8. Limitations and future prospects

This paper illustrates some of the issues and challenges before the MT system that even the AI has not been able to capture despite decades of research and expertise behind the studies.

However, there are studies (Ramos et al., 2008) that draw attention to the concept of Ambient Intelligence (AmI) developed by the European Commission's Information Society Technologies Advisory Group or Istag.

Table 2: Examples tested on online parsing program

	English Sentence	*Arabic Version*
1	Ninety thousand people left Fallujah and Deyala in the last month.	*Tisʿūn alf ŝaxṣ ġādarū al-fallūjah wadiyāla fī al-ŝahr al-māḍi.*
2	Alice, who had been reading about boxing, saw Bob in the ring yesterday.	*Alis allði kān yaqrau ʿan al-mulākamah, raā Bob fī ħalaqat ams.*
3	Google today is announcing the release of version 5.0 of the Google Translate service.	*Gūgil al-yawm tuʕlin al-ifrāj ʕan al-iṣdār 5.0 min xidmat al-tarjamah min Gūgil.*
4	We teach different modules at different colleges.	*Naħnu naʕllimu wiħdāt muxtalifah fī muxtalaf al-kulliyyat.*

Table 3: Examples of showing that inefficiency in MT happens because of incorrect parsing

Example 1a: Original English Sentence (left to right)										
NS Nominal										
(NP) Nsbj (Subject)			VT	Nobj (Object)			PP (prepositional phrase)			
N	N	Nsbj	Root	N	Conj	N	P	Art	Adj	N
Ninety	thousand	People	left	Fallujah	and	Deyala	in	the	last	Month

Example 1b: Google Arabic Translation												
NS Nominal												
(NP) Nsbj			VP		Nobj				PP			
N	N	N	VT Root	Pron	Art	N	Conj	N	P	NP		
										Art	N	Art	Adj
Tisʕūn	*alf*	*ŝaxṣ*	*ġādar*	*ū*	*al*	*fallūjah*	*wa*	*diyāla*	*fī*	*al*	*ŝahr*	*al*	*māḍi*

Example 2a: Original English Sentence (left to right)														
NS Nominal														
(NP) Nsbj									PP					
		Root		Nobj		PP			PP					
N	N	V	V	Art	N	N	V	No	P	Art	N	N	N	
Google	today	is	announcing	the	release	of	Version	5	of	the	Google	Translate	Service.	

Example 2b: Google Arabic Translation (completely equivalent)															
NS Nominal															
Nsbj (NP)			VP												
N	NP		VT	NP		PP				PP				PP	
	Art	N	V	Art	N	P	Art	N	N (No)	P	N	Art	N	P	N
Gūgl	*al*	*yawm*	*tuʕlin*	*al*	*ifrāj*	*ʕan*	*al*	*iṣdār*	*5.0*	*min*	*xidmat*	*al*	*tarjamah*	*min*	*gūgl.*

Example 3a: Original English Sentence (left to right)						
NS Nominal						
(NP) Nsbj	VP			p	PP	
	Root	Nobj			PP	
Pron		Adj	N		Adj	N
We	teach	different	Modules	at	different	colleges

Example 3b: Google Arabic Translation (completely equivalent)							
NS Nominal							
NP	VP			PP (prepositional phrase)			
Nsbj (Pron)	V	Nobj	Adj	Prep	Adj	Art	Pron
Naħnu	*naʕllimu*	*wiħdāt*	*muxtalifah*	*fī*	*muxtalaf*	*al*	*kulliyyat.*

Example 4a: Original English Sentence (left to right)													
NS Nominal													
Nsbj	Independent Clause							VP		PP (prepositional phrase			
N	N	V	V	V	P	N		V	Nobj	P	Art		
Alice,	who	had	been	reading	about	boxing,		saw	Bob	in	the	ring	Yesterday.

Example 4b: Google Arabic Translation (completely equivalent)											
NS Nominal											
NP							VP				
Nsbj	Independent Clause						V (Root)	N	PP		
N	Pron	V	V	P	Art	N	V	Nobj	P	N	N
Alis	*allði*	*kān*	*yaqrau*	*ʕan*	*al*	*mulākamah,*	*raā*	*Bob*	*fi*	*ħalaqati*	*ams.*

Table 4: Problem areas in parsing
(Adapted for this study from Cohen (2015))

Problem Areas (Parsing /Morphology)	Arabic	English
Orthographic ambiguity	More	Less
Orthographic inconsistency	More	Less
Morphological inflections	More	Less
Morpho-syntactic complexity	More	Less
Word order freedom	More	Less
Dialectal variation	More	Less

AmI has been tested to produce results in a digital environment and support concepts like "ubiquitous computing, pervasive computing, context awareness, and embedded systems" that are potential solutions and future of MT rests on these concepts. These concepts work proactively in with smart equipment and micro-electromechanical devices and embedded systems, and most importantly I/O device technology using adaptive software. In the view of the current researcher, and looking ahead it can be maintained that AmI with its intelligence components is capable of both media management and handling computational intelligence (e.g. context awareness, and emotional computing), the two areas where the current MT software are lacking. In some future research, AmI perhaps may find the solutions to the issues raised in this paper.

References

Alhihi N (2015). Lexical problems in English to Arabic translation: A critical analysis of health documents in Australia. Arab World English Journal (AWEJ), 6(2): 316-328.

Ali F and Khaled S (2009). Arabic natural language processing. Challenges and Solutions, 8(4): 1-22.

Alqudsi A, Nazlia O and Khalid S (2012). Arabic machine translation. A Survey Artificial Intelligence Review, 8(3): 549-572.

Bouamor H, Nizar H and Kemal O (2014). A multidialectal parallel corpus of Arabic. 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland: 1240-1245.

Chomsky, Noam (1957). Syntactic Structures. Mouton, The Hague.

Chomsky, Noam (1965). Aspects of the Theory of Syntax. MIT Press, MIT Massachusetts, USA.

Cohen S (2015). Morphology parsing Informatics 2A: Lecture 14. School of Informatics, University of Edinburgh. Available online at: http://www.inf.ed.ac.uk/teaching/courses/inf2a/slides/2015_inf2a_L14_slides.pdf

Darwish K, Hassan S and Hamdy M (2014). Verifiably effective Arabic dialect identification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics: 1465–1468.

Embarki M and Ennaji M (2011). Eds. Modern Trends in Arabic Dialectology. The Red Sea Press, New Jersey, USA.

Farghaly A (2010). Arabic Machine translation: A Developmental Perspective. International Journal of Information and Communication Technology, 3(3): 3-10.

Fatma S (2016). Arabic in danger: Efforts to ensure proper transmission of Arabic continue. Available online at: https://arabizi.wordpress.com/tag/arabic-is-not-dead/

Habash NY (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1): 1-187.

Hijjawi M and Yousef E (2015). Arabic language challenges in text based conversational agents compared to the English language. International Journal of Computer Science and Information Technology (IJCSIT), 7(5): 1-13.

Hjarvard S (2004). The globalization of language how the media contribute to the spread of English and the emergence of medialects. Norcicom Review, 25(1/2): 75-07.

Huang F (2015). Improved Arabic dialect classification with social media data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: 2118-2126.

Meftouh K, Harrat S, Jamoussi S, Abbas M and Smaili K (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In The 29th Pacific Asia Conference on Language, Information and Computation: 26–34.

Safouan M (2007). Why Are the Arabs Not Free? The Politics of Writing. 1st Edition, Blackwell Publishing, Malden, Massachusetts, USA.

Salloum W and Habash N (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Association for Computational Linguistics: 10-21.

Sawaf H (2010). Arabic dialect handling in hybrid machine translation. In Proceedings of the Conference of the Association for Machine Translation in the Americas (amta), Denver, Colorado.

Shaalan K, Rafea A, Moneim AA and Baraka H (2004). Machine translation of English noun phrases into Arabic. International Journal of Computer Processing of Oriental Languages, 17(02): 121-134.

Zaidan OF and Callison-Burch C (2011). The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume. Association for Computational Linguistics, 2: 37-41.

Zaidan OF and Callison-Burch C (2014). Arabic dialect identification. Computational Linguistics, 40(1): 171-202.

Zbib R, Malchiodi E, Devlin J, Stallard D, Matsoukas S, Schwartz R and Callison-Burch C (2012). 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. Association for Computational Linguistics: 49-59.