Articles   Computational Functional Linguistics

2008-5

Towards Arabic to English Machine Translation

Yasser Salem
Technological University Dublin, yasser.salem@ucd.ie

Arnold Hensman
Technological University Dublin, arnold.hensman@tudublin.ie

Brian Nolan
Technological University Dublin, brian.nolan@tudublin.ie

Recommended Citation

2008-05-01

Towards Arabic to English Machine Translation

Yasser Salem
Institute of Technology, Blanchardstown, yasser.salem@ucd.ie

Arnold Hensman
Institute of Technology, Blanchardstown, arnold.hensman@itb.ie

Brian Nolan
Institute of Technology, Blanchardstown, brian.nolan@itb.ie

Recommended Citation

Towards Arabic to English Machine Translation

Yasser Salem, Arnold Hensman and Brian Nolan
School of Informatics and Engineering
Institute of Technology Blanchardstown, Dublin, Ireland
E-mails: {firstname.surname}@itb.ie

Abstract

This paper explores how the characteristics of the Arabic language will effect the development of a Machine Translation (MT) tool from Arabic to English. Several distinguishing features of Arabic pertinent to MT will be explored in detail with reference to some potential difficulties that they might present. The paper will conclude with a proposed model incorporating the Role and Reference Grammar (RRG) technique to achieve this end.

1 Introduction

Arabic is a Semitic language originating in the area presently known as the Arabian Peninsula. It has been spoken in its current form since the 2nd millennium BCE. As a language, Arabic has few irregularities and it is rich in morphological structure. Arabic is also rare in that it is a derivational language rather than concatenative. Words like 'went, go' – ذهب ؛ يذهب can easily be seen as being part of a hierarchy of inheritance from a specific root (in this case ذهب) In English and in many other languages this is not always the case.

The Arabic language is written from right to left. It has 28 letters, many language specific grammar rules and it is a free word order language. Each Arabic letter represents a specific sound so the spelling of words can easily be done phonetically. There is no use of silent letters as in English. Similarly, there is no need to combine letters in Arabic to achieve a certain sound that might be familiar to an English speaker. For example, the 'th' sound in English as in the word 'Thinking' is reduced in Arabic to the character ث .

In addition to the standard challenges involved in developing an efficient translation tool from Arabic to English, the free word order nature of Arabic creates an obstacle unique to the language. The number of possible clause combinations in basic phrasal structures far exceeds that of most languages. There is no copula verb 'to be' in Arabic, resulting in a unique usage of the subject 'I'. The absence of the indefinite article, while not unique to Arabic still poses many difficulties within the context of the language structure. These and other issues are discussed in later sections.

The remainder of this paper is organized in the following manner: Section 2 introduces some common features of Machine Translation and discuses generic problems regardless of language. Section 3 presents the characteristics of the Arabic language. ection 4 will discuss some distinguishing features of Arabic and finally Section 5 will summarize the findings discussed and briefly outline a proposed MT solution.

2 Machine Translation

Machine translation of natural languages, commonly known as MT is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. While semi-automated tools have been applauded in the recent past as the most realistic path to follow, it is no longer the case. The current consensus is that fully automated, efficient translation tools should remain the primary goal. The nature of users of such systems and the type of text involved leave little room for continued dependence on human aids.

The motivation for an Arabic-English translation tool is obvious when one considers that Arabic is the lingua franca of the Middle-Eastern world. Presently, 21 countries with a combined population of 450 million consider standard Arabic as their national language. A simple test case during a study at Abu Dhabi University over three popular Arabic translation tools (Google, Sakhr's Tarjim and Systran) revealed little success in generating the correct meaning [Izwaini, 2006].

2.1 General MT Obstacles

For the purposes of this study, any proposed solution to and Arabic-English translator will be based upon the Interlingua model. A transfer model that directly maps from source language to target language will remove the benefit of similarities between an Arabic translator and others. Arabic is unique in many ways but is not immune to the standard challenges faced in prior developments of MT tools for other languages such as multiple meanings of words, non-verbalisation and insufficient lexicons.
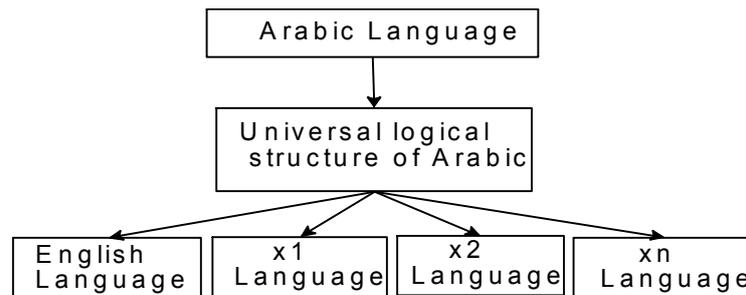
Figure 1: Interlingua model of Arabic MT

An Interlingua model that incorporates source language analysis, thereby creating a so called universal logical structure (in this case Arabic), will facilitate multiple language generation in a more flexible way. An Interlingua model is presented in Figure 1. For the elements of *subject* (S), *verb* (V) and *object* (O), Arabic's free word order allows the combinations of SVO, VSO, VOS, and OVS. The only combinations that do not occur in Arabic are OSV and SOV. Free word order is discussed later in this paper. Our research project investigates into developing a rule-based and lexicon framework for Arabic language processing using the Role and Reference Grammar (RRG) linguistic model. The framework is to be evaluated using a machine translation system that translates an Arabic text as source language into an English text as target language.

3   Characteristic of the Arabic language

The copula verbs 'to be' and 'to have' do not exist in Arabic. Instead of saying 'My name is Zaid', the Arabic equivalent would read like 'Name mine Zaid' - 'esmy zyd*O'[1]. Instead of saying 'She is a student', the Arabic equivalent would be 'She student'. In Arabic 'hEyA *ta:leba:.t*O'. Regarding the verb to have, which in English can also mean 'to own'. Instead of saying 'He has a house', the Arabic equivalent is 'To him a house' - 'la:hO byt*O'. Adjectives in Arabic have both a masculine and a feminine form. The singular feminine adjective is just like the masculine adjective but with the 'ta marbuta' to the end [Ryding, 2005].

Arabic does not ignore the specific case of two nouns, whereas other languages move from the singular to the plural form directly. In Arabic we need only add two letters to the singular form to express the dual form. An example is given in Table 1. The full plural form is obtained using a different mechanism.

Table 1: Dual: Merely add two letters to achieve dual form in Arabic

Arabic	Gloss	English Translation
باب	ba:b	door
بابان	ba:ba:n	two doors

3.1   Characteristics of Arabic words

There are no upper and lower case distinctions. Words are written horizontally from right to left. Most letters change form depending on whether they appear at the beginning, middle or end of a word or on their own. Arabic letters that may be joined are always joined in both hand-written and printed form. An interesting feature of Arabic is its treatment the demonstrative. Where in English one refers to an object that is either near or far away as simply *this* (very near the speaker) or *that* (away from the speaker up to any distance), Arabic has a third demonstrative to specify objects that are in between these points on the distance spectrum. In Arabic, nouns could be feminine or masculine. A feminine noun is formed by adding a special character 'ta marbuta' to the end of the masculine form. The recall feminization of a noun in Arabic graphically appears in the suffix and reflects the gender.

An example is given in Table 2.

---

[1] Arabic examples are written here by using AACP (Arabic Alphabet and the Corresponding Phonetic).

Table 2: Feminine and masculine in Arabic

Arabic	Gloss	English Translation
مُعلِّمٌ	mO3lEm*O	teacher(m)
مُعلِّمة	mO3lEm.t*O	teacher(f)
طالب	*ta:lEb	student(m)
طالبة	ta:lEbA.tO	student(f)

Table 3: feminine is different than masculine

Arabic	Gloss	English Translation
دَجَاجَةٌ	dAjAa:jA.t*O	Chicken
ديك	dyk	Cock

Table 4: Definiteness in Arabic

Arabic	Gloss	English Translation
ال	a:l	the

Table 5: Definiteness example in Arabic

Arabic	Gloss	English Translation
رجل	rAjOl	a man
الرجل	alrAjOl	the man

Sometimes the noun used to refer to the feminine object is different to that of the masculine as indicated in Table 3. The Arabic definite article joins with the word that it precedes. The shape of the definite article is shown in Table 4. The definite noun in Arabic graphically appears in the prefix of an Arabic noun. An example of Arabic definiteness is shown in Table 5.

3.2 Influence of Arabic on other languages

The number of different Arabic words in languages such as Persian, Turkish, Urdu, Malawian or Senegal is more than can be counted. The words derived from Arabic that exist in Spanish, Portuguese, German, Italian, English or French are also numerous [Bateson, 2003].

4 Distinguishing features of the Arabic language

Our classification of the language is illustrated in Figure 2. The syntax of Arabic is primarily classified into particle, noun, verb, clause, verbal sentence and nominal sentence. These categories are explained in the following subsections.
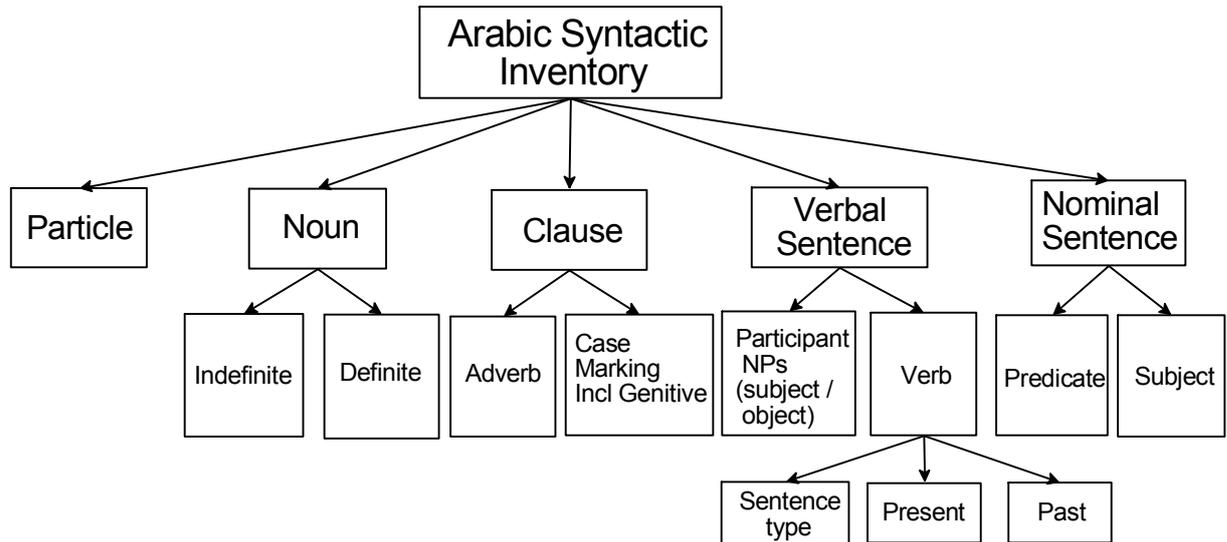
Figure 2: A classification for the Arabic language syntax

4.1   Particles

A particle is a word whose meaning cannot be understood without the context of a noun or verb. For example, *'min'* from, *'ala'* on, *'fi'* in, *'ella'* till or to and *'lan'* Will not [Khan, 2007a]. In the sentence, 'The man went to the school'. 'the' and 'to' are both particles.

Table 6: Noun example in Arabic

Arabic	Gloss	Arabic Meaning	English Translation
لنْ يَذْهَبَ	lAn yA.dhAbA	will not 'he' go	he will not go

The particle 'Lan' is used to negate future events. It is used within the imperfect tense [Versteegh, 2001]. An example is shown in Table 6.

4.2   Noun

A noun describes tangible or intangible objects or persons. Nouns are independent of other words in indicating its meaning. It also does not have any tense. See example in Table 7.

Table 7: Noun example in Arabic

Arabic	Gloss	English Translation
رجل	rAjOl	man
شجرة	ˆsAjArAh	tree

Definite Nouns are of seven categories:
1. Proper nouns e.g. James, Omar.
2. Pronouns e.g. he, you, I.
3. The demonstrative pronoun, e.g. this, that.
4. The relative pronoun, e.g. who.
5. Vocative case, e.g. O man, O boy.
6. The noun having 'al' e.g. the horse, the man.
7. A noun which is related to any of the above-mentioned definite nouns, e.g. Zaids book, this person's book, the book of the man.

4.3   Types of Noun

There are two types, definite and indefinite. Adjectives in Arabic usually follow nouns and agree with them in terms of number, gender, case, and definiteness/indefiniteness.

4.3.1   Definite article
A noun normally can be considered as definite (in Arabic: maarifa) when the speaker and the reader know about it, example in Table 8.

Table 8: Definite example in Arabic

Arabic	الكتاب الذي تبحث عنه فوق الطاوله
Gloss	a:lkEtAbO a:l.dy tAb7ˆtO 3nhO fowq a:l*ta:wElAh
English Translation	The book you are looking for is on the table.

In the example, the word 'book' is definite by using the definite article 'the', since both the speaker and the listener know about what book they are dealing with. The definite article in Arabic is used to introduce and talk about a known subject The Arabic language uses the same article for all nouns, be they male or female, singular or plural. The article is written before the noun it refers to and, graphically, it appears to be attached to it.

4.3.2   Indefinite article
Indefinite (in Arabic: nakira) is translated as 'a' or 'an' in English, e.g. a man, an apple, water. There is no need to translate it everywhere as in the example of water. The absence of the indefinite article is a potential source of problems for Arabic-English machine translation.

Table 9: Definite example in Arabic

Arabic	وجدت كتابا على الطاوله هل هو لك ؟
Gloss	wAja:dtO kEtAba 3a:l.y a:l*tawElAh hAl hOw lAk?
English Translation	I found a book on the table, is it yours?

In table 9 the word 'book' is introduced by the indefinite article 'a', to express the meaning that the speaker doesn't know much about the book he's speaking about. Hence in Arabic there is no letter 'a' before indefinite [Khan, 2007a].

4.4   Verb

A verb on the other hand describes an action. There are only two tenses for verbs that may be used to create the equivalent meaning in other languages:

4.4.1 The perfect tense 'mady' الفعل الماضي

The perfect tense, which indicates that an action has been completed, describe something that happened previously. An example is shown in Table 10.

Table 10: perfect tense 'ma:dy'

Arabic	Gloss	English Translation
كَتَبَ	kAtAbA	he wrote.

4.4.2 The imperfect tense 'mO.da:rE3' الفعل المضارع

The imperfect tense 'moddarea', which indicates that an action has not yet been complete but is being done or will be done. Describe something that is happening at the moment. An example is shown in Table 11.

Table 11: imperfect tense 'moddarea'

Arabic	Gloss	English Translation
يَكْتُبُ	yAktObO	he is writing.

In the Arabic language the verbs can be absolute which means there is no additional letter in the root. That means all letters are original letters. However, in the imperfect tense 'moddarea' there are additional letters on verbs. They are called imperfect letters. Those letters are given in Table 12.

Table 12: imperfect letters in Arabic

Arabic Letter	English Translation
ا	a
ت	t
ن	n
ي	y

Table 13: future tense in Arabic

Arabic	Gloss	English Translation
سوف يكتب	sawfa yaktObO	he will write
سيكتب	sayaktObO	he will write

The imperfect letters must be prefix. Only one letter from the imperfect letter can be attached to a verb, this letter appears graphically as a prefix with the original letters of a verb [Abed, 1990]. The word 'sawfa' if it is before the imperfect tense then the verb is its future meaning. Graphically a word like this will look like [sawfa + imperfect] or [s + imperfect] similar to the example in Table 13.

4.4.3 Imperative

Some morphologists [ibn Ajurum, 1332] regard the imperative as a third category of verb. The imperative (order) is something that may happen in the future and it must be an order from the speaker to the listener. It always has the same characteristic vowel as the jussive [Carl Paul Caspari, 1859]. See example in Table 14. 'Open' is an order tense in the Arabic language.

Table 14: imperative in Arabic

Arabic	Gloss	English Translation
افتح الباب	-eftA7 a:lba:b	open the door

4.5 Sentence

A sentence is a string of words that expresses a semantically complete message. The sentence in the Arabic language has six forms [Ibn-Hisham, 1359]:

4.5.1 Two nouns
Two nouns, such as: Ahmad (is) engineer.

4.5.2 Two clauses
Two clauses, such as: If Ahmad ask, Yasser answer.

4.5.3 Verb and noun
Verb and noun, such as: Ahmad asked.

4.5.4 Verb and two nouns
Verb and two nouns, it is only one type in كان و أخواتها *ka:n wa *a.7wAthA* kan and her sister. They are كان *ka:n* was, صار *sa:r* to become, أصبح *a*sbA7A* to become, أضحى *a.d7.y* to become, أمسى *amsy* to become, ظل *zA2lA* to remain, بات *ba:t* to be, ليس *lAyesA* it is not. In the case of equational sentences, leave the subject in its nominative case, but change the inflected predicate to the accusative case [Khan, 2007a], an example is shown in Table 15.

Table 15: kan and her sister

Arabic	Gloss	English Translation
كان الأكلُ لذيذاً	ka:n a:lAklo la.dy.d*A	The food was delicious.

4.5.5 Verb and three nouns
Verb and three nouns, it is only one type in ظن وأخواتها. *znnA wA *a.7wa:tha* .zan and her sister. They are ظن *zA2nA* to guess, حسب *hAsEbA* to consider, علم *3lEmA* to learn (about), جعل *jA3lA* to make, صير *s2yArA* to make. Usually come before the nominal sentences 'subject and a predicate', an example in Table 16.

Table 16: .zan and her sister

Arabic	ظن أحمد القياده سهلة
Gloss	.zA2nA *a7mAd a:lqya:dAh sAhlA.
English Translation	Ahmad guess leadership easy.

4.5.6   Verb and four nouns

Verb and four nouns, it is only one type in أعلم و أرى *a3lAmA wA*ar.y informed and showed. They are أعلم *a3lAmA informed, أرى *ar.y showed, أنبأ *an2b*aA told. نبأ n2b*aA told, أخبر *a.7bAr told, خبر .7bAr told. حدث 7A2dAˆ talked.

Table 17: informed and showed

Arabic	أعلمتُ عَمْراً خَالِداً تلميذاً
Gloss	*a3lAmtO 3OmAra:*A .7AlEda:*A tElmE.da*A
English Translation	I informed Omer that Khalid (is) student.

a'lam when it has hamza above it can have four nouns [ibn Abd Allah Ibn Malik, 1984], an example is provided in Table 17.

4.6   The Nominal Sentence

The nominal sentence contains two parts (subject and predicate) without any expressed verb. It begins with a noun. In Arabic there is no copula verb 'to be' [Abn-Aqeal, 2007]. The verb 'to be' is understood and then predicate subject. Both the subject and the predicate have to be in the nominative case, an example is shown in Table 18.

Table 18: Nominal Sentence

Arabic	زيد طالب
Gloss	zAyd*O *tAlEb*O.
English Translation	Zaid (is) student.

4.7   The Verbal Sentence

The verbal sentence is the basic sentence. It also contains two main parts (Verb and participant NPs) depending on the verb transitivity; there may be one or more participants. The verbal sentence has the grammatical relations of traditional subject and object etc. The sentence which begins with a verb will have order of a verb (V), subject(S) and object (O) or verb (V), object (O) and subject(S). The only combinations that do not occur in Arabic are OSV and SOV [Attia, 2004].

4.8   Clause

A clause in Arabic may be simple or complex. Conjoining two simple clauses within a coordinator or subordinate relationship forms a complex clause. The adverb in the Arabic language is an element that indicates the place or time of the event action [Khan, 2007b].

4.9   Challenges of Arabic to English Machine Translation

Arabic has a large set of morphological features. These features are normally in the form of prefixes or suffixes that can completely change the meaning of the word. Also, in Arabic there are some words that hold the meaning a full sentence for example سنسافر in English would translate to; we will travel. This means any MT system should have a strong analysis to get the root or to know in one word that there is fact a full sentence in the English equivalent. Arabic has free word order; this is a huge challenge to MT due to the vast possibilities to express the same sentence in Arabic.

    1. Verb Noun Noun
    2. Noun Verb Noun

This means that we have a challenge to identify exactly what are the subject and the object. Table 19 and Table 20.show this challenge. Please note that Table 19 and Table 20 should be read from right to left.

Table 19: Verb Noun Noun first possible.   Table 20: Verb Noun Noun second possible.

يحب زيد أمل		
Zaid loves Amal		
أمل	زيد	يحب
Amal	Zaid	love
noun	noun	verb

يحب أمل زيد		
Zaid loves Amal		
زيد	أمل	يحب
Zaid	Amal	love
noun	noun	verb

The different in Table 19 and Table 20 the position of the actor in Table 19 actor is first argument of the verb. In Table 20 actor is second argument of the verb. Also both sentences have the same meaning.

5. Summaries and Future Work

This paper has taken a subset of the features of Arabic that must be considered for the development of a MT tool from Standard Arabic to English. While some of the features unique to Arabic might have straightforward solution in MT, others will be more complex due the scale of potential ambiguity. Free word order for example will have to be addressed with a complete array of solutions for every case. A rule-based system to infer the meaning from a foundational form such as the SVO combination that English speakers are used to, might be a reasonable starting point. Although in Interlingua system is preferred, the initial target language will be English. Our main challenges for future work is to 1) develop the UniArab system using the Role and Reference

Grammar (RRG) technique which is currently in progress using Java and XML, and 2) evaluate the UniArab language and system by applying them to translate an Arabic text as source language into an English text as target language.

References

[Abed, 1990] Abed, S. (1990). *Aristotelian Logic and the Arabic Language in Alfarabi.* SUNY Press.

[Abn-Aqeal, 2007] Abn-Aqeal (2007). *Sharah Abn-Aqeal ala'lfat Abn-Malek.* Dar Al'alem llmlaien.

[Attia, 2004] Attia, M. (2004). *Report on the introduction of Arabic to Pargram.* Presented at ParGram Fall Meeting, Dublin, Ireland.

[Bateson, 2003] Bateson, M. C. (2003). *Arabic Language Handbook.* Georgetown University Press.

[Carl Paul Caspari, 1859] Carl Paul Caspari, W. W. (1859). *A Grammar of the Arabic Language.* Williams and Norgate.

[ibn Abd Allah Ibn Malik, 1984] ibn Abd Allah Ibn Malik, M. (1984). *Alfiyat Ibn Malik.* Maktabat al-Adab.

[ibn Ajurum, 1332] ibn Ajurum (1332). *Shrh Ajurumiyah.* Dar Alkutab al'elmiah.

[Ibn-Hisham, 1359] Ibn-Hisham, A. (1359). *Qatr Alnada wa Bel Elssada.*

[Izwaini, 2006] Izwaini, S. (23 Oct. 2006). *Problems of Arabic machine translation: evaluation of three systems.* The British Computer Society (BSC), London. Pages 118–148.

[Khan, 2007a] Khan, M. A. S. (2007a). *Arabic Tutor - Volume One.* Madrasah In'aamiyyah.

[Khan, 2007b] Khan, M. A. S. (2007b). *Arabic Tutor - Volume Two.* Madrasah In'aamiyyah.

[Ryding, 2005] Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic.* Cambridge University Press.

[Versteegh, 2001] Versteegh, K. (2001). *The Arabic Language.* Edinburgh University Press; New edition.

Appendices

Appendix 1   Arabic Alphabet and the Corresponding Phonetics (AACP)

	Arabic Letter	Letter Name	Phonetic Value
1	ا	ALEF	a:
2	ب	BEH	b
3	ت	TEH	t
4	ث	THEH	^t
5	ج	JEEM	j
6	ح	HAH	7
7	خ	KHAH	.7
8	د	DAL	d
9	ذ	THAL	.d
10	ر	REH	r
11	ز	ZAIN	z
12	س	SEEN	s
13	ش	SHEEN	^s
14	ص	SAD	*s
15	ض	DAD	.d
16	ط	TAH	*t
17	ظ	ZAH	.z
18	ع	AIN	3
19	غ	GHAIN	.3
20	ف	FEH	f

	Arabic Letter	Letter Name	Phonetic Value
21	ق	QAF	q
22	ك	KAF	k
23	ل	LAM	l
24	م	MEEM	m
25	ن	NOON	n
26	و	WAW	w
27	هـ	HEH	h
28	ي	YEH	y
29	ء	HAMZA	@
30	إ	ALEF WITH HAMZA UNDER	-e
30	أ	ALEF WITH HAMZA ABOVE	*a
31	آ	ALEF WITH MADDA ABOVE	-a
32	ى	YEH	.y
33	ئ	YEH WITH HAMZA ABOVE	^y
34	ؤ	WAW WITH HAMZA ABOVE	^w
35	ة	TEH MARBUTA	.t
36	َ	FATHA	A
37	ُ	DAMMA	O
38	ِ	KASRA	E
39	ً	FATHATAN	*A
40	ٌ	TANWIN ALDAM	*O
41	ٍ	TANWIN ALKASER	*E
42	ْ	SKOON	&
43	ًّ،ٍّ،ُّ	SHADDA	2
44	؟	ARABIC QUESTION MARK	?
45	٪	ARABIC PERCENT SIGN	%
46	؛	ARABIC SEMICOLON	;
47	،	COMMA	,