A SURVEY OF AUTOTAMTED TOOLS FOR TRANSLATING ARAB CHAT ALPHABET INTO ARABIC LANAGUAGE

Lamiaa Mostafa
Business Information System Department
Arab Academy for Science and Technology and Maritime Transport
Alexandria, Egypt
Lamiaa.mostafa31@gmail.com

***Abstract*—Internet is a global phenomenon in our life. Arab internet users are increasing rapidly. However the tools that help those users' needs are not fully tolerated to fulfill their requirements. Arab Chat Alphabet (ACA) is created for this reason. Arab users created a new format of the Arabic language; this format uses English letters to express Arabic words. Usually ACA is used in Instant Messaging (IM) and chatting purposes. For this reason it is important to understand the different available tools that write this language and validate the usage of theses tools. The paper's aim is to report on the survey of the different ACA translation tools and emphasizing the importance of theses tools in other applications like text mining, Information Retrieval and English Arabic translation. An experiment is executed to validate the translation tools, results had shown that Microsoft Maren has the highest accuracy compared to the other tools.**

Keywords- Arab Chat Alphabet; Text Mining ;Translation Tool;

I. INTRODUCTION

On the internet usage, Arabic language was not used as the activity base compared to English language because of huge effort to process the Arabic script since that Arabic language monophonical analysis a very complex task (Hoseini et al.,2011). Arabic language has a very large population all over the world (20 middle east and northern African countries), also Arabic is the official language of Muslims (Beseiso and Ahmad, 2010).The processing of Arabic language has many obstacles as language has many particularities, complex morphological analysis, and hard to identify names or abbreviation since there is no capitalization (Beseiso and Ahmad, 2010).

Arab youth who exposed to the internet usage has five main requirements: chatting, searching, emailing, online discussion and entertainment (Shen,and Shakir, 2009).Arabic Chat Alphabet (ACA) is the way in which Arabic users write their emails or chat. ACA format looks like the format of Roman alphabet and it also uses English letters. An example of ACA is "3elm, (علم)" which means science. A large part of the text written in the internet Webpages and documents is written in English, Different internet activities uses English language as shopping ,games, news, and educational issues; The tendency of all of documents and sites is to be written in English, rather than any linguistic interface. The processing of any text passes by some specific steps: parsing, stopword removal and stemming (Mostafa et al., 2011).For the complexity of the morphological analysis of Arabic and also the many stopwords in its sentences , this make the processing of Arabic language a very exhausting and complex step.

Arabic words are either words appear in the sentences and don't have any meaning or indications about the content such as (so لذلك, with بالإشارة)or a sequence of the sentences like (firstly اولا, secondly ثانيا), or pronouns such as (he هو)or important words that reference the content but with many different shapes: teach is written in Arabic by the four different words (يعلم- يدرس –يلقن - يحاضر) (Bassam et al.,2011). It is important to easily translate from English language to Arabic language and vise versa. The translation tools that will be discussed on the following sections are used for this purpose. These tools help Arabic users and English users to communicate in a better way. The tools solve the complex problem of Arabic language processing by creating a middle language that can be processed easily since it is written in English letters. The rest of this paper will be organized as follows. In section 2, the paper reviews related work on Arab Chat Alphabet tools. In section 3, the experiment is detailed, and section 4 defines the experiment results. The paper concludes with the discussion section and the work summaries.

II. RELATED WORK

A. *Arabic Chat Alphabet*

"Arabic Chat Alphabet (ACA) (also known as: Arabizi, Arabish, Franco-Arab, or Franco) is a writing system for Arabic in which English letters are written instead of Arabic ones. Basically, it is an encoding system that represents every Arabic phenomenon with the English letters that matches the same pronunciation" (Elmahdy et al. ,2011).

"ACA is a natural language that includes short vowel that are missing in traditional Arabic orthography" .There was a comparison between ACA-based approach and Modern Standard Arabic; however the result has shown that 86% of Arabic computer users confirmed that they type faster using ACA and that ACA is more accurate than the graphemic baseline (Elmahdy et al.,2011). The formal Arabic language is the Modern Standard Arabic (MSA).MSA is used in the formal bases like news broadcast, formal speeches, and books. However, MSA is not the Arabic spoken language for speaking in the everyday life. The following figure describes the differences between the Arabic letters and the ACA letters.

Type	IPA	Arabic	SAMPA	ACA
Consonant	ʔ	ء (ق)	?	2, '
	b	ب	b	b
	p	پ	p	p
	t	ت (ث)	t	t
	ʒ	ج	Z	j
	ħ	ح	X\	7
	x	خ	x	kh, 5, 7'
	d	د (ذ)	d	d
	r	ر	r	r
	z	ز (ذ)	z	z
	s	س (ث)	s	s
	ʃ	ش	S	sh
	ṣ	ص	s'	s, 9
	ḍ	ض (ظ)	d'	d, 9'
	ṭ	ط	t'	t, 6
	ẓ	ظ (ض)	D'	z, 6'
	ʕ	ع	?\	3
	ɣ	غ	G	gh, 3'
	f	ف	f	f
	v	ڤ	v	v
	q	ق	q	q, 8, 9
	k	ك	k	k
	l	ل	l	l
	m	م	m	m
	n	ن	n	n
	h	ه	h	h
	w	و	w	w
	j	ي	j	y

Figure 1. Sample utterances with corresponding IPA and ACA transcriptions (Elmahdy et al., 2011).

B. *Arabic Chat Alphabet Translation Tools*

There are many tools that can be used to convert from Arabic to ACA and vice versa. The following table shows a comparison between some of these tools.

TABLE I. ACA TRANSLATION TOOLS

	Type	Creater
Microsoft Maren (Cairo Microsoft Innovation)	Application	2009
Google Transliteration IME ("Google IME")	Both	2009
Yamli ("Yamli")	Webpage	2006
Eiktub ("Eiktub")	Both	2008
Yoolki ("Yoolki ")	Webpage	2005

Microsoft Maren (Cairo Microsoft Innovation) allows internet user to type Arabic in Roman characters and have it converted on the fly to Arabic script.

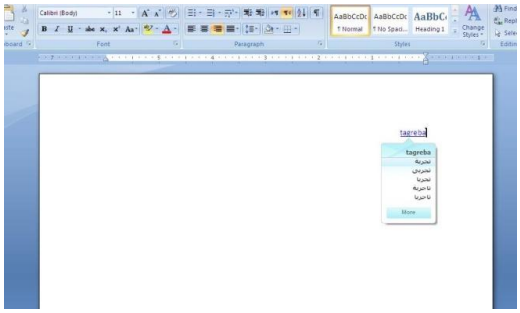

Figure 2. Microsoft Maren interface as a plug in in Microsoft office(Word).

Google Transliteration IME ("Google IME"), one of Google tools that make Arabic users write ACA in a very easy way. It was published in December 2009 for the offline use for the Indian languages, later it was developed for other languages like arabic. Different transliteration applications like Google Transliteration IME facilitate the users to convert Sindhi Scripts into Roman Script.

Yamli ("Yamli"), it is a real time transliteration system, the system translates Arabic using English Characters. Yamli has 2 interfaces the search engine and the Smart Arabic Keyboard. In July 2006 ("Yamli Information"), Habib Haddad launched Yamli.com including the Smart Arabic Keyboard for writing Arabic without an Arabic keyboard. Yamli search engine was published in March 2008 in the form of a free Application Programming Interface (API).

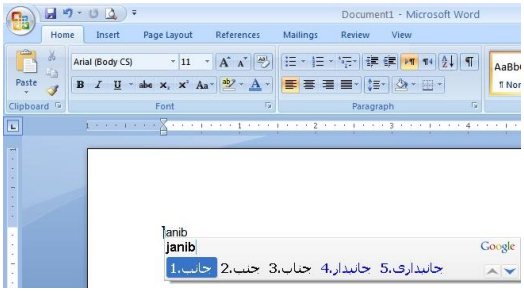

Figure 3. Google Transliterion IME interface.

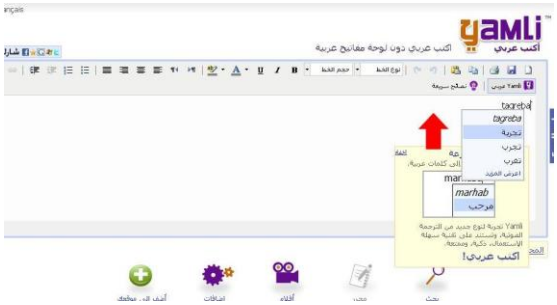

Figure 4. Yamli webinterface of smart arabic keyword.

Eiktub ("Eiktub"), an arabic transliteration text editor and search engine that can be used for ACA translation purpose. Yoolki ("Yoolki"), it is similar to Eiktub, it considers being an online Arabic transliteration, and web based editor and Arabic search engine. Yoolki homepage is divided into 2 partitions, one for the user writing language ACA on the right hand side, and the other textbox is the translated arabic language format on the left hand side.

Figure 5. Eiktub API as tool for writing arabic chat alphabet.

Figure 6. Yoolki homepage.

C. Arabic language Applications

A prototype for English to Arabic Machine Translation (EAMT) system is created by (Aref et al.,1992). The system phases are divided into: Knowledge base of Arabic terms, semantic classification of Arabic words, language independent semantic representation, using Object Oriented Representation (OOR), and finally a generator which generate the Arabic text from the semantic representation.

Text classification is the process of group similar features documents in one group. Arabic text classification was tested using Support Vector Machine (SVM) used with Sequential Minimal Optimization (SMO), Neural Network (Osman,2012), Naïve Bayesian (NB), and J48. The results had chosen that SMO classifier had the highest accuracy and the smallest consumption time (Bassam et al.,2011).A system that classifies Arabic documents is created in (Ghwanmeh et al.,2009).The process of features selection is the most important process in the system; features are used to represent the main document's content.

242 Arabic abstracts from the Saudi Arabian National Computer Conference were the testing set. After the survey and the interview of 74 students from two universities in UAE, the results had shown that there is a positive significant impact on ease use of the language and its usages in the internet applications (Shen,and Shakir, 2009).

In (Palfreyman and Khalil, 2006), researchers' tested the usage of ACA among female university students in the United Arab Emirates, the corpus used was Instant Messenger (IM) conversations. The results had shown the highly usage level of ACA in IM conversations. The researchers in (Beseiso and Ahmad, 2010) evaluated tools for ontology creation not for English language but for Arabic language. They concluded that the technology trend towards the Arabic language is still very limited and it is important to work towards this direction. Natural language processing techniques, titles can be generating on the Arabic paragraphs without extracting words from the document. The results had showed that its better to use the document itself to the training model instead of its metadata (Adwan et al.,2012).

III. EXPERIMENT

The experiment depends on testing the different tools (Microsoft Maren (Cairo Microsoft Innovation), Google Transliteration IME ("Google IME"), Yamli ("Yamli"), Eiktub ("Eiktub"), Yoolki ("Yoolki") on group of Webpages in ACA format and translating these Webpages into Arabic language. The process of translating from English to ACA language is developed by an application that map the English words into ACA words, and this mapping tool's result is tested manually. Each translation tool is fed by the group of Webpages for the translation from ACA to Arabic. All English Webpages is translated using Google translate. There will be a similarity check between each translation tool and Google translate for testing the accuracy of the translation step. The following figure shows the translation experiment steps.

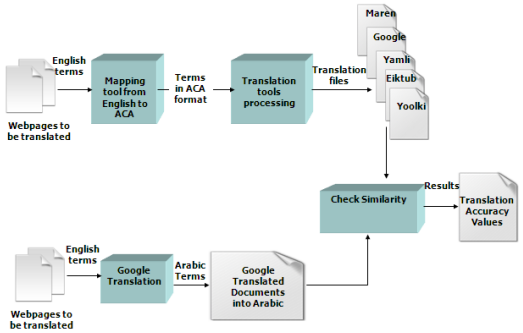

Figure 7. Translation Experiment Results.

To accomplish the experiment design, a data set for Webpages was used. Finding a collection of related Webpages grouped by fields is a difficult step. DMOZ ("DMOZ") data set (Open Directory) can be used for this task. DMOZ directory is edited by human. A group of volunteer editors construct and maintain this directory. The Open Directory is 100% free. The data set used consists of 1000 Webpages of shopping domain extracted from DMOZ dataset.

IV. EVALUATION AND DISCUSSION

The goal of the experiment is to validate the result of the proposed approach by measuring the Accuracy values of each translation tool. The experiment compares each translation tool result file. The following table shows the accuracy level based on the resulted file of each tool. The result had shown that Microsoft Maren has the highest accuracy value.

TABLE II. ACCURACY VALUES

Translation Tool	Accuracy
Microsoft Maren (Cairo Microsoft Innovation)	92.796%
Google Transliteration IME ("Google IME")	87.381%
Yamli ("Yamli ")	86.552%
Eiktub ("Eiktub")	77.113%
Yoolki (Yoolki)	82.429%

V. CONCLUSION AND FUTURE WORK

The huge number of arabic users over the internet burdens the technology creator to invent new tools for fulfilling those user needs. The translation process between languages has many obstacles. Especially the arabic language translation is a complex task since the complexity of the morphological analysis of this language. The spreading of ACA usage between Arab youth in different internet activities like emailing and chatting is a motive for focusing on this language. The paper provides an experiment for testing the accuracy level of different translation tools which are Microsoft Maren (Cairo Microsoft Innovation), Google Transliteration IME ("Google IME"), Yamli ("Yamli"), Eiktub ("Eiktub"), Yoolki ("Yoolki")). Future work could be started in creating a linguistic tool for the translation between English and Arabic chat alphabet. Researchers should provide a formal dictionary for ACA terms and also focusing on using semantic relationships like synonyms relationship in this language.

REFERENCES

Adwan, O., Alnajada, H., Faris, H. Obiedat, R. (2012). Proposing Titles for Paragraphs Written in Arabic Language. *European Journal of Scientific Research ISSN 1450-216X Vol.68 No.1, pp. 110-116.*

Aref,M., Al-Mulhem, M. and Al-Muhtaseb, H.(1992). English to Arabic machine translation: a critical review and suggestions for development. *King Fahd University of Petroleum and Minerals Dhahran, Saudi Arabia.*

Bassam, Al-Shargabi B., AL-Romimah,W.,and Olayah F.(2011). A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination. *ISWSA'11, April 18–20, 2011, Amman, Jordan, ACM*

Beseiso,M., Ahmad,A. (2010). A Survey of Arabic Language Support in Semantic Web. *International Journal of Computer Applications (0975 – 8887) Volume 9– No.1, November 2010.*

Cairo Microsoft Innovation Lab, Retrieved Febraurary, 12,2012 "MicrosoftMaren,"2009,http://www.microsoft.com/middleeast/egypt/cmic/maren

DMOZ, Retreived Febrauary, 12,2012 :http: //www.dmoz.org/ .

Eiktub,Retreieved Febrauray,12,2012 :http://www.eiktub.com/

Elmahdy, M., Gruhn, R., Abdennadher, S.,and Minker, W. (2011).Rapid Phonetic Transcription using everyday life natual chat alphabet orthography for dialectal Arabic speech recognition. *IEEE 978-1-4577-0539-7/11.ICA SSP 2011*

Ghwanmeh, S., Kanaan, G. Al-Shalabi,R. Ababneh, A.(2009).An Enhanced Text-Classification-Based Arabic Information Retrieval System. *IGI Global* .

Google IME,Retrieved Febraurary, 12,2012 :http://www.google.com/ime/transliteration/

Hoseini, M.(2011).Modeling the Arabic language through verb based ontology. *International Journal of Academic Research Vol. 3. No. 3. May, 2011, II Part.*

Mostafa,L. (2011). Webpage Keyword Extraction using Term Frequency. 3rd *IEEE International Conference on Information Management and Engineering (IEEE ICIME 2011).*

Osman,Z. (2012). Classification of Arabic Text using Language Properties. *European Journal of Scientific Research ISSN 1450-216X Vol.68 No.2 (2012), pp. 191-198.*

Palfreyman D., Al Khalil,M. (2006). A Funky Language for Teenzz to Use:" *Representing Gulf Arabic in Instant Messaging*.

Shen, K. & Shakir, M. (2009), 'Internet usage among young Arab students: preliminary findings', *European, Mediterranean and Middle Eastern Conference on Information Systems, pp. 1-10.*

Yamli information, Retreived Febrauary, 12,2012 :http://en.wikipedia.org/wiki/Yamli#mw-head.

Yamli,Retreived Febrauary, 12,2012: http://www.yamli.com/ar/

Yoolki,Retreived Febrauary, 12,2012:http://yoolki.com/