مجلة العدوي للسانيات العرفية و تعليمية اللغات

Issn: 2773/ 3300

http://www.asjp.cirect.dz/en/presentation Revue/804

المجلد:01 العدد: 02 (2021) ص 53-63

The Effect of advanced translation applications on Arabic language identity

أثر تطبيقات الترجمة المتقدمة على هوية اللغة العربية

* Aicha Laachemi

Medea University

laachemi_aicha@yahoo.fr

معلومات المقال

تاريخ الارسال:

2021/08/28

تاريخ القبول:

2021/10/10

الكلمات المفتاحية:

- ✓ اللغة العربية
- ✓ الهوية
- ✓ الترجمة الآلية
- ✓ الأنظمة الذكية
- ✓ الانترنت.

الملخص:

على مرّ نحاول من خلال هذه الورقة البحثية ان نلتفت حول ظاهرة الترجمة الآلية الى اللغة العربية واشكالية الهوية، فقد أدى الانفجار المعلوماتي الذي عرفه العصر الحالي إلى ضرورة اللجوء إلى وسائل التقنية الحديثة في سبيل الإسراع بعملية نقلها وتناقلها بين الشعوب المختلفة واستخدام الحاسوب في عملية الترجمة بصور شتى .ولابد من الاشارة الى أن الترجمة الآلية أبعد ما تكون عن الوفاء بالحق التعبيري البشري، لأنها عادة تكون مبرمجة على الترجمة الحرفية دون غيرها، أي تقريب المعنى إلى المتلقي دون الوفاء إلى القدرات التعبيرية والبلاغية للنص الأصلي وهذا ما يؤثر سلبا على هوية اللغة. وخلال هذه الورقة البحثية سنحاول عرض اشكالية استخدام التطبيقات الحديثة للترجمة وأثرها على اللغة العربية، التي نجدها خلال الاستخدام المكرر للمواقع الإلكترونية، فهل يمكن الحديث عن الهوية الإلكترونية للغة العربية في إطار الترجمة الآلية الناتجة عن استخدام التطبيقات الحديثة لها؟

The Effect of advanced translation applications on Arabic language identity

Abstract :	*Article info*
This paper explores the phenomenon of the Advanced translation applications rise and its effect on the Arabic language ldentity. The rise of the new technologies in the current era led to the need to resort to modern technology in order to speed up the process of transferring it among different peoples and using machine translation process in various forms. It should be pointed out that machine translation is far from fulfilling the human expression right because it is usually programmed on literal translation without any other approximation of meaning to the recipient without satisfying the expressive and rhetorical capacities of the original text. In this paper, we will try to present the problems of machine translation that could rise during the repeated use of websites, through the following problimatic: Can we talk about the electronic identity of the language in the framework of the so-called machine translation?	*Received 28/08/2021* *Accepted 10/10/2021*
	: Keywords. ✓ *Arabic language* ✓ *machine translation* ✓ *smart systems,* ✓ *identity internet*

. *auteur correspondant*

Introduction:

Communication tools make the world like a small village, and consequently people can contact with others, who are from different societies or who speak different languages. This communication cannot happen effectively without translation applications because they can be found anytime and everywhere.

There are a number of studies that have developed translation applications for the English language with so many other languages, except the Arabic it has not been considered yet. Therefore, the aim of this paper is to highlight a roadmap for our proposed translation

applications to provide an enhanced Arabic-English translation based on Semantic and to illustrate its work.

Languages are ways of expressing and recognizing the many social identities people have. They are both acquired naturally and taught formally and both natural acquisition and formal teaching create, strengthen or weaken the links between languages and identities. An important language identity link is the one between ‘national language’ and ‘national identity’. This link may be created, strengthened or weakened by formal teaching in schools, especially in language as subject.

People acquire new identities and new languages or language varieties throughout life; it is a dynamic process[1]. If young people become conscious of this process, they can play with their languages and identities, shifting from one language to another within the same conversation, signaling a change from one identity to another. Young people have been shown to be adept at this as they move from one social situation to another becoming consciously plurilingual.

Arabic is one of the major languages that have been given attention by translation applications researchers since the very early days of Machine Translation and specifically in the U.S. The language has always been considered "due to its morphological, syntactic, phonetic and phonological properties to be one of the most difficult languages for written and spoken language processing. "[2]

Arabic "differs tremendously in terms of its characters, morphology and discretization from other languages." [3] Accordingly, researchers cannot always import solutions from other languages, and today Arabic machine translation still needs more efforts to be improved, mainly in the area of semantic representation systems, which are essential for achieving high quality translation.

In this research paper, we have extend a phrase-based statistical translation applications system in many ways. Phrase-based systems are considered one of the best performing

approaches. Two probabilistic models are used, a translation model and a language model. The translation model is trained on bilingual corpora and is used to model the faithfulness of the translation[4] .

The language model is trained on monolingual corpora and is used to improve the fluency of the translation output. I have been working on improving the translation quality. This is done by focusing on three different aspects:

1- The first aspect is reducing the number of unknown words in the translated output. we concentrate on three types of unknown words:

- First, words which are not correctly morphologically segmented - this can be corrected by using a better segmentation.
- Second, the entities like numbers or dates that can be translated efficiently by some transfer rules.
- Finally, working on the transliteration of named entities.

2-The second aspect of this work is the adaptation of the translation model to the domain or genre of the translation task. This is done by weighting different bilingual subcorpora according to their importance.

3-The third technique is weighting of translation models using perplexity optimization. Another way is using a multi-domain translation model architecture. In this architecture, the computation of the translation model probabilities is delayed until decoding time, allowing dynamic instance weighting using optimized weights[5] .

1- Machine Translation approach related identity challenges

In this category, the problems are specific to the MT approach or method. For example, in corpus-based approaches, we use specific bilingual and monolingual corpora and hence closed vocabulary. This leads to several problems as follows[6] :

1. Some source words will not be translated by the MT system because they are unknown to the translation model. These are called Out-Of-Vocabulary words. Examples of such unknown words are proper nouns, verbs with different morphological form, words with different inflection form and entities like number or dates. Transliteration of proper nouns can be used to decrease the number of Out-Of-Vocabulary in the translation output.

2. Unknown target words to the language model.

3. Mismatch between the domain or the style of the bilingual and monolingual training corpora and the translation task. For example, when the MT system is trained on modern standard Arabic and formal corpora, but it is used to translate Arabic dialectal and informal text.

4. Segmentation errors: words are wrongly segmented instead of being left unprocessed or unsegmented words.

5. Low resource languages: small bilingual corpora mostly will lead to a bad Translation model and many examples, while small monolingual corpora could

lead to non-fluent translations and bad formed target sentences.

6. Pre-ordering and inflection of languages with flexible sentence components is a challenge since several orders can be correct and acceptable but inflection could be different in each order eg:

التفاح أكله احمد .vs اكل احمد التفاح

7. Limited data resources causes data sparseness problem. How often the word occurs in the training data correlates with the machine translation quality. If the word (or phrase) occurs rarely, it causes problems in word alignment, calculation of the translation probabilities and other statistical modeling training.

If the word never occur, this causes the problem of MT, which we discussed in the first point above. The data sparseness problem is generally addressed by using more data, which help in a better word alignment , a better estimation of the words and phrases translation probabilities as well as additional context for PBSMT[7].

Nowadays, one of the common sources of corpora is the web. Internet users who have different backgrounds and education levels write some collected texts[8]. People can make spelling mistakes and they can have their own writing style like stressing on some letters by repeating them or by using some punctuations for other purposes like emotional expression or for text decorations. Therefore, we can have two categories of these problems :

- ✓ Orthographic errors.
- ✓ Writing behavior on digital media.

2- The Entity of translation applications

We focused on number, date, email and URLs entities. Numbers and dates are part of the cultural preference of any language and country. For example, date format in France is different from the date format used in the EAU or the UK (e.g. day/month/year vs. month/day/year). We can also observed a difference in format of numbers (e.g. 2 450, 30 in algeria vs. 2,450.34 in the USA). It is important to translate them by phrase entries in the phrase table, but this is not always possible because usually they have many variations.

Unknown entities are considered as obstacles in the language identity, which their translation should not be a complex task. They can be translated to target language using some rules. This problem is more critical between languages using different writing scripts like Arabic and English than between French and English for example. Since there is no integrated method to handle such entities translation, we developed a procedure to detect numbers, dates and other entities and then transform them from the source language to the target language.

One of the challenges in MT research is dealing with the arabic language words. One way to decrease the arabic language rate is by transliterating proper nouns (or names). In this paper, we will focus on dealing with the challenge of transliteration of Arabic proper nouns into English. Transliteration is the process of writing a word (mainly proper nouns) from one language in the alphabet of another language[9] .

This requires mapping the pronunciation of the word from the original language to the closest possible pronunciation in the target language. Both the word and its transliteration are called a Transliteration Pair (TP).

Since we are using a statistical-based approach throughout this thesis, we will need data to train the system. In this case, the training data should be a bilingual list of TPs in Arabic and English. Since we do not have this training data available, we have to deal with the automatic extraction of these TPs from the available corpora. In this work, I deal with two types of corpora, bilingual corpora and comparable corpora. A comparable corpus is a pair of corpora in two different languages, which come from the same domain. The automatic extraction of TPs from parallel or comparable corpora is called Transliteration Mining (TMI).

This work aims in developing pre-processing and post-processing engines to manage such identity entities, the value of each entity is substituted by a placeholder. The preprocessing engine uses the detection and transformation rules and apply them on the provided text. The separation of the rules in a separate file makes the change of the detection and translation rules more flexible. One post-processing and three preprocessing tools were developed[10] .

The preprocessing tools were applied to all kind of corpora, namely bilingual and monolingual training corpora. Not all preprocessed text contain the entities' values (numbers or dates, etc...), but it only contains the placeholder of each entity. This helps

decreasing data sparseness and decreasing the size of the translation model (i.e. the phrase table) and the language model.

A post-processing tool is responsible for replacing the placeholders in the translation output by their translated values using the source to target alignments provided by the decoder. One advantage of this technique is that we can keep the MT system independent of the source and target languages cultural preferences. At decoding time, we have the flexibility to select the required cultural preference needed for the translation task. For example, the same SMT system can be used to translate text from UK or USA by specifying the input type to the entities handling engine.

Since Arabic is a morphologically rich language, the selection of the suitable morphological segmentation options is one of the important preprocessing steps in MT research. There are many morphological schemes that can be used to segment the Arabic words. I evaluated various Arabic segmentation schemes from full word form to fully segmented form to explore the effect on the system performance and translation quality.

In order to address ambiguous Arabic/Algerian words translation errors, we worked on applying word sense disambiguation technique on them using their context. we integrate this technique into a phrase-based SMT system in order to improve the system performance in translating ambiguous words. Another challenge in translation applications is the dealing with the out-of-vocabulary words. we have performed research on several methods to decrease the out-of-vocabulary rate including proper noun transliteration[11].

3-The challenges of Arabic transliteration

There are several challenges related to Arabic translation One of the challenges is how to perform transliteration of Arabic in order to decrease the number of words in the translation output. This is a challenge because there are some Arabic letters which have no phonically equivalent letters in English (e.g. P and u), and also some English letters do not have phonically equivalent letters in Arabic (e.g. v). Another challenge is the missing of short vowels (i.e. diacritics) in the Arabic text[12] .

While they should be mapped to existing letters in English text during the transliteration process. Additionally, some Arabic letters can be mapped to any letter from a group of phonically close English letters (e.g. H. to p or b), and some Arabic letters can be mapped to a sequence of English letters (e.g. p to ’kh’). There is also a tokenization challenge, since unlike English, sometimes, the Arabic name is concatenated to one clitic (e.g. preposition. or conjunction or both together (e.g. H.), which requires an advanced detection and segmentation for these clitics before performing the transliteration.

The proposed TMI algorithm is based on the following pronunciation (and Hence transliteration) observations in the English language[13] :

1-In most cases, we can sort the letter’s impact on transliteration from low

to high as follows:

- ✓ Phonetically similar vowels have low impact.
- ✓ Phonetically dissimilar vowels have medium impact.
- ✓ Consonants letters have significant impact.

2. Double vowels producing a long vowel sound have more impact on the

Pronunciation of the English word.

3. A sequence of two or more different vowels has a special pronunciation, which has more impact on the pronunciation of the English word.

4. A vowel at the initial position or at the final position in the word has

Significant impact on the pronunciation (e.g. the names: Adham, Samy).

Finally, many languages contain specific entities (like e.g. dates and numbers) which require special treatment, and the Arabic language is one of them. In this work, we addressed the problem of translation of these entities. In this context, since there is no integrated

method to enable the correct translation of numbers and dates, some translation applications developed a method to detect numbers, dates and other entities and then transform them, if needed, from the source language format to the target language format .

In this paper, we have briefly explained an introduction to translation applications and its impact on the arabic language identity, its history and approaches. Since the translation applications is the bases of this paper, we focused on explaining the basics of Machine translation and covered different components of word-based and phrase-based translation applications, including the translation model and the language model. we introduced the current state-of-the-art in language modeling in a full section that covers n-gram back-off and neural network language models.

Conclusion

As an important new medium of human communication, the Internet is bound to have an important long-term effect on language use. It is too early to tell what that impact will be. The trends discussed in this paper could prove to be temporary, if, for example, the development and diffusion of Arabic language software and operating systems bolsters the use of Classical Arabic and stems the tide of online communication in English or in Romanized Arabic dialects. However, language use online, in Jordan and elsewhere, will be shaped not just by the technical capacities that technology enables, but also by the social systems, that technology encompasses.

In addition, as have pointed out, the major social dynamic shaping international media and communication in this age of information is the contradiction between global networks and local identities. In that light, it is worthwhile to consider whether the online use of English and Arabic, which might reflect broader and

[1] Warschauer, M. (2001). Singapore's dilemma: Control vs. autonomy in IT-led development. The Information Society, 17(4), 305-311.
[2] Barber, R. (1995). Jihad vs. McWorld. New York: Ballantine Books.
[3] Castells, M. (1997). The power of identity. Malden, MA: Blackwell.
[4] Ibidem.
[5] Barber R.op.cit.

[6] Takahashi, H. (2000). Dealing with dealing in English: Language skills for Japan's global markets (Report 7A).Washington, DC: Japan Economic Institute.
[7] Hairy, N. (1997). The reproduction of symbolic capital: Language, state, and class. Current Anthropology, 38(1), 795-805.
[8] Stevens, P. B. (1994). The pragmatics of street hustlers' English in Egypt. World English, 13(1), 61-73.
[9] Schaub, M. (2000). English in the Arab World. World English, 19(2), 225-238.
[10] Ibidem.
[11] Ibidem.
[12] Stevens, PB.op.cit.p80-85.
[13] Fandy, M. (2000). Information technology, trust, and social change in the Arab world. Middle East Journal, 54(3),378-394.

References:

1. Barber, R. (1995). Jihad vs. McWorld. New York: Ballantine Books.
2. Castells, M. (1996). The rise of the network society. Malden, MA: Blackwell.
3. Castells, M. (1997). The power of identity. Malden, MA: Blackwell.
4. Crystal, D. (1997). English as a global language. Cambridge: Cambridge University Press.
5. Hairy, N. (1997). The reproduction of symbolic capital: Language, state, and class . Current Anthropology, 38(1), 795-805.
6. McCarty ,S.(2009).Social networking beyond student lines in Japan, in M Thomas (Ed.), Handbook of research on Web 2.0 and second language learning (pp. 181-201), IGI Global. Online. (1998). Chronicle of Higher Education, pp. A27.
7. Schaub, M. (2000). English in the Arab World. World English, 19(2), 225-238.
8. Stevens, P. B. (1994). The pragmatics of street hustlers' English in Egypt. World English, 13(1), 61-73.
9. Takahashi, H. (2000). Dealing with dealing in English: Language skills for Japan's global markets (Report 7A).Washington, DC: Japan Economic Institutue.
10. Warschauer, M. (2001). Singapore's dilemma: Control vs. autonomy in IT-led development. The Information Society, 17(4), 305-311.