Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 73–80

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

Errors and non-errors in English-Arabic machine translation of gender-bound constructs in technical texts

Emad A. S. Abu-Ayyash*

The British University in Dubai, DIAC, Dubai, United Arab Emirates

Abstract

This paper has as its main goal investigating the errors and non-errors made by three machine translation (MT) systems in translating gender-bound constructs from English to Arabic in four selected technical texts. The three MT systems used in this study were Systran's Pure Neural Machine Translation (PNMT), Google Translate (GT) and Microsoft Bing (MB). The target gender-bound constructs were subject-verb agreement, adjectival-noun agreement and pronoun-antecedent agreement, which occurred naturally in four purposefully selected technical texts. The idea behind the choice of technical texts was to reduce the linguistic load found in other genres, such as literary texts, which involve utilising creative linguistic tools that are out of the scope of the present paper. Upon the qualitative examination of the target language texts, the findings revealed that the three MT systems had errors and non-errors in rendering gender-bound constructs from English to Arabic, and that errors transpired in certain co-textual environments.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Machine translation; gender; gender-bound cnstructs; English-Arabic translation

1. Introduction

Attitudes towards machine translation (MT) have been inconsistent and continuously in flux since the 1950s. In essence, adoption of MT as a valid option for translation was tied to the purposes that the field served and the effectiveness of MT in meeting the requirements of different organizational bodies. According to [1] and [2], MT

* Corresponding author. Tel.: 00971-50-7225701; fax: 00971-4-2791471.
E-mail address: emad.ayyash@buid.ac.ae

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.095

loosely passed through a number of stages of acceptance and adoption. When MT was first introduced as a translation option in the 1950s, it was used for military and intelligence purposes, particularly translating the information gathered from and about certain regimes. However, about ten years later, there was an evaluation stage, and MT was criticised heavily in a report written by the National Academy of Sciences ALPAC. In the 1970s, the field of MT lost more battles as MT projects lost government funding. The spiralling rise of MT began in the following decade, and there was an exponentially increasing reliance on MT systems. According to [1], "in 1984, approximately half a million pages of text were translated by machine" (p. 1). Nowadays, with the intricate web of technology, the internet and businesses that need fast and high-quality translations, MT is gaining more and more momentum, but to what effect? In a recent evaluation of neural machine translation, it is reported that fluency has increased, but that the status of adequacy is not yet definitive [16] and that post-editing of MT has become a reality [17].

The issues about adequacy in MT systems are a corollary of the vast differences between language systems at the levels of morphology and syntactic structures. The focus of this study is English and Arabic, which are acknowledged to be substantially different in a number of ways. For example, while the English declarative statement starts with a pronoun or a noun, the Arabic sentence, which, like English, can start with a noun or a pronoun, can also start with a full verb [8]. The pronoun system, which will be discussed further in the present study, is another area of disparity between the two languages.

The present study has as its goal investigating to what extent MT systems are able to render texts from English to Arabic when such texts include forms that have different rules, or grammatical representations, in the two languages. The grammatical constructs that have been selected to serve the purpose of this study were all related to gender, which is acknowledged to be an area of dissimilarity between English and Arabic as far as form is concerned [3]. Therefore, my intention in the current paper was to investigate the types of errors and non-errors perpetrated by MT systems apropos the rendition of gender-bound grammatical constructs from English to Arabic. The target forms have been fed into three MT systems within their natural occurrences, or co-texts, and not as individual, discrete units. That is to say, the selected texts where inserted into the three MT systems as are, yet the analysis focused on the rendition of the gender-bound forms within the produced translations. The transliteration system used in this paper is Qalam [15].

2. Related work

With the exponential rise in the demand for accurate translations, MT is again singled out for critical attention and scrutiny across translation research. Concomitant with MT research, there were attempts to identify the levels of linguistic description pertinent to MT. Levels that were under investigation included phonetics and phonology, morphology, word classes and grammatical categories, syntax, lexicon and semantics, and pragmatics and stylistics [4]. However, a shortcoming of MT systems is that they are 'limited by their relative ignorance of linguistic information' [5] (p. 953) despite the various attempts to involve programs that can identify word structure and sentence structure [6, 16]. A number of studies have investigated this claim with various foci on MT capabilities with different linguistic levels.

At the semantic and syntactic levels of language, [7] examined the notion of ambiguity and how it can be problematic for MT systems. The study identified five divergent areas between English and Arabic at the semantic level as far as ambiguity is concerned. These include category, homograph, transfer, pronoun reference and gender and number. Within the structural divergence between the two languages, [7] identified the areas of word order, tense and aspect, and agreement. These findings dovetail with the de facto assertions that English and Arabic each houses its own distinct linguistic structures and discursive features [8, 9].

At the word classes linguistic level, word ordering and agreement in three MT systems, which were Google, Tarjim and Systran, were investigated [10]. The study investigated the English-Arabic translations of certain structures as per gender, number, case and definiteness. The findings were reported in terms of the errors perpetrated by the MT systems in rendering such structures. The target structures were fed as discrete units into the MT systems, while in reality, these systems are normally fed with longer stretches of texts where the target structures are normally embedded. In fact, the practice of investigating MT systems' effectiveness using separate sentences as discrete input units has even become a tool of evaluating MT systems [11].