A Comparative Study of English-Persian Translation of Neural Google Translation

Mina Zand Rahimi
Shahid Bahonar University of Kerman
mina_zandrahimi@uk.ac.ir

Moein Madayenzadeh
Shahid Bahonar University of Kerman
moein2141373@gmail.com

Mahdi Alizadeh
Shahid Bahonar University of Kerman
mahdi1374.al@gmail.com

Abstract

Many studies abroad have focused on neural machine translation and almost all concluded that this method was much closer to humanistic translation than machine translation. Therefore, this paper aimed at investigating whether neural machine translation was more acceptable in English-Persian translation in comparison with machine translation. Hence, two types of text were chosen to be translated by Google Translate. The inputs have been translated in two distinctive methods. The outputs were investigated by the descriptive-comparative human analysis model of Keshavarz. Consequently, the results revealed that approximately the same errors were found in both methods. However, semantic aspects were improved.

Keywords: Neural Google Translation, Machine Translation, English-Persian Translation, Phrase-Based Machine Translation

1. Introduction

Translation has been done for over a half century in assistance with IT developments. According to Bowker (2002), in 21st century, human has become capable of applying technology in their translations, computer translation, and computer-aided translation. Examples of this technology could be word processors and electronic resources, as well as software used in translation, such as corpus-analysis tools and terminology management systems. They have been applied as an assistant, and in terminology, this method is called 'Machine Translation' (MT). In other words, "MT involves the use of computer program to translate texts from one natural language into another automatically," said Ping (as cited in Baker & Saladanha, 2009, p. 162).

A new more accurate MT system has been presented which is called 'neural machine translation' (NMT). Nowadays, google team applied this system in their new Google Translator application 'Google Neural Machine Translation' (GNMT). Traditional models of MT had the problem with analyzing syntactic aspects of inputs, and they translated literally. While, GNMT has tried to reduce the errors in different outputs. The aim of this paper is to analyze the reduction of those errors between English and Persian Languages.

1.1. Machine Translation

Machine translation has been a research subject although the perfection of translations has not been reached yet. Machine translation mainly categorized as computer-based translation and computer-aided translation (Baker & Saladanha, 2009). Warren Weaver, in 1947, developed the idea of Statistical MT but due to its complexity, it had been dropped (Pilevar & Faili, 2010).

1.2. Neural Machine Translation

In an end-to-end approach of NMT, there was a significant ability of learning the way of connecting input elements to output's. They also suggested two recurrent neural networks (RNNs) were involved in this ability. One got input text sequence and the other generated the translated output elements. Also, there were some relative three advantages and disadvantages for NMT to phrase-based machine translation (PBMT) (Wu et al., 2016). For instance, accuracy of NMT was functionally less than phrase-based translation systems, especially in long inputs. Moreover, its training is time-taking and resource-consuming, it lacks rare word strength in translating and, in some cases, it doesn't translate all input elements. The above-mentioned problems were three disadvantages of NMT. Thus, by applying these solutions to NMT, the relative translation quality would increase (Wu et al., 2016).

PBMT used to consider input sequence (e.g., a language's sentence) as individual words and phrases, therefore its output elements were largely unrelated to each other (Wu et al., 2016). They added, by the mean of RNNs, NMT mapped similar input sequence to output sequence (the same sentence in another language), In other words, there was a significant difference among PBMT and NMT unit of translation. Therefore, in NMT, unit of translation could be a whole sentence (Wu et al., 2016).

2. Methodology

2.1. Materials

The materials considered in this paper included English texts as the input and Persian text as the output data. For this, two types of text were given to the translator, one was a technical text including 115 words and the other, was a

non-technical text consisted of 84 words; then, these inputs were decoded and the relevant outputs were produced. The technical text was selected from a housing industrial article while the non-technical one was chosen from "the Bet" short story. Also, A 267-word text was used as an input and an output was a translated English text consisting 167 words by GNMT as well as translated English text by GMT (Google's previous method), which included 253 words. These data were analyzed and the subsequent errors of this translation were observed.

2.2. Procedures

A descriptive-comparative human analysis model of Keshavarz (1999) were used in the paper by which the translated texts were analyzed and each resulted error were comparatively observed. In this qualitative-descriptive study, the frequent correct and incorrect translated outputs were put into observation as they were classified to different types of errors such as tense errors, wrong preposition, word disordering, wrong collocations, verb-mismatching error, and active or passive voice related errors. According to Keshavarz's (1999) model, the frequencies of individual types of error were counted to reach a certain percentage.

2.3. Data Analysis

The collected data were analyzed unit by unit in each input type. According to Keshavarz's analysis model, the frequency of each type of occurred errors were observed.

Case 1: متغیر مستقل آن ساختار سرمایه بوده...

GNMT: The independent variable of capital structure

GMT: as independent variables that was capital structure

Analysis: In this case, there was a wrong preposition error in GNMT; however, GMT correctly translated this clause. The conjunction "که" in Persian need to be translated as "that, which, etc." so "of" is considered as an error.

Case 2:Specifically, the empirical study analyzes the HSI of low-priced housing in Beijing.

GNMT:به طور خاص، مطالعه تجربی تجزیه و تحلیل HSI مسکن کم قیمت در پکن.

GMT: به طور خاص ، مطالعه تجربی تحلیلHSIاز ارزان مسکن در پکن.

Analysis: In this case, there were word disordering and verb mismatching errors. It was noticeable that in almost all cases GNMT and GMT considered the word "analyze" as a noun, thus it would be translated as a noun as well. Moreover, in Persian, an adjective followed noun while it was the opposite in English. Therefore, another error has occurred in GMT, which was the phrase "low-priced housing". Here, GMT has not considered the target language's linguistic rule, so, it was translated as "ارزان مسکن" instead of "مسکن ارزان".

Case 3: A household satisfaction index (HSI) model based on…

GNMT: یک مدل شاخص رضایت خانگی (HSI) بر اساس...

GMT: یک شاخص رضایت خانواده (HSI) مدل بر اساس...

Analysis: Again, in this case, there was word-disordering error in placement of the word "model" that was placed incorrectly by GMT. In addition, a collocation error has been made through GNMT for translating phrase "household satisfaction" although GMT translated this collocation accurately.

3. Results and Discussion

As previously mentioned, the materials applied in this paper were selected from both academic and non-academic contexts in order to consider two major

text types. These collected data (both input and output) included about 1350 words and among which the percentages of each error has been provided. Moreover, a new error has been discussed, which was categorized as "other errors" which described words that were not translated or those which were mistranslated. These error's percentages were showed in the following table.

Table1. *Frequency Percentage of Errors*

Error Type	GMNT (%)	GMT (%)
Word disordering	63%	67%
Verb-mismatching	3%	4%
Wrong collocation	20%	13%
Wrong preposition	2%	3%
Active and passive	2%	0%
Tense error	0%	0%
Mistranslated*	3%	6%
Not translated	2%	13%

*Mistranslated error has been considered linguistically and not semantically text. Therefore, the meaning of TT (target text) was not considered.

The above table indicated the frequency percentage of errors mostly in linguistic manner rather semantic aspects. As shown in table, GNMT made nearly the same mistakes as GMT did in linguistic aspects.

The semantic adequacies were not considered in this table due to the non-existence of certain criterion. However, according to analyzed data, GNMT transferred the ST (source text) semantics more efficiently; thus, the TT was more fluent than that of GMT. This was even more obvious when translation moved from Persian to English.

4. Conclusion

According to what has been analyzed, the result of the research revealed that GMNT, in comparison to GMT, was more successful in semantic aspects, especially in Persian-to-English translation. Moreover, based on the percentage of the error reduction it was concluded that GNMT system was beneficent for Persian translation in Google Translate.

However, both approaches still need modifications in order to reduce linguistic errors. As the last word, GNMT has been enhanced significantly in comparison to GMT although users still cannot rely much on the machine translation, and if they do, careful modifications need to be applied.

References

Baker, M., & Saladanha, G. (2009). *Routledge Encyclopedia of Translation Studies*: Routledge.

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*: University of Ottawa Press.

Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *CoRR, abs/1603.06147*.

Costa-Jussà, M. R., & Fonollosa, J. A. R. (2016). Character-based neural machine translation. *CoRR, abs/1603.00810*.

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). *Multi-task learning for multiple language translation*. Paper presented at The 53rd Annual Meeting of the Association for Computational Linguistics.

Keshavarz, M. H. (1999). *Brief view on Google Translate machine*. Tehran: Rahnama Publication.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). *Multi-task sequence to sequence learning*. Paper presented at the International Conference on Learning Representations.

Pilevar, M. T., & Faili, H. (2010). *Persian SMT: A first attempt to English-Persian Statistical Machine translation*. Paper presented at the JADT 2010: 10th International Conference on Statistical Analysis of Textual Data Tehran, Iran.

Pilevar, M. T., & Faili, H. (2010). *PersianSMT: A first attempt to English-Persian Statistical Machine translation*. Paper presented at the JADT 2010: 10th International Conference on Statistical Analysis of Textual Data Tehran, Iran.

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2015). *Sequence level training with recurrent neural networks*. Paper presented at the International Conference on Learning Representations.

Sébastien, J., Kyunghyun, C., Memisevic, R., & Bengio, Y. (2015). *On using very large target vocabulary for neural machine translation*. Paper presented at The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units*. Paper presented at The 54th Annual Meeting of the Association for Computational Linguistics.

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). *Minimum risk training for neural machine translation*. Paper presented at The 54th Annual Meeting of the Association for Computational Linguistics.

Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). *Coverage-based neural machine translation.* Paper presented at The 54th Annual Meeting of the Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR, abs/1609.08144*.