Complementary Training Corpora for Arabic Neural Machine Translation

Reem Alfuraih
Princess Nourah bint Abdulrahman University

Abstract

This paper presents a new large cross-lingual resource to be trained to maximize the fluency and the adequacy of Arabic neural machine translation using the neural encoder decoder framework. It suggests the use of the Undergraduate Learner Translator Corpus (ULTC) which contains many cross-lingual complementary subcorpora by learners and professional translators. It assumes that the provided corpus will be valuable in terms of training data and linguistically informed Neural Machine Translation (NMT) models.

Introduction

NMT has started to dominate machine translation research in the last few years. The success of NMT "requires a large parallel corpus to be effective, and is known to fail when the training data is not big enough" (Koehn & Knowles, 2017). Due to the lack of availability of large parallel corpora in many languages, researchers have proposed to use triangulation, pivoting or monolingual corpora. There are a few studies on Arabic NMT (Almahairi et al., 2016). The current situation motivates contributions from cross-disciplinary areas to the advancement of Arabic machine translation.

The Training Corpus

ULTC is an available ongoing error-tagged sentence-aligned parallel corpus of English, Arabic, French and Chinese, with Arabic as its main language. The corpus is unique in terms of combining many complementary subcorpora of cross-lingual data. The preprocessing database is composed of about 7000 files, containing at least 50 million words. The corpus presents a diversity of cross-lingual training data which consist of two or more translations of the same text (e.g. draft and final translations by the same translator, text in different languages, learner and professional translations or multiple translations of the same text by different learners). Multimodal data in which subtitled videos aligned with parallel texts. keystroke data about the translation process and interpreters' recordings are also time-aligned with their transcripts.

Contact

Reem Alfuraih
Rfalfuraih@pnu.edu.sa
Reem.Alfuraih@gmail.com

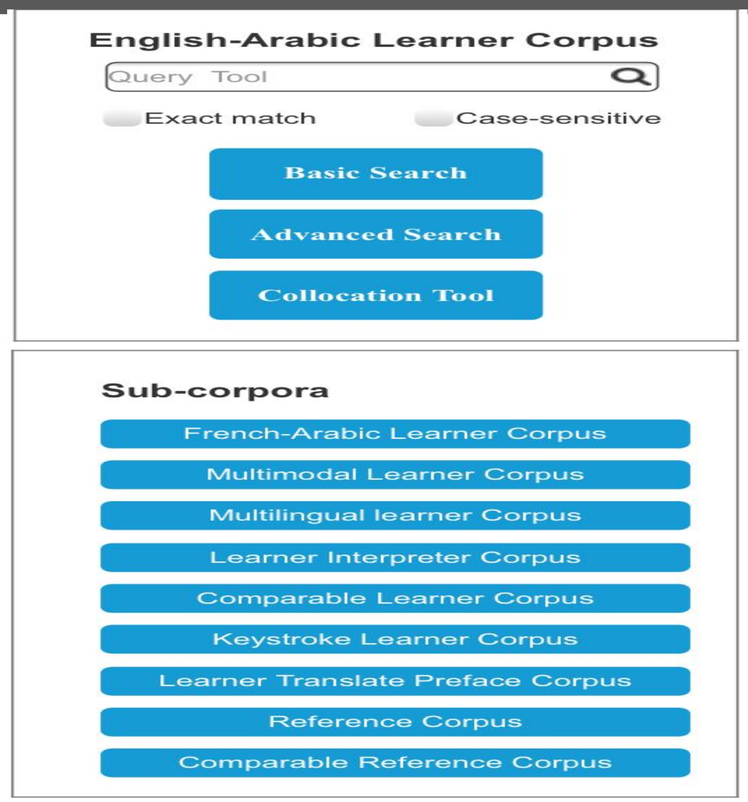

Figure 1: A screenshot of the ULTC interface (beta version)

The Proposed System

The proposed system hypothesizes that for each segment in the source language, the multiple parameter sharing system is trained with many versions of the same segment (i.e. different translations by different translators, draft and edited translations by the same learner, learner and professional translations, multimodal and keystroke aligned data…etc.) The aligned segments can be exploited in an extended context. ULTC has been tagged according to a prepared error taxonomy of errors in the translation from and into Arabic which can have a positive impact on the design and evaluation of NMT systems.

Figure 2: Alternate translations by the same learner translator of the same source text

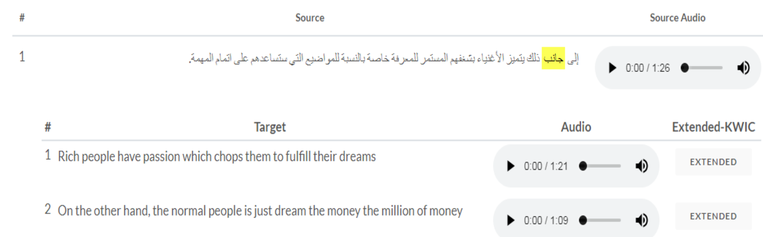

Figure 3: Alternate translations by the different learner interpreters/translators of the same source text

Conclusions

This paper introduces a novel NMT system architecture by compiling ULTC as a representative parallel resource of Arabic and other languages. Specifically, we examine the case when multiple targets are aligned for the same source from the same or different languages. The aligned segments are supported with contextual segments, modalities and annotation. The future work will consider generating multiple source sentences by translating each target segment back.

References

1. Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016). First Result on Arabic Neural Machine. Translation. arXiv preprint arXiv:1606.02680.
2. Philipp Koehn and Rebecca Knowles. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation.

Juan Pino
juancarabina@fb.com
1 Hacker Way, Menlo Park CA 94025
+1 650 785 1527

San Francisco, July 14, 2019

To Whom It May Concern:

This is to certify that Reem Al-Furaih's poster entitled "Complementary Training Corpora for Arabic Neural Machine Translation" was displayed at the first edition of the West Coast NLP summit at Facebook's Headquarter in Menlo Park on September 28, 2018.

Juan Pino