Corpus-Based Neural Machine Translation: Empirical Study Multilingual Corpus for Machine Translation of Opaque Idioms - Cloud AutoML Platform

Authors : Khadija Refouh

Abstract : Culture bound-expressions have been a bottleneck for Natural Language Processing (NLP) and comprehension, especially in the case of machine translation (MT). In the last decade, the field of machine translation has greatly advanced. Neural machine translation NMT has recently achieved considerable development in the quality of translation that outperformed previous traditional translation systems in many language pairs. Neural machine translation NMT is an Artificial Intelligence AI and deep neural networks applied to language processing. Despite this development, there remain some serious challenges that face neural machine translation NMT when translating culture bounded-expressions, especially for low resources language pairs such as Arabic-English and Arabic-French, which is not the case with well-established language pairs such as English-French. Machine translation of opaque idioms from English into French are likely to be more accurate than translating them from English into Arabic. For example, Google Translate Application translated the sentence "What a bad weather! It runs cats and dogs." to "الطقس سيئ جدا !إنها تمطر قطط و كلاب" into the target language Arabic which is an inaccurate literal translation. The translation of the same sentence into the target language French was "Quel mauvais temps! Il pleut des cordes." where Google Translate Application used the accurate French corresponding idioms. This paper aims to perform NMT experiments towards better translation of opaque idioms using high quality clean multilingual corpus. This Corpus will be collected analytically from human generated idiom translation. AutoML translation, a Google Neural Machine Translation Platform, is used as a custom translation model to improve the translation of opaque idioms. The automatic evaluation of the custom model will be compared to the Google NMT using Bilingual Evaluation Understudy Score BLEU. BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Human evaluation is integrated to test the reliability of the Blue Score. The researcher will examine syntactical, lexical, and semantic features using Halliday's functional theory.

Keywords : multilingual corpora, natural language processing (NLP), neural machine translation (NMT), opaque idioms

Conference Title : ICCLMT 2022 : International Conference on Computational Linguistics and Machine Translation

Conference Location : Istanbul, Türkiye

Conference Dates : May 05-06, 2022