语料库驱动的文学性文本与非文学性文本的机器翻译表现研究

A Corpus-driven of Machine Translation Performance on Literary and Non-literary Text

Abstract:In order to investigate the current development of machine translation and assess whether it is capable of translating literary works at the present time, the author conducted a comparative study in this research to test the performance of machine translation on literary and non-literary texts respectively, aiming to detect whether there is a significant difference between machine translation in dealing with the two types of texts. Overall, the performance of machine translation on non-literary texts is better than that of machine translation on literary texts.

Introduction:The development of machine translation has gone through various stages from “rule-based machine translation”, “instance-based machine translation” to “neural machine translation (NMT)”. The translation theories proposed by Chinese translators and the application of corpus in the translation of literary texts are introduced later.

Materials＆Methods:The original text is excerpted from Eileen Chang's novel Love in a Fallen City and Bain China Shopper Report 2019, which represent literary and non-literary texts, respectively. Both are written in Chinese and prepared for translation into English using machine translation.

Results:The three linguistic metrics type token ratio, content word density, and average sentence length were calculated in a python program, and then the calculated data were put into SPSS 20.0 for Pearson correlation analysis.

Discussion:There are still some limitations. The first limitation is to the size of the corpus. In this study, Love in a Fallen City and China Shopping Report (in Chinese) were chosen as literary and non-literary texts, respectively, so the machine translation could only go from Chinese to English. It is expected that in future studies, more classical literary works such as novels, essays and poems, as well as more valuable non-literary texts such as business literature, official reports and public speeches will be categorized and included in the study in order to expand the existing research corpus. In addition, the language of the original texts should be more diverse and not limited to Chinese.