Controlled Document Authoring in a Machine Translation Age

Rei Miyata

First published in 2021

ISBN: 978-0-367-50019-1 (hbk)
ISBN: 978-1-003-04852-7 (ebk)

Preface

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Preface

Every change in the paradigm of machine translation (MT) architectures has lifted the potential of MT technologies, evoking expectations, or dreams, of ‘human parity’ in general-purpose MT. Recent years have witnessed a significant improvement in MT performance due to the advent of neural MT, which makes use of a huge volume of text data to train a deep neural network model to generate target text from source text in a single, unified process. An increasing number of companies, governments and individuals have started to use MT tools to meet their various needs, many of which might not have emerged without the development of MT. Technologies have been expanding the sphere of translation market more than ever before.

However, greatly improved quality of MT—specifically surface fluency of neural MT output—sometimes induces in users a false perception of its reliability. Indeed, fatal misuse of MT has been seen, for example, in the multilingual dissemination of disaster information in Japan. This is attributable to the users’ literacy in MT, not the technologies themselves. MT users, including clients, translation companies and translators, do not necessarily have sufficient knowledge of what MT can and cannot do. Furthermore, MT developers and researchers may not be able to explain what the essential difference between MT and human translation is. More broadly, there is a lack of consensus amongst the various actors on what translation is in the first place. In these circumstances, facing the superficial resemblance of MT output and human translation, we are now in the position to reconsider the proper place of machine and human in translation.

With this fundamental motivation in the background, this book aims to bridge two gaps. The first one is the gap between *document* and *language*. In a nutshell, human translators deal with documents as basic units of translation, while current MT systems chiefly process decontextualised sentences or expressions from language A to language B. Even when human translators handle an individual sentence, phrase or word, this is premised on the existence of the document of which it forms part, and they refer to the document explicitly or inexplicitly for their decision-making at each step of translation. On the other hand, although context-aware MT has been much researched recently, commercially and publicly available MT systems are still sentence-based and do not explicitly take into account the document properties in their translation models.

Here, to address this problem, we employ the notion of controlled language (CL) and extend it to a document level. CLs are artificial languages that are developed by restricting the lexicon, grammar and style of a given natural language, and can be used to reduce the ambiguities and complexity of source text, eventually leading to the enhanced performance of MT. As previous CLs are also sentence-based, we incorporate document-level text properties into the CL rules, which opens a way to contextual MT without directly modifying the MT engines.

The second gap is between *authoring* and *translation*. In this global age, various kinds of information are to be distributed in multiple languages. MT tools have boosted the speed of document multilingualisation, but the output is not always of usable quality and thus needs to be revised by human workers, namely, post-editors. If the authoring process is optimised to the subsequent translation process, we can envisage an enhanced overall productivity. Again, CL is one of the viable devices for bridging the gap between authoring and translation. We pursue both human readability of source texts and their machine translatability by defining human- and machine-oriented CL rules.

In summary, this book proposes to establish *controlled document authoring* that enables contextual MT. To implement controlled document authoring, we orchestrated a wide variety of frameworks and methods covering document formalisation, CL, technical writing, and terminology management. From the practical viewpoint, controlled document authoring is difficult to deploy by non-professional writers. Therefore, we designed, implemented and evaluated an integrated authoring support system to help them write well-structured source documents that are not only easy to read but also easily translated by MT systems.

This book is based on my PhD thesis submitted in 2017, when statistical and rule-based MT systems were widely used in practical situations. While neural MT has become a dominant paradigm in the past few years, previous types of MT are still used in companies and governments, and more importantly, the fundamental problem of the sentence-based MT architecture is yet to be solved. Although some of the technologies used in this work are unavoidably outdated, the idea and frameworks of controlled document authoring and contextual MT continue to be valid in a rapidly changing age of translation technologies.