國立臺灣大學文學院翻譯碩士學位學程

碩士論文

Graduate Program in Translation and Interpretation
College of Liberal Arts
National Taiwan University

Master Thesis

神經機器翻譯於時尚網站在地化之應用

Applying Neural Machine Translation to Fashion Website Localization

廖亭雲
Ting-Yun Liao

指導教授：高照明 博士

Advisor: Zhao-Ming Gao, Ph.D.

中華民國 108 年 7 月

July 2019

doi:10.6342/NTU201903661

Acknowledgment

I would like to give heartful thanks to Professor Zhao-Ming Gao who provided helpful comments and suggestions throughout the writing of this paper. Without his encouragement and guidance, this research would not have materialized. Special thanks also go to Professor Shu-Kai Hsieh, Professor Chao-Lin Liu and Professor Yvonne Tsai whose opinions and information have helped me very much in completion of this study.

I am particularly grateful for the technical support provided by Research Assistant Hou-Wei Lin. My grateful thanks are extended to Anita Fan and Estella Yang who participated in the experiment and offered valuable feedback. Assistance given by Associate Manager Vicky Li was also greatly appreciated.

Finally, I must express my very profound gratitude to my parents and to my friends for their unfailing support and continuous encouragement throughout my years of study.

doi:10.6342/NTU201903661

Abstract

As the number of internet users grows around the world, the demand for website localization also increases. To meet this ever-growing need, a trend has inevitably formed where automatic translation is being integrated into the workflow of localization. In recent years, the development of the internet has provided easy access to various corpora, which advances the technology of machine translation and, in turn, realizes the application of customized neural machine translation (NMT). Nevertheless, previous studies on customized NMT usually center on improving the technology itself. Thus, this research focuses on applying a customized NMT to website localization and validating its performance and effects with automatic evaluation and an experiment involving human translators. To build customized NMT, the bilingual text from shopping websites including H&M, ZARA and Burberry is compiled into a parallel corpus, which is divided into two separate corpora which train and evaluate the customized NMT. In addition to automatic evaluation, this study also carries out an experiment aiming to compare the effectiveness of the customized NMT and that of general MT which was modeled on real localization projects which involved professional translators,. The automatic evaluation result indicates that the NMT built in this research performs well. Moreover, based on the feedback from translators participating in this study, the customized NMT does exhibit positive effects on increasing translation efficiency. Due to the fact that the corpus used in this research is limited to a specific domain, the results might not be applicable to other localization fields, and further research is needed to investigate the effectiveness of customized NMT in other domains.

Keywords: neural machine translation, localization, computer-aided translation, parallel corpora

doi:10.6342/NTU201903661

中文摘要

隨著網路使用人口成長，網站在地化的需求也持續增加，而為了以更快速度提供品質穩定的翻譯成果，將自動翻譯融入在地化專案原有的工作流程已是必然趨勢。另一方面，近年來由於網路發展迅速讓語料庫取得更加容易，也促進了機器翻譯技術的發展，實現自訂神經機器翻譯模型的應用。然而目前自訂神經機器翻譯的相關研究多半仍集中在改善技術層面，因此本研究將會將重心放在將自訂神經機器翻譯系統應用於在地化專案，並且採用自動化方法與專業譯者參與實驗的方式評估其效能與效果。在建立自訂神經機器翻譯系統方面，本研究從風格類似的三個快時尚品牌網站中，分別擷取出繁體中文與英文的產品介紹文字檔，彙整成雙語平行語料庫，再將母語料庫分為訓練與驗證兩個語料庫，用於訓練與檢驗自訂神經機器翻譯系統。而除了採用自動化方法衡量機器翻譯的效能，本研究也招募專業譯者參與模擬實際在地化專案的實驗，比較自訂機器翻譯與一般機器翻譯的效果。研究結果顯示，本研究建立的自訂神經機器翻譯經過自動化方法評估可達到良好效能，而在專業譯者的使用者經驗方面，自訂神經機器翻譯確實有一定的效果，能協助提升譯者的工作效率。儘管本研究指出將神經機器翻譯應用於在地化專案的可行性，由於研究中選擇的語料僅侷限在快時尚購物網站，研究結果可能無法在其他在地化領域上成立，而適用於其他翻譯領域的自訂神經機器翻譯系統則需要將來進一步的研究。

關鍵字：神經機器翻譯、在地化、電腦輔助翻譯、平行語料庫

doi:10.6342/NTU201903661

Table of Contents

doi:10.6342/NTU201903661

doi:10.6342/NTU201903661

List of Figures

doi:10.6342/NTU201903661

List of Tables

doi:10.6342/NTU201903661

Chapter One: Introduction

Thanks to the internet, we live in a world in pursuit of efficiency.We not only ask for everything to be done faster but also better. Websites of fast fashion brands provide the best examples of this. To be competitive, in addition to providing new products every week, these companies have to update their websites to keep up-to-date. Furthermore, in order to tap into foreign markets, this digital content needs to be translated and sometimes adapted (i.e, localized) to service their consumers around the world (Schäler, 2009). In fact, any company which aims to broaden its market around the globe has a need for localization services, which has thus driven the rapid expansion of the localization industry.

Such great demand for localization since the late 20th century has led to the wide application of computer-aided translation (CAT) tools, which are computer programs designed to "facilitating the speed and consistency of human translators" (Garcia, 2014). As localization projects involve great volumes of text and short deadlines, and the content could be highly repetitive or contain coded tags, it is necessary for translators to be equipped with the ability to use CAT tools. In a narrow sense, CAT tools do not include electronics resources such as online dictionaries and corpora, nor do machine translation systems, since, technically speaking, they generate translation rather than assist human translators. Thus, more specifically, the two key functions of CAT tools are to generate translation memories (TMs) and term bases (TBs). While the former can reference previously translated sentences and suggest a translation match with the same or similar sentence, the latter can automatically replace terminology in the source language with corresponding terms in the target language. With CAT tools, human translators do not need to process repeated sentences or terms within

doi:10.6342/NTU201903661

a text, which thus enhances not only work efficiency but also consistency. In theory, the amount and specificity of the data contained in a translation memory or term base determines its usability. However, in reality, it takes time and effort to build and manage domain-specific TMs and TBs, so when translators start a new project from scratch, TM and TB tools provide limited assistance.

Fortunately, the rise of corpus linguistics and machine translation systems has not only revolutionized the usage of CAT tools, but has also helped translators in more efficient ways. Whereas corpora in the past were mainly used in linguistics studies, they are now coupled with concordancing software to serve as translation tools as well (Gao and Chiou, 2017). For instance, parallel corpora consist of original texts and translated texts. In general, this type of corpora can function as domain-specific bilingual dictionaries, and for translators, if the bilingual texts are aligned at the sentence level, the corpora can also be utilized as a translation memory (García, 2014).

Thanks to the internet, nowadays parallel bilingual texts are more accessible than ever, which lays the foundation for the development of statistical machine translation (SMT) systems. For instance, search engines like Google and Microsoft utilize their massive databases as bilingual corpora to build machine translation services such as Google Translate and Microsoft Bing Translator, which are able to pair languages according to their matching probability. Nevertheless, despite the fact that SMT can help translate most words accurately, especially domain-specific terminology, it does not perform well at the sentence level or above (Gao and Chiou, 2017). In recent years, in pursuit of higher quantity, quality and speed in translation work, more research has been done on newer technology andneural machine

doi:10.6342/NTU201903661

translation (NMT), and large service providers like Google Translate and Microsoft Bing are also incorporating NMT systems into their MT service.

NMT can model an artificial neural network to perform machine translation. In contrast to SMT, NMT "has led to remarkable improvements, particularly in terms of human evaluation," (Klein, Kim, Deng, Senellar & Rush, 2017) and thus has the potential to generate usable automated translation without extensive post-editing. Nevertheless, NMT is still in the early development stage, meaning more empirical studies are needed to test and improve its performance. The latest trend in NMT studies is toward customization; that is, using a collection of parallel bilingual text to build an NMT system for domain-specific translation. In 2018, Google released a new service called "AutoML Translation" which allows users to upload data in the form of matching pairs of sentences in a source and target languages, train a customized translation model with the self-prepared data, and then translate specialized content using the trained model. Many expect this new approach on building NMT systems can achieve what general MT cannot.

In terms of MT's application in the localization industry, as Schäler (2009) points out, the localization process will inevitably evolve into more complicated, standardized and automated activities, and translating strings will become less important. Now with the realization of customized NMT, it is possible for human translators to skip to the review and editing stages while machines tackle translation. At the same time, the development of CAT tools has largely improved the technology of translation memories and term bases, which can sufficiently support translator in post-editing automated translations. However, most of the research on NMT which focus on measuring the performance of the system itself and how to improve it tend to deemphasize the feedback from what would potentially be the ultimate

doi:10.6342/NTU201903661

users of the software: profession translators. The question of how helpful is customized NMT for professional translators working on real localization projects in a specific domain when compared to general MT still needs to be answered.

To answer the above question, this research intends to build a customized NMT system by applying an open-source framework, OpenNMT, and then evaluate its performance. The effectiveness of the NMT system is measured by two approaches: automatic (i.e., not supervised by humans) evaluation by using the BLEU method, and an experiment with professional translators as the participants. First, the text for building an English and Traditional Chinese parallel corpus is collected from the official websites of famous fashion brands including H&M, ZARA and Burberry. Then, the bilingual corpus is divided into two corpora, a training corpus and a validating corpus. The training corpus is required so as to let the NMT learn from its example using the framework provided by OpenNMT so the trained NMT can translate an English text from the validating corpus into Traditional Chinese. The automated translation is tested by BLEU, a method for automatic evaluation of machine translation which references the bilingual texts from the validating corpus.

In parallel, a new translation project is created in the CAT tool, Trados 2017, with the English text from the validating corpus serving as the source text, and the bilingual text from the training corpus as the translation memory. The pre-translation function and analysis in Trados can provide some insight into the automatic translation generated from the customized NMT. In addition, to compare the major differences between the performance of general MT (through Google Translate) and the customized NMT, we conduct an experiment which models real translation projects and involves professional translators . Lastly, we collect feedback from the translators on their user experience in the two types of MT through

doi:10.6342/NTU201903661

questionnaires. The whole process of the experiment is also recorded for analysis. It is expected that this research can provide some empirical evidence to support the effectiveness of the customized NMT, yet at the same time also take the first impressions of professional translators into consideration in order to measure the performance of the customized NMT.

doi:10.6342/NTU201903661

Chapter Two: Literature Review

Despite the fact that this research is based on an experiment on a customized NMT, the application of CAT and corpora is also an essential part of the methodology. In Hutchins' research (2005), he outlines a general introduction to machine translation, and he points out that the development of MT is a process of gradually moving from "machine-aided human translation" (Figure 2.1) to "human-aided machine translation" (Figure 2.2). The workflow of the former has human translators equipped with various computer-based tools, while the latter involves three tasks in which MT plays a primary role: machine translation, input, and post-editing. From building a domain-specific MT system, to processing the input to be compatible with the given MT system, and to post-editing the automated translation manually, both corpora and CAT tools are largely relied on in the machine-centered process. From a broader view, corpora and CAT also play a role in the process of evolving from machine-aided translation to human-aided machine translation. As such discussing MT without mentioning its relationship with corpora and CAT would be remiss. In the following sections, previous studies on localization, CAT, Corpora, MT, and the interrelation between them will be reviewed.

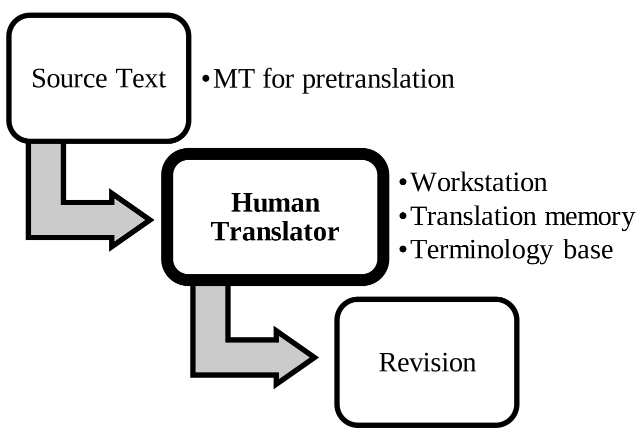


Figure 0.1 Machine-aided human translation

doi:10.6342/NTU201903661

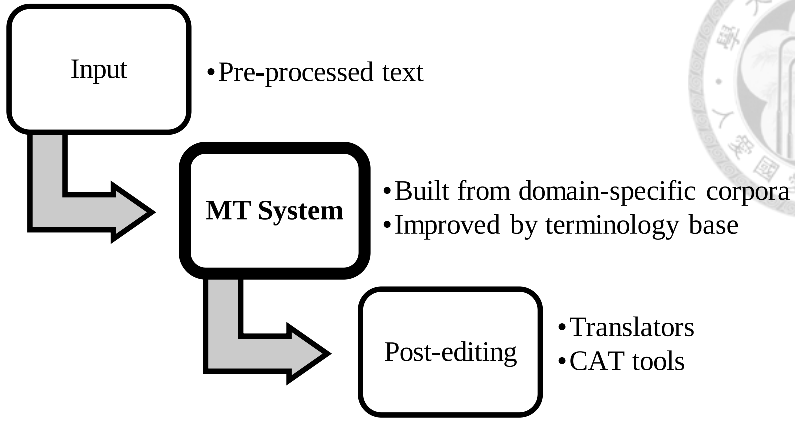


Figure 0.2 Human-aided machine translation

2.1 Localization

According to Schäler (2009), localization is defined as "the linguistic and cultural adaptation of digital content to the requirements and locale of a foreign market" (p.157). The definition also extends to "the provision of service and technologies for the management of multilingualism across the digital global information flow" (Schäler, 2009, p.157). Although it is not a novel concept that sellers customize products for consumers speaking different languages and living in different cultures, Schäler (2009) argues the key divide between localization and conventional translation lies in the types of materials being processed. In short, the localization as it is referred to today is a process of managing digital content. Therefore, localization requires specific technologies, skillsets, procedures and standards, which often vary from the translation process for printed materials. In the following sections, we will introduce the development of the localization industry and the standard working procedures during localization.

In the mid-1980s, software providers began to expand their markets internationally. In

doi:10.6342/NTU201903661

order to sell products to foreign customers, these international corporates decided to translate their products into languages used by their target consumers. As Schäler (2009) points out, localization is used as a means to meet commercial demands. In the early phase of localization, the major challenge faced by translators is that the string which needed translation usually contained both plain text and a large amount of program codes. To save the trouble with codes and help translators working on localization projects increase efficiency, large software publishers have been proactive in advancing "internalization." As described by Schäler (2009), internalization is a process of designing (or modifying) software in a way such that the elements in a program which need to be translated and culturally adapted can be isolated in preparation for localization, and therefore the users who ultimately make use of this software can choose their preferred languages in the program. In terms of dealing with translation, these international companies have also invested in "recycling translation," which utilizes translation memory systems to reuse past translation. Schäler (2009, p.158) indicates the application of TM and other CAT tools can be deemed as "a milestone in the history of localization." In the ideal scenario, the work of translators during localization can be simplified to a kind of "translation" assisted by translation memories, while the sellers provide different language versions of their products, and thus generating greater revenue.

At the present stage, there is a set procedure for most localization projects, regardless of the scale, content, or target of a particular project. The various steps in the procedure are analysis, preparation, translation, engineering/testing and project review. In the analysis step, many aspects need to be considered, some of which are closely related to the work of the translators. These aspects are which parts of the product are to be localized, the word count

doi:10.6342/NTU201903661

and rate, and what kind of tools shall be employed. During the preparation phase, a project manager refers to the analysis report to create a new project and provide a localization kit for translators. The localization kit consists of source documents, CAT tools, a style guide, query instructions, and delivery and contact details. During translation, , professional localizers should be able to leverage CAT tools specified by the manager and apply past translations and terminologies to their translation in addition to translating new content. The major difficulty faced by localizers is that, as Schäler (2009, p.159) indicates, "the pressure to produce high-quality translation within short time frames and at low cost is extremely high." The third phase is when the translation undergoes process engineering/testing, in which quality assurance (QA) is conducted. During this stage, the client's reviewers identify translation issues or errors and document them in the QA report. Finally, to prevent repetitive mistakes , the whole localization project is reviewed by in-house reviewers based on the QA report,.

Considering the short life cycle of a localization project, translation inevitably has stricter time limits. Website localization is no exception. At present, localizers already have various CAT tools at their disposal to increase work efficiency. However, to further expedite localization projects , it seems the best possible solution is automation. As Hutchins (2005) points out, companies that provide information online in multiple languages only have three options: using online machine translation, outsourcing to localization agencies, or integrating automatic localization systems. Hutchins' observation is based on not only the nature of localization, but also the relationship between CAT and MT. The following section introduces the basic features of CAT, as well as its development and its convergence with MT.

doi:10.6342/NTU201903661

2.2 Computer-aided Translation

From the 1990s to the present, the functions of CAT tools have undergone a great deal of innovation and integration due to an increasing demand for localization. However, translation memories, which enable translators to store and reuse past translation, have always been the main feature of CAT tools since their commercialization in the mid-1990s (García, 2014; O'Hagan, 2009). A TM essentially consists of text segmentation and alignment tools. A segmentation tool is used to divide texts into shorter units. In general, these texts are segmented into sentences according to punctuation rules or specific syntactic features such as headings and bullet points. This makes it so that an alignment tool can be utilized to pair source texts and target texts segment by segment. In addition to TMs, every CAT tool is equipped with a terminology management system for building term bases or glossaries, which work similarly to a TM, differing only in that they are restricted to the term level (García, 2014).

Scholars (Bowker and Barlow, 2008; García, 2006; Shuttleworth, 2006) have been studying the pros and cons of CAT tools, especially so for TM and its impact on translators. Although TMs to a certain degree can boost the work efficiency of translators, this technology has clear limitations due to its "unsuitability for non-repetitive texts, the inflexibility of only having matches on the sentence level, the difficulty of retrieving contextual information and the time it takes to produce useful TMs" (Shuttleworth, 2006, p. 3). TBs, which fall short in the same way on the term level, have similar restrictions.

To overcome these limitations, researchers have started to integrate the application of CAT into studies on SMT, as the usability of SMT has been significantly improved in recent years. For instance, Pym (2013) suggests that translators should learn how to operate TM and

doi:10.6342/NTU201903661

MT technology in order to adapt to changes in the role of a translator. Gao and Chiou (2017) have demonstrated how to use SMT, corpora and analysis tools to build a translation memory and term base for a new translation project. In fact, CAT and TM can be seen as having been derived from the same root. In the days when the capability of computers was insufficient to support the continued development of MT, the arrival of CAT tools helped provide a stepping stone to fully automatic translation (García, 2014). Stein (2018) also states while the ultimate goal of research on MT is to achieve "fully automatic high-quality translation," research and development of CAT tools are to be promoted as well. Furthermore, some CAT tools have begun to integrate MT to address "units with lower similarities, the automatic transliteration of numbers, dates and other placeable elements" (Stein, 2018, p. 8-9). Now the evolution of CAT has reached its limit, and as Shuttleworth (2006) indicates, since TM and MT "share common problems, for which the solutions lie in the combination of the two technologies" (p. 6), there will be more and more overlaps between the two.

2.3 Corpus-based Studies

In today's highly competitive localization market, translators also seek assistance from electronic resources in addition to CAT tools. Zanettin (2002) and Bowker and Barlow (2008) propose that a corpus – namely, a "collection of electronic texts assembled according to explicit design criteria which usually aim at representing a larger textual population" (Zanettin, 2002) – can be of use to professional translators. Some large reference corpora such as the British National Corpus (BNC) for British English have gained popularity outside of the linguistics circle, but corpora in such scale are difficult to build and not necessarily suitable for more specific uses. Thus, Zanettin (2002) advises translators to customize small

doi:10.6342/NTU201903661

corpora by retrieving and organizing text information on the internet, which will result in better work efficiency and greater accuracy in terminology and phraseology.

Among all of the types of corpora, bilingual parallel corpora, which are "a collection of texts in one language alongside their translations into another language," are the most pervasive corpora in use during translation (Bowker and Barlow, 2008). However, these corpora are more difficult to come by due to the extra effort required to align bilingual texts at the sentence level. Undoubtedly, parallel corpora can provide ample information, but extra tools are needed to truly benefit from them, such as a bilingual concordancer or translation memory. Unlike a translation memory which pairs texts segment by segment, when the texts are fully aligned in a corpus, , a concordancer can help translators "retrieve all examples of a word or phrase (or part of a word) from the corpus" (Bowker and Barlow, 2008). While translation memories are suitable for localization projects containing highly repetitive sentences, parallel corpora coupled with concordancers can be applied to match sub-sentence elements, thus functioning as a complementary tool.

Apart from serving as references, corpora can also be used to extract lexical information if supplemented with corpus processing software (Hutchins, 2005; Zanettin, 2002). The availability of monolingual comparable corpora, which consist of two or more groups of texts in the same language and belonging to the same domain, have made it possible for researchers to extract terminologies and compile term bases for specific domain. For instance, in Gao and Chiou's studies (2017), segmentation tools, part-of-speech taggers and a corpus processing software are used to identify terminologies in English and Chinese news articles.

As Hutchins (2005) states, recent development in technology has allowed researchers to use simpler rules to build MT; which is to say, MT can be trained by the "information

doi:10.6342/NTU201903661

about word cooccurrences derived from text corpora" (Hutchins, 2005, p.11). Moreover, terminology entries required for MT can also be extract from bilingual corpora with the assistance of corpus analysis tools. On the one hand, the wide application of corpora complements what was lacking in traditional CAT tools; on the other hand, their higher availability provides resources needed for MT research.

2.4 Machine Translation

In Hutchins' study (2005), he provides the brief history of how machine translation systems have evolved over the years. The development of MT can be traced back to around the 1950s, when the first MT conference was held and MT was first demonstrated. Since then, the U.S. government and Georgetown University have installed some of the first MT systems and conducted related research projects. Although the results were not as good as expected, this stage laid the foundation for the progress of MT in the 1980s. Since then, the usage of MT systems has widened, especially with regards to early installations of translation software on personal computers and the introduction of translator workstations to the market. As for the online machine translation services with which the general public is more familiar, this major breakthrough in MT did not made until the 21st century. The subsequent development of the emerging technology of NMT has become a hot research topic as of late. To understand why NMT is a topic worthy of study, it is necessary to understand its differences with respect to past MT systems. According to the definition proposed by Ping (2009), based on the architecture of MT systems, there are generally two types of MT which constitute the main areas of research in MT: rule-based MT and corpus-based MT.

doi:10.6342/NTU201903661

2.4.1 Rule-based machine translation

Rule-based MT is a system built on a variety of linguistics rules determined by developers, and can be classified as “direct” and “indirect” in terms of how they develop (Ping, 2009). Introduced before the 1980s, the direct approach is based on a bilingual dictionary and morphological analysis. In other words, this type of MT translates source text by searching the dictionary at the word level, and then reordering the translated words according to the grammar of the target language. As a result, rule-based MT built on direct approach does not accurately analyze the syntax of source texts nor does it identify the relationship between words.

In the 1980s, developers started to adopt the “indirect” approach when designing MT. This type of MT translates texts in a three-stage method through analysis, transferring and synthesizing. Firstly, the MT analyzes the syntactic structure of source sentences and converts them into “intermediary, abstract representations of the meaning of the original” (Ping, 2009). Then, these representations are transferred into representations which indicate the syntactic structure of the target language. Lastly, the system synthesizes the transferred representations and produces translated sentences. In contrast to the direct approach, the indirect approach can perform additional analysis on the source text and identify both sentence structure and meaning, instead of translating word for word.

However, there are still some challenges in the development of rule-based MT. Hutchins (2006) indicates that for this type of MT to work successfully, developers have to work on complicated grammar rules, and it is difficult to design a model which can apply all the grammar rules perfectly. Issues also arise from the dictionary used up by the MT system,

doi:10.6342/NTU201903661

since every dictionary has its own limitations and cannot cover all meanings. Thus, rule-based MT has a comparatively high entry level in terms of development and application.

2.4.2 Corpus-based machine translation

In the 1990s, researchers discovered it is possible to build MT systems by applying bilingual corpora, especially collections of original and translated versions of texts. There are two types of corpus-based MT systems: example-based MT and statistical MT. The concept behind example-based MT is similar to that of translation memory in CAT tools. Based on a bilingual parallel corpus, the MT first matches the source texts with the most similar examples in the corpus, and then aligns the source text and the examples to find the corresponding parts. Lastly the corresponding parts are reordered and assembled to produce target texts. As Ping points out (2009), the main difference between TM and example-based MT is that the former requires human translators to do the task of reordering and assembling, while the latter can complete all the work automatically.

Thanks to Google Translate and Microsoft Bing Translate, statistical machine translation (SMT) has the method which the general public are most familiar, and there have been numerous studies related to the topic. In Hutchins' presentation (2006), he describes SMT as a system based on bilingual corpora which requires "little or no linguistic 'knowledge.'" Essentially, SMT is built on "word co-occurrences in SL and TL texts (of a corpus), relative positions of words within sentences, and length of sentences" (Hutchins, 2006, p.21). Sentences from the bilingual parallel corpus are aligned by statistical rules such as sentence length and relative positions of words. The translation process then involves two models: translation model and language model. The former model chooses the most probable

doi:10.6342/NTU201903661

translation for a source fragment, which can be a word or a phrase, based on the frequencies of word co-occurrences in the aligned bilingual corpora. The language model organizes target fragments in the most probable order to produce translations based off the frequency of bigrams and trigrams in the target language.

Hutchins (2006) describes SMT as a "direct approach" that replaces a fragment in a source language with a fragment in a target language in the most probable sequence. In other words, it can be said that the size of a training corpora plays an essential role in SMT, and this is where the most obvious advantage of SMT lies. As Stein (2018, p.14) indicates, in contrast to rule-based MT, "SMT systems produce better translations in terms of word choice, disambiguation, etc." Any types of word combinations can be translated, as long as they are included in the corpora in a certain number "to be identified statistically." Thus, Stein (2018, p.14) concludes that the idea behind SMT is that "bigger corpora means better results."

However, although studies on SMT have extended from word-based, phrase-based to syntax-based, there are still some issues mostly arising from training corpora which remain unresolved. As mentioned before, the size of the training corpus for SMT matters, and other scholars such as Ping (2009) also point out the success of SMT is basically determined by the training corpus. For instance, if most of the data used for compiling a training corpus comes from the internet, which is the most efficient way to collect a huge amount of bilingual texts, it is difficult to control the quality, resulting in unstable results when using the SMT program. Apart from this, Stein (2018) also argues that when SMT deals with certain language pairs, many problems stem from differences in grammar like "inflection, word order, use of pronouns, number and kind of temporal forms, etc."

doi:10.6342/NTU201903661

In Gao and Chiou's research (2017), the scholars acknowledge that SMT can serve as a supplement to CAT tools when there is insufficient relevant past translation. Although Google Translate has shown in studies its usefulness for translating terminology and proper nouns, Gao and Chiou discovered the SMT system cannot "identify the beginning and ending of a multiword unit" and often provides translation in simplified Chinses based on probability. As a result, human translators usually need to pay extra attention to pre-editing such as simplifying input, in order to improve the productivity of SMT and achieve the goal of "human-aided machine translation."

Stein (2018, p.15) summarizes the recent development of MT with the observation that, "the use of linguistic information and statistical data, has become one of the most researched fields in MT over the last decade." Combining the strengths of different MT systems not only increases the translation quality of MT but also opens up the possibility for research on MT for rare language pairs; regardless, a real breakthrough has yet to be achieved. This hybrid approach leads to shifting the research focus to identifying appropriate language resources to build MT systems, and especially MT for a specific domain, since "it turns out that the automatic translation of specialized domains is more reliable" (Stein, 2018, p.16).

2.4.3 Neural machine translation

According to Luong, Cho and Manning (2016), nowadays there is a considerable demand for machine translation, especially in the fields of "humanity and commerce." However, although the ultimate goal is to realize "fully automatic high-quality MT," at the present stage, only "user- or platform-initiated low-quality translation" or "author-initiated high quality translation" are available. The first category includes translation service

doi:10.6342/NTU201903661

provided by Google Translate or Bing Translator, while the other category requires post-editing from human translators or MT as supplementary tools for translators. As the technology of MT evolves, , the quality of automatic translation from statistical machine translation to neural machine translation has improved, but there are still obstacles that need to be overcome.

As Luong, Cho and Manning (2016) state, NMT was a “fringe research activity” back in 2014, which then became a widely-acknowledged research approach for general MT in 2016. Luong, Cho, and Manning (2016, p. 14) define NMT as “the approach of modeling the entire MT process via one big artificial neural network”. Simply put, NMT is a two-layered neural encoder-decoder architecture; the encoder network can receive an input source sentence and transform it into a series of vectors, each representing an input word, and then based on this series of vectors, the decoder network generates a translated text.

The emerging technology of neural machine translation is reported to perform better than SMT at the sentence level. Research (Kinoshit, Oshio and Mitsuhashi, 2017) has been conducted on comparing the performance of SMT and NMT with large parallel corpora, and the researchers indicate NMT scores higher both in BLEU – an automatic evaluation framework for machine translation (Papineni, Roukos, Ward and Zhu, 2002) – and human evaluation. The advantages of NMT are explained in the documentation of Google’s AutoML Translation service (2019).When rule-based MT was the mainstream approach to process natural language, it required professional programmers to instruct the computer step by step. Now with large parallel corpora available, it is possible to get the machine learn by itself the language rules from examples using a certain framework. This new approach led to the

doi:10.6342/NTU201903661

application of customized NMT. Trained with a domain-specific corpus, customized NMT can achieve what general MT cannot in the translation of a specific domain.

2.5 Evaluation on machine translation

Although human evaluation on MT can cover all aspects of translation quality, it requires time and effort. To better examine the performance of MT, Papineni, Roukos, Ward and Zhu propose an automatic method called BLEU (bilingual evaluation understudy) to assist developers. Simply put, this method consists of two components: a numerical "translation closeness" metric, and a corpus of good quality human reference translations (Papineni, Roukos, Ward, Henderson and Reeder, 2002). BLEU is able to calculate the similarity of automated translation and reference translation by finding how many matching n-grams exist between the two texts at the sentence level. Apart from n-grams, sentence length is also put into consideration, since amore similar sentence length is an indicator of a better translation. According to the study (Papineni, et al., 2002), BLEU shows the ability to differentiate not only good and bad automated translations, but also ideal and poor human translation. The results of the automatic evaluation are statistically similar to the judgement of human participants in the study.

The score provided by BLEU is between 0 to 1. A score of 1 occurs when the translation is perfectly identical to the reference text. For the sake of convenience, a BLEU score is sometimes not presented as decimals but percentages. It is noteworthy that there is no strict standard to interpret a BLUE score. For instance, the documentation “Evaluating models” published by Google’s AutoML Translation (2019) suggests using the following table (Table

doi:10.6342/NTU201903661

2.1) as a rough guideline, while Lavie (2011) simply proposes that a BLEU score over 30 should be deemed as “understandable translation” and a score over 50 as “good and fluent translation.”

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Table 0.1 Interpretation of BLEU score

However, scholars have pointed out that although BLEU has its advantages in that it saves time and effort, its results do not necessarily correspond to the evaluation of professional translators (Kinoshit and Mitsuhashi, 2017), and that the differences of language pairs might also cause larger variations in BLEU scores (Junczys-Dowmunt, et al., 2016). For instance, all words areweighted equally in BLEU, meaning that it cannot identify the extent of a mistranslation. Also, the automated evaluation is based on matching “exact words,” so synonyms are considered as translation errors. BLEU is a method designed for MT developers to effectively measure and improve their systems, and does not necessarily meet the needs of language analysts. As Lavie (2011, p.40) points out, BLEU scores are not “easily

doi:10.6342/NTU201903661

interpretable by most translation professionals." Lavie (2011, p.50) also argues that compared to other automated metrics such as the Metric for Evaluation of Translation with Explicit Ordering (METEOR), BLEU does not show better "correlation with human judgement."

At this point, most of the studies on machine translation and its automatic evaluation have centered around how to improve the technology itself and the related tools. Feedback from users such as professional translators on machine translation is rarely included in the discussion. Thus, in addition to following the methods of past research on building and evaluating a customized NMT system, this research intends to examine this technology from the perspective of professional translators. By combining the two approaches, it is expected that this paper will be able to investigate into the performance of the customized NMT in a more comprehensive way, yet at the same time also answer the question of whether or not the NMT is useful for professional translators.

doi:10.6342/NTU201903661

Chapter Three: Methodology

According to previous studies, although CAT tools are prevalent in the localization industry, their development has hit a bottleneck. As a result, adding machine translation to a translator's workflow was bound to happen. However, most of the research related to MT focuses on how to improve the technology itself to enhance translation efficiency, yet rarely takes the needs of professional translators into consideration. Thus, this research aims to examine how helpful the latest MT technology is for profession translators who possess the ability to operate CAT tools. For this purpose, it is necessary to answer the following questions: How does the proposed customized NMT perform? How do professional translators perceive the effects of the proposed NMT? What are the empirical effects of the proposed NMT? Lastly, what is the current ideal tool combination for professional translators?

In order to answer these questions, this research combines a collection of tools to conduct an experiment (see Figure 3.1). The first step is to compile a corpus in which English and Traditional Chinese parallel texts are collected from the official websites of the popular fashion brands H&M, ZARA and Burberry. Then, the bilingual corpus is split into two corpora, a training corpus and a validating corpus. The training corpus is applied to build an NMT system from scratch by adopting the framework of OpenNMT. The trained NMT is then used to translate the English text from the validating corpus into Traditional Chinese. The automated translation is tested by BLEU – a method for automatic evaluation of machine translation – which references the bilingual texts from the validating corpus. In addition, the English text from the training corpus is further analyzed with a corpus analysis tool to identify some of the factors which might affect the performance of the customized NMT.

doi:10.6342/NTU201903661

At the same time, a new translation project is created in the CAT tool, Trados 2017, with English text from the validating corpus serving as the source text, and the bilingual text from the training corpus to serve as a translation memory. The pre-translation function and analysis in Trados can provide some insight into the automatic translation generated from the customized NMT. Lastly, in order to compare the usefulness of the most common NMT, Google Translate and the customized NMT, an experiment modeling real translation projects and involving professional translators is conducted. The two groups of source text for the two separate translation projects are selected from the English text from the validating corpus. With the support of the translation memory built from bilingual training corpus and machine translation from NMT or Google Translate, the participating translators have to complete the two projects and answer a feedback questionnaire. It is expected that the result of this experiment can reveal how and why these tools are useful for professional translators.

doi:10.6342/NTU201903661

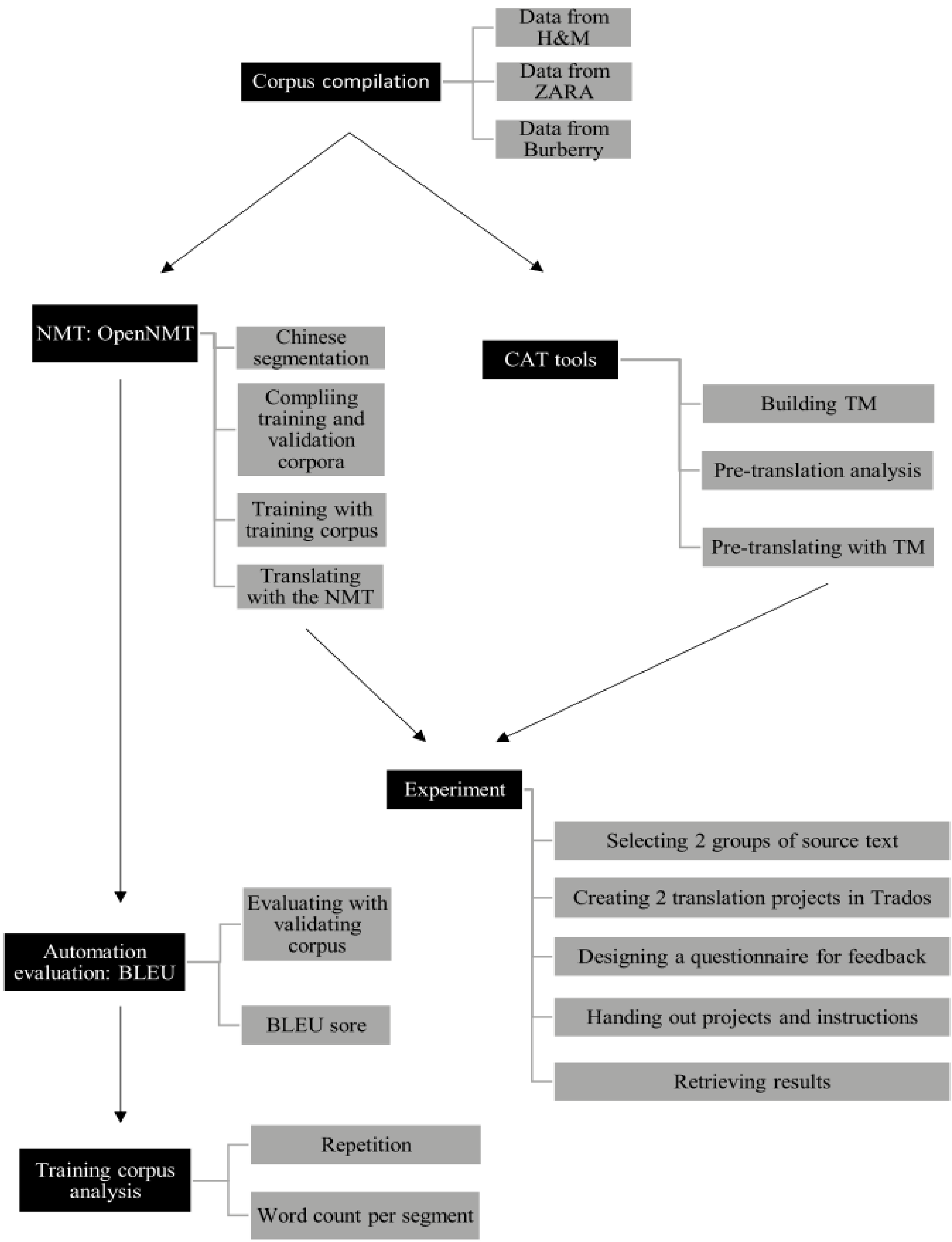


Figure 0.1 Workflow Chart

doi:10.6342/NTU201903661

3.1. Corpus compilation

For the purposes of modeling a real localization project and efficiently collecting sufficient texts to train an NMT system, international fashion brands such as ZARA, H&M and Burberry, which have multiple language versions for their websites and update their webpages frequently, are the ideal sources of data. Since only plain text files are needed to build corpora and train NMT, as well as to compile TM and TB, web crawler tools are used to collect English and Traditional Chinese texts from the shopping websites. The crawler tools can find the English texts' corresponding Traditional Chinese translation by identifying the same elements across webpages in different languages, ensuring that the bilingual texts are extracted from the websites in a fully aligned format. It took about six months (from January 2018 to August 2018) to gather sufficient bilingual text from the three websites.

3.1.1 Data from H&M websites

H&M websites are structured with two layers: product lists and product descriptions. By connecting to the webpage URLs of product lists and analyzing the HTML codes returned by the URLs, every URL from each product, which is in the second layer, can be collected. Then, by connecting to the URL of each product, English and Traditional Chinese product descriptions can be extracted from the websites to compile the corpora for training a machine translation system.

First, the program "get_urls.py[1]" is used to acquire the first layer of URLs, which are four product lists containing different types of product URLs: WOMEN, MEN, DIVIDED,

[1] I would like to thank Research Assistant, Hou-Wei Lin, for providing technical support for operating crawler tools, configuring the proposed NMT and running the BLEU test.

doi:10.6342/NTU201903661

and KIDS. As a result, the first layer of URLs is “http://www2.hm.com/en_asia3/XX/shop-by-product/view-all.html,” in which “XX” can be replaced by “ladies”, “men”, “divided”, or “kids” (see Figure 3.2). The same rule applies to the Traditional Chinese website, but the first layer of URLs needs to be changed into “http://www2.hm.com/zh_asia3/XX/shop-by-product/view-all.html” (see Figure 3.3).

Figure 0.2 The first layer of H&M’s English website

doi:10.6342/NTU201903661

Figure 0.3 The first layer of H&M's Traditional Chinese website

3.1.2 Data from ZARA websites

The websites of ZARA are more complicatedly constructed, so another software, HTTrack WEBSITE COPIER, is used to extract bilingual texts. The software allows users to download selected webpages, and in this case, only URLs which start with "-p" followed by a two-digit number and end with a number are collected. In other words, only webpages containing product information are chosen for downloading. The results are shown in Figure 3.4.

doi:10.6342/NTU201903661

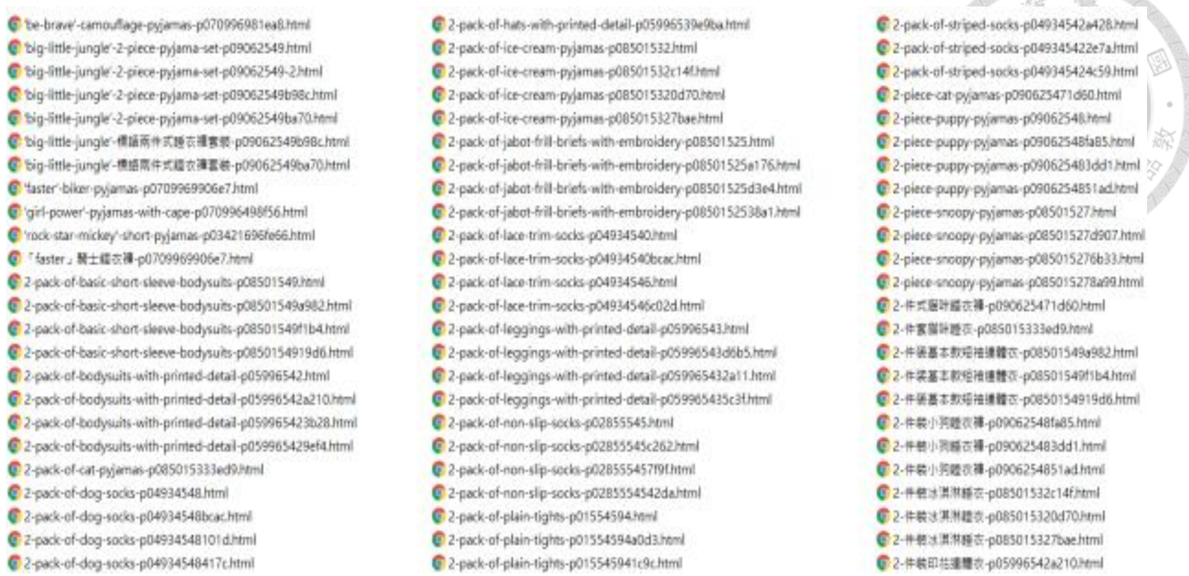


Figure 0.4 Webpages collected from ZARA's website

3.1.3 Data from Burberry websites

After collecting data from the websites mentioned above, it was estimated that more bilingual texts would be needed to build a more effective NMT system. As a result, Burberry is chosen to serve as another data source since it is one of the few international brands which provides adequate traditional Chinese content on its official website. The same methods used in extracting texts from H&M and ZARA websites are both applied to collect 10000 more lines of bilingual texts from the Burberry websites. It should be noted the English text collected from Burberry's website is written in British English, which might result in different spellings for the same word in American English. Moreover, this luxury brand has a different market position from ZARA and H&M in terms of pricing and product features, which could result in varied word choice and text styles in the corpus.

3.1.4 Text pre-processing

doi:10.6342/NTU201903661

Each webpage consists of three to six lines of data, in which there is a line for the product name, one or more lines for product descriptions and a line for the material of the product. In the three websites, the same product in different colors results in repetitive data, so the program “parse_urls.py” is used to delete repeated product pages. Lastly, the text data in each selected webpage is extracted through the program “read_sorted_urls.py,” which can provide plain text files for the usage of this research. In total, this bilingual corpus contains 82815 lines, in which there are 723147 English words and 119251 Traditional Chinese words, both including numbers and punctuations.

3.2. Building customized NMT: OpenNMT

This research adopts the open-source neural machine translation framework, OpenNMT, which provides a complete library that allows users to train and deploy a new NMT model simply by importing the source and target files (OpenNMT, 2017). Google’s AutoML Translation was not available yet since this research was conducted in 2018. Currently, OpenNMT is a more accessible open-source tool. The NMT adopted here is a sequence-to-sequence model, and the mechanism behind this is to “encode the source sequence using some type of encoder and then to output the target sequence with a decoder” (Levin, Dhanuka and Khalilov, 2017). The aim is to train the encoder and decoder together so that they are able to translate source sentences into target sentences. In this research, the encoder and the decoder are two layers of long short-term memory (LSTM) recurrent neural nets. Although it is expected that adding more layers of LSTM can improve the performance of the customized NMT, the preliminary test result shows the opposite, so a two-layer LSTM setting

doi:10.6342/NTU201903661

is applied. It should also be noted that when building the customized NMT, default configurations and parameters are used.

According to the tutorial on GitHub (2018), using OpenNMT requires three steps: pre-processing the data, training the model, and translating sentences. Beforehand, it is necessary to ensure the training files are presented as one sentence per line and every word is separated by a space(i.e., sentence and word segmentation). As the English texts and Traditional Chinese texts are already aligned and paired segment by segment when extracting bilingual data, , there is no need for sentence segmentation. Nevertheless, while English words are naturally separated by a space, Chinese words do not have an equivalent indicator of separation which a computer can recognize. As a result, a word segmentation system for Chinese, NAER Segmentor (National Academy for Educational Research, 2014), is used to solve this problem. As Figure 3.5 shows, the bilingual text is already aligned, and the Chinese text is segmented into word units separated by a space. Although the segmentation is not perfectly correct, it should be noted the system segments Traditional Chinese text in a consistent manner. Thus, repeated similar errors in segmentation are unlikely to cause difficulties for the proposed NMT to learn the relation between Chinese words or between the bilingual word units. There are certainly other segmentation tools for Chinese such as Jeiba and Transformer, but considering the corpus used in this research consists of a smaller amount of Traditional Chinese text, it is assumed that NAER Segmentor can perform better in terms of accuracy.

doi:10.6342/NTU201903661

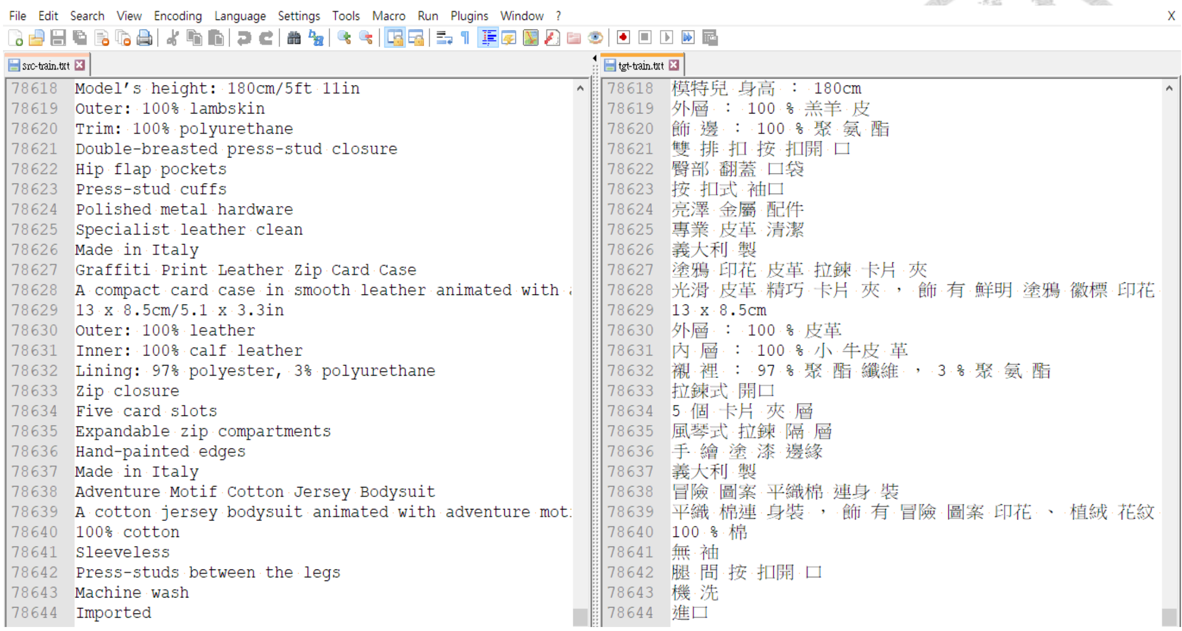


Figure 0.5 Plain text files extracted from the websites of H&M, ZARA and Burberry

By using the tools mentioned above, 82,815 lines of bilingual texts are collected, containing 723,147 English word types and 119,251 Traditional Chinese word types, in which punctuation marks are not calculated. Please note that a Chinese word type does not necessarily equal to a single Chinese character. The parallel corpus compiled previously is then divided into two corpora, the training corpus and the validating corpus. The former is used to train the customized NMT, while the latter serves as a reference translation in BLEU, an automatic evaluation method for machine translation used to assess the effectiveness of the NMT. In other words, after training, the effectiveness of the NMT is measured against the second bilingual parallel corpus. The following table shows the statistics of the two corpora.

doi:10.6342/NTU201903661

		English	Traditional Chinese
Training Corpus	Word Type Count	686,614	77,882
	Line Count	78,644	78,644
Validating Corpus	Word Count	36,533	41,369
	Line Count	4,171	4,171
Total Data Amount	Word Type Count	723,147	119,251
	Line Count	82,815	82,815

Table 0.1 Statistics on training and validating corpus

For the sake of convenience , the trained NMT is then presented as a simple interface for searching English texts' Traditional Chinese translation. As Figure 3.6 shows, the Traditional Chinese translation is segmented word by word.

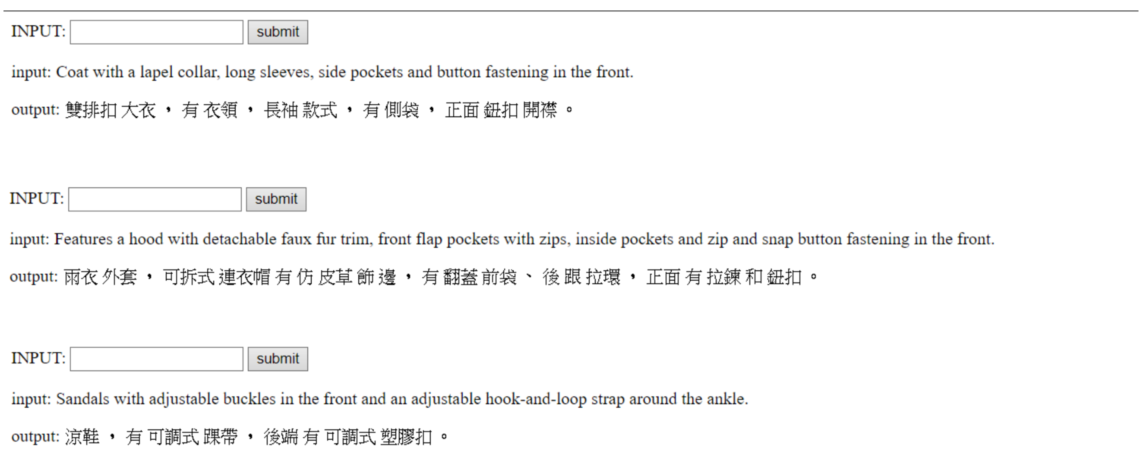


Figure 0.6 Interface of the customized NMT

3.3. Automatic Evaluation: BLEU

After the translation is generated by the trained NMT, BLEU is used to evaluate the results. The core idea behind this method is that "[t]he closer a machine translation is to a

doi:10.6342/NTU201903661

professional human translation, the better it is" (Papineni et al., 2002). That is, BLEU measures the automated translation's similarity to a reference human translation by comparing the n-grams in the two texts and counting the matches sentence by sentence. Theoretically, more matches indicate a better translation. Then, the system provides a numerical metric to determine the quality of the automated translation. The score provided by BLEU is between 0 to 1, but a score of 1 only occurs when the translation is perfectly identical to the reference text.

Papineni et al. (2002) point out that when automatic translation is measured against more reference translations, it scores higher in the BLEU system. Thus, whereas the validating corpus compiled in the previous steps consists of three different sources of texts, the corpus is treated as one single reference when calculating the BLEU score of the NMT-generated translation. This research adopts the MTEval toolkit to run the machine translation evaluation metrics, and the BLEU score achieved was 0.741479. The score represents the total average calculated with the average of 1-4 gram in each segment.

3.4. English Corpora analysis

For the convenience of discussion in the later steps, analysis has been done on the English training and validating corpora in order to identify some of its characteristics which might determine the final results of this research. The tools used in this step include AntConc, Notepad++ and Microsoft Excel. AntConc is a "freeware, multiplatform tool for carrying out corpus linguistics research" (AntConc Readme Help File, 2019) and is utilized to analyze the lexical aspects of the English texts. Some basic calculations on the composition of each line are also made with the help of Excel and Notepad++.

doi:10.6342/NTU201903661

3.4.1 Lexical characteristics

In AntConc, the Word List function "counts all the words in the corpus and presents them in an ordered list" (AntConc Readme Help File, 2019). Although the function is mainly used to "quickly find which words are the most frequent in a corpus" (AntConc Readme Help File, 2019), the goal of this function is to identify the Word Types and Word Tokens in the English corpus. The term "Word Types" refer to the number of unique words, while "Word Tokens" represent the total word count in the corpus. By default, AntConc does not count punctuation as a token. Furthermore, AntConc also shows the frequency of each word type in both corpora. As in Figure 3.7, the word types ranking from 3,194 to the end of the list, 3,859, only appear once in one of the corpora. That is, 676 words among 3,859 word types are either new vocabulary shown in the validating corpus, or one-time information in the training corpus.

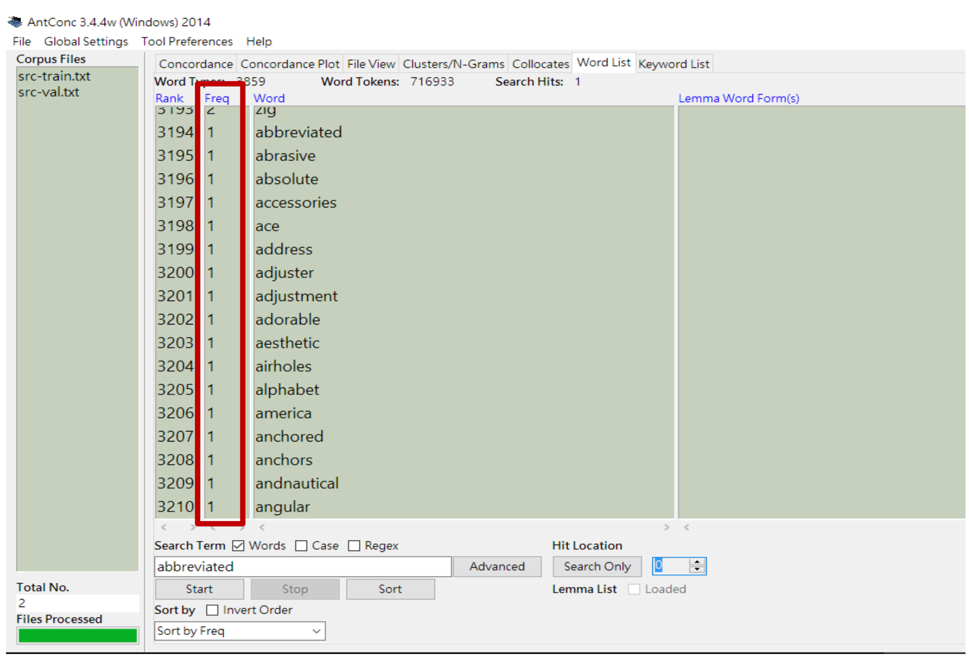


Table 0.2 Frequency of words in AntConc

doi:10.6342/NTU201903661

3.4.2 Composition per line

The "line" referred to here is defined as a segment of information collected from the websites. In other words, a single line might include more than one sentence, which is delineated through the use of a period. The word count per line in both corpora seems to vary, since all the lines are directly copied from the websites and do not undergo further segmentation. In order to understand the composition of each line in the corpora, all the English texts are copied and pasted to Excel, in which each cell contains a single line. The word count without punctuation in each cell can be calculated by using the following function:

```
=IF(LEN(TRIM(text))=0,0,LEN(TRIM(text))-LEN(SUBSTITUTE(text," ",""))+1)
```

The word count is then sorted from high to low in Excel, and the results show the minimum word per line is 1, while the maximum words per line is 93. Nevertheless, further analysis on the average word count per "sentence" cannot be carried out because although there are complete sentences ending with a period, there are also lines including only short phrases or even a single word without any punctuation.

3.5. Computer-aided translation

In order to interpret the performance of the customized NMT in a more comprehensive way, this research also utilizes CAT tools to pre-translate the same text as the NMT. By applying the pre-translation function of the tool, it is expected to gain more understanding on the similarity between the training and validating corpora and how it might affect the performance of the NMT. Also, through this method, the differences between the CAT tools and the customized NMT can be clearly identified.

doi:10.6342/NTU201903661

3.5.1 Building translation memory

The most helpful function a CAT tools provides for a translation project is its translation memory system. For this research, the bilingual training data collected from the websites is also used to compile a translation memory. Since the bilingual text is already segment-aligned, it is easy to convert the text into a translation memory file through various tools. First, the bilingual text is copied and pasted to Excel and saved as an .xlsx file, where the first column is the English text while the second column is Traditional Chinese. Then the .xlsx file is uploaded to the online tools provided by Translatum and converted into a .tmx file. Although .tmx is a format compatible with most of the CAT tools, to make the translation

doi:10.6342/NTU201903661

memory work in a particular CAT system like SDL Trados Studio 2017, a further step of upgrading the .tmx file to a .sdltm file in Trados is needed.

3.5.2 Pre-translation with CAT tools

The next step is to create a new translation project in Trados 2017. The source language

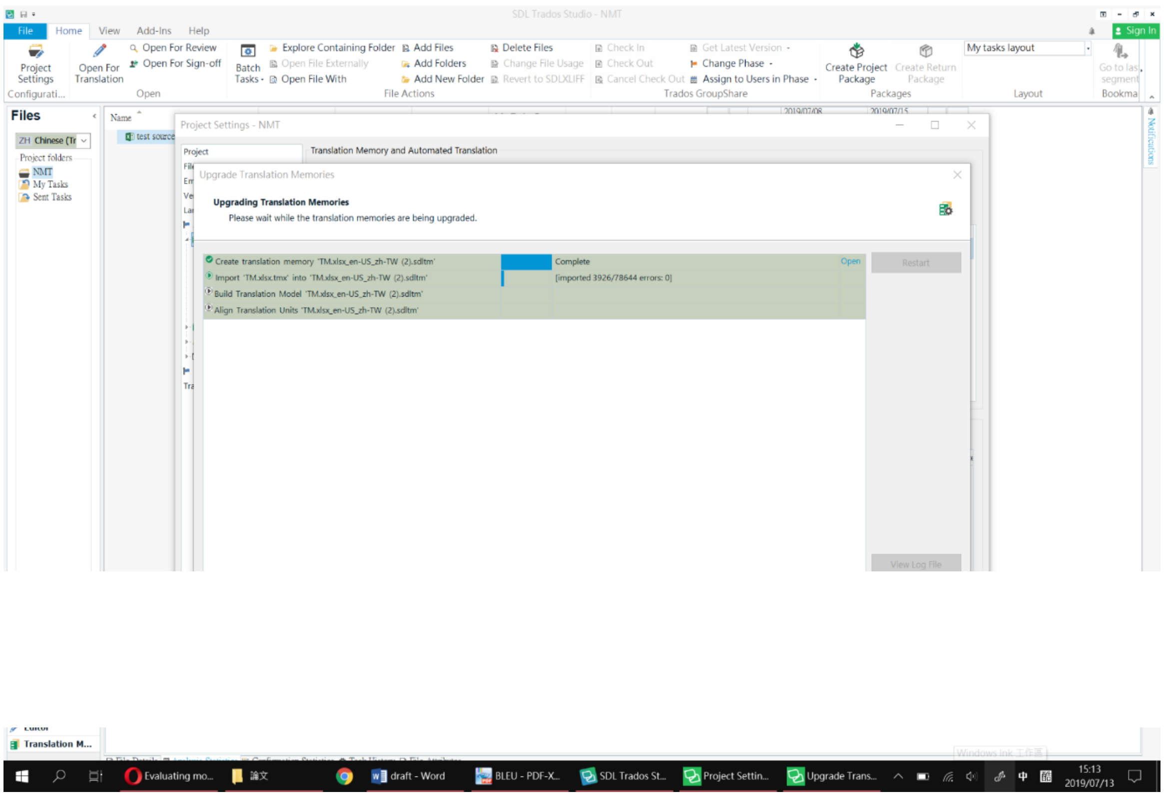


Figure 0.7 Upgrading .tmx file in Trados 2017

and target language are set as English and Traditional Chinese respectively, and then the English text from the validating corpus is imported as the source text. Since the translation memory is already compiled in the previous step, it can be directly added to the new project in Trados. Before the translation and editing start, Trados provides some matching statistics for the translator's references. The feature compares the similarity between the source text and the TM, revealing the word count of the source text that actually needs to be translated.

doi:10.6342/NTU201903661

Trados divides matching rates into seven groups: Repetition, 100%, 95%-99%, 85%-94%, 75%-84%, 50%-74% and No Match. As the following figure shows, Trados provides options for showing the matching rate of "segments" or "words" in percentages, but the way in which the matching rates are calculated in Trados is unknown to the users. The advantages of Trados include providing a clear view for the differences between a source segment and a similar segment from the TM, and it is more convenient for editing. As a result, the following experiment also utilizes Trados as the main tool.

3.6 Experiment: Comparison between NMT, Google Translate and Trados

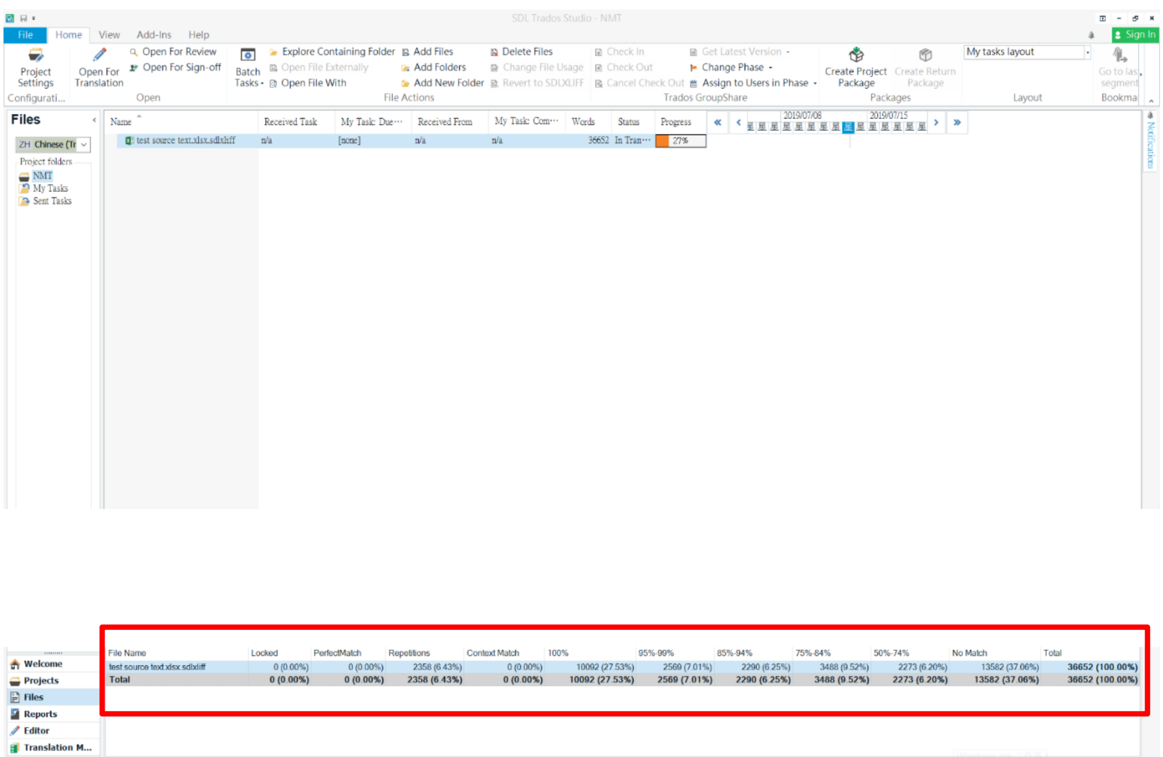


Figure 0.8 Pre-translate statistics provided by Trados 2017

doi:10.6342/NTU201903661

In order to identify how helpful the NMT proposed by this research is for professional translators, it is necessary to integrate the automated translation generated by NMT into a translation project and have professional translators working on the project. In the following experiment, the English texts in the validating corpus are used to create a new project in Trados and serve as the source text. Participants of the experiment, all of whom are professional translators, are asked to translate the selected segments with the help of a translation memory and machine translation. The participants have to complete two different projects. The two projects are created with different groups of selected English text. One project is supplemented with NMT translation, while the other uses Google Translate. The two projects contain segments similar in word count and length, and are equipped with the same translation memory which was built in the previous steps. After completing the two projects, participating translators are asked to answer a question about the roles these tools play in their working processes. The following sections describe in detail how and why this experiment is designed and carried out.

3.6.1 Selecting two groups of source text

The first factor to consider when selecting English segments as the source text is the total word count and segment lengths. Since the participating translators are expected to concentrate continuously on the two projects, the total word count of each project is set at around 200 words, which take roughly half an hour to translate. In addition, longer segments containing more than two sentences, which we determine by the presence of periods, are removed from the experiment to prevent participants from spending too much time on the

doi:10.6342/NTU201903661

same segment. The main purpose is to ensure that even though the two projects are created with different groups of English text, each source segment is almost equally understandable.

Next, to create two translation projects that both require the experiment participants to make the most of MT and translation memory and minimize the usage of other internet resources like search engines, the matching rates provided by Trados' pre-translation analysis are used to exclude segments not suitable for this experiment. For instance, 100% matches and segments with matching rate lower than 30% are not considered, because the former can be confirmed as the correct translation without the help of any tools, while the later contains too much new information that neither CAT nor MT provide adequate assistance. By examining the rest of the text thoroughly, we observed that there are six scenarios which the participants might face: Trados failing to process 100% matches, Trados failing to replace a small part of a segment with the correct translation, Trados contradicting the translator's judgement, NMT combining various sources to provide usable translations, NMT combining various sources to provide translation identical with the texts from, the websites, and NMT mistranslating 100% matches. Further explanations are provided in the next chapter.

	Source text word count (Excel)	Number of segments (Excel)
Group 1	225	13
Group 2	198	12

Table 0.3 Source text word count and segment count

doi:10.6342/NTU201903661

3.6.2 Creating translation projects in Trados

The two groups of text complied in the previous step are saved as two separate Excel files first so they can be imported to Trados as the source text for two different projects. In the process of creating two new projects, the translation memory built in Section 5.1 is also added to the projects. The final step is to copy and paste the translation generated by NMT and by Google Translate to the target segments of the two projects. As the Figure 3.9 and Figure 3.10 show, Project 1 and Project 2 contain different groups of source text and target text for reference, but the applied TM is the same file. Because the translation memory is created directly from the training corpus, there is a space between each Chinese “word unit.” The machine translation provided in Project 1, which is generated from the trained NMT, is also displayed in the same format. Moreover, it is observed that Trados tends to segment text based on periods, so there is a difference in the number of segments in each project in Trados (see Table 5).

doi:10.6342/NTU201903661

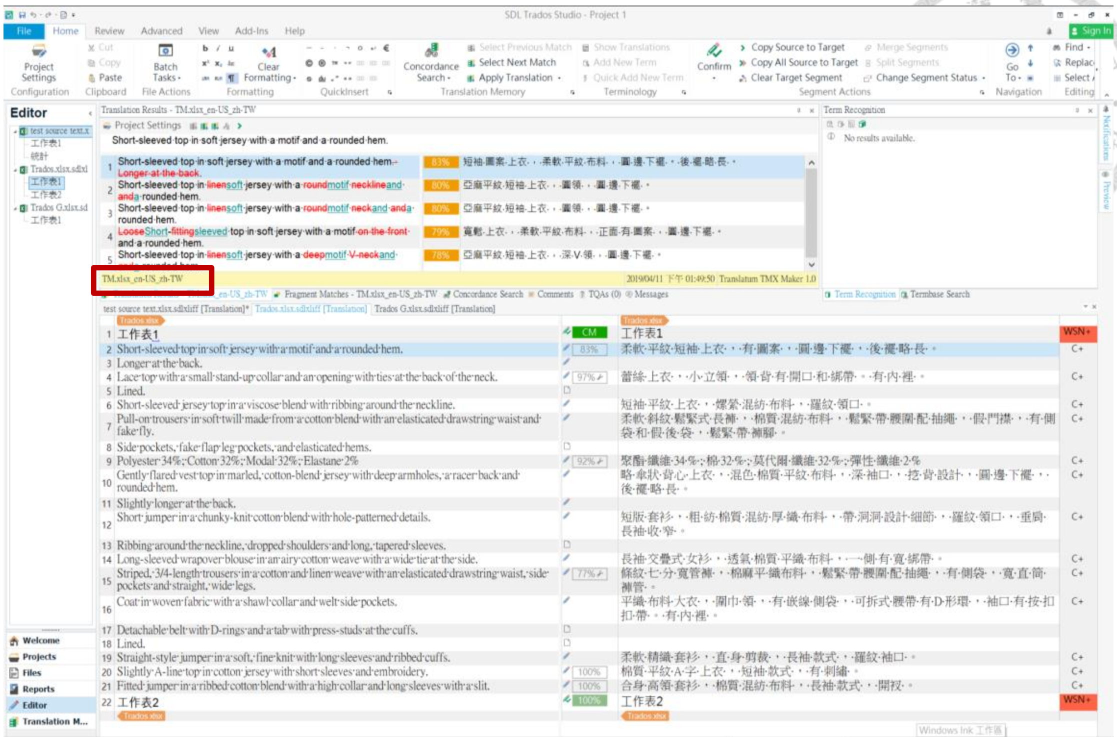


Figure 0.9 Project 1

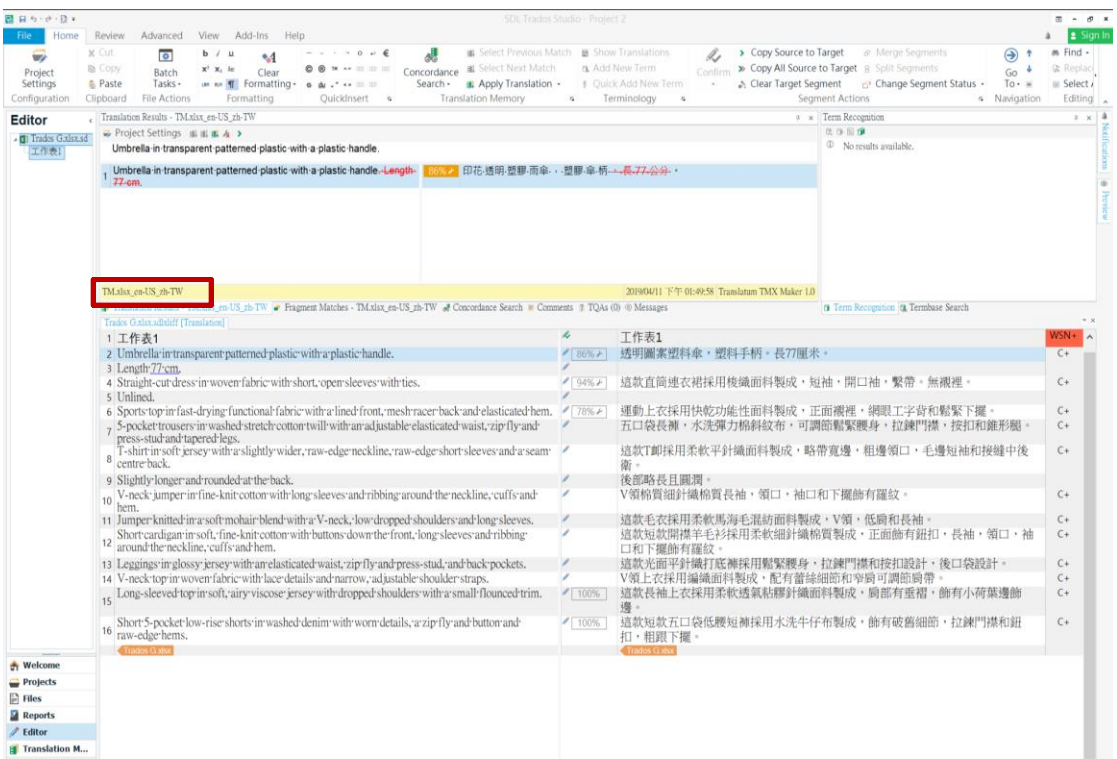


Figure 0.10 Project 2

doi:10.6342/NTU201903661

	Source text word count	Number of segments (Trados)	Machine translation provided
Project 1	225	20	NMT
Project 2	198	15	Google Translate

Table 0.4 Statistics on the two projects in Trados

3.6.3 Designing a questionnaire for feedback

To further understand the thoughts of the participating translators during the translation process, a questionnaire consisting of 10 questions is designed to collect their feedback. All of them are multiple-choice questions, as listed below in order:

The first question is "**In Project 1 and 2, what is "the most" helpful resource for improving translation work efficiency?"** For this question, selectable answers include "Translation memory," "Machine translation," "Other internet resources," and a blank field for other options. The purpose is to understand participants' thoughts on which tool has provided the greatest support for the project. It is expected the answers to the question can help identify whether the translators find using MT a better option than other tools. Upon answering that question, the following questions focus on the performance of the two types of MTs in each project: **Overall, is the machine translation provided in Project 1 coupled with TM helpful for improving translation work efficiency? Overall, is the machine translation provided in Project 2 coupled with TM helpful for improving translation work efficiency?** The selectable answers for these two questions are "Yes, I spent less time on the translation," "No, I spent more time on editing the MT," and "I did not find any

doi:10.6342/NTU201903661

differences." The aim is to confirm whether or not the MT had an effect on the translation process, regardless of if the effect was positive or negative.

The next series of questions is designed to compare the performances of the customized NMT and Google Translate based on the participants' perceived user experience. The comparison is targeted specifically to address the following aspects: **Which machine translation is more usable; in other words, which one requires less editing? In terms of glossary, which machine translation performs better? In terms of grammar and sentence structure, which machine translation performs better? In terms of punctuation, which machine translation performs better? Which machine translation can better compensate for the disadvantages of Trados? Which machine translation has greater disparity from the TM?** For these questions, the participating translators have to choose from "Project 1", "Project 2" or "No differences."

Lastly, to wrap up the questionnaire, the final question intends to clarify if the translators consider either type of MT as providing extra and irreplaceable support even with the availability of the other tools available: **Considering this test, which of the following combinations can result in the best working efficiency?** The answer options are "Only TM and other internet resources," "TM, MT from Project 1 and other internet resources," "TM, MT from Project 2 and other internet resources," and a blank field to fill in other answers. The purpose of the questionnaire is to investigate whether the participants' perceived user experiences are consistent with the observations this research has made.

doi:10.6342/NTU201903661

3.6.4 Handing out projects and instructions

The participants of this experiment are restricted to professional translators who are familiar with Trados 2017 and have experience in localization projects. Due to the limited time, only two qualified translators were found and invited to participate in this study. Since the experiment is not conducted on-site, clear instructions are crucial for participants to finish the test remotely and independently. First of all, for recording the whole translation process, the translators are asked to install "apowersoft-online-launcher" on their computers so they can go to the website https://www.apowersoft.tw/free-online-screen-recorder and use the online tool to record their working process (see Figure 3.11). The translators are than asked to complete the two projects while their every move is recorded. The order of finishing projects does not affect the final result, but there are some details which require clarifications.

The two projects share the same translation memory but provide different machine translations for references. Moreover, the TM serves as the standard for the two projects, while the two groups of MT text only function as references in each project and can be edited based on the judgement of the translators. Additionally, the translators are allowed to use any other internet resource if needed, but we recommended that the translators make the most of the translation memory first. The participants are also instructed to ignore the space between Chinese word units in the TM and MT in Project 1 (NMT) from the Chinese segmentation necessary for the NMT training in order to not waste time removing the spaces.

doi:10.6342/NTU201903661

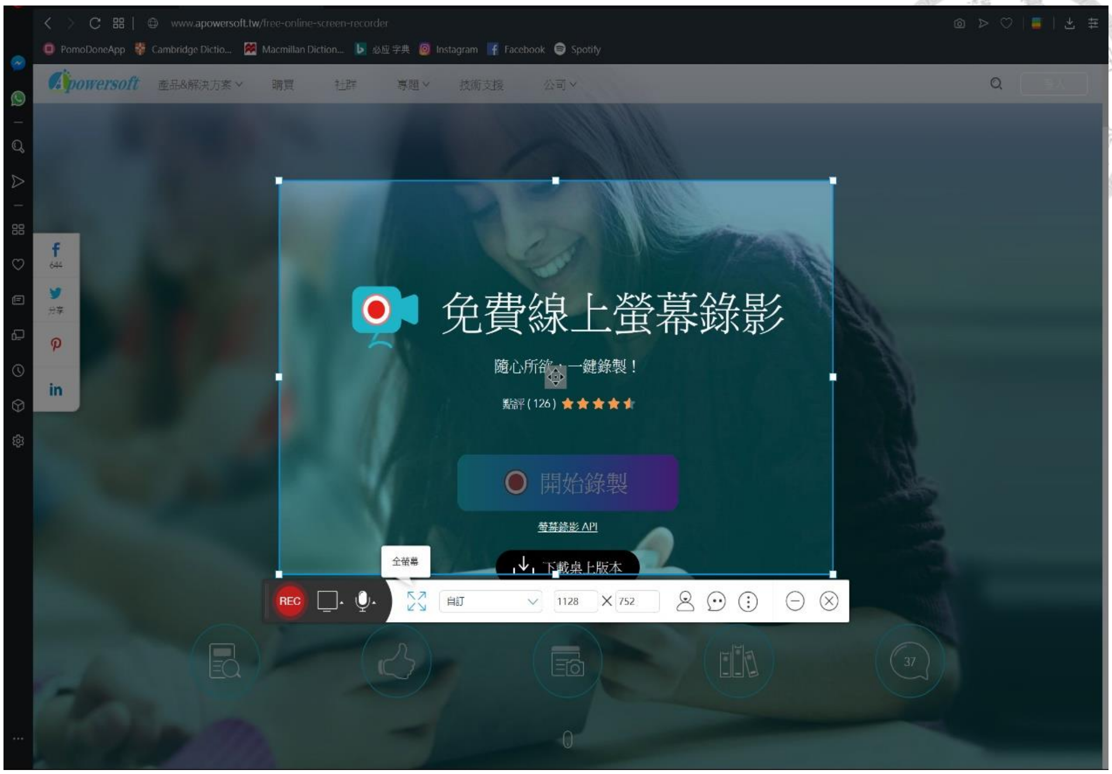


Figure 0.11 Online recording tool: Apowersoft

After translating all the projects, participating translators are asked to send back three files: the target language file for Project 1, the target language file for Project 2, and the video file which recorded their translation process. Lastly, translators are required to fill out the questionnaire from Section 6.3 at https://forms.gle/3zDmHUMDvpAAr9bYA. Only when all these requirements are fulfilled can the experiment come to a valid end. A Word file contains the above information, which is organized under ten bullet points. The participants are required to read through the information before commencing the translation projects. The translators are not informed beforehand which project contained the automatic translation from NMT and which one the translation from Google Translate. In this way, the information does not influence the participants'' feedback.

doi:10.6342/NTU201903661

3.6.5 Retrieving results

At this point, the retrieved files from the two translators include the target language texts of Project 1 and 2, videos recording the translation processes, and the feedback questionnaires. Since the translators expressed their thoughts on the experiment directly through the questionnaires, we will now begin the follow-up analysis. To further understand the opinions of the translators and the reasons for those opinions, the target files and videos are also carefully examined. Based on the data gathered through the methodology described in this chapter, the next chapter will focus on validating the effectiveness of the customized NMT with empirical evidence, as well as examining if the results correspond with the feedback from the professional translators.

doi:10.6342/NTU201903661

Chapter Four: Results and Discussion

This chapter aims at answering the research questions mentioned earlier. The first section addresses the question of how well the customized NMT performs. The evaluation involves interpretations on the final BLEU score, analysis on the training corpus, and examination of the statistics provided by CAT tools. The second part of the discussion centers on the thoughts of the translators on the perceived effects of the proposed NMT, and is largely based on the questionnaires filled out by the translators. The final section studies the empirical evidence discovered in the translation processes and results of the translators in order to verify the effects of the customized NMT.

4.1 Performance of the customized NMT

4.1.1 Automatic Evaluation: BLEU

The NMT proposed in this research has scored 0.741479 on the BLEU test, where a higher score generally indicates a better result. However, without comparison, the score can only be interpreted through referencing the guidelines provided by AutoML Translation (2019) and the scale proposed by Lavie (2011). When converted into a percentage point, the BLEU metric equates to a score of 74. According to the description in AutoML Translation's guideline, a score above 60 indicates "quality often better than human." Also, based on Lavie's scale, a score over 50 "generally reflect[s] good and fluent translation." Thus, the performance of the customized NMT is higher than expected at least in terms of its BLEU score.

doi:10.6342/NTU201903661

To further validate the performance of the customized NMT in comparison to the general MT of Google Translate, BLEU is also used to evaluate the automatic translation generated by the two types of MT. The text translated by the customized NMT and Google Translate is composed of the source texts from Project 1 and Project 2 of the experiment, with 27 segments in total. Since the Chinese translation provided by Google Translate is not segmented into word units, NAER Segmentor is used to pre-process the automatic translation so that it can be evaluated by BLEU test. The reference translation is the content gathered from the official websites of the three fashion brands as described in Chapter Three. The BLEU score of each segment is listed in Table 4.1, and the average and final BLEU scores for the customized NMT and Google Translate are 0.771 and 0.36. The customized NMT clearly performs better when using this evaluation metric.

Segment	NMT	Google Translate
1	0.43391954	0.346966645
2	1	0.274790761
3	1	0.813288281
4	0.869639866	0.421745731
5	1	0.239329796
6	0.536117217	0.419721415
7	0.668960466	0.746381949
8	0.435603381	0.385385692
9	0.872512939	0.335575109
10	1	0.313646099
11	1	0.229328732
12	0.63525139	0.565911926
13	0.547669096	0.801054897
14	0.767250675	0.040197869
15	1	0.224345316
16	0.6495123	0.491677627
17	0.749373194	0.241909484
18	1	0.279663563
19	0.554363965	0.446576991
20	0.801412272	0.299891008

doi:10.6342/NTU201903661

21	0.529868096	0.185438292
22	0.699867292	0.4768956
23	1	0.241179404
24	0.80032032	0.292829801
25	0.719548354	0.400357721
26	0.774075635	0.114343382
27	0.774029862	0.090999927
Sum =	20.81929586	9.719433018
Average =	0.771	0.36

Table 0.1 BLEU score of NMT and Google Translate

The great disparity between the BLEU scores of the two types of MT mainly results from the principles of BLEU and some characteristics in the translation by Google Translate. Since BLEU scores indicate the similarity between the MT text and the reference translation, this automatic evaluation is advantageous to the customized NMT because the training data for the NMT is more similar to the selected source text and the reference translation. Therefore, although Google Translate performs worse when compared to the customized NMT in terms of BLEU scores, it does not necessarily indicate the automatic translation provided by Google Translate is of poor quality. Furthermore, it should be assumed there are major differences between the MT text generated by Google Translate and the reference translation.

As the following example shows, in contrast to Google Translate, the customized NMT can produce a translation more similar to the reference translation in terms of sentence structure and diction. In this particular genre, punctuation marks are used frequently to divide information into smaller chunks to enhance readability. In this regard, Google Translate tends to form a longer chunk with more information as in the following example. Additionally, redundant articles such as “這款” are often found in the translation by Google Translate. As

doi:10.6342/NTU201903661

for diction, Google Translate makes more mistakes through either omission or simple mistranslation, often replacing terms with a synonym.

ST: Long-sleeved wrapover blouse in an airy cotton weave with a wide tie at the side.

Reference Translation (Website): 裹身式 長袖 女衫 ， 透氣 棉質 平織 布料 ， 側邊 有 寬 綁 帶 設計 。

NMT: 長袖 交疊式 女衫 ，透氣 棉質 平織 布料 ， 一 側 有 寬 綁帶 。

Google Translate: 這款長袖上衣採用輕盈棉質梭織面料製成，側面採用寬領帶設計。

Some of the synonyms chosen by Google Translate are not common in Traditional Chinese. In the above example, Google Translate uses the uncommon word choice “面料.” In other instances, Google Translate uses “粘膠” for “viscose” instead of “嫘縈,” and “塑料” for “plastic” instead of “塑膠.” These discrepancies occur because the of differences in expression between Simplified and Traditional Chinese. English-Simplified Chinese bilingual texts account for a significant proportion of English-Chinese training data for Google Translate, and thus Google Translate sometimes provides translations which are unfamiliar to Traditional Chinese readers. This also partly bolsters the argument for using customized NMT in a domain-specific localization project in which translators must adapt language usage to the need of local readers or consumers.

4.1.2 Training corpus analysis

doi:10.6342/NTU201903661

According to the Wordlist provided by AntConc, 676 words among 3,859 word types are not included in both corpora, implying that there is a low chance of the NMT translating unknown vocabulary. For a translator, if the customized NMT can translate most of the glossary correctly, it would be a great advantage, as it is observed the genre of the corpus more likely tolerates changes in word order. In the following example, human translators may find the translation provided by the NMT acceptable. Although the MT text is not perfectly identical to the official translation on the website, translators can choose to leave it unedited to save some time.

Website: V 領 針織 套衫 ， 柔軟 馬海 毛 混紡 布料 ， 低垂 肩 長袖 。

NMT: 柔軟 馬海 毛 混紡 針織 套衫 ， V 領 ， 低垂 肩 長袖 。

There are some features of the corpora that could result in a less effective NMT system. For instance, each line in the training corpus only contains around 9.3 words on average (727,540÷7,864), but the word count per line actually ranges from 1 to 93. The ideal composition of a line should be no more than one sentence in order to maximize the effectiveness of training. However, due to a limited amount of time and the way the bilingual text is aligned, it is difficult to segment lines of content extracted from the websites into smaller units simply based on punctuation.

4.1.3 CAT statistics analysis

According to Table 4.2, almost thirty percent of the words in source text have exact matches in the TM, and less than 10% of the text has no matches. This explains in part why the customized NMT performs well on the evaluation test as there seems to be a fair degree

doi:10.6342/NTU201903661

of similarity between the training corpus and the validating corpus. Furthermore, although way in which Trados 2017 derives its statistics remains unclear, the tool weighs every unmatched word equally regardless of its importance and the word count per segment can affect the matching rate. Thus, the statistics in Table 4.2 might be underestimated. In addition to examining the statistics provided by Trados 2017, the pre-translation is also compared to the MT text generated by the NMT to identify the major differences between the two tools.

	Segments	Words
Repetition	905 (16.3%)	2316 (6.33%)
100%	2335 (41.35%)	10092 (27.59%)
95%-99%	570 (10.09%)	2569 (7.02%)
85%-94%	178 (3.15%)	2272 (6.21%)
75%-84%	316 (5.60%)	3498 (9.56%)
50%-74%	964 (17.07%)	12192 (33.33%)
No Match	379 (6.71%)	3644 (9.96%)
Total	5647 (100%)	36583 (100%)

Table 0.2 Pre-translation Analysis by Trados 2017

First of all, although it is presumed that if a segment has a completely identical match in the training corpus and the TM, both the customized NMT and Trados are able to produce a perfect translation, some of the non-perfect matches in Trados are in fact fragments of a perfect match which has been segmented by the software mostly due to punctuations like periods. As the following examples show, when a segment in the source text contains more than one period, Trados automatically divides it into several segments based on the number of periods, while NMT processes it as a single line of text. In Example 1, NMT is able to produce a translation similar to the website's content without errors. In contrast, Trados's

doi:10.6342/NTU201903661

output contains an untranslated and in fact unnecessary segment, “Unlined,” despite the rest of the translation being identical to that of the website.

Example 1

ST: Straight-cut dress in woven fabric with short, open sleeves with ties. Unlined.

NMT: 平織 布料 洋裝 ， 直 身 剪裁 ，開口式 短袖 有 綁帶 。 無 內 裡 。

Trados: 平織 布料 洋裝 ， 直 身 剪裁 ，開口式 短袖 ，袖口 有 綁 帶 。 無 內 裡 。 Unlined.

Website: 平織 布料 洋裝 ，直 身 剪裁 ，開口式 短袖 ，袖口 有 綁 帶 。 無內 裡 。

Example 2

ST: Pyjamas in soft jersey. Wide, short-sleeved top with a print motif on the front. Short shorts with an elasticated drawstring waist.

NMT: 柔軟 平紋 睡衣 套裝 。 寬鬆 短袖 上衣 ， 正面 有 圖案 ；超短褲 有 鬆緊 帶 腰圍 配 抽 繩 。

Trados: 柔軟 平紋 睡衣 套裝 。，短袖 平紋 連身 睡裙 ， 正面 印 圖案 。柔軟 平紋 超短褲 ， 鬆緊 帶 腰圍 配 由 繩 。

Website: 柔軟 平紋 睡衣 套裝 。 寬版 短袖 上衣 ， 正面 有 圖案 ； 超短褲 有 鬆緊 帶 腰圍 配 抽 繩 。

As for Example 2, the source text contains three units of text. NMT identifies the text as one line and thus generates the same target text as the website’s content, which was used to train the customized NMT. In Trados, the system processes the source text as three separate segments, matches each with the most compatible TM segment, and then pieces them

doi:10.6342/NTU201903661

together to produce the target text. This is why Trados' translation is similar to the correct translation to a degree but requires further editing in these examples. This problem further escalates when it encounters a segment containing more periods. From a translator's perspective, it is a waste of time to correct a segment that could have been identified as a 100% match and translated correctly by a CAT tool.

Apart from failing to provide 100% matches correctly, even for matching rates above 90%, sometimes Trados is unable to replace the unmatched part in a segment with the correct translation that can be found in the TM. In fact, Trados 2017 is equipped with a new feature called "Match Repair," which "identifies the differences between translation suggestions and the source text, and replaces the non-matching text in your translation suggestions with data from the actual source text" (SDL Trados Studio 2017 Professional, Version 14.1.10015.44945). However, in most cases, translators still have to use concordance searches to manually find suitable translations from the TM, and then copy and paste it to the target language segment. On the other hand, if a small part of a line is different from the highly similar line found in the training data but already has a corresponding translation in the training corpus, NMT has no difficulties replacing it with the corresponding translation. In the following examples, Trados provides a 92% match and a 94% match respectively but is still unable to piece together a correct translation.

Example 3

ST: Polyester 34%; Cotton 32%; Modal 32%; Elastane 2%

NMT: 聚酯 纖維 34 % ; 棉 32 % ; 莫代爾 纖維 32 % ; 彈性 纖維 2 %

Trados: <u>貝殼 :</u> 聚 酯 纖維 34 % ; 棉 32 % ; 莫代爾 纖維 32 % ; 彈性 纖維 2 %

Website: 聚酯 纖維 34 % ; 棉 32 % ; 莫代爾 纖維 32 % ; 彈性 纖維 2 %

doi:10.6342/NTU201903661

Example 4

ST: 5-pocket trousers in washed stretch cotton twill with an adjustable elasticated waist, zip fly and press-stud and tapered legs.

NMT: 5 袋式 水洗 彈性 棉質 斜紋 長褲 ， 可 調式 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按扣 ， 褲管 收窄 。

Trados: 5 袋式 長褲 ， 水洗 彈性 棉質 斜紋 布料 ， 可 調式 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 鈕扣 ， 褲管 收窄 。

Website: 5 袋式 水洗 彈性 棉質 斜紋 長褲 ， 可 調式 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按扣 ， 褲管 收窄 。

The matching segment provided by Trados sometimes contradicts translator's judgement (which is based on a concordance search). The example below shows that NMT and Trados choose different translations. Trados finds a 93% match to the source text in the TM, while the NMT provides a different choice. By using the concordance search in Trados to search the source text, it is observed that in fact there is a 95% match in the TM (see Figure 4.1), which is exactly the translation the NMT provides.

Example 5

ST: V-neck jumper in fine-knit cotton with long sleeves and ribbing around the neckline, cuffs and hem.

NMT: 棉質 精 織 套衫 ， V 領 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。

Trados: 蠟染 印花 棉質 精 織 套衫 ， 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。

Website: V 領 棉質 精 織 套衫 ， 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。

doi:10.6342/NTU201903661

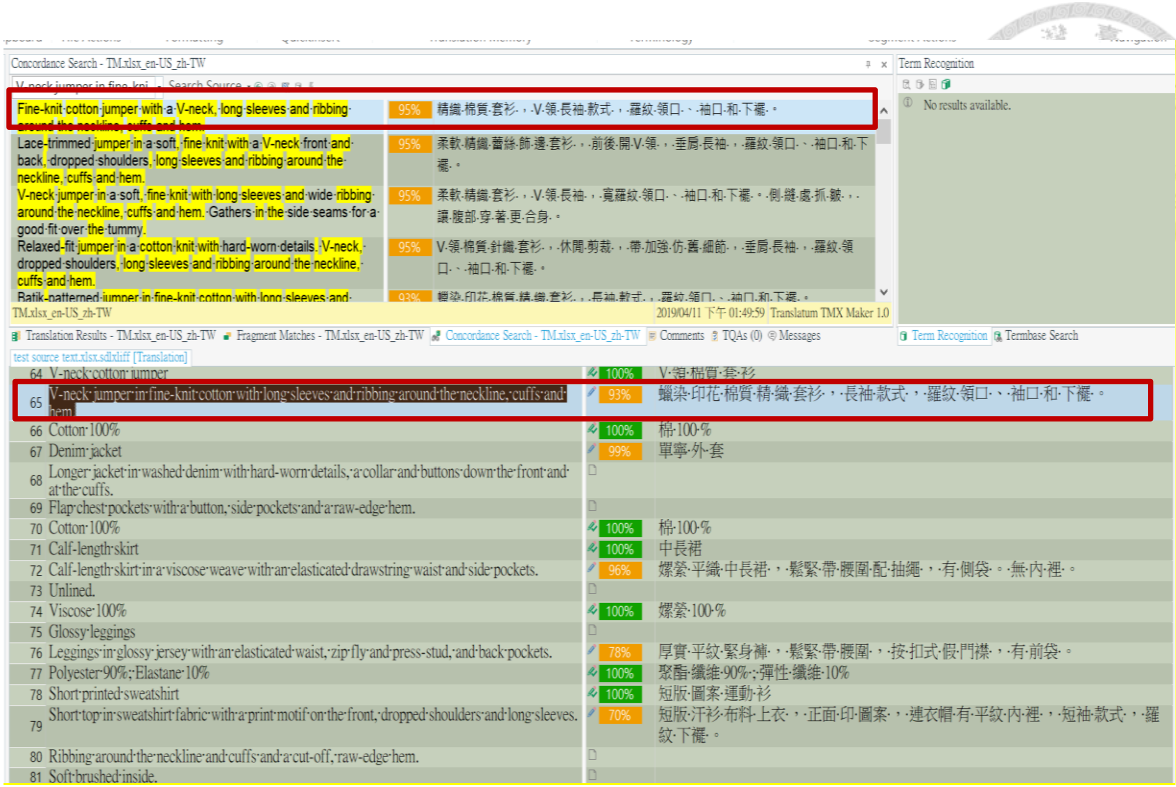


Figure 0.1 Concordance search in Trados 2017

The above analysis is based on observation and cannot fully reflect the actual user experience of the professional translators. To understand the actual opinions of translators on the effectiveness of customized NMT, it is necessary to carry out an experiment and collect both feedback from professional translators and more empirical evidence.

4.2 Perceived effects of the customized NMT

The importance of this section lies in that the participating translators share their impression of using the customized NMT, a topic which rarely mentioned in previous research on the matter. If their feedback is consistent with the results of the evaluation mentioned in the previous section, it would be reasonable to say that the NMT can provide constructive assistance to translators. On the other hand, if the user experience of the

doi:10.6342/NTU201903661

translators contradicts the result of the evaluation of the NMT, it would be a great opportunity to investigate reasons for the gap between the two.

The following table shows the two translators' answers to each question. Please note that machine translation in Project 1 refers to NMT, while machine translation in Project 2 refers to Google Translate. However, the participating translators are not informed which MT text is shown in which project to not influence their feedback with any preconceptions or expectations of the automatic translation. The two participants answer every question with options provided by the questionnaire and do not fill in the blank field to provide additional thoughts.

	Question	**Translator A**	**Translator B**
1	In Project 1 and 2, what is "the most" helpful resource for improving translation work efficiency?	MT	TM
2	Overall, is the machine translation provided in Project 1 coupled with TM helpful for improving translation work efficiency?	Yes, I spent less time on the translation	Yes, I spent less time on the translation.
3	Overall, is the machine translation provided in Project 2 coupled with TM helpful for improving translation work efficiency?	Yes, I spent less time on the translation.	No, I spent more time on editing the MT.

doi:10.6342/NTU201903661

4	Which machine translation is more usable; in other words, which one requires less editing?	No differences	Project 1
5	In terms of glossary, which machine translation performs better?	No differences	Project 1
6	In terms of grammar and sentence structure, which machine translation performs better?	Project 2	Project 2
7	In terms of punctuation, which machine translation performs better?	No differences	No differences
8	Which machine translation can better compensate for the disadvantages of TM?	Project 2	Project 1
9	Which machine translation has greater disparity from the TM?	Project 2	Project 2
10	Considering this test, which of the following combinations can result in the best working efficiency?	TM + Project 2 MT + Internet resources	TM + Project 1 MT + Internet resources

Table 0.3 Translators' answers to the questionnaire

For the first question, the two translators provide different answers. Translator A finds MT the most useful tool when choosing between TMs, MT and other internet resources, while Translator B prefers TMs. Their thoughts in this regard reflect their personal working

doi:10.6342/NTU201903661

preferences: Translator A tends to make as little editing to the MT text as possible, while Translator B prioritizes compliance with the TM and is willing to make as many changes to the MT text as needed. The difference is especially obvious in Project 2, where the MT is generated by Google Translate. Please see the example below:

Google Translate: 這款光面平針織打底褲採用鬆緊腰身，拉鍊門襟和按扣設計，後口袋設計。

Translator A: 這款亮面平織內搭褲採用鬆緊帶褲頭、拉鍊門襟和按扣設計，有後口袋。

Translator B: 亮面平紋內搭褲，鬆緊帶腰圍，拉鍊門襟配按扣，有後口袋。

Next, the two translators agree on the second question. They both find the MT text provided in Project 1 has helped them spend less time on the translation, meaning the customized NMT does have a noticeably positive effect on their translation processes. As for the third question, Translator A still finds the MT text provided in Project 2 useful, but Translator B thinks the opposite. Translator A's answers to these two questions are consistent with her feedback on the first one. Since Translator A acknowledges both of the MT text are effective in shortening translation time, it is reasonable that she decides MT the most useful tool among all resources mentioned. On the other hand, based on Translator B's answer to the third question, it is presumable that in some aspects, editing the text generated by Google Translate is more time-consuming than the text from the NMT. The answers to the following questions might be able to provide further explanation of this phenomenon.

According to Translator A's feedback on the fourth question, there is no obvious difference between editing the translation by the NMT and that by Google Translate.

doi:10.6342/NTU201903661

Furthermore, considering her answers to the previous questions, Translator A most likely thinks both MTs are equally effective in reducing translation time and effort. In contrast, Translator B's answers to Question 2, 3, and 4 show she acknowledges the customized NMT is more effective in increasing translation efficiency than Google Translate In the three questions that follow, the two translators basically share the same idea. The only disagreement is that Translator B thinks the NMT outperforms Google Translate in terms of diction while Translator A does not find either MT better. Interestingly, both translators prefer the performance of Google Translate in terms of grammar and sentence structure. Lastly, it seems both types of MT can process punctuation appropriately. Based on these answers, it can be assumed Translator B spent their time and effort on editing terms in the text generated by Google Translate.

The two translators' views diverge on the eighth question. Translator A finds Google Translate a better support tool when using TM to translate the projects, while Translator B believes the customized NMT provides more additional support. Although it is unsure the reason behind their difference in opinion, their answers to the next question may give some clues. Both translators observe the customized NMT's translation is more similar to the text in the TM than that of Google Translate, even though they are not informed of the fact that the customized NMT and the TM are derived from the same corpus. Considering this, Translator A's answer to the eight question could be based on the idea that Google Translate has successfully provided another version of an acceptable translation and thus can make a good complement to the TM. This idea also corresponds to the translator's answers to the previous questions, showing she trusts the translation by Google Translate. In contrast, Translator B acknowledges the translation of the NMT is close to the text from the TM but

doi:10.6342/NTU201903661

also provides extra help, implying the NMT has suggested better translations than the TM in some cases.

Last of all, the two translators' answers to the final question provide defining evidence for identifying the effectiveness and necessity of either MT system in the experimental tasks. Consistently, Translator A decides that Google Translate, TM and other internet resources are the most ideal tool combination for the translation projects. On the other hand, Translator B shows her preference for the combination of the customized NMT, TM and other internet resources. Although the translators have differing views in this regard, we can confirm that both MT systems are useful for the translators. Therefore, the follow-up discussion will focus on what really makes the customized NMT different from Google Translate.

4.3 Empirical effects of the customized NMT

At first glance, the translator's feedback seems to provide mixed results, so it is still difficult to reach a conclusion on the effectiveness of the customized NMT. To further investigate into the reason behind their answers, the following analysis focuses on two aspects: time and editing. The total time the translators spent on each project is calculated, and how the translators spent their working time is also examined. Next, the editing processes are categorized into different types of action to identify the difficulties the participants may have encountered during translation, as well as the assistance they received from the various translation tools. Through the analysis we will understand how these tools (Trados, NMT, Google Translate or other internet resources) can provide the assistance the translators need, and what kinds of problems remain unsolved even with the help of these tools.

doi:10.6342/NTU201903661

4.3.1 Translators' working time

To start with, the recorded videos are used to measure how much time the two translators spent on each project. As Figure 4.2 shows, the recording tool provides a timer, which is helpful for calculating the total time spent on each project and making a comparison. The total time is counted from when the translators started editing the first segment to the time they confirming the last segment. Due to the fact that the two projects vary in word count, the average time (in seconds) per word and per segment is also calculated (see Table 8).

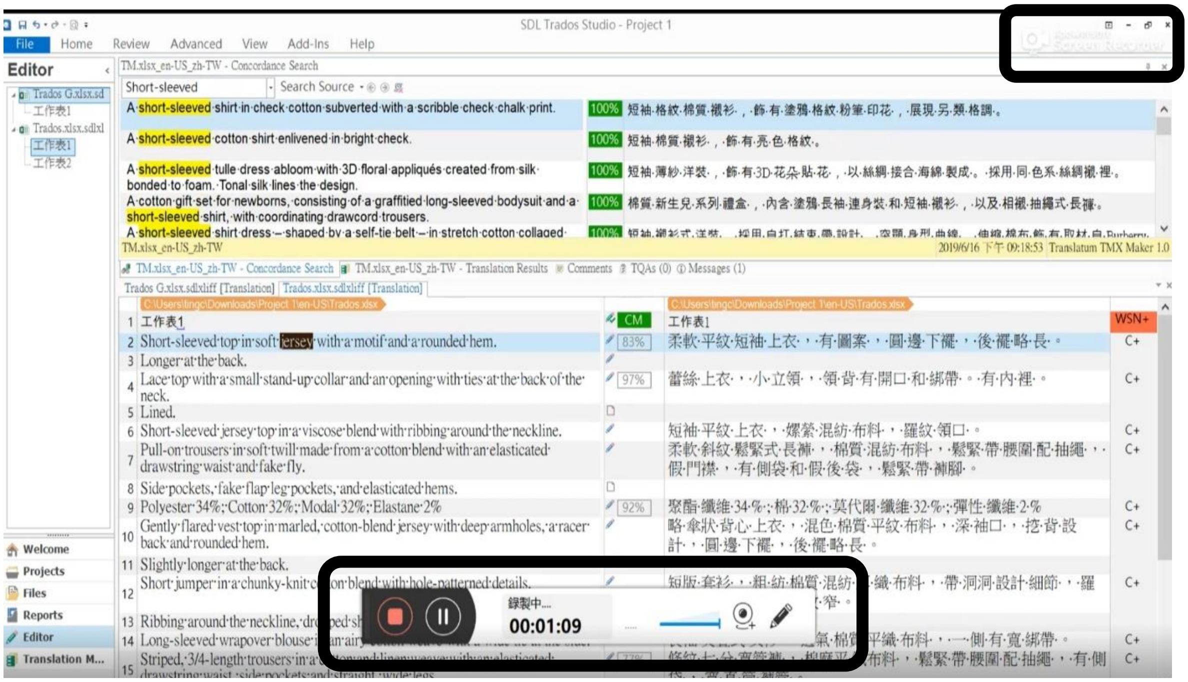


Figure 0.2 Screenshot of the recorded video

doi:10.6342/NTU201903661

	Project 1			Project 2		
	Total time	Sec./Word	Sec./Seg.	Total time	Sec./Word	Sec./Seg.
Translator A	00:16:10	4.31	48.5	00:17:58	5.44	71.87
Translator B	00:53:55	14.38	161.75	00:29:13	8.85	116.87

Table 0.4 Translators' working time spent on each project

There are obvious differences between two translators' total translation time, and consequently the total time spent on each project. It took less than 20 minutes for Translator A to complete one project, but the time she spent on a word and a segment in Project 2 is clearly more than that in Project 1. Interestingly, despite thinking Google Translate provides more assistance in translation than the customized NMT, Translator A's performance on the tasks show the opposite result. As for Translator B, despite being convinced that the customized NMT provides more useful translations than Google Translate, she spent much more time on Project 1. It seems that the translation time cannot fully explain the translators' user experiences.

At this point, the evidence is still insufficient to validate the effectiveness of NMT, and there is some confusion which needs clarification. To further explore whether the translators' performances correspond with or contradict the feedback they gave in the questionnaires, it is necessary to understand how their working time is distributed, and what results in such a disparity between the time it took both translators to complete the projects. Both the two

doi:10.6342/NTU201903661

participants' translations and the videos recording their working processes are thoroughly examined to track and classify their editing actions.

4.3.2 Translators' editing actions

The number of segments in the two projects edited by the translators is recorded in the table below. Please note the segment count here is not based on Trados 2017 but the validating corpus. Considering how many segments have been edited, it is not hard to understand why Translator B spent much more time on translation than their counterpart. Except for correcting mistranslated segments, Translator B also makes changes to segments without errors, especially those from Project 1. The reasons, however, are still unknown. In contrast, Translator A tends to leave the MT text untouched. It is also noteworthy that Translator A left some mistranslated segments unedited, while Translator B has not, implying Translator A holds a more tolerant attitude toward both groups of MT text than Translator B.

Even when taking the translators' personal preferences into consideration, it is still observed that the translation provided by Google Translate requires more editing, since both translators choose to edit segments without translation errors in Project 2. Up to this point, the time spent on translation is essentially determined by their own personal working style. Moreover, the two groups of MT text have obvious differences which result in more editing to be done on the translation by Google Translate. Therefore, in order to pinpoint the key differences, the next analysis needs to further explore the types of editing done by the translators. As a side note, there are some errors in the edited segments and which were left uncorrected by the translators, probably due to negligence and/or a lack of familiarity with

doi:10.6342/NTU201903661

this translation domain. However, this kind of problem occurs only occasionally, and as such is not included in the discussion.

	Translator A		Translator B	
	Project 1	Project 2	Project 1	Project 2
Unedited segments without errors	4	0	1	0
Unedited segments with errors	1	2	0	0
Edited segments	8	10	12	12
Total	13	12	13	12

Table 0.5 Statistics on segment editing

The editing actions of the translators are recorded in Table 10. There are some actions the translators spent time on which were imperceptible in their translation. Thus, we thoroughly reviewed the recordings of their translation process to provide a form of supplementary evidence. The number of these editing actions signifies the number of changes the translators make to the MT text. Therefore, editing one word, rather than editing one Chinese character, is counted as one editing action. As the example below shows, in Translator A's translation, the revisions of "領線" and "飾有" are counted as two editing actions. Moreover, the underlined word "羅紋" indicates a rearrangement to the word order/sentence structure, so one more editing action is counted. As for Translator B's editing, replacing "平紋" with "平織" is counted as one editing action.

doi:10.6342/NTU201903661

NMT: 短袖 平紋 上衣 ， 嫘縈 混紡 布料 ， 羅紋 領口 。

Translator A: 短袖平紋上衣，嫘縈混紡布料，**領線飾有**羅紋。

Translator B: 短袖**平織**短袖上衣，嫘縈混紡布料，羅紋領口。

To start with, the editing actions of the translators are divided into two categories based on whether or not the meaning of the MT text has been altered. Under the category of "meaning unaltered," subcategories are not included because the errors made by the two MT systems in the experiment differ both in number and types. The customized NMT tends to have missing words in its translation, while Google Translate tends to mistranslate terminologies. In this research, terminology is defined as professional terms related to fashion, which are expected to be translated in a consistent way. In this case, the NMT makes far fewer mistakes when compared to Google Translate, which in turn saves the translators effort when editing.

		Translator A		Translator B	
		Project 1	Project 2	Project 1	Project 2
Meaning unaltered	Replacing correct terminologies with synonyms	16	17	10	32
	Adding words for readability	6	16	5	8
	Removing redundant words	1	0	1	7

doi:10.6342/NTU201903661

	Rearranging word order or sentence structure	1	1	5	8
Meaning altered	Replacing MT errors with correct translation	1	12	3	15
Total count		25	46	24	70

Table 0.6 Statistics on editing actions

The most notable part in Table 4.6 is that both translators most frequently resort to "replacing correct terminologies with synonyms" when compared to other editing actions. There are several explanations for this result. First of all, this experiment did not provide a terminology list as a standard, so the participants have to determine which of the available options is most suitable. These options include the translation provided by the MT, varying word usages from the TM which are in fact composed of bilingual text from three different websites, and other synonyms found on the internet. Through examining the videos, we observed due to their unfamiliarity with this domain of translation, the two translators tend to be indecisive when determining whether the automatic translation for terminology is correct and often double check by searching in the TM and/or on the internet. Please see the following example:

ST: Coat in woven fabric with a shawl collar and welt side pockets. Detachable belt with D-rings and a tab with press-studs at the cuffs. Lined.

doi:10.6342/NTU201903661

NMT: 平織 布料 大衣 ， 圍巾 領 ， 有 嵌線 側袋 ， 可拆式 腰帶 有 D 形環，袖口 有 按 扣扣 帶 。 有 內 裡 。

Translator A: 針織大衣，新月領，關邊側袋。可拆式腰帶附有 D 形環，袖口有按扣扣帶。有內裡。

Translator B: 平織布料大衣，圍巾領，西裝側袋。可拆式腰帶，飾有 D 形環，袖口有按扣扣帶。有襯裡。

The above example is selected from Project 1, but in Project 2, this type of editing occurs more frequently because the word choice of Google Translate differs more from the TM than that of the customized NMT. However, due to the fact that Translator A is inclined to leave the MT text intact, the differences in the numbers of edits is not obvious in her case. Interestingly, although the two translators both decide Google Translate performs better on sentence structure and grammar, sometimes the translation by Google Translate is so fragmented and riddled with redundant articles that they have to make more changes to the translation. Translator A adds more connecting words in Project 2, and Translator B rearranges the sentence structure in translation by Google Translate more frequently, as the following example shows:

ST: Jumper knitted in a soft mohair blend with a V-neck, low dropped shoulders and long sleeves.

Google Translate: 這款毛衣採用柔軟馬海毛混紡面料製成，V領，低肩和長袖。

Translator A: 這款套頭衫採用柔軟馬海毛混紡布料製成，採 V 領、落肩和長袖設計。

Translator B: V 領柔軟馬海混紡針織套衫，低垂肩長袖。

doi:10.6342/NTU201903661

To further validate the statistics on editing, a two-way Anova test is applied to analyze if the two variables, (the types of MT and the types of editing) have a statistically significant effect on the number of editing actions. Since this method only takes two variables into consideration, the two translator's editing statistics are tested separately. As Figure 4.3 indicates, the statistical analysis helps bring us to the conclusion that "Blocks" influence the result. This means the translations generated by the customized NMT and Google Translate have affected the number of edits count by Translator A despite the fact that the translator finds the two groups of MT text equally useful (when referencing their answers to the questionnaire). In contrast, Translator B's case is rather different (see Figure 4.4). Statistically, the number of edits by Translator B is not influenced by what type of MT is used but rather by what type of editing is done, implying the translator tends to repeat specific editing actions. Through close examination of the videos, it is observed that Translator B indeed spends more time on checking terminologies multiple times.

doi:10.6342/NTU201903661

	Treatments					
	1	2	3	4	5	6
A	16	6	1	1	1	X
B	17	16	0	1	12	X
C	X	X	X	X	X	X
D	X	X	X	X	X	X

CALCULATE	CLEAR		
Treatment Variation	44.1	Block Variation	92.35
Within Variation	16.85	Total Variation	53.4333333
Treatment Statistic	2.6172107	Its P-Value	0.18752
Block Statistic	5.4807122	Its P-Value	0.07978
Conclusion on Treatments Effects			
Little or no real evidences against the null hypothesis			
Conclusion on Blocks Effects			
Suggestive evidence against the null hypothesis			

Figure 0.3 Two-way ANOVA test on Translator A’s editing

doi:10.6342/NTU201903661

Treatments						
	1	2	3	4	5	6
A	10	5	1	5	3	X
B	32	8	7	8	15	X
C	X	X	X	X	X	X
D	X	X	X	X	X	X

CALCULATE	CLEAR		
Treatment Variation	211.6	Block Variation	90.35
Within Variation	32.35	Total Variation	78.0444444
Treatment Statistic	6.5409583	Its P-Value	0.04859
Block Statistic	2.7928903	Its P-Value	0.1705
Conclusion on Treatments Effects			
Moderate evidence against the null hypothesis			
Conclusion on Blocks Effects			
Little or no real evidences against the null hypothesis			

Figure 0.4 Two-way ANOVA test on Translator B's editing

The two-way Anova test is also applied to analyze the total count of the two translators' editing actions in order to statistically validate if the customized NMT helps reduce the number of editing actions by the two translators. However, the result is that it is unable to confirm the effectiveness of the NMT. The possible reason behind this could be limited samples and/or data. It should be noted that although the effects of the customized NMT cannot be proven by this statistical method, the qualitative analysis still supports the claim that the NMT is more effective than Google Translate in this experiment.

doi:10.6342/NTU201903661

This chapter centers on answering the research questions proposed. First, by interpreting the BLEU score, it is found that the customized NMT's performance is beyond expectations. The analysis on the training corpus and examination of the statistics provided by CAT tools partly explain why the NMT performs well in the automatic evaluation. Secondly, based on the feedback by the translators on the questionnaires, we could identify the perceived effects of the customized NMT, but some results seem to contradict each other. The last part of the discussion goes through the empirical evidence from the participants' translation processes and final translation results. Finally, we confirm the translators' personal working styles and preferences have a substantial impact on both the perceived effects of the customized NMT on their translations and the time spent translating. Though the effectiveness of the proposed NMT is not shown to expedite the editing process, some evidence has proven that it is effective in reducing the number of edits.

doi:10.6342/NTU201903661

Chapter Five: Conclusion and Limitation

5.1 Summary and reflection

This research aims to validate the performance and effects of the proposed NMT. Based on the results of the automatic evaluation and the experiment which involved professional translators, it can be concluded that the customized NMT performs well on automatic evaluation, but the translators' experience using the NMT are not entirely positive, which does limit the perceived effects of the NMT on their work. Despite this, most of the empirical evidence shows the customized NMT is effective in facilitating the experimental localization processes.

When evaluating the performance of the proposed NMT with the widely used automatic evaluation method BLEU, the scores indicate a satisfactory translation. Although BLEU is not the latest nor the most convincing approach to validate a MT system's performance, its accessibility makes it suitable for serving as a baseline for comparing the customized NMT and the general MT of Google Translate. The analysis of BLEU scores for the two types of MT also suggests the customized NMT is more suitable for a domain-specific localization project because issues such as inconsistent sentence structure and uncommon translation for terminology can be avoided through the training data.

To understand the reason behind the high score, the training corpus is examined and the analysis indicates that numerous words are included in both the training and validating corpora, which may enhance the NMT's ability to translate terminology. However, the varied length of lines in the corpora might have a negative impact on the training. By using the pre-translation of Trados 2017, it is further confirmed that almost one-third of the text in the

doi:10.6342/NTU201903661

validating corpus has exact matches in the training corpus, which could have also led to a better training result. The discussion also includes what the NMT can do but the CAT tool cannot. Overall, the training corpus compiled in this research is of high enough quality to build a high-performing customized NMT.

The next part of the analysis centers around the feedback from two professional translators' on working with the translation provided by the customized NMT and Google Translate. Their answers to the questionnaires show mixed results, so it is not easy to reach a definite conclusion. At this point, it is clear that there is a disparity between the automatic evaluation and actual user experience. Although the participants' opinions are not entirely positive, they highlight the importance of this research. The purpose of this experiment is not making the customized NMT scores higher in automatic evaluation, but investigating whether or not it has provided what the software users will need. In spite of the fact that the customized NMT is not optimized in its framework, configurations and evaluation, it does show customized MT is advantageous on a specialized translation task when compared to general MT.

With the professional translators involved, the analysis has become more complicated, and more factors need to be considered. For instance, although it is expected that a translation generated by the proposed NMT would reduce the translation time, in reality, the translator's personal working styles and preferences have a more considerable impact. Thus, the follow-up analysis focuses more on the number of editing actions taken by the translators. Finally, with the combination of statistical tests and qualitative analysis, some empirical evidence is found to support the claim that the customized NMT is more effective than the general MT

doi:10.6342/NTU201903661

in terms of reducing the effort put forth by the professional translators, especially in dealing with terminologies.

5.2 Limitation

The limitations of this research are the domain of translation and the scale of the experiment. The corpora chosen for this research is restricted to a specific genre. On the one hand, this specificity is more beneficial when training a customized NMT. This type of text is also more flexible in terms of what would be deemed a correct translation. On the other hand, the findings in this research might not be applicable to other domains of translation or localization. Apart from that, the scale of the research and the number of samples are limited. Thus, some of the results are not statistically significant enough. Some observations are based on small amounts of data and might not remain true when a larger amount of data is introduced into the study. Moreover, as described in the previous chapters, the tools and related settings applied in this research might not be the optimum choices when different factors are considered, such as language pairs, performance and so on. The objective of this research is to make use of relatively accessible approaches and basic configurations to explore the possibility of applying a customized NMT system to localization projects of a specific domain.

5.3 Future research

This research focuses on utilizing existing and available MT and CAT to carry out an experiment for the purpose of evaluating the effectiveness of the proposed NMT. The methodology can be applied with newer and more optimized MT systems, automatic

doi:10.6342/NTU201903661

evaluation, CAT and experiment design. As the technology of MT continues to evolve, further research investigate how to validate the performance and effect of customized MT in a more comprehensive way, which may include a series of tests and experiments on properly large scale. As for translator-oriented research, it is suggested that the focus can shift to further examining the translator's user experience with MT by conducting follow-up interviews and proposing an ideal combination of tools to help professional translators enhance work efficiency.

doi:10.6342/NTU201903661

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved June 16, 2019, from https://www.laurenceanthony.net/software

Apowersoft 台灣官方網站. (2019). *Apowersoft 免費線上螢幕錄影工具*. [online] Available at: https://www.apowersoft.tw/free-online-screen-recorder [Accessed 10 Jun. 2019].

Arsham, H. (2015). *Two-Way ANOVA*. [online] Home.ubalt.edu. Available at: http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVATwo.htm [Accessed 16 Jul. 2019].

Bowker, L., & Barlow, M. (2008). A comparative evaluation of bilingual concordancers and translation. *Topics in language resources for translation and localisation*, *79*, 1.

Burberry 台灣. (2018). Available at: https://tw.burberry.com/?locale=zf_TW [Accessed 16 Jul. 2018].

Burberry 台灣. (2018). Available at: https://tw.burberry.com/?locale=en_TW [Accessed 16 Jul. 2018].

National Academy for Education Research. (2014). *NAER Segmentor*. [online] Available at: https://coct.naer.edu.tw/Segmentor/ [Accessed 10 Jun. 2018].

Gao, Z., & Chiou, S. (2017). Computer-aided translation. In C. Shei & Z. Gao (Eds.), *The Routledge Handbook of Chinese Translation*. Routledge.

García, I. (2006). Translators on translation memories: a blessing or a curse?. *Translation technology and its teaching*, 97.

doi:10.6342/NTU201903661

Garcia, I. (2014). Computer-Aided Translation. In Chan, S. (Eds.), *The Routledge Encyclopedia of Translation Technology Routledge* (pp. 68-87). Routledge.

GitHub. (2018). *OpenNMT/OpenNMT*. [online] Available at: https://github.com/OpenNMT/OpenNMT [Accessed 1 Jul. 2018].

Google Cloud. (2019). *AutoML Translation Documentation*. [online] Available at: https://cloud.google.com/translate/automl/docs/ [Accessed 13 Jul. 2019].

H&M. (2018). [online] Available at: https://www2.hm.com/en_asia3/index.html [Accessed 24 Apr. 2018].

H&M. (2018). [online] Available at: www2.hm.com/zh_asia3/index.html [Accessed 24 Apr. 2018].

Home.ubalt.edu. (2019). *Two-Way ANOVA*. [online] Available at: http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVATwo.htm [Accessed 16 Jul. 2019].

Httrack.com. (2019). *HTTrack Website Copier - Free Software Offline Browser (GNU GPL)*. [online] Available at: https://www.httrack.com [Accessed 10 Jun. 2018].

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, *17*(1-2), 5-38.

Hutchins, J. (2006). Future prospects in Machine Translation usage and research. *Presentation in February 2006 at the University of Leeds*.

Jiménez-Crespo, M. A. (2009). Conventions in localisation: a corpus study of original vs. translated web texts. *Jostrans: The Journal of Specialized Translation*, *12*, 79-102.

doi:10.6342/NTU201903661

Jiménez-Crespo, M., & Tercedor, M. (2011). Applying corpus data to define needs in web localization training. *Meta: Journal des traducteurs/Meta: Translators' Journal*, *56*(4), 998-1021.

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Kinoshita, S., Oshio, T., & Mitsuhashi, T. (2017). Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)* (pp. 140-145).

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Lavie, A. (2011). Evaluating the Output of Machine Translation Systems.

O'Hagan , M. (2009). Computer-aided translation (CAT). In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (pp. 48-51). Routledge.

OmegaT, omegat.org, 2017, omegat.org. Accessed 24 Apr. 2018.

OpenNMT - Open-Source Neural Machine Translation.. Retrieved January 10, 2018, from http://opennmt.net/

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., Henderson, J., & Reeder, F. (2002). Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and

doi:10.6342/NTU201903661

Spanish results. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 132-137). Morgan Kaufmann Publishers Inc..

Ping, K. (2009). Machine translation. In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (pp. 162-168). Routledge.

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta, 58*(3), 487–503. doi:10.7202/1025047ar.

Schäler, R. (2009). Localization. In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (pp. 48-51). Routledge.

SDL Trados Studio 2017 Professional [Computer software]. (2017). SDL Group.

Shuttleworth, M., & Lagoudaki, E. (2006). ’Translation memory systems: Technology in the service of the translation professional’. In *proceedings of 1st Athens International Conference of Translation and Interpretation*.

Stein, D. (2018). Machine translation: Past, present and future. *Language technologies for a multilingual Europe*, *4*, 5.

Translate.google.com. (2019). *Google Translate*. [online] Available at: https://translate.google.com/ [Accessed 9 Jun. 2019].

Translatum.gr. (2015). *Convert Excel files (xls, xlsx) and tab delimited txt to TMX*. [online] Available at: https://translatum.gr/cgi-bin/excel-to-tmx.pl [Accessed 10 Jul. 2018].

Vilar, D., Xu, J., d’Haro, L. F., & Ney, H. (2006, May). Error analysis of statistical machine translation output. In *Proceedings of LREC* (pp. 697-702).

Zanettin, F. (2002, May). Corpora in translation practice. *Proceedings of the First International Workshop on Language Resources (LR) for Translation Work and Research* (pp. 10-14).

doi:10.6342/NTU201903661

Zara.com. (2018). [online] Available at: https://www.zara.com/tw/ [Accessed 24 April. 2018].

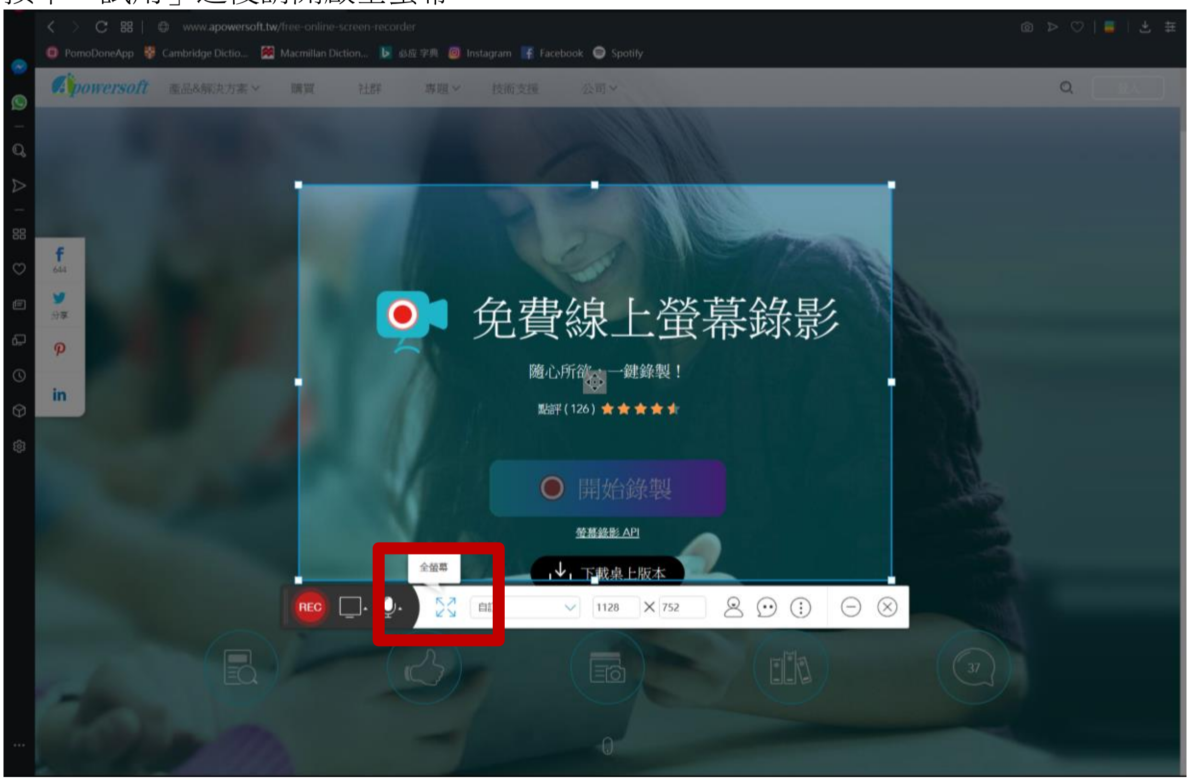
Zara.com. (2018). [online] Available at: https://www.zara.com/tw/en [Accessed 24 April. 2018].

doi:10.6342/NTU201903661

Appendices

Appendix 1: Instruction for the experiment

1. 請先安裝資料中的 apowersoft-online-launcher
2. 前往 https://www.apowersoft.tw/free-online-screen-recorder 按下開始錄製
3. 按下「試用」之後請開啟全螢幕



4. 按下「REC」正式開始錄製
5. 完成資料中的 Project 1 和 Project 2。
6. 專案附有 TM，翻譯風格和用詞皆以 TM 為準。兩個專案的 Target language 欄位中分別提供不同的機器翻譯，僅供參考，可自行判斷是否需要更改或刪除（註：Project 1 機器翻譯的字詞之間有空格可忽略不編輯）。可使用「任何」網路資源協助翻譯，但建議先查詢 TM。
7. 完成翻譯後即可停止錄影。影片檔案會儲存在「文件」資料夾下的「Apowersoft Online Screen Recorder」資料夾中。
8. 請將影片檔案、Project 1 和 Project 2 的 zh-TW 檔案回傳。
9. 填寫問卷：https://forms.gle/3zDmHUMDvpAAr9bYA。
10. 我確認以上步驟完成無誤後，就會將酬金 1000 元匯入您提供的帳戶，非常感謝您的協助！

doi:10.6342/NTU201903661

Appendix 2: Comparison for ST, NMT, Translator A and Translator B's translation in Project 1

Source Text	NMT	Translator A	Translator B
Short-sleeved top in soft jersey with a motif and a rounded hem. Longer at the back.	柔軟 平紋 短袖 上衣 ， 有 圖案 ， 圓 邊 下襬 ， 後 襬 略長 。	柔軟平紋短袖上衣，有圖案，圓邊下襬。後襬略長。	柔軟平織短袖上衣，飾有圖案和圓邊下襬。後襬較長。
Lace top with a small stand-up collar and an opening with ties at the back of the neck. Lined.	蕾絲 上衣 ， 小 立領 ， 領背 有 開口 和 綁帶 。 有 內裡 。	蕾絲上衣，小立領，領背有開口和綁帶。有內裡。	蕾絲小立領上衣，領背有綁帶開口。有襯裡。
Short-sleeved jersey top in a viscose blend with ribbing around the neckline.	短袖 平紋 上衣 ， 嫘縈 混紡 布料 ， 羅紋 領口 。	短袖平紋上衣，嫘縈混紡布料，領線飾有羅紋。	短袖平織短袖上衣，嫘縈混紡布料，羅紋領口。
Pull-on trousers in soft twill made from a cotton blend with an elasticated drawstring waist and fake fly. Side pockets, fake flap leg pockets, and elasticated hems.	柔軟 斜紋 鬆緊式 長褲 ， 棉質 混紡 布料 ， 鬆緊帶 腰圍 配 抽繩 ， 假 門襟 ， 有 側袋 和 假 後袋 ， 鬆緊帶 褲腳 。	柔軟斜紋鬆緊式長褲，棉質混紡布料，鬆緊帶褲頭配有抽繩與假門襟。有側袋和翻蓋假後袋，鬆緊帶褲腳。	柔軟斜紋鬆緊式長褲，棉質混紡布料，鬆緊帶腰圍配抽繩，假門襟。有側袋和翻蓋假後袋，鬆緊帶褲腳。
Polyester 34%; Cotton 32%; Modal 32%; Elastane 2%	聚酯 纖維 34 % ; 棉 32 % ; 莫代爾 纖維 32 % ; 彈性 纖維 2 %	聚酯纖維 34%；棉 32%；莫代爾纖維 32%；彈性纖維 2%	聚酯纖維 34 %、棉 32 %、莫代爾紗線 32 %、彈性纖維 2 %
Gently flared vest top in marled, cotton-blend jersey with deep armholes, a racer back and rounded hem. Slightly longer at the back.	略 傘狀 背心 上衣 ， 混色 棉質 平紋 布料 ， 深 袖口 ， 挖背 設計 ， 圓 邊 下襬 ， 後 襬 略 長 。	略傘狀背心上衣，混色棉質混紡平紋布料，大袖孔，挖背設計，弧形下襬。後襬略長。	混色微傘狀背心上衣，棉質混紡平紋布料，大袖口，挖背圓邊下襬。後襬略長。
Short jumper in a chunky-knit cotton blend with hole-patterned details. Ribbing	短版 套衫 ， 粗 紡 棉質 混紡 厚 織 布料 ， 帶 洞洞 設計	短版套頭衫，棉質混紡粗針織布料，有洞設計細節。領	短版套衫，棉織混紡厚織布料，帶洞洞設計細節。羅紋

doi:10.6342/NTU201903661

around the neckline, dropped shoulders and long, tapered sleeves.	細節，羅紋領口，垂肩長袖收窄。	線飾有羅紋，垂肩設計，長袖收窄。	領口，落肩設計，長袖收窄。
Long-sleeved wrapover blouse in an airy cotton weave with a wide tie at the side.	長袖交疊式女衫，透氣棉質平織布料，一側有寬綁帶。	長袖交疊式女衫，透氣棉質平織布料，一側有寬綁帶。	長袖交疊式女衫，透氣棉質平織布料，一側有寬綁帶。
Striped, 3/4-length trousers in a cotton and linen weave with an elasticated drawstring waist, side pockets and straight, wide legs.	條紋七分寬管褲，棉麻平織布料，鬆緊帶腰圍配抽繩，有側袋，寬直筒褲管。	條紋七分寬管褲，棉麻平織布料，鬆緊帶褲頭配有抽繩，有側袋，寬直筒褲管。	條紋七分寬管褲，棉麻平織布料，鬆緊帶腰圍配抽繩，有側袋和寬直筒褲管。
Coat in woven fabric with a shawl collar and welt side pockets. Detachable belt with D-rings and a tab with press-studs at the cuffs. Lined.	平織布料大衣，圍巾領，有嵌線側袋，可拆式腰帶有 D 形環，袖口有按扣扣帶。有內裡。	針織大衣，新月領，關邊側袋。可拆式腰帶附有 D 形環，袖口有按扣扣帶。有內裡。	平織布料大衣，圍巾領，西裝側袋。可拆式腰帶，飾有 D 形環，袖口有按扣扣帶。有襯裡。
Straight-style jumper in a soft, fine knit with long sleeves and ribbed cuffs.	柔軟精織套衫，直身剪裁，長袖款式，羅紋袖口。	柔軟細針織套頭衫，直筒剪裁，長袖款式，羅紋袖口。	柔軟精織套衫，直身剪裁，羅紋袖口長袖款式。
Slightly A-line top in cotton jersey with short sleeves and embroidery.	棉質平紋 A 字上衣，短袖款式，有刺繡。	棉質平紋 (微) A 字上衣，短袖款式，有刺繡。	短袖刺繡上衣，棉質平紋布料，微 A 字剪裁。
Fitted jumper in a ribbed cotton blend with a high collar and long sleeves with a slit.	合身高領套衫，棉質混紡布料，長袖款式，開衩。	合身 (羅紋) 高領套頭衫，棉質混紡布料，長袖款式，開衩設計。	合身羅紋高領套衫，棉質混紡布料，長袖開單衩。

doi:10.6342/NTU201903661

doi:10.6342/NTU201903661

Appendix 3: Comparison for ST, Google Translate, Translator A and Translator B's translation in Project 2

Source Text	Google Translate	Translator A	Translator B
Umbrella in transparent patterned plastic with a plastic handle. Length 77 cm.	透明圖案塑料傘，塑料手柄。長 77 厘米。	透明圖案塑膠傘，手柄為塑膠材質。長 77 公分。	印花透明塑膠雨傘，塑料傘柄。長 77 公分。
Straight-cut dress in woven fabric with short, open sleeves with ties. Unlined.	這款直筒連衣裙採用梭織面料製成，短袖，開口袖，繫帶。無襯裡。	這款直筒連衣裙採用梭織布料製成，短袖設計，露肩開口袖，附繫帶。無內裡。	平織布料直身短袖洋裝，（開口式）袖口有繫帶。無襯裡。
Sports top in fast-drying functional fabric with a lined front, mesh racer back and elasticated hem.	運動上衣採用快乾功能性面料製成，正面襯裡，網眼工字背和鬆緊下擺。	這款運動上衣採用快乾機能布料製成，正面襯有內裡，搭配網眼挖背和鬆緊帶下擺設計。	快乾機能布料運動上衣，正面有內裡，網部挖背設計，伸縮下擺。
5-pocket trousers in washed stretch cotton twill with an adjustable elasticated waist, zip fly and press-stud and tapered legs.	五口袋長褲，水洗彈力棉斜紋布，可調節鬆緊腰身，拉鍊門襟，按扣和錐形腿。	五口袋長褲，採用水洗彈力棉斜紋布，可調節鬆緊帶褲頭，配有拉鍊門襟與按扣，腿部為錐形剪裁。	5 袋式長褲，水洗彈性棉質斜紋布料，可調式鬆緊帶腰圍，有拉鍊門襟、按扣和褲管收窄。
T-shirt in soft jersey with a slightly wider, raw-edge neckline, raw-edge short sleeves and a seam centre back. Slightly longer and rounded at the back.	這款 T 卹採用柔軟平針織面料製成，略帶寬邊，粗邊領口，毛邊短袖和接縫中後衛。後部略長且圓潤。	這款 T 恤採用柔軟平織布料製成，領線略寬且飾有毛邊，並採毛邊短袖與背部中段縫線設計。後襬略長且成弧狀線。	柔軟平紋 T 恤，領口略寬不收邊，不收邊短袖和背面中間有縫線。圓邊後襬略長。
V-neck jumper in fine-knit cotton with long sleeves and ribbing around the neckline, cuffs and hem.	V 領棉質細針織棉質長袖，領口，袖口和下擺飾有羅紋。	細針織棉質 V 領長袖套頭衫，領線、袖口和下襬飾有羅紋。	V 領棉質精織長袖套衫，羅紋領口、袖口和下襬。

doi:10.6342/NTU201903661

Jumper knitted in a soft mohair blend with a V-neck, low dropped shoulders and long sleeves.	這款毛衣採用柔軟馬海毛混紡面料製成，V 領，低肩和長袖。	這款（針織）套頭衫採用柔軟馬海毛混紡布料製成，採 V 領、落肩和長袖設計。	V 領柔軟馬海混紡針織套衫，低垂肩長袖。
Short cardigan in soft, fine-knit cotton with buttons down the front, long sleeves and ribbing around the neckline, cuffs and hem.	這款短款開襟羊毛衫採用柔軟細針織棉質製成，正面飾有鈕扣，長袖，領口，袖口和下擺飾有羅紋。	這款短款開襟毛衣採用柔軟細針織棉質布料製成，正面飾有鈕扣，採長袖設計，領線、袖口和下擺飾有羅紋。	柔軟棉質精織長袖短版開襟衫，正面鈕扣開襟，羅紋領口、袖口和下襬。
Leggings in glossy jersey with an elasticated waist, zip fly and press-stud, and back pockets.	這款光面平針織打底褲採用鬆緊腰身，拉鍊門襟和按扣設計，後口袋設計。	這款亮面平織內搭褲採用鬆緊帶褲頭、拉鍊門襟和按扣設計，有後口袋。	亮面平紋內搭褲，鬆緊帶腰圍，拉鍊門襟配按扣，有後口袋。
V-neck top in woven fabric with lace details and narrow, adjustable shoulder straps.	V 領上衣採用編織面料製成，配有蕾絲細節和窄肩可調節肩帶。	這款 V 領上衣採用編織布料製成，飾有蕾絲細節和可調整窄版肩帶。	V 領上衣平織布料飾有蕾絲細節和可調式細肩帶。
Long-sleeved top in soft, airy viscose jersey with dropped shoulders with a small flounced trim.	這款長袖上衣採用柔軟透氣粘膠針織面料製成，肩部有垂褶，飾有小荷葉邊飾邊。	這款長袖上衣採用柔軟透氣粘膠針織面料製成，肩部有垂褶，飾有小荷葉邊飾邊。	垂肩長袖上衣，柔軟透氣嫘縈平紋布料，肩部有小荷葉飾邊。
Short 5-pocket low-rise shorts in washed denim with worn details, a zip fly and button and raw-edge hems.	這款短款五口袋低腰短褲採用水洗牛仔布製成，飾有破舊細節，拉鍊門襟和鈕扣，粗跟下擺。	這款短款五口袋低腰短褲採用水洗牛仔布製成，飾有破舊細節，拉鍊門襟和鈕扣，粗跟下擺。	5 袋式水洗丹寧超短褲，低腰剪裁，帶仿舊效果細節，拉鍊門襟配鈕扣，不收邊褲腳。

doi:10.6342/NTU201903661

Appendix 4: Comparison for translation by referenced websites, NMT and Google Translate in BLEU test

Website	NMT	Google Translate
短袖 圖案 上衣 ， 柔軟 平紋 布料 ， 圓 邊 下襬 。 後 襬 略 長 。	柔軟 平紋 短袖 上衣 ， 有 圖案 ， 圓 邊 下襬 ， 後 襬 略 長 。	這款短袖上衣採用柔軟平針織面料製成，飾有圖案和圓形下擺。在後面更長。
蕾絲 上衣 ， 小 立領 ， 領 背 有 開口 和 綁帶 。 有 內 裡 。	蕾絲 上衣 ， 小 立領 ， 領 背 有 開口 和 綁帶 。 有 內 裡 。	蕾絲上衣，有一個小立領和一個開口，後面有領帶。內襯。
短袖 平紋 上衣 ， 嫘縈 混紡 布料 ， 羅紋 領口 。	短袖 平紋 上衣 ， 嫘縈 混紡 布料 ， 羅紋 領口 。	這款短袖平針織上衣採用粘膠混紡面料，領口飾有羅紋。
柔軟 斜紋 鬆緊式 長褲 ， 棉質 混紡 布料 ， 鬆緊 帶 腰圍 配 抽繩 ， 假 門襟 ， 有 側袋 和 假 翻蓋式 褲管 袋 ， 鬆緊 帶 褲腳 。	柔軟 斜紋 鬆緊式 長褲 ， 棉質 混紡 布料 ， 鬆緊 帶 腰圍 配 抽繩 ， 假 門襟 ， 有 側袋 和 假 後 袋 ， 鬆緊 帶 褲腳 。	這款柔軟斜紋布長褲採用棉質混紡面料製成，鬆緊抽繩腰身，假蠅。側袋，假襟腿袋，鬆緊下擺。
聚酯 纖維 34 % ； 棉 32 % ； 莫代爾 纖維 32 % ； 彈性 纖維 2 %	聚酯 纖維 34 % ； 棉 32 % ； 莫代爾 纖維 32 % ； 彈性 纖維 2 %	聚酯 34％;棉 32％;莫代爾 32％;彈性纖維 2％
微 傘狀 背心 上衣 ， 混 色 棉質 混紡平紋 布料 ， 深 袖口 ， 挖 背 設計 ， 圓 邊 下襬 。 後 襬 稍 長 。	略 傘狀 背心 上衣 ， 混色 棉質 平紋 布料 ， 深 袖口 ， 挖 背 設計 ， 圓 邊 下襬 ， 後 襬 略 長 。	這款輕質喇叭背心上衣採用混合棉質混紡平針織面料製成，配有深色袖孔，工字背和圓形下擺。後面略長。
短版 套衫 ， 棉織 混紡 厚 織 布料 ， 帶 洞洞 設計 細節 ， 羅紋 領口 ， 垂肩 長袖 收窄 設計 。	短版 套衫 ， 粗 紡 棉質 混紡 厚 織 布料 ， 帶 洞洞 設計 細節 ， 羅紋 領口 ， 垂肩 長袖 收 窄 。	這款短靴採用粗針織棉質混紡面料製成，飾有孔圖案細節。領口飾有羅紋，肩部垂墜，長袖錐形。
裹身式 長袖 女衫 ， 透氣 棉質 平織 布料 ， 側邊 有 寬 綁 帶 設計 。	長袖 交疊式 女衫 ， 透氣 棉質 平織 布料 ， 一 側 有 寬 綁帶 。	這款長袖上衣採用輕盈棉質梭織面料製成，側面採用寬領帶設計。

doi:10.6342/NTU201903661

條紋 七 分 寬管褲 ， 棉麻平 織布料 ， 鬆緊 帶 腰圍 配 抽繩 ， 有 側袋 ， 直 筒 褲管 。	條紋 七 分 寬管褲 ， 棉麻平 織布料 ， 鬆緊 帶 腰圍 配 抽繩 ， 有 側袋 ， 寬 直 筒 褲管 。	這款條紋長 3/3 長褲採用棉質和亞麻編織而成，鬆緊抽繩腰身，側袋和直筒寬腿。
平織 布料 大衣 ， 圍巾 領 ， 有 嵌線 側袋 ， 可拆式 腰帶 有 D 形環 ， 袖口 有 按 扣扣 帶 。 有 內 裡 。	平織 布料 大衣 ， 圍巾 領 ， 有 嵌線 側袋 ， 可拆式 腰帶 有 D 形環 ， 袖口 有 按 扣扣 帶 。 有 內 裡 。	採用梭織面料製成，配有披肩領和貼邊側袋。帶 D 形環的可拆卸腰帶和袖口處帶按扣的拉環。內襯。
柔軟 精織 套衫 ， 直 身 剪裁 ， 長袖 款式 ， 羅紋 袖口 。	柔軟 精織 套衫 ， 直 身 剪裁 ， 長袖 款式 ， 羅紋 袖口 。	直筒式套頭衫，柔軟精細針織，長袖，羅紋袖口。
短袖 刺繡 上衣 ， 棉質 平紋 布料 ， 微 A 字 剪裁 。	棉質 平紋 A 字 上衣 ， 短袖 款式 ， 有 刺繡 。	略帶 A 字形的棉質平紋針織上衣，短袖和刺繡。
合身 羅紋 高領 套衫 ， 棉質 混紡 布料 ， 長袖 開 單 衩 。	合身 高領 套衫 ， 棉質 混紡 布料 ， 長袖 款式 ， 開衩 。	這款合身毛衣採用羅紋棉質混紡面料製成，高領和長袖設有開衩。
印花 透明 塑膠 雨傘 ， 塑膠 傘 柄 。 長 77 公分 。	印花 透明 塑膠 雨傘 ， 塑膠 傘 柄 。 長 77 公分 。	透明 圖案 塑料傘 ， 塑料 手 柄 。 長 77 厘米 。
平織 布料 洋裝 ， 直 身 剪裁 ， 開口式 短袖 ， 袖口 有 綁 帶 。 無 內裡 。	平織 布料 洋裝 ， 直 身 剪裁 ， 開口式 短袖 有 綁帶 。 無 內 裡 。	這 款 直 筒 連 衣裙 採用 梭織 面料製成 ， 短袖 ， 開口 袖 ， 繫帶 。 無 襯 裡 。
運動 上衣 ， 快 乾機 能 布料 ， 正面 有 內 裡 ， 網 布 挖 背 設計 ， 鬆緊 帶 下緣 。	運動 上衣 ， 快 乾機 能 布料 ， 正面 有 內 裡 ， 挖 背 設計 ， 鬆緊帶 下襬 。	運動 上衣 採用 快 乾 功能性 面料 製成 ， 正面 襯 裡 ， 網 眼 工字背 和 鬆緊 下 擺 。
5 袋式 水洗 彈性 棉質 斜紋 長褲 ， 可 調式 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按 扣 ， 褲管 收窄 。	5 袋式 水洗 彈性 棉質 斜紋 長褲 ， 可 調式 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按 扣 ， 褲管 收窄 。	五 口袋 長褲 ， 水洗 彈力 棉 斜紋布 ， 可 調節 鬆緊 腰身 ， 拉鍊 門襟 ， 按 扣 和 錐形腿 。

doi:10.6342/NTU201903661

柔軟 平紋 T 恤 ， 不 收 邊 領口 略 寬 ， 不 收 邊 短袖 ， 背面 中間 有 縫 線 。 微 圓 後 襬 略 長 。	柔軟 平紋 T 恤 ， 領口 略 寬 ， 不 收 邊 卷 邊 設計 ， 短袖 款式 ， 背 面 中間 有 縫 線 ， 圓 邊 後 襬 略 長 。	這 款 T 卹 採用 柔軟 平 針織 面料 製成 ， 略帶 寬 邊 ， 粗 邊 領口 ， 毛 邊 短袖 和 接縫 中 後衛 。 後部 略 長 且 圓潤 。
V 領 棉質 精 織 套衫 ， 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。	棉質 精 織 套衫 ， V 領 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。	V 領 棉質 細 針織 棉質 長袖 ， 領 口 ， 袖口 和 下 擺飾 有 羅紋 。
V 領 針織 套衫 ， 柔軟 馬海 毛 混紡 布料 ， 低垂 肩 長袖 。	柔軟 馬海 毛 混紡 針織 套衫 ， V 領 ， 低垂 肩 長袖 。	這 款 毛衣 採用 柔軟 馬海 毛 混紡 面料 製成 ， V 領 ， 低 肩 和 長 袖 。
柔軟 棉質 精 織 短 開襟衫 ， 正面 鈕扣 開襟 ， 長袖 款式 ， 羅紋 領 口 、 袖口 和 下襬 。	短版 開襟衫 ， 柔軟 精織 棉質 布 料 ， 正面 鈕扣 開襟 ， 長袖 款式 ， 羅紋 領口 、 袖口 和 下襬 。	這 款 短 款 開襟 羊毛 衫 採用 柔軟 細 針織 棉質 製成 ， 正面 飾 有 鈕 扣 ， 長袖 ， 領口 ， 袖口 和 下 擺 飾 有 羅紋 。
亮面 平紋 內搭褲 ， 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按 扣 ， 有 後 袋 。	亮面 平紋 內搭褲 ， 鬆緊 帶 腰圍 ， 拉鍊 門襟 配 按 扣 ， 有 後 袋 。	這 款 光面 平 針織 打底褲 採用 鬆緊 腰身 ， 拉鍊 門襟 和 按 扣 設計 ， 後 口袋 設計 。
平織 布料 V 領 上衣 ， 有 蕾絲 細 節 。 可 調式 細 肩帶 。	平織 布料 V 領 上衣 ， 有 蕾絲 細 節 ， 可 調式 細 肩帶 。	V 領 上衣 採用 編織 面料 製成 ， 配 有 蕾絲 細節 和 窄 肩 可 調節 肩 帶 。
垂 肩 長袖 上衣 ， 柔軟 透氣 嫘縈 平紋 布料 ， 肩部 有 小 荷葉 飾 邊 。	長袖 上衣 ， 透氣 透氣 嫘縈 平紋 布 料 ， 肩部 有 小 荷葉 飾 邊 。	這 款 長袖 上衣 採用 柔軟 透氣 粘膠 針織 面料 製成 ， 肩部 有 垂褶 ， 飾 有 小 荷葉 邊 飾 邊 。
5 袋式 水洗丹寧 超短褲 ， 低腰 剪 裁 ， 帶 仿 舊 效果 細節 ， 拉鍊 門 襟 配 鈕扣 ， 不 收 邊 褲腳 。	5 袋式 水洗丹寧 低 腰 超短褲 ， 帶 仿 舊 效果 細節 ， 拉鍊 門襟 配 鈕 扣 ， 不 收 邊 褲腳 。	這 款 短 款 五 口袋 低腰 短褲 採用 水洗 牛仔布 製成 ， 飾 有 破舊 細 節 ， 拉鍊 門襟 和 鈕扣 ， 粗 跟 下 擺 。

doi:10.6342/NTU201903661