İstanbul Üniversitesi Çeviribilim Dergisi
Istanbul University Journal of Translation Studies

İstanbul Üniversitesi Çeviribilim Dergisi - Istanbul University Journal of Translation Studies 17, 95-115
DOI: 10.26650/iujts.2022.1182687 Research Article / Araştırma Makalesi

Translation Quality Regarding Low-Resource, Custom Machine Translations: A Fine-Grained Comparative Study on Turkish-to-English Statistical and Neural Machine Translation Systems

Özel Alanlarda Düşük Kaynaklara Sahip Makine Çevirisinde Çeviri Kalitesi: Türkçeden İngilizceye İstatistiksel ve Nöral Makine Çevirisi Üzerine Ayrıntılı Bir Karşılaştırmalı Çalışma

Gökhan Doğru[1]

[1]Dr., Universitat Autònoma de Barcelona, Faculty of Translation, Barcelona, Spain

ORCID: G.D. 0000-0001-7141-2350

Corresponding author/Sorumlu yazar:
Gökhan Doğru,
Universitat Autònoma de Barcelona, Faculty of Translation, Barcelona, Spain
E-mail: gokhan.dogru@uab.cat

Submitted/Başvuru: 30.09.2022
Accepted/Kabul: 09.12.2022

Citation/Atıf: Dogru, G. (2022). Translation Quality Regarding Low-Resource, Custom Machine Translations: A Fine-Grained Comparative Study on Turkish-to-English Statistical and Neural Machine Translation Systems. *İstanbul Üniversitesi Çeviribilim Dergisi - Istanbul University Journal of Translation Studies,* 17, 95-115.
https://doi.org/10.26650/iujts.2022.1182687

ABSTRACT

Corpus-based machine translation (MT) has been the main approach to developing and implementing MT systems in both academia and the industry over the last three decades. In this field, the type and size of the corpus used for training MT engines have presented problems for both statistical MT (SMT) systems as well as neural MT (NMT) systems, being the two dominant corpus-based approaches. Moreover, language pairs such as Turkish-English have been understudied within this framework. This article aims to evaluate the translation quality in Turkish-to-English custom MT systems that have been trained on different corpus sizes and types. Two NMT engines and two SMT engines were trained on the KantanMT platform using two different training corpus types with either only domain-specific cardiology corpus or this corpus plus a mixed-domain corpus. The study conducted both automatic evaluations with metrics including BLEU, F-Measure and TER, as well as a comprehensive human evaluation with metrics including fluency, A/B test, and adequacy. Lastly, the study realized a separate, subjective terminology evaluation in order to investigate how differently MT systems handle terminology, as this is a crucial aspect for specific-domain text types such as cardiology. While the automatic evaluation results suggest the SMT engines to perform better than NMT engines, all human evaluators rated the mixed-domain NMT engine as the highest performing one. However, the terminology evaluation task demonstrated SMT to still be able to perform better and to commit less terminology errors, despite the industry and academia shifting toward NMT engines.

Keywords: Machine translation evaluation, Turkish-to-English machine translation, medical translation, neural machine translation, statistical machine translation

ÖZ

Derlem tabanlı makine çevirisi (MÇ), son otuz yılda hem akademide hem de endüstride MÇ sistemleri geliştirmek ve uygulamak konusunda ana yaklaşım olmuştur. MÇ motorlarını eğitmek için kullanılan derlemin türü ve boyutu, iki

baskın derlem tabanlı yaklaşım olan istatistiksel MÇ (İMÇ) sistemleri ve nöral MÇ (NMÇ) sistemleri için problemler ortaya çıkarmıştır. Ayrıca bu çerçevede Türkçe → İngilizce gibi dil çiftleri üzerinde yeterince çalışma yapılmamıştır. Bu makale, farklı derlem boyutu ve türü üzerinde eğitilmiş Türkçe → İngilizce, özelleştirilmiş MÇ sistemlerinde çeviri kalitesini değerlendirmeyi amaçlamaktadır. İki NMÇ motoru ve iki İMÇ motoru, yalnızca alana özgü kardiyoloji derlemi veya bu derlem artı bir karma alanlı derlem ile iki farklı MÇ eğitme derlemi türü kullanılarak KantanMT platformunda eğitildi. Hem BLEU, F-Measure ve TER gibi metriklerle otomatik değerlendirmeler, hem de akıcılık, A/B testi ve yeterlilik gibi metriklerle kapsamlı bir insan değerlendirmesi yapıldı. Son olarak, kardiyoloji gibi belirli bir alana dayalı metin türleri için çok önemli olduğundan farklı MÇ sistemlerinin terminolojiyi nasıl ele aldığını araştırmak adına ayrı, öznel bir terminoloji değerlendirmesi gerçekleştirildi. Otomatik değerlendirme sonuçları, İMÇ motorlarının NMÇ motorlarından daha iyi performans sergilediğini gösterirken, tüm insan değerlendiriciler, karma alanlı NMÇ motorunu en yüksek performanslı motor olarak değerlendirdi. Yine de terminoloji değerlendirme görevi, endüstri ve akademi NMÇ'ye doğru kaysa da İMÇ'nin yine de daha iyi performans gösterebileceğini ve daha az terminoloji hatası yapabileceğini ortaya koydu.

Anahtar kelimeler: Makine çevirisi değerlendirmesi, Türkçeden İngilizceye makine çevirisi, tıp çevirisi, nöral makine çevirisi, istatistiksel makine çevirisi

1. Introduction[1]

Neural machine translation (NMT) has been replacing statistical machine translation[2] (SMT) in the translation industry and academia since 2015. Many comparative studies (e.g., Bentivogli et al., 2016; Shterionov et al., 2018; Castilho et al., 2018) have shown significant improvements in quality be achieved by NMT engines for different language pairs such as English with German, Portuguese, and French. However, while a few studies are found to have evaluated the quality of English/Turkish NMTs with the general conviction being that NMT performs better with morphologically rich languages such as Turkish (see Ataman, 2018; Oflazer & Saraclar, 2018; Tantuğ & Adalı, 2018), not enough studies are found to have compared the quality of NMT and SMT with regard to Turkish. Hence, the strengths and weaknesses of using each system is not fully known. As of September 2022, big MT providers such as Google,[3] Microsoft, and DeepL provide Turkish NMT; however, the type and number of parallel corpora used for training these NMT systems cannot be known. Due to this lack of access, this study has designed different NMT and SMT training scenarios from scratch using different corpus types and sizes. The objective of these design scenarios is to understand how these parameters influence the automatic and human evaluation results regarding Turkish MT.

The current article aims to provide a fine-grained comparative evaluation of custom Turkish-to-English NMT and SMT systems trained on different corpus types and sizes. Section 2 briefly explains the relevant studies that have been conducted on Turkish NMT, as well as other morphologically rich languages. Section 3 provides the details of the types and sizes of the corpora, as well as the tools and methodology utilized for MT training and evaluation. Section 4 presents the results obtained from the automatic and human evaluation metrics. Section 5 discusses the results and compares them to other studies, while Section 6 concludes the article with its limitations and recommendations for possible future studies.

2. Related Works

Studies on Turkish MT are scarce and have mostly reported automatic evaluation results. An early study on English-to-Turkish SMT with 20,000 sentences reported a Bilingual Evaluation Understudy (BLEU) score of 0.0913 (El-Kahlout & Oflazer, 2006). Tyers & Alperen (2010) conducted an English-to-Turkish SMT study with 208,000 news domain sentences (SETIMES corpus) and achieved a BLEU score of 20.90. Bektaş et al. (2016) trained English-to-Turkish and Turkish-to-English SMT engines with the same corpus (again SETIMES) in addition to

1 The work herein has been adapted from the author's PhD dissertation titled "Terminological Quality Evaluation in Turkish to English Corpus-Based Machine Translation in Medical Domain" and formatted as a stand-alone article. The complete dissertation can be found at the following address: https://ddd.uab.cat/record/251732?ln=ca

2 Phrase-based statistical machine translation (PBMT) has been the dominant statistical machine translation approach, and this article uses PBMT and SMT interchangeably.

3 When Google Translate announced they were transitioning to NMT, Turkish was one of their first 7 languages to work with NMT. See: https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

different configurations using the Moses Toolkit and Turkish morphological analyzer and achieved a maximum BLEU score of 15.06 for the Turkish-to-English pair and a maximum BLEU score of 8.59 for the English-to-Turkish pair. El-Kahlout & Oflazer (2010) conducted a similar study on an English-to-Turkish SMT with 56,000 sentences from mixed domains (news texts and documents from NATO, EU, and foreign ministry sources). They obtained a BLEU score of 25.17 because they had used "a selectively segmented morphemic representation with various additional steps [including re-]ranking the 1000-best outputs" and reordering English phrases to give them a more Turkish-like morpheme structure (p. 1314). Tantuğ et al. (2008) evaluated Oflazer & El-Kahlout's (2007) SMT system with a custom BLEU metric they called BLEU+. BLEU does not perform optimally with agglutinative languages because even a minor suffix addition to a word as compared to the reference word leads to a penalization in the BLEU score. For instance, in the standard definition of BLEU score, if the MT system outputs *kitapların* [of the books] and the reference word is *kitaplar* [the books], the translation will be considered inaccurate. Tantuğ et al.'s (2008) version of BLEU takes into consideration the word roots in the process of word comparison and thus resolved what they called the "all-or-none nature of word comparison" (p. 2) regarding BLEU scoring. Their baseline BLEU score of 27.64 rose to 33.12 once it took word roots into consideration.

In a student survey on Turkish translation, Şahin (2015) reported more than 50% of the students to find the English-to-Turkish MT (Google SMT being in the context of that study) "inadequate" and "only useful for drafting." The study concluded with an expectation that better approaches to Turkish MT would be developed in the future. One year later, NMT started to gain popularity outside of academia. In November 2016, Google announced it had begun transitioning to NMT in its translation platform because NMT has been yielding better results in research. One of the first language pairs on which NMT was implemented was English-Turkish. Furthermore, Empirical Methods in Natural Language Processing (EMNLP) 2018 included English-Turkish language pair for the NMT shared translation task in the news domain, which led to more research papers on English-Turkish NMT. Burlot et al. (2018) compared the results from s shared translation tasks with BLEU scores varying between 24.84 and 48.42. Ataman (2018) conducts a study on English-Turkish NMT using both SETIMES corpus and a custom corpus of 35K sentences, from which she obtains a BLEU score of 13.77. The author also trains a multilingual engine with English, Turkish and Kurdish (adding approximately 14K sentences in English-Kurdish and Kurdish-Turkish language pairs) and achieves a slightly higher score of 13.97.

It can be observed that BLEU scores for English-Turkish language pair fluctuate considerably. This fluctuation may be due to the underlying algorithms, corpus quality, corpus size and corpus type. Furthermore, there is a necessity to compare the results using not only different automatic evaluation metrics but also human evaluation approaches as suggested in (Castilho et al., 2018).

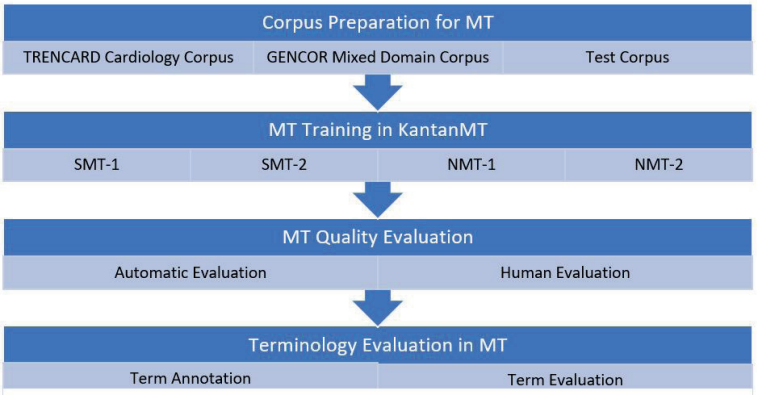

Figure 1. The workflow of the study.

3. Methodology

The study includes two training corpora: a bilingual cardiology corpus and a mixed domain corpus compiled from eight different corpora from Opus Corpus. Using these corpora, the study trains two SMT engines: one with only the cardiology corpus and one with the cardiology corpus plus the mixed domain corpus. Then I repeat the same process with the NMT and train two more engines. The study also uses the KantanMT platform for all training tasks due to it being technically less demanding and supporting both SMT and NMT training with an easy-to-use interface. It also provides the advantage of receiving automatic evaluation results immediately upon completing the training tasks. Three automatic evaluation metrics are used: BLEU, F-Measure, and translation error rate (TER). After the training, a manual evaluation was conducted with five professional translators who evaluate adequacy and fluency, and they ranked these four engines. Lastly, a term annotation and subjective binary (correct/incorrect) terminology evaluation was performed. Figure 1 summarizes the steps implemented in the study. The following subsections provide the details of the corpus statistics, KantanMT, and automatic and human evaluation methods as well as the terminology evaluation technique.

3.1. Description of the corpora

Both SMT and NMT require large amounts of parallel corpora for training. While large corpus repositories and projects such as Opus Corpus (Tiedemann, 2012) and ParaCrawl (Esplà-Gomis et al., 2019) do exist, they may not have the necessary number or type of corpora required for a certain project. Hence, custom parallel corpus preparation may be need to obtain these corpora. Using the procedure described in Dogru et al. (2018), this study has compiled a Turkish-to-English cardiology corpus from the bilingual abstracts of four cardiology-focused

scientific journals[4] in Turkey and saved this corpus in a translation memory exchange (TMX) format. Table 1 shows the corpus statistics.

Table 1. Turkish-to-English Cardiology Corpus Statistics.

Name	**TRENCARD CORPUS**
Domain	**Cardiology**
UNESCO Code	3205.01
Source Word Count	788,046
Target Word Count	907,382
Sentence Count	49,693
Source Word / Sentence Rate	15.85
Target Word / Sentence Rate	18.25

This corpus includes 788,046 source words and 49,693 source sentences; the average source sentence length is 15.85 words. This corpus has been titled "Turkish-to-English Cardiology Corpus (TRENCARD)" and has been shared on GitHub.[5] One SMT engine (SMT-1) and one NMT engine (NMT-1) have been trained using solely this domain-specific corpus.

For the second round of MT training, a mixed domain corpus has been compiled from the openly available parallel corpora in the Opus Corpus repository. This corpus includes 5.7 million source words and 381,322 sentences with an average sentence length of 14.86 words, with Table 2 presenting the corpus statistics.

Table 2. Mixed domain Turkish-to-English Corpus Statistics.

Corpus Name	**Domain**	**Source Word**	**Sentence**	**W/S**
EUBookShop	Information	482,649	22070	21.87
PHP	IT	74,042	9057	8.18
Infopanniki	Information	164,693	13173	12.50
WMT2019 News	News	197,288	10007	19.71
Ubuntu	IT	29,290	7285	4.02
KDE	IT	650,294	130731	4.97
Bianet	News	713,504	31749	22.47
Wikipedia	Information	3,356,369	157250	21.34
Total	**General**	**5,668,129**	**381322**	**14.86**

4 i) Archives of the Turkish Society of Cardiology (https://www.archivestsc.com/about-the-journal), ii) Turkish Journal of Cardiovascular Nursing (http://khd.tkd.org.tr/EN/about), iii) Turkiye Klinikleri Journal of Cardiology (https://turkiyeklinikleri.com/journal/kardiyoloji-dergisi/1300-0314/identity/en-index.html), and vi) Turkish Journal of Thoracic and Cardiovascular Surgery (http://tgkdc.dergisi.org/static.php?id=2).

5 TRENCARD Corpus and other study materials are included here (Links will be shared after journal review process):

The largest portion of this mixed domain corpus is the Wikipedia corpus[6] (Wołk&Marasek, 2014), which has sentences from Wikipedia's informative articles covering a wide variety of subjects. Thus, it is considered to be a proper sub-corpus for a mixed domain corpus. PHP,[7] Ubuntu,[8] and KDE4[9] corpora are from volunteer-translated IT projects. The average sentence lengths are comparatively small in these 3 corpora. EUBookShop[10] and Infopankki[11] have informative content, with these two cases have had their translations been conducted by professional translators. The Bianet corpus[12] (Ataman, 2018) is from a newspaper that publishes news in Turkish, English, and Kurdish. The WMT2019 News[13] corpus also includes news articles from different subjects. Considering these corpora, the terminology can be expected to be quite varied in this mixed domain corpus this study will call the General Domain Corpus (GENCOR). GENCOR and TRENCARD have been used to train the SMT-2 and NMT-2 engines.

Lastly, a test corpus has been created for use in the human and terminology evaluations. The same procedure as used for the TRENCARD corpus has been implemented to compile this corpus from scratch. Issues from the *Archives of the Turkish Society of Cardiology* that were not used for the TRENCARD corpus have been used to compile this corpus, which includes 11,015 source words and 677 sentences (ave. source sentence length = 15.40 words; Table 3).

Table 3. Corpus statistics for test corpus.

Name	**Test Corpus**
Domain	Cardiology
UNESCO Code	3205.01
Source Word Count	11015
Target Word Count	13293
Sentence Count	677
Source Word / Sentence Rate	15,40
Target Word / Sentence Rate	18,38

A sample of 100 sentences for MT human evaluation and terminological quality evaluation has been selected based on sentence length, translation accuracy,[14] and presence of cardiology terms in both the source and target sentences. Table 4 shows the type and number of corpora that have been used for training the four engines.

6 http://opus.nlpl.eu/Wikipedia-v1.0.php (last access: 26.09.2022)
7 http://opus.nlpl.eu/PHP-v1.php (last access: 26.09.2022)
8 http://opus.nlpl.eu/Ubuntu-v14.10.php (last access: 26.09.2022)
9 http://opus.nlpl.eu/KDE4-v2.php (last access: 26.09.2022)
10 http://opus.nlpl.eu/EUbookshop-v2.php (last access: 26.09.2022)
11 http://opus.nlpl.eu/infopankki-v1.php (last access: 26.09.2022)
12 http://opus.nlpl.eu/Bianet-v1.php (last access: 26.09.2022)
13 http://opus.nlpl.eu/WMT-News-v2019.php (last access: 26.09.2022)
14 Due to alignment being made automatically and a light revision being made after this operation, some sentences may still be misaligned. Moreover, the translations in some cases are observed to have been freely extended (in terms of number of words) or summarized (probably to obey word limits in abstracts).

Table 4. Overview of the training corpora and four engines.			
Engines	**Corpus Type**	**Word Count (Source)**	**Sentence Count**
SMT-1	Cardiology	788,046	49,693
SMT-2	Cardiology + Mixed Domain	6,456,175 (788,046 + 5,668,129)	431,015 (49,693 + 381,322)
NMT-1	Cardiology	788,046	49,693
NMT-2	Cardiology + Mixed Domain	6,456,175 (788,046 + 5,668,129)	431,015
Test Corpus	Cardiology	11,015	677

3.2. KantanMT: MT Training and Evaluation Platform

MT training as well as human and automatic evaluation are performed using the proprietary MT platform KantanMT.[15] Both SMT and NMT training are technically complex and resource-intensive (Pérez-Ortiz et al., 2022; Way & Hearne, 2011). KantanMT provides a user-friendly interface for non-technical users and has the same general architecture of NMT and SMT, which Shterionov et al. (2018, p. 224) describes as follows:

> *The training pipeline for both NMTand PBSMT engines follows the same architecture: 1. Instance setup hardware is allocated, software is set up: and data is downloaded; 2. Data pre-processing: data is converted to a suitable format, cleaned and partitioned for training, testing and tuning; for NMT the required dictionaries are prepared; 3. Building of models: for PBSMT, translation, language and recasing models are built; for NMT an encoder–decoder model is built; 4. Engine post-processing: the engine is evaluated, optimised and stored for future use.*

This architecture allows the user to simply configure an MT engine language pair, upload the corpus in the TMX format, and initiate the training. This study has created four engines using KantanMT's default settings for NMT and SMT. Another advantage of KantanMT is that it automatically allocates a test set from the training corpus for automatic evaluation and evaluates the translation quality based on three automatic metrics.

3.3. Automatic Evaluation Metrics

Three automatic evaluation metrics are used in the evaluation: BLEU (Papineni et al., 2002), F-Measure (Melamed et al., 2003), and TER (Snover et al., 2006) scores. These metrics compare a sample of MT outputs to human reference translations based on a specific calculation considering things such as correctly translated words, comparative sentence lengths, omissions, and word order. Each metric gives higher weight to certain parameters. According to Shterionov et al. (2018, p. 223), BLEU concentrates on the translation length, translated words, and word order, while F-Measure concerns translated words without considering word order. TER, on the other hand, aims to measure the number of edits (such as additions, omissions) necessary to transform the MT output into the human reference translation.

15 KantanMT. https://kantanmt.com/ (last access: 26.09.2022)

3.4. Human Evaluation Experiment Design

Human evaluation has been carried out by five professional translators who evaluated a sample of 100 sentences based on three parameters: ranking, adequacy, and fluency. The translators log on to the KantanLQR platform and look at a source sentence and 4 MT translations of the sentence. Adequacy and fluency are evaluated over a 5-point scale, with a 5-star rating being the highest score and 1-star being the lowest. The ranking task involves ordering MT outputs from best to worst where a 4-star rating is the best and 1-star is the worst. Ranking the outputs the same is also permissible if the MT outputs are equivalent.

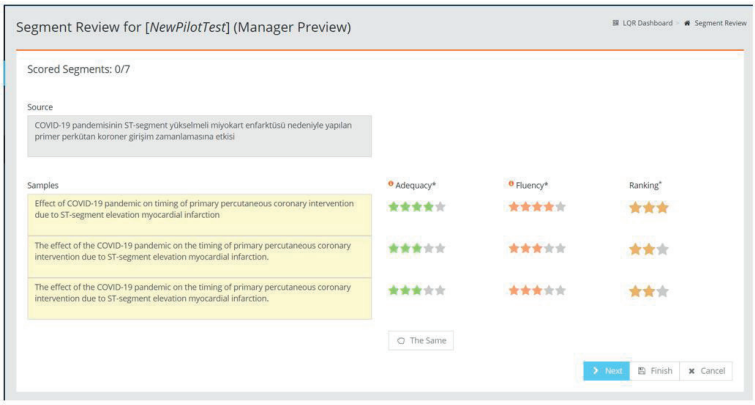

Figure 2. An example of the human evaluation screen displayed for the translators including adequacy, fluency and ranking in KantanMT for 3 different MT outputs. Note that same segments can be ranked at the same level.

Adequacy measures of how much of the meaning in the source sentence is expressed in the target sentences; in other words, it is a measure of accuracy. The translators give five stars when the meaning of the source sentence is expressed completely in the MT output (more stars = better adequacy). Fluency measures the grammaticality and readability of a sentence and focuses more on stylistic aspects. Translators give 5 stars when the target sentence is a fluent sentence in the target language, has no grammatical error, and no problems present in the syntactic structure.

3.5. Terminology Error Annotation and Binary Evaluation

Terminological accuracy plays a crucial role in domain-specific translations such medical translation. The study conducts a binary terminology evaluation (correct/incorrect) of the outputs from the four MT engines. For this evaluation task, a corpus of 100 sentences with at least one cardiology term have been filtered from the test corpus described above. Each cardiological term is then annotated onto a spreadsheet both in the source and target sides. Afterwards, each MT output is compared to the reference translations only to confirm whether the term has been translated correctly or not. Terminological variation (the use of a term translation not equivalent

to the one in the reference corpus) is permitted. The following section reports the results from the automatic and human evaluations, as well as the subjective terminology evaluation.

4. Results

4.1. Automatic Evaluation of TR → EN Specific and Mixed Domain MT Engines

Automatic evaluation metrics (AEM) help one rapidly gain insight into the translation quality of an MT engine and compare different engines by subjecting all of them to exactly the same criteria. These aspects are especially important during the MT development phases, as they require many iterations of quality evaluations, which for repetitive human evaluations might become slow, subjective, and costly. Despite the growing number of criticisms about the effectiviness of these evaluation methods (Way, 2018), they are still widely used in MT studies and paired with human evaluation tasks for increased confidence about the overall quality. This study uses three different evaluation metrics, all of which make a sentence-level comparison between a reference sentence and a machine-translated sentence. A summary of the automatic evaluation scores for the four engines is given in Figure 3.

KantanMT reserves 500 sentences from the training corpora automatically, and these sentences are translated by each engine once the training is done. Then the machine translation outputs are compared to the reference human translations, and the F-Measure, BLEU, and TER scores are calculated based on these 500 sentences. Note that this randomly selected test corpus of 500 sentences is different for each engine.

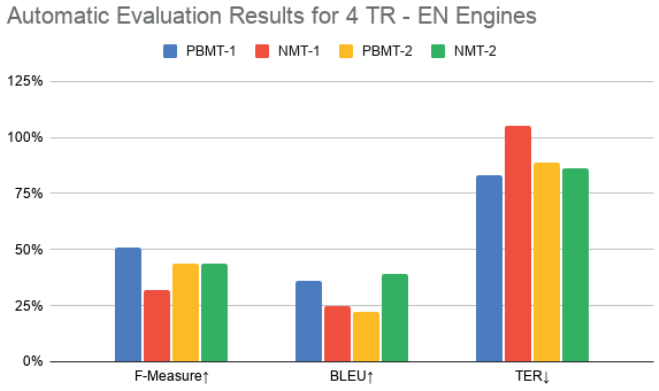

Figure 3. Summary of the automatic evaluation results for the four engines: SMT-1, NMT-1, SMT-2 and NMT-2.[16]

16 Note that while higher score means better translation quality in F-Measure and BLEU, low score means better translation quality in TER. SMT-1 has the highest F-Measure score, NMT-2 has the highest BLEU score and again SMT-1 has the best TER score. TER is calculated based on the ratio between the number of edits (i.e., additions, deletions and substitutions) and the number of words in the reference translation. If the number of edits is more than the number of words in the reference translation, the TER ratio is going to be bigger than 1 and hence the percentage score will be above 100%. That's why NMT-1 has a percentage higher than 100%.

This section will compare the engines, investigate the evolution of the quality based on MT type (i.e., SMT vs. NMT) and corpus size, and determine the best and poorest performing engines.

When considering the overall scores, SMT-1 has the highest F-Measure score with 51%, and NMT-1 has the lowest F-Measure score with 32%. In terms of the BLEU score, NMT-2 is the best performing engine, while SMT-2 is the worst. Finally, SMT-1 again has the highest TER score, while NMT-1 has the lowest. While no one single engine stands out, SMT-1 and NMT-2 can be argued to have performed better than the other two engines according to the automatic metrics. Hence, in terms of MT system type, no significant difference appears to exist in the context of the Turkish-to-English automatic evaluation scores. However, a fine-grained look at the change in corpus size in the MT systems does provide an interesting insight.

Table 5. Automatic evaluation scores for four engines. Best scores are shown in bold.

Engine Name	F-Measure↑	BLEU↑	TER↓
SMT-1	**51%**	36%	**83%**
NMT-1	32%	25%	105%
SMT-2	44%	22%	89%
NMT-2	44%	**39%**	86%

Koehn and Knowles' (2017) study involved incremental training data, and they reported their NMT engine to follow a steeper learning curve compared to their SMT engine. The current study may also expect similar conduct from its engines. Moreover, discussions have also occurred regarding the reliability of these evaluation metrics for measuring the quality of NMT. In a similar study comparing NMT and SMT in terms of the F-Measure, BLEU, and TER for five language pairs, Shterinov et al. (2018, p. 8) hypothesized. "[…] F-measure, BLEU, and TER underestimate the quality of NMT systems" and reported that, while SMT engines yield better automatic scores, the results from the human reviewers indicate the opposite, having given higher scores to the NMT engines. This shows human evaluation to be necessary for a more complete view of the performance of an MT engine. The following section will report on the human evaluation task that was perform for deciding which engine is the best performing engine with regard to the Turkish-to-English language pair.

4.2. Human Evaluation Results

Human evaluations of the four MT engines were conducted on the KantanLQR platform by five professional native Turkish translators. The quality evaluation was performed based on three metrics: adequacy, fluency, and overall ranking.

The four different MT engines translated 100 Turkish segments into English. On the KantanLQR dashboard, an A/B Test project was created alongside the adequacy and fluency evaluations as additional key performance indicators (KPIs). Then, the 100 translated sentences from each engine were imported into the project, and the five translators were invited to connect

to the dashboard to perform the evaluation. All the translators evaluated the same sentences without knowing from which engine the translations had been derived. In each window, the translators looked at a source sentence, four different translations of this sentence (randomly ordered), and the scales for ranking, adequacy, and fluency. The translators were allowed to assign the same score when the quality was the same for two or more segments. Once they completed evaluating all the sentences, the overall scores and a detailed analysis of the results appeared in an analytics dashboard. The study will now firstly describe the profiles of the task participants and then report the findings from this human evaluation task.

Before starting the evaluation task, the evaluators filled out a survey form related to their professional background, with five native Turkish evaluators participating in the study. When considering their educational background, 80% have completed undergraduate studies, and 20% have also completed a master's degree. When looking at their professional profiles, all participants reported performing both translation and review services.

Their translation experience varied between less than one year of experience to 5-10 years of experience, with 60% reported having 3-5 years of experience. Similarly, 60% report using a machine translation in their daily workflow.

Most evaluators (80%) also reported performing postediting tasks. Lastly, since the test sample had been derived from a cardiology corpus, the evaluators were asked whether they have had experience with medical translation, to which 60% stated having previously provided medical translation, editing, and/or postediting services.

This section will now report on the results from the human evaluation task. The five human reviewers completed the evaluation between November 5-12, 2020. Each reviewer evaluated 100 sentences in terms of adequacy and fluency and ranked them from best to worst. With regard to all the evaluation parameters, the NMT-2 engine achieved the best score. The next paragraphs describe the results for each evaluation type.

Ranking. The NMT-2 engine received the best score with a rating of 61.8%, while NMT-1 ranked as the worst engine with a rating of 32.8%. The scores for SMT-1 and SMT-2 are 42.65% and 45.75%, respectively.[17] The SMT-2 engine ranked second, and the SMT-1 engine ranked third.

Table 6. Ranking scores for four engines.

Engine Name	Total Score	Percentage
SMT-1	853/2000	42.65%
SMT-2	915/2000	45.75%
NMT-1	656/2000	**32.8%**
NMT-2	1236/2000	**61.8%**

17 The calculation for this percentage is provided here. There are 100 sentences and five reviewers, and the scores for each sentence vary between 1-4. Hence, if each reviewer gives 4 points to each segment, the highest possible score for an engine is 2,000. The percentages are calculated according to total score of each engine as a percentage of the highest possible score.

Table 6 shows the total ranking score each engine received from the five reviewers. The NMT-2 engine, which was trained on both cardiology and mixed-domain corpora, received a significantly higher score compared to the other three engines. On the other hand, NMT-1 engine, which was trained on only cardiology corpora, received a significantly lower score. The results for each engine are presented in the following starting from the worst engine to the best.

The NMT-1 engine received the lowest score 407 times (81.4%) out of 500 scoring instances.18 It received the best score only 19 times (3.8%). This low score may be due to the low amount of training data, as having strictly specific domain data does not help provide high quality results.

The SMT-1 engine ranked third and was the other engine trained with only cardiology corpora. This engine also ranked very low, ranking the worst a total of 317 times (63.4%) and only ranking the best 54 times (10.4%). These two results above show that the reviewers gave lower rankings to both engines with the fewest resources.

The SMT-2 engine ranked second and was trained with cardiology as well as mixed-domain corpora. This engine received the lowest-ranking score 298 times (59.6%) and the highest-ranking score 67 times (13.4%). These scores are slightly better than those for SMT-1. However, they are still significantly lower than those for NMT-2.

The NMT-2 engine was trained on cardiology and mixed-domain corpora using an NMT system. It received the best overall ranking score. The reviewers selected it as the best engine 167 times (33.4%) and as the worst engine 163 times (32.6%). Note that the ranking task allows the same ranking score to be applied to multiple engines, as well as no ranking to be assigned to any of the four engines.

Lastly, the study will look at the preferences of the reviewers based on their scores as calculated by KantanLQR. Figure 4 shows the percentage distribution of the scores given by each reviewer. Each color represents a different reviewer. The four reviewers can be seen to have given their highest scores to the NMT-2 engine, with only one reviewer (noted in pale pink) giving a slightly higher score to the SMT-2 engine compared to NMT-2 (78% to SMT-2 and 77% to NMT-2). When looking at the lowest ranking engine, the four reviewers again ranked NMT-1 as the lowest performing engine, with only one reviewer (noted in dark brown) giving a slightly lower score to the SMT-1 and SMT-2 engines (30% to NMT-1 and SMT-2, and 29% to SMT-1).

18 100 sentences scored by 5 translators equal to 500 scoring instances.

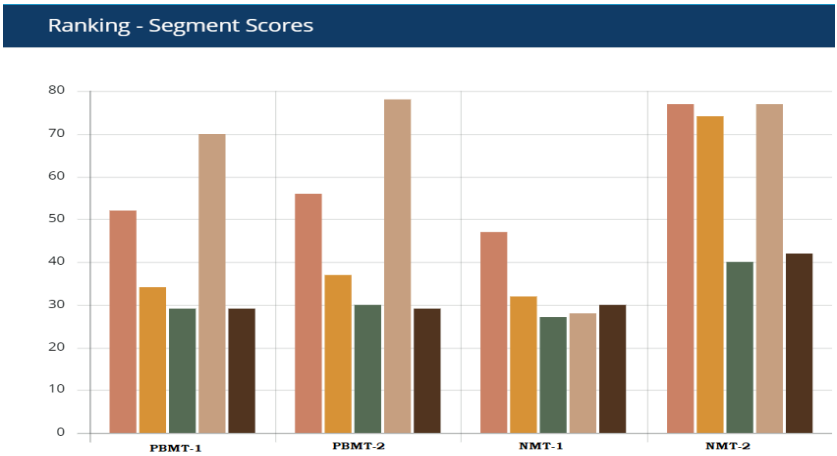

Figure 4. Ranking scores given by each reviewer and shown as percentages. Each color represents a reviewer.

The ranking task showed the reviewers to have ranked the segments coming from the engines trained with greater amounts of data higher, irrespective of engine type. One can also infer from this evaluation that increasing the amount of training data for NMT systems moved the NMT engines from the worst rank to the best and made their results more preferable by reviewers.

Adequacy. The ranking task provided an overview about the MT engines' performance from the perspectives of human reviewers. For a fine-grained analysis, the reviewers were requested to rate the adequacy and fluency of the engines on a 5-point scale regarding the same tasks. Adequacy "is typically defined as the extent to which the translation transfers the meaning of the source-language unit into the target" (Castilho et al. 2018, p. 18). In this study's operalizationalized setting, the human reviwers assigned the segments a score between 1-5, with a score of 1 expressing none of the meaning to have been expressed in the target sentence while a score of 5 meant all the meaning had been expressed in the target sentence.

NMT-2 ranked highest here with an average adequacy score of 2.73 (54.64%).[19] NMT-1 ranked lowest with an average adequacy score of 1.61 (32.24%). SMT-1 and SMT-2 received average scores of 1.90 (38.16%) and 1.99 (39.92%), respectively.

19 The five reviewers assigned adequacy scores on a 5-point scale for the 100 sentences. Hence, the maximum numerical adequacy score that an engine could get is 2,500. The percentages for the four engines were calculated in proportion to this maximum score. For example, NMT-1 received a total score of 1,366, with (1,366 * 100) / 2500 giving the adequacy percentage.

Table 7. Overall adequacy scores and their percentage distribution.		
Engine Name	**Total Score**	**Percentage**
SMT-1	954/2500	38.16%
SMT-2	998/2500	39.92%
NMT-1	806/2500	**32.24%**
NMT-2	1366/2500	**54.64%**

None of the engines were observed to have a score greater than 4 (which means "much of the source segment meaning has been expressed in the target").

Of all the scores given to NMT-2, 7.6% were 5-star, 29.2% were 4-star, 16.6% were 3-star, 22.0% were 2-star, and 24.6% were 1-star ratings. The NMT-2 engine is followed by the SMT-2 engine, of whose scores only 2.6% were 5-star, 12.0% were 4-star, 15.8% were 3-star, 21.8% were 2-star, and 48.0% were 1-star rankings. The most striking score in this engine was 48.0% of its output being ranked as 1-star, which implies that nearly half of the segments had not expressed the meaning in the source segments.

Similar to the SMT-2 engine, the SMT-1 engine also had slightly more than half of its segments receive a 1-star ranking (50.8%), while 2.6% were 5-star (same as SMT-2), 10.8% were 4-star, 12.3% were 3-star, and 23.6% were 2-star rankings.

Lastly, NMT-1 ranked worst with its scores. The human reviewers gave a daunting 68.4% of its output a 1-star ranking, while only 1.0% had a 5-star, 9.8% had a 4-star, 7.0% had a 3-star, and 13.8% had a 2-star ranking.

The difference between the NMT-2's 1-star rankings (24.6%) with NMT-1's 1-star rankings (68.4%) is remarkable. When considering how these two engines varied in terms of the amount of bilingual training corpora, one can conclude that increasing the amount of training data significantly improves the adequacy of NMT systems. When looking at the difference between the SMT-2 and SMT-1 engines, the percentage distribution of their scores was very similar.

Fluency. The fluency tasks focused on the target sentences and considered how well the target language grammar rules had been followed, how good the word choice was, and how appropriate the grammatical structure was. On the KantanLQR scale, the highest score is five stars, which indicates native language fluency, no grammatical errors, good word choices, and proper syntactic structure with no post-editing being required. The lowest score of one star indicates no fluency, complete lack of grammatical structure, and mostly making no sense, with the translation needing to be redone from scratch. Table 8 shows the fluency scores for each engine.

Table 8. Total fluency scores of each engine. Highest and lowest percentages are shown in bold.		
Engine Name	**Total Score**	**Percentage**
SMT-1	913/2500	**36.52%**
SMT-2	959/2500	38.36%
NMT-1	1031/2500	41.24%
NMT-2	1428/2500	**57.12%**

The NMT engines can be observed to have received higher scores compared to the SMT engines, with the NMT-2 engine being the most fluent and SMT-1 the least. The percentage distributions of each engine are reported below.

Although the NMT-2 engine had the highest fluency score, it had a very low percentage of 5-star rankings (8.4%). This implies the number of target sentences with native target language fluency to have been very low. However, the NMT-2 engine had a significantly lower number of target sentences with 1-star rankings (20.8%) compared to the other three engines (46.6% in NMT-1, 49.4% in SMT-2, and 54.0% in SMT-1). Unlike the ranking and adequacy tasks, NMT-1 ranked second for the fluency task, with 1.4% being 5-star rankings, while having more 4-star and 3-star ranking than SMT-2. SMT-2 ranked third with 2/8% being 5-star and 49.45 being 1-star rankings. While the percentages of this engine are close to NMT-1, SMT-2 showed slightly less fluency. The least fluent engine was SMT-1, with more than half of its target sentences (54%) receiving a 1-star ranking. However, SMT-1 had the same percentage of 5-star rankings as SMT-2.

4.3. Final Remarks

In all the human evaluation tasks, NMT-2 can be observed to have performed the best compared to the other engines. Although NMT-2 had mixed-domain corpora, it performed better than the SMT-1 and NMT-1 engines, which had only specific domain (cardiology-based) corpora. Furthermore, NMT-2 performed better than both SMT-1 and SMT-2.

Table 9. An overview of percentages of the human evaluation scores.			
Engine Name	**Ranking**	**Adequacy**	**Fluency**
SMT-1	42.65%	38.16%	**36.52%**
SMT-2	45.75%	39.92%	38.36%
NMT-1	**32.8%**	**32.24%**	41.24%
NMT-2	**61.8%**	**54.64%**	**57.12%**

A few observations in relation to corpus size, system type, and corpus type can be made based on the human evaluations. Concerning adequacy in the context of the customized Turkish-to-

English MT, corpus size appears to be more important than system type and corpus type, as the NMT-2 and SMT-2 engines performed better than the other two engines. Concerning fluency in the same context, system type is more important than corpus size and corpus type, as both NMT engines performed better than both SMT engines. Still, NMT-2 performed better than NMT-1, and SMT-2 performed better than SMT-1, which implies a more voluminous corpus size to lead to better fluency. While the overall rankings were compatible according to the adequacy results, more corpora were seen to lead to better scores, with NMT-2 ranking best, followed by SMT-2.

4.4. Discussion of the MT Evaluation Results

Having conducted automatic and human evaluations, the study can now make overall observations about the evaluation results and discuss them in comparison to other studies.

The first observation is that automatic evaluation scores do not correlate with human evaluation scores in terms of best performing engine. SMT-1 engine was the best performing engine in terms of the F-Measure and TER scores and the second-best performing engine in terms of the BLEU score. However, the SMT-1 engine placed third regarding the ranking and adequacy scores and worst in terms of fluency. The best performing engine out of all the human evaluation tasks was the NMT-2 engine. This shows the study's finding to be compatible with Shterionov et al.'s (2018) hypothesis claiming automatic evaluation scores to underestimate the actual quality of NMT engines. They trained SMT and NMT engines for five language pairs and evaluated the MT quality with human and automatic evaluation metrics. However, their study was conducted with engines that had a corpora size of over 35 million words, which is significantly larger than the current study's corpora. This study's NMT-1 engine had a low volume of corpora and received low scores from both the human and automatic evaluations. For this reason, an exception should be added to Shterionov et al.'s hypothesis: automatic scores correlate with human evaluation scores when the NMT engine is trained with a small volume of corpora (at least in the context of Turkish-to-English MT).

Having the NMT system with more parallel corpora in this study be the best and the NMT system with fewer parallel corpora as the worst indicates that NMT systems are very sensitive to the amount of training data. Koehn & Knowles (2017) trained SMT and NMT engines with different amounts of training data and observed the quality of NMT to follow a steeper curve compared to the SMT in terms of BLEU scores. Both the human and automatic evaluation scores in the current study confirm this observation for the Turkish-to-English MT. When considering the evolution from NMT-1 to NMT-2, the ranking, adequacy, and fluency scores are seen to have respectively increased by 29%, 22%, and 15.88%. An improvement also occurred in the SMT when increasing the corpus size; however, the increase was quite low compared to the NMT's, with SMT ranking, adequacy, and fluency respectively increasing by 3.10%, 1.76%, and 1.84%.

One expectation (i.e., implicit hypothesis) of the study was that engines with strictly narrow domain corpora would have performed better than those with mixed-domain corpora.

However, at least according to the human evaluation scores, this was not the case. Moreover, this study's specific domain engines (SMT-1 and NMT-1) had significantly lower volumes of corpora than the domain engines with mixed-domain corpora. Hence, this study cannot arrive at a final conclusion about the effect of corpus type except for when both scenarios have engines with the same volume of parallel corpora. In other words, this study recommends conducting a future study with the same volume of specific-domain parallel corpora and mixed-domain parallel corpora with regard to the Turkish-English language pair.

4.5. Terminology Annotation and Evaluation Results

During the sentence-by-sentence term annotation process, 231 cardiology-related term pairs were identified. Also, 35 source terms occurred in more than one sentence; hence, the total count of unique (with regard to morphological form, not conceptual meaning) terms was 196. Aside from the 231 terms, 67 acronyms were also identified. Hence, 298 terminological units were studied in total. This study will publish the annotated source and target sentences as a free and open corpus in an open repository[20] as research material for any terminology evaluation in Turkish-to-English MT for use by MT researchers. The 100 sentences including these terms have been translated by the four engines, and the subjective evaluation of the term translations as done by the author of the study was conducted by comparing the MT outputs with human translations. Term translations have been annotated as "correct" or "incorrect." The table 10 shows the percentage of correct and incorrect term translations from the four engines.

Table 10. Percentage of Correct and Incorrect Term Translations by the Four Engines

	Correct Term Translation	Incorrect Term Translation
SMT-1	68.79%	31.20%
SMT-2	70.13%	29.86%
NMT-1	16.77%	83.22%
NMT-2	62.08%	37.91%

When considering the overall terminology evaluation, the SMT-2 engine is seen to have the highest number of correct term translations, with 209 correct term instances, followed by SMT-1 with 204 instances. The NMT-2 engine came in third place with 185 instances, followed in last place by NMT-1 with only 85 correct term translations. In parallel with these results, when considering the term translation errors, NMT-1 had the highest number of errors, with 248 term translation errors. This was followed by NMT-2 with 113 term translation errors, SMT-1 with 94 errors, and SMT-2 in last with 89 errors. In reference to these results, the SMT

20 Turkish English Parallel Corpora and MT Evaluation Results. (Links will be shared after journal review process) (last access: 30.09.2022)

engines can be argued to commit less term translation errors compared to the NMT engines regarding Turkish-to-English corpus-based MT, with SMT-2 performing the best and NMT-1 performing the worst with respect to term translation.

5. Conclusion

Using the specific-domain cardiology corpora, this study trained one SMT engine (SMT-1) and one NMT engine (NMT-1). In this very narrow domain scenario, SMT performed significantly better than the NMT engine with regard to the automatic evaluations. The second scenario involved mixed domain corpora being added to the training set and one more SMT engine (SMT-2) and one more NMT engine (NMT-2) being trained. In this second scenario, the SMT quality decreased while the NMT quality significantly improved with regard to all the automatic metrics. According to the human evaluations for the cardiology sample set, the change from SMT-1 with specific-domain corpora to SMT-2 with mixed-domain corpora resulted in a slight improvement. However, the change from NMT-1 with specific-domain corpora to NMT-2 with mixed-domain corpora resulted in a very significant improvement. In fact, the NMT-2 engine performed the best in terms of ranking, adequacy, and fluency metrics. Collectively, these results imply NMT to have the potential to perform better when translating specific domain content with an engine trained on mixed-domain corpora. Nevertheless, when a low volume of specific corpora is available, SMT may still perform better than NMT, at least in the case of Turkish-to-English MT. When looking at the terminology evaluation, the study observed a slightly different behaviour, with the SMT engines committing fewer terminology errors than the NMT engines. While the change from NMT-1 to NMT-2 significantly decreased the amount of terminology errors, the amount of terminology errors was still greater than those of either SMT-1 or SMT-2.

This study has many limitations due to the decisions taken throughout the process. The objective has been to be able to control all the steps of the MT training process from a translation studies perspective, and this has advantages and disadvantages. While preparing cardiology corpora from scratch instead of benefitting from readily available and open corpora is beneficial for the study and the research community in general, this was time consuming at first, and due to the domain of cardiology being quite narrow, creating a parallel corpora larger than 1 millon source words was impossible. This limited the size of the training corpora used in the specific-domain MT trainings. In the future, I would like to create a less narrow, medical-parallel corpora for large scale medical MT training. Also, one considerable limitation of the study is the subjective terminology evaluation. The use of reference human translations as well as reference terminology resouces minimized subjectivity; yet, the terminology error annotations could have been performed by other human evaluators, just like in the general human evaluation of the MT engines. However, due to four engines needing to be evaluated, no user friendly GUI being available for

terminology evaluation or annotation, and term error annotation task being a complex task involving spreadsheets, I decided to conduct a subjective evaluation. In the future, I aim to benefit from automatic and human terminology evaluation methods for quickly analyzing MT engines' terminology translation qualities.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The author has no conflict of interest to declare.
Grant Support: The work herein is adapted from the PhD study of the author and formatted as a stand-alone article. This work is supported by the European Union-NextGenerationEU grant in the framework of Margarita Salas Postdoctoral Grant. I would like to thank Anna Aguilar Amat and Adria Martin Mor for their invaluable comments during my PhD work and Mr. Joss Moorkens for his feedback during my research visit to Dublin City University as a postdoctoral researcher.

REFERENCES

Ataman, D. (2018). Bianet: A Parallel News Corpus in Turkish, Kurdish and English. Proceedings of the LREC Workshop MLP-Moment, (pp. 14-17).

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, (pp. 257–267). doi:10.18653/v1/D16-1025

Burlot, F., Scherrer, Y., Ravishankar, V., Bojar, O., Gronroos, S.-A., Koponen, M., . . . Yvon, F. (2018). The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. Proceedings of the Third Conference on Machine Translation: Shared Task Papers, (pp. 546–560).

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. (. Doherty (Eds.), Translation Quality Assessment (pp. 9-38). Springer.

Castilho, S., Moorkens, J., & Gaspari, F. e. (2018). Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems. Machine Translation, 32, 255–278. doi:https://doi.org/10.1007/s10590-018-9221-y

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics, 108(1), pp. 109-120. doi:https://doi.org/10.1515/pralin-2017-0013

Doğru, G., Martín-Mor, A., & Aguilar-Amat, A. (2018). Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish-to-English Cardiology Corpora. Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing", (pp. 12 - 15). Miyazaki.

El-Kahlout, İ. D., & Oflazer, K. (2006). Initial Explorations in English → Turkish Statistical Machine Translation. Proceedings of the Workshop on Statistical Machine Translation, (pp. 7-14).

El-Kahlout, İ. D., & Oflazer, K. (2010). Exploiting Morphology and Local Word Reordering in English → Turkish Phrase-based Statistical Machine Translation. IEEE Transactions on Audio, Speech and Language Processing, 1313-1322.

Esplà-Gomis, M., Forcada, M. L., Ramírez-Sánchez, G., & Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. Proceedings of MT Summit XVII, volume 2, (pp. 118 - 119).

Forcada, M. L. (2010). Machine translation today. In Y. Gambier, & L. Doorslaer (Eds.), Handbook of Translation Studies (pp. 215-223). Amsterdam and Philadelphia: John Benjamins.

Koehn, P. (2020). Neural Machine Translation. Cambridge University Press.

Lumeras, M., & Way, A. (2017). On the Complementarity between Human Translators and Machine Translation. HERMES - Journal of Language and Communication in Business(56), 21-42. doi:https://doi.org/10.7146/hjlcb.v0i56.97200

Melamed, I., Green, R., & Turian, J. (2003). Precision and recall of machine translation. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, (pp. 61-63). Edmonton.

Oflazer, K., & El-Kahlout, I. D. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. Proceedings of the Second Workshop on Statistical Machine Translation, (pp. 25-32). Prague.

Oflazer, K., & Saraclar, M. (2018). Turkish Natural Language Processing. Springer International Publishing.

Papineni, K., Roukos, S., Ward, T., & J, Z. W. (2002). BLEU: a method for automatic evaluation of machine. Proceedings of the 40th annual meeting on association for computational linguistics, (pp. 311–318). Philadelphia, Pennsylvania, USA.

Pérez-Ortiz, J. A., Forcada, M. L., & Sánchez-Martínez, F. (2022). How neural machine translation works. In D. Kenny, Machine translation for everyone: Empowering users in the age of artificial intelligence (pp. 141-164). Dublin: Language Science Press.

Şahin, M. (2015). Çevirmen Adaylarının Gözünden İngilizce- Türkçe Bilgisayar Çevirisi ve Bilgisayar Destekli Çeviri: Google Deneyi. Çeviribilim ve Uygulamaları Dergisi, Journal of Translation Studies(21), 43-60.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. Machine Translation, 32, 217–235. doi:https://doi.org/10.1007/s10590-018-9220-z

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. 2006. Proceedings of the 7th conference of the association for machine translation of the Americas. Visions for the future of machine translation, (pp. 223-231). Cambridge,Massachusetts,.

Tantuğ, A. C., & Adalı, E. (2018). Machine Translation Between Turkic Languages. In K. Oflazer, & M. (. Saraçlar, Turkish Natural Language Processing. Springer International Publishing.

Tantuğ, A. C., Oflazer, K., & El-Kahlout, I. D. (2008). BLEU+: a Tool for Fine-Grained BLEU Computation. Proceedings of the International Conference on Language Resources and Evaluation, LREC.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), (pp. 2214 - 2218).

Tyers, F. M., & Alperen, M. S. (2010). South-east European Times: A parallel corpus of Balkan Languages. Proceedings of LREC 2010, Seventh International Conference on Language Resources and Evaluation. Retrieved from http://nlp.ffzg.hr/resources/corpora/setimes/

Way, A. (2018). Quality Expectations of Machine Translation. In S. Castilho, J. Moorkens, & F. &. Gaspari (Eds.), Translation Quality Assessment: From Principles to Practice (pp. 159-178). Springer.

Way, A., & Hearne, M. (2011). On the Role of Translations in State-of-the-Art Statistical Machine Translation. Language and Linguistics Compass, 5/5, 227–248.

Wołk, K., & Marasek, K. (2014). Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs. Procedia Technology, 18, 126-132.