universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Qualitative evaluation of machine translation training data discarded by open-source automatic data cleaning tools"

verfasst von / submitted by

Rujuta Makarand Dixit

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / Degree programme code as it appears on the student record sheet:	UA 070 342 331
Studienrichtung lt. Studienblatt / Degree programme as it appears on the student record sheet	Masterstudium Translation Englisch Deutsch
Betreut von / Supervisor:	Univ.-Prof. Dragoș Ioan Ciobanu, PhD

Acknowledgements

I consider myself very fortunate to have been able to write an entire thesis on a topic that interests me so much. I have always loved languages and always dreamed of pursuing a career in them. Writing this thesis is a big step in that direction, but it would not have been possible without the people who have been my support system throughout this journey.

First and foremost, I would like to thank my supervisor, Professor Dragoș Ioan Ciobanu, for giving me the wonderful opportunity to write this thesis and for the best guidance and support. I am thankful for all the help and advice he gave me, starting from choosing the topic to shaping up the entire thesis. It would not been possible without him. I would also like to thank the University of Vienna and all my other professors who contributed so positively to my entire learning experience.

I would also like to thank Professor Heide Maria Scheidl, whose unwavering support and encouragement got me through some really tough examinations.

I am also grateful to Professor Alexandra Krause for supporting me through my internship and also to Mag.Trisha Kovacic-Young for giving me the opportunity to work as a translation intern at her firm, Young Translations LLC.

A special thanks also to Gema Ramírez-Sánchez, CEO of Prompsit, who very kindly and patiently answered all the questions I had regarding some technicalities in my thesis.

My family away from home and my closest friends, Sonali, Gaurav, Nadja, Marilena, Vineet, Janhavi and Martina, who not only continuously helped me throughout my studies and believed in me when I did not believe in myself, but also made my stay in Vienna so pleasant and so comfortable. Thank you from the bottom of my heart, this entire journey could not have been possible without you.

My dear friend, Janeesh, who is not only a good friend but also a fellow translation student, has been really helpful. Not only did she motivate me to keep going when I was down, she also took the time to proofread my entire thesis. Thank you so much.

To my friends back home in India, Bhavesh, Ishita, Raksha, Nimisha and Nakul, you have been my rock and phone calls with you have saved the day many a time. Thank you so much for being my sounding board and for being calm when I was not.

To my best friend, Kruti, thank you so much for listening to my whining and rants, and taking care of me even from afar.

This list of acknowledgements would not be complete without conveying my thanks to my biggest source of strength, my family. My parents and grandparents have been so wonderful and stood by me every step of the way, and I cannot thank them enough. It is thanks to them that I was able to pursue this degree programme and push through the toughest of times. I hope I did them proud.

Qualitative evaluation of machine translation training data discarded by open-source automatic data cleaning tools

Abstract

Manual cleaning of training data for machine translation systems can be an expensive and time-consuming process. High-quality training data is a must for any machine translation system to produce satisfactory outputs. This research project aims to check the cleaning accuracy of two open-source automatic data cleaning tools, Bicleaner and Moses. Training data is needed for both statistical and neural machine translation systems, but this research project focuses on the latter. A large English-German corpus was cleaned using Bicleaner, which comprises the cleaning algorithms of Moses. I then analysed the data discarded from this corpus for any useful data. As well as the features of the useful data, if any, were analysed. Furthermore, the cleaning process, data analysis and limitations of this research project are discussed in closer detail. The conclusion shows that even though the open-source automatic data cleaning tools seem to be highly accurate, their performance could be improved by making some changes.

Qualitative Bewertung der Trainingsdaten für maschinelle Übersetzung, die von Open-Source-Tools zur automatischen Datenbereinigung verworfen wurden

Zusammenfassung

Die manuelle Bereinigung der Trainingsdaten für maschinelle Übersetzungssysteme kann ein teurer und zeitaufwändiger Prozess sein. Qualitativ hochwertige Trainingsdaten sind für ein maschinelles Übersetzungssystem notwendig, um gute, sinnvolle Ergebnisse zu erzielen. Das Ziel dieses Forschungsprojekts besteht darin, die Reinigungsgenauigkeit von zwei Open-Source-Tools zur automatischen Datenbereinigung, Bicleaner und Moses, zu überprüfen. Trainingsdaten werden sowohl für statistische als auch für neuronale maschinelle Übersetzungssysteme benötigt, aber dieses Forschungsprojekt konzentriert sich auf das Letztere. Ein großes Englisch-Deutsch-Korpus wurde mit Bicleaner bereinigt, welches die Reinigungsalgorithmen von Moses enthält. Anschließend analysierte ich die aus diesem Korpus verworfenen Daten auf nützliche Daten. Außerdem wurden die Merkmale der nützlichen Daten, sofern vorhanden, analysiert. Darüber hinaus werden der Reinigungsprozess, die Datenanalyse und die Beschränkungen dieses Forschungsprojekts näher erläutert. Das Fazit zeigt, dass die automatischen Open-Source-Tools zwar sehr genau zu sein scheinen, ihre Leistung jedoch durch einige Änderungen verbessert werden könnte.

List of Tables

1. Background and Research Questions

1.1 Introduction

Machine translation (hereafter MT) is essentially a system that takes a text from one language (the source language) and translates it into another language (the target language) automatically without any human intervention. The source and target languages are natural languages such as English and German, as opposed to machine languages like C and SQL (Rao, 1998). Neural machine translation (NMT) is the future of translation (Mohamed et al., 2021) and has surpassed statistical machine translation (SMT) in many aspects. Both SMT and NMT systems are based on existing human translations. In both cases, systems need to be trained on relevant and sufficient data (i.e. existing translations) to ensure that they function smoothly and produce satisfactory results. The parallel corpora required for these systems are not only used as training data, but also as testing and development data. As a result, such parallel corpora need to be as clean and as accurate as possible in order for these MT systems to provide the best possible results. The corpora used to train these systems are either cleaned manually by humans, or automatically using automatic data cleaning tools such as Bifixer[1], Bicleaner[2], Moses[3] or TMop (Jalili Sabet et al., 2016). Cleaning data manually can be a time-consuming and expensive task. As such, automatic methods of data cleaning are being explored and employed. This project aims to investigate the cleaning accuracy of two open-source automatic data cleaning tools, namely Bicleaner and Moses. The Bicleaner hard-rules[4] encompass the cleaning algorithms of Moses. The following two research questions are addressed in this project:

RQ1: To what extent do these tools discard genuine training data?

RQ2: What are the features of the discarded training data?

[1] Bifixer: https://github.com/bitextor/bifixer
[2] Bicleaner: https://github.com/bitextor/bicleaner
[3] Moses: http://www2.statmt.org/moses/index.php?n=Main.HomePage
[4] The Bicleaner hard-rules: https://github.com/bitextor/bicleaner-hardrules

The motivation behind this project is to determine whether or not the two open-source automatic data cleaning tools, Bicleaner and Moses, clean training data effectively in order to leave data which can be used to train corpus-based MT systems in an efficient manner. This research has the potential to lead to higher-quality machine-generated translations and could also prevent useful or genuine training data from getting lost as part of future data-cleaning processes. My investigation could help professional translators in utilising MT tools to work more efficiently and produce translations of higher quality.

An English-German (EN-DE) corpus from the EMEA (European Medicines Agency)[5] was cleaned using Bicleaner. The data discarded by Bicleaner following the cleaning process provided the data for this project. The discarded data was analysed for any useful data and the features of the discarded data are also discussed in further detail.

1.2 Types of Machine Translation

The four types of MT are as follows:

- **Rule-Based Machine Translation (RBMT)**
  This is the earliest form of MT and is based on linguistic rules and grammar. It centres around three main factors: morphology, syntax and semantics. An RBMT system relies on a considerable amount of human effort, as language experts develop rules by encoding linguistic knowledge into translation lexicons. In addition, RBMT requires linguistic resources like bilingual dictionaries, transfer rules, morphological analysers, syntactic parsers and part-of-speech taggers. The system then uses all of this knowledge to analyse source-language sentences and translate them into the target language. This form of MT does not use the corpus-based approach. (Sreelekha, 2017; Torregrosa et al., 2019)

[5] European Medicines Agency corpora in 22 language pairs: https://opus.nlpl.eu/EMEA.php; https://www.ema.europa.eu/en

- **Statistical Machine Translation (SMT)**

  SMT relies on parallel corpora to produce translation outputs. The parallel corpora for training SMT systems can be obtained through different ways, for example, by aligning existing human translations (such as domain-specific texts from international organisations like the United Nations, European Union and World Bank), by simply using publicly available corpora (such as the the Europarl[6] corpus), or by building corpora by means of web crawling. The motto for SMT is "the more, the better". While RBMT requires linguistic rules and resources to produce translations, SMT uses machine learning techniques to learn likely translations between language entities (words or phrases, for example). It is, therefore, slightly more advanced and "smarter" than RBMT, but both types of MT experience many of the same problems. (Collins, 2019; Koehn, 2009; Lelner, 2022; Poibeau, 2017; Zbib, 2010)

- **Hybrid Machine Translation (HMT)**

  Hybrid machine translation (HMT) is a combination – or, as the name suggests, a hybrid – of two MT systems. Huang et al.'s 2020 paper was the first to combine RBMT with NMT in a classification-based system which chooses the best translation based on the results of RBMT and NMT. With a translation accuracy of 86.63%, this hybrid model has been shown to outperform both RBMT and NMT. Published in 2017, another paper by Du & Way combined NMT and SMT in such a way that NMT was used to pre-translate the training data and SMT – fine-tuned using the pre-translated training data i.e. a corpus – was used to generate the final target text. HMT combines the core working mechanisms of different MT systems to produce the best translation outputs. It is also an effort to compensate for the deficiencies of individual MT methods.

- **Neural Machine Translation (NMT)**

  NMT is subsequently very different from SMT and RBMT. To date, it is also the most advanced form of MT. NMT uses a single neural network to perform an entire translation process. This neural network mimics the neural network in the human

[6] European Parliament Proceedings Parallel Corpora 1996-2011: https://www.statmt.org/europarl/

> brain "[…]to learn directly, in an end-to-end fashion, the mapping from input text to associated output text".(Shen, 2023, p. 870). A big drawback of NMT, however, can be what is known as noisy data, which the AI can pick up on and use, resulting in mistranslations. (Koehn, 2020a; Tan et al., 2020)

This project focuses on cleaning training data for NMT, more specifically cleaning parallel corpora. The next section comprises the literature review.

2. Literature Review

2.1 History of Machine Translation

It is difficult to determine the origins of MT due to the non-uniform quotation techniques used by MT researchers. MT is said to date back to the 1930s and it was in 1933 that Georges Artsrouni, a French engineer of Armenian extraction, was awarded the patent for what he called the "mechanical brain" (Henisz-Dostert et al., 1979; Koehn, 2020b).

In the same year, P.P. Trojanskij in the USSR proposed a detailed process of translation from one language into another with the help of machines. It was a 3-step process consisting of analysis, transfer and synthesis. In this process, steps 1 and 3 were supposed to be done by humans and step 2 would be carried out by the machine. This was essentially an automatic dictionary that needed pre- and post-editing by humans. Consequently, only the transfer part was done by the machine and the linguistic details of the project had not been worked out. Both the source and target language texts were only translated in the most basic way and details like idiomatic expressions and lexical homonymy were missing. As a result, it was assumed that Trojanskij did not possess the linguistic knowledge required for his process of translation. He maintained, nonetheless, that the entire process could be automated (Henisz-Dostert et al., 1979; Koehn, 2020b).

In 1946, Warren Weaver of the Rockefeller Foundation and Andrew Donald Booth, a British crystallographer, came up with the idea of machine translation during a personal conversation. By this time, the computer had also been born. As a response to Weaver's suggestion that wartime methods of decoding messages can also be applied to languages,

Booth pointed out that an electronic device i.e. a computer can be used to translate word for word using a dictionary. Rather than aiming for perfect syntax or grammar in the target language, the point of MT during this time was simply to make scientific information accessible to researchers (Halliday & Delavenay, 1962; Koehn, 2020b).

A year later in 1947, the method of detailed dictionary coding was developed by A.D. Booth and D. H. V. Britten. In 1948, R.M Richens introduced the concepts of split and full dictionaries, thus providing a way to optimise the limited computer storage of the time. The first MT research was officially announced by Ervin Reifler at the University of Washington in early January of 1950. It was based on Weaver's memorandum, *Translation*, which came out in July 1949 (Halliday & Delavenay, 1962; Henisz-Dostert et al., 1979; Koehn, 2020b). Following this, many universities in the USA followed suit in conducting research in the field of MT and in 1954, the first public demonstration of the viability of MT was given as a collaboration between IBM and Georgetown University (Hutchins, 2006; Koehn, 2020b).

Early MT systems consisted mainly of large bilingual dictionaries. A source-language entry had two or more target-language equivalents which facilitated word-for-word translation. Many subsequent projects were based on this growth in linguistics and gave rise to increased accuracy in machine-generated translations. As optimism regarding the topic grew, however, so did disillusion. Many linguistic problems arose in relation to semantic barriers. There were no viable solutions to the miscommunications that occurred as a result of differences not only in language, but also in culture and education. Some operational systems, such as the Mark II developed by IBM and Washington University, produced output that was unsatisfactory in terms of quality. Concerned by the stagnation and setbacks in MT research, US government sponsors set up an Automatic Language Processing Advisory Committee (ALPAC) in 1964. The ALPAC then famously published a report in 1966 which concluded that MT was slower, more expensive and less accurate than human translation. Furthermore, it stated that MT did not have any immediate use or scope. On account of this, the report deduced that there was no need to continue investing in MT research, and the focus should instead be on the development of machine aids, like automatic dictionaries for human translators, and computational linguistics. Dictation was also cited as a more efficient way of producing translations than MT (Hutchins, 2006; Koehn, 2020b).

The ALPAC report was criticised for being improvident and prejudiced, and for not recognising the full potential of, need for and scope of MT. In spite of this, the report had a tremendous impact. Although MT research stopped in the US, it continued in Canada and parts of Europe. After a period of quiet between the 1960s and the 1970s, MT research gained momentum again in the mid-1970s. The commercial and administrative demands of different communities and international trade in Europe, Canada and Japan led to translation services becoming highly sought after. As these needs went beyond the capability of traditional translation services, cost-effective MT became a necessity for international trade and commerce (Hutchins, 2006; Koehn, 2020b). Elsewhere, it was in 1970 that a system installed by the US Air Force, which produced translations for many years, was replaced by the MT tool SYSTRAN[7] (Hutchins, 1995; Hutchins, 2006; Koehn, 2020b).

MT started supporting even more language pairs in the 1980s. Countries like China and Taiwan came up with their own systems for their own native languages and as a result, MT started growing in many different directions and on many fronts. Besides SYSTRAN, which today translates between many different language pairs, some of the most important early MT systems were Logos (which was first used to translate English aircraft manuals into Vietnamese, but later expanded to the language pairs German-English and English-French) (Scott, 2003), the Pan American Health Organization[8] (English-Spanish and Spanish-English), the METAL system (German-English) (Liu & Liro, 1987), and various systems in Japan for English-Japanese and Japanese-English translations for local computer companies.

By this time, microcomputers and text-processing software were on the market, which opened up a whole new arena for further research into and development of MT. Up until this point, the main framework of MT research had been linguistic rules of various kinds like lexical transfers, lexical rules, and morphological as well as syntactic analysis. This was essentially a strategy of "indirect" translation and some of the major transfer systems of the time, such as GETA-Ariane (Grenoble) (Boitet et al., 1982), Mu (Kyoto) (Nagao & Tsujii, 1986), SUSY (Saarbrücker Übersetzungssystem) (Maas, 1977), METAL (White, 1985) and Eurotra (Maegaard, 1988) were founded on this rule-based approach. Interlingual MT

[7] SYSTRAN Translate: https://www.systran.net/en/translate/
[8] Machine Translation at the Pan American Health Organization: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=14762:machine-translation-at-the-pan-american-health-organization&Itemid=0&lang=en#gsc.tab=0

systems like Distributed Language Translation (DLT) (Utrecht) (Witkam, 1984) and Rosetta (Eindhoven) (Landsbergen, 1989) also used this linguistics-oriented approach. In 1989, IBM broke the rule-based trend by applying new methods and strategies, which are known today as "corpus-based" methods. The corporation performed an experiment which was based entirely on statistical methods and the effectiveness of these methods was surprising to researchers. This study then later contributed to further research on these methods. At the same time, research groups in Japan began experimenting with methods based on corpora of existing translations. This was called "example-based" translation.

The biggest breakthrough, however, came when the statistics-based approach to MT was revived by IBM as part of their Candide project. Statistics-based methods were common in MT-based research during the 1960s, but the results were often disappointing. With the development of speech recognition techniques, IBM approached these methods with a new perspective. A distinctive feature of the Candide project was that statistics-based methods were the only means of analysis, experiment and research, meaning that no linguistic rules were involved. A vast corpus of English and French texts from the Canadian Parliament was used and the method was to align sentences, phrases and individual words in the parallel texts and to calculate the probability of one word or phrase in one language, for instance, corresponding with its aligned word or phrase in the target language (Hutchins, 1995; Koehn, 2020b).

The turning point for MT was the early 1990s, which is when the corpus-based approach – otherwise known as example- or memory-based approach – gathered momentum. Nevertheless, research on the rule-based approach also continued in both transfer or intralingual systems, as well as in interlingual systems. The first translation memory systems (e.g. Trados[9]) came on to the market around this time, giving translators access to previously translated texts. The early 90s also saw the development of practical applications of MT for translators, which allowed them to use the technology in domain-specific and language-specific environments.

The trend continued into the late 1990s, when there was a surge in the use of MT and translation aids (e.g. translator workstations and translation memories). An area of substantial

[9] Trados – Translation Software, CAT Tool & Terminology: https://www.trados.com/

growth was that of software localisation. MT systems also became popular for personal computers and for this reason, "downsized" and improved versions of the mainframe systems started becoming available. MT in online applications, like email applications and other social media, experienced even more rapid growth. Pioneering such MT services was a huge step forward where the technology is concerned, first leading to the translation application Babel Fish[10] and then to Google translate (Hutchins, 2006; Koehn, 2020b).

In the 2000s, SMT became the dominant framework in MT research. Since it is a corpus-based approach, corpus cleaning and classifying systems like Moses, GIZA and Bicleaner also saw substantial development for the training of new MT systems. NMT, however, is the most recent MT framework and the dominant approach used in MT research today. NMT uses a neural network to train MT systems. This neural network can be described as "a machine learning technique which takes in a number of inputs and predicts outputs" (Koehn, 2020b, p. 67). Both SMT and NMT are corpus-based approaches and require data to train the systems. Hybrid MT which combines different MT systems has also had some success, but this approach has now been superseded by NMT (España-Bonet et al., 2011).

MT has since seen rapid growth and major developments. What started as a huge mechanical translation framework with automatic dictionaries is now a software system available on personal computers and a freely available online service. Since the focus of the present project is to see how accurately and efficiently open-source automatic data cleaning tools clean data for training NMT systems, the next section will discuss previous research on various approaches to data cleaning.

2.2 Different Approaches to Data Cleaning

Many studies have been carried out on the topic of cleaning training data for MT systems. Negri et al. (2017) addresses translation memory (TM) cleaning under a number of conditions and the problem of automatically cleaning a TM by identifying disputable translation units (TUs). These problematic TUs contain seemingly useless translations from the point of view of the end user of a computer-aided translation (CAT) tool. A CAT tool is a "translator's aid"

[10] Babel Fish: https://www.babelfish.com/

in the sense that human translators use them to enhance productivity, save time and produce translations of higher quality. The CAT tool splits the input (i.e. source) document up into segments or TUs. If there is an exact or fuzzy match for any of the source segments available in the TM(s) linked to that particular translation project, the tool proposes a translation that the human translator can use and if need be, edit or revise.

Advanced CAT tools come integrated with not only TM technology, but also with MT. TMs are used to store and retrieve previously translated segments and MT technology is used to translate segments automatically. Both TM and MT technology have advanced rapidly, with NMT outperforming every other type of MT system and TM technologies using MT to improve matches (Shen, 2023). Even today, while using a CAT tool, a TM remains the main source of information and suggestions for professional translators.

TM technology has long been a resource used in CAT tools. TMs consist of databases that cache every translated segment of previously translated input documents. Translation fragments can be stored at segment level or paragraph level, for instance, (if segmentation rules are changed), and these together with the corresponding source fragments are what are called TUs. TMs are very useful to translators because such databases help them access previous work that can be used in future tasks, thus assuring consistency. TMs are ever increasing in size. There are also “private TMs”, which vary in size and can be a very valuable resource for both individual translators and large translation companies. Collaborated or public TMs are also beneficial, but these grow in a less controlled manner than private TMs.

The usability of a TU depends on two major factors: the matching process and the quality of the TU. The matching process is basically the source segments in a current project being compared against segments of the same language (i.e. the source language) in the TM. The correspondence is designated a computed “fuzzy match” score (Negri et al., 2017). If the alignment of source segments goes wrong, the match is essentially useless as the translation is wrong. TU quality can be affected by a number of factors like capitalisation, missing spaces, punctuation errors and accuracy. Minor errors can also affect TU quality, but they can be more difficult to spot which can require more time and attention to clean the TM properly.

TMs should benefit users as much as possible. It is always considered easier and better to use a TM to ensure uniformity and consistency in translations. In fact, using a TM ensures the use of client-preferred terminology along with the required quality. TMs should, therefore, regularly be cleaned and maintained. TMs are either cleaned manually – which can be expensive and time-consuming – or automatically by applying filters, which may or may not be accurate. Most CAT tools come with spelling and grammar checks, but some also come with integrated data cleaning methods which can do simple checks for syntax, for instance, like checking for repetition and punctuation errors. These integrated data cleaning methods are, nonetheless, not yet entirely accurate and have a lot of room for improvement.

As mentioned above, cleaning TMs manually can be a time-consuming and costly process. There are, however, quality assurance (QA) and terminology tools like ApSIC Xbench[11] and Verifika[12] (Barbu et al., 2016) that support manual identification and correction of TMs. These paid tools also check for formatting and terminology errors in ongoing translations, allowing users to fix them as they go along. In addition, they check for minor errors like tag mismatches, repetition and spellings mistakes. They also identify complex errors like mistranslations, inconsistencies, number mismatches and URL mismatches. Verifika even checks for grammatical errors (Petrova, 2019).

Barbu et al. (2016) invited six teams of participants from both industry and academia, namely Autodesk (Zwahlen et al., 2016), the University of Edinburgh (Buck & Koehn, 2016), FBK HLT-MT (Ataman et al., 2016), Jadavpur/Saarland University (Nahata et al., 2016), Lingua Custodia (Mandorino, 2016) and the University of South Africa (Wolff, 2016). Every team submitted their automatic TM cleaning systems for evaluation. Data for three language pairs – namely English-Italian, English-Spanish and English-German – was chosen. 3,000 TUs per language pair were taken for evaluation. This data was taken from MyMemory[13], the world's largest public TM. Data was annotated using MT-EQuAI[14], an online toolkit for human assessment of MT output developed by FBK that is accessible via web browser. TUs were annotated based on whether or not the target content of each TU represented an equivalent

[11] ApSIX Xbench – Quality Assurance and Terminology Management: https://www.xbench.net/
[12] Verifika QA: https://e-verifika.com/
[13] My Memory – The World's Largest Translation Memory: https://mymemory.translated.net/doc/
[14] MT-EQuAI toolkit for manual evaluation of MT output: https://mt4cat.fbk.eu/software/mt-equal

translation of the source segment. The annotation was carried out by two native speakers of each target language. A 3-point scale was applied for evaluation, namely correct, semi-correct (i.e. containing a few mistakes amendable by little post-editing), and incorrect.

There were three individual tasks included in the shared task: binary classification I, binary classification II and fine-grained classification. Every participant group used different tools for the shared task. Autodesk used a PoS tagger and an in-house Moses-based MT system; the University of Edinburgh used NMT, a word aligner, a language model and a subword unit splitter; FBK used an open-source TM cleaner and a word aligner; Jadavpur/Saarland University did not use any tools; Lingua Custodia used a word aligner and a stemmer; the University of South Africa used a spell checker, a quality assurance tool and a grammar checker. The results were then compared against two baselines, and it transpired that most of the submitted systems outperformed the baselines in terms of balanced accuracy (BA).

On a related note, Srivastava et al. (2020) discusses two different methods of cleaning parallel corpora for SMT in detail. The first one is a machine-assisted method to clean smaller parallel corpora, and the second is an automatic method to clean large parallel corpora. A machine-assisted method involves a human using a machine to help them clean the data. The effectiveness and performance of these methods is evaluated by experimenting with translations produced by Moses, an open-source toolkit used to build an SMT system. English-to-Indian-language-machine-translation (EILMT) data is used for experimentation, more specifically an English-Hindi corpus. For the purpose of Srivastava et al.'s article, these methods focus on the aforementioned language pair, though they can be extended to other language pairs.

The first method to be evaluated is the machine-assisted data cleaning method. This approach uses two similarity measures (SMs) to detect noisy sentence pairs in the corpus. These SMs are sentence length of the source and target sentences (SM1), and word alignment (SM2). The computation of both these measures is automatic. GIZA++, a freely available software, is used for word alignment. As mentioned above, the machine-assisted method requires human intervention. The noisy data is only cleaned if it takes minimal time and effort to do so, otherwise the sentence pair is removed altogether. 25,000 sentence pairs were used in

Srivastava et al's study. Of these 25,000, 24,000 were used for training, 250 for tuning and 750 for testing. The results were evaluated using the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001) and the National Institute of Standards and Technology (NIST) scores. The results showed, that according to the BLEU and NIST scores as well as human evaluation, an SMT system trained on a clean and filtered corpus provides higher quality translations than an SMT system trained using a noisy corpus. To be more specific, when the SMT engine was trained on a clean corpus rather than a noisy one, the BLEU score improved by 1.8%. The NIST score of the SMT system trained on a noisy corpus was 6.5049, which increased to 6.5482 when the SMT system was trained on a clean corpus.

The second method is the automatic data cleaning method. This method uses two features to evaluate the data, namely difference in length between source and target sentences, and the Viterbi alignment score generated by the Hidden Markov model (HMM) during training. Together, these features are very effective in filtering corpora. Likewise for this method, an SMT system built using Moses was used, as well as GIZA++ for word alignment. The data used here also resembled that used in the first method, i.e. EILMT data for the language pair English-Hindi. Exactly 274,492 sentence pairs were used. These were cleaned based on sentence length, and after cleaning, only 270,545 sentence pairs remained. This clean corpus was then treated as the raw corpus. In this case, the BLEU score improved by 1.4% for the SMT system trained on the clean corpus. The NIST score increased from 6.5049 to 6.5452 for this method.

Having said this, sentence pairs cannot always be said to be either correct or incorrect. Sentences that fall somewhere in between can be referred to as only partially correct. Removing these partially correct sentence pairs is not always useful, as parts of these sentence pairs can still be entirely correct. This is why using an automatic cleaning method to clean a corpus containing such partially correct sentence pairs is difficult. Human involvement is necessary in cases such as these. The overall results show that both cleaning methods are equally efficient in cleaning corpora, although the machine-assisted method is slightly better.

In summary, the machine-assisted method of cleaning corpora performs better than the automatic method. In spite of this, the automatic method can be used for larger corpora. The automatic method also works reasonably well on smaller corpora. Since the automatic method does not involve any human intervention at all, however, the threshold of features and filters need to be set up very carefully, so as to avoid losing useful data.

Negri et al. (2017) and Jalili Sabet et al. (2016) present the use of another tool to clean TMs, namely Translation Memory Open-source Purifier (TMop). This tool is written in Python and designed for users to detect problematic TUs. The actions allowed by TMop can be customised or changed according to specific types of errors and specific language pairs, allowing users to take on a variety of translation errors. One error in one language pair might not be the same in another, so the course of action needed to fix the error would naturally need to be altered and TMop provides the flexibility to do this. Conditions such as external knowledge of other tools, and resources such as labelled training data are also critical to the effectiveness of data cleaning tools.

A TM can be entered into TMop either in TMX format or as a text file containing one TU per line in the form ID, source, target. The output is made up of several files, some of them being good or accepted, and some being bad or rejected. “The tool consists of three parts: core, filters and policy managers”(Negri et al., 2017, p. 99). The core is the part which manages the functionality of filters, policy managers and input and output files. The filters perform the job of detecting “bad” units. Every filter is different, designed to detect a specific type of problem (e.g. punctuation, length or capitalisation), and follows with an evaluation of each TU. According to Jalili Sabet et al. (2016), TMop has 23 filters which have been organised into 4 categories: basic (8), language identification (1), QE-derived (9) and word embedding (5). According to Negri et al. (2017), on the other hand, TMop has 32 filters which are divided into 3 categories. These filters comprise 9 basic filters, 18 QE-derived filters and 5 word-embedding filters. The decision policies of TMop allow the final output of the active filters to be evaluated for each TU. Simple decision-making strategies are usually implemented here, meaning that users can either accept or reject a TU. Users can, however, also replace these decisions with more complex methods, since both filters and policy managers can be modified in TMop. The tool currently has three policies; OneNo, 20%No

and MajorityVoting. The first filter immediately rejects a TU once one of the filters has rejected it. The other two filters only reject a TU if 20% or 50% of the filters have rejected it respectively. As we can see, the first option does not provide any leeway when it comes to rejecting a TU. The third option is very flexible as it takes the decisions of 50% of the filters into account, which can leave more room for error. The second option is an intermediate one. Considering 20% of the filters in its decision-making, it is more forgiving than the first policy, but also more precise than the third. The user's needs and the overall quality of the TM will influence the choice of policy and how controlled the cleaning process is.

The first approach in Negri et al. (2017) is a supervised approach, whereby TMop filters act as feature extractors. The data used in this approach is NLP4TM shared task data and covers three language pairs: English-Italian, English-Spanish and English-German. The TMs used are part of MyMemory, the world's largest public TM (Barbu et al., 2016). The TUs in these TMs were annotated manually and included errors like omission, punctuation errors and capitalisation. Two classifications are used to assess the quality of data: Binary Classification I (BCI) and Binary Classification II (BCII). BCI strictly only accepts the "good" TUs, where the target segment has to be a perfect translation of the source segment. BCII is more flexible and lenient in its quality assessment, whereby TU target segments with minor mistakes – which can be corrected through minimal post-editing – are also accepted. Two sets of data, namely training and test, are used for this approach, and the results are divided into those which are positive, those which are negative, and those which are positive/negative. There is disparity in the data distribution over the two class labels. For BCI, there are 1.6 to 3.5 times the number of positive examples than there are negative, whereas for BCII, there are 3.6 to 5.6 times more positive examples than negative.

To combat this disparity in data distribution, the performance is measured in terms of balanced accuracy (BA), which is calculated by averaging the positive and negative accuracy scores. The results are then compared to two baseline scores, only one of which is mentioned. To be precise, the baseline score for the language pairs English-Spanish and English-German was 0.5, and for English-Italian, it was 0.54. The data score for BCI for English-Spanish was 0.64, for English-German, 0.52, and for English-Italian, 0.65. For BCII, the scores were as follows: English-Spanish achieved 0.73, English-German scored 0.51, and English-Italian obtained 0.75. The differences in the baseline and data scores for both BCI and BCII in

English-Spanish and English-Italian are larger than in English-German. One reason for this may have been unequal instance distribution in the training data. From this, we can deduce that TMop would perform more competitively in different language combinations if more balanced training data were to be provided for learning and tuning the models. The effectiveness of this approach was proved by consistent improvements of the baselines with highly skewed data distributions.

The second approach is the unsupervised TM cleaning approach. In this method, no labelled training data is available hence TMop is used for a very different two-step process. Since there is a lack of supervision, the first step involves using filters to induce reliable binary labels (1 or 0 for good or bad respectively) for some TUs. As part of the second step, the filters are used as feature extractors to train a group of binary classifiers with automatically-labelled data. The purpose of this approach is to see if unsupervised filters provide different and independent views of the data.

To capture the aspects of similarity between source and target texts, three groups of features are organised into groups of two and applied to a subset of unlabelled TUs, which are randomly extracted from the input TM. For each TU in a subset, features belonging to each combination are extracted and the TUs are then rated from best (i.e. close to 1, showing high similarity between source and target), to worst (i.e. close to 0, showing low similarity). After doing this for each combination of features, three ranked lists are obtained. These lists are then processed to get three sets of positive/negative examples. Their size will depend on the number of TUs taken from the top (“good” TUs) and the bottom (“bad” TUs).

As a result, every TM needs to be cleaned periodically, and using automatic cleaning methods is undoubtedly the most time- and cost-effective way to do so. According to Negri et al. (2017), TMop has proven to an effective open-source data cleaning tool, in which new filters are constantly being incorporated to enhance its use.

Defauw et al. (2019) mentions that the motto “the more, the better” applies in the case of SMT, and this refers to the amount of data used to train SMT models. A model for misalignment detection was built and a supervised regression approach was taken – as opposed to a supervised classification approach – to solve the problem of misalignment. Alongside this, a Misalignment Detection (MAD) score was introduced to score the data. The

language pairs for this task were English-French and English-Gaelic. The results obtained were compared to those obtained through Bicleaner, which is a tool that automatically removes noise from a corpus with the help of a set of hard-rules. The difference between Bicleaner and MAD is that Bicleaner cleans other errors and types of noise besides misalignments due to the set of hard-rules it uses. Bicleaner achieved a BLEU score that was 0.3 times better than its MAD score. Ramírez-Sanchez & Zaragoza-Bernabeu's research (2020) also explores and discusses TM and corpora cleaning using two open-source tools which are used to clean parallel data: Bifixer and Bicleaner. Both data cleaning tools are freely available online and are a part of the EU's Paracrawl[15] project.

Sánchez-Cartagena et al. (2018) used Bicleaner to clean data and identify equivalent translations for the pre-processing step. Once it had done so, it used an automatic classifier to detect misalignments. Following this, sentences were scored based on fluency and diversity. Two datasets of 10 million words and 100 million words were tested and the language pair was English-German. Three different methods were used to achieve a training corpus with diverse vocabulary and fluent sentences, namely language model scoring, an active learning inspired data selection algorithm and n-gram saturation. NMT is very sensitive to noisy data. The BLEU scores increased after pre-processing the data with hard-rules and then using the automatic classifier. The increase in BLEU scores also signified an improvement in NMT performance, which outperformed SMT in both datasets.

The data used in Ramírez-Sanchez & Zaragoza-Bernabeu (2020) is in the language combination English-Portuguese. Four publicly available corpora were gathered, namely from the Europarl corpus version 7, OpenSubtitles 2018, JW300 and WikiMatrix. All of the corpora were cleaned with the help of Bifixer and Bicleaner, both of which can be used for over 30 language combinations. Bifixer uses the restorative approach, which differs from the filtering approach used by Bicleaner. With regard to this data, Bifixer was used first, and Bicleaner second. Cleaning the data in this order meant that parallel sentences were obtained before the excess noise was eliminated. Five steps were performed by Bifixer as part of the restoration approach: empty side removal, character-fixing, orthography fixing, re-splitting

[15] ParaCrawl corpora: https://paracrawl.eu/

and duplicates identification. After running the data through the restoration process, it was then ready for filtering.

Bicleaner is a parallel sentence noise filter and classifier tool that first pre-filters data. It used to use 37 rules to pre-filter data, but these have now been reduced to only 20 categories. Some rules are language-dependent, while others consider factors like sentence length, character-usage instead of n-grams, and use of punctuation on both the source and target side. These rules were designed to target noise from online content but can also be used for professionally compiled corpora. It then uses language model fluency scoring to assign scores to the data. It scores every sentence pair against the language model. Sentence pairs with a score of below 0.5, which is the set threshold, are equal to 0, which means that they must be discarded. For the paper in question, this step was not included due to how small the corpora were. The Bicleaner classifier then takes the remaining data and assigns it with scores between 0 (bad) and 1 (good). The EU's Paracrawl corpora only contain sentences with a score of above 0.7. Related studies have shown that data with a score of above 0.5 is also of acceptable quality. Both of these thresholds are explored in this paper.

Before training, the data was cleaned using all of the Bifixer steps and the pre-filtering step from Bicleaner. After running the data through Bifixer, more data was obtained because of the recovery of 1.1% of the sentences after re-splitting. Of this, only the unique data was kept and a better output was achieved. After the pre-filtering step from Bicleaner, most of the data was retained, none of the four corpora matched with all 37 filters and the main source of noise is identical across all four corpora. Finally, this data was scored using the Bicleaner classifier. Application of the more rigid threshold of 0.7 resulted in an average of 22.9% of the data being removed across all four corpora, though mainly from the WikiMatrix and OpenSubtitles corpora. When the threshold was brought down to 0.5, the amount of data removed dropped to an average of only 10.9%.

After this cleaning process, 100 sentences were taken from each corpus and evaluated using KEOPS[16], an open-source tool which proposes a framework for the manual evaluation of

[16] KEOPS – Key Evaluation of Parallel Sentences: https://github.com/paracrawl/keops/blob/master/pm.md

parallel sentences. Every sentence was annotated as either valid or as containing one of the following errors: wrong language identification, wrong alignment, wrong tokenisation, machine translation (which is not discussed in further detail in the paper), translation errors or free translation. Annotation was done according to European Language Resource Coordination's (ELRC) validation guidelines. The annotation process revealed that only two sentences from the Europarl corpus, seven from JW300, 11 from OpenSubtitles and 30 from Wikimatrix had issues.

The data was also evaluated through MT. MT systems were trained before and after data cleaning for both 0.7 and 0.5 thresholds and for both language directions. The results were as follows:

		Europarl		**JW300**		**WikiMatrix**		**OpenSubtitles**	
		EN-PT	**PT-EN**	**EN-PT**	**PT-EN**	**EN-PT**	**PT-EN**	**EN-PT**	**PT-EN**
Before cleaning		26.2	31.5	29.0	34.1	35.8	36.8	31.2	37.9
After cleaning	0.5	26.0	31.5	29.1	34.2	36.2	36.8	31.9	39.5
	0.7	26.2	31.7	29.4	34.4	36.3	37.0	32.2	40.1

Table 1: BLEU scores for NMT systems trained before and after cleaning the corpus (Ramírez-Sanchez & Zaragoza-Bernabeu, 2020, p. 298)

The results indicate that the data did not get worse, but rather consistently improved. In short, data cleaning reduces corpus size but increases quality. Both Bifixer and Bicleaner can be used without any other tools for 30 language combinations to clean corpora for MT systems – especially those using NMT – and perhaps even for other tasks involving NLP. Some of the Bicleaner rules, however, may be slightly too strict and could be relaxed to prevent the loss of useful data.

As of yet, the way in which noisy data affects the quality of NMT outputs has not been the focus of many MT-based studies (Wang et al., 2018). Noisy data has a bigger impact on

NMT than it does on SMT, and is an important issue which needs to be addressed if we are to improve the quality of NMT technology. Khayrallah & Koehn (2018) explores five types of noise that can harm NMT models. These are misaligned sentences, misordered words (i.e. randomly reordered words, disfluent language, or heavily specialised language use), wrong language, short segments, and untranslated sentences. Artificial noise was added to the test data to assess the NMT model's performance. The language pair was English-German. The NMT model was trained using Marian (Junczys-Dowmunt et al., 2018) and the SMT model was trained using Moses (Koehn, 2016). The researchers observed that misalignments – which constituted 41% of the noise within the corpus – harmed the performance of the NMT model considerably, whereas they had almost no effect on the SMT model. In terms of BLEU score, misalignments appeared to reduce NMT performance by 1 to 2.

Parallel corpora can contain a lot of misalignments. If there are 100 sentence pairs in a parallel corpus, and, for instance, the fifth TU is incorrectly aligned with the sixth TU rather than the fifth – which indeed contains the corresponding translation – then there is a possibility of the following sentence pairs also to be misaligned. Such misalignments can easily be corrected when corpora are cleaned manually by a human translator. If automatic cleaning methods are used, however, all of these misaligned pairs will simply be removed, thus reducing the size of the corpus and limiting its ability to improve MT quality.

Lowphansirikul et al. (2021) focused on the low-resource language pair English-Thai. They too believe that insufficient training data or an insufficient number of training examples can deteriorate translation quality. In their project, 1 million sentence pairs gathered from various sources were used to solve the problems associated with quantity and diversity of training data. The MT models performed analogously to Google Translate API on existing data for English-Thai, and outperformed Google when additional parallel sentences from Open Parallel Corpus (OPUS) were added for English-Thai and Thai-English. The MT models were trained based on Transformer (Vaswani et al., 2017) and their performance was compared against that of Google and AI-for-Thai. The Thai-English IWSLT dataset (IWSLT Evaluation 2015 - MT Track, 2015) was the benchmark and BLEU was used for quality evaluation. The data used here was taken from the web by means of web crawling and included documents from English and Thai Wikipedia pages, product reviews, publicly

available datasets and official documents from the Thai government. The results showed that the corpus could in fact be used to train NMT systems. The BLEU scores ranged from 39 to 42 in various configurations and both translation directions.

Furthermore, Bane & Zaretskaya (2021) explored different open-source tools for cleaning bilingual data. The 5 language pairs used in their study were English-Chinese, English-German, English-Japanese, English-Russian and English-Spanish. Professional linguists assessed the quality of the open-source tools using another four evaluation tools. The scores obtained from this quality assessment were then carefully examined and compared. The task was carried out in two phases, phase 1 and phase 2. The results showed that the correct tools can be used to filter bilingual data, even when it comes to smaller datasets.

Zaragoza-Bernabeu et al. (2022) studied Bicleaner AI, an advanced version of Bicleaner, which detects noisy data (e.g. noisy sentences) in parallel corpora. The binary classifier component of Bicleaner AI is based on deep learning techniques. Bicleaner AI has two models, lite and full, and was trained for 33 language pairs in both of these forms. The lite model provides high-speed inference, while the full model provides high-performance inference. Of the many corpora downloaded from MTData[17], OPUS was chosen for the experiments on filtering corpora through Bicleaner AI. The results were as follows:

	EN-FN			**EN-LV**			**EN-RO**		
	5M	**50M**	**100M**	**5M**	**30M**	**60M**	**5M**	**50M**	**100M**
Bicleaner	14.3	21.0	22.7	12.2	17.1	18.5	21.8	28.7	29.5
Bicleaner Lite	14.2	22.4	23.7	12.5	17.7	18.9	21.4	28.8	29.6
Bicleaner Full	13.7	25.6	26.2	15.6	19.5	20.0	25.3	31.0	30.8

Table 2: BLEU scores showing performance comparisons of Bicleaner AI, Bicleaner AI lite and Bicleaner AI full (Zaragoza-Bernabeu et al., 2022, p. 828)

[17] MTData for the collection and preparation of MT datasets: https://github.com/thammegowda/mtdata

The Bicleaner AI full model outperforms both the standard Bicleaner AI model and the Bicleaner AI lite model. This demonstrates that a deep learning-based classifier brings about clean corpora of higher quality, which in turn leads to improved NMT system performance.

The next chapter discusses the methodology of the present project, which will address the research questions outlined in the previous section.

3. Methodology

The methodology for my project is mainly inspired by Ramírez-Sanchez & Zaragoza-Bernabeu's 2020 paper. The aim of my thesis is to see the extent to which automatic cleaning tools discard genuine training data and to analyse the features of said discarded training data. The tools used for the present study are Bicleaner, which is an open-source data cleaning tool, and Moses, which is an SMT system that allows you to automatically train translation models for any language pair. Corpora are cleaned for the purpose of training NMT systems to generate higher quality translation outputs. More training data is needed and this data needs to be cleaned to be most beneficial, especially since neural networks do not allow the incorporation of glossaries as easily as they used to.

This research is going to be performed on a controlled corpus, as opposed to one that has been created using content obtained through web crawling. Bicleaner is arguably most useful when it comes to cleaning data scoured from the internet, since such data is likely to contain a lot of noise, which is perhaps less likely to appear in professionally compiled TMs or corpora. Nevertheless, Bicleaner can still be used to clean professionally produced corpora. An English-German corpus from the European Medicines Agency (EMEA) containing 1,108,752 sentence pairs was cleaned using Bicleaner and the 19,355 discarded sentence pairs form the corpus for this project. Bicleaner's 20 hard-rules are outlined below as they appear on GitHub[18]:

[18] At the time of carrying out the analysis, GitHub displayed 20 hard-rules for Bicleaner. Since then, however, another three rules have been added: https://github.com/bitextor/bicleaner-hardrules

no_empty	Sentence[19] is empty
not_too_long	Sentence is more than 1024 characters long
not_too_short	Sentence is less than 3 words long
length_ratio	The length ratio between the source sentence and target sentence (in bytes) is too low or too high
no_identical	Unwanted literals: "Re:","{{", "%s", "}}", "+++", "***", '=\'"
no_literals	The ratio of non-alphabetic characters in source sentence is more than 90%
no_only_numbers	The ratio of numeric characters in source sentence is too high
no_urls	There are URLs
no_breadcrumbs	There are more than 2 breadcrumb characters in the sentence
no_glued_words	There are words in the sentence containing too many uppercased characters between lowercased characters
no_repeated_words	There are words repeated consecutively
no_unicode_noise	Too many characters from unwanted unicode in source sentence
no_space_noise	Too many consecutive single characters separated by spaces in the sentence (excludes digits)
no_paren	Too many parenthesis or brackets in sentence
no_escaped_unicode	There is unescaped Unicode characters in sentence

[19] Despite the fact that the Bicleaner hard-rules and the Moses cleaning algorithm remove data based on the information in the source or target "sentence", and the fact that I am analysing sentence pairs, I will occasionally refer to the source and target texts in these pairs as "segments" or "fragments". This is because some of them contain more than one individual sentence.

no_bad_encoding	Source sentence or target sentence contains mojibake
no_titles	All words in source sentence or target sentence are uppercased or in titlecase
no_wrong_language	Sentence is not in the desired language
no_porn	Source sentence or target sentence contains text identified as porn
lm_filter	The sentence pair has low fluency score from the language model

Table 3: Bicleaner hard-rules

The data was first passed through the Bicleaner hard-rules (see Table 3), which are pre-filters. After this step, the sentences kept by the hard-rules were then passed through the Bicleaner classifier. The classifier is trained with "good" and "bad" translations, which were then used to give the sentence pairs a score between 0 and 1 based on the language model. In this project, the threshold score for keeping sentence pairs was 0.5. Sentence pairs that scored below 0.5 were discarded by the Bicleaner classifier. The flow chart below illustrates the entire process:

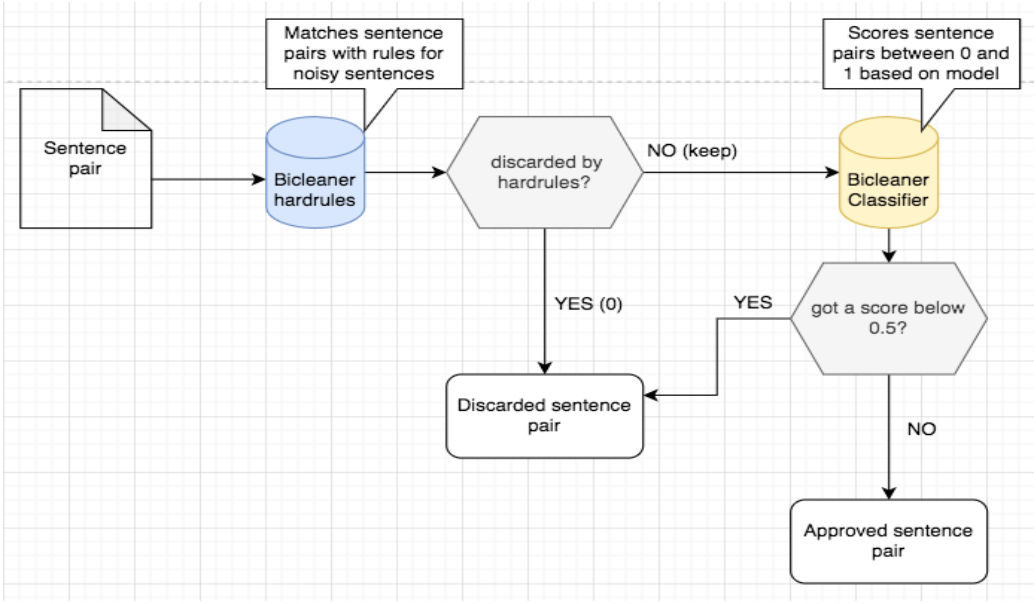

Image 1: Workings of Bicleaner (Image provided by Gema Ramírez-Sánchez, CEO of Prompsit Language Engineering)

This project aims to analyse the data discarded by the Bicleaner classifier. I also want to investigate which hard-rules led to data being discarded. Moreover, I would like to see whether or not any data was discarded by the Bicleaner classifier on account of not matching the translations that it had been trained with.

The 20 Bicleaner hard-rules comprise six out of the seven specifications that are part of the Moses cleaning algorithm. The one Moses rule that Bicleaner does not have is "Sentence aligned (one sentence per line)". The comparison of rules between the two automatic data cleaning tools (the similarities and differences) can be seen in the table below:

Bicleaner hard-rules		**Moses cleaning algorithm specifications**
no_empty	Sentence is empty	- Removes empty lines - One sentence per line, no empty lines
not_too_long	Sentence is more than 1024 characters long	- Drops lines (and their corresponding lines) that are empty, too short, too long or violate the 9-1 sentence ratio limit of GIZA++ - Sentences longer than 100 words (and their corresponding translations) have to be eliminated (note that a shorter sentence length limit will speed up training)
not_too_short	Sentence is less than 3 words long	
length_ratio	The length ratio between the source sentence and target sentence (in bytes) is too low or too high	
no_space_noise	Too many consecutive single characters separated by spaces in the sentence (excludes digits)	Removes redundant space characters

no_titles	All words in source sentence or target sentence are uppercased	Everything lowercased (i.e. sentences containing capital letters are removed)
		Sentence aligned (one sentence per line)

Table 4: Comparison between Bicleaner hard-rules and Moses cleaning algorithm specifications[20]

The two open-source automatic data cleaning tools also differ in two other ways related to how their rules are defined. The Bicleaner hard-rules state that sentences "more than 1024 characters long" will be discarded, while the Moses rules mention that only those sentences "longer than 100 words (and their corresponding translations) have to be eliminated (note that a shorter sentence length limit will speed up training)". Although Bicleaner talks about the maximum sentence length in characters instead of words, it actually specifies the minimum sentence length in words, saying that a sentence should not be less than 3 words long. The second difference is the ratio between the source and target sentences. In Bicleaner, the ratio is 1:1.5, and in Moses, the maximum allowed ratio is 9:1. Moses, therefore, allows for more flexibility than Bicleaner when it comes to word ratio.

The threshold used for cleaning the EMEA corpus was 0.5. The discarded data gave a total of 141 scores, all of which were between 0 and 0.5. The scores were as follows:

0.497	0.44	0.383	0.325	0.268	0.21	0.115
0.495	0.438	0.38	0.323	0.265	0.207	0.11
0.492	0.435	0.378	0.32	0.263	0.205	0
0.49	0.432	0.375	0.318	0.26	0.203	
0.487	0.43	0.372	0.315	0.258	0.2	
0.485	0.427	0.37	0.312	0.255	0.195	
0.482	0.425	0.367	0.31	0.253	0.193	
0.48	0.422	0.365	0.307	0.25	0.19	

[20] The Moses cleaning algorithms listed here are taken from Koehn (2016).

0.477	0.42	0.362	0.305	0.247	0.188	
0.475	0.417	0.36	0.302	0.245	0.185	
0.472	0.415	0.357	0.3	0.242	0.182	
0.47	0.412	0.355	0.297	0.24	0.177	
0.468	0.41	0.352	0.295	0.237	0.175	
0.465	0.407	0.35	0.292	0.235	0.172	
0.463	0.405	0.347	0.29	0.233	0.17	
0.46	0.403	0.345	0.287	0.23	0.168	
0.458	0.4	0.343	0.285	0.228	0.165	
0.455	0.398	0.34	0.282	0.225	0.16	
0.453	0.395	0.338	0.28	0.223	0.158	
0.45	0.393	0.335	0.278	0.22	0.147	
0.448	0.39	0.333	0.275	0.217	0.142	
0.445	0.388	0.33	0.273	0.215	0.135	
0.443	0.385	0.328	0.27	0.212	0.13	

Table 5: Bicleaner scores

All in all, the scores from this dataset can be divided into four score groups, namely those between 0.5 and 0.3, those between 0.3 and 0.2, those between 0.2 and 0, and those with a score of 0. None of the data scored between 0.1 and 0. The first 100 sentence pairs in every score group will be analysed. Of the 19,355 discarded sentence pairs, a total of 515 sentence pairs will be analysed manually to see if they contain genuine data. If the data is found to be genuine and worth keeping – in spite of its score – the reason for keeping it will be clarified. Furthermore, if data is being kept, then its features will also be dissected. Additionally, the data that I did not get round to analysing will also be touched on.

It is safe to say that this research on the Bicleaner hard-rules and Moses cleaning algorithm specifications will help in finding out the extent to which the automatic cleaning tool, Bicleaner, discards useful training data. As a result of my findings, it is possible that more hard-rules will be introduced or that current hard-rules that are too stringent will be removed. In other words, new categories of saving data may be created, which will not only prevent the loss of genuine training data, but will also create better training material for NMT systems.

4. Research Project Findings

As aforementioned, a bilingual parallel unidirectional English-German (EN-DE) corpus from the European Medicines Agency (EMEA), which contained 1,108,752 sentence pairs, was cleaned using Bicleaner as part of my project. 19,355 sentence pairs were discarded, which formed the corpus that was closely analysed in this study. All of these sentence pairs were automatically scored between 0 to 0.5 by the Bicleaner classifier. Of these sentence pairs, a total of 515 sentence pairs were categorised into the following groups:

Score group 1: Scores between 0.5 and 0.3 – 242 sentence pairs

Score group 2: Scores between 0.3 and 0.2 – 124 sentence pairs

Score group 3: Scores between 0.2 and 0 – 48 sentence pairs

Score group 4: Scores of 0 – 101 sentence pairs

The findings of this analysis are discussed in the following subsections.

Score Group 1: Scores between 0.5 and 0.3

The following scores between 0.5 and 0.3 were analysed:

0.497
0.487
0.312
0.31
0.307
0.305
0.302

Table 6: Scores analysed between 0.5 and 0.3

Of the 242 sentence pairs that achieved a score within this range, six of the sentence pairs had data that were worth keeping. This is discussed in the following chapter. The problematic sentence pairs from the mined data were divided into different categories (Moorkens et al.,

2014; Negri et al., 2017; Defauw et al., 2019). The categories, with examples from the corpus, were as follows:

- **Capitalisation**

Acute Kidney Injury (AKI) remains a frequent and serious complication in hospitalized individuals worldwide.	Neuere Beobachtungen lassen vermuten, dass bei behandelten Patienten ein höheres Risiko für die akute Nierenschädigung besteht. ZELLBASIERTE THERAPIEN DER AKI

There is capitalisation in the target text which is not present in the source text. The capitalised text looks like a heading, but it is written as though it continues on from the previous sentence. The source-target length ratio here may also be an issue.

- **Content addition**

Histologically, reactive granulation tissue was confirmed.	Histologisch stellte sich bei beiden Patienten ein reaktiver Granulationsgewebspolyp dar.

Here, we can see that the phrase “bei beiden Patienten” – that is not present in the source text – has been added into the target text. We also notice that the wrong terminology has been used in the target text. The term “granulation tissue” has been translated as “Granulationsgewebspolyp”, which is incorrect. The focus of this example is, nonetheless, on the extra content added into the target text. This could also be rejected by the Moses cleaning algorithm specification “everything lowercased”. In other words, this sentence pair could be rejected on account of containing capital letters.

- **Content omission**

4 days later, the patient presented with distinct patches of urticaria at the	Hieraufhin entwickelte der Patient eine lokale allergische Reaktion in Form von

injection sites, which regressed spontaneously over the course of 2 weeks.	Quaddeln an den Injektionsstellen, die sich im Verlauf zurückbildete.

In this case, the phrases “4 days later” and “over the course of 2 weeks” are missing from the target text. Although the omission does not change the meaning of the target text, an important piece of information has been left out, which renders the target text not worth using. This sentence pair could also be rejected based on the Moses algorithm specification “everything lowercased”.

- **Difference in information in source and target**

17 pelvic and ureteral stones were found.	17 Nierenbecken- und Uretersteine wurden gefunden.

The term “pelvic stone” was translated incorrectly in the target segment above (IQWiG, 2020; Uniklinik, 2016). The mistranslation changes the meaning of the source text and as a result, there is a discrepancy between the information provided by the source and target texts. Another possible reason for the rejection of this sentence pair could, once again, be the Moses cleaning algorithm specification “everything lowercased”.

- **Duplicates**

Different from the previous guideline, the GRADE system was discarded and replaced by the Oxford evidence classification system which allows a more differentiated judgement.	Abweichend von der Vorgängerleitlinie wurde in der aktuellen Überarbeitung nicht mehr das GRADE-System sondern die Oxford Evidenzsystematik mit drei Empfehlungsgraden (A, B, C) verwendet, weil dieses System eine differenziertere Betrachtung erlaubt.
Different from the previous guideline, the GRADE system was discarded and	Abweichend von der Vorgängerleitlinie wurde in der aktuellen Überarbeitung nicht mehr das GRADE-System sondern

replaced by the Oxford evidence classification system which allows a more differentiated judgement.	die Oxford Evidenzsystematik mit drei Empfehlungsgraden (A, B, C) verwendet, weil dieses System eine differenziertere Betrachtung erlaubt.

These two sentence pairs appeared twice in this part of the corpus. Both the source and target segments in both pairs are exactly the same. An interesting observation here is that some content has been added, namely the phrase "mit drei Empfehlungsgraden (A, B, C) verwendet", which is not at all present in the source text. Content addition can be employed as a translation strategy to make the target text more reader-friendly, and this appears to have been the case here. For this reason, this sentence pair could still be useful. The presence of brackets in the target sentence, however, might have triggered the filter and led to this sentence pair being rejected. Use of the Moses cleaning algorithm specification "everything lowercased" could also apply here.

- **Incomplete translations**

This observation may introduce therapeutic options against a novel antimicrobial target in enterococci.	In der vorliegenden Studie sollte der Zusammenhang zwischen der Vancomycinresistenz bei

Here, we can see that the German target text is incomplete. Along with that, the translation is incorrect, meaning that this is both a misalignment and an incomplete translation. Another reason for the rejection of this sentence pair could be the Moses cleaning algorithm specification "everything lowercased".

- **Incorrect grammar**

A stepwise-selective procedure is warranted for the sake of efficiency and public health ethics.	Ein stufenweis-selektives Vorgehen gebietet die Effizienz und die Public Health Ethik.

The verb "warranted" has been translated as "gebietet" in the target text, which is incorrect and changes the meaning of the source sentence. After using minimal human

effort to correct this mistranslation, this sentence pair could be classified as useful data. This sentence pair was probably rejected by the Bicleaner classifier for not being consistent with the translations that it was trained with. This sentence pair may, too, have been rejected by the Moses cleaning algorithm specification "everything lowercased".

- **Incorrect terminology**

2. How would you describe the visibility of the squamocolumnar junction and categorize the transformation zone?	2. Wie sind die Einsehbarkeit der Plattenepithel-Zylinderepithelgrenze und der Typus der Transformationszone zu bewerten?

Here, the term "squamocolumnar junction" has been translated incorrectly as "Plattenepithel-Zylinderepithelgrenze" in the target text. The correct term is "Plattenepithel-Zylinderpithel-Grenze" (Kühn, 2011, p. 497). In the German term, a second hyphen is missing and the word "grenze" should be a separate word that starts with a capital letter. Another lexical error in the target sentence is the term "Einsehbarkeit", a translation of the English term "visibility". A more common and accepted term would be "Sichtbarkeit". Together with this, the source-language verbs "describe" and "categorize" cannot be seen anywhere in the target text. Instead, the verb "bewerten" has been used, which is not at all present in the source text. Just like the previous sentence pair, this sentence pair was probably rejected by the Bicleaner classifier since it did not match the translations that the classifier was trained with. Another possible reason for rejection here could be the Moses cleaning algorithm specification "everything lowercased".

- **Incorrect translation**

After primary stabilisation of the patients, imaging studies should be performed to assess the extent of the injuries and determine the treatment of choice.	Nach der initialen Stabilisierung stellt die Bildgebung eine essenzielle Grundlage der richtigen Versorgung und einer folgenden Operation dar.

A very clear case of an incorrect translation can be seen in the sentence pair above. The target segment more or less literally states, that “after initial stabilization, (medical) imaging is an essential basis for proper care and subsequent surgery”, which is not what the source text says or means. The target text is missing important terms like “stabilization of the patients”, “extent of injuries” and “determination of treatment of choice”, which is why this sentence pair is incorrect. The Bicleaner classifier probably rejected this sentence pair because it did not agree with the translations it had been trained with. The Moses cleaning algorithm specification “everything lowercased” could also be the reason why this sentence pair was rejected.

- **Misalignment**

[CF Lung Disease - a German S3 Guideline: Module 2: Diagnostics and Treatment in Chronic Infection with Pseudomonas aeruginosa].	Mukoviszidose (Cystic Fibrosis, CF) ist die häufigste, autosomal-rezessiv vererbte Multisystemerkrankung.

This is a very clear case of misalignment. The source and target texts do not match at all and are completely unrelated to each other. Furthermore, the lack of numbers in the target text, the different types of brackets used and the difference in punctuation likely triggered the filter. As was the case with the aforementioned examples, the Moses algorithm specification “everything lowercased” may have resulted in the rejection of this sentence pair.

- **Missing terminology**

No complications and/or episodes of gastric volvulus were detected at a 3-month minimum follow-up.	Innerhalb einer Beobachtungszeit von mindestens 3 Monaten nach dem Eingriff traten keine Komplikationen auf.

The source term “gastric volvulus” is missing from the target text. It is also likely that this sentence pair was not comparable with the translations that the Bicleaner classifier was trained with, which may be the reason why it was rejected. What is

more, the Moses cleaning algorithm specification “everything lowercased” may have played a role in the rejection of this sentence pair.

- **Sentence not in the desired language**

Nevertheless, the diagnostic protocol for CD was not changed in the most parts in the new edition of the DSM-5; the addition of a CD specifier with limited emotions is the most relevant change.	Jedoch sind die wesentlichen diagnostischen Kriterien im DSM-5 gleichgeblieben; eine entscheidende Änderung ist die nun mögliche Klassifikation eines CD specifiers with limited prosocial emotions.

The highlighted part of this target segment has been left untranslated and simply copied and pasted from the source text. Some content has also been added here, in that the word “prosocial” does not appear in the source text. As such, not only is the desired language absent in some parts of the target text, but the content that has been added is not in the desired language either. It is also interesting to see that even though the untranslated part is at the end of the target sentence and was not strictly copied word for word from the source segment, the algorithm still managed to detect this error.

- **Wrong source text**

Anamnestic clinical complaints were a syncope associated with paraparesis and weak femoral pules.	Die Anamnese ergab eine Synkope sowie eine Paraparese und schwachen femoralen Puls.

Here, the source text seems to contain a spelling error. Instead of “pulse”, the word included here is “pules”. Nevertheless, this term has been translated as “Puls” in the target text, which is indeed a translation of the English “pulse”. This makes the entire sentence pair ambiguous in its meaning.

As we can see, all of the errors found in the sentence pairs can be divided into different error categories. It can also be said that most of these sentence pairs can be filed under several error categories rather than just one, but the examples given above highlight the critical errors

in each of the chosen segments. The Bicleaner hard-rules and the Moses cleaning algorithm seem to have worked well here, as the sentences were probably rejected based on their filters:

Bicleaner: "The length ratio between the source sentence and target sentence (in bytes) is too low or too high"

Moses: "Everything lowercased"

"Gambling disorder" is the only behavioral addiction added to the DSM. Furthermore, preliminary criteria for "Caffeine Use Disorder" and "Internet Gaming Disorder" have now been defined in the manual.	Als einzige «Verhaltenssucht» wurde die «Glücksspielstörung» (Gambling Disorder) in das DSM-5 aufgenommen, zusätzlich wurden vorläufige Kriterien für eine «Koffeingebrauchsstörung» (Caffeine Use Disorder) sowie für eine «Internetspielstörung» (Internet Gaming Disorder) definiert und in das Kapitel III (Störungsbilder, die weiterer Forschung bedürfen) integriert.

Here, the source text contains 200 characters including spaces, while the target text contains 364 characters including spaces. The target segment is over one and a half times longer than the source segment. The rule "too many parenthesis or brackets in sentence" can also be applied here, since brackets are used several times throughout the target sentence. The Moses cleaning algorithm specification "everything lowercased" could also be applied here, since the translation contains a lot of capital letters. There are also tags and numbers in the target text, which are missing from the source.

Bicleaner: "There are words repeated consecutively"

Moses: "Sentence aligned (one sentence per line)"

[The interparental relationship in families with children with ADHD: Interactions	Zusammenfassung. Der Einfluss der Familie in Bezug auf die Entstehung und

between couple distress and child's symptoms]. The interparental relationship in families with children with ADHD: Interactions between couple distress and child's symptoms Abstract.	Aufrechterhaltung der ADHS bei Kindern ist wissenschaftlich gut fundiert, jedoch hat sich die einschlägige Forschung weitgehend auf das elterliche Erziehungsverhalten oder die Qualität der Eltern-Kind-Beziehung konzentriert.

It can be seen here that the source text contains the title of the article twice. The second time it is mentioned, it is also followed by the subheading “Abstract”. In addition, brackets are used in the source segment, and the content inside these brackets is missing in the target text. Furthermore, according to the Moses cleaning algorithm, there should only be one sentence per line, which is not the case here.

Score Group 2: Scores between 0.3 and 0.2

In the second score group i.e. scores between 0.3 and 0.2, the following scores were analysed:

0.3
0.297
0.295
0.292
0.29

Table 7: Scores analysed between 0.3 and 0.2

5 scores between 0.3 and 0.2 and a total of 124 sentence pairs were analysed. None of these pairs could be considered useful data. Just like in the first score group, these sentence pairs were also divided into different error categories. The four error categories used are listed below with examples:

- **Content missing**

[Major depression and liver disease: the role of microbiome and inflammation]. Depression and liver disease are closely associated.	Zwischen Depressionen und Lebererkrankungen besteht ein enger Zusammenhang.

Here, we can see that the target segment is missing content from the source text, which makes it an incorrect and inaccurate translation of the original. This sentence pair seemingly does not match with the translations that the Bicleaner classifier has been trained with. Along with that, two Moses cleaning algorithm specifications could be the reason as to why this sentence pair was rejected, namely "everything lowercased" and "sentence aligned (one sentence per line)".

- **Incorrect translation**

Yet, despite significant improvements there is a high rate of nonresponse.	Zugleich zeigen sich hohe Nonresponseraten.

This example likewise shows a target segment that is an incorrect translation of the source text. Only one part of the source text (the highlighted text) has been carried over to the target text. Even though the source and target fragments express some of the same meaning, the target segment omits a piece of information from the source, making it an inaccurate translation. This sentence pair seemingly did not line up with the Bicleaner classifier's training material, which is probably why it was rejected. The Moses cleaning algorithm specification "everything lowercased", however, may also have been responsible for the rejection of this sentence pair.

- **Misalignment**

In addition, there is a new inhibitor of von-Willebrand-Polymerisation available.	THERAPIE Bis vor wenigen Jahren waren Plasmapheresen und ggf. immunsuppressive Behandlungen die einzigen wirksamen Behandlungsansätze.

Here, it is clear that the source and target segments are completely unrelated to each other. As well as this, a word has been capitalised in the target text, and no punctuation separates this word from the first sentence.

- **Wrong source text**

Management of a giant perineal condylomata acuminata. A condylomata acuminata infection is caused by human papillomaviridae (HPV).	Condylomata acuminata werden durch humane Papillomaviren verursacht.

In this example, the source term "human papillomaviridae (HPV)" is incorrect. The correct term is actually "human papillomavirus (HPV)" (CDC, 2022). The term "papillomaviridae" on its own would have been correct (Van Doorslaer et al., 2018), but since we know that this source term is supposed to be the full form of the abbreviation HPV (i.e. "human papillomavirus"), we can be certain that the source text is wrong. Moreover, only part of the source text has been translated and included in the target text. Even though this translation is indeed correct, the target text is incomplete as the source content in red has been omitted entirely. Another possible reason for rejection might be that there are two sentences in the source and only one in the target segment. As this means that there is more text in the source than the target, length ratio could be a problem here. Along with that, the Moses cleaning algorithm specifications "everything lowercased" and "sentence aligned (one sentence per line)" could also explain why this sentence pair was rejected.

The Bicleaner hard-rules seem to have worked well here too. Below are some of the common Bicleaner hard-rules and Moses cleaning algorithm specifications on the basis of which sentence pairs in these parts of the corpus may have been rejected:

Bicleaner: "The length ratio between the source sentence and target sentence (in bytes) is too low or too high"

Moses: “Sentence aligned (one sentence per line)”

Management of a giant perineal condylomata acuminata. A condylomata acuminata infection is caused by human papillomaviridae (HPV).	Condylomata acuminata werden durch humane Papillomaviren verursacht.

In this sentence pair, the source text has 131 characters including spaces, while the target text has only 68 characters including spaces. The source segment here is over one and a half times longer than the target segment. The general Bicleaner ratio for this project was r1.5, which means that the target could be up to 1.5 times longer than the source. It is interesting, however, that it is the source that is close to 1.5 times longer than the target text in this case. In relation to the second filter, this sentence pair does not satisfy the Moses cleaning algorithm specification of one sentence per line either.

Bicleaner: “Too many parenthesis or brackets in sentence”

Moses: “Everything lowercased”

In all, 55 cases (83.3%) were available for clinical follow-up examination after an average of 59.0 ± 20.7 months (range: 25-96 months) and of these, 48 (72.7%) patients consented to radiological evaluation to determine healing and position of the greater tuberosity.	Mittels detaillierter Anamnese, 4 klinischen Scores – Constant-Murley Score, WOSI (Western Ontario Shoulder Instability Index), Rowe Score, subjektiver Schulterwert („subjective shoulder value“, SVV) – und Bestimmung des Bewegungsumfangs im Schultergelenk wurden Schulterfunktion und Schulterstabilität erhoben. Eine etwaige Dislokation des Tuberculum majus wurde mittels Röntgenaufnahmen in 3 Ebenen (a.-p., seitlich/Y-View, axial) analysiert.

Brackets are used several times throughout this target sentence. Additionally, this sentence pair has been misaligned, as the target segment is not a translation of the source segment that it has been matched with. The missing numbers in the target text could also be a reason for the rejection of this sentence pair. This sentence pair, moreover, does not satisfy the Moses cleaning algorithm specification “everything lowercased”.

Interestingly, there is only one sentence pair in the entire dataset which contains a URL, and this sentence pair falls into this score group. The sentence pair is as follows:

The brochure is available on the webside oft he swiss society of paraplegia www.ssop.ch.	Das aktualisierte Konzept der Vorsorge und Nachsorge für Querschnittgelähmte, berücksichtigt die spezifischen Probleme dieser Patienten.

Bicleaner gave the example above a fluency score of 0.31. This sentence pair may have been rejected due to the source segment containing a URL, since one of Bicleaner’s rules is “There are URLs”. Additionally, there is a spelling mistake in the source text, which is highlighted in green. As well as this, the source-target length ratio could have been a problem here, but the rejection of this sentence pair may also have been triggered by the Moses cleaning algorithm specification “everything lowercased”.

Score Group 3: Scores between 0.2 and 0

The next subsection analyses sentence pairs with scores between 0.2 and 0. This is one of the smallest score groups in terms of how many sentence pairs it contains, but I analysed more sentence pairs from this group than I did from any of the other three groups. The scores analysed here are as follows:

0.2
0.195
0.193
0.19
0.188

0.185
0.182
0.177
0.175
0.172
0.17
0.168
0.165
0.16
0.158
0.147
0.142
0.135
0.13
0.115
0.11

Table 8: Scores analysed between 0.2 and 0

In this score group, I analysed 21 scores and 48 sentences pairs which did not contain any useful data. All of the sentence pairs were rejected on the basis of five main categories, which were as follows:

- **Citation**

CITATION FORMAT · Kuetting D, Pieper CC . Percutaneous Treatment Options of Lower Urinary Tract Fistulas and Leakages. Fortschr Röntgenstr 2018; 190: 692-700.	· Interventionelle Radiologen sollten mit den gängigen Techniken der transrenalen Ureterokklusion vertraut sein..

Here, the source segment is a citation which does not necessarily need to be translated. The target text, however, is a translation, though not of the citation it has been matched with. As such, this is both a case of misalignment, as well as an error relating to citation. Together with this, numbers are also missing from the target

segment. In addition, some words are capitalised, which neither satisfies the Bicleaner hard-rule regarding capitalisation nor the Moses cleaning algorithm specification "everything lowercased".

- **Incorrect translation**

If a surgical approach is necessary, minimally invasive surgery in the hands of an experienced laparoscopic surgeon is a suitable option.	Bei entsprechender Expertise können Eingriffe laparoskopisch erfolgen.

In this sentence pair, only the highlighted part of the source text has been translated into German. Despite the fact that the source and target sentences are partially related to each other, the translation is still incorrect and inadequate, as it does not capture all of the meaning of the source. Along with this, content is also missing from the target text. Nonetheless, this sentence pair could have been rejected due to the source-target length ratio. As with the aforementioned examples, the Moses cleaning algorithm specification "everything lowercased" could also have been responsible for the rejection of this sentence pair.

- **Incomplete translations**

So chaetocin seems to be no suitable agent for specific targeting ccRCC cells or for the combination therapy with CIK cells in renal cancer.	In 2007 wurde erstmals berichtet, dass Chaetocin potente und selektive

The source and target texts here are completely unrelated to each other. Not only is the German translation incorrect, but the target segment is also incomplete, which affects the source-target length ratio. Furthermore, there are capital letters between lower case letters in the source segment, and there are also numbers present in the target text which have been omitted from the source. In this case, the capital letters do not satisfy the Moses cleaning algorithm specification "everything lowercased" either.

- **Missing content**

The German quality indicators in intensive care medicine 2013--second edition. Quality indicators are key elements of quality management.	Qualitätsindikatoren sind elementare Bestandteile des Qualitätsmanagements.

The content highlighted in the source text is missing from the target text, which means that the target text is incomplete. Once again, this has a misleading impact on the length ratio of the sentence pair. Together with this, the Moses cleaning algorithm specification “everything lowercased” could explain why this sentence pair was rejected.

- **Misalignment**

An evaluation study concluded that this model is feasible and addresses an existing need. The assessment option is experienced as helpful by patients.	Die interdisziplinäre Zusammenarbeit mit Umweltfachpersonen hat sich bewährt.

Here, it can be said that the source and target texts are not related to each other. This is a very clear case of misalignment. Besides this, the length ratio of the sentence pair is affected since there are two sentences in the source segment and only one in the target segment. The Moses cleaning algorithm specifications “everything lowercased” and “sentence aligned (one sentence per line)” could also cause this sentence pair to be rejected.

Most of the sentence pairs in this score group have been misaligned. Hardly any of the translation fragments in this group are exact equivalents of the source fragments they have been matched with, meaning that there is little useful data here. In this score group, only three Bicleaner filters seem to be applicable, namely “The sentence pair has low fluency score from the language model”, “There are words in the sentence containing too many uppercased characters between lowercased characters”, and “The length ratio between the source sentence and target sentence (in bytes) is too low or too high”. Two Moses cleaning

algorithm specifications could also be applied here, namely “everything lowercased” and “sentence aligned (one sentence per line)”. All of the sentence pairs in this group have received low scores and almost all of them are misaligned.

Score Group 4: Scores of 0

The next set of sentence pairs that I analysed were those that were given a score of 0 – the lowest possible score on the Bicleaner scale. The corpus used for this project was scored from 0 to 0.5, so no sentence pairs scored more than 0.5 or less than 0. 108 sentence pairs from this score group were analysed for the purpose of my study and the following error categories were identified:

- **No sentences**

, 2008).	, 2008</citationReference>).

The example above is not a full sentence. There are also tags in the target segment.

- **Incomplete sentences**

, with the aim of helping the patient to carry out activities in his important areas of life and to largely independently create his own life. In rehabilitation clinics or in outpatient settings, occupational therapists pass on these treatment measures, adapting the therapy contents to the status and possibly changed needs of the patient and his goals in the rehabilitation process.	Gemeinsam mit dem Patienten werden Therapieziele festgelegt und die Therapie gestaltet.

In this sentence pair, the source fragment is incomplete, and both the source and target segments are misaligned. Along with that, the source and target texts differ considerably in length. Moreover, the sentence pair does not satisfy two of the Moses

cleaning algorithm specifications, namely “sentence aligned (one sentence per line)” and “everything lowercased”.

- **Incomplete translation**

, the problems of incomplete penetrance, variable expressivity and possible oligogenic inheritance) have to be explained to the families.	Eine nichtkodierende Hexanukleotidrepeat-Expansion des

Here, the German translation is both incomplete and incorrect, meaning that the sentence pair is also misaligned. Together with this, there is a clear difference in length between the source and target segments. The Moses cleaning algorithm specifications “everything lowercased”, however, could also have been responsible for the rejection of this sentence pair.

- **Incorrect terminology**

(total score range: 0-48).	(gesamter Score-Bereich: 0–48).

This sentence pair cannot be considered useful because of the incorrect use of terminology. The highlighted word “range” has been wrongly translated as “Bereich” in the target segment. The correct translation would have been “Weite”. Had it not been for this error, this sentence pair would have been entirely correct. It was likely rejected by the Bicleaner classifier on account of not agreeing with the translations that the classifier was trained with. Furthermore, this sentence pair does not satisfy the Moses cleaning algorithm specification “everything lowercased”. The sentence pair could nevertheless still be viewed as useful, since it only contains one error that would take little human effort to correct.

- **Incorrect translation**

(1) Families with weighted levels of psychosocial burdens reported an enhanced need for help. (2) Midwives	Fortgebildete Hebammen und Pflegefachkräfte betreuen Familien mit psychosozialem Hilfebedarf.

and nurses with additional qualification support more frequently families with high levels of psychosocial burdens.	

This is a case of an incorrect translation, whereby the target segment does not convey all of the source meaning. A few individual words from the source text have indeed been translated but overall, the target text is inadequate. There are also numbers present in the source which are missing from the target fragment. The source-target length ratio could also be a problem here. Besides this, the Moses cleaning algorithm specification of one sentence per line is not satisfied.

- **Misalignment**

[19th century Russian research about collective behavior].	Im 19. Jahrhundert erlebte die Massenpsychologie als wenngleich randständiges sozialpsychologisches und psychiatrisches Forschungsfeld doch eine beachtenswerte Publikationstätigkeit. Heute beschränkt sich die medizin- und psychologiehistorische Darstellung vor allem auf deutsch-, französisch- und englischsprachige Beiträge über induzierte psychische Massenerscheinungen.

Most of the sentence pairs in this group, such as the one above, have been misaligned. Other issues concerning the rejected source and target texts here could be the length ratio and the sentence pairs not complying with the Moses algorithm specification "everything lowercased".

Just as in the other three score groups, the Bicleaner hard-rules and the Moses cleaning algorithm were successful in filtering out unsuitable data. Multiple hard-rules and cleaning

algorithm specifications can be seen to have been involved in rejecting the sentence pairs in this score group i.e. those that scored 0. Some of these specificiations are outlined below:

Bicleaner: "Sentence is less than 3 words long"

: 55, max.	: 55, Max.

This sentence pair violates the Bicleaner hard-rule of a sentence containing more than 3 words.

Bicleaner: "The length ratio between the source sentence and target sentence (in bytes) is too low or too high"

Moses: "Sentences longer than 100 words (and their corresponding translations) have to be eliminated (note that a shorter sentence length limit will speed up training)"

[A Process-Oriented Approach at Current Recommendations for Obstetric Anesthesia and Postoperative Monitoring After C-Section]. The known guidelines before a planned operation on aspiration, fasting and preoperative risk evaluation also apply in obstetrics. Extended measures are only justified under concrete anamnestic or specific symptoms. Neuraxial anesthesia techniques should be offered to the mother as early as possible, as waiting for a certain opening of the cervix is not justified. Catheter procedures offer numerous advantages and are useful for possible emergency situations. Low-dose local anesthetic concentrations in combination	Die von ASA (American Society of Anesthesiologists) und SOAP (Society for Obstetric Anesthesia and Perinatology) für das Jahr 2016 aktualisierten Leitlinien sind eine Handlungsempfehlung in erster Linie für Anästhesisten, die auf das anästhesiologische Management Gebärender, nicht operative und operative Entbindung sowie auf die postpartale Versorgung und Analgesie fokussiert.

with an opioid are still recommended. The benefit of pencil-point spinal needles in minimizing the risk of post-puncture headache has been demonstrated. Predictable emergencies are airway emergencies, hemorrhagic emergencies and cardiopulmonary resuscitation with emergency cesarean if appropriate (> 20 SSW).	

In this example, the source text contains 962 characters including spaces, while the target text only consists of 379 characters including spaces. The difference in length here is prominent as the source segment is over one and a half times longer than the target segment. It is, however, worth noting that the target fragment is missing a lot of source content, which also means that these two fragments are misaligned. The general Bicleaner ratio for this project is r1.5, which means that target segments can be up to 1.5 times longer than the source segment. As a result, it may seem striking that the original is longer than the translation. When we take the omission of source content into consideration, however, the reason for this difference in length becomes clear. Alongside this, the Moses cleaning algorithm specification "sentences longer than 100 words (and their corresponding translations) have to be eliminated (note that a shorter sentence length limit will speed up training)" could also be applied here, since the source segment is 124 words long.

Bicleaner: "There are words in the sentence containing too many uppercased characters between lowercased characters"

Moses: "Everything lowercased"

[Acute Kidney Injury, AKI - Update 2018].	PROPHYLAXE DER KM-INDUZIERTEN AKUTEN NIERENSCHäDIGUNG (AKI) (KONTRASTMITTELNEPHROPATHIE, KMNP): Eine im November 2017

	hochrangig publizierte Studie zeigt, dass die Anwendung von Natriumbikarbonat der Gabe von Natriumchlorid zur Prävention der Kontrastmittelnephropathie nicht überlegen ist. SGLT-2-ANTAGONISTEN UND AKI: SGLT-2-Antagonisten entfalten mutmaßlich protektive Effekte hinsichtlich einer CKD (chronic kidney disease).

The text highlighted in the target segment here shows capital letters surrounded by lower case characters. In this example, the rule "Too many parenthesis or brackets in sentence" has also been broken. The source-target length ratio is, furthermore, likely to have been a deciding factor in the rejection of this sentence pair. In addition to this, the Moses cleaning algorithm specification "everything lowercased" could have been applied here.

A number of Bicleaner hard-rules were employed to clean my dataset. Some, however, were not used at all. The unused Bicleaner hard-rules are outlined below:

no_only_numbers	The ratio of numeric characters in source sentence is too high
no_breadcrumbs	There are more than 2 breadcrumb characters in the sentence
no_unicode_noise	Too many characters from unwanted unicode in source sentence
no_space_noise	Too many consecutive spaces in sentence
no_escaped_unicode	There is unescaped unicode characters in sentence
no_bad_encoding	Source sentence or target sentence contains mojibake
no_titles	All words in source sentence or target sentence are uppercased or in titlecase
no_porn	Source sentence or target sentence contains text identified as porn

These seven rules did not apply to any of the sentence pairs in my dataset.

Likewise, the following two Moses cleaning algorithm specifications were not relevant to any of the sentence pairs in my dataset:

Removes empty lines

Removes redundant space characters

These findings will be discussed in further detail in the next chapter.

5. Discussion of Research Project Findings

The previous chapter described the findings of my research project in detail. This chapter will explore these outcomes more closely in the aim of reaching a conclusion.

The very first thing to note, is that among the 515 sentence pairs that were rejected from my corpus, some of the sentence pairs do indeed contain useful data that is worth saving. The features of the useful data I recovered are discussed in the following subsections.

5.1. Discussion of Analysed Data

5.1.1 Useful Data

Overall, a total of eight rejected sentence pairs were found to be worth keeping. Six of these sentence pairs were given scores between 0.5 and 0.3, and the other two received scores of 0.

Of the sentence pairs that obtained scores between 0.5 and 0.3, I analysed a total of 242 sentence pairs that received seven different scores. The sentence pairs in this score group fit into 13 error categories, which is more error categories than the sentence pairs from the other score groups could be filed under. Despite this, six sentence pairs from this first score group were worth retaining. Two of these sentences were mentioned in the previous chapter. One of the sentence pairs scored 0.497 and included extra content that was not in the source text, which was probably a translation strategy employed for that particular sentence pair (cf. ‘Duplicates’ under ‘Score Group 1: Scores between 0.5 and 0.3’ in Chapter 4). In the second sentence, which also scored 0.497 (cf. ‘Incorrect grammar’ under Score Group 1: Scores

between 0.5 and 0.3' in Chapter 4), changing just one verb would have corrected the sentence pair.

It was mainly the sentence pairs that received the scores 0.497 and 0.487 that were very close to being acceptable, but these may have been rejected either because of the Bicleaner hard-rules, the Moses cleaning algorithm or the Bicleaner classifier. To be more specific, out of the six sentence pairs that were worth saving, five received a score of 0.497 and one scored 0.487. The remaining four sentence pairs worth saving are as follows:

All patients had been previously operated externally.	Alle Patienten waren vorhergehend auswärts operiert worden.

This sentence pair was scored 0.497. In this case, the highlighted term in the source text, "externally", has been translated incorrectly as "auswärts" in the target text. This is an error which would require minimal human effort to amend. A human would be able to recognise that this is a relatively minor lexical error that can easily be corrected by using the available context. As it is, however, this translation unit is unsuitable due to the incorrect use of terminology, which makes little sense in itself and does not convey the meaning of the source text at all. This sentence pair could also have been rejected by the Bicleaner classifier for not matching up with the translations that the classifier was trained with.

Financial burden has a significant influence on every aspect of service use.	Finanzielle Belastungen haben einen bedeutsamen Einfluss auf alle Aspekte der Inanspruchnahme.

This sentence pair was scored 0.497. Here, the highlighted source term "service use" has been translated incorrectly in the target text, which essentially renders this sentence pair useless. Although the rest of the translation indeed seems correct at first glance, it might not be entirely appropriate in the given context. In order to find out whether the translation is suitable for the target situation, human intervention and real-world knowledge would be required. One of these reasons might explain why this sentence pair was rejected by the

Bicleaner classifier. It could also have been discarded by the Moses cleaning algorithm for containing capital letters.

Medical students taught by student peers almost reached the same examination result as the group taught by paediatric teachers (21,7±4,1 vs. 22,6±3,6 of 36 points, p=0,203).	Die Studierenden, die von einem Tutor unterrichtet wurden erreichten dabei im direkten Vergleich eine ähnliche Gesamtpunktzahl, wie die Gruppen, die von einem pädiatrischen Dozenten ausgebildet wurden (21,7±4,1 vs. 22,6±3,6 von 38 Punkten, p=0,203).

The above sentence pair was scored 0.497. In it, we can see that the number "36" in the source text has been incorrectly replaced with the number "38" in the target text. The rest of the target content seems as though it corresponds with the source content, so it is likely that this sentence pair was rejected on the basis of the incorrect value. Nevertheless, the source-target length ratio and unwanted literals could also have presented issues. Additionally, the capital letters in this sentence pair could have triggered the Moses cleaning algorithm. Regardless of this other "noise" which might have caused the sentence pair to be discarded, fixing the value in the target segment using little human effort would make the translation unit useful.

The Isoflurane level on 2 farms was above the Swiss safety limits.	Auf 2 Betrieben wurde der in der Schweiz geltende Grenzwert der Isofluran-Konzentration überschritten.

This pair was scored 0.487. In this case, one wrong term (i.e. the German "Betrieben" for the English "farms", as highlighted in the target and source segments respectively) changes the intended meaning of the source text. This could be the reason as to why the Bicleaner classifier rejected this sentence pair. The differing lengths of the two fragments could also

have been an issue here. Along with that, the Moses cleaning algorithm specification "everything lowercased" might have resulted in this sentence pair being discarded. Changing this one term using minimal human effort would make this sentence pair useful.

Within the group of sentence pairs that scored 0, a total of 108 translation units were analysed. Of these, two sentences were found to be worth retaining. The first one is mentioned in the previous chapter (cf. "Incorrect terminology" under "Score Group 4: Scores of 0" in Chapter 4) and includes a single incorrect term that would take little human effort to rectify.

The second sentence pair worth keeping is an incomplete translation that would likewise be useful following minimal human intervention. The example is as follows:

[Acute scrotal pain in childhood: legal pitfalls]. Acute scrotal pain in childhood is an emergency.	Das akute Scrotum im Kindesalter ist ein Notfall.

The source sentence highlighted in red here has been omitted from the target segment. Even though the rest of the source content has been translated correctly, the fact that some target information has been left out renders the sentence pair incomplete. Bicleaner's rule of "Too many parenthesis or brackets in sentence" could also be applied here, since there are brackets in the source text. These brackets are, however, missing from the target text. This is an example of a sentence pair that could be rescued. While other sentence pairs can be made complete with the slight modification of a single value or term, this translation unit requires an entire sentence, albeit short, to be translated to be made useful.

What five out of six sentence pairs in Score Group 1 (i.e. scores between 0.5 and 0.3) and the first sentence in Score Group 4 (i.e. Scores of 0) have in common, is that you would only have to edit a single value or term in order to make these translation units useful. The remaining sentence pair that was worth saving from Score Group 1 included some extra content that was not in the source segment. As the addition of content did not change the meaning of the original in this case – and content addition could have been a translation strategy employed for this particular sentence – human intervention may not actually be

required to make this sentence pair useful. The length ratio and brackets could have triggered the filters to discard this translation unit.

5.1.2. Non-Useful Data

Score Group 1: Scores between 0.5 and 0.3

This subsection outlines some of the translation units that were discarded due to slightly more severe errors and missing elements which would require a considerable amount of human effort to fix. As a result, these sentence pairs are not worth saving.

The entire mattress costs approximately 80-90 € more than a normal mattress.	Die kompletten Matratzen sind ca. 80–90 € teurer als die normalen Matratzen.

In this example, the source text contains singular nouns, hence the verbs are also singular. The target text, on the other hand, contains plural nouns and verbs. This is both a grammatical and contextual error. With some human intervention, however, this sentence pair could still be rescued. This sentence pair was probably discarded by the Bicleaner classifier for not conforming with the translations that the classifier was trained with. It could also have been rejected by the Moses cleaning algorithm because of the capital letters.

The internal consistency amounted to α = .	Die interne Konsistenz betrug α = 0,94.

When it comes to this example, the target text has a value that the source text does not. The missing value explains why the Bicleaner classifier rejected this sentence pair. Otherwise, the translation is acceptable. The Moses cleaning algorithm specification “everything lowercased” could also have caused this sentence pair to be rejected.

It is worth noting that only one of the analysed sentence pairs in Score Group 1 (i.e. scores between 0.5 and 0.3) was not in the desired language. This sentence pair was discussed in further detail in the previous chapter.

It is also worth mentioning, that the number of errors increases as the scores decrease. The sentence pairs that received scores between 0.3 and 0.2, for instance, contain more errors than those that were given scores between 0.5 and 0.3. Sentence pairs that scored 0.497 and 0.487 include a wide range of errors which could either have been triggered by the Bicleaner or the Moses filters – and hence led to the data being discarded – or they could have been rejected by the Bicleaner classifier. As soon as the scores hit 0.312, there was a major difference in the corpus in the sense that the errors were less varied, but the degree of the errors increased. Most of the errors in the sentence pairs that scored between 0.3 and 0.2 were incorrect translations and misalignments. Upon looking at the sentence pairs, it is clear that the source and target texts progressively become more and more unrelated to each other as the scores decrease. It also, therefore, becomes obvious why this data was rejected by the cleaning tool.

We already know that a total of 13 error categories belong to Score Group 1 (i.e. scores between 0.5 and 0.3). Having an overview of the percentage of sentence pairs belonging to each category makes it easier to evaluate the overall quality of the corpus. The error category that constitutes the most sentence pairs (68%) is “Incorrect translation”. The second largest error category is “Misalignment”, which contains 55% of the sentence pairs. “Content omission” is in third place with 12% of the sentence pairs, followed by “Content addition” (9%,) “Incorrect terminology” (8%), “Incomplete translations” (3%), “Missing terminology” (2%), “Capitalisation” (1.2%), and “Differences in information” and “Duplicates” (0.8% respectively). The smallest error categories are “Sentence not in the desired language” and “Wrong source text”, which each contain 0.4% of the sentence pairs. As aforementioned, the degree of errors increases as the scores decrease.

Consequently, it is safe to say that most of the errors found within this score group are related to incorrect translations.

Score Group 2: Scores between 0.3 and 0.2

In the next score group (i.e. scores between 0.3 and 0.2), a total of 124 sentence pairs with 5 different scores were analysed. Compared to the first score group, far fewer error categories were relevant to the sentence pairs here. To be more precise, errors from only four different

categories were found within this score group. There was no useful data found among these sentence pairs.

Unlike in the previous score group, however, the types of errors are spread more evenly throughout the analysed data here. With the sentence pairs that scored between 0.5 and 0.3, I noticed that more different kinds of errors were present when the score was closer to the decided threshold of 0.5. Furthermore, the higher the score, the closer the sentence pairs were to being acceptable. As the scores declined, the less useful the data became.

Looking at the percentages of data in each of the error categories for this score group, we can see that almost 97% of the sentence pairs contain errors in the categories “Incorrect translation” and “Misalignment”, followed by 2% in “Content missing” and 1% in “Wrong source text”. From this, we can deduce that the biggest error categories in this group are “Incorrect translation” and “Misalignment”. For example:

Degenerative changes were most commonly found, followed by congenital defects and neoplasms.	Letztgenannte waren in sechs Fällen weiter abklärungsbedürftig.

In the example above, it is very clear that the source and target text are unrelated to one another, hence these two segments have been misaligned. This mismatch could have resulted in the sentence pair being rejected by the Bicleaner classifier. The Moses cleaning algorithm specification “everything lowercased” could have been triggered as well.

These two categories constituted most of the errors in the previous score group too. In this group, however, the percentage of data within these error categories has jumped up by 42% for “Incorrect translation” and by 77% for “Misalignment”, demonstrating that as the scores move further away from the threshold, the number of errors increases.

Score Group 3: Scores between 0.2 and 0

Of all the score groups, I analysed the least sentence pairs and most scores from this group, namely 21 different scores. Five error categories were found here, which is one more than

was identified in the previous score group (i.e. scores between 0.3 and 0.2). One error category which was not found in any of the score groups except this one was “citation”. This category refers to the translation of a citation in the source text which was actually supposed to remain untranslated in the target text. Furthermore, there was not any useful data found in this group.

Along with the citation error, there were other errors within this score group, such as “Capitalisation”. For example:

CONCLUSION Opening locked psychiatric wards can help to establish a positive therapeutic atmosphere without changing the therapeutic climate on the other already open wards.	Zwangsmaßnahmen und Entlassungen gegen ärztlichen Rat wurden reduziert. SCHLUSSFOLGERUNG

In the example above, we can see that there is capitalisation in both the source and target texts. Although neither of the words in capital letters are between other words in lower case, they do not correspond with each other because they are not in the correct place in both segments. As well as this, the source and target texts are completely unrelated to each other, hence they are misaligned.

One example of an incomplete sentence has already mentioned in the previous chapter. Another example is outlined below:

An increase of the offer and the additional qualification is recommended for improving the developmental and living conditions of families with psychosocial burdens.	Die Angaben der Mütter mit einer Betreuung durch eine fortgebildete Gesundheitsfachkraft (Gruppe

This example shows the target text being incomplete and the source and target texts being unrelated to each other. As such, this sentence pair contains both the error of an incomplete sentence and the error of misalignment.

This example also illustrates how the quality of the translations decreases as the scores get further away from the acceptable threshold of 0.5. From this, we could also infer that these sentence pairs require a lot of human intervention to turn them into useful training data. Moreover, Bicleaner classifier was trained to discard sentences that do not match up with the translations it was trained with. As such, "The sentence pair has low fluency score from the language model" seems to have been the Bicleaner rule that was used most frequently to eliminate data from this group. The source and target segments in all of these sentence pairs also progressively become more and more unrelated to each other in the sense that the differences begin as only slight mismatches with some content missing, but then become sentences that mean completely different things.

Another peculiar thing I noticed with regard to this group is that it does not contain many sentence pairs, even though it is the score group with the most scores analysed. This is because this score group contained few sentence pairs and few scores overall, hence all of the scores and translation units were analysed. In contrast, the other score groups contained a vast amount of data, so the scores analysed were chosen at random. In this score group, for example, I analysed all 6 sentence pairs that were given a score of 0.2, all 4 pairs that were given a score of 0.195, the 1 that received a score of 0.193, the 2 that got a score of 0.19, and so on.

The percentages of sentence pairs in each of the errors categories for this score group show that the error categories "Content missing", "Incorrect translation" and "Misalignment" each have the highest share of data within them with 98% respectively, followed by "Incomplete sentence" with 4% and "Citation" with 2%. Similar to the other score groups, the majority of errors here are related to the categories "Misalignment" and "Incorrect translation". The only difference here is that there is an additional error category that contains the same percentage of sentence pairs, namely "Content missing".

When the percentages in these error categories are compared with those of the other score groups, there are considerable jumps. To elaborate, the number of sentence pairs in the error category "Incorrect translation" increased by 44% compared to those in Score Group 1 (i.e. scores between 0.5 and 0.3), and by 1% compared to those in Score Group 2 (i.e. scores between 0.3 and 0.2). Similarly, for the error category "Misalignment", there was an increase

of 78% compared to the sentence pairs in Score Group 1, and an increase of 1% compared to those in Score Group 2. It can be observed that the sentence pairs with scores between 0.2 and 0 contain far more errors than those with scores between 0.5 and 0.3, and those with scores between 0.2 and 0. There is a less dramatic difference or jump between the sentence pairs with scores between 0.5 and 0.3 and those with scores between 0.2 and 0.

Score Group 4: Scores of 0

In the next score group, 108 sentence pairs were analysed and a total of five error categories were identified.

0 is the lowest score in this dataset. I noticed that there were a lot of cases of "no sentences" in this score group. This error category was only found in this score group. For example:

# ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ;	# ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ;

Here, we can see that there is no sentence in this sentence pair at all. There are only symbols, which cannot form part of the analysed data. This sentence pair was perhaps rejected on the basis of the Bicleaner rule "Unwanted literals: "Re:","{{", "%s", "}}", "+++", "***", '=\"'". Another example of a "no sentence" is as follows:

# ; ; ; R Buhl; R Bals; X Baur; D Berdel; CP Criée; M Gappa; A Gillissen; T Greulich; P Haidl; E Hamelmann; P Kardos; K Kenn; L Klimek; S Korn; M Lommatzsch; H Magnussen; T Nicolai; D Nowak; O Pfaar; KF Rabe; J Riedler; T Ritz; K Schultz; A Schuster; T Spindler; C Taube; K Taube; C Vogelmeier; A von Leupoldt; F Wantke; S Weise; J Wildhaber; H Worth; A Zacharasiewicz	# ; ; ; R Buhl; R Bals; X Baur; D Berdel; CP Criée; M Gappa; A Gillissen; T Greulich; P Haidl; E Hamelmann; P Kardos; K Kenn; L Klimek; S Korn; M Lommatzsch; H Magnussen; T Nicolai; D Nowak; O Pfaar; KF Rabe; J Riedler; T Ritz; K Schultz; A Schuster; T Spindler; C Taube; K Taube; C Vogelmeier; A von Leupoldt; F Wantke; S Weise; J Wildhaber; H Worth; A Zacharasiewicz

The example above displays symbols and names but no full sentences. It looks like an alphabetical list of authors for a paper. Such sentence pairs were left out of the data analysis as they have no meaning and are irrelevant to the dataset. Another such example of a "no sentence" was discussed in the previous chapter.

Of the entire dataset, this score group contains perhaps the most serious errors. Sentence pairs with a score of 0 contain the most mismatches, for instance. Most of the sentences in this group are complete mismatches. Many sentences may also have been rejected on the basis of the rule "Sentence is less than 3 words long", for example:

["My Home, my Car, my Boat" – Satisfaction with One's Own Standard of Living as a Predictor of Gender-Specific Mortality].	HINTERGRUND

In the example above, the target text consists of only one word. According to the Bicleaner hard-rule, this does not constitute a sentence. Another example of this type of error is given in the previous chapter.

Additionally, this score group also contains incomplete sentences and incomplete translations, which are mentioned in the previous chapter. In fact, most of the sentence pairs in this score group are either "no sentences", incomplete sentences, or "unwanted literals". These incomplete translations were probably rejected by the Bicleaner classifier for not being in accordance with the translations it was trained with. There are, however, also examples of incorrect translations, such as the following:

[A rare cause of heart failure].	Wir berichten über einen 52-jährigen Patienten mit einer dilatativen Kardiomyopathie.

This is a clear example of misalignment which could also be considered an incorrect translation.

The sentence pairs in this score group were rightly rejected by the cleaning tool, as the data would require a lot of human intervention to make it useful. The sentence pairs in the other score groups, in contrast, could have passed through the filters and become useful with minimal human intervention. In this group, however, not a single such sentence pair was found. It is very clear why all of this data has been discarded by the tool and why it received the lowest possible score. Another rule that is likely to have been applied to the sentence pairs in this group is "The sentence pair has low fluency score from the language model".

5.2 Discussion of Some Unanalysed Data

There are also some sentence pairs in the unanalysed[21] part of the dataset which could be considered "useful data" following some human intervention. These are not included in the "useful data" section because they were not a part of 515 sentence pairs analysed for this project. These sentences are as follows:

- (arterial hypertension), E10.	- (arterielle Hypertonie), E10.
- (diabetes mellitus), E78.	- (Diabetes mellitus), E78.
- to E14.	- bis E14.

The source and target segments in the first two sentence pairs correspond with each other. One of the Bicleaner hard-rules is "Sentence is less than 3 words long". Since all four of the segments here are 3 words long, Bicleaner would consider them full sentences. The translations of the source fragments are also correct.

The third sentence pair also looks to be a match, but it has likely been rejected because of the same Bicleaner rule, namely "Sentence is less than 3 words long". It might also have been rejected by the Bicleaner classifier. Since the context of the sentence is unclear, this sentence pair probably does not match with any of those that the classifier was trained with. All of these three sentences belong to Score Group 4 (i.e. scores of 0).

[21] The data in this section is unanalysed in the sense that I did not create any error categories for these sentence pairs. I did, nevertheless, dissect their features.

The unanalysed data contains other sentence pairs which could be considered useful following some human intervention. All of these sentence pairs belong to Score Group 1 and scored between 0.5 and 0.3. A detailed analysis was not required to determine that these sentence pairs contain valuable data, as these source and target segments contain the correct terminology for the most part. Some such sentences are outlined below:

Analyses of correlation were done on levels of sum scores, subscales, and single items.	Korrelationsanalysen erfolgten auf Ebene der Summenwerte, Subskalen und Einzelitems.

The above sentence pair was scored 0.495 by the Bicleaner classifier and was probably rejected either by the classifier itself or by the Moses cleaning algorithm specification "everything lowercased". This is, however, a correct translation and requires no changes (Straub et al., 2014).

8226 patients consented to the survey.	8226 Patientinnen gaben ihr Einverständnis, an der Befragung teilzunehmen.

This sentence pair was scored 0.49 and was probably rejected because of the length ratio or because of the Moses cleaning algorithm specification "everything lowercased". The translation is correct except for the term "Patientinnen". In German, the plural ending "-innen" implies only female patients. However, with minimal human intervention, the term can be made gender-neutral (e.g. PatientInnen or Patient*innen)) and hence, this sentence pair can be useful.

A literature review has also been carried out.	Des Weiteren wird ein Überblick über die Literatur gegeben.

In the sentence pair above, the source and target are a match. This pair was scored 0.487 by the classifier and was probably either rejected by the classifier itself for not being comparable

with the translations it was trained with or by the Moses cleaning algorithm specification "everything lowercased". No changes are required in this sentence pair and it can be classified as useful data.

At the beginning of 2017, IQWiG selected 5, at the beginning of 2018 4 topics for HTA reports.	Anfang 2017 wurden vom IQWiG 5, Anfang 2018 weitere 4 Themen für die Erstellung von HTA-Berichten ausgewählt.

In the example above, it is the source text that is grammatically wrong. It has, nonetheless, been translated perfectly into the target language. The fact that the source sentence contains errors in the first place, however, automatically makes the target sentence wrong. However, this sentence pair was rejected based on the Bicleaner hard-rule "There are words in the sentence containing too many uppercased characters between lowercased characters" and the Moses cleaning algorithm specification "everything lowercased". Compared to the previous example, it would take more human effort to fix this sentence pair, as corrections would be required in both the source and target texts.

40% of the ureteral stones were composed of struvite.	40% der Uretersteine bestanden aus Struvit.

This sentence pair is likewise taken from the unanalysed data. At first glance, it looks like the sentence pair is a match because everything apart from one term in the target segment is a correct translation of the source segment, hence some parts of this translation unit are still valuable. Nevertheless, some human intervention would be required to fix the terminology error before the whole sentence pair can be considered useful data. The most commonly used German term for the English "ureteral stones" is "Harnleitersteine" (Das IQWiG, 2006; IQWiG, 2020). The term „Uretersteine" is also used, though less often (Uniklinik, 2016). As a result, this sentence pair cannot be considered useful in its current form, but it can be improved with some human intervention.

5.3 Ambiguity of the Bicleaner Hard-Rules and the Moses Cleaning Algorithm Specifications

The ambiguity of the Bicleaner hard-rules and the Moses cleaning algorithm specifications will be discussed here in detail in relation to how they have been applied to my dataset.

5.3.1 Bicleaner Hard-Rules

"The length ratio between the source sentence and target sentence (in bytes) is too low or too high"

This rule talks about the source-target length ratio, specifically in bytes. Instead of mentioning a specific metric or measurement, however, it simply refers to length ratios that are either "too low" or "too high", which makes the rule vague. Another reason why this rule is unclear is because the ratio taken for this project was r1.5. This means that the target sentence can contain up to 1.5 times as many characters as the source, and not that the target sentence can be up to 1.5 times longer than the source in bytes. As previously mentioned, this rule only talks about the ratio between the source and target sentences being too high or too low. It does not specify whether the source should be longer than the target, or whether the target should be longer than the source. Even though the r1.5 ratio for this project meant that the target could be 1.5 times longer than the source, the rule was occasionally taken into consideration the other way around. In some of the discarded sentence pairs, for instance, we can see that it was actually the source segment that was 1.5 times longer than the target segment (cf. Score Group 4 – Scores of 0).

Another ambiguous hard-rule that is rather difficult to understand is:

"The ratio of numeric characters in source sentence is too high".

This rule is likewise unclear because it only states "too high" and does not specify a measurement or metric. It is therefore difficult to gauge whether a source sentence contains too many numeric characters or not. Take the following, for example:

Overall, 55 % (n = 44) of patients who underwent LE had positive cervical lymph nodes, compared to 10 % (n = 2) of SLE patients.	Insgesamt ergab sich ein pN+ -Stadium in 55 % (LE, n = 44) bzw. 10 % (SLE, n = 2) der Fälle.

In the sentence pair above, there is a considerable number of numeric characters in both the source and target texts. However, it is difficult to ascertain whether or not this sentence pair was rejected based on this filter, since there is no specific measurement available.

The next limitation when it comes to Bicleaner's hard-rules is the filter:

"There are words in the sentence containing too many uppercased characters between lowercased characters".

Similarly, this rule is very unclear and nonspecific in exactly how it is applied. See the following example:

[Breast reconstruction with the free TRAM or DIEP flap – What is the current standard? Consensus Statement of the German Speaking Working Group for Microsurgery of the Peripheral Nerves and Vessels].	Die Brustrekonstruktion mit freiem Gewebetransfer vom Unterbauch als (muskelsparende) TRAM oder DIEP Lappenplastik stellt das Standardverfahren der autologen Brustrekonstruktion dar.

The rule above states that sentence pairs should be discarded if they contain too many capital letters between characters in lower case. It does not specify how many upper case characters are acceptable between lower case characters. In the example above, it can be seen that there are upper case letters between lower case letters in both the source and the target sentences. On both sides, however, these upper case letters represent abbreviations (TRAM and DIEP) and are not just some random upper case characters with no meaning that have been placed between meaningful characters. Furthermore, the source contains square brackets and punctuation that have not been carried over to the target text. Omitting these characters affects the way that the target sentence should be received, which ultimately makes it wrong. Here, nevertheless, we do not know whether this sentence pair was rejected because of the

upper case letters between the lower case letters, or whether it was because another rule was broken. In fact, if we look closely at the source segment, there are a lot of upper case characters between lower case characters. Almost every word in the second source sentence starts with a capital letter (Consensus Statement of the German Speaking Working Group for Microsurgery of the Peripheral Nerves and Vessels). It is also possible that this sentence was interpreted as a title, hence it was removed by the title filter. This sentence, however, has to be written in this way in order for it to be understood as the full form of an abbreviation. The upper case characters between the lower case characters are therefore required to maintain the source meaning.

The next ambiguous rule to be addressed is:

"There are words repeated consecutively".

What follows is an example of a source segment in which words are repeated, though not consecutively:

[The interparental relationship in families with children with ADHD: Interactions between couple distress and child's symptoms]. The interparental relationship in families with children with ADHD: Interactions between couple distress and child's symptoms Abstract.	Zusammenfassung. Der Einfluss der Familie in Bezug auf die Entstehung und Aufrechterhaltung der ADHS bei Kindern ist wissenschaftlich gut fundiert, jedoch hat sich die einschlägige Forschung weitgehend auf das elterliche Erziehungsverhalten oder die Qualität der Eltern-Kind-Beziehung konzentriert.

In the source segment above, it can be seen that the first sentence inside the brackets (highlighted in red) has been repeated a second time, though without brackets (highlighted in blue). Together with this, the sentence appears incomplete as it seems to make little sense. Consecutive or otherwise, there is no repetition in the German translation of this source segment, showing that these two fragments have been misaligned with the source segment here.

In this case, it is once again unclear whether this sentence pair was rejected because of the repetition or because of the Moses cleaning algorithm specifications “everything lowercased” and “sentence alignment (one sentence per line)”. There are also square brackets and punctuation in the source which are missing from the target segment. The example below is extremely similar in its use of repetition:

[Diagnosis and treatment of motor phenomena in schizophrenia spectrum disorders]. Diagnosis and treatment of motor phenomena in schizophrenia spectrum disorders Abstract.	Zusammenfassung. Motorische Auffälligkeiten gehören zum klinischen Bild der Schizophrenie-Spektrumsstörungen.

Just as with the previous example, it is difficult to say which filter(s) caused this sentence pair to be rejected.

The next rule I would like to touch on is:

“Too many characters from unwanted unicode in source sentence”.

First of all, it is important to establish what Unicode is. Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. Unicode provides a unique number for every character, no matter what platform, device, application, or language you are using. Unicode standard is regularly updated to include new characters, and the most recent version of Unicode contains over 137,000 characters. This includes all of the characters in the ASCII character set, as well as many others. As aforementioned, Unicode allows you to use and manipulate text in any language, including right-to-left scripts like Arabic and Hebrew, and even scripts with complex glyphs like Devanagari and Chinese.

UTF-8, UTF-16 and UTF-32 are some examples of Unicode encoding schemes. UTF-8 is the most widely used Unicode encoding method. It is used on the internet because it supports any Unicode character. It is also backward-compatible with ASCII. Some systems use UTF-16 – such as Windows, most notably – and UTF-32 is used in some specialised applications.

Unicode is important because it allows computers to handle and display text in a consistent way, regardless of the platform or language being used. This makes it possible for software and documents to be understood and used by people all over the world (*IBM Documentation*, 2022).

As my dataset did not contain any Unicode characters, this Bicleaner hard-rule was not used to reject any sentence pairs from my dataset. It is worth noting, however, that the rule does not specify how many Unicode characters constitute "too many".

The next ambiguous Bicleaner rule or filter is as follows:

"Too many consecutive spaces in sentence".

There are spaces in every sentence and sentence pair of this data set, but it is not very clear what is meant by "too many consecutive spaces". Take this sentence pair, for instance:

To analyse if tissue flossing is an effective therapeutic modality to improve regeneration after strength endurance exercises and to reduce delayed-onset muscle soreness (DOMS).	ZIEL: Welche Auswirkungen hat Flossing auf die Regenerationsfähigkeit nach Kraftausdauer-Belastungen und den wahrgenommenen Muskelkaterschmerz?

In the example above, the source text is a normal sentence with spaces in all of the expected places (i.e. one after every word). In the target sentence, however, there is more than one space after the first word "Ziel" and before the second word "Welche". There are no other sentence pairs in this dataset which have any such spacing issues. This sentence pair has also been misaligned though, which could be the reason why it was removed by the Bicleaner classifier. Another possible explanation for the rejection of this translation unit could be the application of the Moses cleaning algorithm specification "everything lowercased". In addition to this, there is punctuation in the target segment, which is not present in the source. Brackets also appear in the source, but seem to be missing from the translation. Overall, there are many possible reasons for the rejection of this sentence pair.

The next rule that is ambiguous is:

“Too many parenthesis or brackets in sentence”.

This is an ambiguous rule because it again does not specify how many instances of parentheses are “too many”.

"Gambling disorder" is the only behavioral addiction added to the DSM. Furthermore, preliminary criteria for "Caffeine Use Disorder" and "Internet Gaming Disorder" have now been defined in the manual.	Als einzige «Verhaltenssucht» wurde die «Glücksspielstörung» (Gambling Disorder) in das DSM-5 aufgenommen, zusätzlich wurden vorläufige Kriterien für eine «Koffeingebrauchsstörung» (Caffeine Use Disorder) sowie für eine «Internetspielstörung» (Internet Gaming Disorder) definiert und in das Kapitel III (Störungsbilder, die weiterer Forschung bedürfen) integriert.

In the sentence pair above, there are no brackets in the source text, but there are quite a few of them in the target text. However, there are two reasons as to why we cannot be sure that this is the reason why this sentence pair got rejected. First of all, the rule does not specify how many instances of parentheses are acceptable. Secondly, it does not mention anything about the position of the parentheses. Brackets might be more or less acceptable if they are spread throughout a sentence than if they are lined up consecutively, for instance.

There are two other rules which have not been applied to this dataset at all, namely “There is unescaped unicode characters in sentence” and “Source sentence or target sentence contains mojibake”. Having already talked about Unicode, we will now see what mojibake is. Mojibake is a problem that can occur when text is displayed or stored using the wrong encoding. It is a term used to describe the garbled or incorrect display of characters, which usually happens when text that was originally encoded in one character set is displayed or interpreted using a different character set.

For example, if a document that was originally written in Russian and encoded using the Unicode UTF-8 character set is displayed or stored using a character set that does not support

Russian characters, the text will appear as a series of garbled or unrecognisable characters. This is because the characters in the original document are being interpreted using the wrong mapping to the characters in the new character set.

Mojibake can be caused by a variety of factors, including incorrect character encoding settings on a computer, the use of an outdated or unsupported character set, and the transfer of a document between systems that use different character sets. It can also be caused by the use of incorrect or outdated font files that do not support the characters being displayed.

To fix mojibake, it is necessary to determine the correct character encoding of the original text and use that encoding to display or store the text correctly. This may require converting the text to a different encoding or updating the character encoding settings on the system that is being used to view the text (A Field Guide to Japanese Mojibake, 2021). Just as the Bicleaner rule relating to Unicode does not specify how many Unicode characters are acceptable in a sentence pair, the rule regarding mojibake does not indicate exactly how much mojibake is acceptable.

The next ambiguous Bicleaner hard-rule is as follows:

“All words in source sentence or target sentence are uppercased or in titlecase”.

The ambiguity of this filter can be illustrated using the following example:

[A complicated clinical course of tick-borne rickettsiosis after safari in South Africa]. HISTORY	ANAMNESE UND KLINISCHER BEFUND

In the sentence pair above, which belongs to Score Group 4 (i.e. scores of 0), it can be seen that the words in the target segment are all in upper case. This is not the case in the source segment, demonstrating that the source and target texts here are completely unrelated to each other and have been misaligned. This is probably why this sentence pair was rejected by the Bicleaner classifier. The Moses cleaning algorithm specification “everything lowercased” could also have been responsible for the removal of this sentence pair.

The next rule which is unclear is:

“Sentence is not in the desired language”.

Once more, this rule does not specify the kind of language i.e. natural or artificial. As well as this, the rule neither specifies if the source can contain the target language, nor explicitly states if the target text can contain the source language. Below is an example of a target sentence that was not in the desired language:

"Gambling disorder" is the only behavioral addiction added to the DSM. Furthermore, preliminary criteria for "Caffeine Use Disorder" and "Internet Gaming Disorder" have now been defined in the manual.	Als einzige «Verhaltenssucht» wurde die «Glücksspielstörung» (Gambling Disorder) in das DSM-5 aufgenommen, zusätzlich wurden vorläufige Kriterien für eine «Koffeingebrauchsstörung» (Caffeine Use Disorder) sowie für eine «Internetspielstörung» (Internet Gaming Disorder) definiert und in das Kapitel III (Störungsbilder, die weiterer Forschung bedürfen) integriert.

The language combination used in my project was English-German. As with all of the sentence pairs in this dataset, the target sentence in the example above should therefore be in German. However, it contains English terminology which has been marked in red. Generally speaking, wherever there is terminology in a source text, this should be translated into the target language and the original source terminology should also be mentioned alongside the translated term. Perhaps this is the translation strategy that was adopted here. Regardless, this sentence pair was rejected, but it remains unclear as to whether it was discarded because of the Bicleaner classifier, the English words in the German translation, the source-target length ratio or content addition.

Another example relates to this rule is as follows:

Nevertheless, the diagnostic protocol for CD was not changed in the most parts in the new edition of the DSM-5; the addition of a CD specifier with limited emotions is the most relevant change.	Jedoch sind die wesentlichen diagnostischen Kriterien im DSM-5 gleichgeblieben; eine entscheidende Änderung ist die nun mögliche Klassifikation eines CD specifiers with limited prosocial emotions.

In the example above, we can see that even though the target text is indeed in the target language, the sentence randomly ends with an English phrase. This makes the target sentence wrong. Here, it can also be said that the target text is not in the desired language. Additionally, this is a very clear case of misalignment, which could also be the reason why this sentence pair got rejected by the Bicleaner classifier. Nevertheless, we cannot be certain as to exactly why this sentence pair was rejected.

The next equivocal Bicleaner hard-rule is:

“The sentence pair has low fluency score from the language model”.

This rule is very cryptic because it does not specify how low a fluency score has to be in order for Bicleaner to consider it a low score. The Bicleaner threshold for this project is 0.5, so we could think of anything below this as a low fluency score. In addition, the “scores from the language model” are unknown. We only know the Bicleaner scores, and the entire dataset has been divided according to these scores. If all of the scores below the threshold score are to be considered a “low fluency score” and rejected based only on this rule, then it is difficult to determine where or if any of the other 19 rules have been applied. As such, this filter remains vague.

The last ambiguous rule being dealt with in this subsection is:

“Sentence is more than 1024 characters long”.

It is not clear whether this is referring to the source segment being over 1,024 characters long or the target segment. There are two sentence pairs in this dataset where only the source side consists of more than 1,024 characters. Below is one of these sentence pairs:

[Amblyopia and refractive error]. Myopia is on the increase worldwide and will become a major challenge over the next decades in terms of secondary ophthalmologic complications. There are effective therapeutic options available to slow or prevent the progression of myopia. So far, it has not been investigated whether there are possible additive effects of these interventions. Further investigations - especially in Caucasian populations - are necessary to verify the study results available from Asia. There is limited data on how long further progression of myopia is preventable. A therapy appears reasonable as long as a progression of myopia is detectable. Consistent childhood amblyopia screening provides a cost-effective measure for the prevention of visual disturbances over the course of life. How this can be best integrated into the existing system of "U-investigations", must be clarified by the cost-bearers and professional associations. This discourse should be supported by close interdisciplinary exchange and further studies on the prevalence of different degrees of amblyopia. In addition, sensitive and specific or even multi-stage tests should be developed in order to implement an early	Der folgende Beitrag stellt zwei aktuelle Themen der Kinderophthalmologie in den Fokus, die vermutlich auch immer wieder in der kinder- und jugendärztlichen Praxis relevant sein werden: die Amblyopie und ihre Früherkennung inklusive Videorefraktometrie und Refraktionsfehler im Schulkindalter mit Fokus auf Myopie, deren Häufigkeit weltweit zunimmt.

detection that is cost-effective and saves resources.	

This sentence pair belongs to Score Group 4, meaning that it received the lowest possible score of 0. We can see that this source segment contains parentheses, but the main thing to note here is that it also has 1,275 characters, while the target segment only has 349 characters. This means that the source segment clearly exceeds the character length specified in the hard-rule. As mentioned above, the rule does not specify which of the two segments (i.e. source or target) the rule applies to. Furthermore, if the r1.5 ratio was applied here – which means that the target fragment can be up to 1.5 times the length of the source – it would have had to be applied in reverse here, since it is the source which is longer than the target segment. The other sentence pair with a source segment that breaks this rule is given below:

[Anastomotic Leakage in the Gastrointestinal Tract: Surgical Versus Nonoperative Management]. Most procedures in gastrointestinal (GI) surgery require reconstruction with an anastomosis. Depending on the location within the GI tract, the perfusion and comorbidities of the patients there is a risk for anastomotic leakage. In case of peritonitis with sepsis usually a surgical treatment is required. A stable patient can be treated nonoperatively. In the following overview different treatment options of anastomotic leakage after surgery of the GI tract are described. In case of a leakage of an esophagojejunal or esophagogastric anastomosis after resection of the esophagus or stomach endoscopic treatment can be successful	Eine Anastomoseninsuffizienz ist nach Resektionen und Rekonstruktionen im Gastrointestinaltrakt eine häufige Komplikation – ihre Folgen sind eine Verlängerung des stationären Aufenthaltes, eine schlechtere Prognose und eine erhöhte Letalität der betroffenen Patienten 1,2. Der folgende Beitrag beleuchtet konservative und operative Therapieoptionen der Anastomoseninsuffizienz und zeigt Strategien zu ihrer Vermeidung und Früherkennung auf.

using either clip or stent or negative pressure therapy (NPT). After surgery of the rectum the use of endoluminal NPT has shown good results in case of anastomotic leakage. Nonoperative management of anastomotic leakage can be successful in a stable patient and requires intensive cooperation in an interdisciplinary team with experts in surgery, endoscopy, radiology and intensive care.	

In this example, the source segment contains 1,122 characters and the target segment contains 441 characters. This sentence pair likewise received the lowest possible score of 0. Besides these two examples, there are no other translation units in the entire analysed data set which exceed the specified character limit.

All of the sentence pairs analysed above show that there is a lot of ambiguity in the Bicleaner rules or filters. Many of them overlap which makes it difficult to determine the exact reason why a sentence pair was rejected. Creating error categories does make classifying the errors in each of the sentence pairs easier, but these categories do not coincide with the existing filters, for example, misalignment or content addition.

5.3.2 Moses Cleaning Algorithm Specifications

The Moses cleaning algorithm specifications are very clear except for the following rule:

“Drops lines (and their corresponding lines), that are empty, too short, too long or violate the 9-1 sentence ratio limit of GIZA++”

This rule is unclear because there is no specific measurement given for “too short” or “too long”. As a result, it is difficult to identify when this rule was applied to the dataset and thus resulted in the rejection of sentence pairs.

The next chapter details the limitations of my research project and sheds light on my anticipated conclusions.

6. Limitations

In the previous two chapters, I discussed the findings of this research project in detail. This chapter addresses the limitations of my thesis.

6.1 Limitations of Bicleaner Hard-Rules and Moses Cleaning Algorithm Specifications

As we saw in the previous chapter, there are 20 Bicleaner hard-rules which were used to clean this entire dataset. These 20 hard-rules are in line with six of the seven Moses cleaning algorithm specifications (cf. Table 4). When it came to checking which rule(s) applied to which sentence pairs, however, there was often no definitive answer. This is because using the Bicleaner hard-rules has a couple of limitations. There is a lot of ambiguity in these rules, for instance, which makes them unclear. Along with this, it is hard to understand their exact scope of application.

6.2 Size of Dataset

The next limitation I faced relates to the size of the dataset. For my project, I used a parallel corpus made up of PDF documents from the European Medicines Agency (EMEA). Altogether, this corpus contains 1,108,752 sentence pairs. 19,355 of these sentence pairs were discarded by Bicleaner. This discarded data formed the corpus for my research. Of these rejected sentence pairs, I only analysed 515 translation units for the purpose of my project. This is a considerably small dataset. In some of the studies referenced (e.g. Barbu et al., 2016, de Souza et al., 2013, Khayrallah & Koehn, 2018, Lowphansirikul et al., 2021, Negri et al., 2017, Ramírez-Sanchez & Zaragoza-Bernabeu, 2020, Sánchez-Cartagena et al., 2018 and Srivastava et al., 2020) a huge amount of data was analysed – going into the thousands and millions – and all of this data was analysed manually. In Ding et al. (2022), Nahata et al.

(2016) and Wang et al. (2018), on the other hand, it was analysed automatically. The data in these works was also used for different purposes like training, tuning and testing. However, in my project, I only analysed the data that was left over after a larger amount of data was passed through the automatic cleaning tool, Bicleaner. This is why the corpus I analysed was relatively small compared to other works. The reason why I only analysed the small subset of leftover data is because the aim of this project was to see if the automatic cleaning tool discarded any useful data from the corpus. Of the 515 sentence pairs analysed, eight were found to be worth keeping. A brief overview of the rest of the dataset showed that the other sentence pairs did not differ much in quality. As a result, analysing a small sample was sufficient, and increasing the quantity of data analysed is unlikely to have yielded vastly different results. For example, Score Group 4 (i.e. scores of 0) largely consisted of sentence pairs such as the one below:

# PMID 23238803	# PMID 23238803

The following types of sentences were also part of Score Group 4:

# O Jenni; C Benz; P Hunkeler; H Werner	# O Jenni; C Benz; P Hunkeler; H Werner

We can deduce that the segments in this score group are not full sentences, but rather just hashtags, names of authors and maybe citations. Data such as this is not really suitable for analysis. A quick glance through the entire dataset shows that these kinds of sentences become more common as the scores decrease. The number of misalignments also increases as the scores decrease. As the sentence pairs with these kinds of errors cannot really be analysed, the group of data available for analysis automatically becomes smaller, even without any prior sorting. The next few pages include examples of major misalignments from my dataset that were clearly noticeable without any detailed analysis. Some of these examples are highlighted and explained in the following paragraph.

Certain very aggressive types (so-called small, blue, round cell sarcomas such as embryonal rhabdomyosarcomas, Ewing	Diese vorübergehende Zulassung basiert auf einer randomisierten Phase-II-Studie. Eine

tumors, PNET and desmoplastic soft tissue sarcomas (desmoplastic small round cell tumors) are primarily treated with systemic treatment in a multimodality setting.	konfirmatorische Phase III zur Bestätigung steht aus.
Most of the tumors are located in the extremities and the pelvis and in about 90% of cases the surgical treatment can be performed by means of a limb-sparing wide resection. An endoprosthetic or biological reconstruction of the resulting defect, depending on several patient- und tumor-related factors, usually is necessary.	Knochensarkome sind im Gegensatz zu den häufig auftretenden benignen Knochentumoren und tumorähnlichen Knochenläsionen sehr selten und verursachen daher häufig diagnostische und therapeutische Schwierigkeiten.
Potential drug-drug interactions with patients' basic medications must be assessed.	Typischerweise werden 2 spezifische Inhibitoren kombiniert, die Therapiedauer liegt für die Mehrheit der Patienten bei 8–12 Wochen.
Soft tissue sarcomas are a diagnostically and therapeutically complex disease.	Niedriger Malignitätsgrad (Grading): keine Änderungen, i.d.R. alleinige Resektion.
The experience of this case report raises the question, if hemoabdomen should be no longer considered as an absolute contraindication for laparoscopy and should be considered as a relative contraindication instead.	Im Bereich des Nabels wurde ein sogenannter „Single Port Access" durchgeführt, gefolgt von der Insufflation mit CO
The main symptoms are recurrent urinary tract infections, post-void dribbling and leakage of urine or purulent discharge by movement, which is caused by the	Selten liegen angeborene Fälle vor. Wir präsentieren einen Fall eines weiblichen UD mit Steinbildung, das im Rahmen einer Inkontinenzabklärung diagnostiziert wurde.

emptying of the diverticular lumen (paradoxical incontinence). As this may imitate stress urinary incontinence, the final diagnosis is a challenge for urologists.	
Therefore, about 60 years after its implementation in Iowa it can be said to be the worldwide golden standard. It is known that Ponseti treated feet are better with regard to function and pain when compared to surgically treated clubfeet.	Im Kontext der verfügbaren Literatur werden die pathoanatomischen Grundlagen, die zu der Entwicklung der einfachen Korrekturhandgriffe führten, beschrieben.
[Sense and nonsense of toxicological analyses in the daily clinical routine]. Screening tests for drugs of abuse are regularly used in the clinical routine.	Drogenscreeningtests mittels Immunoassay werden im klinischen Alltag immer wieder durchgeführt.
After peripheral revascularisation transient dual antiplatelet therapy is widely used although there is only little evidence. Following peripheral bypass surgery most patients are treated with single antiplatelet therapy, in some cases (prostetic bypass grafts) dual antiplated therapy can be useful and selected patients with complex venous grafts might profit from anticoagulation with vitamin K antagonists.	Die aktuellen deutschen und europäischen Leitlinien empfehlen bei Patienten mit einer peripheren arteriellen Verschlusskrankheit (PAVK) die Monotherapie mit einem Thrombozytenaggregationshemmer (ASS 100 mg oder Clopidogrel 75 mg).In der COMPASS (Cardiovascular OutcoMes for People using Anticoagulation StrategieS) - Studie wurde Patienten mit PAVK 2×2,5 mg Rivaroxaban zusätzlich zu ASS 100 mg gegeben.
Based on the answers to a questionnaire, we assessed the readiness of owners to use targeted or targeted selective treatment, and to develop a practical decision tool for the	Die Ergebnisse zeigen, dass die Anzahl anthelminthischer Therapien pro Jahr gegenüber älteren Studien reduziert werden konnte.

indication of treatment of individual animals.	
In practice, supervisors expect that young assistant doctors will be familiar with the correct procedure.	Dadurch entstehen Fehler, die leicht vermeidbar sind.
On the other hand, Kandinsky's hypothesis also implies that no one is exempt from such infection.	Menschen mit starken religiösen Gefühlen, mit einer Neigung zum Mystizismus und mit der Leidenschaft zum Geheimnisvollen und Ungewöhnlichen sah Kandinskij als besonders empfänglich an.
Septic shock is defined by vasopressor-dependent circulatory failure and lactic acidosis.	Vitamin C besitzt multiple biologische Funktionen, die sich im Rahmen einer Sepsis günstig auswirken könnten.
Striking differences were the practice of respiratory support (nasal CPAP vs. HFNC) and the prescription of supportive treatments.	Ein frappierender Unterschied zeigte sich in der Wahl der Atemunterstützung.
The interpretation of lab values must taken into account the changing prevalence of serious bacterial infections. Newer parameters (e. g. Procalcitonin) are upcoming, of which the importance is still debatable.	Unter den nun gegebenen Umständen können Anpassungen und damit eine Vereinfachung im Management dieser Kinder vorgenommen werden.
The new screw design generates interfragmentary compression with use of a compression sleeve.	In dieser Studie wurden die klinischen und radiologischen Ergebnisse nach Schraubenosteosynthese durch die HCS bei Skaphoidfrakturen des mittleren Drittels analysiert.

The passage of the new education law will restructure the entire nursing education system in Germany, including the broad implementation of two former pedagogical pilot programmes - a generalist nursing programme and an academic nursing programme. So far, little research has been done to determine whether these new programmes, particularly the generalist nursing programme, equip students with suitable competencies.	Fragestellung: Inwieweit korrespondieren die beruflichen Orientierungen bisheriger Absolventinnen und Absolventen generalistischer Ausbildungsformen in Baden-Württemberg mit den Erwartungen möglicher Arbeitgeber – unter Berücksichtigung der Altenpflege?
The results show that the number of treatments per year decreased compared to previous studies.	Ein grosser Anteil der Ziegenhalter (73.9%) ist bereit, ihr gegenwärtiges Entwurmungsregime umzustellen.
The samples were prepared for the traction-adhesive strength test (sequence: ceramic disc, silicate and silane layering, protective lacquer ("PolyMA" layer), bone cement, TiAlV probes for the traction-adhesive strength test) and their traction-adhesive strengths were then measured.	Im Stirnzugversuch wurde die Haftfestigkeit bestimmt. Dabei stand die Frage im Vordergrund, bei welcher Mindesttemperatur die Aktivierung durch Ausheizen erfolgen muss.
Through social media social networks, primarily Facebook, were widely used. The use of the network functions differed among socio-demographic groups.	Mädchen (OR 0,55) und Kinder von Arbeitslosen (OR 0,28) hatten reduzierte Chancen auf Vereinssport.
[Suturing an Acute ACL Lesion with Dynamic Intraligamentary Stabilisation]. BACKGROUND	ZIELSETZUNG Der Ersatz des vorderen Kreuzbands (VKB) mit autologem Sehnentransplantat ist derzeit der

	Goldstandard zur Behandlung der VKB-Ruptur.
Additional examples of their use in medical teaching scenarios illustrate and clarify each specific teaching competency.	Anwendungsbeispiele sollen die jeweiligen Kompetenzen verdeutlichen.
All of the recommendations are judged with regard to their evidence-based strength according to the Oxford Centre for Evidence-Based Medicine Levels of Evidence.	Alle Empfehlungen erfolgten gemäß der Evidenz basierten Methodik der Oxford-Klassifikation.
All of the recommendations are judged with regard to their evidence-based strength according to the Oxford Centre for Evidence-Based Medicine Levels of Evidence.	Alle Empfehlungen erfolgten gemäß der Evidenz basierten Methodik der Oxford-Klassifikation.

All of the sentences pairs above are from Score Group 3 (i.e. scores between 0.2 and 0) and have been taken from the unanalysed subset of data. At first glance, we can see that all of these constitute major misalignments. In the sentence pair marked in red, the target segment is incomplete. In the translation unit marked in blue, it can be seen that the source-target length ratio (i.e. the numbers of characters on each side) is almost equal. Despite looking similar lengthwise, however, these two segments are a complete mismatch in terms of content. The final two sentence pairs marked in yellow include duplicate information. The rest of the examples contain errors such as missing values, and different uses of punctuation and brackets. The Bicleaner hard-rules related to length ratio and parentheses, as well as the Moses cleaning algorithm specifications concerned with length ratio, lower case characters and one sentence per line might have been triggered by these sentence pairs, hence they were discarded.

As aforementioned, it is likely that analysing more sentence pairs would not lead to any remarkably different results or conclusions, hence taking a small sample from the dataset was considered sufficient for this project. Nonetheless, the analysed dataset is considerably smaller than those used in similar works. The main reason for this is that some sentence pairs were not suitable for analysis due to their errors. Eliminating these parts of the dataset naturally resulted in less data for my research.

7. Conclusion

The purpose of this project was to investigate whether or not the open-source automatic data cleaning tools Bicleaner and Moses perform their intended functions effectively. This was done by carrying out a qualitative evaluation of the data that was discarded after running a corpus through the named open-source automatic data cleaning tools. The corpus used for this project came from the EMEA and was for the language pair English-German (EN-DE). The automatic open-source data cleaning tool Bicleaner has 20 hard-rules which were used to filter this data. As touched on in Chapter 3, the 20 Bicleaner hard-rules encompass six of the seven Moses cleaning algorithm specifications. The limitations of this project, including those relating to the hard-rules and cleaning algorithm specifications, were also discussed in detail.

My analysis shows that the open-source automatic data cleaning tools Bicleaner and Moses clean training data effectively. However, it is not always entirely clear which filters were involved in rejecting which sentence pairs.

In order to then determine why certain sentence pairs were rejected, I grouped the translation units into different error categories. This allowed me to see which of the Bicleaner or Moses filters were most likely to have resulted in which sentence pairs being discarded. The ambiguity of some of the Bicleaner rules as well as that of one of the Moses cleaning algorithm specifications were also discussed in detail.

As discussed in Chapters 4 and 5, the analysed data contained eight translation units, which were considered to include data worth retaining even though they triggered the Bicleaner hard-rules and the Moses cleaning algorithm, and hence they were discarded. Nonetheless,

the Bicleaner hard-rules can be improved so that they are easier to understand and apply to datasets. One way of improving them would be to specify numbers in the rules, for instance, instead of using vague amounts such as “too many” upper case characters, “too many” parentheses or “too many” consecutive single characters separated by spaces. This would also make it easier to determine why certain sentence pairs were rejected by the tool. The same applies to the Moses cleaning algorithm specification “drops lines (and their corresponding lines), that are empty, too short, too long or violate the 9-1 sentence ratio limit of GIZA++”. It would also be beneficial to add some filters related to grammar and context.

Even though all of the other sentence pairs (except these eight potentially useful pairs) were rightly rejected, the reasons behind their rejection remain unclear. Knowing exactly which rule(s) were applied where would assist in categorising the sentence pairs based on their errors. As I was not able to do so using the current rules, I had to come up with my own error categories.

In conclusion, it is safe to say that the open-source automatic data cleaning tools Bicleaner and Moses were effective in cleaning the entire dataset. The examples in the previous chapters also illustrate that some of the data can definitely be turned into valuable training data with some human intervention.

Bibliography

A Field Guide to Japanese Mojibake. (2021, October 31). https://www.dampfkraft.com/mojibake-field-guide.html

Ataman, D., Sabet, M. J., Turchi, M., & Negri, M. (2016). *FBK HLT-MT Participation in the 1st Translation Memory Cleaning Shared Task*. http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/fbkhltmt-workingnote.pdf

Bane, F. & Zaretskaya, A. (2021). *Selecting the Best Data Filtering Method for NMT Training*, pp. 89-97. https://aclanthology.org/2021.mtsummit-up.9/

Barbu, E., Parra Escartín, C., Bentivogli, L., Negri, M., Turchi, M., Orasan, C., & Federico, M. (2016). The first Automatic Translation Memory Cleaning Shared Task. *Machine Translation*, **30**(3-4), pp. 145–166. https://doi.org/10.1007/s10590-016-9183-x

Boitet, Ch., Guillaume, P., & Quezel-Ambrunaz, M. (1982). Implementation and Conversational Environment of ARIANE 78.4, An Integrated System for Automated Translation and Human Revision. *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, pp. 19–27. https://aclanthology.org/C82-1004

Buck, C. & Koehn, P. (2016). *UEdin participation in the 1st Translation Memory Cleaning Shared Task*. http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/ChristianBuck-TM_Cleaning_Shared_Task.pdf

CDC, C. for D. C. and P. (2022, December 20). *STD Facts—Human papillomavirus (HPV)*. https://www.cdc.gov/std/hpv/stdfact-hpv.htm

Collins, A. (2019, April 2). *A Short Introduction to the Statistical Machine Translation Model*. KantanAI - Machine Translation - Neural Language Technology - AI - Localization Technology - Customer Support Solutions. https://kantanmtblog.com/2019/04/02/a-short-introduction-to-the-statistical-machine-translation-model/

Das IQWiG. (2006). gesundheitsinformation.de. https://www.gesundheitsinformation.de/ueber-uns/das-iqwig/

de Souza, J. G. C., Buck, C., Turchi, M. & Negri, M. (2013). *FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task*, pp. 352–358. https://aclanthology.org/W13-2243/

Defauw, A., Szoc, S., Bardadym, A., Brabers, J., Everaert, F., Mijic, R., Scholte, K., Vanallemeersch, T., Winckel, K. V. & Van den Bogaert, J. (2019). Misalignment Detection for Web-Scraped Corpora: A Supervised Regression Approach. *Informatics*, **6**(3), pp. 1-21. https://doi.org/10.3390/informatics6030035

Ding, L., Peng, K. & Tao, D. (2022). *Improving Neural Machine Translation by Denoising Training* (arXiv:2201.07365). arXiv. http://arxiv.org/abs/2201.07365

Du, J. & Way, A. (2017). *Neural Pre-Translation for Hybrid Machine Translation*, pp. 27–40. https://aclanthology.org/2017.mtsummit-papers.3.pdf

España-Bonet, C., Labaka, G., Díaz de Ilarraza, A. & Màrquez, L. (2011, September 19). Hybrid Machine Translation Guided by a Rule–Based System. *Proceedings of Machine Translation Summit XIII: Papers*. MTSummit 2011, Xiamen, China. https://aclanthology.org/2011.mtsummit-papers.63

Halliday, M. A. K. & Delavenay, E. (1962). An Introduction to Machine Translation. *The Modern Language Review*, **57**(1), pp. 1-144. https://doi.org/10.2307/3721978

Henisz-Dostert, B., Macdonald, R. R. & Zarechnak, M. (1979). *Machine translation*. Mouton, pp. 1-266. https://doi.org/10.1515/9783110816679

Huang, J-X., Lee, K-S. & Kim, Y-K. (2020). Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation. *Electronics*, **9**(2), pp. 1-17. https://doi.org/10.3390/electronics9020201

Hutchins, J. (2006). *The history of machine translation in a nutshell*. https://aclanthology.org/www.mt-archive.info/10/Hutchins-2014.pdf

Hutchins, W. J. (1995). Machine Translation: A Brief History. In: Koerner, E. F. K. & Asher, R. E. eds. Oxford/New York/Tokyo: Pergamon. *Concise History of the Language Sciences*, pp. 431–445. Elsevier. https://doi.org/10.1016/B978-0-08-042580-1.50066-0

IBM Documentation. (2022, June 22). https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/workload-automation/9.5.0?topic=support-what-is-unicode

IQWiG, G. (2020, July 27). *Wie bekommt man Nierensteine oder Harnleitersteine?* Die Techniker. https://www.tk.de/techniker/gesundheit-und-medizin/behandlungen-und-medizin/gynaekologische-und-urologische-erkrankungen/wie-bekommt-man-nierensteine-oder-harnleitersteine-2017338

IWSLT Evaluation 2015—MT Track. (2015). https://sites.google.com/site/iwsltevaluation2015/mt-track

Jalili Sabet, M., Negri, M., Turchi, M., de Souza, J. G. C. & Federico, M. (2016). TMop: A Tool for Unsupervised Translation Memory Cleaning. *Proceedings of ACL-2016 System Demonstrations*, pp. 49–54. https://aclanthology.org/P16-4009.pdf

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T. & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121. https://doi.org/10.18653/v1/P18-4020

Khayrallah, H. & Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 74–83. https://doi.org/10.18653/v1/W18-2709

Koehn. (2009). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, P. (2016). *Statistical Machine Translation System User Manual and Code Guide*. pp. 1-359. http://www2.statmt.org/moses/manual/manual.pdf

Koehn, P. (2020a). *Neural Machine Translation* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108608480

Koehn, P. (2020b). *Neural Machine Translation* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108608480

Kühn, W. (2011). Kolposkopie zur Früherkennung des Zervixkarzinoms. *Der Pathologe*, *32*(6), pp. 497–504. https://doi.org/10.1007/s00292-011-1480-9

Landsbergen, J. (1989). The Rosetta project. *Proceedings of Machine Translation Summit II*, pp. 82–87. https://aclanthology.org/1989.mtsummit-1.15

Lelner, Z. (2022, March 29). *Machine Translation: The Complete Guide*. https://blog.memoq.com/machine-translation-the-complete-guide

Liu, J. & Liro, J. (1987). The METAL English-to-German system: First progress report. *Computers and Translation*, **2**(4), pp. 205–218. https://doi.org/10.1007/BF01682180

Lowphansirikul, L., Polpanumas, C., Rutherford, A. T. & Nutanong, S. (2021). A large English–Thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation*, **56**(2), pp. 477–499. https://doi.org/10.1007/s10579-021-09536-6

Maas, H. D. (1977). *The Saarbrucken automatic translation system (SUSY)*. **1**, pp. 585–592. https://aclanthology.org/www.mt-archive.info/70/CEC-1977-Maas.pdf

Maegaard, B. (1988). Eurotra: The Machine Translation Project of the European Communities. *Literary and Linguistic Computing*, **3**(2), pp. 61–65. https://doi.org/10.1093/llc/3.2.61

Mandorino, V. (2016). *The Lingua Custodia Participation in the NLP4TM2016 TM Cleaning Shared Task*. http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/description_LinguaCustodia.pdf

Mohamed, S. A., Elsayed, A. A., Hassan, Y. F. & Abdou, M. A. (2021). Neural machine translation: Past, present, and future. *Neural Computing and Applications*, **33**(23), pp. 15919–15931. https://doi.org/10.1007/s00521-021-06268-0

Moorkens, J., Doherty, S., Kenny, D. & O'Brien, S. (2014). A virtuous circle: Laundering translation memory data using statistical machine translation. *Perspectives*, **22**(3), pp. 291–303. https://doi.org/10.1080/0907676X.2013.811275

Nagao, H. & Tsujii, J. (1986). The Transfer Phase of the Mu Machine Translation System. *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, pp. 97–103. https://aclanthology.org/C86-1021

Nahata, N., Nayak, T., Pal, S. & Naskar, S. K. (2016). *Rule Based Classifier for Translation Memory Cleaning*. https://www.researchgate.net/publication/303692687_Rule_Based_Classifier_for_Translation_Memory_Cleaning

Negri, M., Ataman, D., Sabet, M. J., Turchi, M. & Federico, M. (2017). Automatic translation memory cleaning. *Machine Translation*, **31**(3), pp. 93–115. https://doi.org/10.1007/s10590-017-9191-5

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association*

for Computational Linguistics - ACL '02, pp. 311-318. https://doi.org/10.3115/1073083.1073135

Petrova, V. (2019). Translation Quality Assessment Tools and Processes in Relation to CAT Tools. *Proceedings of the Second Workshop Human-Informed Translation and Interpreting Technology Associated with RANLP 2019*, pp. 89–97. https://doi.org/10.26615/issn.2683-0078.2019_011

Poibeau, T. (2017). *Machine Translation*. MIT Press. pp. 1-296. https://doi.org/10.7551/mitpress/11043.001.0001

Ramírez-Sanchez, G.,& Zaragoza-Bernabeu, J. (2020). *Bifixer and Bicleaner: Two open-source tools to clean your parallel data*. pp. 291–298. https://aclanthology.org/2020.eamt-1.31/

Rao, D. D. (1998). *Machine translation*. pp. 61-70. https://doi.org/10.1007/bf02837314

Sreelekha, S. (2017). *Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective*. https://doi.org/10.48550/arxiv.1708.04559

Sánchez-Cartagena, V. M., Bañón, M., Ortiz-Rojas, S.,& Ramírez, G. (2018). Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 955–962. https://doi.org/10.18653/v1/W18-6488

Scott, B. (Bud). (2003). The Logos Model: An Historical Perspective. *Machine Translation*, **18**(1), pp. 1–72. https://doi.org/10.1023/B:COAT.0000021745.20402.59

Shen, X. (2023). A Review of Machine Translation: Implications to Human Translators and Translation Teaching. *The Educational Review, USA*, **6**(12), pp. 869–874. https://doi.org/10.26855/er.2022.12.014

Srivastava, J., Sanyal, S. & Srivastava, A. K. (2020). An Automatic and a Machine-assisted Method to Clean Bilingual Corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **19**(1), pp. 1–19. https://doi.org/10.1145/3342351

Straub J, Plener PL, Koelch M, Keller F. (2014). Agreement between self-report and clinician's assessment in depressed adolescents, using the example of BDI-II and CDRS-R. *Zeitschrift fur Kinder- und Jugendpsychiatrie und Psychotherapie*, **42**(4), pp.243-252. https://doi.org/10.1024/1422-4917/a000297

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M. & Liu, Y. (2020). *Neural Machine Translation: A Review of Methods, Resources, and Tools* (arXiv:2012.15515). arXiv. http://arxiv.org/abs/2012.15515

Torregrosa, D., Pasricha, N., Chakravarthi, B. R., Alonso, J., Casas, N., Masoud, M. & Arcan, M. (2019). *Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models*. **2**, pp. 125–133. https://aclanthology.org/W19-6725/

Uniklinik, K. (2016, January 26). *Nieren- und Harnleitersteine*. https://urologie.uk-koeln.de/erkrankungen-therapien/nieren-und-harnleitersteine/

Van Doorslaer, K., Chen, Z., Bernard, H-U., Chan, P. K. S., DeSalle, R., Dillner, J., Forslund, O., Haga, T., McBride, A. A., Villa, L. L., Burk, R. D.,& ICTV Report Consortium. (2018). ICTV Virus Taxonomy Profile: Papillomaviridae. *Journal of General Virology*, **99**(8), pp. 989–990. https://doi.org/10.1099/jgv.0.001105

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wang, W., Watanabe, T., Hughes, M., Nakagawa, T. & Chelba, C. (2018). Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection. *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 133–143. https://doi.org/10.18653/v1/W18-6314

White, J. S. (1985). *Characteristics of the METAL machine translation system at production stage*. pp. 359–369. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fc46bc44f0a555559770459f0ae14bc8ef30fb3c

Witkam, A. P. M. (1984, February 13). Distributed language translation, another MT system. *Proceedings of the International Conference on Methodology and Techniques of Machine Translation: Processing from Words to Language*. BCS 1984, Cranfield University, UK. https://aclanthology.org/1984.bcs-1.34

Wolff, F. (2016). *Unisa system submission at NLP4TM 2016*. http://rgcl.wlv.ac.uk/wp-content/uploads/2016/05/UNISA_working_notes.pdf

Zaragoza-Bernabeu, J., Ramírez-Sánchez, G., Bañón, M. & Rojas, S. O. (2022). *Bicleaner AI: Bicleaner Goes Neural*. pp. 824–831. https://aclanthology.org/2022.lrec-1.87/

Zbib. (2010). Using Linguistic Knowledge in Statistical Machine Translation. https://apps.dtic.mil/sti/pdfs/ADA544288.pdf

Zwahlen, A., Carnal, O.,& Läubli, S. (2016). *Automatic TM Cleaning through MT and POS Tagging: Autodesk's Submission to the NLP4TM 2016 Shared Task* (arXiv:1605.05906). arXiv. http://arxiv.org/abs/1605.05906