Universität des Saarlandes

Fakultät 4 - Philosophische Fakultät II

Fachrichtung 4.7 Allgemeine Linguistik

Computerlinguistik

Bachelor's Thesis

in fulfillment of the requirements for the degree of

„Bachelor of Science“ (B. Sc.)

Exploring Domain Adaptation in Machine Translation for English to Japanese

Submitted by: Torsten Kai Jachmann
Born in: 10.04.1987 in Dudweiler
Matriculation Number: 2523139

Examiners:	Prof. Dr. Hans Uszkoreit
	Dr. Robert Grabowski (thesis advisor)
Thesis advisor:	Yu Chen

SULZBACH, 19.08.2013

Declaration Of Academic Honesty

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Sulzbach, 19.08.2013

Acknowledgment

At this point I want to express my gratitude towards all those, who helped and supported me with and while writing this bachelor's thesis.

First of all I want to express my gratitude towards Prof. Dr. Hans Uszkoreit, who provided me with the possibility to write this thesis. Further I owe the company Acrolinx a debt of gratitude for introducing me to the idea of pre-editing and for supporting me in so many terms. Especially, I want to thank Dr. Robert Grabowski, who always found time to help me and without whom I would not have been able to realize my ideas for this thesis the way I did. I also want to thank Yu Chen for being my advisor, supporting me with many ideas, discussing ideas and realizations and for always helping out with every problem that occurred.

Special thanks also go to all the members of the ACCEPT project, who allowed me to continue and broaden their research.

Also, I would like to thank Shota Arai, Yukako Chiba, Michiko Furukawa, Junki Hitora, Maki Kajimoto, Mayumi Kawada, Nanako Matsumoto, Mika Nakamichi, Kasumi Sakamoto, Akira Takemura and Yoko Toyotani for voluntarily evaluating my work despite their own obligations.

多忙にもかかわらず、本研究に協力して頂いた荒井翔太、梶本万貴、河田真弓、阪本歌澄、竹村映亮、千葉由香子、豊谷陽子、中道海伽、人羅純基、古川実千子、松元那々子の各氏に、この場を借りて感謝の意を表したい。

I also want to express my gratitude towards Hideki Yamaguchi, who deepened my love towards the Japanese language and of whom I think as much more than simply a teacher. Thank you for supporting me over all those years and for being such a great teacher.

Last but not least, I want to thank all my friends all over the world and my family for supporting me morally and emotionally through even the hardest times. I wouldn't be who I am without you. I especially want to thank Nathan Manhanke and Parker Murff for proofreading this thesis.

Speziell möchte ich meinen Eltern, Ilse und Rolf Jachmann danken, die mich bei jeder Entscheidung, die ich treffe, vollkommen unterstützen und mich immer durch gute wie auch schlechte Zeiten begleiten. Ich weiß, dass ich mich immer auf euch verlassen kann.

ABSTRACT

In this thesis, we increase the quality of statistical machine translation (SMT) of forum text for the language pair English – Japanese by exploring three approaches of domain adaptation. The main difficulty the domain of user-generated content poses to SMT is the lack of bilingual in-domain corpora. Therefore, the first approach in this thesis is the use of close-domain and out-of-domain bilingual corpora in combination with monolingual Japanese corpora to train a Moses SMT system. This approach results in a BLEU score of 22.10, which describes a gain of 1.68 points of BLEU in comparison to a SMT system that is only trained on a large close-domain bilingual corpus (20.42 BLEU). The second approach is the synthesis of a bilingual in-domain corpus by translating in-domain forum text that is only available in English with our baseline system. This synthesized corpus is then used to retrain the SMT baseline system by adding it to the translation model training. This system receives a BLEU score of only 20.52, which describes a degradation of 1.58 points compared to the baseline system. The translations contain major grammatical and structural degradation compared to the translations of the baseline system. In this thesis we further introduce pre-editing the text that is to be translated before translating it with the baseline system as the third approach of domain adaptation. This is done by automatically applying rules to the text that are triggered by certain patterns in the text. Those rules then change the text to receive a closer resemblance to the data used in the training of the SMT system. The rules are applied by using the Acrolinx tool in combination to the autoApplyClient. This approach improves the BLEU score by 0.59 points compared to the baseline syste, producing a score of 22.69 BLEU. Furthermore, human evaluators rated most of the translation received by this approach as better compared to the translations of the baseline system without the rules applied. As the approach using the synthesized corpus shows strong differences in the impact on the translations with even small changes to the SMT system it is used in, we combine all three methods in a last step. The monolingual source side text is edited with the pre-editing rules of the third approach before the translation in order to receive a synthesized corpus. This corpus is then used to retrain the baseline system by adding it to the translation model training. The pre-editing rules are further used on the text that is to be translated before translation. This method receives a BLEU score of 22.20, which is a degradation of 0.49 points compared to the approach only using the pre-editing rules. Yet, human evaluators show a strong tendency towards the approach combining both methods. 51% percent of the tested sentences were rated better using this approach and only 33% worse compared to the approach using only the pre-editing rules. This shows that human evaluation is a necessary step in evaluating machine translation instead of merely relying on automatic metrics such as the BLEU score.

List of Abbreviations

CNL	Controlled Natural Language
EMS	Experiment Management System
MERT	Minimal Error Rate Training
MT	Machine Translation
OOV	Out-Of-Vocabulary
POS	Part-Of-Speech
SMT	Statistical Machine Translation
SOV	Subject-Object-Verb
SVO	Subject-Verb-Object
UGC	User-Generated Content

Table of contents

1. Introduction

Machine Translation (MT) is a major task of Computational Linguistics. As such, it was studied and improved over the years by various researchers. A branch of MT is Statistical Machine Translation (SMT), which generates translation based on statistical models, trained on bilingual corpora. This approach has shown to deliver good results on a general base, but encounters problems when aiming for domains, for which bilingual corpora are rare and sometimes even non-existent. To alleviate this problem, a possible solution can be found by using domain adaptation. One approach in domain adaptation is to use small in-domain corpora to train SMT systems and to add larger out-of-domain corpora to improve the performance. Such approaches are mostly being applied to close language pairs such as English–German and English–French in domains for which no sufficient data is available. Concerning distant language pairs, such as English and Japanese, research is still rather rarely done. This holds especially for domains, for which bilingual corpora are not only rare, but also almost not existent. Such domains are for instance represented by User-Generated Content (UGC), such as forum conversations. Those contain many stylistic and constructional differences from standard written language. In comparison, official texts are usually written in standard language. As most translations, which are used as bilingual corpora to train SMT systems, are based on such texts, those differences are unknown to the resulting SMT system.

This thesis will show the advantages of domain adaptation for distant language pairs and will introduce pre-editing as a part of domain adaptation. Pre-editing the source texts by automatically changing certain sentence structures and phrases in order to make them closer in style to the data used in training the SMT system will further improve the translations. Those changes will be achieved by applying automated pre-editing rules on the source text using the Acrolinx tool and the autoApplyTool. Thereby, this work is closely related to the still ongoing ACCEPT project, which mainly focuses mainly on English-to-German and English-to-French translation improvement with a focus on Symantec forum data. Our work will not only explore the general effect of pre-editing rules on translation value, but also demonstrate the need of creating language-specific rule sets. This will be done by comparing results for existing rules, which did perform well on English-to-German translation, with newly created rules, which tackle Japanese-specific problems. In addition to that, language-irrelevant rules are transpired.

In order to evaluate the impact, we conducted human evaluation by native speakers of Japanese to rate the rule-by-rule impact and the overall impact in different steps. In addition to that, we used automatic evaluation metrics (BLEU).

1.1. OVERVIEW

Chapter 2 provides fundamental background information, needed within this thesis. This chapter includes information about (a) machine translation, (b) the Acrolinx tool that was used and (c) the ACCEPT project and (d) the background knowledge on Japanese that is necessary to follow the intended changes and the results of the pre-editing rules.

In chapter 3, hereto-existing papers on domain adaptation and pre-editing are discussed.

Following that, chapter 4 explains the procedure of creating a baseline system with Moses. This chapter further provides insight into the different SMT systems, which were trained for this thesis. That includes explanations for the improvements or degradation from one system to another.

In our work, a synthesized bilingual corpus was created to be used to retrain the baseline system. Chapter 5 gives information on how that corpus was created and used. Furthermore, the results of the retraining in terms of change in BLEU score are described.

The development, evaluation and use of pre-editing rules are handled in chapter 6. The conducted evaluation is explained, alongside the rating system that was used to evaluate the different rules. For each rule, this chapter provides information on the idea behind the rule, the ratings that were assigned by human annotators and an explanation for those results. Finally, information on how the rules are applied on the respective data is provided.

Chapter 7 gives insight into the results that a combination of the pre-editing rules with the synthesis of a corpus used in re-training provides.

In chapter 8 the different results are summarized and discussed. This chapter will additionally provide possible future work, which might further improve the results, provided by this thesis.

Appended to this work, relevant readings can be found under “References” and data, such as tables summarizing the results of the evaluation steps and examples of the data given to the annotators under “Appendix”.

2. Background Information

This chapter provides basic information and background knowledge needed to understand this thesis.

2.1. Statistical Machine Translation

The goal of machine translation (MT) is to translate text from one language to another without the interference of a human and thereby allowing users of MT to understand a text without understanding the language it is originally written in.

The first research in the field of MT was conducted by Yehoshua Bar-Hillel at MIT in 1951. Many other researchers followed until in 1964 the Automatic Language Processing Advisory Committee (ALPAC) was formed to study MT. However, after the ALPAC report that was issued in 1966, funding was severely reduced due to a lack of results and a lackluster progress. Many companies still used MT though and in the late 1980s; with the increase of computational power and a drop of its costs, statistical models for machine translation gained more and more attention. Various MT companies were founded and in 1996 SYSTRAN offered the first free translations of short text in the Internet. In 2007, under the guidance of Philipp Koehn, the Moses MT toolkit was created as an open source project.

There are various approaches in MT, which can roughly be categorized in four larger groups. Those groups are Rule-based MT, Example-based MT, Statistical MT and Hybrid MT that combines the strengths of Rule-based and Statistical MT.

In our work, we use the MT toolkit Moses to train a statistical machine translation (SMT) system. Therefore we focus on an explanation on SMT.

SMT uses statistical probabilities that are derived from analyzing bilingual corpora to translate input. Basically, a translation is done by computing the probability of a string in the source language to be translated with a certain string in the target language. These probabilities are learned from their distribution in the bilingual corpora. This results into a translation model. Additionally, a language model trained on monolingual text in target language is used to compute the probability of a translated string to appear in the target language. The translation is then chosen by applying the Bayes Theorem, which combines the translation model and the language model by choosing the one translation that maximizes the probability. The equation below displays the Bayes rule that is applied.

$$\begin{aligned} argmax_e p(e|f) &= argmax_e \frac{p(f|e) * p(e)}{p(f)} \\ &= argmax_e\, p(f|e) * p(e) \end{aligned}$$

In the equation, “e” stands for the translation and “f” for the original phrase. p(e|f) describes the probability for a certain translation “e” based on the source language phrase “f”. p(f|e) is the probability found in the translation model for that particular phrase and p(e) stands for the probability of the translated phrase to appear in the target language, hence the probability found in the language model.

Originally, SMT was conducted on a word-based level. This means that single words were considered as the fundamental unit of translation. With time, further approaches in SMT were added. This includes syntax-based translation, phrase-based translation and hierarchical phrase-based translation. In our work, we make use of the phrase-based approach.

Phrase-based SMT does not consider single words, but rather phrases that may differ in length as the fundamental unit of translations. Those phrases do not necessarily have to be linguistic phrases, but usually are phrasemes found by statistical methods applied to bilingual corpora.

2.2. Acrolinx

The company Acrolinx is the leading provider of content editing technologies. Their main product, the Acrolinx tool, can be used to provide guidance while creating content.

The tool makes use of a spell checker as well as a grammar checker. Additionally, style and terminology are also checked to achieve an easy-to-understand and conventionalized text that follows customer-defined writing guidelines. This improves the value of thereby created documents.

First, the text that is to be corrected is sent to an Acrolinx server where it is tokenized. Those tokens are then further analyzed and annotated by various components, such as a part-of-speech (POS) tagger, to receive further linguistic information. This tokenized and annotated text is then handed to the implemented rules, which generate flags that mark the position of a potential error. Those flags are categorized by the rule that was applied to extract a specific error.

The rule formalism used in the Acrolinx tool was developed in a research project at the DFKI in 2000, which was conducted by Andrew Bredenkamp, Berthold Crysmann and Mirela Petrea. Instead of validating the text against a positively defined grammar, the rules are phenomena-based. That means that, firstly, triggers for a rule are searched for in the text, which are then verified or dropped. Such triggers are, for example, certain word or character patterns in the text, constellations of POS tags and morphologic annotations. Depending on the target of a rule, not all steps are necessary in this approach. For example, to match the future tense in English, which is often expressed by using the word “will”, it is sufficient to search for this particular word. An additional usage of POS analysis or morphology is not needed for that task. Once such a possible error is found, the surrounding of the flag is further analyzed to eliminate false flags.

The triggers for the rules are formulated in a newly created language that resembles the regular expressions of the programming language Perl in terms of syntax. Yet, the symbols of the expressions in this language can also match tagging and other linguistic information. They can access various taggers that can be used independently from one another to match the needs of each specific rule. Since its beginning in 2000, this formalism was further improved over the years. Among others, in addition to flagging possible errors, suggestions for their corrections were provided.

This addition makes it possible to automatically apply the rules to correct large amounts of texts without manual interaction. The Acrolinx autoApplyClient can be used for this purpose. The autoApplyClient always corrects the found errors with the first suggestion that is proposed by the rules. This tool allows us to use the Acrolinx rules as an automated pre-editing step to improve the value of the automatically received translations done by an MT system.

2.3. ACCEPT

The ACCEPT project deals with MT for the field of web content. This target was chosen as content provided on the Internet, especially forum conversations, gain more and more value with the emergence of Web 2.0. This term describes the use of the Internet where users do not only consume content but also create content themselves. With this change, the exchange of experiences from user to user turned into a prosperous way to tackle various problems, for which official support is not provided or takes too much time. The project therefore concentrates on improving MT in this environment by developing new technologies in order to make such user generated content (UGC) accessible for users of different languages.

The project consists of three main parts:

- Pre-editing
- Post-editing
- Using the insight that is gained in the editing steps in combination with text analytics to improve the SMT engines themselves

2.4. Differences between English and Japanese

A distant language pair describes two languages that share almost no similarities, neither in vocabulary, nor in grammar. Japanese is an isolated language, which means it shares no demonstrable genealogical relationship with other languages. English on the other hand is an Indo-European language. Therefore, this language pair can be considered very distant. Thus, there are many differences between them that pose difficulties for MT.

One major difference between the two languages is the writing, in which different characters are used. Whereas English uses the Latin script, Japanese uses a combination of simplified Chinese characters[1] and two writing systems that exist only in Japanese (Hiragana and Katakana). Therefore, words that are not translated by an MT system pose a bigger problem than they do in languages that use the same writing system and share similar words.

Another major difference is found in the grammar of each language. English is a subject-verb-object (SVO) language, whereas Japanese is a subject-object-verb (SOV) language. This very fundamental difference poses a severe problem as a lot of moving is necessary for correct translations. Furthermore, Japanese uses postpositions contrary to the prepositions in English. A further difference in grammar is the expression of the grammatical cases: In English this is expressed by inflecting the correspondent words, whereas in Japanese the grammatical case is expressed only by particles.

Further differences between the two languages will be explained in chapter 6 while explaining the effect of certain pre-editing rules.

[1] The simplified Chinese characters in Japanese are not the same as the simplified Chinese characters used in Chinese.

3. Related Works

Domain adaptation has been proven to be a successful approach in SMT for domains, in which bilingual corpora are rarely existent or even non-existent. A domain can be understood as text data that breaches the issue of a certain topic. In that case all corpora containing this topic are considered in-domain corpora. Another aspect of domain is style. Therefore, corpora addressing a certain topic are no longer considered in-domain if their style differs vastly from the style of the targeted text data. For example, corpora handling the topic of product support and guidelines, such as manuals, are considered close-domain corpora if the goal is to translate forum conversations on the same topic. A third aspect of domain is the language of the corpora. Even if corpora are handling the same topic and are written in the same style, they cannot be considered in-domain if the language pair they are written in do not match the language pair targeted. All corpora handling topics that are unrelated to the selected topic are considered out-of-domain corpora.

Usually MT systems are trained on corpora consisting of in-domain data. Typically MT systems, trained on a specific domain, do not perform well in other domains. For example, an MT system trained on software manuals may not be able to properly translate a cookbook. Therefore, in order to deal with specific domains for which parallel corpora do almost not exist for certain language pairs, out-of-domain corpora need to be included in the training of an MT system.

Hence, many researchers have proposed solutions for domain adaptation. A commonly used approach for domain adaptation is to use a small amount of accessible in-domain corpora and combine the hereby-trained model with models trained on out-of-domain data. Foster and Kuhn (2007) compared various approaches for mixture-model adaptation in phrase-based Statistical MT. More specifically, they decomposed one large corpus into the different topics addressed in that corpus. Out of those smaller corpora they recieved, they trained a specific MT system for a certain topic by adapting the other parts as out-of-domain data. The difference in this thesis is that we deal with user-generated content (UGC) – forum text – for which no bilingual corpora for English and Japanese exist. In other words, we do not have access to completely in-domain parallel data to apply this method.

Wu et al. (2008) presents a more general approach to improve MT compared to Foster and Kuhn; by adding smaller in-domain dictionaries to systems trained with out-of-domain data. In addition an in-domain monolingual corpus is used to improve the language model. A further addition in their work is the use of the MT system, which is created as described above to automatically translate a monolingual in-domain corpus in source language. This hereby synthetically produced bilingual corpus is then added to retrain the MT system and thereby further improve the performance.

Instead of adding an in-domain dictionary, we are making use of close-domain corpora, namely product manuals and their translations provided by Symantec.

This data is similar to the UGC that we aim to translate. A further addition is the Tatoeba corpus, which contains more colloquial speech, and is therefore closer to the style used in the forum entries. Therefore, this corpus is considered as a part of domain adaptation, regarding the definition of domains we use throughout this thesis.

Banerjee et al. (2011) explored the benefits of domain adaptation in SMT of user-forum data. They describe the problems that occur with only monolingual corpora for a specific domain on a basis of Symantec user-forum data. They explore similar domain adaptation techniques as described by Foster and Kuhn and Wu et al. in order to improve SMT systems aiming for UGC for English-to-German and English-to-French translations. To train their out-of-domain system, they make use of the large Europarl corpus, which exists for many European languages. Although Europarl is clearly out of domain, given the aim for UGC about Symantec products, the much larger size of the corpus is deemed to improve performance. To balance the proportions of in-domain and out-of-domain data, only relevant data from the corpus was used. In order to extract that data, sentence-level perplexity scores in respect to language models trained on monolingual in-domain data were applied on the corpus in order to rank the sentences. A lower perplexity score indicates a closer fit to the forum data. As mentioned above, Europarl exists only for European languages. Therefore, in our case, the target language, Japanese, requires a different bilingual out-of-domain corpus for training. In our work, the kftt corpus is used to achieve this goal. However, the size of the kftt corpus is much smaller than the size of Europarl. Hence, the entire corpus was included, without performing additional selection.

Banerjee et al. (2012) tackled a similar approach as in 2011. However, they shifted the focus to the problem of out-of-vocabulary (OOV) items. OOVs are words or phrases that do not exist in the vocabulary of MT systems and therefore cannot be translated. Such phrases may occur out of various reasons. The obvious reason is an insufficiency of training data, as smaller training data increases the frequency for certain words not to appear in the training data. Special occurrences in the text are another reason for such phrases. Such occurrences are, for instance, maskable tokens such as URLs and email addresses. Other special occurrences are fused words, which consist of two or more valid tokens that are concatenated by punctuation marks. Further there are spelling errors and non-translatable words, like product names and service names. Our focus lies on spelling errors and partially on fused words as a part of grammar correction. This focus is chosen as the other types of OOVs, such as non-translatables and maskable tokens, seem to have only minor adulterating effects. Yet, we do not handle the approach concerning OOVs separately, but as part of the pre-editing procedure applied on the text to be translated.

Pre-editing has a close relation to domain adaptation when it comes to UGC. Both deal with the problem of a lack of bilingual corpora of this very specific domain with far spread topics. Whereas domain adaptation offers the ability of creating a MT system, even for domains that are suffering from a paucity of bilingual data, pre-editing provides the possibility to make the appearance of UGC closer to the appearance of the data used in the training step. Therefore the number of OOVs, especially those that arise because of the special writing styles in UGC, can be reduced.

Roturier at al. (2012) conducted research on the topic of pre-editing in line with the ACCEPT project. They describe the impact of pre-editing rules, which handle the problems that occur while dealing with UGC in MT. Some of those problems are due to the fact that the source content may have been produced by non-professionals as well as by non-native speakers, and thus may contain linguistic and technical errors. Furthermore, UGC tends to be closer to spoken language than to written language in its style. Another characteristic of UGC is that power users follow writing styles that are "guided by attitudes of technological elitism (Leblanc, 2005)". That includes, for example, acronyms, altered spelling, and emoticons. The pre-editing rules created by Roturnier et al. are applied by using the Acrolinx tool in combination with their autoApplyClient[2]. They study the influence of those pre-editing rules on English-to-French and English-to-German translations, whereas we focus on exploring those influences on a distant language pair, namely English-to-Japanese. Furthermore, in this thesis we do not only approach the above-mentioned common problems with UGC, but also try to develop language specific pre-editing rules and compare those to rules that previously worked well on English-to-German translation, thus showing the importance of developing both general and language specific rules.

Lehmann et al. (2012) kept close to the work of Roturier et al., but add some further perspectives and exhibit problems concerning automated pre-editing rule application. For example, some sentences may contain major flaws in the language they are written in or may hold known problems for MT that are not able to be corrected or changed automatically, such as moving longer phrases, etc. To tackle this problem, they suggest guiding the user even while creating their content, resulting in a controlled natural language (CNL). Even though there are certainly benefits in this approach, we do not consider this approach for our project for two reasons. One is that we aim at already existing content, meaning that user guidance cannot be applied at that point. The other is that even though this kind of guidance performs well on English-to-German translations – or other relatively close languages – it may not provide the same effects on distant language pairs, such as English–Japanese. The reason for this assumption is that the structural differences between English and Japanese are more significant than those between English and German. The changes that might need to be done are likely to result in English sentences that are poorly formed. Such suggestions cannot be made to a user while creating content. Therefore this thesis focuses only on pre-editing rules, which can be applied automatically.

[2] Both tools are also used in our work.

4. Statistical Baseline System

For this thesis a Moses baseline system is trained using different bilingual corpora in English and Japanese to train the translation model. In the language model training we use the Japanese side of the same corpora as in the language model training and additionally monolingual corpora in the target language Japanese.

Moses was chosen for this task because of its easy accessibility and use. Furthermore, because it was also used in the ACCEPT project, therefore it allows us to put both works into closer relation and to compare the results. Yet, the corpora used to train the systems in this thesis differ from the ones created in the ACCEPT project as some corpora used in the ACCEPT project are not existent in Japanese. Furthermore, some corpora used in the system trained for this work may provide a certain advantage over the ones used in the ACCEPT project.

4.1. Setup

In order to set up the baseline system, Moses' Experiment Management System (EMS) was used. This EMS combines various steps needed to create a Moses SMT system. In order for it to work, a working directory and a manually written configuration file containing the parameters for the training, have to be created.

The configuration file contains various information such as the paths of the working directory and the tools used for training. In this work, the tools that are delivered with Moses are used in terms of decoder, ttable-binarizer and truecaser. In addition kytea is used to tokenize the Japanese corpora before using them in the EMS and GIZA ++ is used in the EMS to create word alignments.

Firstly, we tokenize all our corpora. These are then cleaned for translation model training. In this step, too-short sentences and sentences containing too many tokens are removed. In our case the maximal length of a sentence was 80 tokens.

The remaining sentences are used to train a "truecasing model". A truecasing model is a list of words and the frequency of their different forms. This model is applied to the corpora in the next step changing the words to their most frequent form. This truecaser is also used on the evaluation corpus and the corpus used in the tuning step.

Afterwards the corpora are prepared for the step in which the alignments are extracted. That means creating two vocabulary files, one for each side of the parallel corpora. Those files contain the words in the corpora and a conversion into numbers for each of them.

GIZA ++ is then run bi-directionally to generate two alignments for each sentence. Those are then symmetrized in order to receive the final alignments for the words in each sentence. We chose to use the commonly used grow-diag-final-and heuristic to combine the two alignments. The maximal length of a phrase was limited to five words.

Those alignments are then further processed in two different steps. One step builds a lexical translation table ordered by maximum likelihood. The other step extracts all phrases found in the aligned sentences and puts them together into one file. The outcome of those steps is then again combined to create a translation table, which contains the probability for each phrase pair.

The extracted phrases are further used to build a reordering model, which contains penalties for skipping words to build phrases. In our case, we chose to use monotone, swap and discontinuous orientation to be taken into consideration in the training of this model. Further we chose a bi-directional orientation on foregoing and following phrases and a base on both source and target language. This setup is also commonly used in machine translation workshops.

Concluding this step of the training, a new configuration file is automatically created that contains all the correct paths to the models trained in this step.

In the language model training, we also use already tokenized corpora in the target language. The model is trained on those corpora using the SRILM Toolkit. Those language models are interpolated and each of them – including the interpolated language model – is then binarized. Those binarized versions of the language models are then also put into the automatic configuration file.

In a tuning step pre-set weights are applied to the previously trained models. Those weights are then automatically changed numerous times until the weights that maximize the performance are found. In our case, this is done by applying the minimal error rate training (MERT) algorithm. In this work we used the BLEU score in this algorithm as reference. The automatic configuration file is then updated with the weights found in the tuning.

This system is then used to decode the evaluation corpus. After removing all markups in a separate step, this translation is then compared to a reference translation and rated with a BLEU score.

All those steps, besides the tokenization are combined in Moses' EMS and are automatically traversed. While doing so, the EMS can display a graph that shows the step it is actually conducting at the moment. An example of such a graph can be found in Figure 1 in the appendix.

4.2. Data

The corpora used to train the baseline system were in the translation model training the Symantec Manuals (about 2M lines), which are considered close-to-domain data as they consist of sentences, which cover the same topic as the to-be-translated UGC – namely the description and guidelines of Symantec products. Therefore, they contain terms and expressions that are vivid for forum conversations on this topic as well. On the other hand, they do not adhere to the same style as UGC does, and therefore cannot be considered as completely in-domain data. In addition, the kftt corpus (about 2.5M lines) is used as a large out-of-domain corpus to supplement our training data in order to improve the translations. Furthermore, the Tatoeba corpus (about 200K lines) is used as a close-to-style supplement as it contains more informal speech as usual official corpora do. Hence, its appearance is closer to UGC than either the Symantec corpus or the kftt corpus to some degree. In the language model training we used all of the above-mentioned corpora and additionally a set of forum data (about 50K lines) only available in Japanese.

The data for tuning and testing both contained 500 lines of Symantec forum data that exist in both, English and Japanese.

4.3. Variations

Various combinations of the corpora in the language model training and translation model training were tested. In the following table selected systems we have built are listed and explained. They are referred to by the number they hold in table 1.

Nr	Translation Model Corpora	Language Model Corpora	BLEU
1	Symantec	Symantec	20.42
2	Symantec	Symantec, Forum	20.83
3	Symantec, kftt	Symantec, Forum, kftt	19.99
5	Symantec, kftt	Symantec, Forum	21.31
9	Symantec, kftt, Tatoeba	Symantec, Forum, Tatoeba	22.10

Table 1: Summary of the various SMT systems trained

In table 1, the number displays the number of the experiment as a reference. The columns "Translation Model Corpora" and "Language Model Corpora" list the corpora used in translation model training and language model training respectively. The column "BLEU" shows the BLEU score for each experiment. A full version of this list that contains all the experiments of this thesis can be found in table 13 in the appendix.

Firstly, an SMT system is trained merely on the close-to-domain Symantec manuals as a very basic system as a starting point. This provides us with a BLEU score of 20.42 as displayed in experiment 1.

By adding the monolingual Japanese forum data to train the language model in experiment 2, these results did improve by 0.41 points, resulting in a BLEU score of 20.83.

By adding the kftt corpus to both language and translation model training in experiment 3, the BLEU score decreased to 19.99. If the kftt corpus is used only in the translation model training as in experiment 5, the BLEU score increases to 21.31. Compared to the system mentioned above, this improvement shows an increase by further 0.48 points of BLEU. This can be explained in respective of the style used in the kftt corpus, as it is very different from the writing style used in UGC. Adding the kftt corpus to the language model, the system will create sentences further away from the appearance of UGC. However, only using the kftt corpus in the language model provides various additions to the dictionary and thereby adds the ability to translate words, which are usually not included in manuals among other additions. For example, the kftt corpus contains personal pronouns and other stylistic differences from mainly descriptive writing styles, from which the system can prosper.

A further addition of the Tatoeba corpus in experiment 9 provides yet another increase of the BLEU score by 0.79 points, resulting into a total of 22.10. Again, this can be explained by the training data, which is again slightly closer to the UGC that is to be translated. Furthermore, the Tatoeba corpus provides more colloquial entries, which cannot usually be found in manuals or the kfft corpus, but are regularly used in UGC. In this case, adding the corpus to the language model improved the systems BLEU score slightly more, compared to using it only in the translation model (22.04 BLEU).

The explanation on why the addition of the kfft corpus to the language model resulted in a decrease, but the addition of the Tatoeba corpus resulted in an increase, even if only a very slight one, may be found in the stylistic differences of those corpora. Where the kftt corpus contains more formal language, the Tatoeba corpus also contains colloquial speech, as mentioned above, and is therefore closer to the style used in UGC, which often containins colloquial expressions and sentences structure.

Considering those results, the system trained in experiment 9 was used as a baseline system with a BLEU score of 22.10. This includes the Symantec Manuals, the Tatoeba corpus and the kftt corpus in the translation model and the Symantec Manuals, the Tatoeba corpus and the Symantec forum data in the language model.

5. Synthesizing a Corpus

The SMT baseline system, which was created in our work, was used to automatically translate a large amount – about 395K lines – of as of yet untranslated English forum data, provided by Symantec. This generates a synthesized corpus, which is used to retrain the SMT system once more.

The idea behind this approach is mainly to find new phrases for the translation. Obviously, using an already existent SMT system to translate text, which is then used to retrain that very same system, will not provide new word translation possibilities. That is because words that were not translated before due to impossibility of translation are more than likely to remain that way. Yet, the translation model is supposed to gain new phrases as new word combinations may appear that can be mapped on the same phrase in the target language and vice versa. Using only in-domain data is supposed to strengthen this effect as domain-specific expressions and phrases are thereby effectively handled.

For example, for a German to English translation, a baseline system might already have the entry to translate "europaeischen" with "in europe", but does not contain any further phrases containing "europaeischen". Using this baseline system to translate a monolingual German corpus that contains a sentence that includes the phrase "im europaeischen", the part "europaeischen" will be translated to "in europe". If this synthesized corpus is then used for retraining, the phrase "im europaeischen" can be extracted with the translation "in europe".

5.1. Retraining the Baseline System

In the retraining of the baseline system, we used the original text of the synthesized corpus together with the automatically translated Japanese version of the text in the translation model training. If used in the language model training, only the automatically translated Japanese version was used.

The synthesized corpus decreases the BLEU score by 1.21 points if used in both, the language and translation model training to 20.89.

If the synthesized corpus is used only in the training of the translation model, the BLEU score decreases even further by 0.37 points, resulting into 20.52 BLEU.

A final retraining in which the synthesized corpus was used only to retrain the language model resulted in a BLEU score of 20.98. This retraining still shows a decrease of 1.12 points compared to the baseline system. Yet it provides an increase in BLEU compared to both variants that use the synthesized corpus in the translation model as well.

This increase is rather unexpected, as a decrease in the value of the language model would be expected in this scenario. The reason for this expectation is that the Japanese side of the synthesized corpus does not mirror the form of well-formed Japanese sentences in most cases. Therefore the use in language model training should worsen the overall outcome as it adds a large amount of incorrect information.

5.2. Evaluation

As the BLEU score merely describes similarity of the automatically translated data and a reference translation, even a translation with a comparably low BLEU score might still offer an improvement in comparison to the translation created by the baseline system. Therefore, the translations of the system using the synthesized corpus in the translation model were manually scanned and compared to the translations of the baseline system.

Some translations of the system using the synthesized corpus did include words or phrases, which were missing in the translations created by the baseline system. On the other hand, word order and grammar in those translations were severely damaged. As the words and phrases, which were newly added by the retrained system, did mainly include words and phrases that are not needed to understand the content of a sentence and did not add further important information, the damage in sentence structure outbalances the gain of those words and phrases.

6. Pre-editing Rules

One of the main problems that tremendously decreases the translation value of UGC are the differences from standard written text in terms of style and error rate, including special or wrong spellings, colloquialisms, etc. Those differences can't be handled by an SMT system, as neither of them is likely to appear in the training data, which is based on standard written language.

In order to achieve better translations, a major task is to bring the source text closer to the corpora used to train the MT system by getting rid of differences and special occurrences that appear only in forum data. This can be achieved through pre-editing the texts. To approach this task, we use the Acrolinx tool to apply pre-editing rules on the source text to change its style and correct its errors without interfering with its meaning. The first section will provide further details on how the rules were created. Following that, detailed insight into the procedure of evaluation is provided. That includes the way the data was prepared and the rating system that was used. Afterwards, the rules will be explained one-by-one, including information about the motivation of the rule, the results and an explanation on why the rule is working well, or not. Concluding this chapter, we will explain at which point the rules were applied to which data. Additionally we give insight into the results of retraining the SMT system with corpora on which the rules are applied.

6.1. Developing the Rules and preparing the Test Data

The rules described in the following sections named "Sentence Splitting" and "Rules Previously Developed for English to German Translations," were used on single sentences of English forum data, provided by Symantec. Then, those changed sentences were translated using the Moses translation system that was trained for this thesis. In order to be able to test the value of a rule without interference any errors, besides the ones that were to be tested, were removed from the sentences before translation. For each rule two lists were created: One containing translations of sentences without the rule applied and one with the rule applied.

The rules described in the section "Rules Developed for English to Japanese Translations" were newly created. In order to do so, the translation of the test data was manually scanned for problematic translations in respect to the original sentences in English. Afterwards, those findings were tested on a set of sentences that were not used in the training process of the SMT system using background knowledge of the Japanese language. Therefore, those ideas were manually applied to sentences of English forum data. In case those changes were deemed to improve the translation compared to the translation of the original sentence, the same procedure as described for the sections "Sentence Splitting" and "Rules Previously Developed for English to German Translations" were applied.

6.2. Evaluation Process

After creating the two lists for each rule, sentences that contained no change at all, were noted as "=" and erased from the lists.

The two lists for each rule with the remaining items were then handed to two annotators separately. Both speak Japanese as their mother tongue and possess at least conversational skills in English. The annotators were instructed to rate the sentences as followed:

++ The sentence is understandable and nearly perfect in the Japanese language.

+ Both sentences are understandable, but the sentence with the "+" is easier to understand or closer to the Japanese language.

△ Both sentence are equally good (understandable), yet contain a difference

▽ Both sentence are equally bad (hard to understand), yet contain a difference

– Both sentences are not easy to understand or contain parts, which are hard to understand, but the sentence rated with an "–" is even worse

– – The better sentence is still not easy to understand but the sentence rated with "– –" is not understood at all, or contains unbearable flaws or mis-translations.

In case both sentences were not understood at all, the annotators were instructed not to rate the sentence.

Those ratings were then transcribed into numbers in order to compute the value of a rule. "++" equals two points, "+" one point, "△" and "▽" was rated zero points, as was the previously noted "=". "–" was rated minus one point and "– –" was rated minus two points. After this step the values for the translations without applying the rule were negated. That means, the "+" ratings for those sentences were counted as "–" ratings and vice versa. In a final step those values were added together and divided by the number of sentences for each rule. Hence, the value of a rule was mirrored by a number between -2, for exceptionally bad rules, and +2, for exceptionally good rules. In this scale, 0 means that a rule had neither good nor bad influence on average.

Rules with an overall rating below 0.5 were discarded.

6.3. Tested rules

The rules that were used in this thesis can be divided into three categories according to the nature of the changes they aim for within a sentence.

6.3.1. Rules addressing sentence length

Even though the idea behind this rule is not necessarily leading to a closer appearance with the corpora used in training, it may still pose an advantage in machine translation as it deals with another, commonly known problem of MT, namely the length of sentences. Usually, the longer a sentence is, the harder it is to create a good translation. This does not only apply to the work at hand, but can be considered a general problem in MT.

6.3.1.1. Sentence Splitting

The main purpose of this rule set is to produce shorter sentences by splitting long sentences at points without changing the meaning of the sentence. Afterwards, the resulting sentences are still within one line, but clearly separated by punctuation marks.

- *Then*

"Then" usually implies a chronological order of events. Therefore, the word occurs only in the right order of the events described in the text in a well-formed sentence. This holds not only for standard text, but also for forum text. Splitting a sentence at this word does not only provide two shorter sentences with an unchanged meaning, but also preserves the structure of well-formed sentences.

before	Please remove the network cable first then try to uninstall/reinstall NSM.
after	Please remove the network cable first. Then try to uninstall/reinstall NSM.

The overall results of this rules evaluation showed a tendency towards slight improvements of the translations. Three out of ten sentences have been considered worse in the translation without the rule by one annotator and four sentences have been rated with improvements with the rule applied.

The second annotator rated one sentence without the rule applied as worse, compared to its counterpart. Two of the translations with the rule applied to the source sentences have been rated as worse. Three were rated as better and one sentence even deemed as almost perfect in Japanese. Furthermore, one sentence was rated as equally good and three as equally bad.

By combining the ratings of both annotators we receive an overall result of 0.55 points according to the rating system described before.

- ***Therefore***

"Therefore" usually introduces a consequence. Splitting a sentence at this word still provides two well-formed sentences with an unaffected meaning.

before	I switch back and forth from one machine to the other, therefore I have different logins for each machine.
after	I switch back and forth from one machine to the other. Therefore, I have different logins for each machine.

The results for this rule were inconclusive. Mostly, the translations neither increased nor decreased in value. This is mirrored by the decisions of the annotators, where the first annotator rated one out of eleven sentences without any changes as worse than the translation of the version with applied rule, one better and two equally bad. In case of the translation with the rule applied, only one was rated better.

The second annotator rated three sentences equally good and three sentences equally bad. In the translation based on the sentences with the rule applied, one was rated worse, one better and one almost perfect, which indicates a slight tendency towards the use of the rule, displayed by an overall rating of 0.15. Even though this result would usually lead to discarding the rule, it is still used as it does not decrease the translation and is believed to provide improvement in combination with other rules. This assumption will be proven later in this thesis.

- ***So***

"So" poses a problem in many cases, as it is often not translated. Therefore the consequential meaning "so" displays is no longer recognizable in the translation. To avoid this problem, "so" not only leads to a sentence splitting, but also is changed into the synonym "therefore". Therefore, the same notes as above are effective.

before	I am a professional photographer and putting together some classes, so if I can provide you with any assistance, let me know.
after	I am a professional photographer and putting together some classes. Therefore, if I can provide you with any assistance, please let me know.

The results of the evaluation of this rule are almost the same as for "therefore". One annotator rated three translations out of forty without the rule applied worse compared to the translation with the rule applied, one better, seven equally bad and one equally good. For the translation with the rule applied, one sentence was rated worse and five were rated better.

On both sides, the other annotator deemed more sentences as at least understandable. Hence, the annotator was able to rate more sentences. However, the average of the ratings does not change dramatically. Out of the sentences provided, 16 sentences were rated. Two sentences without the rule applied were rated unbearable in Japanese, five worse compared to their counterparts, eight equally bad and one equally good. The translations of the sentences after using the rules resulted in ten sentences rated worse and seven better.

Adding those ratings together, this rule receives an overall rating of 0.15. As with "therefore", although no immediate improvement was found, the rule was still used, as an improvement in combination with other rules is supposed to occur.

- ***But***

"But" usually implies a contrary or disagreement in the foregoing sentence or phrase. Even though usually avoided in formal language, it commonly introduces a sentence.

before	It may be that it is working for now but as you can see there has been no input at all from symantec for 10 days
after	It may be that it is working for now. But as you can see there has been no input at all from symantec for 10 days

The results of the evaluation show almost no change in the value of the translations. One annotator rated two out of ten sentences for translation without the rule as worse, one better and one equally bad. In comparison the translations with the applied rule were rated worse in one case.

The second annotator rated in both versions two sentences each as unbearable in Japanese and three equally bad. Only one sentence in the translation without the rule applied was rated worse. This leads to an overall rating of 0.05.

The translation of the version with the rule in use can be translated as a split sentence with "○が" (○ga) in some cases. If the ○ was translated with the copula phrase "です" (desu) or "だ" (da) the translation would still form a well-formed Japanese sentence, but as the translation at hand shows rather a decrease than an increase, this rule was discarded.

A possible reason for this rule to result in worse translations is that even though the Japanese phrases "ですが" (desu ga) and "だが" (da ga) form well-formed and commonly used phrases in Japanese, there is a tendency to use "が" (ga) within a sentence as a coordinating conjunction. Furthermore, as stated above, "but" is usually not used to introduce a sentence in formal language. Therefore it does not appear often in a sentence introducing position in the corpora used to train the system. The distribution of the appearances of “but” within the training data is summarized in table 2.

	sentence introducing	within sentence	overall
but	1630 (6.7%)	12255 (93.3%)	24185

Table 2: Distribution of “but” within the training data

6.3.2. Modification Rules

This rule set mainly focuses on modifying the source sentences in two ways. One is aiming to make the sentences style closer to the style of the sentences that are found in the training data. The second focus is more language-specific in terms of target sentence appearance. The latter is reached by adding or changing words so that they are closer to the desired target word in the target language, resulting in an easier linear translation.

6.3.2.1. Rules Previously Developed for English to German Translations

In the still ongoing ACCEPT project, pre-editing rules for English to German translation were created to improve the translations. Five of those rules were picked out to evaluate the reusability of rules in different languages. We chose only rules that have been shown to improve the translations from English to German.

- *“Going to” → “Will”*

This rule changes “going to” into “will” so that the two types of future occurrences are unified.

before	I'm afraid this is going to be long winded to encompass all the details.
after	I'm afraid this will be long winded to encompass all the details.

For English to Japanese translation the effects of this rule are almost zero. The first annotator rated three out of ten sentences as equally bad and four sentences better than the translation of the sentence using the rule. Only two sentences have been rated better for the translation of the transformed sentences.

The second annotator deemed only two sentences of the translation with the rule applied as better, but deemed three sentences as worse. The translation of the unchanged version of the sentence was rated worse only once. Three translations were rated equally bad. This leads to an overall rating of –0.1. Therefore, this rule is discarded.

Even though there is a strong tendency towards the use of "will" in the training corpora, the rule had almost no influence on the value of the translations. The distribution of "will" in comparison to "going to" in the training data is summarized in table 3.

will	going to	overall
41326 (99.6%)	184 (0.4%)	41510

Table 3: Distribution of the future tense forms "will" and "going to" in the training data

This result may be explained by the fact that Japanese does not have a future tense. Present tense and future tense are expressed by the same sentence structure. The difference of those two forms does not, therefore, affect the translations. On the other hand, German is uses various future tenses. Hence the improvement by applying this rule with German as the target language can be explained.

- ***"not … any" → "no"***

This rule changes every appearance of "not … any" into "no". This also holds for contracted forms of "not", for instance "don't" will be considered as "do not". For example, the phrase "doesn't know of any problems" is changed into "does know of no problems". This rule was introduced because of the German sentence structure. In German "kennt keine Probleme" is more natural than the expression "weiß von keinen Problemen" in respect to the meaning of the sentence.

before	Carbonite does not know of any compatibility issues.
after	Carbonite does know of no compatibility issues.

Despite the positive results in English-to-German translations, this rule has no grounds in Japanese. This difference is confirmed by the ratings of one annotator, where one out of ten sentences was rated worse than the one with the applied rule, three equally bad, but one nearly perfect. For the translations of the changed sentences, one annotator deemed one sentence worse and three sentences better.

The other annotator rated three sentences equally bad. One sentence of the changed translation was rated better than the counterpart, and one nearly perfect in Japanese. On the other hand, two translations of the unchanged version were rated better and another two nearly perfect. Considering that there were few improvements compared to the almost perfect translations of the unchanged sentences, this rule was discarded due to a lack of difference in quality for the translations on average. This decision is explained also by the overall rating of –0.1.

A reason for this slight degradation on average might be that Japanese has a tendency to form sentences with "ない" (nai) as the negation being a part of the verb itself. Therefore sentences using “any” for negation with the negation being closer to the verb might be processed more easily. Furthermore, the lack of tremendous decrease in quality after applying the rule may be explained due to the fact that both versions in English are most likely translated into the same phrase in Japanese. Even with a tendency towards the negation with “no” in the training corpora, the translation should be almost the same Japanese sentence. The distribution in the training data is summarized in table 4.

no	not * any	overall
14952 (85.1%)	2616 (14.9%)	17568

Table 4: distribution of the negation with “no” and “not * any” in the training data

- ***“Have to” → “must” → “need to”***

This rule changes every appearance of “have to” into “must”. Even though this results in unnatural English sentences, this rule was used for English-to-German translation, as the German word “müssen” is closer to “must” than it is to “have to”. Furthermore a change from “have to” and “must” to “need to” was considered in English-to-German translation as it has shown slight advantages over “must” in the translations.

before	I'll have to agree with the Op.
after	I'll need to agree with the Op.

Both versions of this rule showed only slight changes in the translation to Japanese or no changes at all. Changing “have to” to “must” did not change the translation at all. In an evaluation by one annotator in both cases the change into “need to” was rated as equally bad for six out of ten sentences and equally good in two.

Furthermore, the other annotator rated the translation of sentences containing “have to” better in one case. The translations of the sentences with the change into “need to” even contained one translation that was rated as unbearable in Japanese. Further, another translation was rated worse than the original version and three sentences were rated as only slightly better. As a lack of improvement was found, expressed by the overall rating of –0.05, the rule was discarded.

The way to express the need to do something in Japanese differs vastly from the way to express it in English. Whereas German and English both express this meaning within a single verb, Japanese uses an entire phrase, namely “しなければなりません” (shinakereba narimasen)[3]. This phrase can be loosely translated into “if you don't do it, it won’t work out”. Therefore short phrases as provided by a

[3] Variations of this phrase do exist in Japanese.

single verb have the tendency to be translated into the phrase "必要がある" (hitsuyou ga aru), meaning "there is the need to…" rather than the long version shown above, which would need an insertion left and right of the phrase that contains the information about the things that have to be done.

- ***Reformulations***

This rule is composed of several smaller rules. It changes colloquialisms and conventionalized expressions into their standard language counterpart. For example, the word "cuz" is changed into "because" and the sign "&" is changed into the word "and".

before	Cuz I'm really worried, and my dad will be upset if we have to pay for it again.
after	Because I'm really worried, and my dad will be upset if we have to pay for it again.

The first annotator rated five out of ten translations without the rule much worse compared to the translation with the rule in use, one slightly worse, one equally bad and only one slightly better. Four sentences have been deemed slightly worse after translation of the changed sentences.

The same five sentences as above were rated as unbearable in Japanese by the second annotator as well. For the remaining translations four of the sentences with the rule applied were rated as slightly worse and better in another case. Three further sentences were rated equally bad and one equally good. Even though this rule did not result in good translations, it was still able to get rid of mistranslations and very bad sentence structures.

The overall improvement of translation quality obtained by applying this rule, which can also be found in the overall rating of 0.65, is determined by the nature of the task of translating UGC. Whereas the other rules mentioned in this section mainly focus mainly on target-language-specific problems, this rule aims at a language-irrelevant problem that is specific for the forum domain. By applying this rule, the text becomes closer to the training data of the MT system and therefore is easier to translate for the system.

- ***Deletion***

Some words in source texts tend to decrease the quality of a translation without providing necessary information in a sentence, for example "lol". This rule is created to delete such words.

before	Oh and the gadget is like a pink moon at the top lol...
after	Oh and the gadget is like a pink moon at the top

One annotator rated three out of ten translations based on sentences containing “lol” worse, one as equally bad and two as equally good. The translations after the deletion were rated better in two cases and almost perfect in one case.

The second annotator rated one translation of the version containing “lol” worse. The translations after removing “lol” were rated better in three cases and nearly perfect in one case. Two sentences were rated equally good and one equally bad.

Therefore, the overall rating for this rule is 0.65. This improvement is most likely to result from two points. One is the use of “lol” within a sentence as there are no set rules. The word may appear in the beginning, the end or even in the middle of a sentence. Furthermore, “lol” is not always clearly separated from the rest of the sentence and therefore the MT system may not be able to translate the word in the proper position. On top of that, considering the domain of the corpora that were used to train the system, it is more than unlikely that this word will appear at all in the training data. A second problem with “lol” is that in Japanese culture this acronym is usually neither used nor understood.

6.3.2.2. Rules developed for English to Japanese translations

As seen in the section before, some rules work only for certain languages. Hence, this section proposes rules that are especially created for English to Japanese translation.

- ***Add “please” to imperative***

There are different levels of politeness in Japanese. Requests or even imperatives are usually not uttered bluntly, as they are considered rather impolite. This holds not only for formal language, but also for commonly used speech, such as forum conversations. Therefore, in most situations, Japanese people would add the Japanese counterpart of “please” even to express an imperative. As conversations in support forums rely strongly on instructions, it is of major interest to translate correctly the imperative forms.

before	Go in Firefox's Tools Options-Privacy
after	Please go in Firefox's Tools Options-Privacy

One annotator rated the translations without “please” worse in one case, equally bad in one case and slightly better in two cases. The translations after adding “please” were rated better in four cases and almost perfect for three translations.

One translation was rated as unbearable in Japanese among those without “please” by the second annotator, but three slightly better. After adding please to the source side, three sentences were rated nearly perfect and another two better than the counterparts. Three sentences were deemed equally good and one equally bad. Those ratings add together to an overall rating of 0.53.

The original version of the sentences resulted in translations that did not contain an imperative on any politeness level. This might occur because English expresses imperatives not only through the verb form, but also through the position of the verb in the sentence. This difference in position might be insignificant though, as the sentence structure of Japanese is entirely different anyway. By adding “please”, the translation is most likely to result in “ください” (kudasai), which is a typical way to form a request in Japanese.

- ***“not” → “un–”***

The rule changes appearances of the adjectives “fair”, “sure” and “certain” in combination with “not” into the synonym counterpart using the prefix “un”. In some cases, negation is a problem in MT as it is left out in some translations. The idea behind this rule is to combine the negation with the adjective in order to produce a single word, which is supposed to be easier to translate. As in English, there is a way to express the negation of an adjective by adding a prefix – or, more specifically, a kanji-word – without changing the meaning of the expression in Japanese as well. This is done by adding “不” (fu)[4] in front of the respective adjective – or, in some cases, noun.

before	I'm not sure what issues there may be.
after	I'm unsure what issues there may be.

In the evaluation of this rule, the first annotator rated ten out of 17 translations worse in the original version. After applying the rule, two translations were regarded as better than the counterpart and one even as nearly perfect. Four sentences were rated equally bad.

The ratings of the other annotator only contain one slightly worse translation for the sentences with the rule applied, four were rated better and two nearly perfect. Seven other sentences were rated equally bad. This leads to an overall rating of 0.62.

As described before, the negation is translated more easily with this addition because it is now part of the word itself, which might explain the positive results. Another explanation for this might be that there is a tendency in Japanese towards

[4] Depending on the word the reading of this kanji can also be “bu”.

using the in-word negated form of an adjective, rather than negating it within the sentence. Concerning the distribution in the training data, we found a clear tendency towards the use of the in-word negation. Table 5 summarizes the findings.

	"un"-negation	"not"-negation	overall
sure	97 (90.7%)	10 (9.3%)	107
certain	38 (100%)	0 (0%)	38
fair	144 (75%)	48 (25%)	192
overall	279 (82.8%)	58 (17.2%)	337

Table 5: Distribution of negation using "un-" vs. negation using "not" in the training data

- ***Once more (etc.) → again***

The translation of expressions with the meaning "again", namely "once more", "once again" and "one more time", have shown to be problematic in translation with the MT system. Therefore, this rule changes all those appearances into the synonym "again", which has shown the best results throughout the test translations.

before	I clicked <Submit> one more time.
after	I clicked <Submit> again.

One annotator rated the original translations in three cases worse and only once slightly better. The translations received after applying the rules were rated better in two cases. Further, one sentence was rated equally good and three equally bad.

The translations with the rule applied received two nearly perfect ratings by the second annotator and five better ratings. However, one sentence was rated unbearable in Japanese.

Adding those ratings together, we receive an overall result of 0.55. An explanation for these positive results can be found in the test translation: Phrases like "one more time" were often translated into "一つの時間" (hitotsu no jikan), which is a word-by-word translation of "one time" – in more detail, the word time is here used like a countable object. The "more" is not translated in most cases. "Once more" shows similar problems. Even though the translation is close to the right phrase – もう一度 (mou ichi do) – the "more" is usually not translated – except in one case of the sentences used for the evaluation of this rule – resulting in the translation "一度" (ichi do), which means "once" or "one degree". In contrast the word "again" is mostly translated into the correct counterpart "再び" (futatabi).

We were able to find a clear tendency towards the use of "again" compared to the synonym forms. Those findings are summarized in table 6.

	findings	percentage
again	11159	99.9%
once again	96	< 0.1%
one more time	1	< 0.1%
once more	20	< 0.1%
overall	11276	

Table 6: Distribution of "again" and its synonyms in the training data

As the separate tokens for each synonym of "again" are found in the corpus – but can only rarely be found in combination forming the respective phrases – the above described word-by-word translations are bound to happen. Thus, unifying the appearances of all synonyms in the UGC to be translated vastly improves the quality of the translation.

- ***Deletion (Addition to the rule)***

Even though deletion of words was already handled in English-to-German translation, one addition to this rule can be made for Japanese, even though it had no impact or even negative impact on English to German translation. The word "well" – if not used as an adverb or adjective[5] – does not add important information to a sentence. Yet, removing it did not improve the value of the translations for English to German. In translations to Japanese on the other hand, the word was often problematic. After removing "well" from the sentences, the translations improved, contrary to the results of the other language pair.

before	Well, I have them on a USB drive now.
after	I have them on a USB drive now.

The first annotator rated one translation based on a sentence before removing "well" unbearable in Japanese and six additional sentences as worse compared to their counterparts without "well". After removing "well" only one sentence was rated worse. One sentence was rated equally bad.

The translations that still contained "well" received one rating for being worse by the second annotator. After removing "well" from the source sentences, the translations received one rating as being nearly perfect. Two other sentences were rated as better and only one sentence as worse compared to its counterpart. Three sentences were rated equally bad and one equally good.

The positive results for this rule, also expressed in the overall rating of 0.55, follow the same explanation as for the already existing rule for English-to-German. One addition to that is that, even though equivalents to this use of "well" can be

[5] The use of a part-of-speech tagger rules out the mapping of this rule on "well" as an adjective or adverb.

found in Japanese, these equivalents differ in nuance. For example, the two Japanese words "えっと" (etto) and "まあ" (maa) can both[6] be used to translate "well", but differ vastly in nuance so that they cannot be regarded as synonyms and are unlikely to appear in the same context. Furthermore, "well" in this use is very unlikely to appear in the training data.

- ***"Help" / "aid" → "support"***

This rule changes every appearance[7] of "aid" and "help" into "support". In the case of the word "help", this change was restricted to "help" in a noun position. This was assured by using a part-of-speech tagger. The reason for this is that, in verb position, "help" can appear in combination with other words, for instance "help along" and "help out". As those words are sometimes separated, a rule matching all such appearances could not be written.

before	If you need any help regarding your norton program, please come back over here.
after	If you need any support regarding your norton program, please come back over here.

One annotator rated two out of twenty translations based on sentences that still contained help or aid respectively better than the translations based on the version with the changes. Eight sentences were rated worse and one unbearable in Japanese. Four further sentences were rated equally bad. The translations based on the modified sentences were rated nearly perfect in one case, better in another case and worse two times.

The second annotator deemed four translations based on the original versions unbearable in Japanese and another one worse compared to the version with the rule applied. One translation was rated better, two equally good and one equally bad. For the translations after applying the rule, three were rated nearly perfect, one better and three slightly worse.

By those ratings, the rule received an overall rating of 0.53.

Even though these three words cannot precisely be considered synonyms, mapping all appearances of any of those words on "support" increases the value of the translation for different reasons.

One reason is the domain of the UGC that is to be translated. As the UGC consists of conversations of a support forum, many appearances of "help" and "aid" are used in the same context in which "support" can be used as well.

[6] There are further possible translations.

[7] Terms such as „First Aid“ are excluded from this rule.

The second reason is the translation of the different words. “Support” has only a single translation “サポート” (sapooto) in respect to the training data. Whereas “help” and “aid” have many different translations, each containing slightly different nuances that cannot be used in the same context. For example “役立つ” (yakudatsu), “助け” (tasuke) and “支援” (shien). Both annotators stated that the distribution of the translations were very random and only rarely fit the described situation. The translations using “support” instead were largely accepted. In only one case was the translation “サポート” (sapooto) deemed incorrect.

Another reason for the advantages of using “support” can be found in the distribution of those three words in the training data. The findings are summarized in table 7 below.

	findings	percentage
help	15780	37%
aid	427	1%
support	26479	62%
overall	42686	

Table 7: Distribution of “help”, “aid” and “support” in the training data

6.3.3. Error Checking

This rule set addresses a very large difference between common texts and UGC, namely the error rate in the texts. Common texts, which are used primarily to train SMT systems, usually do not contain any spelling or grammar errors. UGC on the other hand is more often than not rich with such errors. Some errors occur accidently, whereas others are used on purpose, such as when users disregard grammatical rules in order to increase their writing speed, etc. By eliminating these errors, the forum texts to be translated become closer to the training data, which usually contains no errors. Hence, the value of the translations should be improved.

Errors are detected using a spell checker and a grammar checker respectively. Both are part of the Acrolinx tool.

6.3.3.1. Spelling

Spelling errors have proven to be untranslatable by the SMT system, as those false words do not appear in the training data. In most cases, they are simply imported into the translation in the same form as they appeared in the original text thereby further destroying the sentence structure of the translations.

before	Subcription Renewal pricing.
after	Subscription Renewal pricing.

Both annotators rated 13 out of 21 translations based on the sentences without the spellchecking as unbearable in Japanese. One annotator even rated one counterpart as nearly perfect after the spellcheck. For the remaining sentences, the first annotator rated one translation on each side better compared to its counterpart and three sentences equally bad.

The second annotator considered two translations of the sentences after the spellchecking as worse, five as equally bad and one as equally good.

Considering those results, spellchecking is by far the most valuable pre-editing rule with an overall rating of 1.19. It is the only rule that received a rating higher than 1.

The 13 sentences rated unbearable by both annotators were all part of two out of three categories of spellchecking tested. These two categories were misspelling and the use of lower case for the pronoun "I". The latter resulted in the same problem as the misspelling did, namely merely importing the misspelled word into the translation and thereby damaging the sentence structure.

The third case tested was the use of lower case at the beginning of a sentence. This error type had only little impact.

6.3.3.2. GRAMMAR

Grammar errors are very common in UGC, regardless if they are on purpose or not. Especially in English, these errors sometimes even change the meaning of sentences or phrases. Such errors are for example "it's – its confusion" or "noun – adjective confusion". Other errors simply worsen the sentences, for example "an – a confusion" or "singular – plural misuse".

To evaluate the impact of grammar checking, five grammar rules have been selected and handed to native speakers within 25 translations of the original and the corrected version each.

before	I see that in most cases Norton Anti-Phishing scans a page only after its been fully downloaded.
after	I see that in most cases Norton Anti-Phishing scans a page only after it's been fully downloaded

One annotator rated four sentences on each side better compared to their respective counterpart and one translation on each side worse. One sentence was rated equally bad and another equally good.

The other annotator deemed one translation of the original sentences worse. The translations after the grammar check received ratings of better in two cases and worse in three cases. Two sentences were rated equally bad and another two equally good.

After inputting the annotations into the rating system, both annotators gave this rule a rating of exactly 0 on average. Furthermore, four out of the 25 sentences did not show any change at all in their translations.

The rules, which did contain the translations without changes, are found in the category addressing grammar errors that simply worsen the sentence in quality in the source language. A possible explanation for the missing impact of "a – an" confusion might be the fact that Japanese does not use any articles at all. In case of misuse of singular and plural, for example the phrase "those option", an explanation might be the usually missing plural form[8] in Japanese.

Slight changes were found in the other category of rules addressing possibly meaning changing errors. In case of "its – it's confusion" and "noun – adjective confusion" the changes did show an impact on the translations. Yet, those impacts were inconclusive as both versions received equally distributed improvement and degradation ratings in the test data.

The last tested grammar rule addressing incorrect verb forms, for example, the phrase "it tell me", did follow the same pattern as found in "its – it's confusion" and "noun – adjective confusion".

Even though the grammar rule set received a rating of 0, meaning no impact, the rule was still used in the larger evaluation for two reasons. One reason is that the rule set did not damage the translations on an average and might still work well in combination with other rules. The other reason is that not all grammar rules could have been tested in the rule evaluation step and those that were tested could be evaluated only on a small scale. Hence, the impact might be different on a larger amount of data.

Table 17 in the appendix shows the ratings of both annotators for each rule, divided into the three sections in addition to their computed value.

6.4. Usage of the Rules

The rules that were decided on after the evaluation step were implemented into the Acrolinx tool and then automatically applied to the test data before translating it with the Moses SMT system that was trained before.

This pre-editing step resulted in a BLEU of 22.69. This expresses an increase by 0.59 points compared to the results of the unchanged test set.

As some of the pre-editing rules do completely eliminate certain sentence structures and phrases in the texts that are to be translated – for example, there will be no sentence containing "therefore" in the middle of a sentence anymore – a further experiment was conducted.

[8] Even though a plural can be formed in Japanese, this is very rarely done and seems unnatural in most cases.

In that experiment, the pre-editing rules were not only applied to the text that was to be translated, but also were used on the source language side of the training corpora before retraining the system. By adding this step, structures that will not appear in the texts to be translated will also not appear in the training data. Even though some rules are supposedly not triggered in the training data – for instance the rule sets addressing errors and those addressing deletion – the whole rule set was used.

This experiment received a BLEU score of only 21.98, which expresses a degradation of 0.71 points compared to the translation of the changed test set with the baseline system.

An explanation for this degradation may be found in the way the rules are applied to the corpora. As the rules are mapped only on the source language side of the corpora, the reference translation remains untouched. For example, in the case of the rule that changes "help" and "aid" into "support", this would mean that one benefit of the rule, namely the one that there is only one translation for support, would vanish. Therefore, this approach could only be beneficial if both sides, source language side and target language side, were changed at the same time.

Concluding the evaluation of the rules, the effect of the "split sentence" rule set was analyzed. This was done by comparing the BLEU scores of different SMT systems that differed only on the point of whether the "split sentence" rule set was used or not.

The translation by the baseline system of the test data with all rules applied received 22.69. Without the use of the "split sentence" rule set, this score decreased by 0.11 points to 22.58.

In the case of the system retrained on the pre-edited corpora, the BLEU score of the translation was 21.98. After removing the "split sentence" rule set from all data – the corpora used for the retraining and the test data – the BLEU score decreased by 0.7 points to 21.28.

These results give evidence for the value of the "split sentence" rule set in combination with the other rules.

6.5. Human Evaluation

In order to gain further insight into the interaction of the rules and the value of the changes applied to the text, we conducted a human evaluation. In this evaluation, native speakers of Japanese were asked to rate the translations generated by using the previously-trained baseline system to translate the original test corpus of the SMT training against the version with the rules applied.

6.5.1. Setup of the Data

Three evaluators were presented with an Excel table containing 100 selected sentences from the 500-sentence test corpus. In order to select these 100 sentences, we first erased all sentences of the test corpus that did not contain any changes[9]. Out of the remaining sentences, we randomly decided on 50 sentences that contained more than one flag and another 50 sentences that contained single flags. It should be noted that the "split sentence" rule was added to the table only if it appeared in combination with another rule.

The first column of the table contains a reference number for each sentence, which was hidden from the evaluators. Those reference numbers were later used to reorder the sentences.

The second column contained the original English sentence as a reference for the evaluators.

The following two columns contained both versions of the Japanese translations in random order. This means that, in 50% of the cases, the original translation was presented first and in the remaining 50%, the translation based on the sentences with the rule applied was presented first. The sentences were also randomly shuffled in order to mix both versions.

The last column was left empty for the ratings of the evaluators. They were asked to rate "1" if they found the translation presented first better, "2" if they deemed the translation presented second better and "0" if both translations were equally good or bad. Additionally, they were asked to rate translations that were entirely the same with "=".

Table 18 in the appendix shows an example for such a table.

6.5.2. Evaluation

First we rearranged the sentences by using the reference number and corrected the ratings[10]. Then, we computed the inter annotator agreement using Fleiss' kappa. This value allows us to compare more than two evaluators with the possibility of more than two ratings per sentence. The inter annotator agreement was rated "fair" (0.34) according to Fleiss' kappa rating scale. It should be noted, though, that this evaluation does not take the distance of the ratings into consideration. The four different possible ratings were "0", meaning that both sentences were equally good or bad, "1", meaning that the first presented sentence was better, "2" meaning that the second presented sentence was better and "=", meaning that there was no change at all. In this rating system "2" and "0" and "1" and "0" are not as far apart as "1" and "2" are. However, the rating "2", "1", "1" for

[9] This means that no rule was triggered in those sentences.

[10] In case the translation based on the sentences with the rules applied was presented first, "1" was changed into "2" and vice versa. Therefore, the rating "1" always means the original version was better and "2" always means the changed version was better.

example, was treated the same way as the rating “0”, “1”, “1” was, although the latter clearly displays a stronger agreement than the first example does. Therefore, the Fleiss’ kappa can be regarded only as an indication for the agreement.

Afterwards, we conducted two evaluation steps on the received ratings. In the first step, each rating of each evaluator was taken into consideration. This means that if a sentence was rated “0” by one annotator and “2” by the other two annotators, one “0” and two “2”s were taken into account.

Out of the possible 300 ratings 31% (92) of the translations were rated “1”. This means that the translation dropped in value in those cases after applying the rules. 42% (127) of the translations were rated “2”, which means they increased in value after applying the rules. 17% (51) of the translations were rated “0” meaning there was no change in value of the translation between the two versions. The remaining translations (10% (30)) were rated “=”. Those results are summarized in table 8.

Rating	Absolute	Percentage
1	92	31%
2	127	42%
0	51	17%
=	30	10%
Overall	300	

Table 8: Distribution of all ratings

In the second step, we combined all three ratings of each sentence into an average rating. For example, if a sentence was rated “1” by two annotators and “0” by the last annotator, the whole sentence was rated “1”.

By doing so, 25 ratings (25%) out of 100 possible ratings were “1”. 46 translations (46%) were rated “2”, 8 translations (8%) were rated “0” and 10 translations (10%) were rated “=”. The remaining 11 translations (11%) were inconclusive. This means they received a different rating from each evaluator. Those results are summarized in table 9.

Rating	Absolute	Percentage
1	25	25%
2	46	46%
0	8	8%
=	10	10%
Inconclusive	11	11%
Overall	100	

Table 9: Distribution of averaged ratings

Both evaluation steps show a tendency towards the translations based on the sentences with the rules applied.

As the sentences rated “0” and “=” neither damaged nor increased the translation value, we can also take a look at the results considering only those translations for which a change in value was recognizable.

For the first evaluation method, this evaluation would describe only a very slight tendency towards the usage of the rules (58%) against not using them (42%). The averaged method, however, shows a much stronger tendency towards the usage of pre-editing rules (65%).

An overview of the results of both evaluation methods and the restriction to certain sentences can be found in table 10 below.

Evaluation Method	Rating	All sentences	Only sentences with value change
Method 1	1	31%	42%
	2	42%	58%
Method 2	1	25%	35%
	2	46%	65%

Table 10: Comparison of the methods and the considered sentences

6.5.3 Explanation

A thorough examination of the presented sentences and the rules that were used in them reveals a correlation between certain rules and rating pairs.

Most translations (7/10) marked with “=” contained a single “Grammar” flag or “Grammar” flags in combination to the spelling rule “capitalize at beginning of sentence”, which also caused two of the ten “=” ratings as a single flag. The remaining “=” rated translation was caused by the single flag of the reformulation rule “might → may”, which also showed no remarkable influence in combination with other rules.

The grammar correction further appeared in the context of damaged translations six times. In cases where the grammar correction appeared in a sentence that showed an improvement, it always appeared in combination with other flags that showed a good influence on the translations. Therefore, it can be said that grammar correction does not improve the translations from English to Japanese, but contrarily tends to damage the translation if any change happens at all. This observation supports the ratings for this rule set as described in chapter “6.3.3.2. Grammar”. The reason for this is most likely the major difference between English and Japanese grammar.

The spelling rule “capitalize at beginning of sentence” can also be erased from the rule-set, as this should usually show no changes due to the truecaser used in Moses.

Further the reformulation rule “might → may” can also be discarded, as it did not show any influence on the translations. This is possibly due to the translations of both phrases in Japanese, which are most likely to be “かもしれません” (kamoshiremasen) in both cases.

Concerning the translations rated as “1”, which was the case for 25 out of 100 sentences after averaging the ratings, correlations besides the grammar rule-set and the spelling rule “capitalize at beginning of sentence” were also found.

Four sentences out of the 25 sentences rated “1” contained a wrong spelling correction. This is a major problem with this kind of rule, as there is usually more than one suggestion to correct a spelling error. The first suggestion is always chosen while automatically applying this rule, which does not necessarily lead to the correct choice. Furthermore, the rule applies also to words that are user-names or colloquial speech but that are not adapted to the reformulation rule-set – for example the abbreviation “prob” for “problem”. On the other hand, in case a spelling correction was correct, the sentence was always rated as better. One way to avoid this problem would be to improve further the spell checker in terms of suggestions, so these suggestions would be more likely to be correct in the given context. Another way would be to extend the “Reformulation” rule-set. The most straightforward solution though would be to refrain from automatizing the spell checking and to use it instead to guide users while creating content.

The last outstanding correlation to translation value decrease was the rule “add please to infinitive” in case of a wrong flag. This could be observed in six out of the 25 sentences rated “1”.

Yet, it should be noted that this rule was the most frequent rule in the 100 sentences with 45 out of 212 flags. Out of those 45 flags only ten damaged the translation of six sentences[11] when incorrectly used[12]. Two of those four sentences also contained major mistakes in spelling correction, which may also be the reason for those sentences having been rated with “1”. In three other sentences the rule was severely misused by flagging nouns or adjectives instead of imperative forms of verbs, which could be avoided by refining the rules implementation:

- [...] in Step 1(where clearly Norton protection is down), please step 3 and please step 4, there [...]
- [...] please port (4172) Remote IP address,please port: [...]
- [...] Symantec Resources, please like the Update Center or the [...]

28 flags of this rule in 22 sentences increased the value of the translation or appeared in a sentence marked as better. This contains correct flags as well as incorrect flags. The remaining six flags appeared in three sentences that were inconclusive and in one sentence that was rated “0”. This is summarized in table 11.

[11] In some sentences the rule was used multiple times.

[12] Only one correct flag appeared in the context of a sentence rated “1”. In this sentence it appeared alongside two incorrect flags of the same rule.

Rating	Correct	Incorrect	Overall
1	1	10	11
2	17	11	28
0	0	1	1
Inconclusive	0	5	5
Overall	18	27	45

Table 11: Correlation between correct/incorrect flags and evaluation for the rule "add please to imperative"

If we took out the all the sentences that contained only those rules that were found to be damaging the translation or not changing it at all, we would lose seven sentences that were rated better after applying the rules, but also get rid of all the sentences without any change (10), twelve sentences rated as worse, four sentences rated as equally good or bad and another four sentences with inconclusive ratings. This is summarized in table 12.

Rating	Original	Without "bad" rules
1	25 (25%)	13 (20.6%)
2	46 (46%)	39 (61.9%)
0	8 (8%)	4 (6.4%)
Inconclusive	11 (11%)	7 (11.1%)
=	10 (10%)	0 (0%)
Overall	100	63

Table 12: Distribution of ratings for all sentences against sentences containing only rules with bad or no impact

It should be taken into consideration that some of these rules that are left out may still appear in some of the sentences that were not discarded. Therefore, these sentences are likely to change in their ratings as well.

7. Combining Synthesized Corpus and Pre-editing Rules

Even though using a synthesized corpus for retraining did not improve the translations in an earlier step[13], we decided to combine the methods presented in this thesis. This was done because synthesizing a corpus did show vastly different impacts with even small changes to the SMT system or the data used to train it throughout the development stage of our baseline system.

7.1. Conduction

The corpus that was synthesized in this step was the same monolingual English forum text that was also used in chapter 5. Before translating the data we applied the pre-editing rules, which we developed in chapter 6, to the entire monolingual corpus automatically. Although pre-editing the training corpora did resolve in a degradation of the performance earlier[14], the reasons for the degradation were not present in this case. This is because there is no target side of the corpus that would be left unaffected. Furthermore, the style and domain of the monolingual corpus is exactly the target of the rules we developed. Additionally the pre-editing rules were also applied to the test data of the SMT system in the same way as described in chapter 6.4.

This synthesized corpus was then used to retrain the SMT baseline system by adding it only to the translation model training data. This resolved in a BLEU score of 22.20, which is a slight improvement compared to the BLEU score of the baseline system (22.10). Compared to the system trained in chapter 5, which used the synthesized corpus in the same way as it was used here[15] (20.52), the BLEU score increased by 1.68 points.

7.2. Evaluation

The data that was presented to three human evaluators was prepared the same way as described in chapter 6.5.1. To be able to compare the results, the same set of 100 sentences was chosen and we conducted the same two methods for evaluation. The inter annotator agreement for this annotation set was “fair” (0.26) on the scale of Fleiss’ kappa. Here again we face the same problem as described before concerning the fact that the distance between the ratings is not considered in this way of computing the inter annotator agreement.

[13] See chapter 5. “Synthesizing a Corpus” for comparison

[14] See chapter 6.4. “Usage of the Rules” for comparison

[15] Only to retrain the translation model

The first method of evaluation, which takes every single rating of each annotator into consideration showed only a slight tendency towards the usage of the rules in combination with a synthesized corpus with 103 "better" ratings out of 300 possible annotations (34.3%). 77 ratings were "1", meaning a degradation of the translations. The majority of ratings in this scenario (40%) fell on "0", meaning that there is no change in value of the translation. In this case, every rule influenced the translations in some way[16]. These results are summarized in table 13 below.

Rating	Absolute	Percentage
1	77	25.7%
2	103	34.3%
0	120	40%
Overall	300	

Table 13: Distribution of all ratings

The second method, which uses the averaged ratings for each translation pair, did not change those results vastly as shown in table 14.

Rating	Absolute	Percentage
1	26	26%
2	32	32%
0	42	42%
Overall	100	

Table 14: Distribution of averaged ratings

It should be mentioned though that no translation pair was inconclusive amongst the annotators[17].

When we compare the averaged annotations per sentences of this evaluation with the averaged annotations of the evaluation of chapter 6, we find only 15 true contradictions. This means only in 15 cases the annotators of one group rated with "1" while the other group rated with "2". Differences between "0" and "1" or "2" as well as differences that included an "=" rating or were inconclusive in one group were left out.

Those 15 translation pairs mostly contained the spelling rule "capitalize at beginning of sentence" (4 cases) or any rule of the grammar rule set (5 cases). The latter also was rated as damaging to the translations in those cases where the other group rated with an "=". This result backs the conclusion made in chapter 6 that suggested removing grammar checking from pre-editing for English-to-Japanese translations.

[16] No translation pair was rated "=", meaning that in every case of translation a change did occur.

[17] No translation pair received every possible rating (0, 1, 2) once.

We also compared only those ratings of both groups that were for translations that did not contain only rules that were rated as having no influence or mostly bad influence according to chapter 6.5.3. in combination with the support for those decisions we gained in this evaluation step.

By ignoring the same set of sentences in both averaged evaluations we get rid of nine true contradictions. The whole results of this comparison are summarized in table 15.

Rating	Group 1	Group 2
1	13 (20.6%)	15 (23.8%)
2	39 (61.9%)	18 (28.6%)
0	4 (6.4%)	30 (47.6%)
Inconclusive	7 (11.1%)	0 (0%)
Overall	63	
True Contradictions	6 (9.5%)	

Table 15: Comparison of the evaluations of chapter 6 and chapter 7 without rules marked as "bad"

This evaluation supports and strengthens the conclusions drawn in chapter 6 regarding the deletion of certain rules and rule sets such as the grammar rule set. Those tendencies can be trusted, as the true contradictions between the two evaluations are rather rare.

Even though this group expresses a tendency towards rating with "0", which means that the changes do not influence the translations that much, the remaining ratings display a slight tendency towards the usage of a synthesized corpus in combination with the pre-editing rules.

7.3. Comparison of the two Approaches

As the evaluation of both groups did show a tendency towards the usage of the respective approach, we set up a control group that compared the version with only the pre-editing rules applied (chapter 6) with the version that also uses the synthesized corpus (this chapter). This comparison was done to find out which approach is more beneficial. The control group evaluated the same set of 100 sentences. The data was prepared in the same way as for the other two groups.

First, we estimated the Fleiss' kappa to receive the inter annotator agreement for this group. According to the rating scale, we received a "moderate" agreement (0.42). Yet, again we face the same problems as in the other two groups concerning the disregarding of the rating distance.

To evaluate this group we used the same two methods used for the two foregoing groups. The first method, taking every single rating into consideration, showed a strong tendency (45%) towards the usage of the combination of pre-editing rules with a synthesized corpus over the usage of only the pre-editing rules (27.3%). 27.7% (83) of the 300 possible ratings deemed both approaches as equally good.

The second evaluation method, which averages the rating for each sentence first, strengthened this tendency. Here, 51% of the translations received an average rating in favor of the usage of synthesized corpus in combination with the pre-editing rules. 33% of the translations were rated better if only the pre-editing rules were applied and the remaining 16% were equally good. The results of both methods are summarized in table 16 below.

Tendency	Method 1	Method 2
Combination of pre-editing rules and synthesized corpus	135 (45%)	51 (51%)
Pre-editing rules only	82 (27.3%)	33 (33%)
Equal	83 (27.7%)	16 (16%)
Overall	300	100
Circularities		3 (3%)

Table 16: Summary of the evaluations of the control group

Within this evaluation we found three circularities. This means, for example, that the first group rated a changed translation for a certain sentence as better and group two rated the changed translation for the very same sentence as worse. However, the control group favored the supposedly worse translation over the supposedly better one. Those circularities appeared in one case together with a grammar rule in combination with the spelling rule “capitalize at beginning of sentence”. The two other circularities appeared with two translations that both contained only falsely flagged “add please to imperative” rules.

Despite the drop in BLEU score from 22.69 points for the version that uses only the pre-editing rules, to 22.20 points for the version using a synthesized corpus in combination with the rules, the control group clearly favors the latter. This preference shows the importance of human evaluation in comparison to relying merely on automatically-applied scores.

8. Discussion

In this thesis, we combined various approaches of domain adaptation for the domain of forum texts. This specific domain has been gaining more and more interest since the introduction of Web 2.0. One approach in this work is the usage of large out-of-domain corpora in combination with small in-domain corpora to train an SMT system using the Moses tool-kit. This approach is rather straightforward for domains for which bilingual corpora are rare or even not existent. Such an SMT system served as a baseline model.

The second approach that we used in this thesis is the synthesis of a bilingual corpus by translating monolingual in-domain texts with our baseline system. This synthesized bilingual corpus was then used to retrain the translation model of our baseline system.

The last approach, which was also the main focus of this thesis, is the creation and usage of pre-editing rules in order to make the texts that are to be translated closer to the data used in the SMT training.

In this chapter we will discuss the results of those three approaches and their combination. We will discuss positive and negative aspects of this work and present possible future work that could be based on this thesis.

8.1. Analysis

Within this thesis we demonstrated the benefits of domain adaptation for distant language pairs, such as English-to-Japanese, in the domain of support forum text.

While creating the SMT baseline system, we showed the importance of the usage of at least close-domain corpora in the training. Although out-of-domain corpora did clearly improve the performance of the translation model, in-domain corpora and close-domain corpora are essential for the language model training. This necessity is because the structure and style of different domains vary vastly between one another. In our case, adding the totally out-of-domain kftt corpus to the language model dropped the performance of the whole system by 0.48 points of BLEU in comparison to the same setup without using the kftt corpus in language model training. On the other hand, using the out-of-domain kftt corpus only in the translation model training together with the close-domain corpora on the Symantec manuals and the Tatoeba corpus combined with the in-domain monolingual forum data, as well as the two close-domain corpora in the language model training increased the BLEU score by 1.68 points in comparison to a very basic system trained only on the Symantec manual corpus.

We were also able to show the benefits of automatically applied pre-editing rules. In terms of BLEU, this application was especially beneficial while applying those rules only on the text that we want to translate. By doing so, the BLEU score increased by 0.59 points in comparison to the baseline system. It is to be said though, that a human evaluation showed that this pre-editing step does not improve the translations in all cases. Some translations with the rules applied did perform worse compared to the original translations. On an average though, the usage of automated pre-editing did improve the translations in almost half of the evaluated sentences (46%) and degraded them in one fourth (25%) of the sentences.

Furthermore, we were able to carve out rules that have a poor effect on the translations from English to Japanese though they showed good results for close language pairs, such as English-to-German and English-to-French. On the other hand, newly developed rules that are specific for the task of translating English-to-Japanese forum texts did show a very positive effect on the translations. Here we can highlight the "add please to imperative" rule, which mostly had a largely positive influence on the translations even if used wrongly. These two findings clearly show the need to create rules that are targeting certain language pairs.

A further interesting finding in this step was the influence of grammar checking on the translations. Although grammar checking did show good results on English-to-German and English-to-French translations in the ACCEPT project, this positive effect did not occur for English-to-Japanese translations. To the contrary, it typically influenced the translations negatively, if at all.

Another rule set should also be mentioned, namely the spell checking. This rule set did perform exceptionally well while testing the rules in an evaluation step where the rules were applied manually. However, if applied automatically, this rule set did worsen the translations in some cases. This was mainly because of wrong flags or wrong corrections. The reason for this negative influence is the poor performance of automated spell checking, as always the first suggestion in a list of various suggestions is chosen while applying this rule automatically. Those lists are not ordered by likelihood for a word to be correct in the given context, but by the computational cost of the correction. Thus, it must be said that either the spell checker used here must be improved in performance, or the spell checking should be done manually. If applied manually, it would be best to do so while creating content by guiding users simultaneously while writing.

We also tested the influence of the use of a synthesized corpus on SMT performance. This was done in two different steps. In one step, we used only the baseline system to translate a monolingual source language corpus and use that thereby-received synthesized bilingual data to retrain our SMT system. This approach not only reduced the BLEU score by 1.58 points, but also severely damaged the translations in terms of structure and grammar. In the second step, we combined the synthesis of such a bilingual corpus with the pre-editing rules. The rules were applied to the monolingual source language data before translating it with our baseline system to receive the synthesized bilingual data. This bilingual corpus was then again used to retrain the baseline system. This system was then used to translate the pre-edited test set. Such an approach of combining pre-editing rules with a synthesized corpus has not been done before. In comparison to the baseline system, the BLEU score increased by 0.1 points when using this combined method. On the other hand, compared to the BLEU score received by applying the rules to the test set without using the synthesized corpus at all, the score dropped by 0.49 points. Despite the drop in terms of BLEU score, human evaluators favored the usage of the synthesized corpus in the SMT system over the use of the pre-editing rules. Considering these two very different results for both approaches including a synthesized corpus, it seems as if this approach was very unstable and differed vastly in results depending on even small changes in the data.

To conclude, it can be said that this thesis clearly shows the benefits of domain adaptation for distant language pairs, such as English-to-Japanese, though certain exceptions must be made. Also, the combination of the three approaches that were handled in this thesis (usage of close-domain and out-of-domain corpora to train a SMT system; automated pre-editing; synthesizing a bilingual corpus) was proven to be conductive to effective translation.

8.2. Future Work

This work and its evaluation show various points that provide the possibility for future work.

In terms of the pre-editing rules, different adjustments could be made. For example, the deletion of certain rules that did influence the translations negatively such as the grammar rule set. By erasing those rules the performance should further improve.

Another adjustment that could be made is refining the rules. Some rules did trigger many false flags. An example of such a rule was the “add please to imperative” rule, which still performed well even in case of a wrong flag. The degradation of the translations that appeared in combination with this rule occurred mainly in cases where the flag was severely violating its task. One example of this type of violation is if the “please” were set in front of a noun instead of a verb. Refining those rules implementation could reduce such false flags, and should also improve the performance.

Furthermore, the spell checker needs to be refined as well, or totally deleted from the automated pre-editing. In case of the latter, it would be necessary to provide guidance to the users while creating their content, as a correct spell-checking did prove to be extremely beneficial for the task of translating UGC.

As the rules developed for English-to-Japanese translations in this thesis performed exceptionally well, the development of further rules that tackle the differences between the two languages should be targeted.

Another rule set that seems to be worth expanding is the split sentence rule set. All evaluators of each group stated that the main problem within the test data was, that the longer a sentence was, the worse the translations were. Translations based on sentences that were changed by the split sentence rule typically received better ratings. Besides the rules in this thesis, there are many more words that can be used to split a sentence. Furthermore, refining the rule set further might be beneficial as well. For example, the effect of splitting not only the sentence in the way that the both resulting sentences remain on the same line, but also inserting a line break between the two sentences could be interesting while using Moses.

The pre-editing of the training data as described in Chapter 6.4. "Usage of the Rules" might also be successful if the pre-editing step would be extended to affect not only the source language side of the training corpora, but also the target language side.

Other possible future work could be conducted to tackle the grammatical differences between English and Japanese. The word order differences that exist between an SVO and SOV language pose especially severe problems for SMT. To alleviate this problem, the texts to be translated could be automatically preordered prior to the translation. For this approach the training corpora would have to be reordered before training as well in order to achieve the right alignments. Furthermore, if this method were combined with pre-editing, the pre-editing step must be done before reordering the text, as the preordering might destroy patterns for the pre-editing rules or might lead to wrong changes.

In this thesis we used the Moses toolkit to train a phrase-based SMT system. Moses can also be used to train SMT systems based on other approaches. Therefore, it might be interesting to train and compare different systems that use different approaches. For example, hierarchical phrase-based SMT systems showed slight improvements for English-to-Japanese translations in the ACCEPT project[18,19]. On the other hand, this would also mean that some pre-editing rules might change in terms of effect on the translations.

[18] See table 19 in the appendix for a slight overview of the BLEU scores reached in the ACCEPT project for English – Japanese SMT systems.

[19] All work on English to Japanese translations were dropped in the ACCEPT project.

References

Takako Aikawa, Lee Schwartz, Ronit King, Mo Corston-Oliver and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. Proceedings of MT Summit XI, pages 1-7. Copenhagen, Denmark.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way and Josef van Genabith. 2011. Domain Adaptation in Statistical Machine Transla- tion of User-Forum Data using Component-Level Mixture Modelling. In Proceedings of the Thirteenth Machine Translation Summit, pages 285–292. Xia- men, China.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way and Josef van Genabith. 2012. Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data? Proceedings of EAMT 2012. Trento, Italy.

Nicola Bertoldi and Frederico Marcello. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. Proceedings oft he 4th EACL Workshop on Statistical Machine Translation, pages 182–189. Athens, Greece.

Arianna Bisazza, Nick Ruiz and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In Proc. of the International Workshop on Spoken Language Translation, San Francisco, USA.

Andrew Bredenkamp, Berthold Crysmann and Mirela Petrea. 2000. Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking. Proceedings of the 2nd International Conference on Language Resources and Evaluation, pages 667-673. Athens, Greece.

George Foster and Roland Kuhn. 2007. Mixture- model adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 128–135, Prague, Czech Republic.

Dimitry Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. Proceedings of the 23rd International Conference on Computational Linguistics, pages 376–384. Association for Computational Linguistics.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada and Kevin Duh. 2010. Head finalization: A simple re-ordering rule for sov languages. InProc.of WMT-MetricsMATR, pages 244–251.

Philipp Koehn. 2010. Statistical Machine Translation. Cambridge University Press.

Ralf Kühnlein. 2013. Nutzerverhalten-basierte Optimierung einer linguistischen KI. Diplomarbeit. Freie Universität Berlin, Fachbereich Mathematik und Informatik. Berlin, Germany.

Sabine Lehmann et al. 2012. Applying CNL Authoring Support to Improve Machine Translation of Forum Data (invited paper). Workshop on Controlled Natural Language (CNL). Zurich, Switzerland.

Sharon O'Brien and Johann Roturier. 2007. How Portable are Controlled Languages Rules ? A Comparison of Two Empirical MT Studies. Proceedings of MT Summit XI, pages 345-352. Copenhagen, Denmark.

Johann Roturier, Linda Mitchell, Robert Grabowski and Melanie Siegel. 2012. Using Automatic Machine Translation Metrics to Analyze the Impact of Source Reformulations. Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA). San Diego, USA.

Hua Wu, Haifeng Wang and Chengqing Zong. 2008. Domain adap- tation for statistical machine translation with domain dictionary and monolingual corpora. In Coling 2008, 22nd International Conference on Computational Linguistics, pages 993–1000. Manchester, UK.

Appendix

	Rule	Annotator	In use?	+ +	+	△	△	–	– –	Val	Av. Val
Sentence length											
	Split sentence										
	Then	1	✖					3		0,7	0,55
			✔		4						
		2	✖			1	3	1		0,4	
			✔	1	3			2			
	Therefore	1	✖		1		2	1		0,1	0,15
			✔		1						
		2	✖			3	3			0,2	
			✔	1	1			1			
	So	1	✖		1	1	7	3		0,15	0,15
			✔		5			1			
		2	✖			1	8	5	2	0,15	
			✔		7			10			
	But	1	✖		1		1	2		0	0,05
			✔					1			
		2	✖				3	1	2	0,1	
			✔						2		
Sentence appearance											
	Reuse of previously developed rules of English to German										
	“going to” → “will”	1	✖		4		3			– 0,2	– 0,1
			✔		2						
		2	✖				3	1		0	
			✔		2			3			
	“not … any” → “no”	1	✖	1			3	1		0,1	– 0,1
			✔		3			1			
		2	✖	2	2		3			– 0,3	
			✔	1	1						
	“have to” → “must” →“need to”	1	✖			2	6			0	– 0,05
			✔								
		2	✖		1					– 0,1	
			✔		3			1	1		
	Reformulation	1	✖		1		1	1	5	0,6	0,65
			✔					4			
		2	✖			1	3		5	0,7	
			✔		1			4			
	Deletion (lol)	1	✖			2	1	3	0	0,7	0,65
			✔	1	2						
		2	✖			2	1	1		0,6	
			✔	1	3						
	Newly developed rules for English to Japanese										
	Add “please”	1	✖		2		1	1		0,6	0,53
			✔	3	4						
		2	✖		2	3	1		1	0,47	
			✔	3	2						
	“not” → “un-“	1	✖				4	10		0,82	0,62
			✔	1	2						
		2	✖				7			0,41	
			✔	2	4			1			
	“once more” (etc.) → “again”	1	✖		1	1	3	3		0,4	0,55
			✔		2						
		2	✖							0,7	
			✔	2	5				1		
	Deletion (well)	1	✖				1	6	1	0,7	0,55
			✔					1			
		2	✖			1	3	1		0,4	
			✔	1	2			1			
	“help” / “aid” → “support”	1	✖		2		4	8	1	0,45	0,53
			✔	1	1			2			
		2	✖		1	2	1	1	4	0,6	
			✔	3	1			3			
Errors											
	Spelling										
	Spelling	1	✖		1		3		13	1,24	1,19
			✔		1						
		2	✖			1	5		13	1,14	
			✔	(1)				2			
	Grammar										
	Grammar	1	✖		4	1	1	1		0	0
			✔		4			1			
		2	✖			2	2	1		0	
			✔		2			3			

Table 17: Summary of all ratings of both annotators for the single rule evaluation (chapter 6.3.)

	英語	日本語１	日本語２	評価
RF298	I had used Norton 2003 ghost with no problems but with limited USB support I wanted to get out of the DOS environment.	私は Norton2004 使った怨霊が問題なくを限定したUSB サポートを取得するには DOS 環境が不足しています。	私は Norton2004 怨霊を使った問題が限定されていませんが、USB サポートを取得したかったのは DOS 環境です。	
SU116	CK, if you need any help regarding your norton program, please come back over here, we all will be happy to help.	CK サポートが必要な場合は、製品については、ここに戻ってきてください、私たちは喜んだろう。	CK が必要な場合は、製品に関するのに役立ちましたが、ここに戻ってきてください、私たちは喜んだろう。	
SS427	It will remove any Trojan Services and Registry Entries that it finds then prompt you to press any key to reboot.	トロイの木馬が削除されているサービスとレジストリエントリが見つかるように指示されます。次に、任意のキーを押します。	トロイの木馬が削除されているサービスとして検出されたレジストリエントリを要求するには、任意のキーを押します。	
MF197	also might just download program again from here and the install should pick up your license with no interaction needed from you.	でも、プログラムを再びダウンロードする可能性があり、インストールはここからライセンスを必要としていませんでした。	でも、プログラムを再びダウンロードする可能性があり、インストールはここからライセンスを必要としないからです。	
RF037	Currently, to deslect a folder, you have to manually and painstakingly deselect every file under that folder by browsing all the directory trees, this is not on!	現在、フォルダを手動で painstakingly、選択を解除する必要があり、そのフォルダの下にあるすべてのファイルを参照によって、すべてのディレクトリではない。	現在、フォルダ deslect するには、手動でインストールする必要があるすべてのファイルを選択解除 painstakingly 参照によってそのフォルダの下のすべてのディレクトリツリーではない。	
MF389	It might be useful to add a sticky FAQ at the top of each help forum with some standard posting suggestions as to what to include in any posts requesting help.	役に立つ場合があり粘り気を追加するにはよくある質問にサポートフォーラムヒントをいくつかの標準の投稿に含める内容についての投稿をサポートを要求しています。	追加すると便利でねっとりよくある質問に関するのに役立つヒントをいくつかの標準の投稿しないで何をして含めるのに役立ちました。	

Table 18: Extraction of the first six entries of the table handed to the human evaluators.

Nr	Translation Model Corpora	Pre-editing Rules	Language Model Corpora	Pre-editing Rules	BLEU
1	Symantec	✖	Symantec	✖	20.42
2	Symantec	✖	Symantec, Forum	✖	20.83
3	Symantec, kftt	✖	Symantec, Forum, kftt	✖	19.99
4	Symantec	✖	Symantec, Forum, kftt	✖	19.95
5	Symantec, kftt	✖	Symantec, Forum	✖	21.31
6	Symantec, kftt, Tatoeba	✖	Symantec, Forum, kftt, Tatoeba	✖	20.96
7	Symantec, Tatoeba	✖	Symantec, Forum, Tatoeba	✖	21.26
8	Symantec, Tatoeba	✖	Symantec, Forum	✖	21.84
9	Symantec, kftt, Tatoeba	✖	Symantec, Forum, Tatoeba	✖	22.10
10	Symantec, kftt, Tatoeba	✖	Symantec, Forum	✖	22.04
11	Symantec, kftt, Tatoeba	✖	Forum	✖	21.29
12	Symantec, kftt, Tatoeba	✖	Forum, Tatoeba	✖	20.88
13	Symantec, kftt, Tatoeba, SYN-forum	✖	Symantec, Forum, Tatoeba, SYN-forum	✖	20.89
14	Symantec, kftt, Tatoeba	✖	Symantec, Forum, Tatoeba, SYN-forum	✖	20.98
15	Symantec, kftt, Tatoeba, SYN-forum	✖	Symantec, Forum, Tatoeba	✖	20.52
16	Symantec, kftt, Tatoeba	✖	Symantec, Forum, Tatoeba	✔	22.69
17	Symantec, kftt, Tatoeba	✖	Symantec, Forum, Tatoeba	✔ (no split sentence)	22.58
18	Symantec, kftt, Tatoeba	✔	Symantec, Forum, Tatoeba	✔	21.98
19	Symantec, kftt, Tatoeba	✔ (no split sentence)	Symantec, Forum, Tatoeba	✔ (no split sentence)	21.28
20	Symantec, kftt, Tatoeba, SYN-forum (rules applied)	✖	Symantec, Forum, Tatoeba	✖	21.63
21	Symantec, kftt, Tatoeba, SYN-forum (rules applied)	✖	Symantec, Forum, Tatoeba	✔	22.20
ACCEPT	Symantec		Symantec		19.10
ACCEPT	Symantec		Symantec, Forum		19.81
ACCEPT	Symantec		Symantec, Forum	HIER	20.37

Table 19: Summary of all experiments conducted in this work with their respective BLEU score.

Table 19 contains all the different experiments conducted in this thesis including their respective BLEU score. In the first column the experiment number is given as a reference. The experiments 1 through 12 show the various experiments concerning the baseline system (chapter 4). The experiments 13 to 15 are part of the synthesized forum tests without the pre-editing rules and the remaining experiments (16 – 21) include the pre-editing rule (chapters 6 and 7). The highlighted experiments are the ones showing the best results in their respective chapter. The last three rows named “ACCEPT” show the results of SMT systems for English-Japanese translation in the ACCEPT project.

The columns “Translation Model Corpora” and “Language Model Corpora” contain the corpora used in each experiment in translation model training and language model training respectively.

The column named “Pre-editing Rules” following the “Translation Model Corpora” column shows whether the pre-editing rules were applied (✔) on the corpora used in training or not (✖). Exceptions are named in braces. An exception to this was made in experiments 20 and 21, where the pre-editing rules were applied only on the monolingual forum data before translating it to receive the synthesized bilingual corpus.

The second column named “Pre-editing Rules” expresses whether the pre-editing rules were applied to the test set or not. Again, exceptions are named in braces. In the last row, this columns entry expresses that in the ACCEPT project a hierarchical phrase-based system was trained.

The last column named “BLEU” contains the BLEU score of the experiment.

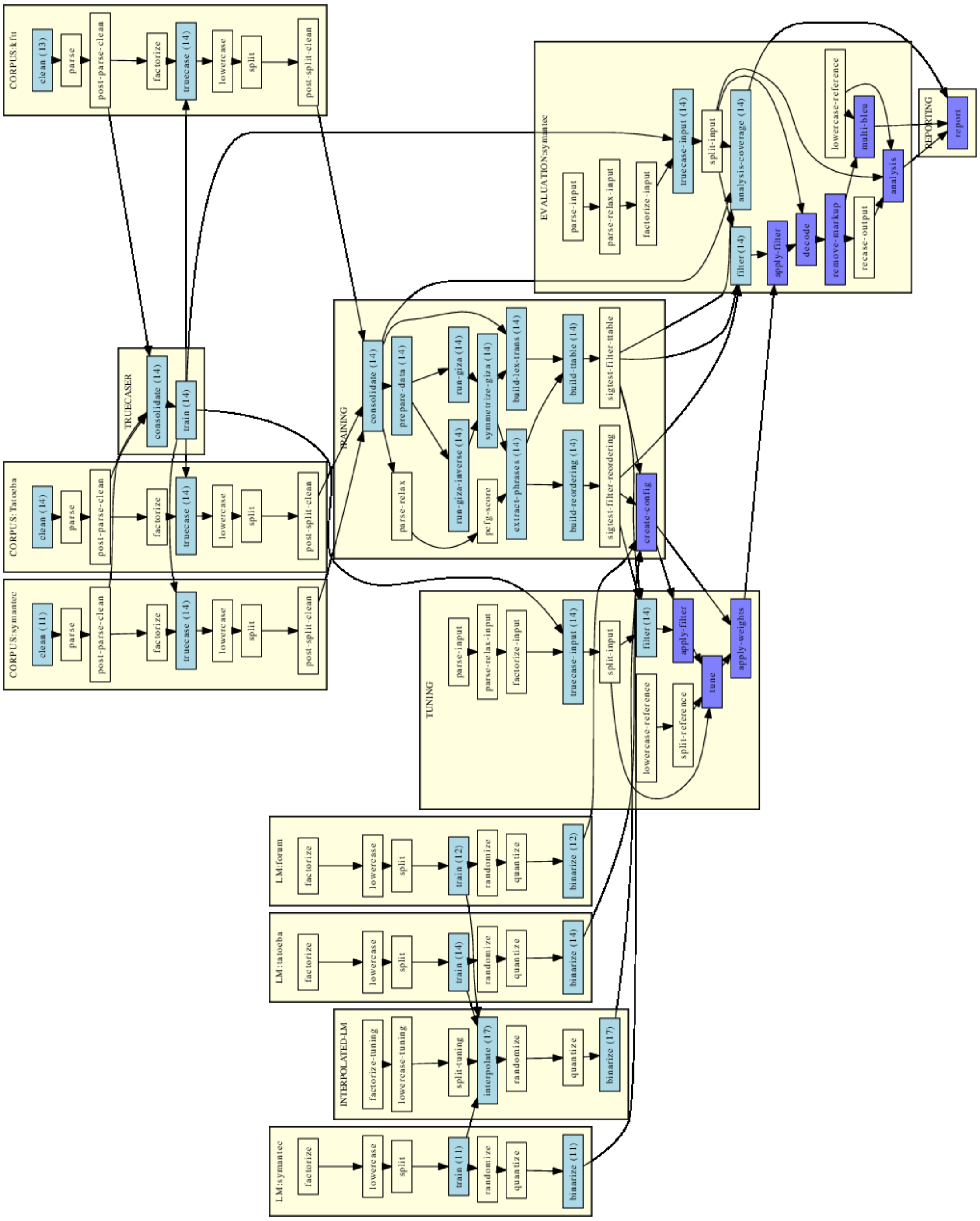

Figure 1: Graph for an experiment as displayed by the Moses EMS

Figure 1 shows a graph as it is displayed by the Moses EMS. The colored steps are the ones that are done in the experiment. Light blue means, that a step was reused from an earlier experiment. Dark blue means that this step was done in the experiment at hand.