Machine-Translating English Forum Posts to Japanese: On Pre-editing Rules as Part of Domain Adaptation

Torsten Jachmann
Saarland University
Saarbrücken, Germany
torsten@jachmann.de

Robert Grabowski　Mayo Kudo
Acrolinx GmbH
Berlin, Germany
{robert.grabowski,mayo.kudo}@acrolinx.com

January 21, 2014

Abstract

We increase the quality of statistical machine translation of English support forum posts to Japanese by exploring two approaches of domain adaptation: combining in-domain monolingual data with close-domain bilingual corpora; and developing pre-editing rules to reformulate phrases that are difficult for machine translation to Japanese. Automatic and human evaluations show that a combination of these measures improve the translations.

1 Introduction

In the modern Web, an increasing amount of useful information is created by community users. In a support forum, for example, users share their experiences and help each other out. A large amount of valuable information is added every day. To share it with users who do not speak the same language, machine translation is the only feasible approach.

Our aim is to use and improve a statistical machine translation (SMT) system to translate posts from the English Norton community forum to its Japanese counterpart. However, the type of content poses a number of challenges to SMT systems.

A typical problem is the lack of parallel training data that covers the same semantic domain as the text to be translated. The Norton support forum contains many contributions from subject-matter experts on Symantec products and software in general, for which only limited parallel data exists.

For user-generated content (UGC) such as forum posts, the lack of in-domain data also reaches to the language and style level. Forum posts usually contain colloquialism, spelling and grammar errors, ellipses, and many other aspects which are difficult to translate for typical SMT systems, as they are usually trained on error-free standard written texts. At the same time, corpora of colloquial language are rare to non-existent.

Another challenge is that English and Japanese share almost no similarities, neither in vocabulary, nor in grammar. They have different writing systems, different sentence structures, different ways to indicate modifications, inflections and grammatical cases, and so on.

In this paper, we respond to these challenges in two steps:

1. We combine close-domain and out-of-domain bilingual corpora with monolingual in-domain corpora to train a Moses SMT system. This method addresses the lack of bilingual in-domain corpora for the user-generated content we aim to translate.

2. For the Moses system, we develop pre-editing rules with the content optimization software, Acrolinx, to automatically reformulate patterns in the source text that constitute fundamental difficulties for the SMT translation to Japanese. This includes the removal of colloquial expressions and Internet slang, simplification of the sentence structure and adding "Japanese-like" phrases.

This paper proceeds as follows: in section 2, we describe related work. Sections 3 and 4 then explain the two steps above before we end with a summary.

2 Related Work

This paper summarizes the Bachelor thesis by Torsten Jachmann [2]. More detailed information on the presented results can be found in the thesis.

The work has emerged from the EU research project, ACCEPT [1], which develops new technologies in the fields of SMT systems, pre-editing, post-editing, and text analytics to support English, French, and German communities translate their content. The project partner, Symantec, provided us with the manual and forum data sets.

Set name	Contents	Proximity to UGC Domain	Languages	#segments
Forum	Symantec forum posts	in-domain	monolingual: JA	50K
Manuals	Symantec product manuals, localization strings	close-domain: similar topics	parallel: EN-JA	2M
Tatoeba	Full sentence dictionary containing colloquial speech	close-domain: similar language	parallel: EN-JA	200K
kftt	Wikipedia articles related to Kyoto	out-of-domain	parallel: EN-JA	2.5M
Dev/Test	manually translated support forum posts	in-domain	parallel: EN-JA	500/500

Table 1: List of used data sets and corpora

Domain adaptation has been proven to be a successful approach in SMT for domains, in which bilingual corpora are rarely existent. [4]

Previous research on pre-editing within the ACCEPT project [3] has not yet handled distant language pairs such as English-Japanese. In this paper, we, therefore, look into the applicability of pre-editing for such language pairs.

3 Baseline System

We use the statistical machine translation system Moses throughout our work. Moses combines a translation model that is trained on bilingual corpora with a language model that only needs monolingual target language corpora for training. The translation model is used to translate phrases, whereas the language model orders those phrases according to the sentence structure of the target language.

To train a Moses system for translating English Norton forum posts to Japanese, we faced the problem of the lack of bilingual in-domain data. We, therefore, used a combination of bilingual close-domain and out-of-domain data for training the translation model: a corpus of product manuals provided by Symantec within ACCEPT, the Tatoeba corpus[1] and the kftt corpus[2]. Additionally, forum posts extracted from the Japanese Symantec forum acted as an in-domain monolingual corpus for training the language model. The development and test sets each consist of 500 manually translated in-domain segments. Table 1 summarizes the data sets.

We used different combinations of these corpora for creating both a non-hierarchical translation model and a language model using the standard Moses training pipeline. The systems were tuned with the development set, and tested with the test set. The reported BLEU scores for the systems are summarized in table 2.

Within the ACCEPT project, a number of prototypical EN-JA systems had already been trained based on the Symantec data sets only. The best of these systems received a BLEU score of 20.37 and is shown for comparison in the first row of the table.

As expected, the choice of the corpora significantly determines the performance of the system. Out-of-domain corpora had a negative impact when used for the language model, but still helped the translation model. The monolingual in-domain corpus improved the performance of the language model. The close-domain sets had a positive impact on either model.

In total, the combination of the product manuals, kftt and Tatoeba corpora for the translation model and the product manuals, forum posts and Tatoeba corpora for the language model produced the best result (a BLEU score of 22.10). We took this system as the baseline system for the following domain adaption techniques. The baseline is about 1.73 points better than the prototypical ACCEPT system, which shows that a careful combination of close-domain and out-of-domain data can make up for the lack of in-domain bilingual data to a certain extent.

4 Pre-editing Rules

Forum posts and user-generated content differ from standard written text in terms of style and error rate, including special or wrong spellings, colloquialism, and others. Neither of them is likely to appear in the training data of typical SMT systems, as the data is typically based on standard written language. In

Translation Model	Language Model	BLEU
M	M, Forum	20.37
M	M	20.42
M	M, Forum	20.83
M, kftt	M, Forum, kftt	19.99
M, kftt	M, Forum	21.31
M, kftt, Tatoeba	M, Forum,Tatoeba	22.10

Table 2: Selected SMT systems with data sets used for training the models (M=Manuals)

[1] http://tatoeba.org/eng/

[2] http://www.phontron.com/kftt/

Rule	English sentence	Japanese translation
Sentence splitting	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.	ネットワークケーブルを削除してください。最初にアンインストールして再試行してください：//NSM です。 まず、ネットワークケーブルを削除してください。それからをアンインストールして再インストールしてください：//NSM です。
Add "please"	Before: Click on Add After: Please click on Add	[追加] をクリックします [追加] をクリックしてください。

Table 3: Examples for effects of pre-editing rules

order to achieve better translations, a major task is to bring the source text closer to the training corpora by removing differences and special occurrences that appear only in forum data. This can be achieved by pre-editing the texts before translating them.

To this end, we develop rules for the Acrolinx software to automatically correct and change the style of the source text without interfering with its meaning.

4.1 Acrolinx

Acrolinx is content optimization software that provides guidance while creating content. Among others, it features a customizable rule-based component that checks documents for a consistent writing style, and also includes a spell and grammar checker. The software marks found issues in the text, and often provides one or more replacement suggestions.

While Acrolinx is normally used in an interactive context that lets authors review the found issues and select a suggestion, we developed special style rules for English texts that always provide one suggestion, and which can therefore be applied automatically.

4.2 Investigated Rules

We developed the rules according to problematic patterns in English text that we manually found in test translations to Japanese. The rules aim to remove ambiguous language, adjust the domain of the input text to the training corpus and also shorten the gap between English and Japanese. Additionally, we took rules that had been created in the ACCEPT project for translations to German, and investigate whether they also improve translations to Japanese. In total, there were 16 pre-editing rules that implemented a large number of pre-editing patterns whose effect can be classified into 4 categories:

Reformulating slang and punctuation: "cuz" → "because", "guess so" → "I guess so", "&" → "and", "AFAIK" → "as far as I know", deletion of "LOL", and many others

Splitting sentences: long sentences are split at "then", "therefore", "so" and "but".

Shortening phrases & removing ambiguities: "going to"→"will", "not"→"un–" before adjectives, "might"→"may", "have to"→"must", "not. . . any"→"no", "help"→"support", etc.

Modifying English towards Japanese: add "please" to imperatives, delete discourse marker "well"

We conducted a small pre-evaluation with two annotators to filter out individual rules that have a clearly negative influence on the translations. Details can be found in the thesis [2]. In total, 13 out of 16 tested rules passed this pre-evaluation step; the 3 eliminated rules were re-used rules from ACCEPT.

4.3 Automatic Evaluation

To evaluate the overall effect of the pre-editing rules, we used the Acrolinx "autoApplyClient" tool. The tool automatically replaces each issue found by a rule with its (unique) suggestion. We also used the tool to auto-correct spelling and grammar issues. In case these check components provide multiple suggestions for an issue, the tool would choose the first one.

We automatically pre-edited the 500 sentences of the test set in this way. On some sentences, multiple editing operations were carried out. With the pre-edited test set, the BLEU score of the baseline system improves to 22.69, which is 0.59 points better than the evaluation with the unmodified test set.

4.4 Human Evaluation

The BLEU improvement shows that the coverage of the machine-translated sentences with respect to the reference translations increases when the input is pre-edited. However, as our rules sometimes slightly change the meaning or remove words, we additionally conducted a human evaluation with three Japanese native speakers to confirm that the pre-editing indeed preserves the adequacy of the translations.

Preferred version	All ratings		Averaged ratings		Averaged + filtered	
Original	92	31%	25	25%	13	21%
Pre-edited	127	42%	46	46%	39	62%
Similar	51	17%	8	8%	4	6%
Identical	30	10%	10	10%	0	0%
Disagree	-	-	11	11%	7	11%
Total	300	100%	100	100%	63	100%

Table 4: Results of human evaluation

We extracted and chose 100 sentences from the test set where at least one rule was applicable. This gave us 100 sentence pairs, consisting of the unmodified original sentence and the sentence in which all matching pre-editing rules had been applied. All sentences were then translated with the baseline system.

The judges were given a table where each row contained the original English sentence, its Japanese translation, and the Japanese translation of the pre-edited sentence. The two translations were randomly swapped to prevent a bias. The task was to compare the quality of both translations with respect to the original English sentence on a three-point scale (first translation better / both equally good or bad / second better). Additionally, judges could also indicate that two translations are entirely the same. All three judges rated the same 100 sentence pairs.

4.5 Results

We first examined the distribution of all 300 individual ratings (see "All ratings" in table 4). The results show that the pre-editing rules have an overall positive effect: they improve the translation in 42% of the rated instances, while they degraded the quality in only 31% of the cases. In the remaining cases, they had no impact on the quality (17%) or did not change the translation at all (10%).

In a second step, we investigated the inter-annotator agreement between the three judges. The agreement was rated "fair" (0.34) according to Fleiss' kappa rating scale. Although this number does not take the "rating distance" into account[3], we found the agreement good enough to combine all three ratings for each sentence into an average rating (see "Averaged ratings" in table 4). The result shows an even stronger tendency towards the sentences with the pre-editing rules applied.

On detailed inspection, we found that spell and grammar checking indeed degraded the translation quality, because they often provided multiple replacement suggestions. The autoApplyClient would always choose the first one, which was not necessarily correct or the best. Besides, the "might" → "may" reformulation usually had no impact. If we filter out the sentences with these reformulations, the distribution of averaged ratings clearly improves as illustrated in table 4 in the right-most column.

[3]E.g., a "first better"/"equal" disagreement has the same impact as a "first better"/"second better" disagreement.

5 Summary

We explored two approaches of domain adaptation for the domain of forum texts. First, we combined large out-of-domain corpora with small in-domain corpora to train a Moses SMT system. The approach is rather straight-forward for domains for which bilingual corpora are rare or even non-existent. Second, we developed and applied pre-editing rules in order to make the forum posts that are to be translated closer to the data used in the SMT training.

Both approaches show clear BLEU improvements. A human evaluation also showed that the developed pre-editing rules improve the translation in a majority of instances.

The evaluations also showed that simply applying the first suggestion for spelling and grammar issues with no human review can be harmful. Improving these components, for example by statistical re-ranking of suggestions with a language model, is part of ongoing work at Acrolinx.

In the thesis [2], a third approach has been explored: the use of a synthesized parallel corpus created by pre-editing a monolingual English in-domain corpus and translating it with the baseline system. Despite a degrading BLEU score, human evaluators rated this system as the best of the three approaches, which makes future research in this area promising.

References

[1] ACCEPT: Automated Community Content Editing PorTal. `www.accept-project.eu`.

[2] Torsten Jachmann. Exploring Domain Adaptation in Machine Translation for English to Japanese. Bachelor's thesis, Saarland University, 2013.

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

[4] Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. *22nd International Conference on Computational Linguistics (Coling)*, 2008.