3-step Parallel Corpus Cleaning using Monolingual Crowd Workers

Toshiaki Nakazawa Sadao Kurohashi
Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
nakazawa@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

Hayato Kobayashi Hiroki Ishikawa Manabu Sassano
Yahoo Japan Corporation
Midtown Tower, 9-7-1 Akasaka,
Minato-ku, Tokyo 107-6211, Japan
{hakobaya, hishikaw, msassano}@yahoo-corp.jp

***Abstract*—A high-quality parallel corpus needs to be manually created to achieve good machine translation for the domains which do not have enough existing resources. Although the quality of the corpus to some extent can be improved by asking the professional translators to translate, it is impossible to completely avoid making any mistakes. In this paper, we propose a framework for cleaning the existing professionally-translated parallel corpus in a quick and cheap way. The proposed method uses a 3-step crowdsourcing procedure to efficiently detect and edit the translation flaws, and also guarantees the reliability of the edits. The experiments using the fashion-domain e-commerce-site (EC-site) parallel corpus show the effectiveness of the proposed method for the parallel corpus cleaning.**

***Keywords*-parallel corpus cleaning; crowdsourcing; machine translation;**

I. Introduction

Bilingual sentence-aligned parallel corpora are essential language resources for corpus-based machine translation systems. The translation quality highly depends on the quality and quantity of the parallel corpora. Parallel sentences can be extracted from the Web [1] for general domain translations. For some domains such as patent documents [2] or parliamentary proceedings [3], parallel sentences can be extracted from the existing resources. Some methods of extracting parallel sentences [4], [5] or fragments [6] from comparable corpora have also been proposed. However, for the domains without existing language resources, researchers have to create parallel corpora manually.

Constructing a parallel corpus by hand is both time-consuming and expensive. A number of studies have been done recently in the direction of reducing the translation costs by post-editing the output of machine translation systems [7], [8], but this kind of framework may not work well when constructing a parallel corpora for a new domain. Another solution to reduce the translation cost is using crowdsourcing [9], [10]. Crowdsourcing workers basically are not professional translators, and some of them "cheat" on completing the task by using online translation services, which is why it is difficult to guarantee the translation quality. Some researchers have tried to predict the hidden reliability of translators and translations to choose the more appropriate translations [11], but still it is difficult to achieve the quality level of professional translators.

Another important issue, which is the main target of our study, is detecting and editing translation flaws in human-translated parallel corpora. Although we ask professionals to perform translation, the outcome occasionally contains translation flaws for various reasons. If the target text size is small, we can reduce the number of mistakes by making several reviewers check the translation. However, high-quality machine translation requires tens of thousands of parallel sentences to hundreds of thousands of parallel sentences; thus it is almost impossible to check the whole corpus. In this paper, we propose a framework to detect and edit the translation flaws contained in the existing manually-translated parallel corpus. The framework uses crowdsourcing in 3 steps: Step 1 detects the translation flaws, Step 2 edits the flaws and Step 3 validates the edits. By using crowdsourcing, corpus cleaning process can be done quicker and cheaper compared to professional cleaning. In addition, by dividing the cleaning into 3 steps, the quality of cleaning can be guaranteed.

The organization of the present paper is as follows: In Section II, we briefly describe the fashion-domain e-commerce-site (EC-site) parallel corpus which we use in our experiments. We explain the way of constructing this corpus, and the translation flaws it contains. Section III explains the proposed framework for the parallel corpus cleaning. Section IV and V show the experimental results of parallel corpus cleaning and translation, and Section VI summarizes this paper.

II. Fashion-domain EC-site Parallel Corpus

Yahoo! JAPAN was running an e-commerce site named "Yahoo! China Mall"[1] where customers could purchase Chinese items using Japanese interface. Originally, the descriptions of items were automatically translated into Japanese using a rule-based machine translation system. However the quality of translation was quite poor. We launched a joint project to improve the translation quality by changing the translation paradigm from rule-based to corpus-based. The

[1]Unfortunately, this service has been closed now.

Chinese-Japanese Fashion-Domain EC-site parallel corpus (we call it FDEC corpus) containing 1.2M sentences (6.3M Chinese words, 8.7M Japanese words) was created during the project.

The FDEC corpus was created by manual translation of the Chinese sentences from the fashion item pages. The pages are basically composed of 3 sections, *Title*, *Feature* and *Description*. Although longer sentences can be extracted from the Description section, the sentences in the Title and Feature sections are shorter or sometimes containing only one word. The parallel corpus construction of EC-site is different from that of novels [12] and newspapers [13]. In this section, we present some of the issues we have discovered so far and describe the know-how which we acquired during the corpus construction.

A. Translation Company Selection

The translation company should be carefully chosen because the quality of the machine translation highly relies on the quality of the parallel corpus. We first prepared trial sentences to check the translation quality of each company, and asked 3 different companies to translate the trial sentences. After considering the translation quality and price per unit, we have chosen two translation companies as contractors. Choosing multiple companies provides some flexibility in case of unexpected matters such as decrease of the translation quality or increase of the unit price. Moreover, we can acquire various translation choices because each company has its own characteristics in the translations, and also two companies can uses their own translation technology, which can balance the drawbacks of each company's translation.

B. Notes for Chinese-Japanese EC-site Translation

It is important to pay attention to the technical terms, ambiguities of words, and the difference of cultures to create a high quality parallel corpus in a specific domain. Below we describe some examples which we took care of during the corpus creation.

1) Domain-specific expressions: Some of the basic words have different meanings in a specific domain. For example, the Chinese word "不规则 (*disorder*)" is also used in Japanese "不規則 (*disorder*)". However, in fashion domain, it is used like " 不规则的下摆 (*wavy skirt*)". In this case, it is meaningless to translate it as "不規則 (*disorder*)", but it should be translated as "波打った (*wavy*)" or "フレアの (*flare*)".

Similarly, "木耳" originally means "wood ear" in both Chinese and Japanese. However it should be translated as "フリル (*frill*)" in Japanese.

2) EC-specific expressions: Some expressions appear in all EC-sites (not only those belonging to the fashion domain): for example, "秒殺 (*sold out in no time*)" or "淘金币 (*bargain sale*)". Many EC-sites have a seller-ranking system. In our case, there are ranking names "钻 (*diamond*)", "皇冠 (*silver crown*)", "金冠, 金皇冠 (*gold crown*)", and so on. It is important to take this fact into consideration in order to provide correct translation.

We also need to identify items which should not be translated (user IDs in the review posts, for example). These are proper nouns, which is why it is better not to translate them.

3) Cultural difference: Blatant expressions are more commonly used in Chinese, while euphemistic expressions are favorable in Japanese. This holds true for EC-sites. For example, "大きめサイズのレディース (*larger size ladies' wear*)" in Japanese is expressed as "胖女人 (*fat ladies*)". If Japanese female customers see the direct translation of the Chinese, they will displeased. Chinese descriptions often contain words like "我们 (*we*)" and "您 (*you*)". However the corresponding Japanese expressions are "当店 (*our shop*)" and "お客様 (*customers*)".

Some slang words are also used in the EC-site. For example, "MM" and "GG" in Chinese mean "girls" and "boys", respectively. These come from the Chinese pronunciation of "妹妹 (MeiMei/*girls*)" and "哥哥 (GeGe/*boys*)". These words had better to be translated properly to convey the intent correctly.

4) Unnatural or unsuitable compound nouns in Japanese: Chinese and Japanese share Chinese characters, and some of the Chinese compound nouns make sense in Japanese as they are. In the Chinese-to-Japanese translation, translators tend to preserve Chinese compound nouns as they are without consideration. However, in some cases, they are unnatural or unsuitable in Japanese. For example, the Chinese compound noun "特别 (*special*) 強調 (*emphasis*)" is understandable in Japanese but "注意事項 (*caution*)" is more natural. Another example is "着用 (*wear*) 効果 (*effect*) 图 (*figure*)": it means not "figure of effect to wear", but "picture of wearing".

5) Technical terms, proper nouns: Technical terms and proper nouns are often difficult to translate, which also holds for the case of EC-site translation. General item names such as "磨毛 (*fleece*)" and "风衣 (*windbreaker*)" have their corresponding Japanese translations, but some items such as "开裆裤 [2]" do not have corresponding translations in languages, other than Chinese. In addition, company names are often not translated into other languages. The rules for handling these kinds of words should be defined beforehand. In our project, only proper nouns that have corresponding Japanese expressions were translated into Japanese.

6) Extremely long Chinese sentences: Chinese sentences tend to be long because Chinese sub-sentences are often joined by commas. When performing Chinese-Japanese translation, it is better to divide translations of sub-sentences, if there is no strict relation between the sub-sentences. For example, the sentence in Figure 1 is easy to understand if it

[2]Pants for children without the inside of a thigh being sewn up

is translated after being divided into three sub-sentences at the || marks.

7) Repetitions: There are many fixed expressions repeatedly used in the EC-site such as sales copies, material names and so on. If we translate whole item page every time, we cannot increase the coverage of the parallel corpus because of the repetitions. It is necessary to carefully choose the sentences to be translated so as not to repeat the translation process for the sentences which have already been translated before.

C. Translation Specification for Parallel Corpus Construction

Parallel corpus construction for MT has certain specific requirements, which are different from those for usual publishing translation:

- Avoid liberal translations
  Liberal translations are hard to be correctly handled by the majority of the current MT systems. We requested the translators to translate obediently rather than finically.
- Prohibit omissions and additions
  Omissions and additions (adding explanations of some technical terms using parentheses, like this) decreases the machine translation quality.
- Stick to one-to-one sentence translation
  Most of the current MT systems assume that the sentences in the parallel corpus have one-to-one correspondences.
- Respect the sections
  We requested to the translators to pay attention to the characteristics of each section: the Title section should be translated as a noun phrase, and the Feature section should be translated as a sequence of nouns or numerals. For example, Chinese expression "到货 ！ " should be translated as "入荷！ (*Arrival!*)" in the Title section, and "入荷しました！ (*is arrived!*)" in the Description section.
- Divide the long sentences into appropriate units
  The background of this request is as follows:
  - The original Chinese sentences are automatically extracted from Web pages; thus they contain errors of sentence boundary detection.
  - Chinese sentences tend to be joined by commas, which results in generating very long sentences (see Section II-B6). However it is natural to divide them into smaller parts in other languages.

We documented the translation guidelines to correctly convey these requests along with the notes for Chinese-Japanese translation (Section II-B) to the translation companies. However, some mistakes are still present, even after using the guidelines' recommendations. Table I shows examples of translation flaws found in the sampling survey. In addition, there are some sentences forcibly translated as one sentence by joining with commas as in Figure 1. Translation companies have many translation workers and it is difficult to ask all the workers to thoroughly obey the guidelines; thus translation flaws are unavoidable.

To reduce the number of translation mistakes to the minimum and keep the quality of the parallel corpus high, we conducted sampling survey of the translations by Japanese-native observers who can understand Chinese. The low-quality translations and translation flaws were sent to the translation companies as feedback to improve translation in future. The translation companies have also sent feedback to us which points out the unclear or ambiguous parts of the guidelines. We can improve the guidelines by modifying the imperfections and augmenting it to handle new phenomena.

However, this kind of solution cannot modify the sentences which have been already translated. Taking into consideration the high costs, it would have been unwise to send the completed translation to the companies back for additional post-editing. Therefore, we propose using crowdsourcing to clean the existing parallel corpus in a comparatively quick and cheap way.

III. Parallel Corpus Cleaning using Crowdsourcing

Although the percentage of the sentences which include translation mistakes is small, it is difficult to automatically detect them. We need to check the whole corpus in order to correct all the translation flaws, which is quite expensive.

To solve this problem, we propose a framework of cleaning an existing corpus efficiently and cheaply using crowdsourcing. The framework is composed of 3 steps:

1) Fluency Judgement
2) Edit of Unnatural Sentences
3) Verification of Edits

In the crowdsourcing, any number of workers paticipate the task, and each worker completes the very small part of it. Each step is basically conducted by the monolingual workers of the target language (in our case, Japanese workers). The number of monolingual workers is much greater than that of bilingual workers; thus the tasks can be done efficiently. This framework mainly aims at correcting the unnatural sentences as in Table I. In the following sections, each step is explained in detail.

A. Step 1: Fluency Judgement

The first step detects the translation flaws by asking the crowd workers to judge if the sentences are natural and grammatically correct. This task is done by only showing the translated sentences. Some technical words and proper nouns remain in the translated sentences as they are in the source sentences, and the workers may judge them as unnatural. The workers are instructed to ignore such special words.

Zh: 上面的刺绣和亮片均为原厂工人原厂设备精心缝制, || 挑剔的姐妹们在看到货品之后会发现绝对可以和专柜货品比肩, 而且绣工精细清晰, || 精棉质地, 密度高, 手感好, 穿着舒适, 质量超好。

Ja: 上の刺繍とスパンコールは、全てオリジナル工場の作業員とオリジナル工場の設備で心を込めて作成したものです、|| あら捜しをするお客様も、この商品を見れば、専門店の商品に匹敵するほど、作りが精細で、はっきりしたものだと思われるはずです、|| 精綿生地で、密度が高く、手触りも良く、着用すると快適で、品質もとても良いです。

En: *The embroidery and spangles above are all made with care by original factory workers with our original factory equipment,* || *even the most demanding customers should rank this product with one from a specialist shop and consider it finely and exactly made,* || *it is made of fine cotton, is dense, has a good feel, is comfortable to wear and the quality is very high.*

Figure 1. Very long Chinese sentence joined by commas, and its Japanese translation provided by the translation company (it is natural to divide the sentence at || marks).

Table I
EXAMPLES OF TRANSLATION FLAWS.

Category	**Input Chinese**	**Translation**	**Reference**
Omission	看看有没有其他合适的商品	**看看有没有**その他合適的商品	他に良いものがないかご覧ください
Mistranslation	加湿器功能:	除湿器の機能： (*functions of dehumidifier:*)	加湿器の機能: (*functions of humidifier:*)
Mistranslation	买家秀身上穿的是两件, 一口价是一件的价格!	お客様ショーの体に着ているのは 2 点、 ワンプライスは一枚の値段です！	モデルが着ているものは２着で、 価格は 1 着の値段です！
Insertion	不要随便拍下一种	随意にに 1 種類だけ注文するのではなく	随意に 1 種類だけ注文するのではなく
Chinese Character	精神焕发之效果	元**气**あふれるという効果があります	元気があふれるという効果があります
Unnatural	在清洁保养时应切断电源, 拔下插头防止意外事故发生。	お手入れの時、電源を切れ、 プラグを抜いてください。	お手入れの時は、電源を切り、 プラグを抜いてください。

This is a choice-based task. If we ask two or more workers to answer the same task, we can increase the reliability of the judgement by putting all decisions together.

B. Step 2: Edit of Unnatural Sentences

In the second step, the workers are asked to edit the translated sentences. This task is also done by only showing the translated sentences. However it is possible to show the source sentence as well for the reference[3]. The bilingual workers, if they are available, would edit the translations more precisely with the reference source sentence, and monolingual workers just ignore them.

This is a free writing task. If we ask two or more workers to answer the same task, we can acquire a variety of edits. Different from the studies to create a parallel corpus using crowdsourcing (see Section I), this task is just editing, not translating.

C. Step 3: Verification of Edits

In the last step, each edit made by each worker is validated by asking the workers to judge if the edited translation is better than the original one. This step is important to further improve the quality of the outcome because the edits are not necessarily correct.

This is a choice-based task; thus we can increase the reliability of the judgement by asking two or more workers to answer the same task.

IV. CORPUS CLEANING EXPERIMENTS

To evaluate the effectiveness of the proposed framework, we conducted corpus cleaning experiments using the FDEC corpus introduced in Section II. We used Yahoo! Crowdsourcing[4] as the crowdsourcing service. We can carry out several styles of crowdsourcing tasks such as Yes/No questions and free writings with this service. In the following sections, we explain the experimental settings and discuss the results. The service is run in Japan; therefore most of the workers are Japanese. In addition we cannot select the workers by their abilities, and the workers who participated in our experiments do not necessarily understand Chinese (perhaps almost all of them does not).

A. Step 1

We used 358,085 sentences from the FDEC corpus with length between 10 and 130 characters excluding numerals, Roman characters, symbols and white spaces. We asked 5 different workers to answer the same question. Table II shows the results. 108,340 sentences (30.2%) are flawed translations if we set the threshold of the flawed translation at 3 or more, and 48,104 sentences (13.4%) are flawed if we set the threshold at 4 or more. Below are examples of the results.

- 5 workers judged as unnatural
  お支払終了後、値切ことは承りません。
- 4 workers judged as unnatural
  もし同僚やガードマンが代印されるなら、事前に確認作業をされて下さい。
- 3 workers judged as unnatural
  商品を受け取ったら、すぐ評価をご確認ください！ ！
- 2 workers judged as unnatural
  2010 年 3 月、春はぽかぽかと花が満開になる季節。
- 1 worker judged as unnatural
  当店ではすべての商品の実物写真をご用意しております。

[3]In our experiments, we showed both source and translated sentences.

[4]http://crowdsourcing.yahoo.co.jp

Table II
EXPERIMENTAL RESULT OF FLUENCY JUDGEMENT.

# unnatural judegement	# sentences	percentage
5	13,056	(3.6%)
4	35,048	(9.8%)
3	60,200	(16.8%)
2	83,150	(23.2%)
1	93,187	(26.0%)
0	73,444	(20.5%)

- 0 worker judged as unnatural
  最後には、具体的な状態で検討すべきです。

We asked Japanese native speakers to check the results and confirmed that the results are reasonable. It is surprising that the parallel corpus is constructed manually and yet contains 30% incorrect translations. One reason for this is that most of the sentences are translated by native Chinese speakers, not Japanese speakers. It is often said that translations should be done by native speakers of the target language. However, native speakers of the source language are very knowledgeable about the source sentences including culture and background, and this is an advantage for correctly translating the input sentences.

B. Step 2

From the results of Step 1, we used 47,420 sentences which were judged as unnatural by 4 or more workers[5] in Step 2. We asked 3 different workers to edit the translations. The workers can skip the task if they think that the sentences do not need to be edited. The original Chinese sentences are also shown to the workers. However the workers do not necessarily understand the Chinese.

The results are shown in Table III. 34,542 sentences (72.8%) are edited and a total number of 54,550 edits are acquired. The following are examples of edits.

- edited by 3 workers
  Original: 100%適するとは言えないので最終的な決めるのはご自身になります。
  Edit1: 100%適するとは断言できませんので最終的に決めるのはご自身になります。
  Edit2: 100%適するとは言えないので最終的に決めるのはご自身となります。
  Edit3: 100%適してるとは言えないので最終的に決めるのはご自身になります。
- edited by 2 workers
  Original: 100%実物写真、実際の物品は絶対にいっそうきらめいて、更に心や目を楽しませます。
  Edit1: 100%実物写真です、実際の物品はよりいっそうきらめいて、更に心や目を楽しませます。
  Edit2: 100%実物写真です、実際の物品は絶対にいっそうきらめいて、更に心や目を楽しませます。
- edited by 1 worker
  Original: お支払終了後、値切ことは承りません。
  Edit1: お支払終了後、値切ることは承りません。

[5]We excluded some sentences which are garbled.

Table III
STATISTICS OF THE EDITS OF UNNATURAL SENTENCES.

# workers edited	# sentences	percentage
3	3,755	(7.9%)
2	12,498	(26.4%)
1	18,289	(38.6%)
0	12,878	(27.2%)

We asked Japanese native speakers to check the edits and confirmed that the edits are reasonable and correct.

C. Step 3

The quality of a total number of 54,550 edits were verified. The workers were asked to judge which of the original and edited translations is more natural. The original Chinese sentences were also shown along with the two translations. We asked 5 different workers to answer the same question.

Table IV shows the results of the validation looking at each edit independently. 49,237 edits (90.3%) were judged to be better than the original translations by the majority of the workers, which is much greater number than the other. This result clearly shows that the proposed parallel corpus cleaning framework works well. Looking at the result by the original sentence, 32,244 sentences (93.3%) among 34,542 edited sentences have one or more better edits. The following are examples of the validations.

- 5 workers judged the edit is more natural
  Original: お支払終了後、値切ことは承りません。
  Edited: お支払終了後、値切ることは承りません。
- 4 workers judged the edit is more natural
  Original: 定期な得意先への連絡と潜在な忠実な取引先の掘り起こし
  Edited: 定期的な得意先への連絡と潜在的に忠実な取引先の掘り起こし
- 3 workers judged the edit is more natural
  Original: 10 元追加すると、ノートブックの放熱台座を差し上げます。
  Edited: 10 元追加すると、ノートブックの放熱パッドを差し上げます。
- 2 workers judged the edit is more natural
  Original: 100%のゼロリスク、頑張ってくださいね
  Edited: 100、リスクなし。頑張ってくださいね。
- 1 worker judged the edit is more natural
  Original: 24K ゴールドの新鮮なバラで、永遠にしおれないバラ
  Edited: 24K ゴールドは新鮮なバラで、永遠にしおれないバラ
- 0 worker judged the edit is more natural
  Original: 写真の説明通りでした（少し異臭がしますが、理解できます）
  Edited: 写真の説明通りでした（少しがしますが、理解できます）

Although the edited sentences are natural as Japanese sentences, they might be incorrect as translations. We reviewed the 100 edits randomly sampled from the ones which are judged to be more natural than the original sentence by 5

Table IV
VALIDATION RESULTS OF EACH EDIT.

# judged better	# sentnece	percentage
5	25,053	(45.9%)
4	16,478	(30.2%)
3	7,706	(14.1%)
2	3,338	(6.1%)
1	1,462	(2.7%)
0	513	(0.9%)

workers. We found three types of inequalities: 1) deletion of symbols, 2) omission and 3) mistranslation, and the number of each inequality was 8, 13 and 5 respectively. The following are examples of the inequalities.

1) deletion of symbols
   Chinese: 亲们拍下后联系客服修改价格就好呢 ～～～
   Original: お客様にはご購入後にカスタマーサービスオペレーターに連絡し価格を訂正してください ～～～
   Edited: お客様はご購入後、カスタマーサービスオペレーターへ連絡し価格を訂正してください
2) omission
   Chinese: 但是实际颜色稍微深点，衣服素雅大方
   Original: 実際の色はちょっと深くて、衣裳もさっぱりとしていて大方です。
   Edited: 実際の色はちょっと深くて、衣裳もすっきりとしています。
3) mistranslation
   Chinese: 引用一位资深**黄钻**买家买这款衣服时对我说的话：
   Original: あるかなり経歴のあるイエローダイヤモンドのお客様がその商品を買った話によると：
   Edited: 歴史あるイエローダイアモンドを買ったお客様の話によると：

In the first example, the symbols at the end of the sentence are removed. This effect can be avoided by correctly instructing the workers to keep the symbols. In the second example, the Chinese word "大方" is omitted. Actually this is a very complicated problem. The Chinese word "大方" has several meanings such as *generous*, *liberal* and *stylish*. There is the same word in Japanese, but it means *almost* or *nearly* which is completely different from the Chinese meanings. The professional translators left the word in the Japanese sentence. However it is completely unnatural, and the crowd workers removed it.

In the third example, "黄钻 (*yellow diamond*)" is the name of a rank in the rating system of the EC-site. However, the crowd workers thought it as the real diamond, and edited the sentence incorrectly. The second and third effects are difficult to prevent, and this is left as future work.

D. Crowdsourcing Cost

In our experiments, Step 1 costs 2 million Japanese Yen (JPY), Step 2 costs 310 thousand JPY and Step 3 costs 280 thousand JPY, in total 2.6 million JPY. Of course the fee varies depending on the number of workers for each question (this time 5, 3 and 5 workers respectively). We cannot directly compare with professional editing, but one editing company costs at least 6 JPY per English word[6]. If we apply this rate to our Chinese-to-Jpanaese translation editing, all the sentences containing 6.8M words costs about 40 million JPY, which is 15 times larger than using crowdsourcing.

As for the editing time, Step 1 took 115 hours, Step 2 took 35 hours and Step 3 took 36 hours, in total 186 hours. Note that this is not the sum of the active working time of all the workers, but the time from when we submit the task until we get the results. The professional edits 4000 words per day; thus it takes 1700 days to edit all the sentences. Using crowdsourcing, we can greatly reduce both the time and cost.

V. TRANSLATION EXPERIMENT

To evaluate the crowdsourcing cleaning extrinsically, we also conducted a translation experiment. We used the original FDEC corpus as the baseline and divided it into training, development and test sets. Then, part of the Japanese sentences were replaced by the edits which were judged to be reasonable by the majority of the workers in Step 3. For the sentences which have more than one edits, we duplicated the sentences to use all the edits (cleaned 1) or randomly chose one (cleaned 2) for only development and test sets. We did not use cleaned 2 for the training data because bigger training data basically makes the translation quality better. Table V shows the statistics of the corpus.

We used a dependency tree based alignment model [14] for word alignment and KyotoEBMT system [15] for decoding with the default settings and evaluated the translation quality by BLEU [16] score. The results are shown in Table VI. The baseline (setting 1) score was 21.39 and it was improved by 0.3 points BLEU score in setting 2 where only the training data is cleaned. The p-value calculated by the bootstrap resampling [17] was 0.052. From this result we conclude that the proposed framework actually cleans the parallel corpus, and it contributes to improve the translation quality.

In other settings where one or both of the development and test data sets were cleaned, the BLEU scores slightly decreased. We think this is due to the inequalities between the original Chinese and the edited Japanese (See Section IV-C). The effect of the inequalities in the training data can be moderated during word alignment by handling them as NULL aligned words. However those in the development and test data are not negligible because all the automatic evaluation scores suppose the content of the input and output are strictly equal.

VI. CONCLUSION

This paper proposed a framework of cleaning existing corpora efficiently and cheaply using crowdsourcing. The

[6]http://www.editage.com

Table V
THE NUMBER OF SENTENCES FOR THE TRANSLATION EXPERIMENTS.

	original (OR)	cleaned 1 (CL1)	cleaned 2 (CL2)
train	1,220,597	1,256,908	-
dev	11,186	11,489	11,186
test	11,200	11,495	11,200

Table VI
EXPERIMENTAL RESULTS.

setting	1(base)	2	3	4	5	6
train	OR	CL1	CL1	CL1	CL1	CL1
dev	OR	OR	CL1	CL1	CL2	CL2
test	OR	OR	OR	CL1	OR	CL2
BLEU	21.39	**21.69**	21.34	21.12	21.37	21.09

framework is composed of 3 steps and is able to clean existing parallel corpora containing noise reliably. The experimental results show the effectiveness of the proposed method.

As stated in Section IV-C, there still remain translation flaws which are not easy to prevent and correct, and solving this problem is future work. One possible solution is to ask the workers to give confidence scores of their edits. By only passing the edits with low confidence to the professional checkers, we might clean the corpus more reliably while keeping the cost low.

Another remained issue is that this framework can improve the translation fluency, but not able to improve the translation accuracy. We need to come up with a new idea to effectively improve the translation accuracy of the existing parallel corpora.

ACKNOWLEDGMENTS

This work is supported by the Yahoo Japan Corporation. We want to thank the anonymous reviewers for many very useful comments.

REFERENCES

[1] J. Uszkoreit, J. Ponte, A. Popat, and M. Dubiner, "Large scale parallel document mining for machine translation," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 1101–1109.

[2] M. Utiyama and H. Isahara, "A Japanese-English patent parallel corpus," in *MT summit XI*, 2007, pp. 475–482.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, 2005, pp. 79–86.

[4] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 403–411.

[5] C. Chu, T. Nakazawa, and S. Kurohashi, "Chinese-Japanese parallel sentence extraction from quasi-comparable corpora," in *Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC 2013)*, 2013, pp. 34–42.

[6] C. Chu, T. Nakazawa, and S. Kurohashi, "Accurate parallel fragment extraction from quasi-comparable corpora using alignment model and translation lexicon," in *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, 2013, pp. 1144–1150.

[7] N. Aranberri, G. Labaka, A. D. de Ilarraza, and K. Sarasola, "Comparison of post-editing productivity between professional translators and lay users," in *Proceedings of the Third Workshop on Post-editing Technology and Practice*, 2014, pp. 20–33.

[8] L. Schwartz, "Monolingual post-editing by a domain expert is highly effective for translation triage," in *Proceedings of the Third Workshop on Post-editing Technology and Practice*, 2014, pp. 34–44.

[9] V. Ambati and S. Vogel, "Can crowds build parallel corpora for machine translation systems?" in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 62–65.

[10] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[11] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 1220–1229.

[12] D. Cao, H. Nakano, Y. Xu, and H. Kumai, "Development of "Chinese-Japanese bilingual corpus" and its remaining tasks," *IPSJ SIG Notes*, vol. 99, no. 95, pp. 1–8, nov 1999.

[13] Y. Zhang, K. Uchimoto, Q. Ma, and H. Isahara, "Building an annotated Japanese-Chinese parallel corpus - a part of NICT multilingual corpora," in *Proceedings of 2nd International Joint Conference on Natural Language Processing*, 2005, pp. 85–90.

[14] T. Nakazawa and S. Kurohashi, "Alignment by bilingual generation and monolingual derivation," in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 1963–1978. [Online]. Available: http://www.aclweb.org/anthology/C12-1120

[15] J. Richardson, F. Cromières, T. Nakazawa, and S. Kurohashi, "KyotoEBMT: An example-based dependency-to-dependency translation framework," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 79–84.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation." in *ACL*, 2002, pp. 311–318.

[17] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP 2004*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 388–395.