Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems

Rei Miyata[†] **Atsushi Fujita**[‡]
[†]Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
miyata@nuee.nagoya-u.ac.jp
[‡]National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
atsushi.fujita@nict.go.jp

Abstract

Machine translation (MT) systems are not able to always produce translations of human-level quality. As a practical means of such MT systems, we investigated the potential of pre-editing strategy, by collecting actual pre-edit instances using a human-in-the-loop protocol. In our study, targeting Japanese-to-English translation on four different datasets and using an off-the-shelf MT system, we collected a total of 12,687 pre-edit instances for 400 source sentences and showed promising results; more than 85% of source sentences turned out to be accurately translated by the MT system. We also found that the pre-edited Japanese source sentences were better translated into Chinese and Korean, confirming the usefulness of pre-editing strategy in a multilingual setting. Through decomposing the collected pre-edit instances, we built a typology of primitive edit operations comprising 53 types, which unveils the subjects for further research.

1 Introduction

Given the improved quality of machine translation (MT) and the increased demand for rapid delivery of translations, a number of off-the-shelf MT systems have become available. However, none of them can guarantee that their raw outputs are always of sufficient quality. When we consider embedding such MT systems in computer-aided translation (CAT) settings, it is indispensable to explore practical means to obtain high-quality translations without configuring the MT systems.

One option to make better use of such MT systems is to edit source text (ST) so that it is amenable to the targeted MT system, i.e., *pre-editing*. As demonstrated in the literature, pre-editing ST leads to improved MT quality (Bernth and Gdaniec, 2001; Miyata et al., 2015) and reduced post-editing effort (Pym, 1988; O'Brien and Roturier, 2007; Aikawa et al., 2007). Controlling ST is particularly effective in multilingual settings (Ó Broin, 2009).

Several studies have examined human-in-the-loop protocols that include pre-editing ST in order to improve MT quality. Uchimoto et al. (2006) have used back translation as a means to spot non-machine-translatable spans in ST, which are subsequently served to humans to be edited. Resnik et al. (2010) have taken advantage of monolingual human knowledge of the target language to identify spans of ST that are likely to cause translation errors. Mirkin et al. (2013) have devised an interactive tool for monolingual authors. It suggests appropriate alternatives along with confidence scores for MT outputs.

In this paper, we investigate the capability of the pre-editing strategy and provide an overview of possible edit operations used for pre-editing. First, we empirically demonstrate the potential usefulness of the pre-editing strategy, i.e., how often STs turn out to be accurately translated by a targeted MT system. To this end, we designed a human-in-the-loop protocol, in which human editors incrementally edit given STs (Section 2), and experimented with Japanese-to-English translation tasks on four different datasets (Section 3). Using the original and the best-edited STs, we also

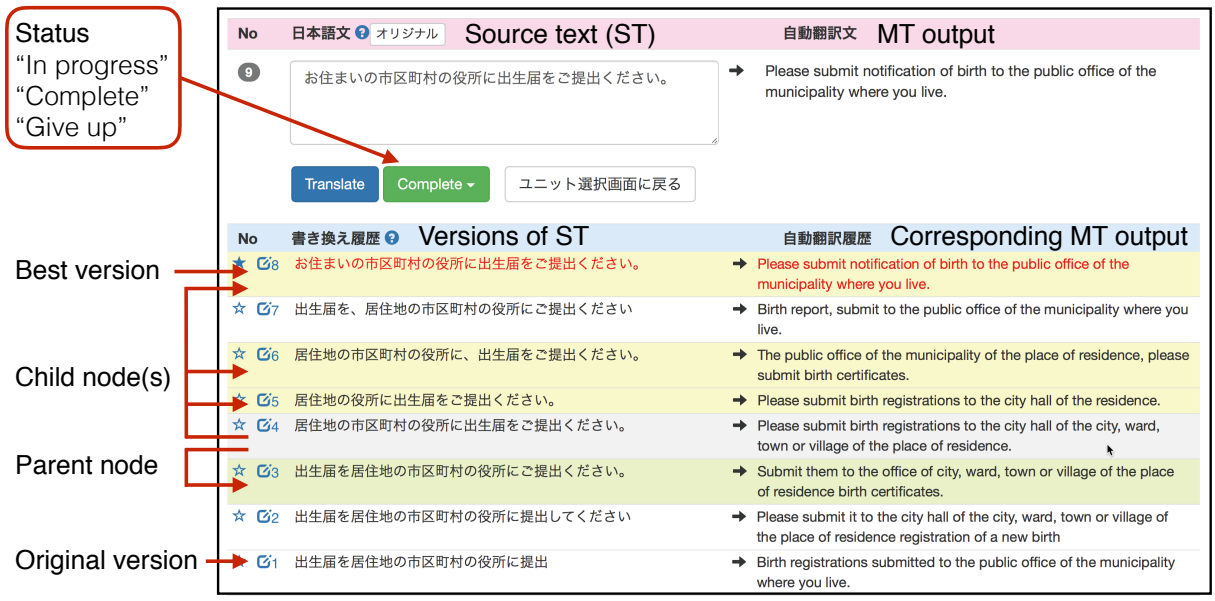

Figure 1: Our platform for collecting pre-edit instances: when an edited ST in the upper pane is submitted, it is registered with its MT output in a sequential order as shown in the bottom pane.

investigated the usefulness of the pre-edited STs in translating Japanese STs into Chinese and Korean (Section 4). To give an overall picture of pre-editing, we built a typology of edit operations upon actual *pre-edit instances*, i.e., pairs of STs before/after minimal pre-editing, collected through the above protocol followed by manual decomposition (Section 5). The typology can act as a guidepost to determine useful operations, such as those having the largest impact on the MT quality and those that are easy to automate.

2 Protocol for Collecting Pre-editing Instances

As in Miyata et al. (2015), we ask human editors to incrementally edit STs relying on their introspection, so that improved MT quality is achieved. Miyata et al. (2015) collected only the final versions of edited STs and directly compared them with the originals. In contrast, we aim to observe the trials and errors of editors and to achieve translations of satisfactory quality as much as possible. To that end, we developed a Web-based platform, shown in Figure 1, with the following two features.

- We record ST after every minimal edit is performed in order to capture the detailed process of pre-editing.
- We allow editors to resume editing from any given past version of ST in order to facilitate their trial and error.

Editors are asked to follow the iterative procedure given below for each original ST. We refer to the set of collected versions of STs for the same original ST as a *unit*.

Step 1. Assess the MT output for the present ST according to the 5-point scale criterion in Table 1. Go to Step 4, if it has satisfactory quality;[1] otherwise, go to Step 2.

Step 2. Select one version of ST to be edited, from the past versions of STs, referring to the corresponding MT outputs, and go to Step 3; if none is likely to achieve satisfactory quality even if edited, go to Step 4.

Step 3. Minimally edit the selected version of ST, while keeping the meaning of the ST, referring to the MT output for it, so that the MT system would be able to generate a better translation. When the edited ST is submitted, its MT output is automatically generated and registered together. Go back to Step 1.

Step 4. Choose one version of ST that achieves the best MT quality among all the versions registered in the unit (called the *Best ST*), and terminate the procedure for this ST.

To observe fine-grained edit operations, we instructed editors to make edits primitive as much as possible in Step 3, showing some examples. Table 2 shows an example; a phrase reordering for sentence (a) makes sentence (b), and a passiviza-

[1] “Perfect” or “Good” quality in our criterion in Table 1.

5. Perfect	Information of the original text has been completely translated. There are no grammatical errors in the translation. Word choice and phrasing is natural even from a native speaker's point-of-view.
4. Good	Word choice and phrasing is slightly unnatural, but the information of the original text has been completely translated and there are no grammatical errors in the translation.
3. Fair	There are some minor errors in the translation of less important information of the original text, but the meaning of the original text can be easily understood.
2. Acceptable	Important parts of the original text are omitted or incorrectly translated, but the core meaning of the original text can still be understood with some efforts.
1. Incorrect/nonsense	The meaning of the original text is incomprehensible.

Table 1: Criterion for evaluating MT quality.

(a) 来院しなくても十日前後で登録のクレジットカードから引き落としを行います。
(b) 来院しなくても登録のクレジットカードから十日前後で引き落としを行います。
(c) 来院しなくても登録のクレジットカードから十日前後で引き落としが行われます。

Table 2: Examples of primitive edits on a Japanese sentence whose meaning is "You've registered your credit card. We will charge on that card in around 10 days regardless of your visit."

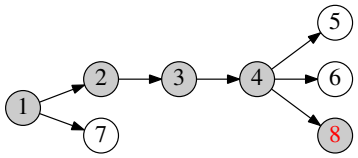

Figure 2: Tree representation of versions of STs shown in Figure 1.

tion of sentence (b) leads to sentence (c). In this case, the edit from sentence (a) to sentence (c) is not considered as primitive.

Also in Step 3, we prohibited editors from registering an ST identical to any past versions of ST. With this constraint, versions of ST in each unit form a tree structure, as illustrated in Figure 2. Each node comprises a version of ST accompanied by the MT output for it; the number in the node stands for the chronological order of the version in each unit, with one node, No. 8 in this example, labeled the Best ST. Every node, except for the original one (No. 1), is derived from a parent node. It is guaranteed that the path between the Best ST and the original one (henceforth, *best path*) in each unit, e.g., gray nodes in Figure 2, contains edit operations effective in improving MT quality.

3 Pilot Run

Using our protocol presented in Section 2, we collected versions of STs and pre-edit instances in Japanese-to-English translation of four sets of STs in three domains: hospital conversation[2] (*hosp*), living information provided by municipalities[3] (*muni*), and two types of news articles, Japanese-origin ones from BCCWJ[4] (*bccwj*) and English-origin ones from Reuters[5] (*reuters*). While *hosp* is spoken, the others are written; sentence length is markedly diverse (see also Table 3). These domains are so different from each other that we expect that the applicability of our proposed protocol can be evaluated from diverse points of view. For each dataset, we randomly sampled 100 Japanese sentences and used them as original STs.

As the off-the-shelf MT system, we used TexTra,[6] a freely-available, state-of-the-art phrase-based statistical MT system, through its REST API. We assigned the pre-editing task to one native Japanese speaker who has a good command of English and ample experience in evaluating the quality of various types of MT systems according to the criterion in Table 1, while she has no prior knowledge of TexTra.

As a result, 13,087 versions of STs and thus 12,687 pre-edit instances were collected; see Table 3 for statistics. As shown in the rightmost column, more than 85% of the STs were ended with MT outputs of satisfactory quality. This demonstrates the high potential of the MT system when proper human intervention is incorporated. In general, the longer the original ST was, the more edit operations were required to attain satisfactory quality. Table 4 shows an example of the Best ST of a unit in *reuters*, which was obtained after 25 consecutive edits in the best path and the MT output of which met satisfactory quality.

[2]An in-house speech transcription corpus of conversational utterances in a hospital.

[3]Excerpts from websites of municipalities in Japan (Miyata et al., 2015).

[4]http://pj.ninjal.ac.jp/corpus_center/bccwj/

[5]http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/reuters/

[6]https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

Dataset	Mode	Avg. num. of tokens in original ST (s.d.)	Num. of pre-edit instances				Num. of units	
			Total	Avg.	Med.	Max	Original=Best	Satisfactory quality
hosp	spoken	12.1 (4.5)	1199	12.0	3	105	40	97
muni	written	21.3 (12.0)	2119	21.2	14	89	3	97
bccwj	written	26.9 (16.0)	3823	38.2	26	209	0	86
reuters	written	34.8 (12.6)	5546	55.5	45	258	4	93

Table 3: Statistics of the collected data.

	ST	MT output
Original	同国は、前年の過剰輸出と、今年の減産によって、穀物不足に直面しており、大量の小麦輸入の計画を表明している。	Excess exports in the previous year, and reduced production this year, is facing a shortage of grain, a large amount of wheat imports plan.
Best	当年の減産と前年の過剰輸出による穀物の不足をふまえ、この国は小麦を大量に輸入する計画を表明している。	Based on the shortage of grain due to production cuts in the current year and excessive exports last year, this country has announced plans to import a large amount of wheat.
Reference	The country, currently battling an acute grain shortage due to excessive exports last year, faces a poor harvest this year and intends to import large quantities of wheat.	

Table 4: An example of Best ST with satisfactory MT quality.

	ST	MT output
Original	ＷＳＣによると、４日には弱い複数の降雨の可能性があるものの、５–６日には全般に乾燥した天候が戻る見通し。	WSC, although the possibility of weak more rainfall within 4 days, the weather in general dry return to 5-6 days.
Best	ＷＳＣによると、４日には弱い降雨の可能性が存在する一方で、５日から６日にかけては、乾燥した天候が全般に戻ってくる見込み、とのこと。	WSC said, while the possibility of a weak rain exists on **June** 4, from **June** 5 to 6, the dry weather comes back, in general.
Reference	WSC said the outlook was for a chance of a few light showers on 4th, and generally dry conditions on 5th and 6th.	

Table 5: An example of Best ST for which our protocol cannot achieve satisfactory MT quality.

It should also be noted that 27 out of 400 units did not attain satisfactory quality in our human-in-the-loop protocol. Among these "Give up" cases, we identified that mis-translation of proper nouns and incorrect lexical choices were the most difficult types of MT errors to rectify. For example, the Best ST in Table 5 contains expressions for dates, "4 日," "5 日," and "6 日," proper translations of which are "4th," "5th," and "6th," respectively. The MT system specified "June" improperly. This error stems from the wrong phrase alignment in the statistical model. These types of errors should be addressed during training the models and/or post-editing, rather than pre-editing. Our protocol enables us to identify MT errors that are difficult to amend only by the pre-editing strategy. This will eventually help us streamline the overall translation workflow using off-the-shelf MT systems.

4 Machine Translatability into Different Languages

We examined the effectiveness of the pre-editing strategy in a multilingual translation setting, i.e., whether an ST, edited so that it is better translated into one target language, can also be better translated into other languages. First, all the original and the Best STs in the four datasets (800 sentences in total) were translated into Chinese and Korean using the corresponding models of TexTra. Then, for each set of Chinese and Korean translations, one human evaluator was asked to assess the MT quality using the 5-point scale in Table 1.

As shown in Table 6, the MT quality for the Best STs was, on average, higher than that for the original STs for all the datasets, indicating that edit operations that improved English-translatability of Japanese STs are portable to Chinese- and Korean-translatability to a certain degree. For both languages, the MT quality for the Best STs in *hosp* and *bccwj* well surpassed that for the original STs, while there were no significant improvements in *muni* and *reuters*. Further scrutiny into the language dependency of machine-translatability is important to justify the pre-editing approach to other target languages and domains.

5 Typology of Edit Operations

We analyzed the diversity of edit operations exhibited during our pre-editing exercise. As mentioned in Section 2, it is likely that the best path contains edit operations effective in improving MT quality. We therefore focused on pre-edit instances in the

Chinese	Avg. score		Num. of units (Org vs. Best)		
	Org	Best	>	=	<
hosp	2.73	2.93**	7	70	23
muni	2.84	2.89	32	31	37
bccwj	2.39	2.75**	13	42	45
reuters	2.61	2.77	22	45	33
Korean	Avg. score		Num. of units (Org vs. Best)		
	Org	Best	>	=	<
hosp	3.32	3.56**	12	57	31
muni	3.58	3.67	32	29	39
bccwj	3.37	3.60*	18	47	35
reuters	3.31	3.36	24	47	29

Table 6: Results of human evaluation of MT quality: "$*$" and "$**$" indicate significant differences over "Org(inal ST)" tested by Wilcoxon signed-rank test with $p < 0.05$ and $p < 0.01$, respectively.

Dataset	Num. of instances in best path		
	(a) raw	(b) decomposed	(b)/(a)
hosp	97	185	1.91
muni	106	186	1.75
bccwj	174	340	1.95
reuters	191	268	1.40

Table 7: The number of decomposed pre-editing instances in the best path of 10 sampled units.

best path of 10 randomly sampled units for each of the four datasets. First, we decomposed each of the sampled 568 instances into a sequence of primitive edit operations, because our editor might not strictly seek the primitiveness. Indeed, as shown in Table 7, this process increased the number of instances by from 1.40 to 1.95 times, resulting a total of 979 instances of primitive edit operations. We then manually created a typology of edit operations, by categorizing each instance, regarding surface-level differences of each pair of STs as clues. Table 8 shows the resulting typology with 53 types of edit operations that cover all of the analyzed pre-edit instances, with their frequency in each dataset. We observed an extended variety of edit operations in our collection, ranging from ones at surface-level, such as insertion/deletion of punctuation and word reordering, to various types of syntactic alternation.

The most frequent type across the datasets was **C01 (Alternative lexical choice)**, including edit operations such as replacing "一度" with "一回" (both mean *once*), and "習得する" (*acquire*) with "学ぶ" (*learn*). This type of edit operations would be automated by constructing lexical resources tailored to particular MT systems. We also identified several frequent types of edit operations that are likely to be effective for improving MT quality. For example, **S05 (Phrase reordering)** and **S07 (Insertion/deletion of punctuation)** can help MT systems parse the input sentences correctly, which subsequently leads to better MT outputs.

Some types of edit operations were observed only in specific domains. For example, we observed **S15 (Use/disuse of clause-ending noun)** and **S20 (Use/disuse of nominal/verbal suffix)** only in the news domain (*bccwj* and *reuters*). Both types reflect the fact that the elliptic expressions often used in news articles could degrade the MT quality. Our method is also useful to unveil these kinds of domain-specific issues.

Last but not least, let us describe a less frequent type of edit operations, i.e., **S13 (Head-switching of verb phrase)**:

[before] 懸念を強め (*strengthen anxiety*)

[after] 強い懸念を抱き (*have strong anxiety*)

This type of edit operation has not been covered by existing controlled language rule sets in Japanese, such as (Ogura et al., 2010; Hartley et al., 2012; Miyata et al., 2015), nor even by a comprehensive typology of paraphrases.[7] It is worth exploring to what extent these types of edit operations are effective in improving MT quality.

6 Conclusion and Future Work

In this paper, we have presented our human-in-the-loop protocol for collecting pre-edit instances. Using this protocol, we collected 12,687 pre-edit instances for four different datasets, demonstrating that most of the source sentences can be edited into machine-translatable ones. Human evaluation revealed that, for some datasets, English-translatable Japanese STs significantly improved the quality of translations into Chinese and Korean. We also built a typology comprising a wide range of edit operations, and found that alternating lexical choice was the most frequent one taken by our editor.

Based on this study, we plan to develop an automatic pre-editor. One approach to this is controlled language formulation by assessing the effectiveness of each type of edit operation (Bernth and Gdaniec, 2001; Miyata et al., 2015). Another is to build a statistical model. It is worth investigating data-driven methods based on our collection of pre-edit instances, although this data do not guar-

[7] http://paraphrasing.org/paraphrase.html

ID	Type	Freqency			
		H	M	B	R
S01	Division/synthesis of sentence(s)	4	1	7	2
S02	Use of line break	0	3	0	0
S03	Use of compound/complex sentence	0	0	0	1
S04	Split of phrase	0	0	1	0
S05	Phrase reordering	24	6	22	13
S06	Insertion/deletion of subject	0	2	2	2
S07	Insertion/deletion of punctuation	24	5	27	27
S08	Change of scope of subject	0	0	1	1
S09	Use of nominative case "*ga*" or topic marker "*wa*"	0	1	3	2
S10	Change of marked element	0	2	11	0
S11	Change of voice	3	1	13	3
S12	Change of restrictive/continuous modification	2	0	12	13
S13	Head-switching of verb phrase	0	0	0	3
S14	Indication of conditional clause	2	7	2	0
S15	Use/disuse of clause-ending noun	0	1	3	5
S16	Change of subject in noun phrase	0	0	1	0
S17	Use of noun phrase or verb phrase	3	4	9	0
S18	Use/disuse of compound verb	2	0	2	0
S19	Use/disuse of compound noun	2	7	5	8
S20	Use/disuse of nominal/verbal suffix	2	1	10	5
S21	Change of connective expression	6	16	12	13
S22	Change of parallel expression	2	3	1	0
S23	Change of apposition expression	0	0	0	5
S24	Change of specification expression	0	0	0	3
S25	Change of locative expression	0	0	0	2
S26	Change of hearsay expression	0	0	0	4
S27	Change of expression for indirect question	0	0	0	1
S28	Change of *sahen* noun expression	1	2	7	4
S29	Change of formal noun expression	0	1	3	5
S30	Change of substantive verb expression	1	0	0	1
S31	Change of *ni-/to-naru* expression	0	0	0	11
C01	Alternative lexical choice	29	36	69	33
C02	Lexical elaboration	5	3	2	1
C03	Lexical simplification	0	5	0	0
C04	Change of reference expression	0	0	0	1
C05	Use of redundant expression	0	1	0	1
F01	Use of honorific expression	19	11	14	4
F02	Change of tense	0	3	1	2
F03	Change of conjunctive word	4	4	0	1
F04	Change of auxiliary verb	1	0	0	0
F05	Insertion/deletion of particle	4	9	24	9
F06	Use of particle	4	3	3	10
F07	Use of compound particle	0	1	1	5
T01	Change of named entity	0	0	3	6
O01	Orthographical change	1	7	7	4
O02	Change of sentence-ending expression	0	1	2	0
O03	Insertion/deletion/change of symbol	0	6	0	0
O04	Insertion of omitted element	0	0	3	2
O05	Specification of chunk with brackets	0	5	3	1
I01	Change of content	18	20	27	16
I02	Change of nuance	0	7	17	6
E01	Grammatical errors	3	1	4	6
E02	Other errors	19	0	6	6

Table 8: Our typology of edit operations (H: *hosp*, M: *muni*, B: *bccwj*, R: *reuters*): The first letter of ID indicates seven major categories: **S** (**S**tructure), **C** (**C**ontent word), **F** (**F**unctional word), **T** (**T**erminology), **O** (**O**rthography), **I** (**I**nformation), and **E** (**E**dit that causes/resolves error in ST).

antee to improve MT quality as directly addressed by post-editing (Simard et al., 2007).

Acknowledgments

We are deeply grateful to the anonymous reviewers for their valuable comments on the earlier version of this paper. This work was partly supported by JSPS KAKENHI Grant Numbers 25730139 and 25240051. One of the corpora used in our study was created under a MIC program "Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology."

References

Aikawa, Takako, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proc. of MT Summit*, pages 1–7.

Bernth, Arendse and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16(3):175–218.

Hartley, Anthony, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata. 2012. Readability and translatability judgments for 'Controlled Japanese'. In *Proc. of EAMT*, pages 237–244.

Mirkin, Shachar, Sriram Venkatapathy, Marc Dymetman, and Ioan Calapodescu. 2013. SORT: An interactive source-rewriting tool for improved translation. In *Proc. of ACL: System Demonstrations*, pages 85–90.

Miyata, Rei, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proc. of MT Summit*, pages 90–103.

O'Brien, Sharon and Johann Roturier. 2007. How portable are controlled language rules? A comparison of two empirical MT studies. In *Proc. of MT Summit*, pages 345–352.

Ó Broin, Ultan. 2009. Controlled authoring to improve localization. *Multilingual*, Oct./Nov.:12–14.

Ogura, Eri, Mayo Kudo, and Hideo Yanagi. 2010. Simplified technical Japanese: Writing translation-ready Japanese documents. *IPSJ SIG Technical Reports*, 2010-DD-78(5):1–8. (in Japanese).

Pym, Peter. 1988. Pre-editing and the use of simplified writing for MT: An engineer's experience of operating an MT system. In *Proc. of Translating and the Computer 10*, pages 80–96.

Resnik, Philip, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving translation via targeted paraphrasing. In *Proc. of EMNLP*, pages 127–137.

Simard, Michel, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proc. of NAACL-HLT*, pages 508–515.

Uchimoto, Kiyotaka, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2006. Automatic detection and semi-automatic revision of non-machine-translatable parts of a sentence. In *Proc. of LREC*, pages 703–708.