Improving Translation Fluency with Search-Based Decoding and a Monolingual Statistical Machine Translation Model for Automatic Post-Editing

Jing-Shin Chang
Department of Computer Science
& Information Engineering
National Chi Nan University
1, Univ. Road, Puli, Nantou 545, TAIWAN
jshin@csie.ncnu.edu.tw

Sheng-Sian Lin
Department of Computer Science
& Information Engineering
National Chi Nan University
1, Univ. Road, Puli, Nantou 545, TAIWAN
s94321509@ncnu.edu.tw

Abstract

The BLEU scores and translation fluency for the current state-of-the-art SMT systems based on IBM models are still too low for publication purposes. The major issue is that stochastically generated sentences hypotheses, produced through a stack decoding process, may not strictly follow the natural target language grammar, since the decoding process is directed by a highly simplified translation model and n-gram language model, and a large number of noisy phrase pairs may introduce significant search errors. This paper proposes a statistical post-editing (SPE) model, based on a special monolingual SMT paradigm, to "translate" disfluent sentences into fluent sentences. However, instead of conducting a stack decoding process, the sentence hypotheses are searched from fluent target sentences in a large target language corpus or on the Web to ensure fluency. Phrase-based local editing, if necessary, is then applied to correct weakest phrase alignments between the disfluent and searched hypotheses using fluent target language phrases; such phrases are segmented from a large target language corpus with a global optimization criterion to maximize the likelihood of the training sentences, instead of using noisy phrases combined from bilingually word-aligned pairs. With such search-based decoding, the absolute BLEU scores are much higher than automatic post editing systems that conduct a classical SMT decoding process. We are also able to fully correct a significant number of disfluent sentences into completely fluent versions. The BLEU scores are significantly improved. The evaluation shows that on average 46% of translation errors can be fully recovered, and the BLEU score can be improved by about 26%.

Keywords: Translation Fluency, Fluency-Based Decoding, Search-Based Decoding, Statistical Machine Translation, Automatic Post-Editing

1 Introduction and Motivation

1.1 Fluency Problems with Statistical Machine Translations

Translation fluency of Machine Translation systems is a serious issue in the current SMT research works. With the research efforts for the past tens of years, the performances are still far from satisfactory. In translating English to Chinese, for instance, the BLEU scores [16] range only between 0.21 and 0.29 [22, 5, 17], depending on test sets and numbers of reference translations. Such translation quality is extremely disfluent for human readers. We therefore propose a statistical post-editing (SPE) model, based on a special monolingual SMT framework, for improving the fluency and adequacy of translated sentences.

The classical IBM SMT models [1, 2] formulate the translation problem of a source sentence F as finding the best translation E* from some stack decoded hypotheses, E, such that:

$$\begin{aligned} E^* &= \arg\max_E \Pr(E \mid F) \\ &= \arg\max_E \Pr(F \mid E) \times \Pr(E) \end{aligned} \quad (1)$$

where $\begin{cases} E\text{:target sentence} \\ F\text{:source sentence} \end{cases}$ and $\begin{matrix} \Pr(F \mid E) : Translation\ \ Model\ \ (TM) \\ \Pr(E) : Language\ \ Model\ \ (LM) \end{matrix}$

The $\arg\max_E$ operation implies to generate candidate target sentences E of F so that the SMT model can score each one, based on the TM and LM scores and select the best candidate. The process of candidate generation is known as the decoding process. The conventional decoding process is significantly affected by the TM and LM scores; only those candidates that satisfy the underlying criteria of the TM and LM will receive high scores. Unfortunately, to make the SMT computationally feasible, the TM and LM are highly simplified. Therefore, the candidates are not really generated based on target language grammar, but based on the model constraints. For instance, the classical SMT model does not prefer word re-ordering with long distance movement. Such candidates are then not generated regardless of the possibility that the target grammar might prefer them.

1.2 LM and Decoding

There are three directions to improve the translation fluency with the classical SMT model, Equation (1). Firstly, we can improve the Translation Model (TM) to fit the source-target transfer process. Secondly, we can improve the Language Model (LM) to respect the target language grammar. Finally, we could try to generate better and much more fluent candidates in the decoding process so that the TM and LM can select the real best one from fluent candidates, rather than from junk sentences.

The research communities normally focus on the TM and LM components by assuming that there are good ways to generate good candidates for scoring. Actually, most attention is paid to the Translation Model (TM); LM and decoding were not gaining the same weight. In particular, people tend to think that the candidate generation process guided by the highly simplified TM and LM will eventually generate good candidates.

Unfortunately, to make the computation feasible, the classical SMT models have very low expressive power in the Translation Model (TM) and Language Model (LM) components. It formulates the TM in terms of the *fertility* probability, lexical *translation* probability and *distortion* probability [1, 2]. A word-based 3-gram model is usually used as the language model (LM). Longer n-grams are used at higher training cost and severe data sparseness.

In fact, the candidates of the target sentence, which are hidden in the $\arg\max_E$ operator, are generated as a stochastic process in most SMT today. Starting from a particular state, the next word is predicted based on a local n-gram window within a distance allowed by the distortion criterion; the possible paths are exploited using stack decoding, beam search or other searching algorithms. The candidates generated in this way thus may be only "piecewise" consistent with the target language grammar, but may not be really globally

grammatical or fluent. This means that the TM and LM are not scoring a complete sentence but some segments pasted by the n-gram LM. It is then not likely to be fluent all the time.

This decoding process therefore sometimes falls into the "garbage-in and garbage-out" situation. No matter how well-formulated the TM and LM may be, if the stochastically generated candidates do not include the correct and fluent translation, the system will eventually deliver a garbage output, that is, a disfluent sentence, as the *best* one. This kind of error is known as searching error. Because the TM and LM have limited expressive power to describe the real criteria that carry the generation process, the decoding process might only generate noisy sentence segments and thus disfluent sentences for scoring. This could lead to bad performance in terms of BLEU score or human judgments.

Phrase-based SMT had partially resolved the expressive power issue of TM and LM by using longer word sequences. However, the acquisition of "phrases" has its own problems. In particular, most phrase-based SMT acquires the phrase pairs by conducting bilingual word alignment first. Adjacent words are then connected in some heuristic ways [12, 13, 14, 15], which do not have direct link with the source or target grammar, to form the "phrases". The phrases generated in this way normally do not satisfy any global optimization criteria related to the target grammar, such as maximizing the likelihood of the target language sentences. The quality of such phrases is therefore greatly affected by the word alignment accuracy; and, the phrases for the target language side may not really respect the target grammar. Under such circumstances, a huge number of noisy "phrases" will be introduced and significantly enlarge the searching space. The stochastically generated phrase sequences thus may not correspond to good candidate sentences either.

To summarize, the application of word-for-word or phrase-to-phrase translation (with "noisy" phrases) plus a little bit local word/phrase re-ordering in classical SMT might not generate fluent target sentences that respect the target grammar. In particular, many target specific lexical items and morphemes cannot be generated through this kind of models. If they do, they may be generated in very special ways. This could be a significant reason why the SMT models do not work well after the long period of research.

The implication is that we might have to examine the $\arg\max_{E}$ operation, that is, the *decoding* or *searching* process, in the classical SMT models more carefully. We should try decoding method that respect target grammar more, instead of following the criteria set forth by the TM and LM of the SMT model, which encode highly simplified version of the target grammar. Only with a decoding process that respect the target grammar, will the system generate fluent candidates at the first place before submitting the candidates to the TM and LM for scoring.

Furthermore, a phrase-based language model, instead of word-based n-gram model for the target side may improve the fluency of machine translation further since more context words can be consulted, if the "phrases" are not noisy. To avoid a huge number of noisy source-dependent phrases that might be harmful for fluency and searching, such phrases may better be trained from a target corpus, instead of being acquired from bilingually word-aligned chunks.

1.3 Statistical Post-Editing Model Based on Monolingual SMT

Instead of developing new models for the TM and LM, an alternative to improve the translation fluency is to cascade an Automatic Post-Editing (APE) module to the translation output of an MT/SMT system. While the classical SMT models may not be suitable for directly *generating* fluent translation, due to the limited expressive power of the TM and LM and search errors of the decoding process, an SMT or its variant may be sufficient for re-ranking hypotheses in the automatic post editing purposes, if appropriate hypotheses generation mechanism is available. Actually, we can regard a post-editing process as a translation process from disfluent sentence to fluent sentence. This is particularly true if the disfluency is limited to local editing operations like *insertion* of target specific morphemes, *deletion* of source-specific function words, and *lexical substitution* from many possible lexical choices. These kinds of errors are often seen in MT/SMT systems. Inspired by the above ideas, this paper propose a statistical post-editing (SPE) model based on a monolingual SMT paradigm for improving the translation fluency of an MT system, instead of improving the TM directly.

In this SPE model, the searching or decoding is a fluency-based search. We search fluent translations, based on the lexical hints of the disfluent sentence, from a large target text corpus or from the Web. Therefore, all candidates will be fluent ones. The best hypotheses re-ranked best by the SPE model will then serve as the post-edited version of the disfluent sentence. Sometimes, a searched sentence may not have a high translation score to justify itself as an appropriate translation. For instance, the target sentence pattern may be correct but different lexical choices have been made. In this case, automatic local editing is applied to the weakest alignments to incrementally patch the target sentence pattern with right target lexical items. By combining the grammatical (and fluent) sentence pattern of the searched sentence and the right lexical items from the disfluent sentence, the disfluent translation could be repaired to a fluent one incrementally. This may include some local insertion, deletion and lexical substitution operations over phrase pairs that are unlikely to be translation of each other.

To really improve the fluency incrementally, the local editing process is applied in a manner that will monotonically increase the likelihood of the incrementally repaired sentence. To respect the target grammar further, the repair is phrase-based. In other words, phrase-based n-gram language model (n=1) is used in the translation score so that the likelihood of the repaired target sentence is incrementally increased during the local editing process.

In parallel with the development of our work, a few APE systems were also proposed [7, 20, 21, 8] with good results. Publicly available SMT systems (like Portage PBMT, Moses, etc.) are used directly as the post-editing module. They are trained using human post-edited target sentences with their un-edited MT outputs to learn the translation knowledge between disfluent ('source') and fluent ('target') sentences [20]. Alternatively, they may be trained using standard parallel corpora (Europarl, News Commentary, Job Bank, Hansard, etc.) where the disfluent sentences are generated using a rule-based MT (like SYSTRAN) or other SMT [21].

Therefore, these works require substantial human post-editing costs to train the SMT. Or they need a sizable parallel corpus for training, which may not be available to many language pairs. In addition, it requires an RBMT or SMT pre-trained for translating the source corpus, which may not be available to many language pairs. Most importantly, these frameworks use

the same decoding process as well as the TM and LM of the original SMT to generate their post-editing hypotheses. Therefore, the previously discussed performance issues that apply to classical SMT will also apply to such APE modules. The cascade of an SMT as an APE module might imply the use of a system with low BLEU performance to correct the outputs with low BLEU scores. The improvement could thus be substantially limited. This may be seen from the fact that the contribution of the APE becomes negligible as the training data is increased [21].

In contrast, we discard the stochastic decoding process, which might generate disfluent hypotheses, but search a large corpus for highly similar sentences to the disfluent sentence, and thus will have raw hypotheses with high BLEU scores. Additional local editing will further improve the fluency. Furthermore, our proposal can generate interesting error patterns automatically using the target language corpus alone. Therefore, the APE module can be constructed without a real MT system (although it would be better to have one in order to correct the specific errors of a specific system.). The following sections will discuss the formulation in more details.

2 Problem Formulation for SPE

In our work, we propose to adopt a Statistical Post-Editing (SPE) Model to translate disfluent sentences into fluent versions. Such a system can be regarded as a "disfluent-to-fluent" SMT. As will be seen later, it can be trained with a Monolingual SMT Model. Given a disfluent sentence E' translated form a source sentence F, the automatic post-editing problem can be formulated as finding the most fluent sentence E* from some candidate sentences E such that:

$$\begin{aligned} E^* &= \arg\max_E \Pr(E \mid E') \\ &= \arg\max_E \Pr(E' \mid E)\Pr(E) \end{aligned} \quad (2)$$

As usual, we will refer Pr(E'|E) as the translation model (TM), and Pr(E) as the language model (LM) of the SPE model. We thus encountered the same SMT problems to formulate the TM, LM and the decoding (or searching) process.

2.1 Order-Preserved Translation Model

The automatic post-editing problem is intuitively easier than SMT since we can assumes that the disfluency is due to some local editing errors, such as mis-insertion or mis-deletion of function words, and wrong lexical choices. Under this assumption, we can formulate the TM as:

$$\begin{aligned} &\Pr(E' \mid E) \\ &= \sum_A \Pr(E', A \mid E) \\ &\approx \max_A \Pr(E', A \mid E) \\ &\approx \Pr(E', A_s \mid E) \\ &= \prod_{E_p = A_s(E'_p)} \Pr(E'_p \mid E_p) \end{aligned} \quad (3)$$

In Eqn. (3), phrase-aligned phrase pairs are represented by E'p and Ep for the disfluent and fluent versions, respectively. We assume that the most likely alignment A_s, among all generic alignment pattern A, between E' and E is an "order-preserved" or "sequential" alignment between their constituents. We further assume that this most likely alignment has much higher probability than other alignments such that we don't have to sum over all generic alignment patterns. In the post-editing context, this assumption may be reasonable if the disfluency results from simple local editing operations. In particular, if we are using phrase-based alignment, the word order within the phrases can be ignored. The order preservation assumption will be even more reasonable. We therefore assume that the TM is the product of the probabilities of sequentially aligned target phrase pairs. The phrase segmentation model for dividing E or E' into phrases will be further detailed later when discussing the target phrase-based LM. Given the segmented phrases, the best sequential alignment can easily be found using a standard dynamic programming algorithm for finding the "shortest path".

The TM for the SPE model is special in that the training corpus can be easily acquired from a large monolingual corpus with fluent target sentences. Generating a disfluent version of the fluent monolingual corpus automatically based on some error model of the translation process will make this possible. One can then easily acquire the model parameters for translating disfluent sentences into fluent ones through a similar training process for a standard SMT. In comparison with standard SMT training, which requires a parallel bilingual corpus, the monolingual corpus is much easier to acquire.

2.2 Target Phrase-Based Language Model

To respect the fluency of the target language in the decoding process, the language model score Pr(E) should be evaluated based on long target language phrases, Ep, instead of target words. The "phrases" should also be defined independent of source-language in order not to introduce a huge number of noisy phrases as PBSMT normally did. The proposed LM for the current SPE, which is responsible for selecting fluent target segments, is therefore a phrase-based *unigram* model, instead of the widely used word-based *n-gram* model. In other words, we have

$$\Pr(E) = \prod_{Ep \in E} \Pr(Ep).$$

To avoid source-language dependency, we also decided not to define target phrases in terms of chunks of bilingually aligned words. Instead, the best target phrases are directly trained from the monolingual target corpus by optimizing the phrase-based unigram model. In other words, the best phrase sequence $\vec{p}^*$ for an n-word sentence w_1^n, will be the sequence, among all possible phrase segmentation, p_1^m, such that:

$$\vec{p}^* = \arg\max_{p_1^m} \Pr\left(p_1^m \mid w_1^n\right) = \arg\max_{p_1^m} \prod_i \Pr\left(p_i\right).$$

Fortunately, extracting monolingual phrases using the phrase-based uni-gram model can be done easily. The training method is just like the word based uni-gram word segmentation model [4], which was frequently used in Chinese word segmentation tasks. Unsupervised training is easy for this. Upon convergence, a set of well-formed phrases can be acquired. (This set of phrases will be called a phrase example base, PEB. Phrases in the PEB will be used later in the Local Editing Algorithm for post-editing.)

Since a phrase trained in this way can be longer than a 3-gram pattern, the modeling error could be reduced to some extend. Furthermore, the number of such phrases will be much smaller than those randomly combined phrases acquired from word-aligned word chunks. As a result, the estimation error due to data sparseness will be significantly reduced too. Unlike the rare parallel bilingual training corpus, the amount of such target language corpora is extremely large. Therefore, fluent phrases can be extracted easily. With phrases as the basic lexical unit, SPE model will reduces to

$$E^* = \arg\max_{E} \prod_{Ep = As(Ep')} \Pr(Ep' \mid Ep) \Pr(Ep) \quad (4).$$

Since a phrase can cover more than 3 words, the selected phrases might be more fluent than word trigrams. Such phrases will fit target grammar better and therefore will prefer more fluent target sentences in general.

2.3 Search-Based Decoding for Fluency

One key issue that causes disfluency is the decoding process used in classical SMT. Most decoding process regard target sentence generation as a stochastic process, and only local context of finite length window is consulted while decoding. Therefore, the target sentences generated in this way are usually not fluent. Our work proposes to search fluent translation candidates from a huge target sentence base or from web documents, instead of using traditional decoding methods to generate the translation candidates. Since the large corpus and the Web documents are produced by native speakers, the target sentences thus searched are most likely fluent with high BLEU scores.

Our current work simply used a heuristic matching score to extract a set of candidate sentences for a disfluent sentence. The candidates are then re-ranked using the translation score defined by the SPE model. The best candidate will be regarded as the post-edited version of the disfluent sentence if the translation score is higher than a threshold. Otherwise, it will be locally edited to incrementally increase its translation score. The matching score is simply the number of identical word tokens in two sentences, which is normalized by the average length of the two sentences. In other words, it is the percentage of word matches between two sentences.

We searched the candidate translations from the Academia Sinica Word Segmentation Corpus, ASWSC-2001 [6], as well as Chinese webpages indexed by Google. (We assume that the target language is Chinese.) Different query strings will result in different returned pages. Totally, we have tried 4 models for searching:

(1) Model **C**: search the corpus (only) for Top-N hypotheses (N=20). (The length difference must not be greater than two words.)
(2) Model **C+W**: search the corpus and the web for additional N hypotheses by submitting the complete disfluent target sentence as-is to Google.
(3) Model **C+W+P**: including partial matches against substrings of the disfluent target sentence, where 1~L-1 words in the disfluent sentence are successively deleted and then submitted as query strings to the search engine. (L: number of words in disfluent sentence)
(4) Model **C+W+Q**: adjacent words in the deleted disfluent sentence are quoted as a single query token before submission so that the search engine will match more exactly.

Even with such a heuristic search, a substantial number of fluent sentences similar to the disfluent sentences can be found for re-ranking and local editing.

2.4 Local Editing

If exact translation is found during searching, the searching process itself is exactly a perfect translation process. If highly similar sentences are found, simple lexical substitution or automatic post-editing [9, 11] might patch the searched fluent sentences into correct translations. Some previous works for automatic post editing have been restricted to special function words, such as the English article 'the/a' [9, 10], the Japanese case markers and Chinese classifier or particle 'de' [18]. The automatic post-editing model here is intended to resolve general editing errors that are frequently made by a machine translation system.

Briefly, the best sentence E^*_{eb} in the searched candidates will be output as the translation of the disfluent translation E' if the translation score associated with the SPE model is higher than a threshold. (The set of candidate translation sentences is called its example base, thus the subscript 'eb'.) Otherwise, the automatic local editing algorithm will find the weakest phrase alignments and fix them one-by-one to maximize the translation score.

An alignment phrase pair <Ep', Ep> is said to be "weak" if its local alignment score Pr(Ep'|Ep) x Pr(Ep) is small and thus contributes little to the global translation score for the sentence pair <E', E>. When the weakest pair, (Ep'- | Ep-) with the lowest local alignment score is identified, we should try to replace Ep-, the "most questionable phrase" in the fluent (yet incorrect) example sentence E, with some candidates that would make the patched example sentence more likely to be the translation of E'.

There are some reasons why the alignment (Ep'- | Ep-) is the weakest. First of all, Ep- might not be the right phrase, and should be replaced by Ep'- to make the fluent sentence E also the correct translation of E'. Second, Ep'- might not be the correct translation of some source phrase. In this case, the most likely translation(s) of Ep'-, called Ep+, should be used to replace Ep-. Third, Ep- is a more appropriate phrase than Ep+. In this case, it should be retained and next weakest alignment pair be repaired.

As a result, potential candidates for replacing Ep- will include Ep'-, Ep+ and Ep- itself. The best substitution will be the phrase that maximizes Pr(Ep'|Ep) x Pr(Ep). Actually, many phrases in the PEB can be a more fluent version of Ep'-. Currently, the 20 best matches will play the role of Ep+ during local editing. And the local editing algorithm will successively edit weaker alignments until the (monotonically increasing) translation score is above some threshold. The algorithm is outlined as follows.

Local Editing Algorithm

Input : E' and E^*_{eb}

Step 1 : Find the weakest alignment entry in E' from the $< E', E^*_{eb} >$ alignment.

$$Ep'- = \arg\min_{Ep' \in E'} \Pr(Ep' \mid Ep) \Pr(Ep)$$

Step 2 : Identify $Ep-$ that is the phrase in E^*_{eb} aligned with $Ep'-$.

$$Ep- = \text{align}(Ep'-)$$

Step 3 : Find the fluent phrase $Ep+$ of $Ep'-$ from PEB.

$$Ep+ = \text{PEB}(Ep'-)$$

Step 4 : Select the best substitution among Ep'-, Ep+ and Ep- which maximize the translation score:

$$E^*_{ps} = \arg\max_{E_{ps} \in \{Ep'-, Ep+, Ep-\}} \Pr(E' \mid E) \Pr(E)$$

Step 5 : Cut $Ep-$ from E^*_{eb} and paste Eps to E^*_{eb}.

$$E^*_{eb} = E^*_{eb} - (Ep-) + (Eps)$$

(Repeat until the translation score Pr(E'|E)xPr(E) reaches some threshold.)

Constrained Decoding

Note that, local editing is applied only to a local region of the example sentence based on the disfluent sentence. Intuitively, those sentences searched from a text corpus or from the Web corpus will be much more fluent than stochastically combined sentences from the SMT decoding module. Even if local editing is required, the repair will be quite local. The search space for repairing will be significantly constrained by words in the most likely example sentence. Such a searching and local editing combination can thus be regarded as a *constrained decoding*. The searching error can thus be reduced significantly in comparison with the large search space of the decoding process of a typical SMT.

2.5 Generating Faulty Sentences

The TM parameters can actually be trained from an E'-to-E monolingual Machine Translation System, where E' can be derived by applying to E some commonly found editing operations in the SMT translation process. The operations might include the insertion of target specific lexicon, deletion of source specific lexicon, local reordering of words and substitution of lexical items.

In the current work, we apply three kinds of editing operations to the fluent sentences in a monolingual corpus to simulate frequently found errors in an MT system. The fluent and its disfluent versions are then phrase segmented so that the sentences are represented by phrase tokens (instead of word tokens). Such fluent-disfluent (E-E') target sentence pairs are then trained using the GIZA++ alignment tools [12, 13, 14, 15]. Upon convergence, the translation model between the sentences to be post-edited and their correct translation can readily be acquired.

The three editing operations include:

(1) **Insertion**: The insertion errors will occur when an MT system translates a source word into a target word while it should not be translated. For instance, the English infinitive "to" need not be translated into any Chinese word most of the time. But the bilingual dictionary may indicate the possibility to translate it into "去" (chu). We therefore automatically insert the Chinese words to simulate such an error.

(2) **Deletion**: The deletion error occurs when a target specific word is not generated in the translation. For instance, the Chinese classifiers have no correspondence in the English language. We therefore delete the following classifiers from fluent Chinese sentences to create instances with deletion errors: '個', '隻', '枝', '位', '顆', '棵'.

(3) **Substitution**: When a translation system chooses a wrong lexical item, a typical substitution error will occur. To simulate the substitution errors, Chinese words in the fluent sentences are lookup against an English-Chinese dictionary. Chinese words that are also the translation of the English word are then substituted to simulate the substitution error. For instance, '問題' is a Chinese translation for the English word 'problem'. But 'problem' also has other translations, like '習題' and '疑難'. These words are therefore used to simulate the substitution errors. In our simulation, the top-30 most frequently used Chinese words are adopted to simulate the substitution errors.

With disfluent sentences created from fluent sentences with the above frequently encountered translation errors, an automatic statistical post-editing model can readily be trained using state-of-the-art alignment tools.

3 Experiments

To see the performance of the current SMT-based SPE model, about 300,000 word segmented Chinese sentences from the Academia Sinica [6] was used as our target sentence corpus. The corpus has about 2,450,000 word tokens, and the vocabulary size is about 83,000 word types. 10% of the sentences are used as the test set and 90% are used for training. The 3 types of errors are applied to the testing sentences independently. For each error type, 100 sentences are randomly selected for evaluating automatic post editing.

The performance is evaluated in terms of two criteria. The first criterion is the number (percentage) of fully corrected disfluent sentences from the test set. By fully corrected, we mean that the sentence corrected by the statistical post editing (SPE) system is completely the same as its original fluent version. Table 1 indicates the performance in terms of the error correction capability.

Error types	Searching Models			
	C	C+W	C+W+P	C+W+Q
Substitution	21	23	32	34
Deletion	28	39	46	62
Insertion	40	43	47	47
Average	30	35	42	48

Table 1. Number of fully corrected sentences with different searching models (N=100)

Note that, even with the very simple minded searching method, the SPE was able to correct, on average, about 48% of the faulty sentences to their fluent version if the search space is sufficiently large (with the C+W+Q searching model). The performance increases with the search space. And the performance is increased at most by 62%, 121% and 17.5 %, respectively for the substitution, deletion and insertion errors when the Web corpus is included to the search space. Obviously, the substitution is the hardest to resolve while insertion error seems to be easier to resolve.

The second evaluation criterion is the improvement in the BLUE score with respect to the un-corrected test sentence. Table 2 shows the BLEU scores for the various searching models. The first column labeled as E'(ts) lists the BLEU scores for the test sentences that has not been post-edited. By searching for fluent translation and applying local editing, the BLEU scores are improved with increasing search space. The best performance is to increase the BLEU scores by 15%, 38% and 26% respectively for the three types of errors. On average, the improvement is about 26%, which is substantial. On the other hand, the absolute changes are 9.4, 22.8 and 16.9 points in BLEU score, respectively.

Error Types	BLEU Scores				
	E'(ts)	C	C+W	C+W+P	C+W+Q
Substitution	0.637	0.656	0.676	0.737	0.731
Deletion	0.598	0.686	0.750	0.781	0.826
Insertion	0.646	0.762	0.780	0.810	0.815

Table 2. BLEU Scores for Various Searching Models

Note that, with search-based decoding, the absolute BLEU scores are much higher than automatic post editing systems that simply cascade a classical SMT module to the output of an MT/SMT [20, 21, 8]. Although the experiment settings are not the same and thus cannot be compared directly, the results to have higher absolute BLEU scores can be expected since searched sentences are almost always fluent, whether they are post-edited or not.

Obviously, with the same training corpus, the search space and the searching method play important roles in improving the performance. The inclusion of the web corpus does improve the performance significantly. It was reported in [19] that well formulated query strings can effectively improve searching accuracy. Therefore, by using better searching strategy, part of the translation problems for fluent translation might be resolved as a searching and automatic post-editing problems. Currently, a statistical searching model specific for the fluency-based decoding is being developed.

4 Concluding Remarks

In this paper, we propose not to generate sentence hypotheses for APE systems by using conventional SMT decoding process, since such a decoding process tends to lead to an open-ended search space. It is not easy to generate fluent sentence hypotheses under such circumstances due to the large search error. We propose to search sentence hypotheses, from a large target text corpus or from the web, based on the words in the disfluent translations, since the potential candidates will mostly be fluent. A statistical post-editing model is also proposed to re-rank the searched sentences, and a local editing algorithm is proposed to automatically recover the translation errors when the searched sentence is not a good

translation. With the SPE, the local editing algorithm tries to maximize the translation score for each local editing. It therefore improves the translation fluency incrementally. Since the TM can be trained from an automatically generated fluent-disfluent parallel corpus, training such a system is easy. The evaluation shows that, on average, 46% of translation errors can be fully recovered, and the BLEU score can be improved by about 26%. The absolute BLEU is also high with the search-based decoding process in comparison with conventional decoding process.

References

[1] Brown, Peter F., J. Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation." *Computational Linguistics*, 16(2):79–85, 1990.

[2] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation." *Computational Linguistics*, 19(2):263–311, 1993.

[3] Chang, Jing-Shin and Chun-Kai Kung, "A Chinese-to-Chinese Statistical Machine Translation Model for Mining Synonymous Simplified-Traditional Chinese Terms," *Proceedings of Machine Translation Summit XI*, pages 81-88, Copenhagen, Denmark, 10-14, September, 2007.

[4] Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su, "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pp. 123-146, Taipei, Taiwan, R.O.C., 1992.

[5] Chiang, David, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," *Proc. ACL-2005*, pages 263–270, 2005.

[6] CKIP 2001, *Academia Sinica Word Segmentation Corpus*, ASWSC-2001, (中研院中文分詞語料庫), Chinese Knowledge Information Processing Group, Acdemia Sinica, Tiapei, Taiwan, ROC, 2001.

[7] Dugast, L., J. Senellart, P. Koehn, "Statistical Post-Editing on SYSTRANS's Rule-Based Translation System," *Proceedings of the Second Workshop on Statistical Machine Translation*, 2nd WSMT, pp. 220-223, Prague, Czech Republic, June 2007.

[8] Isabelle, P., G. Goutter, M. Simard, "Domain Adaptation of MT Systems through Automatic Post-Editing," *Proceedings of MT Summit XI*, pp. 255-261, Copenhagen, Denmark, 10-14 Sept. 2007.

[9] Knight, Kevin, and Ishwar Chander, "Automated Post-Editing of Documents," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 779-784, CA, USA, 1994.

[10] Lee, J., "Automatic Article Restoration," in *Proc. HLT-NAACL 2004 Student Research Workshop*, Boston, MA, 195-200, May, 2004.

[11] Llitjós, Ariadna Fontós, and Jaime Carbonell, "Automating Post-Editing to Improve MT Systems," in *Automated Post-Editing Workshop*, AMTA, Boston, USA, August 12, 2006.

[12] Och, Franz Josef, Christoph Tillmann, and Hermann Ney, "Improved Alignment Models for Statistical Machine Translation," in *Proc. EMNLP/WVLC*, 1999.

[13] Och, Franz Josef and Hermann Ney, “A comparison of alignment models for statistical machine translation.” In *Proc. COLING ’00: The 18th International Conference on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany, August, 2000.

[14] Och, Franz Josef and Hermann Ney, “Improved statistical alignment models.” In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, 2000.

[15] Och, Franz Josef and Hermann Ney, “The alignment template approach to statistical machine translation.” *Computational Linguistics*, 30:417–449, 2004.

[16] Papineni, K., S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” In *Proceedings of ACL-2002*, 40th Annual Meeting of the Association for Computational Linguistics pp. 311—318, 2002.

[17] Shen, Wade, Brian Delaney and Tim Anderson, “The MIT-LL/AFRL IWSLT-2006 MT System,” *Proc. of the International Workshop on Spoken Language Translation (IWSLT) 2006*, pp. 71-76, Kyoto, Japan, 27 November 2006.

[18] Shia, Min-Shiang, *Using Phrase Structure and Fluency to Improve Statistical Machine Translation*, Master Thesis, Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC, June, 2006.

[19] Shih, Shu-Fan, *A Query Augmentation Model for Answering Well-Defined Questions*, Master Thesis, Department of Computer Science and Information Engineering, National Chi Nan University, Taiwan, ROC, July, 2007.

[20] Simard, M., G. Goutter, P. Isabelle, “Statistical Phrase-Based Post-Editing”. *Proceedings of NAACL-HLT 2007*, pp. 508-515, Rochester, NY, April 2007.

[21] Simard, M., N. Ueffing, P. Isabelle, R. Kuhn, “Rule-Based Translation with Statistical Phrase-Based Post-Editing”. *Proceedings of the Second Workshop on Statistical Machine Translation*, 2nd WSMT, pp. 203-206, Prague, Czech Republic, June 2007.

[22] Zhou, Yu, Chengqing Zong, and Bo Xu, “Bilingual Chunk Alignment in Statistical Machine Translation,” In *Proceedings of IEEE International Conference on Systems, Man & Cybernetics (SMCC2004)*, Hague, Netherlands, 2004.