An Empirical Study in Source Word Deletion for Phrase-based Statistical Machine Translation

Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou
Microsoft Research Asia
Beijing, China
chl, dozhang@microsoft.com
muli, mingzhou@microsoft.com

Hailei Zhang
Northeastern University of China
Shenyang, China
hailei.zh@gmail.com

Abstract

The treatment of 'spurious' words of source language is an important problem but often ignored in the discussion on phrase-based SMT. This paper explains why it is important and why it is not a trivial problem, and proposes three models to handle spurious source words. Experiments show that any source word deletion model can improve a phrase-based system by at least 1.6 BLEU points and the most sophisticated model improves by nearly 2 BLEU points. This paper also explores the impact of training data size and training data domain/genre on source word deletion.

1 Introduction

It is widely known that translation is by no means word-to-word conversion. Not only because sometimes a word in some language translates as more than one word in another language, also every language has some 'spurious' words which do not have any counterpart in other languages. Consequently, an MT system should be able to identify the spurious words of the source language and not translate them, as well as to generate the spurious words of the target language. This paper focuses on the first task and studies how it can be handled in phrase-based SMT.

An immediate reaction to the proposal of investigating source word deletion (henceforth SWD) is: Is SWD itself worth our attention? Isn't it a trivial task that can be handled easily by existing techniques? One of the reasons why we need to pay attention to SWD is its significant improvement to translation performance, which will be shown by the experiments results in section 4.2. Another reason is that SWD is not a trivial task. While some researchers think that the spurious words of a language are merely function words or grammatical particles, which can be handled by some simple heuristics or statistical means, there are in fact some tricky cases of SWD which need sophisticated solution. Consider the following example in Chinese-to-English translation: in English we have the subordinate clause "according to NP", where NP refers to some source of information. The Chinese equivalent of this clause can sometimes be "ACCORDING-TO/根据 NP EXPRESS/表示"; that is, in Chinese we could have a clause rather than a noun phrase following the preposition ACCORDING-TO/根据. Therefore, when translating Chinese into English, the content word EXPRESS/表示 should be considered spurious and not to be translated. Of course, the verb EXPRESS/表示 is not spurious in other contexts. It is an example that SWD is not only about a few function words, and that the solution to SWD has to take context-sensitive factors into account. Moreover, the solution needed for such tricky cases seems to be beyond the scope of current phrase-based SMT, unless we have a very large amount of training data which covers all possible variations of the Chinese pattern "ACCORDING-TO/根据 NP EXPRESS/表示".

Despite the obvious need for handling spurious source words, it is surprising that phrase-based SMT, which is a major approach to SMT, does not well address the problem. There are two possible ways for a phrase-based system to deal with SWD. The first one is to allow a source

Proceedings of the Third Workshop on Statistical Machine Translation, pages 1–8,
Columbus, June 2008.

language phrase to translate to nothing. However, no existing literature has mentioned such a possibility and discussed the modifications required by such an extension. The second way is to capture SWD within the phrase pairs in translation table. That is, suppose there is a foreign phrase $\tilde{F} = (f_A f_B f_C)$ and an English phrase $\tilde{E} = (e_A e_C)$, where f_A is aligned to e_A and f_C to e_C, then the phrase pair $(\tilde{F}, \tilde{E})$ tacitly deletes the spurious word f_B. Such a SWD mechanism fails when data sparseness becomes a problem. If the training data does not have any word sequence containing f_B, then the spurious f_B cannot associate with other words to form a phrase pair, and therefore cannot be deleted tacitly in some phrase pair. Rather, the decoder can only give a phrase segmentation that treats f_B itself as a phrase, and this phrase cannot translate into nothing, as far as the SMT training and decoding procedure reported by existing literature are used. In sum, the current mechanism of phrase-based SMT is not capable of handling all cases of SWD.

In this paper, we will present, in section 3, three SWD models and elaborate how to apply each of them to phrase-based SMT. Experiment settings are described in section 4.1, followed by the report and analysis of experiment results, using BLEU as evaluation metric, in section 4.2, which also discusses the impact of training data size and training data domain on SWD models. Before making our conclusions, the effect of SWD on another evaluation metric, viz. METEOR, is examined in section 5.

2 Literature Review

Research work in SMT seldom treats SWD as a problem separated from other factors in translation. However, it can be found in different SMT paradigms the mechanism of handling SWD. As to the pioneering IBM word-based SMT models (Brown et al., 1990), IBM models 3, 4 and 5 handle spurious source words by considering them as corresponding to a particular EMPTY word token on the English side, and by the fertility model which allows the English EMPTY to generate a certain number of foreign words.

As to the hierarchical phrase-based approach (Chiang, 2007), its hierarchical rules are more powerful in SWD than the phrase pairs in conventional phrase-based approach. For instance, the "ACCORDING-TO/根据 NP EXPRESS/表示" example in the last section can be handled easily by the hierarchical rule

$$X \rightarrow < \text{根据 X 表示, according to X} > .$$

In general, if the deletion of a source word depends on some context cues, then the hierarchical approach is, at least in principle, capable of handling it correctly. However, it is still confronted by the same problem as the conventional phrase-based approach regarding those words whose 'spuriousness' does not depend on any context.

3 Source Word Deletion Models

This section presents a number of solutions to the problem of SWD. These solutions share the same property that a specific empty symbol ϵ on the target language side is posited and any source word is allowed to translate into ϵ. This symbol is invisible in every module of the decoder except the translation model. That is, ϵ is not counted when calculating language model score, word penalty and any other feature values, and it is omitted in the final output of the decoder. It is only used to delete spurious source words and refine translation model scores accordingly.

It must be noted that in our approach phrases comprising more than one source word are not allowed to translate into ϵ. This constraint is based on our subjective evaluation of alignment matrix, which indicates that the un-alignment of a continuous sequence of two or more source words is far less accurate than the un-alignment of a single source word lying within aligned neighbors. Consequently, in order to treat a source word as spurious, the decoder must give a phrase segmentation that treats the word itself as a phrase.

Another important modification to the phrase-based architecture is a new feature added to the log-linear model. The new feature, ϵ-penalty, represents how many source words translate into ϵ. The purpose of this feature is the same as that of the feature of word penalty. As many features used in the log-linear model have values of logarithm of probability, candidate translations with more words have, in general, lower scores, and

Model 1	$P(\epsilon)$
Model 2	$P(\epsilon\|f)$
Model 3	$P_{CRF}(\epsilon\|\vec{F}(f)$

Table 1: Summary of the Three SWD Models

therefore the decoder has a bias towards shorter translations. Word penalty (in fact, it should be renamed as word *reward*) is used to neutralize this bias. Similarly, the more source words translate into ϵ, the shorter the translation will be, and therefore the higher score the translation will have. The ϵ-penalty is proposed to neutralize the bias towards shorter translations.

The core of the solutions is the SWD model, which calculates $P(\epsilon|f)$, the probability distribution of translating some source word f to ϵ. Three SWD models will be elaborated in the following subsections. They differ from each other by the conditions of the probability distribution, as summarized in Table 1. Model 1 is a uniform probability distribution that does not take the source word f into account. Model 2 is a simple probability distribution conditioned on the lexical form of f only. Model 3 is a more complicated distribution conditioned on a feature vector of f, and the distribution is estimated by the method of Conditional Random Field.

3.1 Model 1: Uniform Probability

The first model assumes a uniform probability of translation to ϵ. This model is inspired by the HMM-based alignment model (Och and Ney, 2000a), which posits a probability P_0 for alignment of some source word to the empty word on the target language side, and weighs all other alignment probabilities by the factor $1 - P_0$. In the same style, SWD model 1 posits a probability $P(\epsilon)$ for the translation of any source word to ϵ. The probabilities of normal phrase pairs should be weighed accordingly. For a source phrase containing only one word, its weight is simply $P(\bar{\epsilon}) = 1 - P(\epsilon)$. As to a source phrase containing more than one word, it implies that every word in the phrase does not translate into ϵ, and therefore the weighing factor $P(\bar{\epsilon})$ should be multiplied as many times as the number of words in the source phrase. In sum, for any phrase pair $< \tilde{F}, \tilde{E} >$, its probability is

$$P(\tilde{E}|\tilde{F}) = \begin{cases} P(\epsilon) & if \tilde{E} = (\epsilon) \\ P(\bar{\epsilon})^{|\tilde{F}|} P_T(\tilde{E}|\tilde{F}) & otherwise \end{cases}$$

where $P_T(\tilde{E}|\tilde{F})$ is the probability of the phrase pair as registered in the translation table, and $|\tilde{F}|$ is the length of the phrase $\tilde{F}$. The estimation of $P(\epsilon)$ is done by MLE:

$$P(\epsilon) = \frac{\text{number of unaligned source word tokens}}{\text{number of source word tokens}}.$$

3.2 Model 2: EMPTY as Normal Word

Model 1 assumes that every word is as likely to be spurious as any other word. Definitely this is not a reasonable assumption, since certain function words and grammatical particles are more likely to be spurious than other words. Therefore, in our second SWD model the probability of translating a source word f to ϵ is conditioned on f itself.

This probability, $P(\epsilon|f)$, is in the same form as the probability of a normal phrase pair, $P(\tilde{E}|\tilde{F})$, if we consider ϵ as some special phrase of the target language and f as a source language phrase on its own. Thus $P(\epsilon|f)$ can be estimated and recorded in the same way as the probability of normal phrase pairs. During the phase of phrase enumeration, in addition to enumerating all normal phrase pairs, we also enumerate all unaligned source words f and add phrase pairs of the form $< (f), (\epsilon) >$. These special phrase pairs, TO-EMPTY phrase pairs, are fed to the module of phrase scoring along with the normal phrase pairs. Both types of phrase pairs are then stored in the translation table with corresponding phrase translation probabilities. It can be seen that, since the probabilities of normal phrase pairs are estimated in the same procedure as those of TO-EMPTY phrase pairs, they do not need re-weighing as in the case of SWD model 1.

3.3 Model 3: Context-sensitive Model

Although model 2 is much more informative than model 1, it is still unsatisfactory if we consider the problem of SWD as a problem of *tagging*. The decoder can be conceived as if it carries out a tagging task over the source language sentence: each source word is tagged either as "spurious" or "non-spurious". Under such a perspective, SWD model 2 is merely a unigram tagging model, and it uses only one feature template, viz. the lexical form of the source word in hand. Such a model can by no means encode any contextual information, and therefore it cannot handle the "ACCORDING-TO/根据 NP EXPRESS/表示" example in section 1.

An obvious solution to this limitation is a more powerful tagging model augmented with context-sensitive feature templates. Inspired by research work like (Lafferty et al., 2001) and (Sha and Pereira, 2003), our SWD model 3 uses first-order Conditional Random Field (CRF) to tackle the tagging task.[1] The CRF model uses the following feature templates:

1. the lexical form and the POS of the foreign word f itself;
2. the lexical forms and the POSs of f_{-2}, f_{-1}, f_{+1}, and f_{+2}, where f_{-2} and f_{-1} are the two words to the left of f, and f_{+1} and f_{+2} are the two words to the right of f;
3. the lexical form and the POS of the *head* word of f;
4. the lexical forms and the POSs of the *dependent* words of f.

The lexical forms are the major source of information whereas the POSs are employed to alleviate data sparseness. The neighboring words are used to capture local context information. For example, in Chinese there is often a comma after verbs like "*said*" or "*stated*", and such a comma is not translated to any word or punctuation in English. These spurious commas are therefore identified by their immediate left neighbors. The head and dependent words are employed to capture non-local context information found by some dependency parser. For the "ACCORDING-TO/根据 NP EXPRESS/表示" example in section 1, the Chinese word ACCORDING-TO/根据 is the head word of EXPRESS/表示. The spurious token of EXPRESS/表示 in this pattern can be distinguished from the non-spurious tokens through the feature template of head word.

[1] Maximum Entropy was also tried in our experiments but its performance is not as good as CRF.

The training data for the CRF model comprises the alignment matrices of the bilingual training data for the MT system. A source word (token) in the training data is tagged as "non-spurious" if it is aligned to some target word(s), otherwise it is tagged as "spurious". The sentences in the training data are also POS-tagged and parsed by some dependency parser, so that each word can be assigned values for the POS-based feature templates as well as the feature templates of head word and dependency words.

The trained CRF model can then be used to augment the decoder to tackle the SWD problem. An input source sentence should first be POS-tagged and parsed for assigning feature values. The probability for f being spurious, $P(\epsilon|f)$, is then calculated by the trained CRF model as

$$P_{CRF}(\text{spurious}|\vec{F}(f)).$$

The probability for f being non-spurious is simply $1 - P(\epsilon|f)$. For a normal phrase pair $< \tilde{F}, \tilde{E} >$ recorded in the translation table, its phrase translation probability and the lexical weight should be re-weighed by the probabilities of non-spuriousness. The weighing factor is

$$\prod_{f_i \epsilon \tilde{F}} (1 - P(\epsilon|f_i)),$$

since the translation of $\tilde{F}$ into $\tilde{E}$ means the decoder considers every word in $\tilde{F}$ as non-spurious.

4 Experiments

4.1 Experiment Settings

A series of experiments were run to compare the performance of the three SWD models against the baseline, which is the standard phrase-based approach to SMT as elaborated in (Koehn et al., 2003). The experiments are about Chinese-to-English translation. The bilingual training data is the one for NIST MT-2006. The GIGAWORD corpus is used for training language model. The development/test corpora are based on the test sets for NIST MT-2005/6.

The alignment matrices of the training data are produced by the GIZA++ (Och and Ney, 2000b) word alignment package with its default settings. The subsequent construction of translation table was done in exactly the same way as explained in (Koehn et al., 2003). For SWD model 2, the phrase enumeration step is modified as described in section 3.2. We used the Stanford parser (Klein and Manning, 2003) with its default Chinese grammar for its POS-tagging as well as finding the head/dependent words of all source words. The CRF toolkit used for model 3 is CRF++[2]. The training data for the CRF model should be the same as that for translation table construction. However, since there are too many instances (every single word in the training data is an instance) with a huge feature space, no publicly available CRF toolkit can handle the entire training set of NIST MT-2006.[3] Therefore, we can use at most only about one-third of the NIST training set (comprising the FBIS, B1, and T10 sections) for CRF training.

The decoder in the experiments is our reimplementation of HIERO (Chiang, 2007), augmented with a 5-gram language model and a reordering model based on (Zhang et al., 2007). Note that no hierarchical rule is used with the decoder; the phrase pairs used are still those used in conventional phrase-based SMT. Note also that the decoder does not translate OOV at all even in the baseline case, and thus the SWD models do not improve performance simply by removing OOVs.

In order to test the effect of training data size on the performance of the SWD models, three variations of training data were used:

FBIS Only the FBIS section of the NIST training set is used as training data (for both translation table and the CRF model in model 3). This section constitutes about 10% of the entire NIST training set. The purpose of this variation is to test the performance of each model when very small amount of data are available.

BFT Only the B1, FBIS, and T10 sections of the NIST training set are used as training data. These sections are about one-third of the entire NIST training set. The purpose of this variation is to test each model when medium size of data are available.[4]

NIST All the sections of the NIST training set are used. The purpose of this variation is to test each model when a large amount of data are available.

Data	baseline	model 1	model 2	model 3
FBIS	28.01	29.71	29.48	29.64
BFT	29.82	31.55	31.61	31.75
NIST	29.77	31.39	31.33	31.71

Table 2: BLEU scores in Experiment 1: NIST'05 as dev and NIST'06 as test

(Case-insensitive) BLEU-4 (Papineni et al., 2002) is used as the evaluation metric. In each test in our experiments, maximum BLEU training were run 10 times, and thus there are 10 BLEU scores for the test set. In the following we will report the mean scores only.

4.2 Experiment Results and Analysis

Table 2 shows the results of the first experiment, which uses the NIST MT-2005 test set as development data and the NIST MT-2006 test set as test data. The most obvious observation is that any SWD model achieves much higher BLEU score than the baseline, as there is at least 1.6 BLEU point improvement in each case, and in some case the improvement of using SWD is nearly 2 BLEU points. This clearly proves the importance of SWD in phrase-based SMT.

The difference between the performance of the various SWD models is much smaller. Yet there are still some noticeable facts. The first one is that model 1 gives the best result in the case of using only FBIS as training data but it fails to do so when more training data is available. This phenomenon is not strange since model 2 and model 3 are conditioned on more information and therefore they need more training data.

The second observation is about the strength of SWD model 3, which achieves the best BLEU score in both the BFT and NIST cases. While its improvement over models 1 and 2 is marginal in the case of BFT, its performance in the NIST

[2]http://crfpp.sourceforge.net/

[3]Apart from CRF++, we also tried FLEX-CRF (http://flexcrfs.sourceforge.net) and MALLET (http://mallet.cs.umass.edu).

[4]Note also that the BFT data set is the largest training data that the CRF model in model 3 can handle.

case is remarkable. A suspicion to the strength of model 3 is that in the NIST case both models 1 and 2 use the entire NIST training set for estimating $P(\epsilon)$, while model 3 uses only the BFT sections to train its CRF model. It may be that the BFT sections are more consistent with the test data set than the other NIST sections, and therefore a SWD model trained on BFT sections only is better than that trained on the entire NIST. This conjecture is supported by the fact that in all four settings the BLEU scores in the NIST case are lower than those in the BFT case, which suggests that other NIST sections are noisy. While it is impossible to test model 3 with the entire NIST, it is possible to restrict the data for the estimation of $P(\epsilon|f)$ in model 1 to the BFT sections only and check if such a restriction helps.[5] We estimated the uniform probability $P(\epsilon)$ from only the BFT sections and used it with the translation table constructed from the complete NIST training set. The BLEU score thus obtained is 31.24, which is even lower than the score (31.39) of the original case of using the entire NIST for both translation table and $P(\epsilon|f)$ estimation. In sum, the strength of model 3 is not simply due to the choice of training data.

The test set used in Experiment 1 distinguishes itself from the development data and the training data by its characteristics of combining text from different *genres*. There are three sources of the NIST MT-2006 test set, viz. "newswire", "newsgroup", and "broadcast news", while our development data and the NIST training set comprises only newswire text and text of similar style. It is an interesting question whether SWD only works for some genres (say, newswire) but not for other genres. In fact, it is dubious whether SWD fits the test set to the same extent as it fits the development set. That is, perhaps SWD contributes to the improvement in Experiment 1 simply by improving the translation of the development set which is composed of newswire text only, and SWD may not benefit the translation of the test data at all. In order to test this conjecture, we ran Experiment 2, in which the SWD models were still applied to the development data during training, but all SWD models stopped working when translating the test data with the trained parameters. The results are shown in Table 3. These results are very discouraging if we compare each cell in Table 3 against the corresponding cell in Table 2: in all cases SWD seems harmful to the translation of the test data. It is tempting to accept the conclusion that SWD works for newswire text only.

Data	model 1	model 2	model 3
FBIS	29.85	29.91	29.95
BFT	31.73	31.84	32.08
NIST	31.70	31.82	32.05

Table 3: BLEU scores in Experiment 2, which is the same as Experiment 1 but no word is deleted for test corpus. Note: the baseline scores are the same as the baselines in Experiment 1 (Table 2).

To scrutinize the problem, we split up the test data set into two parts, viz. the newswire section and the non-newswire section, and ran experiments separately. Table 4 shows the results of Experiment 3, in which the development data is still the NIST MT-2005 test set and the test data is the newswire section of NIST MT-2006 test set. It is confirmed that if test data shares the same genre as the training/development data, then SWD does improve translation performance a lot. It is also observed that more sophisticated SWD models perform better when provided with sufficient training data, and that model 3 exhibits remarkable improvement when it comes to the NIST case.

Of course, the figures in Table 5, which shows the results of Experiment 4 where the non-newswire section of NIST MT-2006 test set is used as test data, still leave us the doubt that SWD is useful for a particular genre only. After all, it is reasonable to assume that a model trained from data of a particular domain can give good performance only to data of the same domain. On the other hand, the language model is another cause of the poor performance, as the GIGAWORD corpus is also of the newswire style.

While we cannot prove the value of SWD with respect to training data of other genres in the mean time, we could test the effect of using development data of other genres. In our last experiment, the first halves of both the newswire

[5] Unfortunately this way does not work for model 2 as the estimation of $P(\epsilon|f)$ and the construction of translation table are tied together.

	apply SWD for test set			no SWD for test set		
Data	model 1	model 2	model 3	model 1	model 2	model 3
FBIS	30.81	30.81	30.68	29.23	29.61	29.46
BFT	33.57	33.74	33.71	31.88	31.87	32.25
NIST	33.65	34.01	34.42	32.14	32.59	32.87

Table 4: BLEU scores in Experiment 3, which is the same as Experiments 1 and 2 but only the **newswire** section of NIST'06 test set is used. Note: the baseline scores are the same as the baselines in Experiment 1 (Table 2).

	apply SWD for test set			no SWD for test set		
Data	model 1	model 2	model 3	model 1	model 2	model 3
FBIS	29.19	28.86	29.16	30.07	29.67	30.08
BFT	30.62	30.64	30.86	31.66	31.83	32.00
NIST	30.34	30.10	30.46	31.50	31.45	31.66

Table 5: BLEU scores in Experiment 4, which is the same as Experiments 1 and 2 but only the **non-newswire** section of NIST'06 test set is used. Note: the baseline scores are the same as the baselines in Experiment 1 (Table 2).

Data	baseline	model 1	model 2	model 3
FBIS	26.87	27.79	27.51	27.61
BFT	29.11	30.38	30.49	30.41
NIST	29.34	30.63	30.95	31.00

Table 6: BLEU scores in Experiment 5: which is the same as Experiment 1 but uses half of NIST'06 as development set and another half of NIST'06 as test set.

and non-newswire sections of NIST MT-2006 test set are combined to form the new development data, and the second halves of the two sections are combined to form the new test data. The new development data is therefore consistent with the new test data. If SWD, or at least a SWD model from newswire, is harmful to the non-newswire section, which constitutes about 60% of the development/test data, then it will be either that the parameter training process minimizes the impact of SWD, or that the SWD model will make the parameter training process fail to search for good parameter values. The consequence of either case is that the baseline setting should produce similar or even higher BLEU score than the settings that employ some SWD model. Experiment results, as shown in Table 6, illustrate that SWD is still very useful even when both development and test sets contain texts of different genres from the training text. It is also observed, however, that the three SWD models give rise to roughly the same BLEU scores, indicating that the SWD training data do not fit the test/development data very well as even the more sophisticated models are not benefited from more data.

5 Experiments using METEOR

The results in the last section are all evaluated using the BLEU metric only. It is dubious whether SWD is useful regarding recall-oriented metrics like METEOR (Banerjee and Lavie, 2005), since SWD removes information in source sentences. This suspicion is to certain extent confirmed by our application of METEOR to the translation outputs of Experiment 1 (c.f. Table 7), which shows that all SWD models achieve lower METEOR scores than the baseline. However, SWD is not entirely harmful to METEOR: if SWD is applied to parameter tuning only but not for the test set, (i.e. Experiment 2), even higher METEOR scores can be obtained. This puzzling observation may be because the parameters of the decoder are optimized with respect to BLEU score, and SWD benefits parameter tuning by improving BLEU score. In future experiments, maximum METEOR training should be used instead of maximum BLEU training so as to examine if SWD is really useful for parameter tuning.

	Experiment 1				Experiment 2		
	SWD for both dev/test				SWD for dev only		
Data	baseline	model 1	model 2	model 3	model 1	model 2	model 3
FBIS	50.07	47.90	49.83	49.34	51.58	51.08	51.17
BFT	52.47	50.55	51.89	52.10	54.72	54.43	54.30
NIST	52.12	49.86	50.97	51.59	54.14	53.82	54.01

Table 7: METEOR scores in Experiments 1 and 2

6 Conclusion and Future Work

In this paper, we have explained why the handling of spurious source words is not a trivial problem and how important it is. Three solutions, with increasing sophistication, to the problem of SWD are presented. Experiment results show that, in our setting of using NIST MT-2006 test set, any SWD model leads to an improvement of at least 1.6 BLEU points, and SWD model 3, which makes use of contextual information, can improve up to nearly 2 BLEU points. If only the newswire section of the test set is considered, SWD model 3 is even more superior to the other two SWD models.

The effect of training data size on SWD has also been examined, and it is found that more sophisticated SWD models do not outperform unless they are provided with sufficient amount of data. As to the effect of training data domain/genre on SWD, it is clear that SWD models trained on text of certain genre perform the best when applied to text of the same genre. While it is infeasible for the time being to test if SWD works well for non-newswire style of training data, we managed to illustrate that SWD based on newswire text still to certain extent benefits the training and translation of non-newswire text.

In future, two extensions of our system are needed for further examination of SWD. The first one is already mentioned in the last section: maximum METEOR training should be implemented in order to fully test the effect of SWD regarding METEOR. The second extension is about the weighing factor in models 1 and 3. The current implementation assumes that all source words in a normal phrase pair need to be weighed by $1 - P(\epsilon)$. However, in fact some source words in a source phrase are tacitly deleted (as explained in the Introduction). Thus the word alignment information within phrase pairs need to be recorded and the weighing of a normal phrase pair should be done in accordance with such alignment information.

References

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. 1990. A Statistical Approach to Machine Translation *Computational Linguistics, 16(2)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of Workshop on Evaluation Measures for MT and/or Summarization at ACL 2005*.

David Chiang. 2007. Hierarchical Phrase-based Translation. *Computational Linguistics, 33(2)*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings for ACL 2003*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. *Proceedings for HLT-NAACL 2003*.

John Lafferty, Andrew McCallum, and Fernando Pereira 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings for 18th International Conf. on Machine Learning*.

Franz J. Och, and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. *Proceedings of COLING 2000*.

Franz J. Och, and Hermann Ney. 2000. Improved Statistical Alignment Models. *Proceedings for ACL 2000*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings for ACL 2002*.

Fei Sha, Fernando Pereira. 2003. Shallow parsing with conditional random fields. *Proceedings of NAACL 2003*.

Dongdong Zhang, Mu Li, Chi-Ho Li and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. *Proceedings for EMNLP 2007*.