Example-based Rescoring of Statistical Machine Translation Output

Michael Paul[⋆†]**, Eiichiro Sumita**[⋆] **and Seiichi Yamamoto**[⋆†]

[⋆]ATR Spoken Language Translation Labs
Keihanna Science City
619-0288 Kyoto, Japan

[†]Kobe University
Graduate School of Science and Technology
657-8501 Kobe, Japan

{ Michael.Paul , Eiichiro.Sumita , Seiichi.Yamamoto }@atr.jp

Abstract

Conventional statistical machine translation (SMT) approaches might not be able to find a good translation due to problems in its statistical models (due to data sparseness during the estimation of the model parameters) as well as search errors during the decoding process. This paper[1] presents an example-based rescoring method that validates SMT translation candidates and judges whether the selected decoder output is good or not. Given such a validation filter, defective translations can be rejected. The experiments show a drastic improvement in the overall system performance compared to translation selection methods based on statistical scores only.

1 Introduction

The statistical machine translation framework (SMT) formulates the problem of translating a sentence from a source language S into a target language T as the maximization problem of the conditional probability:

$$\textbf{TM}\cdot\textbf{LM} = \text{argmax}_T \; p(S|T) * p(T), \qquad (1)$$

where $p(S|T)$ is called a *translation model* (TM), representing the generation probability from T into S, $p(T)$ is called a *language model* (LM) and represents the likelihood of the target language (Brown et al., 1993). The TM and LM probabilities are trained automatically from a parallel text corpus (*parameter estimation*). They represent the general translation knowledge used to map a sequence of words from the source language into the target language. During the translation process (*decoding*) a statistical score based on the probabilities of the translation and the language models is assigned to each translation candidate and the one with the highest TM·LM score is selected as the translation output.

However, the system might not be able to find a good translation due to parameter estimation problems of the statistical models (due to data sparseness during the estimation of the model probabilities) and search errors during the translation process. Moreover, conventional SMT approaches use words as the translation unit. Therefore, the optimization is carried out locally generating the translation word-by-word.

In the framework of example-based machine translation (EBMT), however, a parallel text corpus is used directly to obtain the translation (Nagao, 1984). Given an input sentence, translation examples from the corpus that are best matched to the input are retrieved and adjusted to obtain the translation. Thus the translation unit used in EBMT approaches is a complete sentence, providing a larger context for the generation of an appropriate translation. However, this approach requires appropriate translation examples to achieve an accurate translation.

A combination of statistical and example-based MT approaches shows some promising perspectives for overcoming the shortcomes of each approach. In this paper, we propose an example-based rescoring method (EBRS) for selecting translation candidates generated by a statistical decoder, as illustrated in Figure 1.

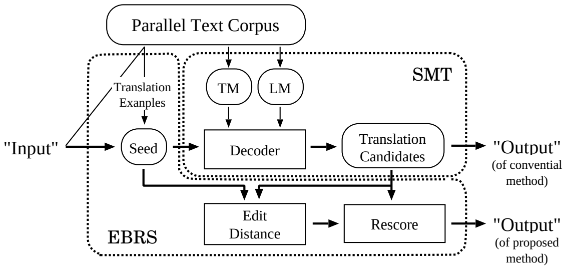

Figure 1: Outline

It retrieves translation examples that are similar to the input from a parallel text corpus (cf. Section 2). The target parts of these examples (*seed*) paired with the input form the input of a statistical decoder (cf. Section 3). The statistical scores of each generated translation candidate are rescored using information about how much the seed sentence is modified during decoding. It measures the distance between the word sequences of the decoder output and its seed sentence based on the costs of *edit distance* operations (cf. Section 4). We combine the distance measure with the statistical scores of the SMT engine, resulting in a reliability measure to identify modeling problems in statistically optimized translation candidates and to reject inappropriate solutions (cf. Section 5).

[1] The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, *"A study of speech dialogue translation technology based on a large corpus"*.

2 Translation Example Retrieval

Translation examples consist of pairs of pre-translated sentences, either by humans (high quality) or automatically using MT systems (reduced quality). A collection of translation examples can be used directly to obtain a translation of a given input sentence. The similarity of the input to the source part of the translation examples enables us to identify translation candidates that might be close to the actual translation.

A common approach to measure the distance between sequences of words is the *edit distance* criteria (Wagner, 1974). The distance is defined as the sum of the costs of *insertion* (INS), *deletion* (DEL), and *substitution* (SUB) operations required to map one word sequence into the other. The edit distance can be calculated by a standard *dynamic programming* technique.

$$\text{ED}(s_1,s_2) = |\text{INS}| + |\text{DEL}| + |\text{SUB}|$$

An extension of the edit-distance-based retrieval method is presented in (Watanabe and Sumita, 2003). It incorporates the *tf·idf* criteria as seen in the information retrieval framework by treating each translation example as a document. For each word of the input, its term frequency $\text{tf}_{i,j}$ is combined with its document frequency df_i into a single weight $w_{i,j}$, which is used to select the most relevant ones out of N documents (= example targets).

Another possibility for obtaining translation examples is simply to utilize available (off-the-shelf) MT systems by pairing the input sentence with the obtained MT output. However, the quality of those translation examples might be much lower than manually created translations.

3 Statistical Decoding

(Germann et al., 2001) presents a greedy approach to search for the translation that is most likely according to previously learned statitistical models. An extension of this approach that can take advantage of translation examples provided for a given input sentence is proposed in (Watanabe and Sumita, 2003). Instead of decoding and generating an output string word-by-word as is done in the basic concept, this greedy approach slightly modifies the target part of the translation examples so that the pair becomes the actual translation.

The advantage of the example-based approach is that the search for a good translation starts from the retrieved translation example, not a guessed translation resulting in fewer search errors. However, since it uses the same greedy search algorithm as the basic method, search errors cannot be avoided completely. Furthermore, the parameter estimation problem still remains.

The experiment discussed in Section 5.1 indeed shows a large degradation in the system performance when the greedy decoder is applied to already perfect translations, indicating that the decoder may modify translations wrongly based on its statistical models (IBM model 4).

4 Example-based Rescoring

Therefore we have to validate the quality of translation candidates selected by the decoder and judge whether problems in the SMT models or search errors resulted in an inaccurate translation or not.

Our approach extends the example-based concept of (Watanabe and Sumita, 2003). It compares the decoder output with the *seed* sentence, i.e., the target part of the translation example that forms the input of the decoder. Given a translation example whose source part is quite similar to the input, we can assume that the fewer the modifications that are necessary to alter the corresponding example target to the translation candidate during decoding, the less likely it is that there will be a problem in the statistical models.

The decision on translation quality is based on the edit distance criteria, as introduced in Section 2. For each translation candidate, we measure the edit distance between the word sequence of the decoder output and the seed sentence. The proposed method rescores the translation candidates of the SMT decoder by combining the statistical probabilities of the translation and language models with the example-based translation quality hypothesis and selects the translation candidate with the highest revised score as the translation output.

The rescoring function $rescore$ has to be designed in such a way that almost unaltered translation candidates with good translation and language model scores are preferred over those with the highest statistical scores that required lots of modifications to the seed sentence.

For the experiments described below we defined two different rescoring functions. First, the edit distance of the seed sentence s_d and the decoder output d is used as a weight to decrease the statistical scores. The larger the edit distance score, the smaller the revised score of the respective translation candidate. The scaling factor *scale* depends on the utilized corpus and can be optimized on a development set reserved for parameter tuning.

$$\mathbf{TM}\cdot\mathbf{LM}\cdot\mathbf{ED_W}(d) = \frac{\text{TM}\cdot\text{LM}(d)}{\exp(\ scale * \text{ED}(s_d,d)\ )} \tag{2}$$

The second rescoring function assigns a probability to each decoder output that combines the exponential of the sum of log probabilities of TM and LM and the scaled negative ED scores of all translation candidates TC as follows.

$$\mathbf{TM}\cdot\mathbf{LM}\cdot\mathbf{ED_P}(d) = \tag{3}$$

$$\frac{\exp(\log \text{TM}(d) + \log \text{LM}(d) - scale * \text{ED}(s_d,d))}{\sum\limits_{(s_{tc},tc)\in TC} \exp(\log \text{TM}(tc) + \log \text{LM}(tc) - scale * \text{ED}(s_{tc},tc))}$$

5 Evaluation

The evaluation of our approach is carried out using a collection of Japanese sentences and their English translations that are commonly found in phrasebooks for tourists going abroad (Takezawa et al., 2002). The *Basic Travel Expression Corpus* (BTEC) contains 157K sentence pairs and the average lengths in words of Japanese and English sentences are 7.7 and 5.5, respectively. The corpus was split randomly into three parts for training (155K), parameter tuning (10K), and evaluation (10K) purposes. The experiments described below were carried out on 510 sentences selected randomly as the test set.

For the evaluation, we used the following automatic scoring measures and human assessment.

- Word Error Rate (WER), which penalizes the edit distance against reference translations (Su et al., 1992)
- BLEU: the geometric mean of n-gram precision for the translation results found in reference translations (Papineni et al., 2002)
- Translation Accuracy (ACC): subjective evaluation ranks ranging from A to D (A: perfect, B: fair, C: acceptable and D: nonsense), judged blindly by a native speaker (Sumita et al., 1999)

In contrast to WER, higher BLEU and ACC scores indicate better translations. For the automatic scoring measures we utilized up to 16 human reference translations.

5.1 Downgrading Effects During Decoding

In order to get an idea about how much degradation is to be expected in the translation candidates modified by the statistical decoder, we conducted an experiment using the reference translations of the test set as the input of the example-based decoder. These seed sentences are already accurate translations, thus simulating the "optimal" translation example retrieval case resulting in an upper boundary of the statistical decoder performance.

Table 1: Downgrading Effects During Decoding

scoring	automatic		subjective (ACC)			
scheme	WER	BLEU	A	A+B	A+B+C	gain
TM·LM	0.255	0.744	0.660	0.790	0.854	–
TM·LM·ED$_{\rm P}$	0.179	0.814	0.745	0.854	0.898	0.044
TM·LM·ED$_{\rm W}$	0.010	0.984	0.903	0.968	0.982	0.128

The results summarized in Table 1 show a large degradation (WER=25.5%, BLEU=0.744) in the reference translations when modified by the statistical decoder (TM·LM). Only 66.0% of the decoder output are still perfect and 14.6% even result in unacceptable translations. The rescoring function TM·LM·ED$_{\rm P}$ enables us to recover some of the decoder problems gaining 4.4% in accuracy compared to the statistical decoder. The best performance is achieved by the weight-based rescoring function TM·LM·ED$_{\rm W}$. However, around 10% of the selected translations are not yet perfect.

5.2 Baseline Comparison

In the second experiment, we used two types of retrieval methods (*tf·idf*-based, MT-based), as introduced in Section 2, and compared the results with the baseline system TM·LM, i.e., the example-based decoding approach of (Watanabe and Sumita, 2003) using the *tf·idf* criteria for the retrieval of translation examples and only the statistical scores for the selection of the translation.

For the MT-based retrieval method we used eight machine translation systems for Japanese-to-English. Three of them were in-house EBMT systems which differ in the translation unit (sentence-based vs. phrase-based). They were trained on the same corpus as the statistical decoder. The remaining five systems were (off-the-shelf) general-purpose translation engines with quite different levels of performance (cf. Table 2).

Table 2: MT System Performance

	MT$_1$	MT$_2$	MT$_3$	MT$_4$	MT$_5$	MT$_6$	MT$_7$	MT$_8$
WER	0.320	0.408	0.419	0.580	0.584	0.588	0.600	0.646
BLEU	0.604	0.489	0.424	0.222	0.252	0.237	0.205	0.200

The results of our experiments are summarized in Table 3. The baseline system TM·LM seems to work best when used in combination with the *tf·idf*-based retrieval method, achieving around 80% translation accuracy. Moderate improvements of around 2% can be seen when the proposed rescoring functions are used together with the seed sentences obtained for the baseline system.

However, the largest gain in performance is achieved when the decoder is applied to the output of multiple machine translation systems and the translation is selected using the weight-based rescoring function.

Table 3: Baseline Comparison

tf·idf-based	automatic		subjective (ACC)			
retrieval	WER	BLEU	A	A+B	A+B+C	gain
TM·LM	**0.313**	**0.655**	**0.629**	**0.743**	**0.808**	–
TM·LM·ED$_{\rm P}$	0.297	0.668	0.668	0.766	0.823	0.015
TM·LM·ED$_{\rm W}$	0.289	0.639	0.676	0.749	0.815	0.007

MT-based	automatic		subjective (ACC)			
retrieval	WER	BLEU	A	A+B	A+B+C	gain
TM·LM	0.338	0.630	0.627	0.731	0.796	-0.012
TM·LM·ED$_{\rm P}$	0.292	0.673	0.719	0.811	0.854	0.046
TM·LM·ED$_{\rm W}$	**0.272**	**0.661**	**0.809**	**0.890**	**0.927**	**0.119**

Table 4 compares the evaluation results of the baseline and the TM·LM·ED$_{\rm W}$ system. 67.5% of the translations are assigned to the same rank, out of which 29.2% of the translations are identical. TM·LM·ED$_{\rm W}$ achieves higher grades for 27% of the sentences, whereas 5.5% of the baseline system translations are better. In total, the translation accuracy improved by 11.9% to 92.7%. Examples of differing translation ratings are given in Table 5.

One of the reasons for the improved performance is

Table 4: Change in Translation Accuracy

		TM·LM·ED$_W$				
		A	B	C	D	Σ
TM·LM	A	0.592	0.012	0.015	0.010	0.629
	B	0.080	0.024	0.004	0.006	0.114
	C	0.035	0.012	0.010	0.008	0.065
	D	0.102	0.033	0.008	0.049	0.192
	Σ	0.809	0.081	0.037	0.073	

Table 5: Translation Examples

input:	Zutsuu ga shimasu asupirin wa arimasu ka
TM·LM	[D] aspirin do i have a headache
TM·LM·ED$_W$	[A] i have a headache do you have any aspirin
input:	kore wa nani de dekiteimasu ka
TM·LM	[C] what is this made
TM·LM·ED$_W$	[A] what is this made of
input:	nanjikan no okure ni narimasu ka
TM·LM	[B] how many hours are we behind schedule
TM·LM·ED$_W$	[A] how many hours are we delayed
input:	watashi wa waruku arimasen
TM·LM	[A] it ’s not my fault
TM·LM·ED$_W$	[B] I ’m not bad
input:	omedetou onnanoko ga umareta sou desu ne
TM·LM	[A] i hear you had a baby girl congratulations
TM·LM·ED$_W$	[C] congratulations i heard you were born a boy or a girl
input:	ima me o akete mo ii desu ka
TM·LM	[A] is it all right to open my eyes now
TM·LM·ED$_W$	[D] do you mind opening the eye

that the seed sentences obtained by the *tf·idf*-based retrieval method are not translations of the input sentence.

Moreover, the translations of the MT-based retrieval method cover a large variation of expressions due to different MT output styles, whereby the reduced quality of these seed sentences seems to be successfully compensated by the statistical models. In contrast, the translation examples retrieved by the *tf·idf*-based method are quite similar to each other. Thus, local optimization might result in the same decoder output.

In addition, the statistical decoder has the tendency to select shorter translations (4.8 words/sentence for TM·LM and 5.5 words/sentence for TM·LM·ED$_W$, which might indicate some problems in the utilized translation models as well as the language model.

(Watanabe and Sumita, 2003) try to overcome these problems by skipping the decoding process of seed sentences whose *tf·idf*-score indicates an *exact match* and output the obtained seed sentence instead. However, this shortcut method (WER=0.295, BLEU=0.641, ACC=0.898) is out-performed by the proposed rescoring method by 2.9% in translation accuracy, because our method takes advantage of translations successfully modified by the decoder and is able to identify and reject wrongly modified ones.

Moreover, the rescoring function is language-independent and thus can be easily applied to other language-pairs as well.

6 Conclusion

In this paper, we proposed an example-based method for selecting translation candidates generated by a statistical decoder. It utilizes translation examples that are similar to the source sentence as the input and validates the decoder output against its seed sentences in order to identify defective translations. The revised scoring scheme achieved a translation accuracy of 92.7%, an improvement of 11.9% over the baseline system.

So far, we treated the statistical decoder as a black-box. However, further investigations will have to separate *modeling errors* and *search errors* during decoding and compare our findings to advanced statistical modeling approaches (*phrase-based*) and other search strategies. Future work will also focus on the integration of the proposed rescoring formula in the decoding process.

References

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proc. of ACL 2001*, Toulouse, France.

M. Nagao. 1984. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*. A. Elithorn and R. Banerji (eds), Artificial and Human Intelligence, Amsterdam, North-Holland.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, USA.

K. Su, M. Wu, and J. Chang. 1992. A new quantitative quality measure for machine translation systems. In *Proc. of the 14th COLING*, pages 433–439, Nantes, France.

E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. 1999. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *Proc. of the Machine Translation Summit VII*, pages 229–235, Singapore.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd LREC*, pages 147–152, Las Palmas, Spain.

R.W. Wagner. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):169–173.

T. Watanabe and E. Sumita. 2003. Example-based decoding for statistical machine translation. In *Proc. of the Machine Translation Summit IX*, pages 410–417, New Orleans, USA.