Repairing Incorrect Translation with Examples

Junguo Zhu, Muyun Yang, Sheng Li, Tiejun Zhao
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
{ymy, jgzhu}@mtlab.hit.edu.cn; {lisheng,tjzhao}@hit.edu.cn

Abstract

This paper proposes an example driven approach to improve the quality of MT system outputs. Specifically, We extend the system combination method in SMT to combine the examples by two strategies: 1) estimating the confidence of examples by the similarity between source input and the source part of examples; 2) approximating target word posterior probability by the word alignments of the bilingual examples. Experimental results show a significant improvement of 0.64 BLEU score as compared to one online translation service (Google Translate).

1 Introduction

Statistical Machine Translation (SMT), state-of-the-art solution, has remarkable success with the support of the large-scale bilingual corpora to boost the translation quality at present. However, due to the long tail effect of human language, statistical anomalies in the training data can cause that tons of desired translation knowledge could not be statistically learned from the large-scale bilingual corpora. As a result, bulks of the specific translation requirements not well addressed still perplex machine translation academia and industry.

Combining the examples with machine translations output is a good solution to improve translation quality for this issue. Several methods have been proposed in recent years. One approach tries to replace relevant chunks, taking advantage of Translation Memory (TM). Its motivation is to store and to retrieve similar translation examples for a given input, then to avail of examples to replace the similar chunks into the input by the threshold of similar score (Smith and Clark, 2009; Koehn and Senellart,2010) or by the decision of an automatic classifier (He at al., 2010; Ma et al., 2011). Another approach tries to enhance phrase table of SMT, integrating collected bilingual pairs into the phrase table (Biçici and Clark, 2009; Simardand Isabelle, 2010; DauméIII and Jagarlamudi, 2011).

Different to the above studies in which the EBMT and SMT function in a pipeline style, the work in this paper tries to integrate the SMT results and translation examples in a unified framework. In parallel to the system combination in SMT, we try to integrate the translation examples into the confusion network, allowing each word in both SMT results and examples to compete for the optimal output. In order to achieve the goal, the proposed method introduces some new features to bridge the statistical and example translation.

This paper presents an approach to repairing the translation errors via retrieving translation examples from examples corpus. The effectiveness of our method is validated on the standard test set of Olympics task in IWSLT 2012 Evaluation Campaign. Experimental results show that an absolute increase of 0.64 BLEU score is observed after repairing original translations. This significant improvement suggests the proposed strategy as a promising solution to the subtle task of integrating example knowledge into statistical model outputs, as well as a practical way to boost current MT service.

2 Repairing Translations with Improved Confusion Network

Repairing translation can be viewed as a process of translation knowledge fusion. As illustrated in Fig.1, the proposed approach consists of following steps. We first obtain an online translation system output E_0 for a given input sentence F. Then we retrieve the top-*n* examples $\{< ex\_F_i, ex\_E_i >) \mid (i \in \{1, 2, \ldots, n\}\}$ from bilingual examples corpus which are most similar to F. Then taking the translation E_0 as the initial skele-

International Joint Conference on Natural Language Processing, pages 967–971,
Nagoya, Japan, 14-18 October 2013.

ton, we construct a confusion network by adding the top-n examples into the skeleton incrementally by the word alignments relation between the current skeleton and the *i-th* example ex_E$_i$. The key step is to estimating the word confidence. In this work, we design a feature based on example confidence and word posterior probability by word alignment of examples. Finally, we decode the confusion network by the classic features used in MT combination and new features via a log-linear model.

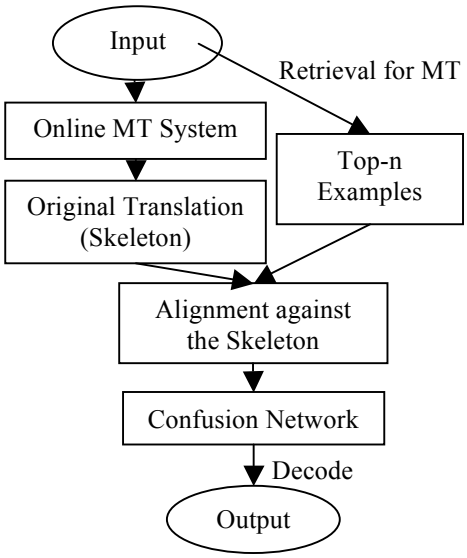

Figure 1. Framework of Repairing Incorrect Translation by Examples

2.1 Estimating Example Confidence

We use a word-based vector space model to retrieve examples from bilingual corpus, by comparing the deviation of angles between the source part of each example vector ex_F$_i$ and the source input vector F. The Cosine Similarity of the vectors is applied to measure the similarity between the source input sentence F and the each pleex_F$_i$, as calculated by:

$$CosSim(ex\_F_i, F) = \frac{ex\_F_i \cdot F}{\|ex\_F_i\| * \|F\|} \quad (1)$$

where ex_F$_i \cdot$ F is the intersection between the vector ex_F$_i$ and the vector F. $\|$ex_F$_i\|$is the norm of the vector ex_F$_i$ and $\|$F$\|$ is the norm of the vector F.To balance the word recall and precision of examples, we filter the examples by simply keeping top-n similar examples for fusion.

Obviously, a reasonable assumption is that the target example has a higher confidence to occur in the final output if the corresponding source part of example has a higher similarity with the source input sentence. Therefore, we estimate the confidence C$_i$ of each target example ex_E$_i$(i $\in$ $\{1, 2, \dots, n\}$)by the Cosine Similarity score between the source part of example ex_F$_i$ and the source input F.

2.2 Estimating Word Posterior Probability by Word Alignment of Examples

To penalize the irrelevant information from examples, we estimate word posterior probability by word alignment between source words and target words in examples. For word alignment between bilingual pairs in SMT, the most popularly used is the IBM model (Brown et al., 1993) in the toolkit GIZA++ (Och et al., 1999), combined with symmetric heuristics.

We estimate the word posterior probability according to word alignments of examples. We create a counter for each word, which might involve in the final translation. The words can come from target parts of examples or the skeleton translation.

For each word, its counter works as follows: 1) Initialize the counter as 0. 2) Keep the counter unchanged if the word either comes from the original translation or does not appear in alignments. 3) Increase the counter by one if its corresponding source word in alignments also appears in the source input sentence. 4) Decrease the counter by one if its corresponding source word in alignments does not disappear in the source input sentence.

Then we estimate posterior probability of the word wfor each fusing translationE$_i$by the counter value as following formula:

$$p(w|E_i) = \frac{1}{1 + e^{-c}} \quad (2)$$

where c is the counter value. The value of E$_i$ is defined as follows:

$$E_i = \begin{cases} ex\_E_i, & i = 1,2..n \\ E_0\ , & i = 0 \end{cases} \quad (3)$$

2.3 Features for Fusion

In the practice of translation fusion under SMT system combination framework, six common features are used to guide the decoding:

Language model: probability from an N-gram language model.

Word penalty: penalty depending on the size (in words) of the hypothesis.

Null-arc penalty: penalty depending on the number of null-arcs crossed in the confusion network to obtain the hypothesis.

N-gram agreement: the value which is equal to the counts of N-gram matches between fusing translations (examples and original translations) and the hypothesis divided by the number of the fusing translations.

N-gram probability: a kind of like language model trained on the top-n examples.

Word confidence: the production of word posterior probability and the confidence of the fusing translation where the word come from.

The practical effect is that the word posterior probability is computed with a simple method at the cost of estimating the word confidence from original translation roughly. To solve this problem, an original word penalty feature is introduced into our method.

Original word penalty: penalty depending on the number of words from the original translation. The feature indicates the degree of repairing original translation.

In addition, although the incremental TER alignment is used in constructing the confusion network to avoid most of alignment errors, overcoming the noise from the examples is critical. So we adopt repetitive word penalty to debilitate this effect.

Repetitive word penalty: penalty depending on the number of repetitive words in the hypothesis.

3 Experiments

3.1 Experimental Settings

Our experiments are carried out on the HIT dataset in the OLYMPICS task of IWSLT 2012 Evaluation Campaign (Federico et al., 2012). We take the training dataset as examples corpus, which contains 52,603 pairs of Chinese-English sentences. Development and test dataset provided by the task contain 2,057 and 998 pairs of Chinese-English sentences, respectively. The Chinese text is segmented by Stanford Word Segmentation (Chang et al., 2008). Detailed statistics of the corpus are shown in Table 1.

	sent	Segment(zh) Token(en)
Example corpus	52,603	495,638 (zh) 527,599 (en)
Dev	2,057	19,457(zh) 20,782(en)
Test	998	10,047(zh) 11,004(en)

Table 1. The Description of HIT dataset

The original MT outputs of develop set and test set come form Google Translate services. The 5-gram English target language model has been trained on Example corpus using SRILM (Stolcke, 2002). The model parameters are trained by MERT (Och, 2003). The 500-best list is created at each MERT iteration and is appended to the n-best lists created at previous iterations. The results are evaluated by BLEU-4 (Papineni et al., 2002) score.

To grasp the distribution of test set on similar score, the composition of test subsets based on similar scores is calculated， which is shown in Table 2.

	Sent	Segment	Segment/Sent
[0.9,1.0)	36	235	6.53
[0.8,0.9)	209	1,394	6.67
[0.7,0.8)	423	3,218	7.61
[0.6,0.7)	720	6,407	8.90
[0.5,0.6)	931	9,121	9.80
[0.4,0.5)	923	9,462	10.25
(0.0,0.4)	570	5,917	10.38

Table 2. Composition of test subsets based on similar scores

3.2 Evaluating Translation Quality

In our experiments, firstly we re-rank the retrieval examples corpus by the Cosine Similarity score and empiricaly retrieve the top-15 similar examples for each source sentence in development and test dataset. Secondly, using the Google Translate services to translate the source sentence in test and development set, we obtain the results of its translations (Original). Then we tune the model parameter on the development dataset, and decode on the test dataset to generate new translations (Repaired). For the comparison, we list two results of our baseline method: One is the result of original translation (Original); the other baseline is the result of replacing original translations by the example with max similar score (Replaced). When combining the top-15 similar examples and original translation, the BLEU score of word-level oracle system (Oracle) is shown in Table 3, and the best system (IWSLT12_Best) on the dataset in IWSLT 2012 Evaluation Campaign is also listed.

As we can see from Table 3, we still obtain significantly inferior results compared to the original translation if we replace all the Google translations by the most similar examples, which is reflected by an absolute 8.55 point drop on the test set in BLEU score. On the other hand, our repairing method, which can repair original

translation result automatically in word-level, leads to an increase of 0.64 absolute BLEU point on the test set.

Model	BLEU%
Original	18.77
Replaced	10.22
Repaired	**19.41***
IWSLT12_Best	19.17

Table 3. Comparison with others on BLEU score (* significant at 0.005-level compared with the score of Original)

The experimental results show that our retrieval examples driven method appears to be effective in repairing incorrect translation with significant improvement in translation quality. Replacing by the most similar example cannot improve the translation quality when the similar score is low. Combing the examples with original translation improves the translation quality. In this sense, it is promising to correct translation by the examples via the proposed method.

3.3 The Effect of Example Similarity

We compare our method (Repaired) with two baselines (Original and Replaced) in different similar score region. We evaluate the translations by BLEU score. The results are listed in Table 4.

	Original	*Replaced*	*Repaired*
(0.9, 1.0)	17.23	**36.99**	22.98
(0.8, 0.9]	20.92	**27.20**	21.44
(0.7, 0.8]	19.90	15.03	**20.23***
(0.6, 0.7]	1 8.55	9.12	17.71
(0.5, 0.6]	18.38	4.22	17.50
(0.4, 0.5]	18.76	1.78	17.54
(0.0, 0.4]	18.25	0.80	17.20

Table 4. BLEU in different similar score region (*significant at 0.005-level compared with the score of Original)

From Table 4, we can see when the similar score is greater than 0.8, replacing the original translation by the most similar example has a serious advantage on BLEU score. When the similar score declines, the BLEU score also drops sharply. When the similar score region is (0.7, 0.8], our method has a significant improvement of absolute 0.33 BLEU score compared with original. And when the similar score declines bellow 0.7, the original translation is better. But it is remarkable that the result is generated by un-tuned parameters model.

3.4 Feature Analysis

In the experiment, we investigated the contribution of our different feature sets. After removing one feature, we retune the weights of features on the development set and re-decode on the test set. We evaluate the outputs of these models by BLEU, and list the results in Table 5.

Model	BLEU%
Repaired	19.41
Without word penalty	19.39
Without N-gram agreement	19.33
Without language model	19.23^
Without repetitive word penalty	19.21^
Without null-arc penalty	19.10^
Without original word penalty	19.01^
Without word confidence	18.98^
Without N-gram probability	18.71^

Table 5. Contribution of Features (^significant at 0.05-level compared with the repaired score)

As shown in Table 5, the performance drops significantly ($p<0.05$) when language model, repetitive word penalty, null-arc penalty, original word penalty, word confidence, N-gram probability is removed from the feature set respectively. And word penalty and N-gram agreement have weak effects on the results. It is remarkable that our specific features repetitive word penalty, original word penalty, and word confidence can bring the improvement of 0.20, 0.40, and 0.43 BLEU point than that without them respectively.

4 Conclusion

In this paper, we introduce statistical confusion network for translation example fusion to improve the current online MT quality We estimate the posterior of the word translation by the example similarity and introduce some new features to enhance the log-linear model optimization for the best translation. We check our method on the HIT dataset in the OLYMPICS task of IWSLT 2012 Evaluation Campaign. The Experimental results indicate that proposed method enhance the Chinese-English translations by Google, with a significant 0.64 absolute improvement according to BLEU score.

Acknowledgments

This work is supported by the NSF China (No. 61272384 & 61105072) and the National High Technology Research and Development Program of China (863 Program, No. 2011AA01A207).

References

Biçici Ergun and Marc Dymetman. 2008. Dynamic translation memory: using statistical machine translation to improve translation memory fuzzy matches. In Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, pages 454-465.

Brown, Peter F., Stephen A. Della-Pietra, Vincent J.Della-Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation. Computational Linguistics, 19(2): 263–313.

Chang Pi-Chuan, Michel Galley, and Christopher D. Manning.2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224-232.

DauméIII Hal and Jagadeesh Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of ACL11-HLT*, pages 407-412.

He Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 622-630.

Och Franz Josef. 2003. Minimum error rate training in statistical machine translation. *In Proceedings of ACL03*, pages 160-167.

Och Franz Josef, Christoph Tillman, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC), pages 20–28.

Koehn Philipp and Jean Senellart. 2010.Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Researchand the Translation Industry*, pages 21–31.

Ma Yanjun, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent Translation using Discriminative Learning - A Translation Memory-inspired Approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1239-1248.

Simard Michel and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. The Twelfth Machine Translation Summit, pages 120-127.

Smith James and Stephen Clark. 2009. EBMT for SMT: a new EBMT--SMT hybrid. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation*, pages 3-10.

Stolcke Andreas. 2002 SRILM - An Extensible Language Modeling Toolkit, in Proc. Intl. Conf. Spoken Language Processing.

Papineni Kishore, SalimRoukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311-318.