Augmenting a Statistical Translation System with a Translation Memory

Sanjika Hewavitharana, Stephan Vogel, Alex Waibel

Language Technologies Institute
Carnegie Mellon University, Pittsburgh
U.S.A.
{sanjika, vogel+, ahw}@cs.cmu.edu

Abstract. In this paper, we present a translation memory (TM) based system to augment a statistical translation (SMT) system. It is used for translating sentences which have close matches in the training corpus. Given a test sentence, we first extract sentence pairs from the training corpus, whose source side is similar to the test sentence. Then, the TM system modifies the translation of the sentences by a sequence of substitution, deletion and insertion operations, to obtain the desired result. Statistical phrase alignment model of the SMT system is used for this purpose. The system was evaluated using a corpus of Chinese-English conversational data. For close matching sentences, the translations produced by the translation memory approach were compared with the translations of the statistical decoder.

1. Introduction

Spoken language translation has received more attention in recent times. Some of the notable attempts include Verbmobil (Wahlster, 2000) and Nespole (Metze et at., 2002). Many corpora have been compiled for this purpose covering various domains, including conversations in travel and medical domains. Typically, these corpora contain shorter sentences. For example, in the Basic Travel Expression Corpus (BTEC) corpus (Takezawa et al., 2002), the sentences have 6-7 words on average. Another noticeable feature is that they have sentence with similar patterns, as shown in Figure 1 with Spanish-English sentence pairs from the BTEC corpus.

en qué tipo de trabajo estás interesado ? what kind of job are you interested in ?
en qué tipo de cosas estás interesado ? what kind of things are you interested in ?
en qué tipo de excursiones estás interesado ? what kind of tour are you interested in ?

Figure 1: Similar patterns in sentences

These three sentences differ only in one word in both Spanish sentences as well as their English translations. For a given test sentence, we often find in the training corpus, a very similar sentence with few mismatching words; sometimes even an exact matching sentence. Translation memory (TM) systems typically work well in these situations. In its pure form, a TM system is simply a database of past translations, stored as sentence pairs in source and target languages. Whenever an exact match is found for a new sentence to be translated, the desired translation is extracted from the translation memory. TM systems have been successfully used in Computer Aided Translations (CAT) as a tool for human translators.

There have been attempts to combine translation memory with other machine translation approaches. In (Marcu, 2001) an automatically derived TM is used along with a statistical model to obtain translations of higher probability than those found using only a statistical model. Sumita (2001) describes an example-based technique which extracts similar translations and modifies them using a bilingual dictionary. Watanabe and Sumita (2003) proposed an example-based decoder that start with close matching example translations, and then modify them using a greedy search algorithm. Instead of extracting complete sentences from the TM, Langlais and Simard, (2002) work on sub sentential level. Translations for word sequences are extracted from a TM and then fed into a statistical engine to generate the desired translation.

In this paper, we present an experiment where we attempted to augment a statistical translation system with a translation memory. For a sentence which has a close match in the training corpus, the idea is to start with the available translation and apply specific modifications to produce the desired translation. By a close match, we mean a very similar sentence with only a few mismatching words.

Given a test sentence, we extract sentence pairs from the bilingual training corpus, whose source side is similar to the test sentence. If a close matching sentence is found, we use our TM system to translate it. For each mismatching word in the source side of the close matching pair, we identify its translation in the target side. Then a sequence of substitution, deletion and insertion operations is applied to the target side to produce the correct translation. If a close match is not found in the training corpus, we use a statistical translation system to generate the translation.

The system was evaluated using a subset of the Chinese-English BTEC corpus. For those close matching sentences, the translations produced by the TM system were compared with the translations produced by the statistical decoder. In our current experiments TM system did not show an improvement in terms of automatic evaluation metrics. However, a subjective human evaluation found that, in several instances, the TM system produced better translations than the statistical decoder.

In the following section we explain the TM system in detail. We also describe the phrase extraction method we used to identify alignments between source words and target words, which is a modified version of the IBM1 alignment model (Brown et al. 1993). In Section 3, we present the experimental setting and the results of the evaluation. It is followed by a discussion in section 4, and conclusions in section 5. We have identified a number of improvements to the current system, some of which are already in progress.

2. Translation Memory System

2.1. Extracting Similar Sentences

For each new test sentence *F*, we find a set of similar source sentences {F_1, F_2, …} from the training corpus. The similarity is measured in terms of the standard edit distance criterion with equal penalties for insertion, deletion and substitution operations. The corresponding set of translations {E_1, E_2,…} is also extracted from the bilingual training corpus.

Following are some close matching sentences we extracted for the Spanish sentence *estoy nerviosa*.

i. estoy resfriado (i have a cold)
ii. estoy cansada (i am tired)
iii. estoy resfriado (i feel chilled)

If we select the first match as input to the TM system, it will generate the translation, *i have nervous*. If instead we select the second match, we get, *I am nervous*, which is the correct translation. Selecting the first best does not always produce better results. Therefore, for each test sentence, we select the 10 best matching sentence pairs as candidates for the next step.

If we found an exact match among the extracted sentences, we terminate the search and output the translation of it as the desired translation of the test sentence. In the case of multiple exact matches (which might have different meanings in the target side), we score each sentence pair (F_k, E_k) using a translation model and a language model and select the best one.

2.2. Modifying Translations of Close Matching Sentences

If an exact match is not found, but a close matching sentence pair (F_k, E_k) is found, then the translation E_k is slightly altered using a statistical translation model to produce the correct result. We start by identifying the words in F_k that have to be changed and the sequence of substitution, deletion, or insertion operations[1] required to make it the same as *F*. For each of these words, we then identify its alignment in the target side E_k. Finally, the aligned words are modi-

[1] Since there can be many such sequences with the same edit distance, the sequence is not unique.

fied with the identified operations, to produce the desired translation. Figure 2 illustrates the substitution operation for a single word. Details of how each operation is performed are explained in section 2.4.

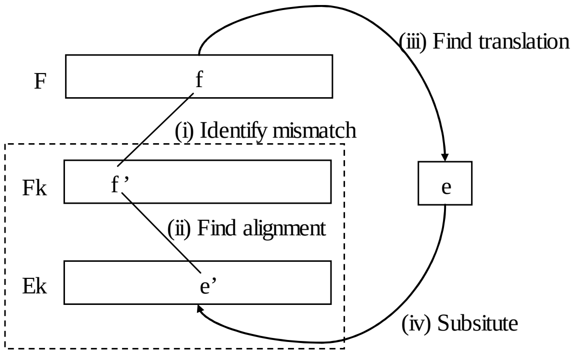

Figure 2: Steps in the substitution operation

The underlying assumption here is that the same sequence of operations that resolves the mismatch between the test sentence *F* and the source sentence F_k, would produce the correct translation *E* from E_k. Therefore, it is important to reliably identify the alignments of the words in the source sentence. Our initial experiments with word-to-word alignment did not produce correct translations since a word in the source side, sometimes, corresponds to more than one word in the target side.

Therefore, we used a phrase-to-phrase alignment method which allows us to do phrase level operations. The term *phrase* is used throughout the paper to indicate any sequence of words, not necessarily in the linguistic sense. We used the same method to identify the candidate translations of the mismatching words in *F*. In the next section, we explain the PESA phrase extraction method (Vogel et al., 2004) used in the experiments.

2.3. Phrase Extraction via Sentence Alignment (PESA)

Suppose we are searching for a good translation for the source phrase $f = f_1 \ldots f_k$, and that we found a sentence in the bilingual corpus, containing the same word sequence. We are now interested in identifying a sequence of words $e = e_1 \ldots e_l$ in the target sentence, which is an optimal translation of the source phrase. Although any sequence of words in the target sentence can be a candidate translation, most of them would be deemed incorrect. Some of them would be partial translations while a small number of candidates would be acceptable or good translations. We want to find these good candidates.

The IBM1 word alignment model aligns each word in the source phrase to all the words in the target phrase with varying probabilities. Typically, only one or two words will have high alignment probability, which for IBM1 model is just the lexicon probability. We now modify the IBM1 alignment with the following constrains:

- for words inside the source phrase we sum probabilities only over the words inside the candidate target phrase, and for words outside the source phrase we sum probabilities only over the words outside the candidate target phrase.
- the position alignment probability, which for the standard IBM1 alignment is $1/I$, where I is the number of words in the target sentence, is modified to $1/l$ inside the source phrase and to $1/(I-l)$ outside the source phrase.

More formally we calculate the constrained alignment probability:

$$p_{i_1,i_2}(f \mid e) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} p(f_j \mid e_i) \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} p(f_j \mid e_i) \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} p(f_j \mid e_i)$$

and optimize over the target side boundaries i_1 and i_2,

$$(i_1, i_2) = \arg\max_{i_1,i_2} \{ p_{i_1,i_2}(f \mid e) \}$$

where J is the number of words in the source sentence.

Since word alignment models are asymmetric with respect to aligning one-to-many words, it gives better results when the alignments are calculated for both directions. Similarly we calculate the alignment probabilities for the other direction:

$$p_{i_1,i_2}(e \mid f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1..j_2)} p(e_i \mid f_j) \times \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} p(e_i \mid f_j) \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1..j_2)} p(e_i \mid f_j)$$

To find the optimal target phrase we interpolate both alignment probabilities and take the pair (i_1, i_2) which gives the highest probability.

$$(i_1, i_2) = \arg\max_{i_1, i_2}\{(1-c)\log(p_{i_1,i_2}(f \mid e)) + c\log(p_{i_1,i_2}(e \mid f))\}$$

The phrase pairs are extracted from the bilingual corpus at decoding time. We treat the single source words in the same way as a phrase of length 1. The target translation can then be either one or several words.

Most phrase pairs $(f, e) = (f_{j_1}... f_{j_2}, e_{i_1}...e_{i_2})$ are seen only a few times, even in very large corpora. Therefore, probabilities based on occurrence counts have little discriminative power. We calculate phrase translation probabilities based on statistical lexicon, i.e. on the word translation probabilities *p(f, e)*:

$$p(f \mid e) = \prod_j \sum_i p(f_j \mid e_i)$$

2.4. Modification Operations

For a given test sentence *F* and a close matching sentence pair (*F'*, *E'*) with an edit distance of one, the three repair operations are handled as follows. Boldface letters are used to indicate phrases.

1. Substitution of word *f'* in *F'* with word *f* in *F*:
   - i. Find all possible phrase alignments ***e'*** in *E'* for the word *f'*.
   - ii. Find all possible translations **e** of word *f*.
   - iii. Replace ***e'*** with ***e*** to produce *E*.
   - iv. Score the resulting translation (*E*, *F*) with the translation and language models.
   - v. Iterate over all ***e'*** and ***e*** and choose the best *E* as the desired translation.

2. Deletion of word f' from *F'*:
   - i. Find the possible phrase alignments ***e'*** in *E'* for the word *f'*.
   - ii. Remove ***e'*** from *E'* to produce *E*.
   - iii. Score the resulting translation (*E*, *F*) with the translation and language model.
   - iv. Iterate over all ***e'*** and choose the best *E* as the desired translation.

3. Insertion of word f into *F'*:
   - i. Find all possible translations ***e*** of word *f*.
   - ii. Insert ***e*** into a position *i* in *E'* to produce *E*.
   - iii. Score the resulting translation (*E*, *F*) with the translation and language model.
   - iv. Iterate over all translations ***e*** and all word positions *i* in *E'* and choose the best *E* as the desired translation.

When more than one close matching sentence is found, the above process is iteratively applied on all of them and the best one is selected as the resultant translation.

3. Evaluation

3.1. Corpus

For the evaluation we used a subset of the BTEC which contains travel conversations in Chinese and English. The corpus was originally created in Japanese and English by ATR (Takezawa et al., 2002) and was later extended to other languages including Chinese. Our training set contained 20,000 sentence pairs, where the Chinese sentences were already word segmented. Table 1 summarizes the statistics of the training set.

	Chinese	English
Sentences	20,000	20,000
Words	182,902	188,935
Vocabulary	7,645	7,181
LM PP	—	68.6

Table 1: Training data statistics

We used a development set (Dev) to tune the parameters of the system and a final test set (Test) to evaluate the tuned system. It was assumed that the word segmentation of the test data matches the word segmentation of the training data. 16 reference translations per sentence were used for the evaluation. Table 2 gives the details of the two test sets.

	Chinese	
	Dev	Test
Sentences	506	500
Words	3515	4108
Vocabulary	870	893
Unknown Words	160	104

Table 2: Test data statistics

3.2. Language Model

A standard trigram language model was used to evaluate the translations produced by TM system, as well as in the statistical decoder. We used the SRI language model toolkit (SRI_LM Toolkit) to build the language model using English data of the training set. Table 1 also contains the language model perplexity (LM PP).

3.3. Statistical Translation System

We used a statistical machine translation (SMT) decoder which allows phrase-to-phrase translation using the phrase extraction method explained in section 2.3. The decoding process works in two states: First, the word-to-word and phrase-to-phrase translations and, if available, other specific information like named entity translation tables are used to build a translation lattice. A standard n-gram language model is then applied to find the best path in this lattice. Standard pruning strategies are employed to keep the decoding time within reasonable bounds. Details of the system are described in (Vogel et al., 2003) and (Vogel, 2003).

Our SMT system and the TM system are closely connected, since we use the same IBM1 translation lexicon, language model and the phrase extraction method in both systems. This contrasts our approach with a multi engine approach where results from different, often independent, translation systems are integrated.

3.4. Evaluation

We extracted similar sentences from the training data using the edit distance criterion. Table 3 gives the similarity statistics for both development and test set, based on the best match.

	Dev	Test
Exact match	27	30
1 mismatch	103	104
> 1 mismatch	376	366

Table 3: Best matching cases

For the 506 sentence development set, 5% of the sentences had an exact match in the training corpus. Another 20% of the sentences could be matched with one substitution, deletion or insertion. For the 500 sentence test set, these values were 6% and 20% respectively.

We tested the TM system for the test sentences that have exact matches or close matches with only one mismatching word. There are 130 sentences in the development set which holds this condition, and in the test set there are 134 sentences. Translation results are reported in Bleu (Papineni, 2001) and NIST mteval (Mteval, 2002) scores. NIST mteval script version 11a was used to calculate both the NIST and Bleu scores.

We used the SMT system to generate translations for the complete data set. Parameters of the SMT system were tuned to generate translations with high NIST scores[2]. The translations corresponding to exact matches or one mismatch were then replaced by those produced by the TM system. We tested the system with two different settings; only considering the single best matching sentence (TM 1–Best), and considering up to 10 best matching sentences (TM n–Best). Table 4 gives the final translation results.

	Dev		Test	
	Bleu	NIST	Bleu	NIST
TM 1–Best	38.8	7.84	36.8	8.16
TM n–Best	39.3	7.86	37.8	8.27
SMT Alone	39.1	7.90	37.9	8.31

Table 4: Translation results

4. Discussion

As it can be seen in Table 4, the translation memory did not produce improved results in terms of NIST score. For the development set, it has slightly better results with respect to Blue score. Use of the n-best list of close matching sentences, rather than only the best matching sentence did produce better results. Still, there is a small drop in NIST scores. However, the differences are not statistically significant[3].

When the translations of the two methods are compared, in several instances, the TM system produced better quality translations than the SMT

[2] This discrepancy between Blue and NIST scores is due to the different method used to calculate length penalty for Blue metric in the current mteval script. This problem arises only, when several reference translations are available, and when they are very short, as is the case with BTEC data.

[3] 95% confidence levels for the data set are:
For Bleu: [-3.0,+3.0] (i.e. ± 8% relative difference)
For NIST:[- 0.4,+0.4] (i.e. ± 5% relative difference)

system. Some of the notable examples are given in Figure 3.

Ref	how much does it cost to send this to japan
SMT	please send this to japan how much is it
TM	what is the cost for sending this to japan
Ref	do i have to transfer to get there
SMT	i'd like to change trains to get there
TM	do i have to change buses to get there
Ref	could you repeat that please
SMT	would you please say it again please
TM	Would you say it again please
Ref	what is today's date
SMT	what is today's number
TM	what 's the date today
Ref	i don't know my size
SMT	i don't know my size
TM	nobody knows size
Ref	where's the ladies' restroom
SMT	where's ladies' bathroom
TM	where are the restrooms

Figure 3: Sample translations

For each sentence, one reference translation, the result of the SMT system and the result of the TM system are provided. Top part of Figure 3 contains examples where the TM system generated better translations than the SMT system. In the last 2 examples, the SMT translation is better than the TM translation.

We conducted a subjective evaluation, to compare the quality of the translations. Two persons were asked to compare the translations of the TM system and the SMT system for those 130 sentences with exact matches/only one mismatch. They were asked to mark each sentence with one of the following:

A – Translation 1 is better than Translation 2.
B – Translation 2 is better than Translation 1.
C – Both translations are comparable in quality.

Evaluators were not aware of which system generated which result. We also shuffled the translations for each sentence so as to further remove the bias towards a particular system. Table 5 gives the subjective evaluation results for evaluators E1 and E2.

	E1		E2	
	#	%	#	%
SMT Better	11	8	17	13
TM Better	37	29	37	29
Comparable	82	63	76	58

Table 5: Subjective evaluation results

According to the subjective evaluations, for 29% of the sentences, the TM system produced better quality translations than the SMT system. On average, 11% of the sentences are better translated by the SMT system. Nearly 60% of the time both systems produced translations of comparable quality.

When the full dataset is considered, 5% of the sentences have better translations after combining the TM system with the SMT system.

Why is this improvement not reflected in the automatic evaluation scores? A possible explanation can be as follows: The differences we observe in the subjective evaluations are at the sentence level whereas automatic metrics work on the word level. Therefore, these metrics might not be able to capture the subtle differences in quality between the two systems, as in the cases listed in Figure 3. For example, let's consider the n-gram precision for Bleu metric for the example 1 in Figure 3. The TM system translation has 1 trigram, 2 bigram and 4 unigram matches. The SMT system translation has 1 4-gram, 2 trigram, 4 bigram and 7 unigram matches. This would give a higher Bleu score for the SMT translation than the TM translation, although the TM translation is clearly better than the SMT translation.

The phrase extraction method used in the SMT system allows alignments up to any length. For sentences that are close or exact matching to those found in the training corpus, this allows the extraction of longer phrases, or even the full translation. Therefore, the SMT system can generate the desired translation fairly easily with less re-orderings. In other words, the SMT system in these situations works as a translation memory. This makes it a stronger baseline. Further, scoring the translations produced by the TM system using an SMT based translation model might have a bias towards translations that are closer to those produced by the SMT system.

5. Conclusions and Future Work

In this paper we presented a translation memory system, which can enhance the translations of a statistical machine translation system. We also presented a phrase alignment approach, which finds the target phrase for a given source phrase by optimizing the alignment probability for the entire sentence pair. The TM system did not show an improvement over the SMT baseline in terms of automatic evaluation metrics. However, a subjective evaluation found that the TM system generate better quality translation, resulting in a 5% overall improvement over the combined system. We plan to extend this work in a number of directions: 1. Allow more than one mismatch between the test sentence and the sentences in the training corpus, esp. for longer sentences. 2. Using additional information, such as parts of speech, to have a more discriminative matching between sentences. 3. Integrating the SMT system and the TM system using a better criterion than just on the number of mismatches.

Perhaps a more interesting direction would be to use the TM system within the phrase search itself. The current phrase search only extracts exact matching phrases. Using the same repair operations we use in our TM system, we would be able to extract close matching phrases, repair them and use them in the SMT decoder.

6. References

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

Philippe Langlais and Michel Simard (2002). Merging Example-Based and Statistical Machine Translation: An Experiment. *In Proceedings of the 5th Conference of Association for Machine Translation in the Americas (AMTA)*, pp. 104-114, Tiburon, California, October.

Daniel Marcu (2001). Towards a Unified Approach to Memory- and Statistical-Based Machine Translation. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 378-385, Toulouse, France, July.

Florian Metze, J. McDonough, H. Soltau, C. Langley, A. Lavie, L. Levin, T. Schultz, A. Waible, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, E. Pianta (2002), The NESPOLE! Speech-to-Speech Translation System, In Proceedings of HLT, San Diego, California U.S, March.

MTeval (2002). NIST MT Evaluation Kit Version 11a. Available at:

http://www.nist.gov/speech/test/mt/.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. *Technical Report RC22176 (W0109-022)*, IBM Research Division, T.J. Watson Research Center.

SRI-LM. The SRI Language Modeling Toolkit. SRI Speech Technology and Research Laboratory. http://speech.sri.com/projects/srilm/

Eiichiro Sumita (2001). Example-based machine translation using DP-matching between word sequences, *DDMT workshop of 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1-8.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, &Seiichi Yamamoto. (2002). Towards a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings. of LREC 2002*, pp. 147-152, Las Palmas, Canary Islands, Spain, May.

Stephan Vogel (2003). SMT Decoder Dissected: Word Reordering. *In Proceedings of 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 561-566, Beijing, China, October.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel (2003). The CMU Statistical Translation System. *In Proceedings of MT Summit IX*, New Orleans, LA, USA, September.

Stephan Vogel, Sanjika Hewavitharana, Muntsin Kolss and Alex Waibel (2004). The ISL Statistical Translation System for Spoken Language Translation. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 65-72, Kyoto, Japan, September.

Wolfgang Wahlster, ed. (2000). Verbmobil: Foundations of Speech-to-Speech Translation, Springer.

Taro Watanabe and Eiichiro Sumita (2003). Example-based Decoding for Statistical Machine Translation. *In Proceedings of MT Summit IX*, New Orleans, LA, USA, September.