Using Noisy Bilingual Data for Statistical Machine Translation

Stephan Vogel
Interactive Systems Lab
Language Technologies Institute
Carnegie Mellon University
vogel+@cs.cmu.edu

Abstract

SMT systems rely on sufficient amount of parallel corpora to train the translation model. This paper investigates possibilities to use word-to-word and phrase-to-phrase translations extracted not only from clean parallel corpora but also from noisy comparable corpora. Translation results for a Chinese to English translation task are given.

1 Introduction

Statistical machine translation systems typically use a translation model trained on bilingual data and a language model for the target language, trained on perhaps some larger monolingual data. Often the amount of clean parallel data is limited. This leads to the question of whether translation quality can be improved by using additional noisier bilingual data.

Some approaches, like (Fung and MxKeown, 1997), have been developed to extract word translations from non-parallel corpora. In (Munteanu and Marcu, 2002) bilingual suffix trees are used to extract parallel sequences of words from a comparable corpus. 95% of those phrase translation pairs were judged to be correct. However, no results where reported if these additional translation correspondences resulted in improved translation quality.

2 The SMT System

Statistical translation as introduced in (Brown et al., 1993) is based on word-to-word translations. The SMT system used in this study relies on multi-word to multi-word translations. The term phrase translations will be used throughout this paper without implying that these multi-word translation pairs are phrases in some linguistic sense. Phrase translations can be extracted from the Viterbi alignment of the alignment model.

Phrase translation pairs are seen only a few times. Actually, most of the longer phrases are seen only once in even the larger corpora. Using relative frequency to estimate the translation probability would make most of the phrase translation probabilities 1.0. This would lead to two consequences: First, phrase translation would always be preferred over a translation generated using word translations from the statistical and manual lexicons, even if the phrase translation is wrong, due to misalignment. Secondly, two translations would often have the same probability. As the language model probability is larger for shorter phrases this will usually result in overall shorter sentences, which sometimes are too short.

To make phrase translations comparable to the word translations the translation probability is calculated on the basis of the word translation probabilities resulting from IBM1-type alignment.

$$p(f_m^n|e_k^l) = \prod_{j=m}^{n} \sum_{i=k}^{l} p(f_j|e_i) \qquad (1)$$

This now gives the desired property that longer

translations get higher probabilities. If the additional word should not be part of the phrase translation then these additional probabilities $p(f_j|e_i)$ which go into the sum will be small, i.e. the phrase translation probabilities will be very similar and the language model gives a bias toward the shorter translation. If, however, this additional word is actually the translation of one of the words in the source phrase then the additional probabilities going into the summation are large, resulting in an overall larger phrase translation probability.

More importantly, calculating the phrase translation probability on the basis of word translation probabilities increases the robustness. Wrong phrase pairs resulting from errors in the Viterbi alignment will have a low probability.

3 What's in the Training Data

3.1 The Corpora

To train the Chinese-to-English translation system 4 different corpora were used: 1) Chinese tree-bank data (LDC2002E17): this is a small corpus (90K words) for which a tree-bank has been built. 2) Chinese news stories, collected and translated by the Foreign Broadcast Information Service (FBIS). 3) Hong Kong news corpus distributed through LDC (LDC2000T46). 4) Xinhua news: Chinese and English news stories publish by the Xinhua news agency.

The first three corpora are truly bilingual corpora in that the English part is actually a translation of the Chinese. Together, the form the clean corpus which has 9.7 million words.

The Xinhua news corpus (XN) is not a parallel corpus. The Chinese and English news stories are typically not translations of each other. The Chinese news contains more national news whereas the English news is more about international events. Only a small percentage of all stories is close enough to be considered as comparable. Identification of these story pairs was done automatically at LDC using lexical information as described in (Xiaoyi Ma, 1999). In this approach a document B is considered an approximate translation of document A if the similarity between A and B is above some threshold, where similarity is defined as the ratio of tokens from A for which a translation appears in document B in a nearby position. The document with the highest similarity is selected. For the Xinhua News corpus less then 2% of the entire news stories could be aligned. Inspection showed that even these pairs can not be considered to be true translations of each other.

In our translation experiments we also used the LDC Chinese English dictionary (LDC2002E27). This dictionary has about 53,000 Chinese entries with on average 3 translations each.

The FBIS, Hong Kong news and Xinhua news corpora all required sentence alignment. Different sentence alignment methods have been proposed and shown to give reliable results for parallel corpora. For non-parallel but comparable corpora sentence alignment is more challenging as it requires – in addition to finding a good alignment – also a means to distinguish between sentence pairs which are likely to be translations of each other and those which are aligned to each other but can not be considered translations.

An iterative approach to sentence alignment for this kind of noisy data has been described in (Bing Zhao, 2002). This approached was used to sentence align the Xinhua News stories. Sentence length and lexical information is used to calculate sentence alignment scores. The alignment algorithm allows for insertions and deletions. These sentences are removed as are sentence pairs which have a low overall sentence alignment score. About 30% of the sentence pairs were deleted to result in the final corpus of 2.7 million words.

The test data used in the following analysis and also in the translation experiments is a set of 993 sentences from different Chinese news wires, which has been used in the TIDES MT evaluation in December 2001.

3.2 Analysis: Vocabulary Coverage

To get good translations requires first of all that the vocabulary of the test sentences is well covered by the training data. Coverage can be expressed in terms of tokens, i.e. how many of the tokens in the test sentences are covered by the vocabulary of the training corpus, and in terms of types, i.e. how many of the word types in the test sentences have been seen in the training data.

Table 1: Corpus coverage (C-Voc) and vocabulary coverage of the test data given different training corpora.

Corpus	Voc	C-Cov	V-Cov
Clean	46,706	99.51	97.89
Clean + XN	69,269	99.80	98.88
Clean + XN + LDC	74,014	99.84	99.10

A problem with Chinese is of course that the vocabulary depends heavily on the word segmentation. In a way the vocabulary has to be determined first, as a word list is typically used to do the segmentation. There is a certain trade-off: a large word list for segmentation will result in more unseen words in the test sentences with respect to a training corpus. A small word list will lead to more errors in segmentation. For the experiments reported in this paper a word list with 43,959 entries was used for word segmentation.

Table 1 gives corpus and vocabulary coverage for each of the Chinese corpora.

3.3 Analysis: N-gram coverage

Our statistical translation system uses not only word-to-word translations but also phrase translations. The more phrases in the test sentences are found in the training data, the better. And longer phrases will generally result in better translations, as they show larger cohesiveness and better word order in the target language. The n-gram coverage analysis takes all n-grams from the test sentences for n=2, n=3, ... and finds all occurrences of these n-grams in the different training corpora. From Table 2 we see that the Xinhua news corpus, which is only about a quarter of the size of the clean data, contains a much larger number of long word sequences occurring also in the test data. This is no surprise, as part of the test sentences come from Xinhua news, even though they date from a year not included in the training data. Adding this corpus to the other training data therefore gives the potential to extract more and longer phrase to phrase translations which could result in better translations.

Many of the detected n-grams are actually overlapping, resulting from a very small number of very long matches was detected. And each n-gram contains m (n-m+1)-grams. The longest matching n-grams in the Xinhua news corpus were 56, 53, 43, 34, 31, 28, 24, 21 words long, each occurring once.

Table 2: Number of n-grams from test sentences found in the different corpora.

n	Clean	XN	Clean + XN
2	12621	11503	13683
3	6990	6525	8663
4	2396	2735	3628
5	810	1283	1611
6	314	745	884
7	123	486	545
8	53	368	395
9	29	310	321
10	18	275	281

3.4 Training the Alignment Models

IBM1 alignments (Brown et al., 1993) and HMM alignments (Vogel et al., 1996) were trained for both the clean parallel corpus and for the extended corpus with the noisy Xinhua News data. The alignment models were trained for Chinese to English as well as English to Chinese. Phrase-to-phrase translations were extracted from the Viterbi path of the HMM alignment. The reverse alignment, i.e. English to Chinese, was used for phrase pair extraction as this resulted in higher translation quality in our experiments. The translation probabilities, however, where calculated using the lexicon trained with the IBM1 Chinese to English alignment.

Table 3: Training perplexity for clean and clean plus noisy data.

Model	Clean	Clean + XN
IBM1	123.44	142.85
IBM1-rev	105.72	120.48
HMM	101.34	121.34
HMM-rev	78.61	92.79

Table 3 gives the alignment perplexities for the different runs. English to Chinese alignment gives

lower perplexity than Chinese to English. Adding the noisy Xinhua news data leads to significantly higher alignment perplexities. In this situation, the additional data gives us more and longer phrase translations, but the translations are less reliable. And the question is, what is the overall effect on translation quality.

4 Translation Results

The decoder uses a translation model (the LDC glossary, the IBM1 lexicon, and the phrase translation) and a language model to find the best translation. The first experiment was designed to amplify the effect the noisy data has on the translation model by using an oracle language model built from the reference translations. This language model will pick optimal or nearly optimal translations, given a translation model. To evaluate translation quality the NIST MTeval scoring script was used (MTeval, 2002). Using word and phrase translations extracted form the clean parallel data resulted in an MTeval score of 8.12. Adding the Xinhua News corpus improved the translation significantly to 8.75. This shows that useful translations have been extracted from the additional noisy data.

The next step was to test if this improvement is also possible when using a proper language model. The language model used was trained on a corpus of 100 million words from the English news stories published by the Xinhua News Agency between 1992 and 2001. Unfortunately, the MTeval score dropped from 7.59 to 7.31 when adding the noisy data. Restricting the lexicon, however, to a small number of high probabilty translations, thereby reducing the noise in the lexiocn, the score improved only marginally for the clean data system, but considerably for the noisy data system. The noisy data system then outperformed the clean data system. These results are summarized in Table 4. A t-test run on the sentence level scores showed that the difference between 7.62 and 7.69 is statistically significant at the 99% level.

Table 4: Translation results.

System Setup	Clean	Noisy
LM-Oracle	8.12	8.75
LM-100m	7.59	7.31
LM-100m, lexicon prunded	7.62	7.69

5 Summary

Initial translation experiments have shown that using word and phrase translations extracted from noisy parallel data can improve translation quality. A detailed analysis will be carried out to see how the different training corpora contributed to the translations. This will include a human evaluation of the quality of phrase translations extracted from the noisier data. Next steps will include training the statistical lexicon on clean data only and using this to filter the phrase-to-phrase translations extracted from comparable corpora.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

Pascale Fung and Kathleen McKeown. 1997. A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups. *In Machine Translation*, volume 12, numbers 1-2 (Special issue), Kluwer Academic Publisher, Dordrecht, The Netherlands, pp. 53-87.

Xiaoyi Ma and Mark Y. Lieberman. 1999. BITS: A Method for Bilingual Text Search over the Web.. *Machine Translation Summit VII*.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing Comparable Corpora With Bilingual Suffi x Trees. *Empirical Methods in Natural Language Processing* , Philadelphia, PA.

NIST MT evaluation kit version 9. Available at: http://www.nist.gov/speech/tests/mt/.

Stephan Vogel, Hermann Ney, and Christoph Tillmann, HMM-based Word Alignment in Statistical Translation, in *COLING '96: The 16th Int. Conf. on Computational Linguistics*, Copenhagen, August 1996, pp. 836–841.

Bing Zhao and Stephan Vogel, 2002. Adaptive Parallel Sentence Mining from Web Bilingual News Collection. *ICDM '02: The 2002 IEEE International Conference on Data Mining* , Maebashi City, Japan, December 2002.