7th International Conference on Spoken
Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

IMPROVING STATISTICAL MACHINE TRANSLATION FOR A SPEECH-TO-SPEECH TRANSLATION TASK

Stephan Vogel, Alicia Tribble

Interactive Systems Laboratory
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper we investigate if statistical machine translation (SMT) is possible when only a small bilingual corpus is available for training the system. Using additional knowledge sources which are not domain-specific improves the performance of the system considerably. We present results on a speech translation task for German to English. Automatic and human evaluation are used to compare the performance of the SMT system to an interlingua-based translation system.

1. INTRODUCTION

"There is no data like more data!" This saying characterizes the need for large domain-specific training corpora for statistical systems. Data sparseness is often used as an argument against statistical systems and in favor of hand-crafted, knowledge-based systems. In this paper we investigate whether statistical machine translation (SMT) is possible when only a small bilingual corpus is available for training the system. What can be done to improve system performance, especially by adding knowledge sources which are not domain specific.

The context of our evaluation is a speech translation task; training data and evaluation results were collected in the Nespole! speech translation project. We develop and test a statistical system trained on this data and compare it to the interlingual Nespole! system. We present results from both automatic and human evaluations.

2. STATISTICAL MACHINE TRANSLATION

Statistical machine translation has been advocated by the IBM research group from the early 90s [1]. The approach is based on Bayes' decision rule: given a source sentence f_1^J of length J, the translation e_1^I is given by:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{p(e_1^I) \cdot p(f_1^J|e_1^I)\} \quad . \qquad (1)$$

Here, $p(e_1^I)$ is the language model of the target language, and $p(f_1^J|e_1^I)$ is the string translation model. The argmax operation denotes the search problem.

The language model is typically an n-gram language model. The translation model we use is an HMM-based alignment model introduced in [2]. Using the 'hidden' alignments $a_1^J := a_1 ... a_j ... a_J$ for a sentence pair $[f_1^J; e_1^I]$ the translation probability can be rewritten as:

$$\begin{aligned} p(f_1^J|e_1^I) &= \sum_{a_1^J} p(f_1^J, a_1^J|e_1^I) \\ &= \sum_{a_1^J} \prod_{j=1}^{J} p(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \end{aligned}$$

Assuming a first-order dependence on the alignments a_j only and that the translation probability depends only on e_{a_j} we have the following HMM-based model:

$$p(f_1^J|e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})]$$

This model can be trained using the Forward-Backward algorithm.

To translate a new sentence f_1^J amounts to a search problem. Using the lexicon and the n-gram language model, the sentence e_1^I is generated which has the highest probability $p(e_1^I|f_1^J)$. The translation program used in this study is based on weighted finite state transducers [3]. The statistical lexicon as well as the phrase-to-phrase alignments produced by the alignment modeling software can be converted into such transducers. Additional knowledge sources can easily be converted to transducer format and added to the translation program in this way.

3. OPTIMIZING THE STATISTICAL SYSTEM

In the following section we describe several extensions to the basic SMT system defined above. Performance results for each of these changes are given in Section 4.

3.1. Phrase-level translation

One drawback of the basic alignment model is that it produces a lexicon containing only one-to-one, word-level mappings. This poses some problems if one language uses compound words. For example, the German word 'Skipiste' might be translated as 'ski slope', generating two entries in the statistical lexicon:

Skipiste # ski # p('Skipiste' | 'ski')
Skipiste # slope # p('Skipiste' | 'slope')

The lexicon does not distinguish between this situation and the situation where a word has several translations, i.e. synonyms.

This is one reason why the SMT system discussed in our experiments uses not only the word-to-word lexicon but also phrase-to-phrase alignments generated during HMM alignment. We also take advantage of the asymmetry of the word alignment models, which allow one target word to align to several source words but not vice versa, by training in both directions and combining the alignment results.

3.2. Large Language Model

In developing data-driven translation systems the bottle-neck is usually the bilingual corpus. Monolingual data is more easily available in most cases, and using a larger monolingual corpus to train the target language model can significantly improve the translation performance. The benefit will be stronger if this additional data is comparable to the bilingual training data.

3.3. Adding a General Purpose Lexicon

Our system is tuned for translating spontaneous speech in the travel domain, but the cost of collecting, transcribing, and translating a corpus of speech data for this domain generally prevents a large amount of training material from being available when developing a new system. In such a case, the major obstacle to translation quality is the high number of unseen words.

One method of lessening this effect is to add a large, general-purpose background lexicon like an online bilingual dictionary. This lexicon can be reformatted as an additional transducer with uniform probability distribution. Alternatively, the lexicon can be added to the training corpus. The second approach allows probabilities for the lexicon entries to be set according to their combined distribution in the new lexicon and original training corpus.

3.4. Leveraging Existing MT Knowledge Sources

Another resource for world knowledge that can be added to the SMT system in our experiments is a body of analysis grammars that have been written for an existing interlingual MT (IL-MT) system in the same domain. These grammars are composed of hand-written context-free rules that transform natural language text into a language-independent semantic representation. They have been created to cover the same training data that is seen by the SMT system, but because they are written by hand they also encode some world knowledge from the human grammar writers.

A new statistical transducer can be created from the parallel vocabulary in two of these grammars in the following way: given analysis grammars in English (E) and German (G), we search for all rules that have matching LHSs, E_i and G_m. Assuming that rules generating the same IF Value may be taken as translations for each other, we create a transducer entry for every pair (E_{ij}, G_{mn}) where j and n index over the RHSs of the rules E_i and G_m, respectively.

In this transducer each rule has equal weight normalized over the target words or phrases. An additional transducer can be constructed by hypothesizing that the first rule in any list of RHS alternatives would be the most common translation, and assigning unequal rule weights accordingly. Results reported for transducer I refer to a weighted transducer where the first rule in a list receives $\frac{1}{2}$ of the translation probability for that target word.

4. EXPERIMENTS

4.1. The Corpus

The training data for the SMT system was originally collected in the Nespole! speech-to-speech MT project [4]. Several dialogs were recorded from telephone conversations between an Italian tourist office and native English- and German-speaking clients. The agents, native speakers of Italian, spoke English or German for the data collection.

Table 1 shows that the corpus is very small. Nearly 50% of the German vocabulary and about 40% of the English Vocabulary occurs only once in the corpus.

Table 1. Training corpus statistics.

	German	English
Sentences	3182	3182
Words	14992	15572
Vocabulary	1367	1041
Singletons	645	410

For testing the translation systems, a number of the dialogs were held out. The results reported here are for three of the held-out dialogs originally recorded in German. One dialog (70 sentences) was used as cross-evaluation data to run our optimization experiments on the SMT system. Two dialogs (82 sentences) were then used as test data in a comparative evaluation between the SMT system and the Nespole! IL-MT system. The training data fails to cover 29% of the types in this test set, giving a token OOV rate of 11%.

4.2. Evaluation Methods

In our experiments we applied both automatic and manual evaluation. To evaluate our SMT optimization efforts, we used the automatic evaluation metric Bleu score as proposed in [5]. The Bleu score is based on n-gram precisions when comparing the system translation with several human reference translations. As precision without recall favors short translations, a length penalty is combined with the weighted average of those precisions for the final result.

Human evaluation was carried out for the comparative evaluation of the IL-MT and the SMT systems. The evaluators were presented with the German turn and the two translations. Grading was done on a 3-point scale:

- Good: for translations which give the required information and which are easy to understand, i.e. no critical syntactic errors.
- Okay: for translations which give useful information, even if they are syntactically not correct.
- Bad: for missing translations or for translations which give no useful or even misleading information.

For long turns, information units were identified beforehand and the turns segmented accordingly. Human graders then assigned quality scores on a per-segment basis.

4.3. SMT Optimization Experiments

4.3.1. Transducer Configurations

We experimented with five transducers, $\{L, P, P_2, R, M, I\}$. L is the statistical lexicon as it is produced by the HMM alignment program. It contains only word-to-word translations. P represents phrase-level alignments. P_2 is the phrase-level product of bidirectional HMM alignment. R is a transducer for some fixed number and date expressions that was hand-coded for German-English translation. It is domain-independent and reusable. M is constructed from an online German-English lexicon. I is the transducer extracted from the interlingual analysis grammars.

Table 2 shows the effect of combining these transducers on system performance. For each configuration of the translation system the Bleu score is given. The last two columns in the table give corpus coverage, i.e. how many words from the test corpus were translated, and the vocabulary coverage, i.e. how many word types from the test corpus were translated.

Table 2. Evaluation results for cross-evaluation set: text input.

Configuration	Bleu Score	C-Cov	V-Cov
L	0.1893	89.18	70.90
LR	0.1903	89.83	72.12
LM	0.1926	93.27	81.21
LRP	0.2350	90.32	72.72
$LRPI$	0.2434	90.49	73.33
$LRMPI$	0.2432	95.08	85.45
LRP_2	0.2654	90.81	73.93
$LRMP_2$	0.2522	94.91	84.24
LRP_2I	0.2714	90.98	74.54
$LRMP_2I$	0.2613	95.24	85.45

The baseline result of 0.1893 comes from translating with transducer L alone. Adding transducer R gave a small improvement. Transducers P and P_2 gave more significant improvements of 23% and 40%, respectively, over L and R alone. Adding transducer I gave no improvement when added to the baseline system, but accounted for small improvements when used in conjunction with phrase transducers $\{P, P_2\}$.

Transducer M, the background lexicon, gave a large boost in type and token coverage, but translation quality as measured by the Bleu score went down. This points to a problem with adding a general-purpose lexicon: all translation probabilities in the lexicon are equal, and do not reflect the distribution of the training data.

4.3.2. Effect of the Large Language Model

Improvements to the language model were made by retraining it on a larger monolingual corpus. First, the English side of the background lexicon was added. In addition we used data from in the Verbmobil project [6], which is composed of recorded dialogs like the Nespole! data. The Verbmobil corpus is about 500,000 words in size.

Table 3. Effect of large language model.

Configuration	Small LM	Large LM
L	0.1893	0.1782
LM	0.1926	0.2298
$LRMP$	0.2334	0.2703
$LRMP_2$	0.2522	0.3141
$LRMP_2I$	0.2613	0.3172

The results of using this larger language model can be seen in Table 3. For convenience, the results from using the small language model are repeated in this table. The larger language model almost always helped to improve translation quality. The effect is most pronounced in those configurations which use the background lexicon transducer as well.

4.3.3. Background Lexicon as Training Data

In the final experiment the large background lexicon was added to the training corpus for the alignment model. In this way the vocabulary covered in the general-purpose lexicon becomes part of the statistical lexicon transducer L, and the separate background lexicon transducer M is left out.

Results for some transducer configurations are represented in Table 4 and show a clear improvement. Again, the results when translating with the background lexicon as a separate transducer are repeated for comparison.

Table 4. Effect of adding background lexicon to training corpus.

Configuration	Separate	Integrated
LM	0.2298	0.2050
$LRMP$	0.2703	0.2813
$LRMP_2$	0.3141	0.3275
$LRMP_2I$	0.3172	0.3300

4.4. Comparing SMT and IL-MT

To put the performance of the SMT system into perspective we compared it to an existing IL-MT system [4] which was developed as part of the Nespole! project. The Bleu scores and the results from human evaluation are given in Table 5 for text (human transcribed) and speech (speech recognizer transcribed) input. The numbers for ‘Good’, ‘Okay’ and ‘Bad’ translations are the sum of two evaluators. To condense those numbers an average score for the human evaluation was calculated by giving each good translation a score of 1, each okay translation a score of 0.5 and each bad translation a score of 0.0.

Table 5. Evaluation results for IL-MT and SMT.

		Bleu	Good	Okay	Bad	Score
Text	IF	0.068	77	104	227	0.32
	SMT	0.333	124	80	205	0.40
Speech	IF	0.059	64	101	243	0.28
	SMT	0.262	95	83	227	0.34

The Bleu score is much higher for the SMT system than IL-MT system. The human evaluation revealed the same ordering of the systems but with much closer scores. This indicates that the perceptible difference in translation quality is not as large as the Bleu score suggests.

5. CONCLUSIONS

Statistical machine translation is possible even with a small corpus for domain-specific training, provided that additional general-purpose knowledge sources such as manually compiled lexica and larger monolingual data are available. An advantage of the SMT approach is that these sources can be added to the system with very little human effort. Comparing the results from the SMT system with the results of an IL-MT system shows that statistical translation is at least competitive, yielding comparable translation quality in significantly less development time.

6. REFERENCES

[1] 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.

[2] 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.

[3] Stephan Vogel and Hermann Ney, “Translation with cascaded finite state transducers,” in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hongkong, China, October 2000, pp. 23–30.

[4] A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei, and F. Balducci, “Architecture and design considerations in nespole!: a speech translation system for e-commerce applications,” in *Proceedings of HLT: Human Language Technology*, San Diego, CA, March 2001.

[5] Kishore Papinini, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” Tech. Rep. RC22176(W0109-022), September 17, 2001, IBM, 2001.

[6] Wolfgang Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag: Berlin, Heidelberg, New York, 2000.