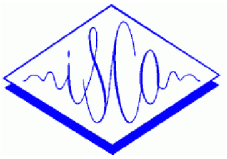

International Workshop on Spoken Language Translation (IWSLT) 2005
Pittsburgh, PA, USA
October 24-25, 2005

The NTT Statistical Machine Translation System for IWSLT2005

Hajime Tsukada, Taro Watanabe, Jun Suzuki, Hideto Kazawa, and Hideki Isozaki

NTT Communication Science Laboratories
{tsukada,taro,jun,kazawa,isozaki}@cslab.kecl.ntt.co.jp

Abstract

This paper reports the NTT statistical translation system participating in the evaluation campaign of IWSLT 2005. The NTT system is based on a phrase translation model and utilizes a large number of features with a log-linear model. We studied the various features recently developed in this research field and evaluate the system using supplied data as well as publicly available Chinese, Japanese, and English data. Despite domain mismatch, additional data helped improve translation accuracy.

1. Introduction

Recently, phrase-based translation combined with other features by log-linear models has become the standard technique for statistical machine translation. Shared task based workshops of machine translation including IWSLT and NIST Machine Translation Workshops showed which features effectively improve translation accuracy. However, it remains unclear whether using of these features all together with our system is helpful.

One unavoidable problem with statistical approaches is training data preparation. Since the amount of training data is generally limited, how to utilize similar monolingual or bilingual resources is an important research topic in statistical machine translation.

In this evaluation campaign, we studied the use of a large number of reportedly effective features with our system and also evaluated both additional monolingual and bilingual corpus to improve translation accuracies.

2. Log-linear Models

Our system adopts the following log-linear decision rule to obtain the maximum likely translation:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \frac{1}{Z(f_1^J)} \exp\left(\sum_j \lambda_j f_j(e_1^I, f_1^J)\right) \quad ,$$

where $f_j(e_1^I, f_1^J)$ represents a feature function and $Z(f_1^J)$ denotes a normalization term. Feature function scaling factors λ_j are efficiently computed based either on the maximum likelihood criterion [1] or the minimum error rate criterion [2]. Our system adopts the latter criterion in the experiments.

One advantage of log-linear models is the ability to easily combine various features relating to translation models, language models, and lexical reordering models. The feature details are described in the following sections.

3. Features

3.1. Language Model Features

In statistical machine translation, improving language models strongly impacts translation accuracy. Especially recently, the power of long-span n-grams and the use of huge amounts of training data have been reported [3]. In this evaluation campaign, we combined several long n-gram language models as features.

To train language models from various corpora in different domains, corpus weighting is necessary to fit the trained language models to the test set domain. Log-linear models naturally provide weighting of a language model trained by each corpus.

We used the following long n-gram language models:

- 6-gram
- Class-based 9-gram
- Prefix-4 9-gram
- Suffix-4 9-gram

All n-grams are based on mixed casing. The prefix-4 (suffix-4) language model takes only 4-letter prefixes (suffixes) of English words. Prefix-4 (suffix-4) roughly means the word stem (inflectional endings). For example, “Would it be possible to ship it to Japan” becomes “Woul+ it be poss+ to ship it to Japa+” by prefix-4, and “+ould it be +ible to ship it to +apan” by suffix-4, where “+” at the end or beginning of a word denotes deletion. Prefix-4 and suffix-4 are likely to contribute to word alignment and language modeling, respectively.

3.2. Phrase-based Features

Our system adopts a phrase-based translation model represented by phrase-based features, which are based on phrase

translation pairs extracted by the method proposed by Och and Ney [4].

First, many-to-many word alignment is set by using both one-to-many and many-to-one word alignments generated by GIZA++ toolkit. In the experiment, we used prefix-4 for word-to-word alignment. Using prefix-4 produced better translations than the original form in preliminary experiments.

Next, phrase pairs consistent with word alignment are extracted. The words in a legal phrase pair are only aligned to each other and not to words outside. Hereafter, we use $\mathrm{count}(\tilde{e})$ and $\mathrm{count}(\tilde{f}, \tilde{e})$ to denote the number of extracted phrase $\tilde{e}$ and extracted phrase pair $(\tilde{f}, \tilde{e})$, respectively.

We used the following features based on extracted phrase pairs:

- Phrase translation probability $\phi(\tilde{e}|\tilde{f})$ and $\phi(\tilde{f}|\tilde{e})$, where
$$\phi(\tilde{e}|\tilde{f}) = \frac{\mathrm{count}(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} \mathrm{count}(\tilde{f}, \tilde{e})}$$
- Frequency of phrase pairs $\mathrm{count}(\tilde{e}, \tilde{f})$, $\mathrm{count}(\tilde{e})$, and $\mathrm{count}(\tilde{f})$
- χ^2 value and Dice coefficient of $\tilde{f}$ and $\tilde{e}$
- Phrase extraction probability of source/target, i.e.,
$$\frac{\text{\# of extracted source/target phrases}}{\text{\# of source/target phrases appearing in the corpus}}$$
- Phrase pair extraction probability, i.e.,
$$\frac{\text{\# of sentences phrase pairs extracted}}{\text{\# of sentences phrase pairs appearing in the corpus}}$$
- Adjusted Dice coefficient, which is an extension of the measure proposed in [5], i.e.,
$$\mathrm{Dice}(\tilde{f}, \tilde{e}) log(\mathrm{count}(\tilde{f}, \tilde{e}) + 1)$$

3.3. Word-level Features

We used the following word-level features, where
$$w(f|e) = \frac{\mathrm{count}(f, e)}{\sum_{f'} \mathrm{count}(f', e)},$$
I is the number of words in the translation and J is the number of words in the input sentence.

- Lexical weight $p_w(\tilde{f}|\tilde{e})$ and $p_w(\tilde{e}|\tilde{f})$ [6], where
$$p_w(\tilde{f}|\tilde{e}) = max_a \prod_{j=1}^{J} \frac{1}{|\{i|(i,j) \in a)\}|} \cdot \sum_{\forall (i,j) \in a} w(f_j|e_i)$$
- IBM Model 1 score $p_{M1}(\tilde{f}|\tilde{e})$ and $p_{M1}(\tilde{e}|\tilde{f})$, where
$$p_{M1}(\tilde{f}|\tilde{e}) = \frac{1}{(\tilde{I}+1)^{\tilde{J}}} \prod_{j}^{\tilde{J}} \sum_{i}^{\tilde{I}} w(\tilde{f}_j|\tilde{e}_i)$$
- Viterbi IBM Model 1 score $p_{M1'}(\tilde{f}|\tilde{e})$ and $p_{M1'}(\tilde{e}|\tilde{f})$, where
$$p_{M1'}(\tilde{f}|\tilde{e}) = \frac{1}{(\tilde{I}+1)^{\tilde{J}}} \prod_{j}^{\tilde{J}} \max_{i} w(\tilde{f}_j|\tilde{e}_i)$$
- Noisy OR gate $p_{NOR}(\tilde{f}|\tilde{e})$ and $p_{NOR}(\tilde{e}|\tilde{f})$ [7], where
$$p_{NOR}(\tilde{f}|\tilde{e}) = \prod_{j}(1 - \prod_{i}(1 - w(\tilde{f}_j|\tilde{e}_i)))$$
- Deletion penalty $p_{del}(\tilde{e}, \tilde{f})$ where
$$p_{del}(\tilde{e}, \tilde{f}) = \sum_{j} del(\tilde{e}_1^{\tilde{I}}, \tilde{f}_j)$$
$$del(\tilde{e}_1^{\tilde{I}}, \tilde{f}_j) = \begin{cases} 1 & i \text{ does not exist s.t.} \\ & w(\tilde{e}_i|\tilde{f}_j) > threshold \\ 0 & \text{otherwise.} \end{cases}$$

3.4. Lexical Reordering Features

We used the following features to control the reordering of phrases:

- Distortion model $d(a_i - b_{i-1}) = \exp^{-|a_i - b_{i-1} - 1|}$, where a_i denotes the starting position of the foreign phrase translated into the i-th English phrase, and b_{i-1} denotes the end position of the foreign phrase translated into the $(i-1)$-th English phrase [6].
- Right monotone model $P_R(\tilde{e}, \tilde{f})$ (and left monotone model $P_L(\tilde{e}, \tilde{f})$) inspired by Och's scheme [8], where
$$P_R(\tilde{f}, \tilde{e}) = \frac{\mathrm{count}_R}{\mathrm{count}(\tilde{f}, \tilde{e})},$$
and count_R denotes the number of right connected monotone phrases.

3.5. Other Features

The following additional features are used.

- number of words that constitute a translation
- number of phrases that constitute a translation

4. Decoder

The decoder is based on word graph [9] and uses a multi-pass strategy to generate n-best translations. It generates hypothesized translations in a left-to-right order by combining phrase translations for a source sentence. The first pass of our decoding algorithm generates a word graph, a compact representation of hypothesized translations, using a breadth-first beam search, as in [10][11][12][13]. Then, n-best translations are extracted from the generated word graph using A^* search.

The search space for a beam search is constrained by restricting the reordering of source phrases. We have window size constraints that restrict the number of words skipped before selecting a segment of the source sequence [6][12]. An ITG-constraint [14] is also implemented that prohibits the extension of a hypothesis that violates ITG constraints, which will be useful for language pairs with drastic reordering, such as Japanese-to-English and Korean-to-English translations.

During the beam search stage, three kinds of pruning are performed to further reduce the search space [11]. First, observation pruning limits the number of phrase translation candidates to a maximum of N candidates. Second, threshold pruning is performed by computing the most likely partial hypothesis and by discarding hypotheses whose probability is lower than the maximum score multiplied with a threshold. Third, histogram pruning is carried out by restricting the number of hypotheses to a maximum of M candidates. Observation and threshold pruning are also applied to the back pointer to reduce the size of the word graph. In pruning hypotheses, future cost is also estimated on the fly and then integrated with the preceding score for beam pruning.

We estimated future cost as described in [13]. Exact costs for the phrase-based features and word level features can be calculated for each extracted phrase pair. For the language model features, their costs were approximated by using only output words contained by each phrase pair. The upper bound of lexical reordering feature costs can be computed beforehand by considering the possible permutations of phrase pairs for a given input.

After generating a word graph, it is then pruned using the posterior probabilities of edges [15] to further reduce the number of duplicate translations for A^* search. An edge is pruned if its posterior score is lower than the highest posterior score in the graph by a certain amount.

5. Experiments

To validate the use of the reportedly effective features, we conducted translation experiments using all features introduced in Section 3. Also, we conducted comparable experiments in both supplied and unrestricted data tracks to study the effectiveness of additional language resources.

English data sets	Corpus size (words)
IWSLT (supplied)	190,177
ATR	1,100,194
WEB	8,482,782
Gigaword	1,799,531,558

Table 1: Monolingual corpora for unrestricted data track

5.1. Corpus Preparation

To obtain comparable results for all source and target language pairs, we concentrated on tracks generating English, i.e., Japanese-to-English, Chinese-to-English, Arabic-to-English, and Korean-to-English.

The English parts of the corpora are tokenized using LDC's standards. For Arabic, it is simply tokenized by splitting punctuation and then removing Arabic characters denoting "and". For other languages, supplied segmentation is used. For unrestricted data tracks, Japanese is segmented using ChaSen [1], and Chinese is segmented using an LDC segmenter with lexicon entries gathered from supplied data and an LDC corpus. Test sets including ASR 1-best are also re-segmented in the same manner to maintain segmentation consistency.

We used mixed casing and prefix-4 form for word-to-word alignment in the phrase extraction. Also, mixed casing was used for training n-grams.

5.2. Language Models

6-gram language models and class-based/prefix-4/suffix-4 9-gram models trained by the SRI language modeling toolkit [16] were used in both supplied and unrestricted data tracks.

We used the following additional monolingual corpora for language models of unrestricted data tracks: (i) ATR Spoken Language Database publically available from ATR[2]; (ii) Web pages crawled from discussion groups and FAQs about travel; and (iii) English Gigaword corpus from LDC.

As additional bilingual corpora for translation models of unrestricted data tracks, we used the ATR Spoken Language Database for Japanese-to-English translation and the two largest corpora in the LDC collection, LDC2004T08 and LDC2005T10, for Chinese-to-English translation. No additional resources were used for other language pairs. Tables 1 and 2 illustrate the data size of each corpus.

Using the monolingual corpora, a total of 10 n-grams were trained and used as a feature of log-linear models when decoding. Table 3 shows the output language perplexity of each n-gram used in the decoder. On the other hand, Table 4 shows the input language perplexity of the trigram trained by the supplied corpora. Tables 3 and 4 suggest that the ATR

[1] http://chasen.naist.jp

[2] http://www.red.atr.jp/product/index.html

Data sets	Language pairs	Corpus size (English words)
IWSLT (supplied)	JE/CE/AE/KE	190,177
ATR	JE	1,334,852
LDC	CE	76,939,292
-	AE	-
-	KE	-

Table 2: Bilingual corpora for unrestricted data track

Test sets	Japanese		Chinese	
	IWSLT	ATR	IWSLT	LDC
devset1	16.9	29.5	56.6	462
devset2	17.6	32.9	56.1	449
testset	24.5	28.6	50.7	432

Table 4: Input language perplexity of trigram trained by supplied corpora

and IWSLT datasets are similar, WEB is closer to IWSLT than Gigaword, and that LDC is very different from IWSLT.

Since the collection is enormous in Gigaword, the vocabulary set is first limited to that observed in the English part of supplied corpus and the ATR database. Then for decoding, an actual n-gram language model is estimated on the fly by constraining the vocabulary set to that observed in a given test set.

5.3. Other Setups

Following one of the best systems [17] in IWSLT 2004, feature function scaling factors λ_j are trained using NIST scores [18] in a loss function of minimum error rate training, and development set 1 (CSTAR) was used for it.

For Japanese and Korean, ITG constraints of lexical reordering were applied, and for Arabic and Chinese, simple window size constraints up to 7 were used.

5.4. Results

Table 5 summarizes the overall results of the supplied/unrestricted data tracks. The scores of the table are obtained by the comparable conditions for each language pair while some are not the same as those released by the organizer.

"^{m}unrestricted" denotes that monolingual corpora are unrestricted but bilingual corpora are restricted; "^{m}bunrestricted" denotes that both monolingual and bilingual corpora are unrestricted.

The table shows that unrestricted data tracks consistently outperform restricted data tracks except for Japanese-to-English with ASR output. This may be because resegmentation of the ASR output produces bad segmentation because of ASR errors.

"^{mb}unrestricted" is inferior to "^{m}unrestricted" in Chinese-to-English translation whereas the former is better than the latter in Japanese-to-English translation. This may be because the additional bilingual resources are similar in Japanese-to-English but are different in Chinese-to-English as shown in Tables 3 and 4.

The overall results suggest that our feature design could not deal with domain mismatch of bilingual corpora but could deal with small mismatch of monolingual corpora. While Gigaword differs most from the supplied corpus in terms of perplexity, as shown in Table 3, its n-gram surprisingly contributes more to translation than other n-grams in terms of feature function scaling factors of log-linear models. It would be interesting to study this finding in more detail.

6. Conclusion

The NTT statistical machine translation system in the evaluation campaign is reported. A log-linear model naturally enabled weighting of various features including language models. As a result, we obtained competitive accuracies. The log-linear model effectively utilized n-grams trained by out-of-domain corpora, and we improved the translation accuracy of the supplied data.

These experiments simply used all available features. However, feature extraction may additionally improve translation accuracy. It is worth studying.

Compared to other sites, our results are better in terms of NIST scores but inferior in terms of BLEU scores. This is because feature function scaling factors are trained by a loss function based on NIST scores. We also doubt the overfitting of feature function scaling factors. We need to continue studying both training methods of the scaling factors and loss functions to improve other translation metrics as well as NIST scores.

7. Acknowledgements

Our training tool for phrase translation models is an extension of that provided by Philipp Koehn under a contract of MIT-NTT collaboration.

8. References

[1] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2002, pp. 295–302.

[2] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the 41th Annual Meet-*

Test sets	6-gram				class-9-gram		prefix4-9-gram		suffix4-9-gram	
	ATR	IWSLT	WEB	GIGA	IWSLT	IWSLT+ATR	IWSLT	IWSLT+ATR	IWSLT	IWSLT+ATR
devset1	37.7	41.0	81.2	93.8	40.7	41.1	41.5	34.0	40.0	32.6
devset2	41.1	44.3	88.4	92.0	45.0	44.8	44.5	36.4	42.9	35.2

Table 3: Output language perplexity of n-grams for decoding

Language pairs	Translation input	Training data	BLEU scores	NIST scores
AE	transcription	supplied	0.4350	9.1821
		^{m}unrestricted	0.4764	9.3674
CE	transcription	supplied	0.3275	8.0768
		^{m}unrestricted	0.4112	8.8418
		^{mb}unrestricted	0.3943	8.6804
	ASR 1-best	supplied	0.2739	6.5185
		^{mb}unrestricted	0.2965	6.9416
JE	transcription	supplied	0.3669	7.9669
		^{m}unrestricted	0.3679	8.1207
		^{mb}unrestricted	0.3932	8.6442
	ASR 1-best	supplied	0.3881	8.3855
		^{mb}unrestricted	0.3762	8.3502
KE	transcription	supplied	0.3218	7.8489
		^{m}unrestricted	0.3497	8.0160

Table 5: NTT results of evaluation campaign

ing of the Association for Computational Linguistics (ACL), July 2003, pp. 160–167.

[3] ——, "The Google statistical machine translation system for the 2005 NIST MT evaluation (unpublished)," in *Machine Translation Workshop*, 2005.

[4] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, December 2004.

[5] M. Kitamura and Y. Matsumoto, "Automatic extraction of translation patterns in parallel corpora," *IPSJ Transactions*, vol. 38, no. 4, pp. 727–736, 1997.

[6] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, May–June 2003, pp. 127–133.

[7] R. Zens and H. Ney, "Improvements in phrase-based statistical machine translation," in *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, May 2004, pp. 257–264.

[8] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "A smorgasbord of features for statistical machine translation," in *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, May 2004, pp. 161–168.

[9] N. Ueffing, F. J. Och, and H. Ney, "Generation of word graphs in statistical machine translation," in *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, July 2002, pp. 156–163.

[10] Y.-Y. Wang and A. Waibel, "Decoding algorithm in statistical machine translation," in *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[11] C. Tillmann and H. Ney, "Word reordering and a dynamic programming beam search algorithm for statistical machine translation," *Computational Linguistics*, vol. 29, no. 1, pp. 97–133, March 2003.

[12] S. Vogel, Y. Zhang, F. Huang, A. Venugopal, B. Zhao, A. Tribble, M. Eck, and A. Waibel, "The CMU statistical machine translation system," in *Proc. of MT Summit IX*, September 2003.

[13] P. Koehn, *PHARAOH: User manual and description for version 1.2*, UCS Information Science Institute, August 2004.

[14] R. Zens, E. Matusov, and H. Ney, "Improved word alignment using a symmetric lexicon model," in *Proc. of 20th International Conference on Computational Linguistics (COLING)*, August 2004, pp. 36–42.

[15] R. Zens and H. Ney, "Word graphs for statistical machine translation," in *Proc. of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, June 2005, pp. 191–198.

[16] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of 7th International Conference on Spoken Language Processing*, 2002.

[17] O. Bender, R. Zens, E. Matusov, and H. Ney, "Alignment templates: the RWTH SMT system," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 79–84.

[18] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. of HLT 2002*, 2002.