A Phrase-Based Unigram Model for Statistical Machine Translation

Christoph Tillmann and Fei Xia
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{ctill,feixia}@us.ibm.com

Abstract

In this paper, we describe a phrase-based unigram model for statistical machine translation that uses a much simpler set of model parameters than similar phrase-based models. The units of translation are blocks - pairs of phrases. During decoding, we use a block unigram model and a word-based trigram language model. During training, the blocks are learned from source interval projections using an underlying word alignment. We show experimental results on block selection criteria based on unigram counts and phrase length.

1 Phrase-based Unigram Model

Various papers use phrase-based translation systems (Och et al., 1999; Marcu and Wong, 2002; Yamada and Knight, 2002) that have shown to improve translation quality over single-word based translation systems introduced in (Brown et al., 1993). In this paper, we present a similar system with a much simpler set of model parameters. Specifically, we compute the probability of a block sequence b_1^n. The block sequence probability $Pr(b_1^n)$ is decomposed into conditional probabilities using the chain rule:

$$\begin{aligned} Pr(b_1^n) &\approx \prod_{i=1}^{n} Pr(b_i|b_{i-1}) \qquad (1) \\ &= \prod_{i=1}^{n} p^{\alpha}(b_i|b_{i-1}) \cdot p^{(1-\alpha)}(b_i|b_{i-1}) \\ &\approx \prod_{i=1}^{n} p^{\alpha}(b_i) \cdot p^{(1-\alpha)}(b_i|b_{i-1}) \end{aligned}$$

We try to find the block sequence that maximizes $Pr(b_1^n)$: $b_1^n = \arg\max_{b_1^n} Pr(b_1^n)$. The model proposed is a joint model as in (Marcu and Wong, 2002), since target and source phrases are generated jointly. The approach is illustrated in Figure 1. The source phrases are given on the x-axis and the target phrases are given on the y-axis.

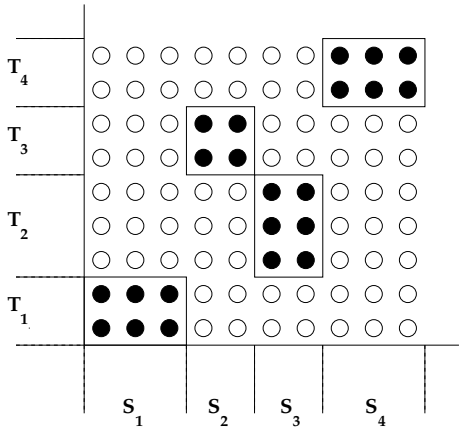

Figure 1: A block sequence that jointly generates 4 target and source phrases.

The two types of parameters in Eq 1 are defined as:

- **Block unigram model $p(b_i)$:** we compute unigram probabilities for the blocks. The blocks are simpler than the alignment templates in (Och et al., 1999) in that they do not have any internal structure.
- **Trigram language model:** the probability $p(b_i|b_{i-1})$ between adjacent blocks is computed as the probability of the first target word in the target clump of b_i given the final two words of the target clump of b_{i-1}.

The exponent α is set in informal experiments to be 0.5. No other parameters such as distortion probabilities are used.

To select blocks b from training data, we compute unigram block co-occurrence counts $N(b)$. $N(b)$ cannot be

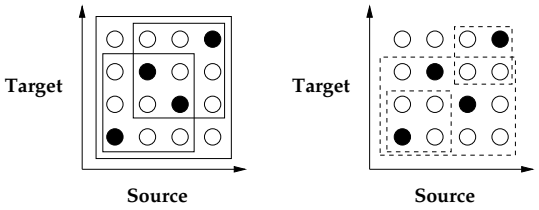

Figure 2: The left picture shows three blocks that are learned from projecting three source intervals. The right picture shows three blocks that cannot be obtain from source interval projections .

computed for all blocks in the training data: we would obtain hundreds of millions of blocks. The blocks are restricted by an underlying word alignment. The word alignment is obtained from an HMM Viterbi training (Vogel et al., 1996). The HMM Viterbi training is carried out twice with English as target language and Chinese as source language and vice versa. We take the intersection of the two alignments as described in (Och et al., 1999). To generate blocks from the intersection, we proceed as follows: for each source interval $[j, j']$, we compute the minimum target index i and maximum target index i' of the intersection alignment points that fall into the interval $[j, j']$. The approach is illustrated in Figure 2. In the left picture, for example, the source interval $[1, 3]$ is projected into the target interval $[1, 3]$. The pair $([j, j'], [i, i'])$ together with the words at the corresponding positions yields a block learned from this training sentence pair. For source intervals without alignment points in them, no blocks are produced. We also extend a block corresponding to the interval pair $([j, j'], [i, i'])$ by elements on the union of the two Viterbi HMM alignments. A similar block selection scheme has been presented in (Och et al., 1999). Finally, the target and source phrases are restricted to be equal or less than 8 words long. This way we obtain 23 millions blocks on our training data including blocks that occur only once. This baseline set is further filtered using the unigram count $N(b)$: Nk denotes the set of blocks b for which $N(b) \geq k$. Blocks where the target and the source clump are of length 1 are kept regardless of their count.[1] We compute the unigram probability $p(b)$ as relative frequency over all selected blocks.

We also tried a more restrictive projection scheme: source intervals are projected into target intervals and the reverse projection of the target interval has to be included in the original source interval. The results for this symmetrical projection are currently worse, since some blocks with longer target intervals are excluded. An example of 4 blocks obtained from the training data is shown in

[1]To apply the restrictions exhaustively, we have implemented tree-based data structures to store the 23 million blocks with phrases of up to length 8 in about 1.6 gigabyte of RAM.

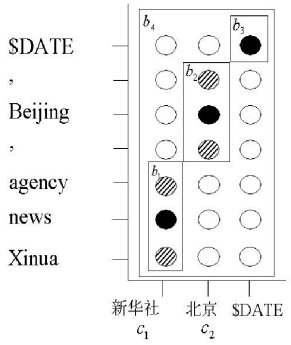

Figure 3: An example of 4 recursively nested blocks b_1, b_2, b_3, b_4.

Figure 3. '$DATE' is a placeholder for a date expression. Block b_4 contains the blocks b_1 to b_3. All 4 blocks are selected in training: the unigram decoder prefers b_4 even if b_1,b_2, and b_3 are much more frequent. The solid alignment points are elements from the intersection, the striped alignment points are elements from the union. Using the union points, we can learn one-to-many block translations; for example, the pair (c_1,'Xinhua news agency') is learned from the training data.

We use a DP-based beam search procedure similar to the one presented in (Tillmann, 2001). We maximize over all block segmentations b_1^n for which the source phrases yield a segmentation of the input source sentence, generating the target sentence simultaneously. In the current experiments, decoding without block re-ordering yields the best translation results. The decoder translates about 180 words per second.

2 Experimental Results

The translation system is tested on a Chinese-to-English translation task. The training data come from several news sources. For testing, we use the DARPA/NIST MT 2001 dry-run testing data, which consists of 793 sentences with $20,333$ words arranged in 80 documents.[2] The training data is provided by the LDC and labeled by NIST as the Large Data condition for the MT 2002 evaluation. The Chinese sentences are segmented into words. The training data contains 23.7 million Chinese and 25.3 million English words.

Experimental results are presented in Table 1 and Table 2. Table 1 shows the effect of the unigram threshold. The second column shows the number of blocks selected. The third column reports the BLEU score (Papineni et al., 2002) along with 95% confidence interval. We use IBM

[2]We did not use the first 25 documents of the 105-document dry-run test set because they were used as a development test set before the dry-run and were subsequently added to our training data.

Table 1: Effect of the unigram threshold on the BLEU score. The maximum phrase length is 8.

Selection Restriction	# blocks selected	BLEUr4n4
IBM1 baseline	1.23M	0.11 ± 0.01
N2	4.23 M	0.18 ± 0.02
N3	1.22 M	0.18 ± 0.01
N4	0.84 M	0.17 ± 0.01
N5	0.65 M	0.17 ± 0.01

Table 2: Effect of the maximum phrase length on the BLEU score. The unigram threshold is $N(b) \geq 2$.

maximum phrase length	# blocks selected	BLEUr4n4
8	4.23 M	0.18 ± 0.02
7	3.76 M	0.17 ± 0.02
6	3.26 M	0.17 ± 0.01
5	2.73 M	0.17 ± 0.01
4	2.16 M	0.17 ± 0.01
3	1.51 M	0.16 ± 0.01
2	0.77 M	0.14 ± 0.01
1	0.16 M	0.12 ± 0.01

Model 1 as a baseline model which is similar to our block model: neither model uses distortion or alignment probabilities. The best results are obtained for the $N2$ and the $N3$ sets.

The $N3$ set uses only 1.22 million blocks in contrast to $N2$ which has 4.23 million blocks. This indicates that the number of blocks can be reduced drastically without affecting the translation performance significantly. Table 2 shows the effect of the maximum phrase length on the BLEU score for the $N2$ block set. Including blocks with longer phrases actually helps to improve performance, although length 4 already obtains good results.

We also ran the N2 on the June 2002 DARPA TIDES Large Data evaluation test set. Six research sites and four commercial off-the-shelf systems were evaluated in Large Data track. A majority of the systems were phrase-based translation systems. For comparison with other sites, we quote the NIST score (Doddington, 2002) on this test set: N2 system scores 7.44 whereas the official top two systems scored 7.65 and 7.34 respectively.

3 Conclusion

In this paper, we described a phrase-based unigram model for statistical machine translation. The model is much simpler than other phrase-based statistical models. We experimented with different restrictions on the phrases selected from the training data. Longer phrases which occur less frequently do not help much.

Acknowledgment

This work was partially supported by DARPA and monitored by SPAWAR under contract No. N66001-99-2-8916.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the Second International Conference of Human Language Technology Research*, pages 138–145, March.

Daniel Marcu and William Wong. 2002. A Phrased-Based, Joint Probability Model for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP 02)*, pages 133–139, Philadelphia, PA, July.

Franz-Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 99)*, pages 20–28, College Park, MD, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of machine translation. In *Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Christoph Tillmann. 2001. *Word Re-Ordering and Dynamic Programming based Search Algorithm for Statistical Machine Translation*. Ph.D. thesis, University of Technology, Aachen, Germany.

Stefan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM BasedWord Alignment in Statistical Machine Translation. In *Proc. of the 16th Int. Conf. on Computational Linguistics (COLING 1996)*, pages 836–841, Copenhagen, Denmark, August.

Kenji Yamada and Kevin Knight. 2002. A Decoder for Syntax-based Statistical MT. In *Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02)*, pages 303–310, Philadelphia, PA, July.