Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT

Mei Yang
Dept. of Electrical Engineering
University of Washington, Seattle, WA, USA
yangmei@u.washington.edu

Jing Zheng
SRI International
Menlo Park, CA, USA
zj@speech.sri.com

Abstract

We investigate the use of Fisher's exact significance test for pruning the translation table of a hierarchical phrase-based statistical machine translation system. In addition to the significance values computed by Fisher's exact test, we introduce compositional properties to classify phrase pairs of same significance values. We also examine the impact of using significance values as a feature in translation models. Experimental results show that 1% to 2% BLEU improvements can be achieved along with substantial model size reduction in an Iraqi/English two-way translation task.

1 Introduction

Phrase-based translation (Koehn et al., 2003) and hierarchical phrase-based translation (Chiang, 2005) are the state of the art in statistical machine translation (SMT) techniques. Both approaches typically employ very large translation tables extracted from word-aligned parallel data, with many entries in the tables never being used in decoding. The redundancy of translation tables is not desirable in real-time applications, e.g., speech-to-speech translation, where speed and memory consumption are often critical concerns. In addition, some translation pairs in a table are generated from training data errors and word alignment noise. Removing those pairs could lead to improved translation quality.

(Johnson et al., 2007) has presented a technique for pruning the phrase table in a phrase-based SMT system using Fisher's exact test. They compute the significance value of each phrase pair and prune the table by deleting phrase pairs with significance values smaller than a threshold. Their experimental results show that the size of the phrase table can be greatly reduced with no significant loss in translation quality.

In this paper, we extend the work in (Johnson et al., 2007) to a hierarchical phrase-based translation model, which is built on synchronous context-free grammars (SCFG). We call an SCFG rule a phrase pair if its right-hand side does not contain a nonterminal, and otherwise a rewrite rule. Our approach applies to both the phrase table and the rule table. To address the problem that many translation pairs share the same significance value from Fisher's exact test, we propose a refined method that combines significance values and compositional properties of surface strings for pruning the phrase table. We also examine the effect of using the significance values as a feature in translation models.

2 Fisher's exact test for translation table pruning

2.1 Significance values by Fisher's exact test

We briefly review the approach for computing the significance value of a translation pair using Fisher's exact test. In Fisher's exact test, the significance of the association of two items is measured by the probability of seeing the number of co-occurrences of the two items being the same as or higher than the one observed in the sample. This probability is referred to as the p-value. Given a parallel corpus consisting of N sentence pairs, the probability of seeing a pair of phrases (or rules) $(\tilde{s}, \tilde{t})$ with the joint frequency $C(\tilde{s}, \tilde{t})$ is given by the hypergeometric distribution

$$P_h(C(\tilde{s},\tilde{t})) = \frac{C(\tilde{s})!(N - C(\tilde{s}))!C(\tilde{t})!(N - C(\tilde{t}))!}{N!C(\tilde{s},\tilde{t})!C(\tilde{s},\neg\tilde{t})!C(\neg\tilde{s},\tilde{t})!C(\neg\tilde{s},\neg\tilde{t})!}$$

where $C(\tilde{s})$ and $C(\tilde{t})$ are the marginal frequencies of $\tilde{s}$ and $\tilde{t}$, respectively. $C(\tilde{s}, \neg\tilde{t})$ is the number of sentence pairs that contain $\tilde{s}$ on the source side

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 237–240,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

but do not contain $\tilde{t}$ on the target side, and similar for the definition of $C(\neg\tilde{s}, \tilde{t})$ and $C(\neg\tilde{s}, \neg\tilde{t})$. The p-value is therefore the sum of the probabilities of seeing the two phrases (or rules) occur as often as or more often than $C(\tilde{s}, \tilde{t})$ but with the same marginal frequencies

$$P_v(C(\tilde{s}, \tilde{t})) = \sum_{c=C(\tilde{s},\tilde{t})}^{\infty} P_h(c)$$

In practice, p-values can be very small, and thus negative logarithm p-values are often used instead as the measure of significance. In the rest of this paper, the negative logarithm p-value is referred to as the *significance value*. Therefore, the larger the value, the greater the significance.

2.2 Table pruning with significance values

The basic scheme to prune a translation table is to delete all translation pairs that have significance values smaller than a given threshold.

However, in practice, this pruning scheme does not work well with phrase tables, as many phrase pairs receive the same significance values. In particular, many phrase pairs in the phrase table have joint and both marginal frequencies all equal to 1. Such phrase pairs are referred to as *triple-1* pairs. It can be shown that the significance value of triple-1 phrase pairs is $log(N)$. Given a threshold, triple-1 phrase pairs either all remain in the phrase table or are discarded entirely.

To look closer at the problem, Figure 1 shows two example tables with their percentages of phrase pairs that have higher, equal, or lower significance values than $log(N)$. When the threshold is smaller than $log(N)$, as many as 35% of the phrase pairs can be deleted. When the threshold is greater than $log(N)$, at least 90% of the phrase pairs will be discarded. There is no threshold that prunes the table in the range of 35% to 90%. One may think that it is right to delete all triple-1 phrase pairs as they occur only once in the parallel corpus. However, it has been shown in (Moore, 2004) that when a large number of singleton-singleton pairs, such as triple-1 phrase pairs, are observed, most of them are not due to chance. In other words, most triple-1 phrase pairs are significant and it is likely that the translation quality will decline if all of them are discarded. Therefore, using significance values alone cannot completely resolve the problem of phrase table pruning. To further discriminate phrase pairs of the same significance values, particularly the triple-1 phrase pairs, more information is needed.

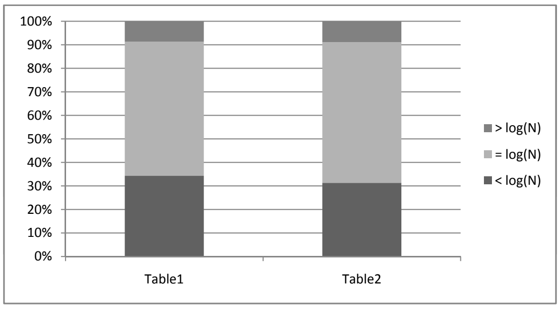

Figure 1: Percentages of phrase pairs with higher, equal, and lower significance values than $log(N)$.

The Fisher's exact test does not consider the surface string in phrase pairs. Intuitively, some phrase pairs are less important if they can be constructed by other phrase pairs in the decoding phase, while other phrase pairs that involve complex syntactic structures are usually difficult to construct and thus become more important. This intuition inspires us to explore the compositional property of a phrase pair as an additional factor. More formally, we define the compositional property of a phrase pair as the capability of decomposing into subphrase pairs. If a phrase pair $(\tilde{s}, \tilde{t})$ can be decomposed into K subphrase pairs $(\tilde{s}_k, \tilde{t}_k)$ already in the phrase table such that

$$\tilde{s} = \tilde{s}_1\tilde{s}_2 \dots \tilde{s}_K$$
$$\tilde{t} = \tilde{t}_1\tilde{t}_2 \dots \tilde{t}_K$$

then this phrase pair is compositional; otherwise it is noncompositional. Our intuition suggests that noncompositional phrase pairs are more important as they cannot be generated by concatenating other phrase pairs in order in the decoding phase. This leads to a refined scheme for pruning the phrase table, in which a phrase pair is discarded when it has a significance value smaller than the threshold and it is not a noncompositional triple-1 phrase pair. The definition of the compositional property does not allow re-ordering. If re-ordering is allowed, all phrase pairs will be compositional as they can always be decomposed into pairs of single words.

In the rule table, however, the percentage of triple-1 pairs is much smaller, typically less than 10%. This is because rules are less sparse than phrases in general, as they are extracted with a shorter length limit, and have nonterminals that match any span of words. Therefore, the basic pruning scheme works well with rule tables.

3 Experiment

3.1 Hierarchical phrase-based SMT system

Our hierarchical phrase-based SMT system translates from Iraqi Arabic (IA) to English (EN) and vice versa. The training corpus consists of 722K aligned Iraqi and English sentence pairs and has 5.0M and 6.7M words on the Iraqi and English sides, respectively. A held-out set with 18K Iraqi and 19K English words is used for parameter tuning and system comparison. The test set is the TRANSTAC June08 offline evaluation data with 7.4K Iraqi and 10K English words, and the translation quality is evaluated by case-insensitive BLEU with four references.

3.2 Results on translation table pruning

For each of the two translation directions IA-to-EN and EN-to-IA, we pruned the translation tables as below, where α represents the significance value of triple-1 pairs and ε is a small positive number. Phrase table *PTABLE3* is obtained using the refined pruning scheme, and others are obtained using the basic scheme. Figure 2 shows the percentages of translation pairs in these tables.

- *PTABLE0*: phrase table of full size without pruning.
- *PTABLE1*: pruned phrase table using the threshold $\alpha - \varepsilon$ and thus all triple-1 phrase pairs remain.
- *PTABLE2*: pruned phrase table using the threshold $\alpha + \varepsilon$ and thus all triple-1 phrase pairs are discarded.
- *PTABLE3*: pruned phrase table using the threshold $\alpha + \varepsilon$ and the refined pruning scheme. All but noncompositional triple-1 phrase pairs are discarded.
- *RTABLE0*: rule table of full size without pruning.
- *RTABLE1*: pruned rule table using the threshold $\alpha + \varepsilon$.

Since a hierarchical phrase-based SMT system requires a phrase table and a rule table at the same time, performance of different combinations of phrase and rule tables is evaluated. The baseline system will be the one using the full-size tables of *PTABLE0* and *RTABLE0*. Tables 2 and 3 show the BLEU scores for each combination in each direction, with the best score in bold.

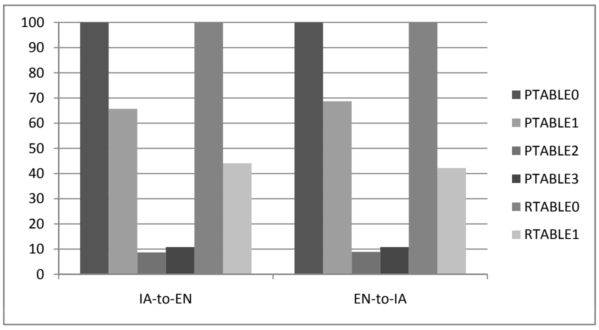

Figure 2: The percentages of translation pairs in phrase and rule tables.

It can be seen that pruning leads to a substantial reduction in the number of translation pairs. As long phrases are more frequently pruned than short phrases, the actual memory saving is even more significant. It is surprising to see that using pruned tables improves the BLEU scores in many cases, probably because a smaller translation table generalizes better on an unseen test set, and some translation pairs created by erroneous training data are dropped. Table 1 shows two examples of discarded phrase pairs and their frequencies. Both of them are incorrect due to human translation errors.

We note that using the pruned rule table *RTABLE1* is very effective and improved BLEU in most cases except when used with *PTABLE0* in the direction EN-to-IA. Although using the pruned phrase tables had mixed effect, *PTABLE3*, which is obtained through the refined pruning scheme, outperformed others in all cases. This confirms the hypothesis that noncompositional phrase pairs are important and thus suggests that the proposed compositional property is a useful measure of phrase pair quality. Overall, the best results are achieved by using the combination of *PTABLE3* and *RTABLE1*, which gave improvement of 1% to 2% BLEU over the baseline systems. Meanwhile, this combination is also twice faster than the baseline system in decoding.

3.3 Results on using significance values as a feature

The p-value of each translation pair can be used as a feature in the log-linear translation model, to penalize those less significant phrase pairs and rewrite rules. Since component feature values cannot be zero, a small positive number was added to p-values to avoid infinite log value. The results of using p-values as a feature with different combinations of phrase and rule tables are shown in

Iraqi Arabic phrase	English phrase in data	Correct English phrase	Frequencies
إحنا خمسة	there are **four** of us	there are **five** of us	1, 29, 1
الشباب ثلاثة أو أربع	young men three **of** four	young men three **or** four	1, 1, 1

Table 1: Examples of pruned phrase pairs and their frequencies $C(\tilde{s}, \tilde{t})$, $C(\tilde{s})$, and $C(\tilde{t})$.

	RTABLE0	RTABLE1
PTABLE0	47.38	48.40
PTABLE1	47.05	48.45
PTABLE2	47.50	48.70
PTABLE3	47.81	**49.43**

Table 2: BLEU scores of IA-to-EN systems using different combinations of phrase and rule tables.

	RTABLE0	RTABLE1
PTABLE0	29.92	29.05
PTABLE1	29.62	30.60
PTABLE2	29.87	30.57
PTABLE3	30.62	**31.27**

Table 3: BLEU scores of EN-to-IA systems using different combinations of phrase and rule tables.

	RTABLE0	RTABLE1
PTABLE0	47.72	47.96
PTABLE1	46.69	48.75
PTABLE2	47.90	48.48
PTABLE3	47.59	**49.50**

Table 4: BLEU scores of IA-to-EN systems using the feature of p-values in different combinations.

	RTABLE0	RTABLE1
PTABLE0	29.33	30.44
PTABLE1	30.28	30.99
PTABLE2	30.38	31.44
PTABLE3	30.74	**31.64**

Table 5: BLEU scores of EN-to-IA systems using the feature of p-values in different combinations.

Tables 4 and 5. We can see that the results obtained by using the full rule table with the feature of p-values (the columns of *RTABLE0* in Tables 4 and 5) are much worse than those obtained by using the pruned rule table without the feature of p-values (the columns of *RTABLE1* in Tables 2 and 3). This suggests that the use of significance values as a feature in translation models is not as efficient as the use in translation table pruning. Modest improvement was observed in the direction EN-to-IA when both pruning and the feature of p-values are used (compare the columns of *RTABLE1* in Tables 3 and 5) but not in the direction IA-to-EN. Again, the best results are achieved by using the combination of *PTABLE3* and *RTABLE1*.

4 Conclusion

The translation quality and speed of a hierarchical phrase-based SMT system can be improved by aggressive pruning of translation tables. Our proposed pruning scheme, which exploits both significance values and compositional properties, achieved the best translation quality and gave improvements of 1% to 2% on BLEU when compared to the baseline system with full-size tables. The use of significance values in translation table pruning and in translation models as a feature has a different effect: the former led to significant improvement, while the latter achieved only modest or no improvement on translation quality.

5 Acknowledgements

Many thanks to Kristin Precoda and Andreas Kathol for valuable discussion. This work is supported by DARPA, under subcontract 55-000916 to UW under prime contract NBCHD040058 to SRI International.

References

Philipp Koehn, Franz J. Och and Daniel Marcu. 2003. *Statistical phrase-based translation.* Proceedings of HLT-NAACL, 48-54, Edmonton, Canada.

David Chiang. 2005. *A hierarchical phrase-based model for statistical machine translation.* Proceedings of ACL, 263-270, Ann Arbor, Michigan, USA.

J Howard Johnson, Joel Martin, George Foster and Roland Kuhn. 2007. *Improving Translation Quality by Discarding Most of the Phrasetable.* Proceedings of EMNLP-CoNLL, 967-975, Prague, Czech Republic.

Robert C. Moore. 2004. *On Log-Likelihood-Ratios and the Significance of Rare Events.* Proceedings of EMNLP, 333-340, Barcelona, Spain