Phrasetable Smoothing for Statistical Machine Translation

George Foster and **Roland Kuhn** and **Howard Johnson**
National Research Council Canada
Ottawa, Ontario, Canada
firstname.lastname@nrc.gc.ca

Abstract

We discuss different strategies for smoothing the phrasetable in Statistical MT, and give results over a range of translation settings. We show that any type of smoothing is a better idea than the relative-frequency estimates that are often used. The best smoothing techniques yield consistent gains of approximately 1% (absolute) according to the BLEU metric.

1 Introduction

Smoothing is an important technique in statistical NLP, used to deal with perennial data sparseness and empirical distributions that overfit the training corpus. Surprisingly, however, it is rarely mentioned in statistical Machine Translation. In particular, state-of-the-art phrase-based SMT relies on a *phrasetable*—a large set of ngram pairs over the source and target languages, along with their translation probabilities. This table, which may contain tens of millions of entries, and phrases of up to ten words or more, is an excellent candidate for smoothing. Yet very few publications describe phrasetable smoothing techniques in detail.

In this paper, we provide the first systematic study of smoothing methods for phrase-based SMT. Although we introduce a few new ideas, most methods described here were devised by others; the main purpose of this paper is not to invent new methods, but to compare methods. In experiments over many language pairs, we show that smoothing yields small but consistent gains in translation performance. We feel that this paper only scratches the surface: many other combinations of phrasetable smoothing techniques remain to be tested.

We define a phrasetable as a set of source phrases (ngrams) $\tilde{s}$ and their translations $\tilde{t}$, along with associated translation probabilities $p(\tilde{s}|\tilde{t})$ and $p(\tilde{t}|\tilde{s})$. These conditional distributions are derived from the joint frequencies $c(\tilde{s}, \tilde{t})$ of source/target phrase pairs observed in a word-aligned parallel corpus.

Traditionally, maximum-likelihood estimation from relative frequencies is used to obtain conditional probabilities (Koehn et al., 2003), eg, $p(\tilde{s}|\tilde{t}) = c(\tilde{s}, \tilde{t}) / \sum_{\tilde{s}} c(\tilde{s}, \tilde{t})$ (since the estimation problems for $p(\tilde{s}|\tilde{t})$ and $p(\tilde{t}|\tilde{s})$ are symmetrical, we will usually refer only to $p(\tilde{s}|\tilde{t})$ for brevity). The most obvious example of the overfitting this causes can be seen in phrase pairs whose constituent phrases occur only once in the corpus. These are assigned conditional probabilities of 1, higher than the estimated probabilities of pairs for which much more evidence exists, in the typical case where the latter have constituents that co-occur occasionally with other phrases. During decoding, overlapping phrase pairs are in direct competition, so estimation biases such as this one in favour of infrequent pairs have the potential to significantly degrade translation quality.

An excellent discussion of smoothing techniques developed for ngram language models (LMs) may be found in (Chen and Goodman, 1998; Goodman, 2001). Phrasetable smoothing differs from ngram LM smoothing in the following ways:

- Probabilities of individual unseen events are not important. Because the decoder only proposes phrase translations that are in the phrasetable (ie, that have non-zero count), it never requires estimates for pairs $\tilde{s}, \tilde{t}$ having

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 53–61,
Sydney, July 2006.

$c(\tilde{s}, \tilde{t}) = 0$.[1] However, probability mass is reserved for the *set* of unseen translations, implying that probability mass is subtracted from the seen translations.

- There is no obvious lower-order distribution for backoff. One of the most important techniques in ngram LM smoothing is to combine estimates made using the previous $n-1$ words with those using only the previous $n-i$ words, for $i = 2 \ldots n$. This relies on the fact that closer words are more informative, which has no direct analog in phrasetable smoothing.

- The predicted objects are word sequences (in another language). This contrasts to LM smoothing where they are single words, and are thus less amenable to decomposition for smoothing purposes.

We propose various ways of dealing with these special features of the phrasetable smoothing problem, and give evaluations of their performance within a phrase-based SMT system.

The paper is structured as follows: section 2 gives a brief description of our phrase-based SMT system; section 3 presents the smoothing techniques used; section 4 reviews previous work; section 5 gives experimental results; and section 6 concludes and discusses future work.

2 Phrase-based Statistical MT

Given a source sentence $\mathbf{s}$, our phrase-based SMT system tries to find the target sentence $\hat{\mathbf{t}}$ that is the most likely translation of $\mathbf{s}$. To make search more efficient, we use the Viterbi approximation and seek the most likely combination of $\mathbf{t}$ and its alignment $\mathbf{a}$ with $\mathbf{s}$, rather than just the most likely $\mathbf{t}$:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{s}) \approx \operatorname*{argmax}_{\mathbf{t},\mathbf{a}} p(\mathbf{t}, \mathbf{a}|\mathbf{s}),$$

where $\mathbf{a} = (\tilde{s}_1, \tilde{t}_1, j_1), ..., (\tilde{s}_K, \tilde{t}_K, j_K)$; $\tilde{t}_k$ are target phrases such that $\mathbf{t} = \tilde{t}_1 \ldots \tilde{t}_K$; $\tilde{s}_k$ are source phrases such that $\mathbf{s} = \tilde{s}_{j_1} \ldots \tilde{s}_{j_K}$; and $\tilde{s}_k$ is the translation of the kth target phrase $\tilde{t}_k$.

[1] This is a first approximation; exceptions occur when different phrasetables are used in parallel, and when rules are used to translate certain classes of entities.

To model $p(\mathbf{t}, \mathbf{a}|\mathbf{s})$, we use a standard loglinear approach:

$$p(\mathbf{t}, \mathbf{a}|\mathbf{s}) \propto \exp\left[\sum_i \lambda_i f_i(\mathbf{s}, \mathbf{t}, \mathbf{a})\right]$$

where each $f_i(\mathbf{s}, \mathbf{t}, \mathbf{a})$ is a feature function, and weights λ_i are set using Och's algorithm (Och, 2003) to maximize the system's BLEU score (Papineni et al., 2001) on a development corpus. The features used in this study are: the length of $\mathbf{t}$; a single-parameter distortion penalty on phrase reordering in $\mathbf{a}$, as described in (Koehn et al., 2003); phrase translation model probabilities; and trigram language model probabilities $\log p(\mathbf{t})$, using Kneser-Ney smoothing as implemented in the SRILM toolkit (Stolcke, 2002).

Phrase translation model probabilities are features of the form:

$$\log p(\mathbf{s}|\mathbf{t}, \mathbf{a}) \approx \sum_{k=1}^{K} \log p(\tilde{s}_k|\tilde{t}_k)$$

ie, we assume that the phrases $\tilde{s}_k$ specified by $\mathbf{a}$ are conditionally independent, and depend only on their aligned phrases $\tilde{t}_k$. The "forward" phrase probabilities $p(\tilde{t}|\tilde{s})$ are not used as features, but only as a filter on the set of possible translations: for each source phrase $\tilde{s}$ that matches some ngram in $\mathbf{s}$, only the 30 top-ranked translations $\tilde{t}$ according to $p(\tilde{t}|\tilde{s})$ are retained.

To derive the joint counts $c(\tilde{s}, \tilde{t})$ from which $p(\tilde{s}|\tilde{t})$ and $p(\tilde{t}|\tilde{s})$ are estimated, we use the phrase induction algorithm described in (Koehn et al., 2003), with symmetrized word alignments generated using IBM model 2 (Brown et al., 1993).

3 Smoothing Techniques

Smoothing involves some recipe for modifying conditional distributions away from pure relative-frequency estimates made from joint counts, in order to compensate for data sparsity. In the spirit of ((Hastie et al., 2001), figure 2.11, pg. 38) smoothing can be seen as a way of combining the relative-frequency estimate, which is a model with high complexity, high variance, and low bias, with another model with lower complexity, lower variance, and high bias, in the hope of obtaining better performance on new data. There are two main ingredients in all such recipes: some probability distribution that is smoother than relative frequencies (ie, that has fewer parameters and is thus less

complex) and some technique for combining that distribution with relative frequency estimates. We will now discuss both these choices: the distribution for carrying out smoothing and the combination technique. In this discussion, we use $\tilde{p}()$ to denote relative frequency distributions.

Choice of Smoothing Distribution

One can distinguish between two approaches to smoothing phrase tables. *Black-box* techniques do not look inside phrases but instead treat them as atomic objects: that is, both the $\tilde{s}$ and the $\tilde{t}$ in the expression $p(\tilde{s}|\tilde{t})$ are treated as units about which nothing is known except their counts. In contrast, *glass-box* methods break phrases down into their component words.

The black-box approach, which is the simpler of the two, has received little attention in the SMT literature. An interesting aspect of this approach is that it allows one to implement phrasetable smoothing techniques that are analogous to LM smoothing techniques, by treating the problem of estimating $p(\tilde{s}|\tilde{t})$ as if it were the problem of estimating a bigram conditional probability. In this paper, we give experimental results for phrasetable smoothing techniques analogous to Good-Turing, Fixed-Discount, Kneser-Ney, and Modified Kneser-Ney LM smoothing.

Glass-box methods for phrasetable smoothing have been described by other authors: see section 3.3. These authors decompose $p(\tilde{s}|\tilde{t})$ into a set of lexical distributions $p(s|\tilde{t})$ by making independence assumptions about the words s in $\tilde{s}$. The other possibility, which is similar in spirit to ngram LM lower-order estimates, is to combine estimates made by replacing words in $\tilde{t}$ with wildcards, as proposed in section 3.4.

Choice of Combination Technique

Although we explored a variety of black-box and glass-box smoothing distributions, we only tried two combination techniques: linear interpolation, which we used for black-box smoothing, and loglinear interpolation, which we used for glass-box smoothing.

For black-box smoothing, we could have used a backoff scheme or an interpolation scheme. Backoff schemes have the form:

$$p(\tilde{s}|\tilde{t}) = \begin{cases} p_h(\tilde{s}|\tilde{t}), & c(\tilde{s},\tilde{t}) \geq \tau \\ p_b(\tilde{s}|\tilde{t}), & \text{else} \end{cases}$$

where $p_h(\tilde{s}|\tilde{t})$ is a higher-order distribution, $p_b(\tilde{s}|\tilde{t})$ is a smooth backoff distribution, and τ is a threshold above which counts are considered reliable. Typically, $\tau = 1$ and $p_h(\tilde{s}|\tilde{t})$ is version of $\tilde{p}(\tilde{s}|\tilde{t})$ modified to reserve some probability mass for unseen events.

Interpolation schemes have the general form:

$$p(\tilde{s}|\tilde{t}) = \alpha(\tilde{s},\tilde{t})\tilde{p}(\tilde{s}|\tilde{t}) + \beta(\tilde{s},\tilde{t})p_b(\tilde{s}|\tilde{t}), \quad (1)$$

where α and β are combining coefficients. As noted in (Chen and Goodman, 1998), a key difference between interpolation and backoff is that the former approach uses information from the smoothing distribution to modify $\tilde{p}(\tilde{s}|\tilde{t})$ for higher-frequency events, whereas the latter uses it only for low-frequency events (most often 0-frequency events). Since for phrasetable smoothing, better prediction of unseen (zero-count) events has no direct impact—only seen events are represented in the phrasetable, and thus hypothesized during decoding—interpolation seemed a more suitable approach.

For combining relative-frequency estimates with glass-box smoothing distributions, we employed loglinear interpolation. This is the traditional approach for glass-box smoothing (Koehn et al., 2003; Zens and Ney, 2004). To illustrate the difference between linear and loglinear interpolation, consider combining two Bernoulli distributions $p_1(x)$ and $p_2(x)$ using each method:

$$p_{linear}(x) = \alpha p_1(x) + (1-\alpha)p_2(x)$$

$$p_{loglin}(x) = \frac{p_1(x)^\alpha p_2(x)}{p_1(x)^\alpha p_2(x) + q_1(x)^\alpha q_2(x)}$$

where $q_i(x) = 1 - p_i(x)$. Setting $p_2(x) = 0.5$ to simulate uniform smoothing gives $p_{loglin}(x) = p_1(x)^\alpha/(p_1(x)^\alpha + q_1(x)^\alpha)$. This is actually *less* smooth than the original distribution $p_1(x)$: it preserves extreme values 0 and 1, and makes intermediate values more extreme. On the other hand, $p_{linear}(x) = \alpha p_1(x) + (1-\alpha)/2$, which has the opposite properties: it moderates extreme values and tends to preserve intermediate values.

An advantage of loglinear interpolation is that we can tune loglinear weights so as to maximize the true objective function, for instance BLEU; recall that our translation model is itself loglinear, with weights set to minimize errors. In fact, a limitation of the experiments described in this paper is that the loglinear weights for the glass-box techniques were optimized for BLEU using Och's algorithm (Och, 2003), while the linear weights for

black-box techniques were set heuristically. Obviously, this gives the glass-box techniques an advantage when the different smoothing techniques are compared using BLEU! Implementing an algorithm for optimizing linear weights according to BLEU is high on our list of priorities.

The preceding discussion implicitly assumes a single set of counts $c(\tilde{s}, \tilde{t})$ from which conditional distributions are derived. But, as phrases of different lengths are likely to have different statistical properties, it might be worthwhile to break down the global phrasetable into separate phrasetables for each value of $|\tilde{t}|$ for the purposes of smoothing. Any similar strategy that does not split up $\{\tilde{s}|c(\tilde{s}, \tilde{t}) > 0\}$ for any fixed $\tilde{t}$ can be applied to any smoothing scheme. This is another idea we are eager to try soon.

We now describe the individual smoothing schemes we have implemented. Four of them are black-box techniques: Good-Turing and three fixed-discount techniques (fixed-discount interpolated with unigram distribution, Kneser-Ney fixed-discount, and modified Kneser-Ney fixed-discount). Two of them are glass-box techniques: Zens-Ney "noisy-or" and Koehn-Och-Marcu IBM smoothing. Our experiments tested not only these individual schemes, but also some loglinear combinations of a black-box technique with a glass-box technique.

3.1 Good-Turing

Good-Turing smoothing is a well-known technique (Church and Gale, 1991) in which observed counts c are modified according to the formula:

$$c_g = (c+1)n_{c+1}/n_c \tag{2}$$

where c_g is a modified count value used to replace c in subsequent relative-frequency estimates, and n_c is the number of events having count c. An intuitive motivation for this formula is that it approximates relative-frequency estimates made by successively leaving out each event in the corpus, and then averaging the results (Nádas, 1985).

A practical difficulty in implementing Good-Turing smoothing is that the n_c are noisy for large c. For instance, there may be only one phrase pair that occurs exactly $c = 347,623$ times in a large corpus, and no pair that occurs $c = 347,624$ times, leading to $c_g(347,623) = 0$, clearly not what is intended. Our solution to this problem is based on the technique described in (Church and Gale, 1991). We first take the log of the observed (c, n_c) values, and then use a linear least squares fit to $\log n_c$ as a function of $\log c$. To ensure that the result stays close to the reliable values of n_c for large c, error terms are weighted by c, ie: $c(\log n_c - \log n'_c)^2$, where n'_c are the fitted values.

Our implementation pools all counts $c(\tilde{s}, \tilde{t})$ together to obtain n'_c (we have not yet tried separate counts based on length of $\tilde{t}$ as discussed above). It follows directly from (2) that the total count mass assigned to unseen phrase pairs is $c_g(0)n_0 = n_1$, which we approximate by n'_1. This mass is distributed among contexts $\tilde{t}$ in proportion to $c(\tilde{t})$, giving final estimates:

$$p(\tilde{s}|\tilde{t}) = \frac{c_g(\tilde{s}, \tilde{t})}{\sum_s c_g(\tilde{s}, \tilde{t}) + p(\tilde{t})n'_1},$$

where $p(\tilde{t}) = c(\tilde{t}) / \sum_{\tilde{t}} c(\tilde{t})$.

3.2 Fixed-Discount Methods

Fixed-discount methods subtract a fixed discount D from all non-zero counts, and distribute the resulting probability mass according to a smoothing distribution (Kneser and Ney, 1995). We use an interpolated version of fixed-discount proposed by (Chen and Goodman, 1998) rather than the original backoff version. For phrase pairs with non-zero counts, this distribution has the general form:

$$p(\tilde{s}|\tilde{t}) = \frac{c(\tilde{s}, \tilde{t}) - D}{\sum_{\tilde{s}} c(\tilde{s}, \tilde{t})} + \alpha(\tilde{t}) p_b(\tilde{s}|\tilde{t}), \tag{3}$$

where $p_b(\tilde{s}|\tilde{t})$ is the smoothing distribution. Normalization constraints fix the value of $\alpha(\tilde{t})$:

$$\alpha(\tilde{t}) = D\, n_{1+}(*, \tilde{t}) / \sum_{\tilde{s}} c(\tilde{s}, \tilde{t}),$$

where $n_{1+}(*, \tilde{t})$ is the number of phrases $\tilde{s}$ for which $c(\tilde{s}, \tilde{t}) > 0$.

We experimented with two choices for the smoothing distribution $p_b(\tilde{s}|\tilde{t})$. The first is a plain unigram $p(\tilde{s})$, and the second is the Kneser-Ney lower-order distribution:

$$p_b(\tilde{s}) = n_{1+}(\tilde{s}, *) / \sum_{\tilde{s}} n_{1+}(\tilde{s}, *),$$

ie, the proportion of unique target phrases that $\tilde{s}$ is associated with, where $n_{1+}(\tilde{s}, *)$ is defined analogously to $n_{1+}(*, \tilde{t})$. Intuitively, the idea is that source phrases that co-occur with many different

target phrases are more likely to appear in new contexts.

For both unigram and Kneser-Ney smoothing distributions, we used a discounting coefficient derived by (Ney et al., 1994) on the basis of a leave-one-out analysis: $D = n_1/(n_1 + 2n_2)$. For the Kneser-Ney smoothing distribution, we also tested the "Modified Kneser-Ney" extension suggested in (Chen and Goodman, 1998), in which specific coefficients D_c are used for small count values c up to a maximum of three (ie D_3 is used for $c \geq 3$). For $c = 2$ and $c = 3$, we used formulas given in that paper.

3.3 Lexical Decomposition

The two glass-box techniques that we considered involve decomposing source phrases with independence assumptions. The simplest approach assumes that all source words are conditionally independent, so that:

$$p(\tilde{s}|\tilde{t}) = \prod_{j=1}^{\tilde{J}} p(s_j|\tilde{t})$$

We implemented two variants for $p(s_j|\tilde{t})$ that are described in previous work. (Zens and Ney, 2004) describe a "noisy-or" combination:

$$\begin{aligned} p(s_j|\tilde{t}) &= 1 - p(\bar{s}_j|\tilde{t}) \\ &\approx 1 - \prod_{i=1}^{\tilde{I}} (1 - p(s_j|t_i)) \end{aligned}$$

where $\bar{s}_j$ is the probability that s_j is *not* in the translation of $\tilde{t}$, and $p(s_j|t_i)$ is a lexical probability. (Zens and Ney, 2004) obtain $p(s_j|t_i)$ from smoothed relative-frequency estimates in a word-aligned corpus. Our implementation simply uses IBM1 probabilities, which obviate further smoothing.

The noisy-or combination stipulates that s_j should not appear in $\tilde{s}$ if it is not the translation of any of the words in $\tilde{t}$. The complement of this, proposed in (Koehn et al., 2005), to say that s_j *should* appear in $\tilde{s}$ if it is the translation of at least one of the words in $\tilde{t}$:

$$p(s_j|\tilde{t}) = \sum_{i \in A_j} p(s_j|t_i)/|A_j|$$

where A_j is a set of likely alignment connections for s_j. In our implementation of this method, we assumed that $A_j = \{1, \ldots, \tilde{I}\}$, ie the set of all connections, and used IBM1 probabilities for $p(s|t)$.

3.4 Lower-Order Combinations

We mentioned earlier that LM ngrams have a naturally-ordered sequence of smoothing distributions, obtained by successively dropping the last word in the context. For phrasetable smoothing, because no word in $\tilde{t}$ is a priori less informative than any others, there is no exact parallel to this technique. However, it is clear that estimates made by replacing particular target (conditioning) words with wildcards will be smoother than the original relative frequencies. A simple scheme for combining them is just to average:

$$p(\tilde{s}|\tilde{t}) = \sum_{i=\tilde{I}} \frac{c_i^*(\tilde{s}, \tilde{t})}{\sum_{\tilde{s}} c_i^*(\tilde{s}, \tilde{t})} / \tilde{I}$$

where:

$$c_i^*(\tilde{s}, \tilde{t}) = \sum_{t_i} c(\tilde{s}, t_1 \ldots t_i \ldots t_{\tilde{I}}).$$

One might also consider progressively replacing the least informative remaining word in the target phrase (using tf-idf or a similar measure).

The same idea could be applied in reverse, by replacing particular source (conditioned) words with wildcards. We have not yet implemented this new glass-box smoothing technique, but it has considerable appeal. The idea is similar in spirit to Collins' backoff method for prepositional phrase attachment (Collins and Brooks, 1995).

4 Related Work

As mentioned previously, (Chen and Goodman, 1998) give a comprehensive survey and evaluation of smoothing techniques for language modeling. As also mentioned previously, there is relatively little published work on smoothing for statistical MT. For the IBM models, alignment probabilities need to be smoothed for combinations of sentence lengths and positions not encountered in training data (García-Varea et al., 1998). Moore (2004) has found that smoothing to correct overestimated IBM1 lexical probabilities for rare words can improve word-alignment performance. Langlais (2005) reports negative results for synonym-based smoothing of IBM2 lexical probabilities prior to extracting phrases for phrase-based SMT.

For phrase-based SMT, the use of smoothing to avoid zero probabilities during phrase induction is reported in (Marcu and Wong, 2002), but no details are given. As described above, (Zens and

Ney, 2004) and (Koehn et al., 2005) use two different variants of glass-box smoothing (which they call "lexical smoothing") over the phrasetable, and combine the resulting estimates with pure relative-frequency ones in a loglinear model. Finally, (Cettollo et al., 2005) describes the use of Witten-Bell smoothing (a black-box technique) for phrasetable counts, but does not give a comparison to other methods. As Witten-Bell is reported by (Chen and Goodman, 1998) to be significantly worse than Kneser-Ney smoothing, we have not yet tested this method.

5 Experiments

We carried out experiments in two different settings: broad-coverage ones across six European language pairs using selected smoothing techniques and relatively small training corpora; and Chinese to English experiments using all implemented smoothing techniques and large training corpora. For the black-box techniques, the smoothed phrase table replaced the original relative-frequency (RF) phrase table. For the glass-box techniques, a phrase table (either the original RF phrase table or its replacement after black-box smoothing) was interpolated in loglinear fashion with the smoothing glass-box distribution, with weights set to maximize BLEU on a development corpus.

To estimate the significance of the results across different methods, we used 1000-fold pairwise bootstrap resampling at the 95% confidence level.

5.1 Broad-Coverage Experiments

In order to measure the benefit of phrasetable smoothing for relatively small corpora, we used the data made available for the WMT06 shared task (WMT, 2006). This exercise is conducted openly with access to all needed resources and is thus ideal for benchmarking statistical phrase-based translation systems on a number of language pairs.

The WMT06 corpus is based on sentences extracted from the proceedings of the European Parliament. Separate sentence-aligned parallel corpora of about 700,000 sentences (about 150MB) are provided for the three language pairs having one of French, Spanish and German with English. SRILM language models based on the same source are also provided for each of the four languages. We used the provided 2000-sentence dev-sets for tuning loglinear parameters, and tested on the 3064-sentence test sets.

Results are shown in table 1 for relative-frequency (RF), Good-Turing (GT), Kneser-Ney with 1 (KN1) and 3 (KN3) discount coefficients; and loglinear combinations of both RF and KN3 phrasetables with Zens-Ney-IBM1 (ZN-IBM1) smoothed phrasetables (these combinations are denoted RF+ZN-IBM1 and KN3+ZN-IBM1).

It is apparent from table 1 that any kind of phrase table smoothing is better than using none; the minimum improvement is 0.45 BLEU, and the difference between RF and all other methods is statistically significant. Also, Kneser-Ney smoothing gives a statistically significant improvement over GT smoothing, with a minimum gain of 0.30 BLEU. Using more discounting coefficients does not appear to help. Smoothing relative frequencies with an additional Zens-Ney phrasetable gives about the same gain as Kneser-Ney smoothing on its own. However, combining Kneser-Ney with Zens-Ney gives a clear gain over any other method (statistically significant for all language pairs except en→es and en→de) demonstrating that these approaches are complementary.

5.2 Chinese-English Experiments

To test the effects of smoothing with larger corpora, we ran a set of experiments for Chinese-English translation using the corpora distributed for the NIST MT05 evaluation (www.nist.gov/speech/tests/mt). These are summarized in table 2. Due to the large size of the out-of-domain UN corpus, we trained one phrasetable on it, and another on all other parallel corpora (smoothing was applied to both). We also used a subset of the English Gigaword corpus to augment the LM training material.

corpus	use	sentences
non-UN	phrasetable1 + LM	3,164,180
UN	phrasetable2 + LM	4,979,345
Gigaword	LM	11,681,852
multi-p3	dev	993
eval-04	test	1788

Table 2: Chinese-English Corpora

Table 3 contains results for the Chinese-English experiments, including fixed-discount with unigram smoothing (FDU), and Koehn-Och-Marcu smoothing with the IBM1 model (KOM-IBM1)

smoothing method	fr ⟶ en	es ⟶ en	de ⟶ en	en ⟶ fr	en ⟶ es	en ⟶ de
RF	25.35	27.25	20.46	27.20	27.18	14.60
GT	25.95	28.07	21.06	27.85	27.96	15.05
KN1	26.83	28.66	21.36	28.62	28.71	15.42
KN3	26.84	28.69	21.53	28.64	28.70	15.40
RF+ZN-IBM1	26.84	28.63	21.32	28.84	28.45	15.44
KN3+ZN-IBM1	**27.25**	**29.30**	**21.77**	**29.00**	**28.86**	**15.49**

Table 1: Broad-coverage results

as described in section 3.3. As with the broad-coverage experiments, all of the black-box smoothing techniques do significantly better than the RF baseline. However, GT appears to work better in the large-corpus setting: it is statistically indistinguishable from KN3, and both these methods are significantly better than all other fixed-discount variants, among which there is little difference.

Not surprisingly, the two glass-box methods, ZN-IBM1 and KOM-IBM1, do poorly when used on their own. However, in combination with another phrasetable, they yield the best results, obtained by RF+ZN-IBM1 and GT+KOM-IBM1, which are statistically indistinguishable. In constrast to the situation in the broad-coverage setting, these are not significantly better than the best black-box method (GT) on its own, although RF+ZN-IBM1 *is* better than all other glass-box combinations.

smoothing method	BLEU score
RF	29.85
GT	30.66
FDU	30.23
KN1	30.29
KN2	30.13
KN3	30.54
ZN-IBM1	29.55
KOM-IBM1	28.09
RF+ZN-IBM1	**30.95**
RF+KOM-IBM1	30.10
GT+ZN-IBM1	30.45
GT+KOM-IBM1	30.81
KN3+ZN-IBM1	30.66

Table 3: Chinese-English Results

A striking difference between the broad-coverage setting and the Chinese-English setting is that in the former it appears to be beneficial to apply KN3 smoothing to the phrasetable that gets combined with the best glass-box phrasetable (ZN), whereas in the latter setting it does not. To test whether this was due to corpus size (as the broad-coverage corpora are around 10% of those for Chinese-English), we calculated Chinese-English learning curves for the RF+ZN-IBM1 and KN3-ZN-IBM1 methods, shown in figure 1. The results are somewhat inconclusive: although the KN3+ZN-IBM1 curve is perhaps slightly flatter, the most obvious characteristic is that this method appears to be highly sensitive to the particular corpus sample used.

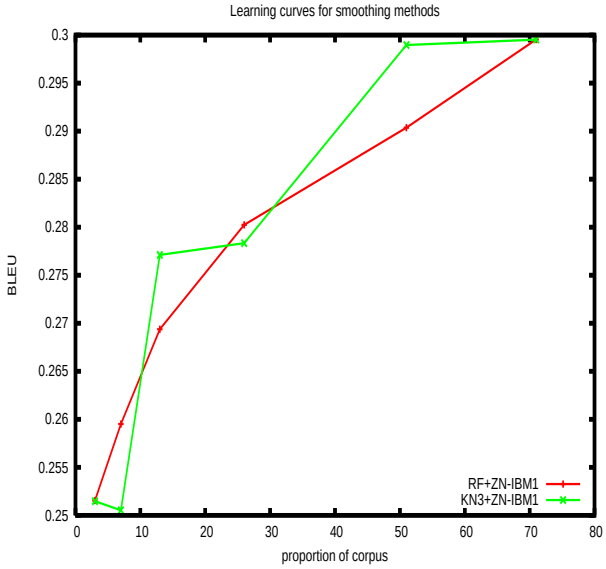

Figure 1: Learning curves for two glass-box combinations.

6 Conclusion and Future Work

We tested different phrasetable smoothing techniques in two different translation settings: European language pairs with relatively small corpora, and Chinese to English translation with large corpora. The smoothing techniques fall into two

categories: black-box methods that work only on phrase-pair counts; and glass-box methods that decompose phrase probabilities into lexical probabilities. In our implementation, black-box techniques use linear interpolation to combine relative frequency estimates with smoothing distributions, while glass-box techniques are combined in loglinear fashion with either relative-frequencies or black-box estimates.

All smoothing techniques tested gave statistically significant gains over pure relative-frequency estimates. In the small-corpus setting, the best technique is a loglinear combination of Kneser-Ney count smoothing with Zens-Ney glass-box smoothing; this yields an average gain of 1.6 BLEU points over relative frequencies. In the large-corpus setting, the best technique is a loglinear combination of relative-frequency estimates with Zens-Ney smoothing, with a gain of 1.1 BLEU points. Of the two glass-box smoothing methods tested, Zens-Ney appears to have a slight advantage over Koehn-Och-Marcu. Of the black-box methods tested, Kneser-Ney is clearly better for small corpora, but is equivalent to Good-Turing for larger corpora.

The paper describes several smoothing alternatives which we intend to test in future work:

- Linear versus loglinear combinations (in our current work, these coincide with the black-box versus glass-box distinction, making it impossible to draw conclusions).
- Lower-order distributions as described in section 3.4.
- Separate count-smoothing bins based on phrase length.

7 Acknowledgements

The authors would like to thank their colleague Michel Simard for stimulating discussions. The first author would like to thank all his colleagues for encouraging him to taste a delicacy that was new to him (shredded paper with maple syrup). This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

Peter F. Brown, Stephen A. Della Pietra, Vincent Della J. Pietra, and Robert L. Mercer. 1993. The mathematics of Machine Translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312, June.

M. Cettollo, M. Federico, N. Bertoldi, R. Cattoni, and B. Chen. 2005. A look inside the ITC-irst SMT system. In *Proceedings of MT Summit X*, Phuket, Thailand, September. International Association for Machine Translation.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

K. Church and W. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer speech and language*, 5(1):19–54.

M. Collins and J. Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora (WVLC)*, Cambridge, Massachusetts.

Ismael García-Varea, Francisco Casacuberta, and Hermann Ney. 1998. An iterative, DP-based search algorithm for statistical machine translation. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP) 1998*, volume 4, pages 1135–1138, Sydney, Australia, December.

Joshua Goodman. 2001. A bit of progress in language modeling. *Computer Speech and Language*.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1995*, pages 181–184, Detroit, Michigan. IEEE.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Eduard Hovy, editor, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133, Edmonton, Alberta, Canada, May. NAACL.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, D. Talbot, and M. White. 2005. Edinburgh system description for the 2005 NIST MT evaluation. In *Proceedings of Machine Translation Evaluation Workshop*.

Philippe Langlais, Guihong Cao, and Fabrizio Gotti. 2005. RALI: SMT shared task system description.

In *Proceedings of the 2nd ACL workshop on Building and Using Parallel Texts*, pages 137–140, University of Michigan, Ann Arbor, June.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA.

Robert C. Moore. 2004. Improving IBM word-alignment model 1. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, July.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 10:1–38.

Arthur Nádas. 1985. On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, ASSP-33(6):1415–1417, December.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of Machine Translation. Technical Report RC22176, IBM, September.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado, September.

WMT. 2006. *The NAACL Workshop on Statistical Machine Translation (www.statmt.org/wmt06)*, New York, June.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of Human Language Technology Conference / North American Chapter of the ACL*, Boston, May.