Why Generative Phrase Models Underperform Surface Heuristics

John DeNero, Dan Gillick, James Zhang, Dan Klein
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94705
{denero, dgillick, jyzhang, klein}@eecs.berkeley.edu

Abstract

We investigate why weights from generative models underperform heuristic estimates in phrase-based machine translation. We first propose a simple generative, phrase-based model and verify that its estimates are inferior to those given by surface statistics. The performance gap stems primarily from the addition of a hidden segmentation variable, which increases the capacity for overfitting during maximum likelihood training with EM. In particular, while word level models benefit greatly from re-estimation, phrase-level models do not: the crucial difference is that distinct word alignments cannot all be correct, while distinct segmentations can. Alternate segmentations rather than alternate alignments compete, resulting in increased determinization of the phrase table, decreased generalization, and decreased final BLEU score. We also show that interpolation of the two methods can result in a modest increase in BLEU score.

1 Introduction

At the core of a phrase-based statistical machine translation system is a *phrase table* containing pairs of source and target language phrases, each weighted by a conditional translation probability. Koehn et al. (2003a) showed that translation quality is very sensitive to how this table is extracted from the training data. One particularly surprising result is that a simple heuristic extraction algorithm based on surface statistics of a word-aligned training set outperformed the phrase-based generative model proposed by Marcu and Wong (2002).

This result is surprising in light of the reverse situation for word-based statistical translation. Specifically, in the task of word alignment, heuristic approaches such as the Dice coefficient consistently underperform their re-estimated counterparts, such as the IBM word alignment models (Brown et al., 1993). This well-known result is unsurprising: re-estimation introduces an element of *competition* into the learning process. The key virtue of competition in word alignment is that, to a first approximation, only one source word should generate each target word. If a good alignment for a word token is found, other plausible alignments are explained away and should be discounted as incorrect for that token.

As we show in this paper, this effect does not prevail for phrase-level alignments. The central difference is that phrase-based models, such as the ones presented in section 2 or Marcu and Wong (2002), contain an element of *segmentation*. That is, they do not merely learn correspondences between phrases, but also segmentations of the source and target sentences. However, while it is reasonable to suppose that if one alignment is right, others must be wrong, the situation is more complex for segmentations. For example, if one segmentation subsumes another, they are not necessarily incompatible: both may be equally valid. While in some cases, such as idiomatic vs. literal translations, two segmentations may be in true competition, we show that the most common result is for different segmentations to be recruited for different examples, overfitting the training data and overly determinizing the phrase translation estimates.

In this work, we first define a novel (but not radical) generative phrase-based model analogous to IBM Model 3. While its exact training is intractable, we describe a training regime which uses word-level alignments to constrain the space of feasible segmentations down to a manageable number. We demonstrate that the phrase analogue of the Dice coefficient is superior to our generative model (a result also echoing previous work). In the primary contribution of the paper, we present a series of experiments designed to elucidate what re-estimation learns in this context. We show that estimates are overly determinized because segmentations are used

Proceedings of the Workshop on Statistical Machine Translation, pages 31–38,
New York City, June 2006.

in unintuitive ways for the sake of data likelihood. We comment on both the beneficial instances of segment competition (idioms) as well as the harmful ones (most everything else). Finally, we demonstrate that interpolation of the two estimates can provide a modest increase in BLEU score over the heuristic baseline.

2 Approach and Evaluation Methodology

The generative model defined below is evaluated based on the BLEU score it produces in an end-to-end machine translation system from English to French. The top-performing *diag-and* extraction heuristic (Zens et al., 2002) serves as the baseline for evaluation.[1] Each approach – the generative model and heuristic baseline – produces an estimated conditional distribution of English phrases given French phrases. We will refer to the distribution derived from the baseline heuristic as ϕ_H. The distribution learned via the generative model, denoted ϕ_{EM}, is described in detail below.

2.1 A Generative Phrase Model

While our model for computing ϕ_{EM} is novel, it is meant to exemplify a class of models that are not only clear extensions to generative word alignment models, but also compatible with the statistical framework assumed during phrase-based decoding.

The generative process we modeled produces a phrase-aligned English sentence from a French sentence where the former is a translation of the latter. Note that this generative process is opposite to the translation direction of the larger system because of the standard noisy-channel decomposition. The learned parameters from this model will be used to translate sentences from English to French. The generative process modeled has four steps:[2]

1. Begin with a French sentence **f**.
2. Segment **f** into a sequence of I multi-word phrases that span the sentence, $\bar{f}_1^I$.
3. For each phrase $\bar{f}_i \in \bar{f}_1^I$, choose a corresponding position j in the English sentence and establish the alignment $a_j = i$, then generate exactly one English phrase $\bar{e}_j$ from $\bar{f}_i$.
4. The sequence $\bar{e}_j$ ordered by a describes an English sentence **e**.

The corresponding probabilistic model for this generative process is:

$$\begin{aligned} P(\mathbf{e}|\mathbf{f}) &= \sum_{\bar{f}_1^I, \bar{e}_1^I, a} P(\mathbf{e}, \bar{f}_1^I, \bar{e}_1^I, a|\mathbf{f}) \\ &= \sum_{\bar{f}_1^I, \bar{e}_1^I, a} \sigma(\bar{f}_1^I|\mathbf{f}) \prod_{\bar{f}_i \in \bar{f}_1^I} \phi(\bar{e}_j|\bar{f}_i) d(a_j = i|\mathbf{f}) \end{aligned}$$

where $P(\mathbf{e}, \bar{f}_1^I, \bar{e}_1^I, a|\mathbf{f})$ factors into a segmentation model σ, a translation model ϕ and a distortion model d. The parameters for each component of this model are estimated differently:

- The segmentation model $\sigma(\bar{f}_1^I|\mathbf{f})$ is assumed to be uniform over all possible segmentations for a sentence.[3]
- The phrase translation model $\phi(\bar{e}_j|\bar{f}_i)$ is parameterized by a large table of phrase translation probabilities.
- The distortion model $d(a_j = i|\mathbf{f})$ is a discounting function based on absolute sentence position akin to the one used in IBM model 3.

While similar to the joint model in Marcu and Wong (2002), our model takes a conditional form compatible with the statistical assumptions used by the Pharaoh decoder. Thus, after training, the parameters of the phrase translation model ϕ_{EM} can be used directly for decoding.

2.2 Training

Significant approximation and pruning is required to train a generative phrase model and table – such as ϕ_{EM} – with hidden segmentation and alignment variables using the expectation maximization algorithm (EM). Computing the likelihood of the data

[1] This well-known heuristic extracts phrases from a sentence pair by computing a word-level alignment for the sentence and then enumerating all phrases compatible with that alignment. The word alignment is computed by first intersecting the directional alignments produced by a generative IBM model (e.g., model 4 with minor enhancements) in each translation direction, then adding certain alignments from the union of the directional alignments based on local growth rules.

[2] Our notation matches the literature for phrase-based translation: e is an English word, $\bar{e}$ is an English phrase, and $\bar{e}_1^I$ is a sequence of I English phrases, and **e** is an English sentence.

[3] This segmentation model is deficient given a maximum phrase length: many segmentations are disallowed in practice.

for a set of parameters (the e-step) involves summing over exponentially many possible segmentations for each training sentence. Unlike previous attempts to train a similar model (Marcu and Wong, 2002), we allow information from a word-alignment model to inform our approximation. This approach allowed us to directly estimate translation probabilities even for rare phrase pairs, which were estimated heuristically in previous work.

In each iteration of EM, we re-estimate each phrase translation probability by summing fractional phrase counts (soft counts) from the data given the current model parameters.

$$\phi_{new}(\bar{e}_j|\bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i)} =$$

$$\sum_{(\mathbf{f},\mathbf{e})} \frac{\sum_{\bar{f}_1^I:\bar{f}_i\in\bar{f}_1^I} \sum_{\bar{e}_1^I:\bar{e}_j\in\bar{e}_1^I} \sum_{a:a_j=i} P(\mathbf{e}, \bar{f}_1^I, \bar{e}_1^I, a|\mathbf{f})}{\sum_{\bar{f}_1^I:\bar{f}_i\in\bar{f}_1^I} \sum_{\bar{e}_1^I} \sum_a P(\mathbf{e}, \bar{f}_1^I, \bar{e}_1^I, a|\mathbf{f})}$$

This training loop necessitates approximation because summing over all possible segmentations and alignments for each sentence is intractable, requiring time exponential in the length of the sentences. Additionally, the set of possible phrase pairs grows too large to fit in memory. Using word alignments, we can address both problems.[4] In particular, we can determine for any aligned segmentation $(\bar{f}_1^I, \bar{e}_1^I, a)$ whether it is compatible with the word-level alignment for the sentence pair. We define a phrase pair to be compatible with a word-alignment if no word in either phrase is aligned with a word outside the other phrase (Zens et al., 2002). Then, $(\bar{f}_1^I, \bar{e}_1^I, a)$ is compatible with the word-alignment if each of its aligned phrases is a compatible phrase pair.

The training process is then constrained such that, when evaluating the above sum, only compatible aligned segmentations are considered. That is, we allow $P(\mathbf{e}, \bar{f}_1^I, \bar{e}_1^I, a|\mathbf{f}) > 0$ only for aligned segmentations $(\bar{f}_1^I, \bar{e}_1^I, a)$ such that a provides a one-to-one mapping from $\bar{f}_1^I$ to $\bar{e}_1^I$ where all phrase pairs $(\bar{f}_{a_j}, \bar{e}_j)$ are compatible with the word alignment.

This constraint has two important effects. First, we force $P(\bar{e}_j|\bar{f}_i) = 0$ for all phrase pairs not compatible with the word-level alignment for some sentence pair. This restriction successfully reduced the total legal phrase pair types from approximately 250 million to 17 million for 100,000 training sentences. However, some desirable phrases were eliminated because of errors in the word alignments.

Second, the time to compute the e-step is reduced. While in principle it is still intractable, in practice we can compute most sentence pairs' contributions in under a second each. However, some spurious word alignments can disallow all segmentations for a sentence pair, rendering it unusable for training. Several factors including errors in the word-level alignments, sparse word alignments and non-literal translations cause our constraint to rule out approximately 54% of the training set. Thus, the reduced size of the usable training set accounts for some of the degraded performance of ϕ_{EM} relative to ϕ_H. However, the results in figure 1 of the following section show that ϕ_{EM} trained on twice as much data as ϕ_H still underperforms the heuristic, indicating a larger issue than decreased training set size.

2.3 Experimental Design

To test the relative performance of ϕ_{EM} and ϕ_H, we evaluated each using an end-to-end translation system from English to French. We chose this non-standard translation direction so that the examples in this paper would be more accessible to a primarily English-speaking audience. All training and test data were drawn from the French/English section of the Europarl sentence-aligned corpus. We tested on the first 1,000 unique sentences of length 5 to 15 in the corpus and trained on sentences of length 1 to 60 starting after the first 10,000.

The system follows the structure proposed in the documentation for the Pharaoh decoder and uses many publicly available components (Koehn, 2003b). The language model was generated from the Europarl corpus using the SRI Language Modeling Toolkit (Stolcke, 2002). Pharaoh performed decoding using a set of default parameters for weighting the relative influence of the language, translation and distortion models (Koehn, 2003b). A maximum phrase length of three was used for all experiments.

To properly compare ϕ_{EM} to ϕ_H, all aspects of the translation pipeline were held constant except for the parameters of the phrase translation table. In particular, we did not tune the decoding hyperparameters for the different phrase tables.

[4] The word alignments used in approximating the e-step were the same as those used to create the heuristic *diag-and* baseline.

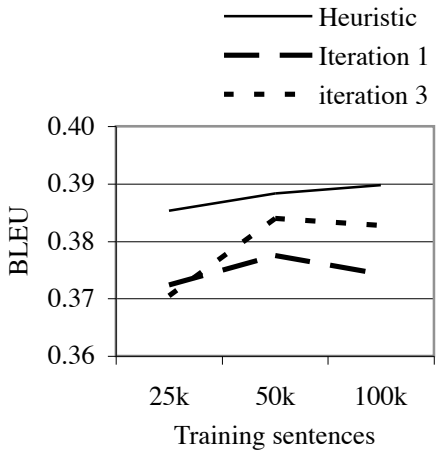


Figure 1: Statistical re-estimation using a generative phrase model degrades BLEU score relative to its heuristic initialization.

3 Results

Having generated ϕ_H heuristically and ϕ_{EM} with EM, we now compare their performance. While the model and training regimen for ϕ_{EM} differ from the model from Marcu and Wong (2002), we achieved results similar to Koehn et al. (2003a): ϕ_{EM} slightly underperformed ϕ_H. Figure 1 compares the BLEU scores using each estimate. Note that the expectation maximization algorithm for training ϕ_{EM} was initialized with the heuristic parameters ϕ_H, so the heuristic curve can be equivalently labeled as iteration 0.

Thus, the first iteration of EM increases the observed likelihood of the training sentences while simultaneously degrading translation performance on the test set. As training proceeds, performance on the test set levels off after three iterations of EM. The system never achieves the performance of its initialization parameters. The pruning of our training regimen accounts for part of this degradation, but not all; augmenting ϕ_{EM} by adding back in all phrase pairs that were dropped during training does not close the performance gap between ϕ_{EM} and ϕ_H.

3.1 Analysis

Learning ϕ_{EM} degrades translation quality in large part because EM learns overly determinized segmentations and translation parameters, overfitting the training data and failing to generalize. The primary increase in richness from generative word-level models to generative phrase-level models is due to the additional latent segmentation variable. Although we impose a uniform distribution over segmentations, it nonetheless plays a crucial role during training. We will characterize this phenomenon through aggregate statistics and translation examples shortly, but begin by demonstrating the model's capacity to overfit the training data.

Let us first return to the motivation behind introducing and learning phrases in machine translation. For any language pair, there are contiguous strings of words whose collocational translation is non-compositional; that is, they translate together differently than they would in isolation. For instance, *chat* in French generally translates to *cat* in English, but *appeler un chat un chat* is an idiom which translates to *call a spade a spade*. Introducing phrases allows us to translate *chat un chat* atomically to *spade a spade* and vice versa.

While introducing phrases and parameterizing their translation probabilities with a surface heuristic allows for this possibility, statistical re-estimation would be required to learn that *chat* should never be translated to *spade* in isolation. Hence, translating *I have a spade* with ϕ_H could yield an error.

But enforcing competition among segmentations introduces a new problem: true translation ambiguity can also be spuriously explained by the segmentation. Consider the french fragment *carte sur la table*, which could translate to *map on the table* or *notice on the chart*. Using these two sentence pairs as training, one would hope to capture the ambiguity in the parameter table as:

French	English	$\phi(e\|f)$
carte	*map*	0.5
carte	*notice*	0.5
carte sur	*map on*	0.5
carte sur	*notice on*	0.5
sur	*on*	1.0
...	...	...
table	*table*	0.5
table	*chart*	0.5

Assuming we only allow non-degenerate segmentations and disallow non-monotonic alignments, this parameter table yields a marginal likelihood $P(\mathbf{f}|\mathbf{e}) = 0.25$ for both sentence pairs – the intuitive result given two independent lexical ambigu-

ities. However, the following table yields a likelihood of 0.28 for both sentences:[5]

French	English	$\phi(e\|f)$
carte	*map*	1.0
carte sur	*notice on*	1.0
carte sur la	*notice on the*	1.0
sur	*on*	1.0
sur la table	*on the table*	1.0
la	*the*	1.0
la table	*the table*	1.0
table	*chart*	1.0

Hence, a higher likelihood can be achieved by allocating some phrases to certain translations while reserving overlapping phrases for others, thereby failing to model the real ambiguity that exists across the language pair. Also, notice that the phrase *sur la* can take on an arbitrary distribution over any english phrases without affecting the likelihood of either sentence pair. Not only does this counterintuitive parameterization give a high data likelihood, but it is also a fixed point of the EM algorithm.

The phenomenon demonstrated above poses a problem for generative phrase models in general. The ambiguous process of translation can be modeled either by the latent segmentation variable or the phrase translation probabilities. In some cases, optimizing the likelihood of the training corpus adjusts for the former when we would prefer the latter. We next investigate how this problem manifests in ϕ_{EM} and its effect on translation quality.

3.2 Learned parameters

The parameters of ϕ_{EM} differ from the heuristically extracted parameters ϕ_H in that the conditional distributions over English translations for some French words are sharply peaked for ϕ_{EM} compared to flatter distributions generated by ϕ_H. This determinism – predicted by the previous section's example – is not atypical of EM training for other tasks.

To quantify the notion of peaked distributions over phrase translations, we compute the entropy of the distribution for each French phrase according to the standard definition.

$$H(\phi(\bar{e}|\bar{f})) = \sum_{\bar{e}} \phi(\bar{e}|\bar{f}) \log_2 \phi(\bar{e}|\bar{f})$$

The average entropy, weighted by frequency, for the most common 10,000 phrases in the learned table was 1.55, comparable to 3.76 for the heuristic table. The difference between the tables becomes much more striking when we consider the histogram of entropies for phrases in figure 2. In particular, the learned table has many more phrases with entropy near zero. The most pronounced entropy differences often appear for common phrases. Ten of the most common phrases in the French corpus are shown in figure 3.

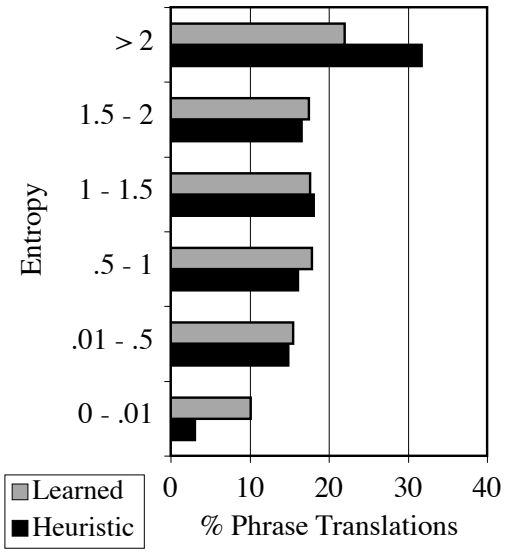


Figure 2: Many more French phrases have very low entropy under the learned parameterization.

As more probability mass is reserved for fewer translations, many of the alternative translations under ϕ_H are assigned prohibitively small probabilities. In translating 1,000 test sentences, for example, no phrase translation with $\phi(\bar{e}|\bar{f})$ less than 10^{-5} was used by the decoder. Given this empirical threshold, nearly 60% of entries in ϕ_{EM} are unusable, compared with 1% in ϕ_H.

3.3 Effects on Translation

While this determinism of ϕ_{EM} may be desirable in some circumstances, we found that the ambiguity in ϕ_H is often preferable at decoding time.

[5]For example, summing over the first translation expands to $\frac{1}{7}(\phi(map \mid carte)\phi(on\ the\ table \mid sur\ la\ table) + \phi(map \mid carte)\phi(on \mid sur)\phi(the\ table \mid la\ table))$.

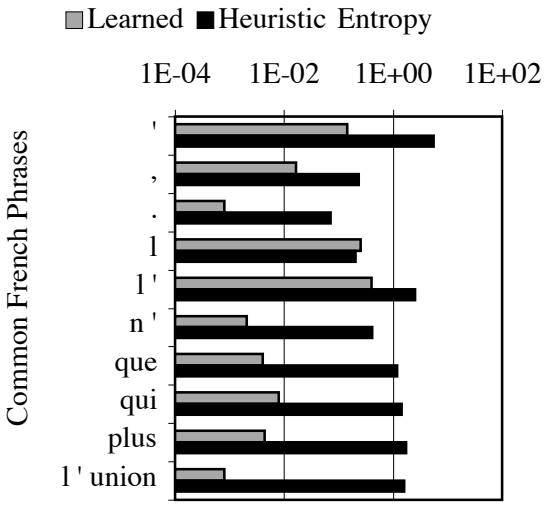


Figure 3: Entropy of 10 common French phrases. Several learned distributions have very low entropy.

In particular, the pattern of translation-ambiguous phrases receiving spuriously peaked distributions (as described in section 3.1) introduces new translation errors relative to the baseline. We now investigate both positive and negative effects of the learning process.

The issue that motivated training a generative model is sometimes resolved correctly: for a word that translates differently alone than in the context of an idiom, the translation probabilities can more accurately reflect this. Returning to the previous example, the phrase table for *chat* has been corrected through the learning process. The heuristic process gives the incorrect translation *spade* with 61% probability, while the statistical learning approach gives *cat* with 95% probability.

While such examples of improvement are encouraging, the trend of spurious determinism overwhelms this benefit by introducing errors in four related ways, each of which will be explored in turn.

1. Useful phrase pairs can be assigned very low probabilities and therefore become unusable.

2. A proper translation for a phrase can be overridden by another translation with spuriously high probability.

3. Error-prone, common, ambiguous phrases become active during decoding.

4. The language model cannot distinguish between different translation options as effectively due to deterministic translation model distributions.

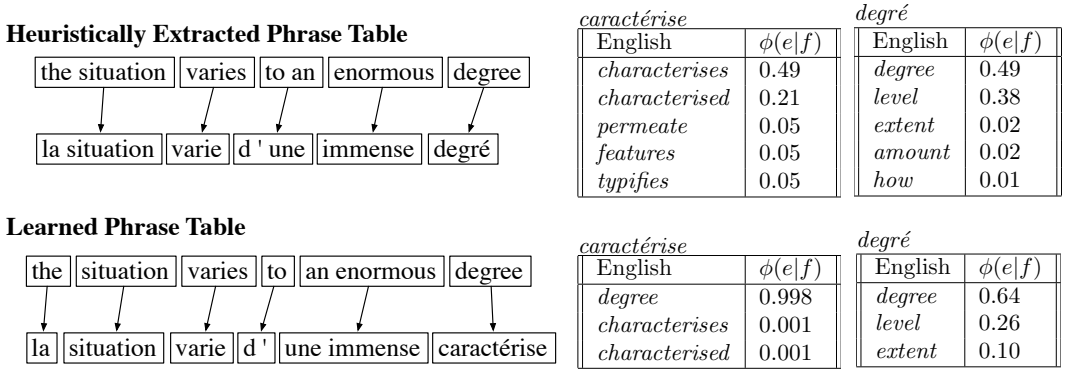
The first effect follows from our observation in section 3.2 that many phrase pairs are unusable due to vanishingly small probabilities. Some of the entries that are made unusable by re-estimation are helpful at decoding time, evidenced by the fact that pruning the set of ϕ_{EM}'s low-scoring learned phrases from the original heuristic table reduces BLEU score by 0.02 for 25k training sentences (below the score for ϕ_{EM}).

The second effect is more subtle. Consider the sentence in figure 4, which to a first approximation can be translated as a series of cognates, as demonstrated by the decoding that follows from the heuristic parameterization ϕ_H.[6] Notice also that the translation probabilities from heuristic extraction are non-deterministic. On the other hand, the translation system makes a significant lexical error on this simple sentence when parameterized by ϕ_{EM}: the use of *caractérise* in this context is incorrect. This error arises from a sharply peaked distribution over English phrases for *caractérise*.

This example illustrates a recurring problem: errors do not necessarily arise because a correct translation is not available. Notice that a preferable translation of degree as *degré* is available under both parameterizations. *Degré* is not used, however, because of the peaked distribution of a competing translation candidate. In this way, very high probability translations can effectively block the use of more appropriate translations at decoding time.

What is furthermore surprising and noteworthy in this example is that the learned, near-deterministic translation for *caractérise* is not a common translation for the word. Not only does the statistical learning process yield low-entropy translation distributions, but occasionally the translation with undesirably high conditional probability does not have a strong surface correlation with the source phrase. This example is not unique; during different initializations of the EM algorithm, we noticed such pat-

[6] While there is some agreement error and awkwardness, the heuristic translation is comprehensible to native speakers. The learned translation incorrectly translates *degree*, degrading the translation quality.



caractérise

English	$\phi(e\|f)$
characterises	0.49
characterised	0.21
permeate	0.05
features	0.05
typifies	0.05

degré

English	$\phi(e\|f)$
degree	0.49
level	0.38
extent	0.02
amount	0.02
how	0.01

caractérise

English	$\phi(e\|f)$
degree	0.998
characterises	0.001
characterised	0.001

degré

English	$\phi(e\|f)$
degree	0.64
level	0.26
extent	0.10

Figure 4: Spurious determinism in the learned phrase parameters degrades translation quality.

terns even for common French phrases such as *de* and *ne*.

The third source of errors is closely related: common phrases that translate in many ways depending on the context can introduce errors if they have a spuriously peaked distribution. For instance, consider the lone apostrophe, which is treated as a single token in our data set (figure 5). The shape of the heuristic translation distribution for the phrase is intuitively appealing, showing a relatively flat distribution among many possible translations. Such a distribution has very high entropy. On the other hand, the learned table translates the apostrophe to *the* with probability very near 1.

Heuristic

English	$\phi_H(e\|f)$
our	0.10
that	0.09
is	0.06
we	0.05
next	0.05

Learned

English	$\phi_{EM}(e\|f)$
the	0.99
,	$4.1 \cdot 10^{-3}$
is	$6.5 \cdot 10^{-4}$
to	$6.3 \cdot 10^{-4}$
in	$5.3 \cdot 10^{-4}$

Figure 5: Translation probabilities for an apostrophe, the most common french phrase. The learned table contains a highly peaked distribution.

Such common phrases whose translation depends highly on the context are ripe for producing translation errors. The flatness of the distribution of ϕ_H ensures that the single apostrophe will rarely be used during decoding because no one phrase table entry has high enough probability to promote its use. On the other hand, using the peaked entry $\phi_{EM}(the|')$ incurs virtually no cost to the score of a translation.

The final kind of errors stems from interactions between the language and translation models. The selection among translation choices via a language model – a key virtue of the noisy channel framework – is hindered by the determinism of the translation model. This effect appears to be less significant than the previous three. We should note, however, that adjusting the language and translation model weights during decoding does not close the performance gap between ϕ_H and ϕ_{EM}.

3.4 Improvements

In light of the low entropy of ϕ_{EM}, we could hope to improve translations by retaining entropy. There are several strategies we have considered to achieve this. Broadly, we have tried two approaches: combining ϕ_{EM} and ϕ_H via heuristic interpolation methods and modifying the training loop to limit determinism.

The simplest strategy to increase entropy is to interpolate the heuristic and learned phrase tables. Varying the weight of interpolation showed an improvement over the heuristic of up to 0.01 for 100k sentences. A more modest improvement of 0.003 for 25k training sentences appears in table 1.

In another experiment, we interpolated the output of each iteration of EM with its input, thereby maintaining some entropy from the initialization parameters. BLEU score increased to a maximum of 0.394 using this technique with 100k training sentences, outperforming the heuristic by a slim margin of 0.005.

We might address the determinization in ϕ_{EM} without resorting to interpolation by modifying the

training procedure to retain entropy. By imposing a non-uniform segmentation model that favors shorter phrases over longer ones, we hope to prevent the error-causing effects of EM training outlined above. In principle, this change will encourage EM to explain training sentences with shorter sentences. In practice, however, this approach has not led to an improvement in BLEU.

Another approach to maintaining entropy during the training process is to smooth the probabilities generated by EM. In particular, we can use the following smoothed update equation during the training loop, which reserves a portion of probability mass for unseen translations.

$$\phi_{new}(\bar{e}_j|\bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i) + k^{l-1}}$$

In the equation above, l is the length of the French phrase and k is a tuning parameter. This formulation not only serves to reduce very spiked probabilities in ϕ_{EM}, but also boosts the probability of short phrases to encourage their use. With $k = 2.5$, this smoothing approach improves BLEU by .007 using 25k training sentences, nearly equaling the heuristic (table 1).

4 Conclusion

Re-estimating phrase translation probabilities using a generative model holds the promise of improving upon heuristic techniques. However, the combinatorial properties of a phrase-based generative model have unfortunate side effects. In cases of true ambiguity in the language pair to be translated, parameter estimates that explain the ambiguity using segmentation variables can in some cases yield higher data likelihoods by determinizing phrase translation estimates. However, this behavior in turn leads to errors at decoding time.

We have also shown that some modest benefit can be obtained from re-estimation through the blunt instrument of interpolation. A remaining challenge is to design more appropriate statistical models which tie segmentations together unless sufficient evidence of true non-compositionality is present; perhaps such models could properly combine the benefits of both current approaches.

Estimate	BLEU
ϕ_H	0.385
ϕ_H phrase pairs that also appear in ϕ_{EM}	0.365
ϕ_{EM}	0.374
ϕ_{EM} with a non-uniform segmentation model	0.374
ϕ_{EM} with smoothing	0.381
ϕ_{EM} with gaps filled in by ϕ_H	0.374
ϕ_{EM} interpolated with ϕ_H	0.388

Table 1: BLEU results for 25k training sentences.

5 Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback on this paper.

References

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*, 19(2), 1993.

Philipp Koehn. *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*. USC Information Sciences Institute, 2002.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. *HLT-NAACL*, 2003.

Philipp Koehn. *Pharaoh: A Beam Search Decoder for Phrase-Based Statisical Machine Translation Models*. USC Information Sciences Institute, 2003.

Daniel Marcu and William Wong. A phrase-based, joint probability model for statistical machine translation. *Conference on Empirical Methods in Natual Language Processing*, 2002.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. *ACL Workshops*, 1999.

Andreas Stolcke. Srilm – an extensible language modeling toolkit. *Proceedings of the International Conference on Statistical Language Processing*, 2002.

Richard Zens, Franz Josef Och and Hermann Ney. Phrase-Based Statistical Machine Translation. *Annual German Conference on AI*, 2002.