Hallucinating Phrase Translations for Low Resource MT

Ann Irvine
Center for Language and Speech Processing
Johns Hopkins University

Chris Callison-Burch
Computer and Information Science Dept.
University of Pennsylvania

Abstract

We demonstrate that "hallucinating" phrasal translations can significantly improve the quality of machine translation in low resource conditions. Our hallucinated phrase tables consist of entries *composed* from multiple unigram translations drawn from the baseline phrase table and from translations that are induced from monolingual corpora. The hallucinated phrase table is very noisy. Its translations are low precision but high recall. We counter this by introducing 30 new feature functions (including a variety of monolingually-estimated features) and by aggressively pruning the phrase table. Our analysis evaluates the intrinsic quality of our hallucinated phrase pairs as well as their impact in end-to-end Spanish-English and Hindi-English MT.

1 Introduction

In this work, we augment the translation model for a low-resource phrase-based SMT system by automatically expanding its phrase table. We "hallucinate" new phrase table entries by composing the unigram translations from the baseline system's phrase table and translations learned from comparable monolingual corpora. The composition process yields a very large number of new phrase pair translations, which are high recall but low precision. We filter the phrase table using a new set of feature functions estimated from monolingual corpora. We evaluate the hallucinated phrase pairs intrinsically as well as in end-to-end machine translation. The augmented phrase table provides more coverage than the original phrase table, while being high quality enough to improve translation performance.

We propose a four-part approach to hallucinating and using new phrase pair translations:

1. Learn potential translations for out-of-vocabulary (OOV) words from comparable monolingual corpora
2. "Hallucinate" a large, noisy set of phrase translations by composing unigram translations from the baseline model and from the monolingually-induced bilingual dictionary
3. Use comparable monolingual corpora to score, rank, and prune the huge number of hallucinated translations
4. Augment the baseline phrase table with hallucinated translations and new feature functions estimated from monolingual corpora

We define an algorithm for generating *loosely compositional* phrase pairs, which we use to hallucinate new translations. In oracle experiments, we show that such loosely compositional phrase pairs contribute substantially to the performance of end-to-end SMT, beyond that of component unigram translations. In our non-oracle experiments, we show that adding a judiciously pruned set of automatically hallucinated phrase pairs to an end-to-end baseline SMT model results in a significant improvement in translation quality for both Spanish-English and Hindi-English.

2 Motivation

Translation models learned over small amounts of parallel data suffer from the problem of *low coverage*. That is, they do not include translations for many words and phrases. Unknown words, or out-of-vocabulary (OOV) words, have been the focus of previous work on integrating bilingual lexicon induction and machine translation (Daumé and Jagarlamudi, 2011; Irvine and Callison-Burch, 2013a; Razmara et al., 2013). Bilingual lexicon induction is the task of learning translations from monolingual texts, and typical approaches compare projected distributional signatures of words in the source language with distributional signatures representing target language words (Rapp, 1995; Schafer and Yarowsky, 2002; Koehn and Knight, 2002; Haghighi et al., 2008). If the source and target language each contain, for example, $100,000$ words, the number of pairwise comparisons is about 10 billion, which is significant but computationally feasible.

In contrast to unigrams, the difficulty in inducing a comprehensive set of *phrase* translations is that the number of both source and target phrases is immense. For example, there are about 83 million unique phrases up to length three in the English Wikipedia. Pairwise comparisons of two sets of 100 million phrases corresponds to 1 x 10^{16}. Thus, even if we limit the task to short phrases, the number of pairwise phrase comparisons necessary to do an exhaustive search is infeasible. However, multi-word translation units have been shown to improve the quality of SMT dramatically (Koehn et al., 2003). Phrase translations allow translation models to memorize local context-dependent translations and reordering patterns.

3 Approach

Rather than compare *all* source language phrases with *all* target language phrases, our approach efficiently proposes a smaller set of hypothesis phrase translations for each source language phrase. Our method builds upon the notion that many phrase translations can be composed from the translations of its component words and subphrases. For example Spanish *la bruja verde* translates into English as *the green witch*. Each Spanish word corresponds to exactly one English word. The phrase pair could be memorized and translated as a unit, or the English translation could be composed from the translations of each Spanish unigram.

Zens et al. (2012) found that only 2% of phrase pairs in German-English, Czech-English, Spanish-English, and French-English phrase tables consist of multi-word source and target phrases and are non-compositional. That is, for these languages, the vast majority of phrase pairs in a given phrase table could be composed from smaller units. Our approach takes advantage of the fact that many phrases can be translated compositionally.

We describe our approach in three parts. In Section 3.1, we begin by inducing translations for unknown unigrams. Then, in 3.2, we introduce our algorithm for composing phrase translations. In order to achieve a high recall in our set of hypothesis translations, we define compositionality more loosely than is typical. Finally, in 3.3, we use comparable corpora to prune the large set of hypothesis translations for each source phrase.

3.1 Unigram Translations

In any low resource setting, many word translations are likely to be unknown. Therefore, before moving to phrases, we use a bilingual lexicon induction technique to identify translations for unigrams. Specifically, because we assume a setting where we have some small amount of parallel data, we follow our prior work on supervised bilingual lexicon induction (Irvine and Callison-Burch, 2013b). We take examples of good translation pairs from our word aligned training data (described in Section 4) and use random word pairs as negative supervision. We use this supervision to learn a log-linear classifier that predicts whether a given word pair is a translation or not. We pair and score all source language unigrams in our tuning and test sets with target language unigrams that appear in our comparable corpora. Then, for each source language unigram, we use the log-linear model scores to rerank candidate target language unigram translations. As in our prior work, we include the following word pair features in our log-linear classifier: contextual similarity, temporal similarity, topic similarity, frequency similarity, and orthographic similarity.

3.2 Loosely Compositional Translations

We propose a novel technique for *loosely composing* phrasal translations from an existing dictionary of unigram translations and stop word lists. Given a source language phrase, our approach considers all *combinations* and all *permutations* of all unigram translations for each source phrase content word. We ignore stop words in the input source phrase and allow any number of stop words anywhere in the output target phrase. In order to make the enumeration efficient, we precompute an inverted index that maps sorted target

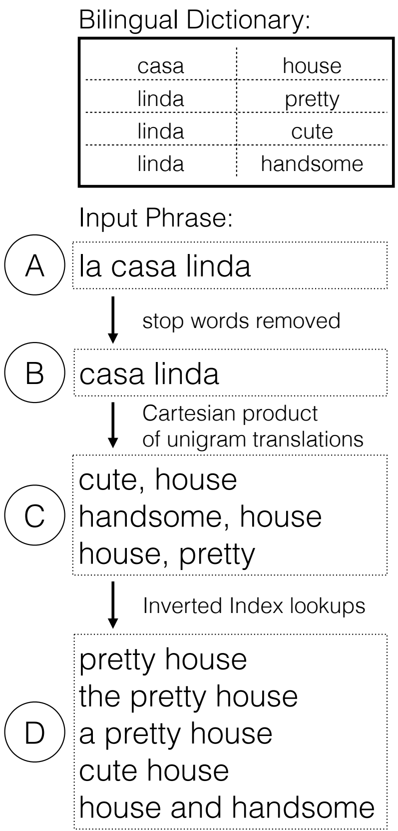

Figure 1: Example of loosely composed translations for the Spanish input in A, *la casa linda*. In B, we remove the stop word *la*. Then, in C, we enumerate the cartesian product of all unigram translations in the bilingual dictionary and sort the words within each alphabetically. Finally, we look up each list of words in C in the inverted index, and corresponding target phrases are enumerated in D. The inverted index contains all phrasal combinations and permutations of the word lists in C which also appear monolingually with some frequency and with, optionally, any number of stop words.

Input: A set of source language phrases of interest, $\mathbf{S}$, each consisting of a sequence of words $s_1^m, s_2^m, ...s_i^m$; A list of all target language phrases, $targetPhrases$; Source and target stop word lists, $Stop_{src}$ and $Stop_{trg}$; Set of unigram translations, $\mathbf{t}_{s_i^m}$, for all source language words $s_i^m \notin Stop_{src}$; monolingual target language phrase frequencies, $Freq_T$; Monolingual frequency threshold θ_{Freq_T}

Output: $\forall\ S^m \in \mathbf{S}$, a set of candidate phrase translations, $T_1^m, T_2^m, ...T_k^m$

Construct TargetInvertedIndex:
for $T \in targetPhrases$ **do**
  if $Freq_T(T) \geqslant \theta_{Freq_T}$ **then**
    T' ←words $t_j \in T$ if $t_j \notin Stop_{trg}$
    $T'_{sorted} \leftarrow$ sorted(T')
    append T to TargetInvertedIndex$[T'_{sorted}]$
  end
end

for $S^m \in \mathbf{S}$ **do**
  S' ←words $s_i^m \in S^m$ if $s_i^m \notin Stop_{src}$
  Combs$_{S'} \leftarrow \mathbf{t}_{s'_1} \times \mathbf{t}_{s'_2} \times ... \times \mathbf{t}_{s'_k}$
  $T \leftarrow [\,]$
  for $c_{s'} \in Combs_{S'}$ **do**
    $c_{s'_{sorted}} \leftarrow$ sorted$(c_{s'})$
    $T \leftarrow T+$TargetInvertedIndex$(c_{s'_{sorted}})$
  end
  $T^m = T$
end

Algorithm 1: Computing a set of candidate compositional phrase translations for each source phrase in the set S. An inverted index of target phrases is constructed that maps sorted lists of content words to phrases that contain those content words, as well as optionally any stop words, and have a frequency of at least θ_{Freq_T}. Then, for a given source phrase S^m, stop words are removed from the phrase. Next, the cartesian product of all unigram translations is computed. Each element in the product is sorted and any corresponding phrases in the inverted index are added to the output.

language content words to sets of phrases containing those words in any order along with, optionally, any number of stop words. Our algorithm for composing candidate phrase translations is given in Algorithm 1, and an example translation is composed in Figure 1. Although in our experiments we compose translations for source phrases up to length three, the algorithm is generally applicable to any set of source phrases of interest.

Algorithm 1 yields a set of target language translations for any source language phrase for which all content unigrams have at least one known translation. For most phrases, the resulting set of hypothesis translations is very large and the majority are incorrect. In an initial pruning step, we add a monolingual frequency cutoff to the composition algorithm and only add target phrases that have a frequency of at least θ_{Freq_T} to the inverted index. Doing so eliminates improbable target language constructions early on, for example *house handsome her* or *cute a house*.

3.3 Pruning Phrase Pairs Using Scores Derived from Comparable Corpora

We further prune the large, noisy set of hypothesized phrase translations before augmenting a seed translation model. To do so, we use a supervised setup very similar to that used for inducing unigram translations; we estimate a variety of signals that indicate translation equivalence, including temporal, topical, contextual, and string similarity. As we showed in Klementiev et al. (2012), such signals are effective for identifying phrase translations as well as unigram translations. We add ngram length, alignment, and unigram translation features to the set, listed in Appendix A.

We learn a log-linear model for combining the features into a single score for predicting the quality of a given phrase pair. We extract training data from the seed translation model. We rank hypothesis translations for each source phrase using classification scores and keep the top-k. We found that using a score threshold sometimes improves precision. However, as experiments below show, the recall of the set of phrase pairs is more important, and we did not observe improvements in translation quality when we used a score threshold.

4 Experimental Setup

In all of our experiments, we assume that we have access to only a small parallel corpus. For our Spanish experiments, we randomly sample $2,000$ sentence pairs (about $57,000$ Spanish words) from the Spanish-English Europarl v5 parallel corpus (Koehn, 2005). For Hindi, we use the parallel corpora released by Post et al. (2012). Again, we randomly sample $2,000$ sentence pairs from the training corpus (about $39,000$ Hindi words). We expect that this amount of parallel text could be compiled for a single text domain and any pair of modern languages. Additionally, we use approximately $2,500$ and $1,000$ single-reference parallel sentences each for tuning and testing our Spanish and Hindi models, respectively. Spanish tuning and test sets are newswire articles taken from the 2010 WMT shared task (Callison-Burch et al., 2010).[1] We use the Hindi development and testing splits released by Post et al. (2012).

4.1 Unigram Translations

Of the $16,269$ unique unigrams in the source side of our Spanish MT tuning and test sets, 73% are OOV with respect to our training corpus. 21% of unigram tokens are OOV. For Hindi, 61% of the $8,137$ unique unigrams in the tuning and test sets are OOV with respect to our training corpus, and 18% of unigram tokens are OOV. However, because automatic word alignments estimated over the small parallel training corpora are noisy, we use bilingual lexicon induction to induce translations for *all* unigrams. We use the Wikipedia and online news web crawls datasets that we released in Irvine and Callison-Burch (2013b) to estimate similarity scores. Together, the two datasets contain about 900 million words of Spanish data and about 50 million words of Hindi data. For both languages, we limit the set of hypothesis target unigram translations to those that appear at least 10 times in our comparable corpora.

We use $3,000$ high probability word translation pairs extracted from each parallel corpus as positive supervision and $9,000$ random word pairs as negative supervision. We use Vowpal Wabbit[2] for learning. The top-5 induced translations for each source language word are used as both a baseline set of new translations (Section 6.3) and for composing phrase translations.

4.2 Composing and Pruning Phrase Translations

There are about 183 and 66 thousand unique bigrams and trigrams in the Spanish and Hindi tuning and test sets, respectively. However, many of these phrases do not demand new hypothesis translations. We do not translate those which contain numbers or punctuation. Additionally, for Spanish, we exclude names, which are typically translated identically between Spanish and English.[3] We exclude phrases which are sequences of stop words only. Additionally, we exclude phrases that appear more than 100 times in the small training corpus because our seed phrase table likely already contains high quality translations for them. Finally, we exclude phrases that appear fewer than 20 times in our comparable corpora as our features are unreliable when estimated over so few tokens. We hypothesize translations for the approximately 15 and 6 thousand Spanish and Hindi phrases, respectively, which meet these criteria. Our approach for inducing translations straightforwardly generalizes to any set of source phrases.

In defining loosely compositional phrase translations, we use both the induced unigram dictionary (Section 3.1) and the dictionary extracted from the word aligned parallel corpus. We expand these dictionaries further by mapping unigrams to their five-character word prefixes. We use monolingual corpora of Wikipedia articles[4] to construct stop word lists, containing the most frequent 300 words in each language, and indexes of monolingual phrase frequencies. There are about 83 million unique phrases up to length three in the English Wikipedia. However, we ignore target phrases that appear fewer than three times, reducing this set to 10 million English phrases. On

[1] *news-test2008* plus *news-syscomb2009* for tuning and *newstest2009* for testing.

[2] `http://hunch.net/~vw/`, version 6.1.4. with standard learning parameters

[3] Our names list comes from page titles of Spanish Wikipedia pages about people. We iterate through years, beginning with 1AD, and extract names from Wikipedia 'born in' category pages, e.g. '2013 births,' or 'Nacidos en 2013.'

[4] All inter-lingually linked source language and English articles.

average, our Spanish model yields $7,986$ English translations for each Spanish bigram, and $9,231$ for each trigram, or less than 0.1% of all possible candidate English phrases. Our Hindi model yields even fewer candidate English phrases, 826 for each bigram and $1,113$ for each trigram, on average.

We use the same comparable corpora used for bilingual lexicon induction to estimate features over hypothesis phrase translations. The full feature set is listed in Appendix A. We extract supervision from the seed translation models by first identifying phrase pairs with multi-word source strings, that appear at least three times in the training corpus, and that are composeable using baseline model unigram translations and induced dictionaries. Then, for each language pair, we use the $3,000$ that have the highest $p(f|e)$ scores as positive supervision. We randomly sample $9,000$ compositional phrase pairs from those not in each phrase table as negative supervision. Again, we use Vowpal Wabbit for learning a log linear model to score any phrase pair.

4.3 Machine Translation

We use GIZA++ to word align each training corpus. We use the Moses SMT framework (Koehn et al., 2007) and the standard phrase-based MT feature set, including phrase and lexical translation probabilities and a lexicalized reordering model. When we augment our models with new translations, we use the average reordering scores over all bilingually estimated phrase pairs. We tune all models using batch MIRA (Cherry and Foster, 2012). We average results over three tuning runs and use approximate randomization to measure statistical significance (Clark et al., 2011).

For Spanish, we use a 5-gram language model trained on the English side of the complete Europarl corpus and for Hindi a 5-gram language model trained on the English side of the complete training corpus released by Post et al. (2012). We train our language models using SRILM with Kneser-Ney smoothing. Our baseline models use a phrase limit of three, and we augment them with translations of phrases up to length three in our experiments.

5 Oracle Experiment

Before moving to the results of our proposed approach for composing phrase translations, we present an oracle experiment to answer these research questions: Would a low resource translation model benefit from composing its unigram translations into phrases? Would this be further improved by adding unigram translations that are learned from monolingual texts? We answer these questions by starting with our low-resource Spanish-English and Hindi-English baselines and augmenting each with (1) phrasal translations composed from baseline model unigram translations, and (2) phrasal translations composed of a mix of baseline model unigram translations and the monolingually-induced unigrams.

Figure 2 illustrates how our hallucinated phrase-table entries can result in improved translation quality for Spanish to English translation. Since the baseline model is trained from such a small amount of data, it typically translates individual words instead of phrases. In our augmented system, we compose a translation of *was no one* from *habia nadie,* since *habia* translates as *was* in the baseline model, *nadie* translates as *one,* and *no* is a stop word. We are able to monolingually-induce translations for the OOVs *centros* and *electorales* before composing the phrase translation *polling stations* for *centros electorales*.

In our oracle experiments, composed translations are only added to the phrase table if they are contained in the reference. This eliminates the huge number of noisy translations that our compositional algorithm generates. We augment baseline models with translations for the same sets of source language phrases described in Section 4. We use GIZA++ to word align our tuning and test sets[5] and use a standard phrase pair extraction heuristic[6] to identify oracle phrase translations. We add oracle translations to each baseline model *without* bilingually estimated translation scores[7] because such scores are not available for our automatically induced translations. Instead, we score the oracle phrase pairs using the 30 new phrase table features described in Section 3.3.

Table 1 shows the results of our oracle experiments. Augmenting the baselines with the subset of oracle translations which are *composed* given the unigram translations in the baseline models themselves (i.e. in the small training sets) yields

[5] For both languages, we learn an alignment over our tuning and test sets and complete parallel training sets.

[6] grow-diag-final

[7] We use an indicator feature for distinguishing new composed translations from bilingually extracted phrase pairs.

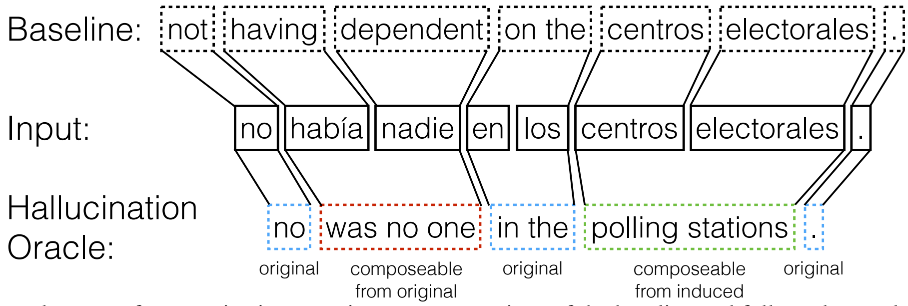

Figure 2: Example output from motivating experiment: a comparison of the baseline and full oracle translations of Spanish *no había nadie en los centros electorales*, which translates correctly as *there was nobody at the voting offices*. The full oracle is augmented with translations composed from the seed model as well as induced unigram translations. The phrase *was no one* is composeable from *había nadie* given the seed model. In contrast, the phrase *polling stations* is composeable from *centros electorales* using induced translations. For each translation, the phrase segmentations used by the decoder are highlighted.

Experiment	BLEU Baseline Features	BLEU Monolingually Estimated Feats.
Spanish		
Low Resource Baseline	13.47	13.35
+ Composeable Oracle from Initial Model	14.90	15.18
+ Composeable Oracle w/ Induced Unigram Trans.	15.47	15.94
Hindi		
Low Resource Baseline	8.49	8.26
+ Composeable Oracle from Initial Model	9.12	9.54
+ Composeable Oracle w/ Induced Unigram Trans.	10.09	10.19

Table 1: Motivating Experiment: BLEU results using the baseline SMT model and composeable oracle translations with and without induced unigram translations.

a BLEU score improvement of about 1.4 points for Spanish and about 0.6 for Hindi. This finding itself is noteworthy, and we investigated the reason for it. A representative example of a compositional oracle translation that was added to the Spanish model is *para evitarlos*, which translates as *to prevent them*. In the training corpus, *para* translates far more frequently as *for* than *to*. Thus, it is useful for the translation model to know that, in the context of *evitarlos*, *para* should translate as *to* and not *for*. Additionally, *evitarlos* was observed only translating as the unigram *prevent*. The small model fails to align the adjoined clitic *los* with its translation *them*. However, our loose definition of compositionality allows the English stop word *them* to appear anywhere in the target translation.

In the first result, composeable translations do not include those that contain new, induced word translations. Using the baseline model *and* induced unigram translations to compose phrase translations results in a 2 and 1.6 BLEU point gain for Spanish and Hindi, respectively.

The second column of Table 1 shows the results of augmenting the baseline models with the same oracle phrase pairs as well as the new features estimated over *all* phrase pairs. Although the features do not improve the performance of the baseline models, this diverse set of scores improves performance dramatically when new, oracle phrase pairs are added. Adding all oracle translations and the new feature set results in a total gain of about 2.6 BLEU points for Spanish and about 1.9 for Hindi. These gains are the maximum that we could hope to achieve by augmenting models with our hallucinated translations and new feature set.

6 Experimental Results

6.1 Unigram Translations

Table 2 shows examples of top ranked translations for several Spanish words. Although performance is generally quite good, we do observe some instances of false cognates, for example the top ranked translation for *aburridos*, which translates correctly as *bored*, is *burritos*. Using automatic word alignments as a reference, we find that 44% of Spanish tuning set unigrams have a correct translation in their top-10 ranked lists and 62% in the top-100. For Hindi, 31% of tuning set unigrams have a correct translation in their top-10 ranked lists and 43% in the top-100.

6.2 Hallucinated Phrase Pairs

Before moving to end-to-end SMT experiments, we evaluate the goodness of the hallucinated and pruned phrase pairs themselves. In order to do so, we use the same set of oracle phrase translations described in Section 5.

Table 3 shows the top three English translations for several Spanish phrases along with their model scores. Common, loose translations of some phrases are scored higher than less common but literal translations. For example, *very obvi-*

Spanish	abdominal	abejorro	abril	aburridos	accionista	aceite	actriz
Top 5 English Translations	**abdominal**	**bumblebees**	**april**	burritos	actionists	adulterated	**actress**
	abdomen	bombus	march	**boredom**	actionist	iooc	actor
	bowel	xylocopa	june	agatean	telmex	olive	award
	appendicitis	ilyitch	july	burrito	**shareholder**	milliliters	american
	acute	bumble	december	poof	antagonists	canola	singer

Table 2: Top five induced translations for several source words. Correct translations are bolded. *aceite* translates as *oil*.

Spanish	English	Score
ambos partidos	**two parties**	5.72
	both parties	5.31
	and parties	3.16
había apoyado	were supported	4.80
	were members	4.52
	had supported	4.39
ministro neerlandès	finnish minister	4.76
	finnish ministry	2.77
	dutch minister	1.31
unas cuantas semanas	over a week	4.30
	a few weeks	3.72
	few weeks	3.22
muy evidentes	**very obvious**	1.88
	very evident	1.87
	obviously very	1.84

Table 3: Top three compositional translations for several source phrases and their model scores. Correct translations are bolded.

ous scores higher than *very evident* as a translation of Spanish *muy evidentes*. Similarly, *dutch minister* is scored higher than *netherlands minister* as a translation for *ministro neerlandès*.

We use model scores to rerank candidate translations for each source phrase and keep the top-k translations. Figure 3 shows the precision and type-based recall (the percent of source phrases for which at least one correct translation is generated) as we vary k for each language pair. At $k = 1$, precision and recall are about 27% for Spanish and about 25% for Hindi.[8] At $k = 200$, recall increases to 57% for Spanish and precision drops to 2%. For Hindi, recall increases to 40% and precision drops to 1%.

Moving from $k = 1$ to $k = 200$, precision drops at about the same rate for the two source languages. However, recall increases less for Hindi than for Spanish. We attribute this to two things. First, Hindi and English are less related than Spanish and English, and fewer phrases are translated compositionally. Our oracle experiments showed that there is less to gain in composing phrase translations for Hindi than for Spanish. Second, the accuracy of our induced unigram translations is lower for Hindi than it is for Spanish. Without accurate unigram translations, we are unable to compose high quality phrase translations.

[8] Since we are computing type-based recall, and at k=1, we produce exactly one translation for each source phrase, precision and recall are the same.

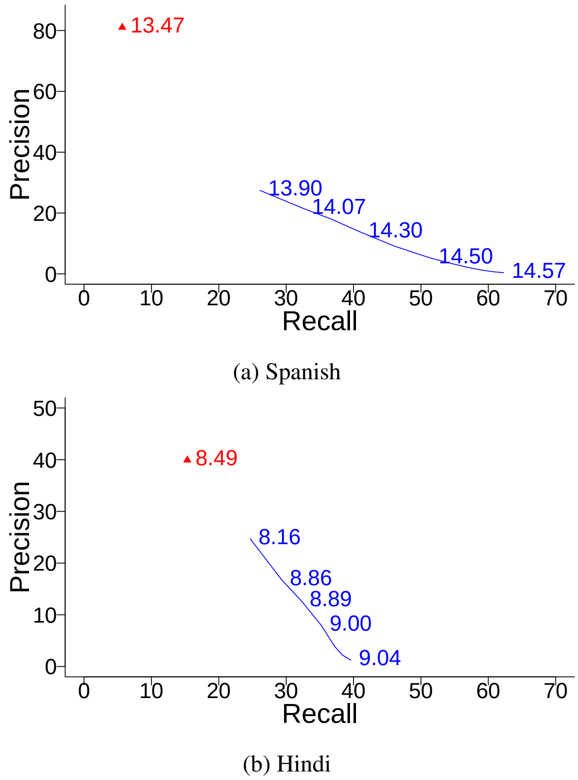

Figure 3: Precision Recall curve with BLEU scores for the top-k scored hallucinated translations. k varies from 1 to 200. Baseline model performance is shown with a red triangle.

Because we hallucinate translations for source phrases that appear in the training data up to 100 times, our baseline models include some of the oracle phrase translations. Not surprisingly, the bilingually extracted phrase pairs have relatively high precision (81% and 40% for Spanish and Hindi, respectively) and low recall (6% and 15% for Spanish and Hindi, respectively).

6.3 End-to-End Translation

Table 4 shows end-to-end translation BLEU score results (Papineni et al., 2002). Our first baseline SMT models are trained using only 2,000 parallel sentences and no new translation model features. Our Spanish baseline achieves a BLEU score of 13.47 and our Hindi baseline a BLEU score of 8.49. When we add the 30 new feature functions estimated over comparable monolingual corpora, performance is slightly lower, 13.35 for

Experiment	BLEU Spanish	BLEU Hindi
Baseline	13.47	**8.49**
+ Mono. Scores	13.35	8.26
+ Mono. Scores & OOV Trans	**14.01**	8.31
+ Phrase Trans, k=1	13.90	8.16
+ Phrase Trans, k=2	14.07	8.86*
+ Phrase Trans, k=5	14.30*	8.89*
+ Phrase Trans, k=25	14.50*	9.00*
+ Phrase Trans, k=200	**14.57***	**9.04***

Table 4: Experimental results. First, the baseline models are augmented with monolingual phrase table features and then also with the top-5 induced translations for all OOV unigrams. Then, we append the top-k hallucinated phrase translations to the third baseline models. BLEU scores are averaged over three tuning runs. We measure the statistical significance of each +Phrase Trans model in comparison with the highest performing (bolded) baseline for each language; * indicates statistical significance with $p < 0.01$.

Spanish and 8.26 for Hindi. Our third baselines augment the second with unigram translations for all OOV tuning and test set source words using the bilingual lexicon induction techniques described in Section 3.1. We append the top-5 translations for each,[9] score both the original and the new phrase pairs with the new feature set, and retune. With these additional unigram translations, performance increases to 14.01 for Spanish and 8.31 for Hindi.

We append the top-k composed translations for the source phrases described in Section 4 to the third baseline models. Both original and new phrase pairs are scored using the new feature set. BLEU score results are shown at different values of k along the precision-recall plots for each language pair in Figure 3 as well as in Table 4. We would expect that higher precision and higher recall would benefit end-to-end SMT. As usual, a tradeoff exists between precision and recall, however, in this case, improvements in recall outweigh the risk of a lower precision. As k increases, precision decreases but both recall and BLEU scores increase. For both Spanish and Hindi, BLEU score gains start to taper off at k values over 25.

In additional experiments, we found that **without** the new features the same sets of hallucinated phrase pairs hurt performance slightly in comparison with the baseline augmented with unigram translations, and results don't change as we vary k.[10] Thus, the translation models are able to effectively use the higher recall sets of new phrase pairs because we also augmented the models with 30 new feature functions, which help them distinguish good from bad translations.

7 Discussion

Our results showed that including a high recall set of "hallucinated" translations in our augmented phrase table successfully improved the quality of our machine translations. The algorithm that we proposed for hypothesizing translations is flexible, and in future work we plan to modify it slightly to output even more candidate translations. For example, we could retrieve target phrases which contain at least one source word translation instead of all. Alternatively, we could identify candidates using entirely different information, for example the monolingual frequency of a source and target word, instead of unigram translations. This type of inverted index may improve recall in the set of hypothesis phrase translations at the cost of generating a much bigger set for reranking.

Our new phrase table features were informative in distinguishing correct from incorrect phrase translations, and they allowed us to make use of noisy but high recall supplemental phrase pairs. This is a critical result for research on identifying phrase translations from non-parallel text. We also believe that using fairly strong target (English) language models contributed to our models' ability to discriminate between good and bad hallucinated phrase pairs. We leave research on the influence of the language model in our setting to future work.

In this work, we experimented with two language pairs, Spanish-English and Hindi-English. While Spanish and English are very closely related, Hindi and English are less related. Our oracle experiments showed potential for composing phrase translations for both language pairs, and, indeed, in our experiments using hallucinated phrase translations we saw significant translation quality gains for both. We expect that improving the quality of induced unigram translations will yield even more performance gains.

The vast majority of prior work on low resource MT has focused on Spanish-English (Haghighi et al., 2008; Klementiev et al., 2012; Ravi and Knight, 2011; Dou and Knight, 2012; Ravi, 2013; Dou and Knight, 2013). Although such experiments serve as important proofs of concept, we found it important to also experiment with a more

[9]The same set used for composing phrase translations.

[10]For all values of k between 1 and 100, without the new features, BLEU scores are about 13.70 for Spanish

truly low resource language pair. The success of our approach that we have seen for Spanish and Hindi suggests that it is worth pursuing such directions for other even less related and resourced language pairs. In addition to language pair, text genre and the degree of looseness or literalness of given parallel corpora may also affect the amount of phrase translation compositionality.

8 Related Work

Phrase-based SMT models estimated over very large parallel corpora are expensive to store and process. Prior work has reduced the size of SMT phrase tables in order to improve efficiency without the loss of translation quality (He et al., 2009; Johnson et al., 2007; Zens et al., 2012). Typically, the goal of pruning is to identify and remove phrase pairs which are likely to be inaccurate, using either the scores and counts of a given pair itself or those relative to other phrase pairs. Our work, in contrast, focuses on low resource settings, where training data is limited and provides incomplete and unreliable scored phrase pairs. We begin by dramatically *increasing* the size of our SMT phrase table in order to expand its coverage and then use non-parallel data to rescore and filter the table.

In the decipherment task, translation models are learned from comparable corpora without any parallel text (Ravi and Knight, 2011; Dou and Knight, 2012; Ravi, 2013). In contrast, we begin with a small amount of parallel data and take a very different approach to learning translation models. In our prior work (Irvine and Callison-Burch, 2013b), we showed how effective even small amounts of bilingual data can be for learning translations from monolingual texts.

Garera and Yarowsky (2008) pivot through bilingual dictionaries in several language pairs to compose translations for compound words. Zhang and Zong (2013) construct a set of new, additional phrase pairs for the task of domain adaptation for machine translation. That work uses two dictionaries to bootstrap a set of phrase pair translations: one probabilistic dictionary extracted from 2 million words of bitext and one manually created new-domain dictionary of $140,000$ word translations. Our approach to the construction of new phrase pairs is somewhat similar to Zhang and Zong (2013), but we don't rely on a very large manually generated dictionary. Additionally, we focus on the low resource language pair setting, where a large training corpus is not available.

Deng et al. (2008) work in a standard SMT setting but use a discriminative framework for extracting phrase pairs from parallel corpora. That approach yields a phrase table with higher precision and recall than the table extracted by standard world alignment based heuristics (Och and Ney, 2003; Koehn et al., 2003). The discriminative model combines features from word alignments and bilingual training data as well as information theoretic features estimated over monolingual data into a single log-linear model and then the phrase pairs are filtered using a threshold on model scores. The phrase pairs that it extracts are limited to those that appear in pairs of sentences in the parallel training data. Our work takes a similar approach to that of Deng et al. (2008), however, unlike that work, we *hallucinate* phrase pairs that did *not* appear in training data in order to augment the original, bilingually extracted phrase table.

Other prior work has used comparable corpora to extract parallel sentences and phrases (Munteanu and Marcu, 2006; Smith et al., 2010). Such efforts are orthogonal to our approach. We use parallel corpora, when available, and hallucinates phrase translations without assuming any parallel text in our comparable corpora.

9 Conclusions

We showed that "hallucinating" phrasal translations can significantly improve machine translation performance in low resource conditions. Our hallucinated translations are *composed* from unigram translations. The translations are low precision but high recall. We countered this by introducing new feature functions and pruning aggressively.

10 Acknowledgements

This material is based on research sponsored by DARPA under contract HR0011-09-1-0044 and by the Johns Hopkins University Human Language Technology Center of Excellence. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Hal Daumé, III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Yonggang Deng, Jia Xu, and Yuqing Gao. 2008. Phrase table training for precision and recall: What makes a good phrase and a good phrase pair? In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Qing Dou and Kevin Knight. 2013. Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nikesh Garera and David Yarowsky. 2008. Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Zhongjun He, Yao Meng, and Hao Yu. 2009. Discarding monotone composed rule for hierarchical phrase-based statistical machine translation. In *Proceedings of the 3rd International Universal Communication Symposium*.

Ann Irvine and Chris Callison-Burch. 2013a. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Ann Irvine and Chris Callison-Burch. 2013b. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*.

Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Dragos Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Sujith Ravi. 2013. Scalable decipherment for machine translation via hash sampling. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Richard Zens, Daisy Stanton, and Peng Xu. 2012. A systematic comparison of phrase table pruning techniques. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Jiajun Zhang and Chengqing Zong. 2013. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Appendix A: Phrase pair filtering features

The first ten features are similar to those described by Irvine and Callison-Burch (2013b). Stop words are defined as the most frequent 300 words in each language's Wikipedia, and content words are all non-stop words.

- Web crawl phrasal context similarity score
- Web crawl lexical context similarity score, averaged over aligned unigrams
- Web crawl phrasal temporal similarity score
- Web crawl lexical temporal similarity score, averaged over aligned unigrams
- Wikipedia phrasal context similarity score
- Wikipedia lexical context similarity score, averaged over aligned unigrams
- Wikipedia phrasal topic similarity score
- Wikipedia lexical topic similarity score, averaged over aligned unigrams
- Normalized edit distance, averaged over aligned unigrams
- Absolute value of difference between the logs of the source and target phrase Wikipedia monolingual frequencies
- Log target phrase Wikipedia monolingual frequency
- Log source phrase Wikipedia monolingual frequency
- Indicator: source phrase is longer
- Indicator: target phrase is longer
- Indicator: source and target phrases same length
- Number of source content words higher than target
- Number of target content words higher than source
- Number of source and target content words same
- Number of source stop words higher than target
- Number of target stop words higher than source
- Number of source and target stop words same
- Percent of source words aligned to at least one target word
- Percent of target words aligned to at least one source word
- Percent of source content words aligned to at least one target word
- Percent of target content words aligned to at least one source word
- Percent of aligned word pairs aligned in bilingual training data
- Percent of aligned word pairs in induced dictionary
- Percent of aligned word pairs in stemmed induced dictionary