Effective Phrase Translation Extraction from Alignment Models

Ashish Venugopal
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ashishv@cs.cmu.edu

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
vogel+@cs.cmu.edu

Alex Waibel
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ahw@cs.cmu.edu

Abstract

Phrase level translation models are effective in improving translation quality by addressing the problem of local re-ordering across language boundaries. Methods that attempt to fundamentally modify the traditional IBM translation model to incorporate phrases typically do so at a prohibitive computational cost. We present a technique that begins with improved IBM models to create phrase level knowledge sources that effectively represent local as well as global phrasal context. Our method is robust to noisy alignments at both the sentence and corpus level, delivering high quality phrase level translation pairs that contribute to significant improvements in translation quality (as measured by the BLEU metric) over word based lexica as well as a competing alignment based method.

1 Introduction

Statistical Machine Translation defines the task of translating a source language sentence ($s = s_1 \cdots s_I$) into a target language sentence ($t = t_1 \cdots t_J$). The traditional framework presented in (Brown et al., 1993) assumes a generative process where the source sentence is passed through a noisy stochastic process to produce the target sentence. The task can be formally stated as finding the $\hat{t}$ s.t $\hat{t}$ = $argmax\ p(t|s)$ where the search component is commonly referred to as the decoding step (Wang and Waibel, 1998). Within the generative model, the Bayes reformulation is used to estimate $p(t|s) \sim p(t)p(s|t)$ where $p(t)$ is considered the language model, and $p(s|t)$ is the translation model; the IBM (Brown et al., 1993) models being the de facto standard. Direct translation approaches (Foster, 2000) consider estimating $p(t|s)$ directly, and work by (Och and Ney, 2002) show that similar or improved results are achieved by replacing $p(s|t)$ in the optimization with $p(t|s)$, at the cost of deviating from the Bayesian framework. Regardless of the approach, the question of accurately estimating a model of translation from a large parallel or comparable corpus is one of the defining components within statistical machine translation.

Re-ordering effects across languages have been modeled in several ways, including word-based (Brown et al., 1993), template-based (Och et al., 1999) and syntax-based (Yamada, Knight, 2001). Analyzing these models from a generative mindset, they all assume that the atomic unit of lexical content is the word, and re-ordering effects are applied above that level. (Marcu, Wong, 2002) illustrate the effects of assuming that lexical correspondence can only be modeled at the word level, and motivate a joint probability model that explicitly generates phrase level lexical content across both languages. (Wu, 1995) presents a bracketing method that models re-ordering at the sentence level. Both (Marcu, Wong, 2002; Wu, 1995) model the reordering phenomenon effectively, but at significant computational expense, and tend to be difficult to scale to long sentences. Reasons to introduce phrase level translation knowledge sources have been adequately shown and confirmed by (Och, Ney, 2000), and we focus on methods to build these sources from existing, mature components within the translation process.

This paper presents a method of phrase extraction from alignment data generated by IBM Models. By working directly from alignment data with appropriate measures taken to extract accurate translation pairs, we try to avoid the computational complexity that can result from methods that try to create globally consistent alignment model phrase segmentations.

We first describe the information available within alignment data, and go on to describe a method for extracting high quality phrase translation pairs from such data. We then discuss the implications of adding phrasal translation pairs to the decoding process, and present evaluation results that show significant improvements when applying the described extraction technique. We end with a discussion of strengths and weaknesses of this method and the potential for future work.

2 Motivation

Alignment models associate words and their translations at the sentence level creating a translation lexicon across the language pair. For each sentence pair, the model also presents the maximally likely association between each source and target word across the sentence pair, forming an alignment map for each sentence pair in the training corpus. The most likely alignment pattern between a source and target sentence under the trained alignment model will be referred to as the maximum approximation, which under HMM alignment (Vogel et al., 1996) model corresponds to the Viterbi path. A set of words in the source sentence associated with a set of words in the target sentence is considered a phrasal pair and forms a partition within the alignment map. Figure 1. shows a source and target sentence pair with points indicating alignment points.

A phrasal translation pair within a sentence pair can be represented as the 4-tuple hypothesis $H_p(i, l_s, j, l_t)$ representing an index (i, j) and length (l_s, l_t) within the source and the target sentence pair p, respectively. The phrasal extraction task involves selecting phrasal hypotheses based on the alignment

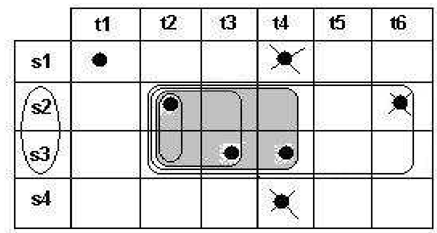

Figure 1: Sample source S_i and target T_i aligment map. Partitions/Potential translations for source phrase s2s3 are shown by rounded boxes.

model (both the translation lexicon as well as the maximal approximation). The maximal approximation captures context at the sentence level, while the lexicon provides a corpus level translation estimate, motivating the alignment model as a starting point for phrasal extraction. The extraction technique must be able to handle alignments that are only partially correct, as well as cases where the sentence pairs have been incorrectly matched as parallel translations within the corpus. Accommodating for the noisy corpus is an increasingly important component of the translation process, especially when considering languages where no manually aligned parallel corpus is available.

Building a phrasal lexicon involves Generation, Scoring, and Pruning steps, corresponding to generating a set of candidate translation pairs, scoring them based on the translation model, and pruning them to account for noise within the data as well as the extraction process.

3 Generation

The generation step refers to the process of identifying source phrases that require translations and then extracting translations from the alignment model data. We begin by identifying all source language n-grams upto some n within the training corpus. When the test sentences that require translation are known, we can simply extract those n-grams that appear in the test sentences. For each of these n-grams, we create a set of candidate translations extracted from the corpus. The primary motivation to restrict the identification step to the test sentence n-grams is savings in computational expense, and the result is a phrasal translation source that extracts translation pairs limited to the test sentences. For each source language n-gram within the pool, we have to find a set of candidate translations. The generation task is formally defined as finding $\hat{H}_g$ in Equation (1)

$$\hat{H}_g \; s.t \, \forall \, H_p(i, l_s, j, l_t) \; \in \; \hat{H}, \; p_i \cdots p_{i+l_s} = g \quad (1)$$

where g is the source n-gram for which we are extracting translations, $\hat{H}$ is the set of all partitions, and p_i refers to the word at position i in the source sentence p. $\hat{H}_g$ is then the set of all translations for source n-gram g, and h is a specific translation hypothesis within this set. When considering only those hypothesis translation extracted from a particular sentence pair p, we use $\hat{H}_g(p)$.

We extract these candidates from the alignment map by examining each sentence pair where the source n-gram occurs, and extracting all possible target phrase translations using a sliding window approach. We extract candidate translations of phrase length 1 to I, starting at offset 0 to $I - 1$. Figure 1. shows circular boxes indicating each potential partition region. One particular partition is indicated by the shading.

Over all occurrences of the n-gram within the sentences as well as across sentences, a sizeable candidate pool is generated that attempts the cover the translated usage of the source n-gram g within the corpus. This set is large, and contains several spurious translations, and does not consider other source side n-grams within each sentence. The deliberate choice to avoid creating a consistent partitioning of the sentence pairs across n-grams reflects the ability to model partially correct alignments within sentences. This sliding window can be restricted to exclude word-word translations, ie $l_s \neq 1$, $l_t \neq 1$ if other sources are available that are known to be more accurate. Now that the candidate pool has been generated, it needs to be scored and pruned to reflect relative confidence between candidate translations and to remove spurious translations due to the sliding window approach.

4 Scoring

The candidate translations for the source n-gram now need to be scored and ranked according to some measure of confidence. Each candidate translation pair defines a partition within the sentence map, and this partitioning can be scored for confidence in translation quality. We estimate translation confidence by measures from three models; the estimation from the maximum approximation (alignment map), estimation from the word based translation lexicon, and language specific measures. Each of the scoring methods discussed below contributes to the final score under (2)

$$FinalScore(h \in \hat{H}_g) \;=\; \prod_i (Score_i(h \in \hat{H}_g))^{w_i} \quad (2)$$

where $\sum_i w_i = 1$ and h refers to a translation hypothesis for a given source n-gram g. From now on we will refer to a $Score$ with regard to a particular g implicitly.

4.1 Alignment Map

We define two kinds of scores, within sentence consistency and across sentence consistency from the alignment map, in order to represent local and global context effects.

4.2 Within Sentence

The partition defined by each candidate translation pair imposes constraints over the maximum approximation hypothesis for sentences in which it occurs. We evaluate the partition by examining its consistency with the maximum approximation hypothesis by considering the alignment hypothesis points within the sentence. An alignment point $A_p(x, y)$ (source, target) is said to be consistent if it occurs within the partition defined by $H_p(i, l_s, j, l_t)$. $A_{x,y}$ is considered inconsistent in two cases.

$$i \;\leq\; x \;\leq\; i + l_s \text{ and} (\; y \;< j \text{ or } y > j + l_t) \quad (3)$$

$$j \;\leq\; y \;\leq\; j + l_t \text{ and } (x < i \text{ or } x > i + l_s) \quad (4)$$

Each $H_p(i, l_s, j, l_t)$ in $\hat{H}_g(p)$ ($i \cdots i{+}l_s$ defines g) determines a set of consistent and inconsistent points. Figure 1. shows inconsistent points with respect to the shaded partition by drawing an X over the alignment point. The within sentence consistency scoring metric is defined in Equation (5).

$$Score_{ws}(H_p(i, l_s, j, l_t)) = \frac{\#cons}{\#incons + \#cons} \quad (5)$$

This measure represents consistency of $H_p(i, l_s, j, l_t)$ within the maximal approximation alignment for sentence pair p.

4.3 Across Sentence

Several hypothesis within $\hat{H}_g(p)$ are similar or identical to those in $\hat{H}_g(q)$ where $p \neq q$. We want to score hypothesis that are consistent across sentences higher than those that occur rarely, as the former are assumed to be the correct translations in context. We want to account for different contexts across sentences; therefore we want to highlight similar translations, not simply exact matches. We use a word level Levenstein distance to compare the target side hypotheses within $\hat{H}_g$. Each element h within $\hat{H}_g$ (the complete candidate translation list for g) is assigned the average Levenstein distance with all other elements as its across sentence consistence score; effectively performing a single pass average link clustering to identify the correct translations.

$$Score_{as}(h) = \frac{1}{\frac{1}{N}\sum_{\acute{h} \in H} LD(h, \acute{h})} \quad (6)$$

where LD calculates the Levenshein distance between the target phrases within two hypothesis h and $\acute{h}$, N is the number of elements in $\hat{H}_g$.

The higher the $Score_{as}$, the more likely the hypothesis pair is a correct translation. The clustering approach accounts for noise due to incorrect sentence alignment, as well as the different contexts in which a particular source n-gram can be used. As predicted by the formulation of this method, preference is given towards shorter target translations. This effect can be countered by introducing a phrase length model to approximate the difference in phrases lengths across the language boundary. This will be discussed further as a language specific scoring method.

4.4 Alignment Lexicon

The methods presented above used the maximum approximation to score candidate translation hypotheses. The translation lexicon generated by the IBM models provides translation estimates at the word level built on the complete training corpus. These corpus level estimates can be integrated into our scoring paradigm to balance the sentence level estimates from the alignment map methods.

The translation lexicon provides a conditional probability estimate $p(s_x|t_y)$ for each $A_p(x, y)$ (s_x refers to the word at position x in sentence p) within the maximum approximation. Depending on the direction in which the traditional IBM models are trained, we can either condition on the source or target side, while joint probability models can give us a bidirectional estimate. These translation probability estimates are used to weight the $A_p(x, y)$ within the methods described above. Instead of simply counting the number of consistent/inconsistent $A_p(x, y)$, we sum the probability estimates $p(s_x|t_y)$ for each $A_p(x, y)$. So far we have only considered the points within the partition where alignment points are predicted by the maximal approximation. The translation lexicon provides estimates at the word level, so we can construct a scoring measure for the complete region within $H_p(i, l_s, j, l_t)$ that models the complete probability of the partition. The lexical scoring equation below models this effect.

$$Score_{lex}(H_p(i, l_s, j, l_t)) = \prod_{i \le x \le l_s} \sum_{j \le y \le l_t} p(s_x|t_y) \quad (7)$$

This method prefers longer target side phrases due to the sum over the target words within the partition. Although it would also prefer short source side phrases, we are only concerned with comparing hypothesis partitions for a given source n-gram g.

4.5 Language Specific

The nature of the phrasal association between languages varies depending on the level of inflexion, morphology as well as other factors. The predominant language specific correction to the scoring techniques discussed above models differences in phrase lengths across languages. For example, when comparing English and Chinese translations, we see that on average, the English sentence is approximately 1.3 times longer (under our current segmentation in the small data track). To model these language specific effects, we introduce a phrase length scoring component that is based on the ratio of sentence length between languages. We build a sentence length model based on the DiffRatio statistic defined as $DiffRatio = \frac{I-J}{J}$ where I is the source sentence length and J is the target sentence length. Let μ_{DR} be the average $DiffRatio$ over the sentences in the corpus, and σ^2_{DR} be the variance; thereby defining a normal distribution over the DiffRatio statistic. Using the standard Z normalization technique under a normal distribution parameterized by μ_{DR}, σ^2_{DR}, we can estimate the probability that a new DiffRatio calculated on the phrasal pair can be generated by the model, giving us the scoring estimate below.

$$Score_{len}(H_p(i, l_s, j, l_t)) = P(l_s, l_t | \{\mu_{DR}, \sigma^2_{DR}\}) \tag{8}$$

To improve the model we might consider examining known phrase translation pairs if this data is available. We explore the language specific difference further by noting that English phrases contain several function words that typically align to the empty Chinese word. We accounted for this effect within the scoring process by treating all target language (English) phrases that only differed by the function words on the phrase boundary as the same translation. The burden of selecting the appropriate hypothesis within the decoding process is moved towards the language model under this corrective strategy.

5 Pruning

The list of candidate translations for each source n-gram g is large, and must be pruned to select the most likely set of translations. This pruning is required to ensure that the decoding process remains computationally tractable. Simple threshold methods that rank hypotheses by their final score and only save the top N hypotheses will not work here, since phrases differ in the number of possible correct translations they could have when used in different contexts. Given the score ordered set of candidate phrases $\hat{H}_g$, we would like to label some subset as incorrect translations and remove them from the set. We approach this task as a density estimation problem where we need to separate the distribution of the incorrectly translated hypothesis from the distribution of the likely translations. Instead of using the maximum likelihood criteria, we use the maximal separation criteria ie. selecting a splitting point within the scores to maximize the difference of the mean score between distributions as shown below.

$$SplitScore = argmax_p(\mu_{h<p} - \mu_{h \geq p}) \tag{9}$$

where $\mu_{h<p}$ is the mean score of those hypothesis with a score less than p, and $\mu_{h \geq p}$ is the mean score of those hypothesis with a greater than or equal to p. Once pruning is completed, we convert the scores into a probability measure conditioned on the source n-gram g and assign the probability estimate as the translation probability for the hypothesis h as shown below.

$$p(t \in h | s \equiv g) = \frac{FinalScore(h)}{\sum_{\acute{h} \in \hat{H}_g} FinalScore(\acute{h})} \tag{10}$$

(10) calculates direct translation probabilities, ie $p(t|s)$. As mentioned earlier, (Och and Ney, 2002), show that using direction translation estimates in the decoding process as compared with calculating $p(s|t)$ as prescribed by the Bayesian framework does not reduce translation quality. Our results corroborate these findings and we use (10) as the phrase level translation model estimate within our decoder.

6 Integration

Phrase translation pairs that are generated by the method described in this paper are finally scored with estimates of translation probability, which can be conditioned on the target language if necessary. These estimates fit cleanly into the decoding process, except for the issue of phrase length. Traditional word lexicons propose translations for one source word, while with phrase translations, a single hypothesis pair can span several words in the source or target language. Comparing between a path that uses a phrase compared to one that uses multiple words (even if the constituent words are the same) is difficult. The word level pathway involves the product of several probabilities, whereas the phrasal path is represented by one probability score. Potential solutions are to introduce translation length models or to learn scaling factors for phrases of different lengths. Results in this paper have been generated by empirically determining a scaling factor that was inversely proportional to the lenth of the phrase, causing each translation to have a score comparable to the product of the word to word translations within the phrase.

7 HMM Phrase Extraction

In order to compare our method to a well understood phrase baseline, we present a method that ex-

$Name$	$Pairs$	$Chinese$	$English$
Small	3540	90K	115K
Large	77558	2.46M	2.69M
Testing	993	27K	NA

Table 1: Corpus figures indicating no. of sentence pairs, no. of Chinese and English words

tracts phrases by harvesting the Viterbi path from an HMM alignment model (Vogel et al., 1996). The HMM alignment model is computationally feasible even for very long sentences, and the phrase extraction method does not have limits on the length of extracted target side phrase. For each source phrase ranging from positions i_1 to i_2 the target phrase is given by $j_{min} = min_i\{j = a(i)\}$ and $j_{max} = max_i\{j = a(i)\}$, where $i = i_1...i_2$ and j refers to an index in the target sentence pair. We calculate phrase translation probabilities (the scores for each extracted phrase) based on a statistical lexicon for the constituent words in the phrase. As the IBM1 alignment model gives the global optimum for the lexical probabilities, this is the natural choice. This leads to the phrase translation probability

$$p(\tilde{s}|\tilde{t}) = \frac{1}{J^I} \prod_i \sum_j p(s_i|t_j) \tag{11}$$

where J and I denotes the length of the target phrase $\tilde{t}$, source phrase $\tilde{s}$, and the word probabilities $p(s_i|t_j)$ are estimated using the IBM1 word alignment model. The phrases extracted from this method can be used directly within our in-house decoder without the significant changes that other phrase based methods could require.

8 Experimentation

IBM alignment models were trained up to model 4 using GIZA (Al Onaizan et al., 1999) from Chinese to English and Chinese to English on two tracks of data. Figures describing the characteristics of each track as well as the test sentences are shown in Table (1). All the data were extracted from a newswire source. We applied our in house segmentation toolkit on the Chinese data and performed basic preprocessing which included; lowercasing, tagging dates, times and numbers on both languages. Translation quality is evaluated by two metrics, (MTEval, 2002) and BLEU (Papeneni et al., 2001), both of which measure n-gram matches between the translated text and the reference translations. NIST is more sensitive to unigram precision due to its emphasis toward high perplexity words. Four reference translations were available for each test sentence. We first compare against a system built using word level lexica only to reiterate the impact of phrase translation, and then show gains by our method over a system that utilizes phrase extracted from the HMM method. The word level system consisted of a hand crafted (Linguistics Data Consortium) bilingual dictionary and a statistical lexicon derived from training IBM model 1. In our experiments we found that although training higher order IBM models does yield lower alignment error rates when measured against manually aligned sentences, the highest translation quality is achieved by using a lexicon extracted from the Model 1 alignment. Experiments were run with a language model (LM) built on a 20 million word news source corpus using our in house decoder which performs a monotone decoding without reordering. To implement our phrase extraction technique, the maximum approximation alignments were combined with the union operation as described in (Och et al., 1999), resulting in a dense but inaccurate alignment map as measured against a human aligned gold standard. Since bi-directional translation models are available, scoring was performed in both directions, using IBM Model 1 lexica for the within sentence scoring. The final phrase level scores computed in each direction were combined by a weighted average before the pruning step. Source side phrases were restricted to be of length 2 or higher since word lexica were available. Weights for each scoring metric were determined empirically against a validation set (alignment map scores were assigned the highest weighting). Table (2) shows results on the small data track, while Table (3) shows results on the large data track. The technique described in this paper is labelled $Phrases$ in the tables. The results show that the phrase extraction method described in this paper contribute to statistically significant improvements over the baseline word and phrase level(HMM) systems. When compared against the HMM phrases, our technique show statistically significant improvements. Statistical significance is evaluated by con-

$Method$	$BLEU$	$NIST$
Baseline-Word	0.135	6.19
Baseline-Word+Phrases	0.167	6.71
Baseline-HMM	0.166	6.49
Baseline-HMM+Phrases	0.174	6.71

Table 2: Small track results

$Method$	$BLEU$	$NIST$
Baseline-Word	0.147	6.62
Baseline-Word+Phrases	0.190	7.48
Baseline-HMM	0.187	7.42
Baseline-HMM+Phrases	0.197	7.60

Table 3: Large track results

sidering deviations in sentence level NIST scores over the 993 sentence test set with a NIST improvement of 0.05 being statistically significant at the 0.01 alpha level. In combination with the HMM method, our technique delivers further gains, providing evidence that different kinds of phrases have been learnt by each method. The improvements caused by our methods is more apparent in the NIST score rather than the BLEU score. We predict that this effect is due to the language specific correction that treats target phrases with function words at the boundaries as the same phrase. This correction cause the burden to be placed on the language model to select the correct phrase instance from several possible translations. Correctly translating function words dramatically boosts the NIST measure as it places emphasis on high perplexity words ie. those with diverse contexts.

9 Conclusions

We have presented a method to efficiently extract phrase relationships from IBM word alignment models by leveraging the maximum approximation as well as the word lexicon. Our method is significantly less computationally expensive than methods that attempt to explicitly model phrase level interactions within alignment models, and recovers well from noisy alignments at the sentence and corpus level. The significant improvements above the baseline carry through when this method is combined with other phrasal and word level methods. Further experimentation is required to fully appreciate the robustness of this technique, especially when considering a comparable, but not parallel, corpus. The language specific scoring methods have a significant impact on translation quality, and further work to extend these methods to represent specific characteristics of each language, promises to deliver further improvements. Although the method performs well, it lacks an explanatory framework through the extraction process; instead it leverages the well understood fundamentals of the traditional IBM models.

Combining phrase level knowledge sources within a decoder in an effective manner is currently our primary research interest, specifically integrating knowledge sources of varying reliability. Our method has shown to be an effective contributing component within the translation framework and we expect to continue to improve the state of the art within machine translation by improving phrasal extraction and integration.

References

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

George Foster 2000. *A Maximum Entropy Minimum Divergence Translation Model*, Proc. of the 38th Annual Meeting of the Association for Computational Linguistics

Daniel Marcu and William Wong 2002. *A Phrase-Based, Joint Probability Model for Statistical Machine Translation*, Proc. of the Conference on Empirical Methods in Natural Language Processing , Philadelphia, PA

NIST 2002. *MT Evaluation Kit Version 9*, www.nist.gov/speech/tests/mt/

Franz Josef Och, Hermann Ney 2002. *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation*, Proc. North American Association for Computational Linguistics

Franz Josef Och and Hermann Ney 200. *A Comparison of Alignment Models for Statistical Machine Translation*, Proc. of the 18th International Conference on Computational Linguistics. Saarbrucken, Germany

Franz Josef Och, Christoph Tillmann, Hermann Ney 1999. *Improved Alignment Models for Statistical Machine Translation*, Proc. of the Joint Conference of

Empirical Methods in Natural Language Processing, p20-28, MD.

Al' Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah H. Smith and David Yarowsky 1999. *Statistical Machine Translation*, Final Report, JHU Summer Workshop

Kishore Papeneni, Salim Roukos, Todd Ward 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*, IBM Research Report, RC22176

Stephan Vogel, Hermann Ney, and Christoph Tillmann 1996. *HMM-based Word Alignment in Statistical Translation*, Proc. of COLING '96: The 16th International Conference on Computational Linguistics, pp. 836-841. Copenhagen, Denmark

Yeyi Wang, Alex Waibel 1998. *Fast Decoding for Statistical Machine Translation*, Proc. of the International Conference in Spoken Language Processing

Dekai Wu 1995. *Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora*, Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), pp. 1328-1335. Montreal

Kenji Yamada and Kevin Knight 2001. *A syntax-based statistical translation model*, Proc. of the 39th Annual Meeting of the Association for Computational Linguistics, France