A Generalized Alignment-Free Phrase Extraction

Bing Zhao
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA-15213
bzhao@cs.cmu.edu

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

Abstract

In this paper, we present a phrase extraction algorithm using a translation lexicon, a fertility model, and a simple distortion model. Except these models, we do not need explicit word alignments for phrase extraction. For each phrase pair (a block), a bilingual lexicon based score is computed to estimate the translation quality between the source and target phrase pairs; a fertility score is computed to estimate how good the lengths are matched between phrase pairs; a center distortion score is computed to estimate the relative position divergence between the phrase pairs. We presented the results and our experience in the shared tasks on French-English.

1 Introduction

Phrase extraction becomes a key component in today's state-of-the-art statistical machine translation systems. With a longer context than unigram, phrase translation models have flexibilities of modelling local word-reordering, and are less sensitive to the errors made from preprocessing steps including word segmentations and tokenization. However, most of the phrase extraction algorithms rely on good word alignments. A widely practiced approach explained in details in (Koehn, 2004), (Och and Ney, 2003) and (Tillmann, 2003) is to get word alignments from two directions: source to target and target to source; the intersection or union operation is applied to get refined word alignment with pre-designed heuristics fixing the unaligned words. With this refined word alignment, the phrase extraction for a given source phrase is essentially to extract the target candidate phrases in the target sentence by searching the left and right projected boundaries.

In (Vogel et al., 2004), they treat phrase alignment as a sentence splitting problem: given a source phrase, find the boundaries of the target phrase such that the overall sentence alignment lexicon probability is optimal. We generalize it in various ways, esp. by using a fertility model to get a better estimation of phrase lengths, and a phrase level distortion model.

In our proposed algorithm, we do not need explicit word alignment for phrase extraction. Thereby it avoids the burden of testing and comparing different heuristics especially for some language specific ones. On the other hand, the algorithm has such flexibilities that one can incorporate word alignment and heuristics in several possible stages within this proposed framework to further improve the quality of phrase pairs. In this way, our proposed algorithm is more generalized than the usual word alignment based phrase extraction algorithms.

The paper is structured as follows: in section 2, The concept of blocks is explained; in section 3, a dynamic programming approach is model the width of the block; in section 4, a simple center distortion of the block; in section 5, the lexicon model; the complete algorithm is in section 6; in section 7, our experience and results using the proposed approach.

2 Blocks

We consider each phrase pair as a block within a given parallel sentence pair, as shown in Figure 1.

The y-axis is the source sentence, indexed word by word from bottom to top; the x-axis is the target sentence, indexed word by word from left to right. The block is defined by the source phrase and its projection. The source phrase is bounded by the *start* and the *end* positions in the source sentence. The projection of the source phrase is defined as the left and right boundaries in the target sentence. Usually, the boundaries can be inferred according to word alignment as the left most and right most aligned positions from the words in the source phrase. In

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 141–144,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

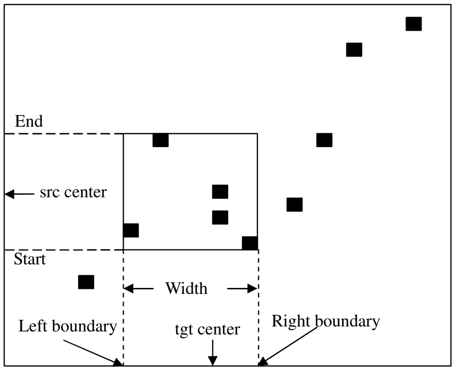

Figure 1: Blocks with "width" and "centers"

this paper, we provide another view of the block, which is defined by the *centers* of source and target phrases, and the *width* of the target phrase.

Phrase extraction algorithms in general search for the left and right projected boundaries of each source phrase according to some score metric computed for the given parallel sentence pairs. We present here three models: a phrase level fertility model score for phrase pairs' length mismatch, a simple center-based distortion model score for the divergence of phrase pairs' relative positions, and a phrase level translation score to approximate the phrase pairs' translational equivalence. Given a source phrase, we can search for the best possible block with the highest combined scores from the three models.

3 Length Model: Dynamic Programming

Given the word fertility definitions in IBM Models (Brown et al., 1993), we can compute a probability to predict *phrase length*: given the candidate target phrase (English) e_1^I, and a source phrase (French) of length J, the model gives the estimation of $P(J|e_1^I)$ via a dynamic programming algorithm using the source word fertilities. Figure 2 shows an example fertility trellis of an English trigram. Each edge between two nodes represents one English word e_i. The arc between two nodes represents one candidate non-zero fertility for e_i. The fertility of zero (i.e. generating a NULL word) corresponds to the direct edge between two nodes, and in this way, the NULL word is naturally incorporated into this model's representation. Each arc is associated with a English word fertility probability $P(\phi_i|e_i)$. A path ϕ_1^I through the trellis represents the number of French words ϕ_i generated by each English word e_i. Thus, the probability of generating J words from the English phrase along the Viterbi path is:

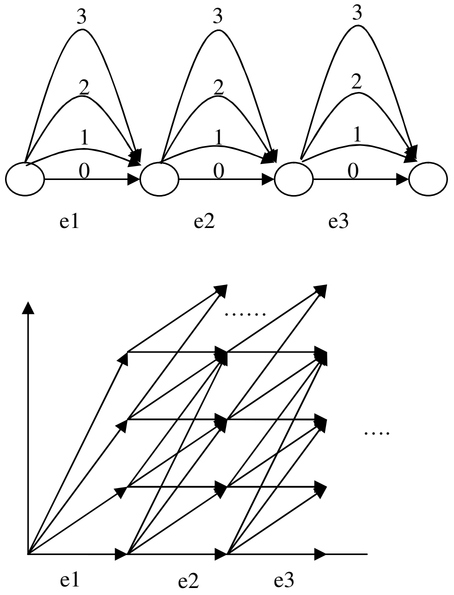

Figure 2: An example of fertility trellis for dynamic programming

$$P(J|e_1^I) = \max_{\{\phi_1^I, J=\sum_{i=1}^I \phi_i\}} \prod_{i=1}^{I} P(\phi_i|e_i) \qquad (1)$$

The Viterbi path is inferred via dynamic programming in the trellis of the lower panel in Figure 2:

$$\phi[j,i] = max \begin{cases} \phi[j,i-1] + \log P_{NULL}(0|e_i) \\ \phi[j-1,i-1] + \log P_{\phi}(1|e_i) \\ \phi[j-2,i-1] + \log P_{\phi}(2|e_i) \\ \phi[j-3,i-1] + \log P_{\phi}(3|e_i) \end{cases}$$

where $P_{NULL}(0|e_i)$ is the probability of generating a NULL word from e_i; $P_{\phi}(k=1|e_i)$ is the usual word fertility model of generating one French word from the word e_i; $\phi[j,i]$ is the cost so far for generating j words from i English words $e_1^i : e_1, \cdots, e_i$.

After computing the cost of $\phi[J,I]$, we can trace back the Viterbi path, along which the probability $P(J|e_1^I)$ of generating J French words from the English phrase e_1^I as shown in Eqn. 1.

With this phrase length model, for every candidate block, we can compute a phrase level fertility score to estimate to how good the phrase pairs are match in their lengthes.

4 Distortion of Centers

The centers of source and target phrases are both illustrated in Figure 1. We compute a simple distortion score to estimate how far away the two centers are in a parallel sentence pair in a sense the block is close to the diagonal.

In our algorithm, the source center $\odot_{f_j^{j+l}}$ of the phrase f_j^{j+l} with length $l+1$ is simply a normalized relative position defined as follows:

$$\odot_{f_j^{j+l}} = \frac{1}{|F|} \sum_{j'=j}^{j'=j+l} \frac{j'}{l+1} \tag{2}$$

where $|F|$ is the French sentence length.

For the center of English phrase e_i^{i+k} in the target sentence, we first define the expected corresponding relative center for every French word $f_{j'}$ using the lexicalized position score as follows:

$$\odot_{e_i^{i+k}}(f_{j'}) = \frac{1}{|E|} \cdot \frac{\sum_{i'=i}^{(i+k)} i' \cdot P(f_{j'}|e_{i'})}{\sum_{i'=i}^{(i+k)} P(f_{j'}|e_{i'})} \tag{3}$$

where $|E|$ is the English sentence length. $P(f_{j'}|e_i)$ is the word translation lexicon estimated in IBM Models. i is the position index, which is weighted by the word level translation probabilities; the term of $\sum_{i=1}^{I} P(f_{j'}|e_i)$ provides a normalization so that the expected center is within the range of target sentence length. The expected center for e_i^{i+k} is simply a average of $\odot_{e_i^{i+k}}(f_{j'})$:

$$\odot_{e_i^{i+k}} = \frac{1}{l+1} \sum_{j'=j}^{j+l} \odot_{e_i^{i+k}}(f_{j'}) \tag{4}$$

This is a general framework, and one can certainly plug in other kinds of score schemes or even word alignments to get better estimations.

Given the estimated centers of $\odot_{f_j^{j+l}}$ and $\odot_{e_i^{i+k}}$, we can compute how close they are by the probability of $P(\odot_{e_i^{i+k}}|\odot_{f_j^{j+l}})$. To estimate $P(\odot_{e_i^{i+k}}|\odot_{f_j^{j+l}})$, one can start with a flat gaussian model to enforce the point of $(\odot_{e_i^{i+k}}, \odot_{f_j^{j+l}})$ not too far off the diagonal and build an initial list of phrase pairs, and then compute the histogram to approximate $P(\odot_{e_i^{i+k}}|\odot_{f_j^{j+l}})$.

5 Lexicon Model

Similar to (Vogel et al., 2004), we compute for each candidate block a score within a given sentence pair using a word level lexicon $P(f|e)$ as follows:

$$\begin{aligned} P(f_j^{j+l}|e_i^{i+k}) &= \prod_{j' \in [j,j+l]} \sum_{i' \in [i,i+k]} \frac{P(f_{j'}|e_{i'})}{k+1} \\ &\cdot \prod_{j' \notin [j,j+l]} \sum_{i' \notin [i,i+k]} \frac{P(f_{j'}|e_{i'})}{|E|-k-1} \end{aligned}$$

6 Algorithm

Our phrase extraction is described in Algorithm 1. The input parameters are essentially from IBM Model-4: the word level lexicon $P(f|e)$, the English word level fertility $P_\phi(\phi_e = k|e)$, and the center based distortion $P(\odot_{e_i^{i+k}}|\odot_{f_j^{j+l}})$.

Overall, for each source phrase f_j^{j+l}, the algorithm first estimates its normalized relative center in the source sentence, its projected relative center in the target sentence. The scores of the phrase length, center-based distortion, and a lexicon based score are computed for each candidate block A local greedy search is carried out for the best scored phrase pair (f_j^{j+l}, e_i^{i+k}).

In our submitted system, we computed the following *seven* base scores for phrase pairs: $P_{ef}(f_j^{j+l}|e_i^{i+k})$, $P_{fe}(e_i^{i+k}|f_j^{j+l})$, sharing similar function form in Eqn. 5.

$$\begin{aligned} P_{ef}(f_j^{j+l}|e_i^{i+k}) &= \prod_{j'} \sum_{i'} P(f_{j'}|e_{i'}) P(e_{i'}|e_i^{i+k}) \\ &= \prod_{j'} \sum_{i'} \frac{P(f_{j'}|e_{i'})}{k+1} \end{aligned} \tag{5}$$

We compute phrase level relative frequency in both directions: $P_{rf}(f_j^{j+l}|e_i^{i+k})$ and $P_{rf}(e_i^{i+k}|f_j^{j+l})$. We compute two other lexicon scores which were also used in (Vogel et al., 2004): $S_1(f_j^{j+l}|e_i^{i+k})$ and $S_2(e_i^{i+k}|f_j^{j+l})$ using the similar function in Eqn. 6:

$$S(f_j^{j+l}|e_i^{i+k}) = \prod_{j'} \sum_{i'} P(f_{j'}|e_{i'}) \tag{6}$$

In addition, we put the *phrase level fertility score* computed in section 3 via dynamic programming to be as one additional score for decoding.

Algorithm 1 A Generalized Alignment-free Phrase Extraction

```
Input: Pre-trained models: $P_{\phi}(\phi_e = k|e)$ , $P(\odot_E|\odot_F)$ , and $P(f|e)$.
Output: PhraseSet: Phrase pair collections.
Loop over the next sentence pair
for $j : 0 \rightarrow |F|-1$,
  for $l : 0 \rightarrow$ MaxLength,
    foreach $f_j^{j+l}$
      compute $\odot_f$ and $\odot_E$
      left = $\odot_E \cdot |E|$-MaxLength,
      right= $\odot_E \cdot |E|$+MaxLength,
    for $i$ : left $\rightarrow$ right,
      for $k : 0 \rightarrow$ right,
        compute $\odot_e$ of $e_i^{i+k}$,
        score the phrase pair $(f_j^{j+l}, e_i^{i+k})$, where
      score = $P(\odot_e|\odot_f)P(l|e_i^{i+k})P(f_j^{j+l}|e_i^{i+k})$
  add top-n $\{(f_j^{j+l}, e_i^{i+k})\}$ into PhraseSet.
```

7 Experimental Results

Our system is based on the IBM Model-4 parameters. We train IBM Model 4 with a scheme of $1^7 2^0 h^7 3^0 4^3$ using GIZA++ (Och and Ney, 2003). The maximum fertility for an English word is 3. All the data is used as given, i.e. we do not have any preprocessing of the English-French data. The word alignment provided in the workshop is not used in our evaluations. The language model is provided by the workshop, and we do not use other language models.

The French phrases up to 8-gram in the development and test sets are extracted with top-3 candidate English phrases. There are in total 2.6 million phrase pairs [1] extracted for both development set and the unseen test set. We did minimal tuning of the parameters in the pharaoh decoder (Koehn, 2004) settings, simply to balance the length penalty for Bleu score. Most of the weights are left as they are given: [ttable-limit]=20, [ttable-threshold]=0.01, [stack]=100, [beam-threshold]=0.01, [distortion-limit]=4, [weight-d]=0.5, [weight-l]=1.0, [weight-w]=-0.5. Table 1 shows the algorithm's performance on several settings for the *seven* basic scores provided in section 6.

settings	Dev.Bleu	Tst.Bleu
s_1	27.44	27.65
s_2	27.62	28.25

Table 1: Pharaoh Decoder Settings

In Table 1, setting s_1 was our submission without using the inverse relative frequency of $P_{rf}(e_i^{i+k}|f_j^{j+l})$. s_2 is using all the seven scores.

8 Discussions

In this paper, we propose a generalized phrase extraction algorithm towards word alignment-free utilizing the fertility model to predict the width of the block, a distortion model to predict how close the centers of source and target phrases are, and a lexicon model for translational equivalence. The algorithm is a general framework, in which one could plug in other scores and word alignment to get better results.

References

P.F. Brown, Stephen A. Della Pietra, Vincent. J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19(2), pages 263–331.

Philip Koehn. 2004. Pharaoh: a beam search decoder for phrase-based smt. In *Proceedings of the Conference of the Association for Machine Translation in the Americans (AMTA).*

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Stephan Vogel, Sanjika Hewavitharana, Muntsin Kolss, and Alex Waibel. 2004. The ISL statistical translation system for spoken language translation. In *Proc. of the International Workshop on Spoken Language Translation*, pages 65–72, Kyoto, Japan.

[1] Our phrase table is to be released to public in this workshop