To appear in Proc. COLING 2008

Phrasal Segmentation Models for Statistical Machine Translation

Graeme Blackwood, Adrià de Gispert, William Byrne
Machine Intelligence Laboratory
Department of Engineering, Cambridge University
Trumpington Street, Cambridge, CB2 1PZ, U.K.
{gwb24|ad465|wjb31}@cam.ac.uk

Abstract

Phrasal segmentation models define a mapping from the words of a sentence to sequences of translatable phrases. We discuss the estimation of these models from large quantities of monolingual training text and describe their realization as weighted finite state transducers for incorporation into phrase-based statistical machine translation systems. Results are reported on the NIST Arabic-English translation tasks showing significant complementary gains in BLEU score with large 5-gram and 6-gram language models.

1 Introduction

In phrase-based statistical machine translation (Koehn et al., 2003) phrases extracted from word-aligned parallel data are the fundamental unit of translation. Each phrase is a sequence of contiguous translatable words and there is no explicit model of syntax or structure.

Our focus is the process by which a string of words is segmented as a sequence of such phrases. Ideally, the segmentation process captures two aspects of natural language. Firstly, segmentations should reflect the underlying grammatical sentence structure. Secondly, common sequences of words should be grouped as phrases in order to preserve context and respect collocations. Although these aspects of translation are not evaluated explicitly, phrases have been found very useful in translation. They have the advantage that, within phrases, words appear as they were found in fluent text. However, reordering of phrases in translation can lead to disfluencies. By defining a distribution over possible segmentations, we hope to address such disfluencies. A strength of our approach is that it exploits abundantly available monolingual corpora that are usually only used for training word language models.

Most prior work on phrase-based statistical language models concerns the problem of identifying useful phrasal units. In (Ries et al., 1996) an iterative algorithm selectively merges pairs of words as phrases with the goal of minimising perplexity. Several criteria including word pair frequencies, unigram and bigram log likelihoods, and a correlation coefficient related to mutual information are compared in (Kuo and Reichl, 1999). The main difference between these approaches and the work described here is that we already have a definition of the phrases of interest (i.e. the phrases of the phrase table extracted from parallel text) and we focus instead on estimating a distribution over the set of possible alternative segmentations of the sentence.

2 Phrasal Segmentation Models

Under the generative model of phrase-based statistical machine translation, a source sentence s_1^I generates sequences $u_1^K = u_1, \ldots, u_K$ of source language phrases that are to be translated. Sentences cannot be segmented into phrases arbitrarily: the space of possible segmentations is constrained by the contents of the phrase table which consists of phrases found with translations in the parallel text. We start initially with a distribution in which segmentations assume the following dependencies:

$$P(u_1^K, K|s_1^I) = P(u_1^K|K, s_1^I)P(K|I). \quad (1)$$

The distribution over the number of phrases K is chosen to be uniform, i.e. $P(K|I) = 1/I$, $K \in \{1, 2, \ldots, I\}$, and all segmentations are considered equally likely. The probability of a particular segmentation is therefore

$$P(u_1^K|K, s_1^I) = \begin{cases} C(K, s_1^I) & \text{if } u_1^K = s_1^I \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $C(K, s_1^I)$ is chosen to ensure normalisation and the phrases $u_1, \ldots, u_K$ are found in the phrase table. This simple model of segmentation has been found useful in practice (Kumar et al., 2006).

Our goal is to improve upon the uniform segmentation of equation (2) by estimating the phrasal segmentation model parameters from naturally occuring phrase sequences in a large monolingual training corpus. An order-n phrasal segmentation model assigns a probability to a phrase sequence u_1^K according to

$$P(u_1^K|K, s_1^I) = \prod_{k=1}^{K} P(u_k|u_1^{k-1}, K, s_1^I) \approx$$
$$\begin{cases} C(K, s_1^I) \prod_{k=1}^{K} P(u_k|u_{k-n+1}^{k-1}) & \text{if } u_1^K = s_1^I \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where the approximation is due to the Markov assumption that only the most recent $n-1$ phrases are useful when predicting the next phrase. Again, each u_k must be a phrase with a known translation. For a fixed sentence s_1^I, the normalisation term $C(K, s_1^I)$ can be calculated. In translation, however, calculating this quantity becomes harder since the s_1^I are not fixed. We therefore ignore the normalisation and use the unnormalised likelihoods as scores.

2.1 Parameter Estimation

We focus on first-order phrasal segmentation models. Although we have experimented with higher-order models we have not yet found them to yield improved translation.

Let $f(u_{k-1}, u_k)$ be the frequency of occurrence of a string of words w_i^j in a very large training corpus that can be split at position x such that $i < x \leq j$ and the substrings w_i^{x-1} and w_x^j match precisely the words of two phrases u_{k-1} and u_k in the phrase table. The maximum likelihood probability estimate for phrase bigrams is then their relative frequency:

$$\hat{P}(u_k|u_{k-1}) = \frac{f(u_{k-1}, u_k)}{f(u_{k-1})}. \quad (4)$$

These maximum likelihood estimates are discounted and smoothed with context-dependent backoff such that

$$P(u_k|u_{k-1}) =$$
$$\begin{cases} \delta(u_{k-1}, u_k)\hat{P}(u_k|u_{k-1}) & \text{if } f(u_{k-1}, u_k) > 0 \\ \alpha(u_{k-1})P(u_k) & \text{otherwise} \end{cases} \quad (5)$$

where $\delta(u_{k-1}, u_k)$ discounts the maximum likelihood estimates and the context-specific backoff weights $\alpha(u_{k-1})$ are chosen to ensure normalisation.

3 The Transducer Translation Model

The Transducer Translation Model (TTM) (Kumar and Byrne, 2005; Kumar et al., 2006) is a generative model of translation that applies a series of transformations specified by conditional probability distributions and encoded as Weighted Finite State Transducers (Mohri et al., 2002).

The generation of a target language sentence t_1^J starts with the generation of a source language sentence s_1^I by the source language model $P_G(s_1^I)$. Next, the source language sentence is segmented according to the uniform phrasal segmentation model distribution $P_W(u_1^K, K|s_1^I)$ of equation (2). The phrase translation and reordering model $P_\Phi(v_1^R|u_1^K)$ generates the reordered sequence of target language phrases v_1^R. Finally, the reordered target language phrases are transformed to word sequences t_1^J under the target segmentation model $P_\Omega(t_1^J|v_1^R)$. These component distributions together form a joint distribution over the source and target language sentences and their possible intermediate phrase sequences as $P(t_1^J, v_1^R, u_1^K, s_1^I)$.

In translation under the generative model, we start with the target sentence t_1^J in the foreign language and search for the best source sentence $\hat{s}_1^I$. Encoding each distribution as a WFST leads to a model of translation as a series of compositions

$$L = G \circ W \circ \Phi \circ \Omega \circ T \quad (6)$$

in which T is an acceptor for the target language sentence and L is the word lattice of translations obtained during decoding. The most likely translation $\hat{s}_1^I$ is the path in L with least cost.

The above approach generates a word lattice L under the unweighted phrasal segmentation model of equation (2). In the initial experiments reported here, we apply the weighted phrasal segmentation model via lattice rescoring. We take the word lattice L and compose it with the unweighted transducer W to obtain a lattice of phrases $L \circ W$; this lattice contains phrase sequences and translation scores consistent with the initial translation. We also extract the complete list of phrases relevant to each translation.

We then wish to apply the phrasal segmentation model distribution of equation (3) to this phrase lattice. The conditional probabilities and backoff structure defined in equation (5) can be encoded as a weighted finite state acceptor (Allauzen et al., 2003). In this acceptor, Ψ, states encode histories and arcs define the bigram and backed-off unigram phrase probabilities. We note that the raw counts of equation (4) are collected prior to translation and the first-order probabilities are estimated only for phrases found in the lattice.

The phrasal segmentation model is composed with the phrase lattice and projected on the input to obtain the rescored word lattice $L' = (L \circ W) \circ \Psi$. The most likely translation after applying the phrasal segmentation model is found as the path in L' with least cost. Apart from likelihood pruning when generating the original word lattice, the model scores are included correctly in translation search.

4 System Development

We describe experiments on the NIST Arabic-English machine translation task and apply phrasal segmentation models in lattice rescoring.

The development set *mt02-05-tune* is formed from the odd numbered sentences of the NIST MT02–MT05 evaluation sets; the even numbered sentences form the validation set *mt02-05-test*. Test performance is evaluated using the NIST subsets from the MT06 evaluation: *mt06-nist-nw* for newswire data and *mt06-nist-ng* for newsgroup data. Results are also reported for the MT08 evaluation. Each set contains four references and BLEU scores are computed for lower-case translations.

The uniformly segmented TTM baseline system is trained using all of the available Arabic-English data for the NIST MT08 evaluation[1]. In first-pass translation, decoding proceeds with a 4-gram language model estimated over the parallel text and a 965 million word subset of monolingual data from the English Gigaword Third Edition. Minimum error training (Och, 2003) under BLEU optimises the decoder feature weights using the development set *mt02-05-tune*. In the second pass, 5-gram and 6-gram zero-cutoff stupid-backoff (Brants et al., 2007) language models estimated using 4.7 billion words of English newswire text are used to generate lattices for phrasal segmentation model rescoring. The phrasal segmentation model parameters are trained using a 1.8 billion word subset of the same monolingual training data used to build the second-pass word language model. A phrasal segmentation model scale factor and phrase insertion penalty are tuned using the development set.

[1] http://www.nist.gov/speech/tests/mt/2008/

	mt02-05-tune	mt02-05-test
TTM+MET	48.9	48.6
+6g	51.9	51.7
+6g+PSM	52.7	52.7

Table 2: BLEU scores for phrasal segmentation model rescoring of 6-gram rescored lattices.

5 Results and Analysis

First-pass TTM translation lattices generated with a uniform segmentation obtain baseline BLEU scores of 48.9 for *mt02-05-tune* and 48.6 for *mt02-05-test*. In our experiments we demonstrate that phrasal segmentation models continue to improve translation even for second-pass lattices rescored with very large zero-cutoff higher-order language models. Table 1 shows phrasal segmentation model rescoring of 5-gram lattices. The phrasal segmentation models consistently improve the BLEU score: +1.1 for both the development and validation sets, and +1.4 and +0.4 for the in-domain newswire and out-of-domain newsgroup test sets. Rescoring MT08 gives gains of +0.9 on *mt08-nist-nw* and +0.3 on *mt08-nist-ng*.

For a limited quantity of training data it is not always possible to improve translation quality simply by increasing the order of the language model. Comparing tables 1 and 2 shows that the gains in moving from a 5-gram to a 6-gram are small. Even setting aside the practical difficulty of estimating and applying such higher-order language models, it is doubtful that further gains could be had simply by increasing the order. That the phrasal segmentation models continue to improve upon the 6-gram lattice scores suggests they capture more than just a longer context and that they are complementary to word-based language models.

The role of the phrase insertion penalty is to encourage longer phrases in translation. Table 3 shows the effect of tuning this parameter. The upper part of the table shows the BLEU score, brevity penalty and individual n-gram precisions. The lower part shows the total number of words in the output, the number of words translated as a phrase of the specified length, and the average number of words per phrase. When the insertion

	mt02-05-tune	mt02-05-test	mt06-nist-nw	mt06-nist-ng	mt08-nist-nw	mt08-nist-ng
TTM+MET	48.9	48.6	46.1	35.2	48.4	33.7
+5g	51.5	51.5	48.4	36.7	49.1	36.4
+5g+PSM	52.6	52.6	49.8	37.1	50.0	36.7

Table 1: BLEU scores for phrasal segmentation model rescoring of 5-gram rescored lattices.

PIP	-4.0	-2.0	0.0	2.0	4.0
BLEU	48.6	50.1	51.1	49.9	48.7
BP	0.000	0.000	0.000	-0.034	-0.072
1g	82.0	83.7	84.9	85.7	86.2
2g	57.3	58.9	59.9	60.5	61.1
3g	40.8	42.2	43.1	43.6	44.2
4g	29.1	30.3	31.1	31.5	32.0
words	70550	66964	63505	60847	58676
1	58840	46936	25040	15439	11744
2	7606	12388	18890	19978	18886
3	2691	4890	11532	13920	14295
4	860	1820	5016	6940	8008
5	240	450	1820	2860	3500
6+	313	480	1207	1710	2243
w/p	1.10	1.21	1.58	1.86	2.02

Table 3: Effect of phrase insertion penalty (PIP) on BLEU score, brevity penalty (BP), individual n-gram precisions, phrase length distribution, and average words per phrase (w/p) for *mt02-05-tune*.

penalty is too low, single word phrases dominate the output and any benefits from longer context or phrase-internal fluency are lost. As the phrase insertion penalty increases, there are large gains in precision at each order and many longer phrases appear in the output. At the optimal phrase insertion penalty, the average phrase length is 1.58 words and over 60% of the translation output is generated from multi-word phrases.

6 Discussion

We have defined a simple model of the phrasal segmentation process for phrase-based SMT and estimated the model parameters from naturally occurring phrase sequence examples in a large training corpus. Applying first-order models to the NIST Arabic-English machine translation task, we have demonstrated complementary improved translation quality through exploitation of the same abundantly available monolingual data used for training regular word-based language models.

Comparing the in-domain newswire and out-of-domain newsgroup test set performance shows the importance of choosing appropriate data for training the phrasal segmentation model parameters. When in-domain data is of limited availability, count mixing or other adaptation strategies may lead to improved performance.

Acknowledgements

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Allauzen, Cyril, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 557–564.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, pages 858–867.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Kumar, Shankar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of the conference on HLT and EMNLP*, pages 161–168.

Kumar, Shankar, Yonggang Deng, and William Byrne. 2006. A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering*, 12(1):35–75.

Kuo, Hong-Kwang Jeff and Wolfgang Reichl. 1999. Phrase-based language models for speech recognition. In *Sixth European Conference on Speech Communication and Technology*, pages 1595–1598.

Mohri, Mehryar, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. In *Computer Speech and Language*, volume 16, pages 69–88.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA.

Ries, Klaus, Finn Dag Bu, and Alex Waibel. 1996. Class phrase models for language modeling. In *Proceedings of the 4th International Conference on Spoken Language Processing*.