Phrasal Segmentation Models for Statistical Machine Translation

Graeme Blackwood, Adrià de Gispert, and William Byrne

Machine Intelligence Laboratory, Cambridge University Engineering Department

INTRODUCTION

- In phrase-based SMT phrases are the fundamental unit of translation
- Each phrase is a string of contiguous words found in parallel data
- Advantage is that words within phrases are as found in fluent text
- BUT: reordering of phrases in translation leads to disfluencies

Phrasal segmentation models address reordering disfluencies
- Define a mapping from word strings to translatable phrase sequences
- Estimate a distribution over possible segmentations of monolingual data
- Exploitation of monolingual data usually only for word-based LMs
- Complementary gains even with large 5-gram and 6-gram word-based LMs

- Segmentations ideally capture two aspects of natural language:
  - Reflect the underlying grammatical sentence structure
  - Group words to preserve context and collocations
- We do not address the problem of identifying useful phrasal units
- Already defined by phrase table extracted from parallel data

PHRASAL SEGMENTATION MODELS

- Generative model: words $x_i$ generate phrase sequences $a^K_i$ of length $K$
- Source word string cannot be segmented arbitrarily
- Segmentation constrained by contents of the phrase table
- Assume distribution has the following dependencies:

$$P(a^K, K|x_i) = P(a^K|K, x_i) P(K|x_i)$$

- First-order phrasal segmentation model:
  - Estimate parameters from phrases observed in monolingual corpus
  - Order-$n$ PSM assigns probability to phrase sequence $a^n_i$ according to:

$$P(a^n, K|x_i) = \prod_{k=1}^n P(a^n_k|K, a^{k-1}_i)$$

  - Normalisation $C(K, x_i)$ can be computed for fixed $x_i$
  - In translation $x_i$ vary and computing the normalisation is harder
  - Ignore normalisation and use unnormalised likelihoods as scores

- Phrasal segmentation model parameter estimation
  - Phrase pair bigram probabilities computed by relative frequency

$$\hat{P}(a^n_k|a^{k-1}_i) = \frac{f(a^n_k|a^{k-1}_i)}{f(a^{k-1}_i)}$$

- ML estimates smoothed with context-dependent back-off:

$$P(a^n_k|a^{k-1}_i) = \frac{\hat{P}(a^n_k|a^{k-1}_i) + \alpha f(a^{k-1}_i)}{\hat{P}(a^{k-1}_i) + \alpha}$$

LATTICE RESCORING WITH PSMs

- Noisy channel model of statistical machine translation

$$a_i' = \arg\max_{a_i} P(x_i'|a_i) = \arg\max_{a_i} P(x_i'|a_i') P(x_i')$$

- TTM phrasal model based with Weighted Finite State Transducers

$$P(a_i'W) = \max_{a_i} P(a_i') P(x_i')$$

- Source language model acceptor
- $W$: unweighted source phrase segmentation transducer
- $\Omega$: translation and reordering transducer
- Search for most likely translation under joint distribution:

$$a_i = \arg\max_{a_i} P(a_i') P(x_i')$$

- Translation decoding and lattice generation

- $L = G \circ W \circ \Phi \circ \Omega \circ T$

- Most likely translation $a_i'$ is path in lattice $L$ with least cost

PSM Rescoring of $2^{nd}$ pass 5-gram lattices

- Phrase insertion penalty (PIP), BLEU translation score, brevity penalty (BP),
  - Parameters estimated using 1.8 billion word subset of LM data

- Phrase lengths distribution analysis
  - Phrase insertion penalty (PIP) adds fixed cost to each phrase arc
  - Role is to encourage longer phrases in translation
  - Single-word phrases dominate when the PIP is too low
  - Advantage of phrase-internal fluency and longer context lost
  - 1.58 words/phrase and > 60% multi-word phrases at optimal PIP

SYSTEM DEVELOPMENT

- NIST Arabic-English machine translation task
  - Four reference translations for each set
  - NIST BLEU scores reported for lower-case translations
  - mt02-05-tune: odd numbered sentences from MT02 – MT05
  - mt02-05-test: even numbered sentences from MT02 – MT05

- Baseline system configuration
  - OpenFST TTM baseline with uniform segmentation distribution
  - Decoder feature weights optimised using MET under BLEU
  - 1st pass translation decoding with KN 4-gram LM
  - 1st pass LM trained on parallel text + 965m words of GigaWord v3
  - 2nd pass rescoring with large zero-cutoff 5-gram and 6-gram LMs
  - 2nd pass LM trained on 4.7 billion words of mostly newswire data

- Phrasal segmentation models training
  - PSMs applied in 2nd pass lattice rescoring
  - Parameters estimated using 1.8 billion word subset of LM data
  - PSM scale factor and insertion penalty tuned using mt02-05-tune

RESULTS AND ANALYSIS

- Phrasal Rescoring of $2^{nd}$ pass 5-gram lattices
  - Gains of +1.1 BLEU on mt02-05-tune and mt02-05-test
  - Newsrew test performance also good: +0.9 on mt08-nist-nw
  - Less effective for out-of-domain mt08-nist-ng data (data mismatch)

- Phrasal Rescoring of $2^{nd}$ pass 6-gram lattices
  - 6-gram provides only small gains of +0.4 and +0.2 over 5-gram
  - 6-gram vs. 5-gram suggests further gains from increased LM order unlikely
  - Larger gains of PSM show more than just a longer context is captured

Phrase insertion penalty (PIP), BLEU translation score, brevity penalty (BP), number of words translated as part of a phrase of the specified lengths 1-6+, and average number of words per phrase for mt02-05-tune.