

# 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 used 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  $s_1^j$  generate phrase sequences  $u_1^K$  of length  $K$ 
  - Source word string cannot be segmented arbitrarily
  - Segmentation constrained by contents of the phrase table
  - Assume distribution has the following dependencies:

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

- First-order phrasal segmentation model:
  - Estimate parameters from phrases observed in monolingual corpus
  - Order- $n$  PSM assigns probability to phrase sequence  $u_1^K$  according to:

$$P(u_1^K | K, s_1^j) = \prod_{k=1}^K P(u_k | u_{k-1}^{k-1}, K, s_1^j) \approx \begin{cases} C(K, s_1^j) \prod_{k=1}^K P(u_k | u_{k-n+1}^{k-1}) & \text{if } u_1^K = s_1^j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- Normalisation  $C(K, s_1^j)$  can be computed for fixed  $s_1^j$
- In translation  $s_1^j$  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

$$\hat{P}(u_k | u_{k-1}) = \frac{f(u_{k-1}, u_k)}{f(u_{k-1})} \quad (3)$$

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

$$P(u_k | u_{k-1}) = \begin{cases} \delta(u_{k-1}, u_k) \hat{P}(u_k | u_{k-1}) & \text{if } f(u_{k-1}, u_k) > 0 \\ \alpha(u_{k-1}) P(u_k) & \text{otherwise} \end{cases} \quad (4)$$

## LATTICE RESCORING WITH PSMs

- Noisy channel model of statistical machine translation

$$\hat{s}_i^j = \operatorname{argmax}_{s_1^j} P(s_1^j | t_1^j) = \operatorname{argmax}_{s_1^j} P(t_1^j | s_1^j) P(s_1^j) \quad (5)$$

- TTM phrase-based translation with Weighted Finite State Transducers

$$\begin{array}{ccccccc} t_1^j & \leftarrow & v_1^R & \leftarrow & u_1^K & \leftarrow & s_1^j \\ & & P_\Omega(t_1^j | v_1^R) & & P_\Phi(v_1^R | u_1^K) & & P_W(u_1^K | s_1^j) & & P_G(s_1^j) \\ & & \Omega & & \Phi & & W & & G \end{array}$$

- $G$ : source language model acceptor
- $W$ : unweighted source phrasal segmentation transducer
- $\Phi$ : translation and reordering transducer
- $\Omega$ : unweighted target phrasal segmentation transducer

- Translation decoding and lattice generation

- Search for most likely translation under joint distribution:

$$\hat{s}_i^j = \operatorname{argmax}_{s_1^j, u_1^K, v_1^R} P(t_1^j, v_1^R, u_1^K, s_1^j) \quad (6)$$

- Translation decoding and lattice generation with WFSTs:

$$L = G \circ W \circ \Phi \circ \Omega \circ T \quad (7)$$

- Most likely translation  $\hat{s}_i^j$  is path in lattice  $L$  with least cost

- Phrasal segmentation model lattice rescoring

- Compose 1<sup>st</sup> pass word lattice with  $W$  to get phrase lattice
- PSM parameters of equation (4) encoded as WFST  $\Psi$
- PSM lattice rescoring applied through composition:  $L' = (L \circ W) \circ \Psi$

## 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
- 1<sup>st</sup> pass translation decoding with KN 4-gram LM
- 1<sup>st</sup> pass LM trained on parallel text + 965m words of GigaWord v3
- 2<sup>nd</sup> pass rescoring with large zero-cutoff 5-gram and 6-gram LMs
- 2<sup>nd</sup> pass LM trained on 4.7 billion words of mostly newswire data

- Phrasal segmentation models training

- PSMs applied in 2<sup>nd</sup> 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

- PSM Rescoring of 2<sup>nd</sup> pass 5-gram lattices

- Gains of +1.1 BLEU on mt02-05-tune and mt02-05-test
- Newswire test performance also good: +0.9 on mt08-nist-nw
- Less effective for out-of-domain mt08-nist-ng data (data mismatch)

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

- PSM Rescoring of 2<sup>nd</sup> 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

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

- 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

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
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

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.