Large-Scale Statistical Machine Translation with Weighted Finite State Transducers

Graeme Blackwood, **Adrià de Gispert**, Jamie Brunning, William Byrne

Department of Engineering
University of Cambridge

7th Int. Workshop on Finite-State Methods and Natural Language Processing

Ispra, Lago Maggiore, (Italy)

11-12 Sep 2008

Source-channel model of Statistical Machine Translation

- Task is to recover the source sentence $\mathbf{S} = s_1, s_2, \ldots, s_I$ that generated the observed target $\mathbf{T} = t_1, t_2, \ldots, t_J$
- Typically decomposed into translation and language model probabilities:

$$\widehat{\mathbf{S}} = \underset{\mathbf{S}}{\operatorname{argmax}}\, P(\mathbf{S}|\mathbf{T}) = \underset{\mathbf{S}}{\operatorname{argmax}}\, P(\mathbf{T}|\mathbf{S})P(\mathbf{S})$$

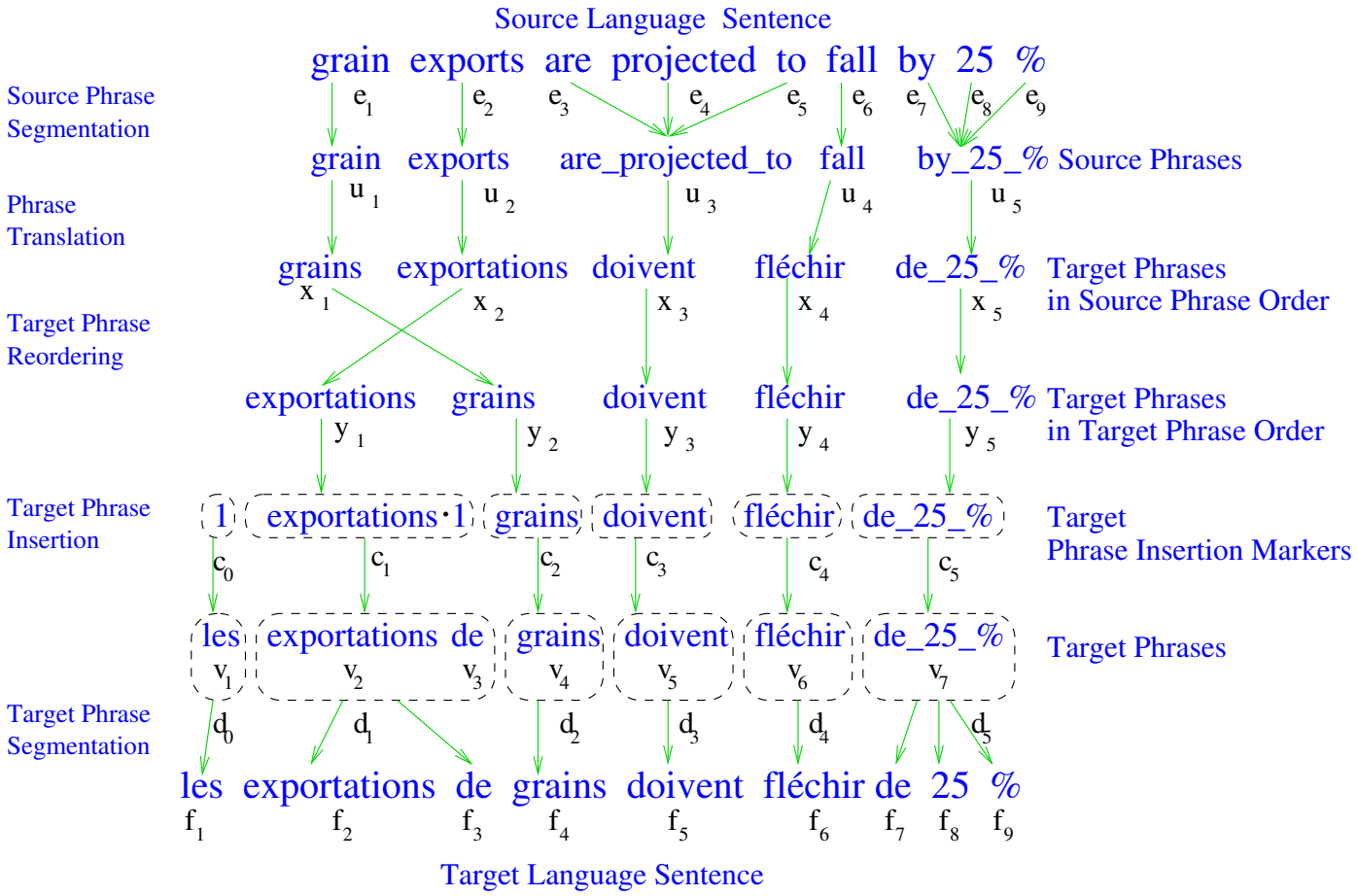
- These probability distributions are estimated from (word-aligned) parallel and monolingual corpora
- Minimal translation unit is the phrase

Transducer Translation Model (TTM)

- The Cambridge University Engineering Department phrase-based SMT system follows the Transducer Translation Model (TTM)
- Generative model of translation
- Implemented with Weighted Finite State Transducers (WFST)
  - WFSTs used for word alignment, language model, word-to-phrase segmentation, phrase translation and reordering
  - Translation is performed using libraries of standard FST operations
  - No special-purpose decoder required
  - Modularity. Easy to work on translation components in isolation
  - Open Source WFST Toolkit [1] – www.openfst.org/
- Incorporates various second-pass lattice rescoring stages

[1] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut , and M. Mohri (2007), OpenFst: A General and Efficient Weighted Finite-State Transducer Library. CIAA.

Transducer Translation Model (TTM)



- Transformations via stochastic models implemented as WFSTs
- Built with standard WFST operations such as composition and best-path search

TTM Component Models

Basic models:

- Source first-pass language model G
- Source phrase segmentation (unweighted) W
- Phrase translation and reordering R
- Target phrase insertion Φ
- Target phrase segmentation (unweighted) Ω
- Word penalty and phrase penalty

Decoding: A translation lattice is obtained through the series of compositions:

$$L = G \circ W \circ R \circ \Phi \circ \Omega \circ \mathbf{T}$$

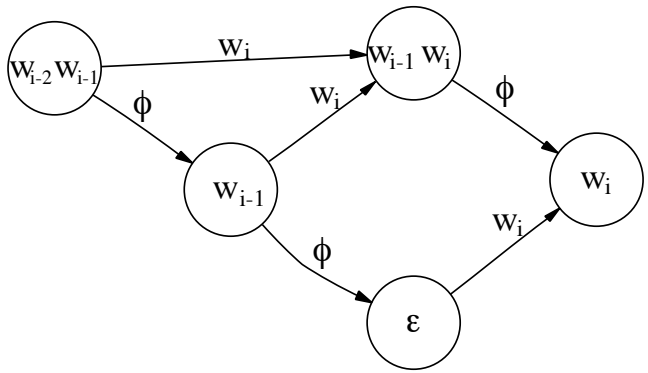
where $\mathbf{T}$ is the target sentence to translate.

$\Rightarrow$ the most likely translation $\widehat{\mathbf{S}}$ is the path in L with least cost (i.e. minimum negative log-likelihood in tropical semiring)

First-pass Language Model (word *n*-Gram)

Backoff *n*-gram approximation : $P(s_1^l) \approx \prod_i P(s_i|s_{i-n+1}^{i-1})$

$$P(s_i|s_{i-n+1}^{i-1}) = \begin{cases} \rho(s_{i-k+1}^{i}) & \text{if } c(s_{i-k+1}^{i}) > \tau \\ \lambda(s_{i-k+1}^{i-1})\, P(s_i|s_{i-n+2}^{i-1}) & \text{otherwise} \end{cases}$$



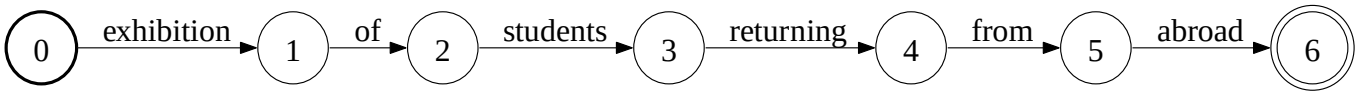
WFSA Trigram [2]

- each probability and back-off weight is encoded as a cost on an arc in the grammar WFST
- ρ and λ can be pre-computed and stored for reasonable sized language models
- WFST implements backoff *n*-gram exactly (ϕ is a failure arc) or approximately

For 'reasonable' sized LM training sets, WFST implementations work well

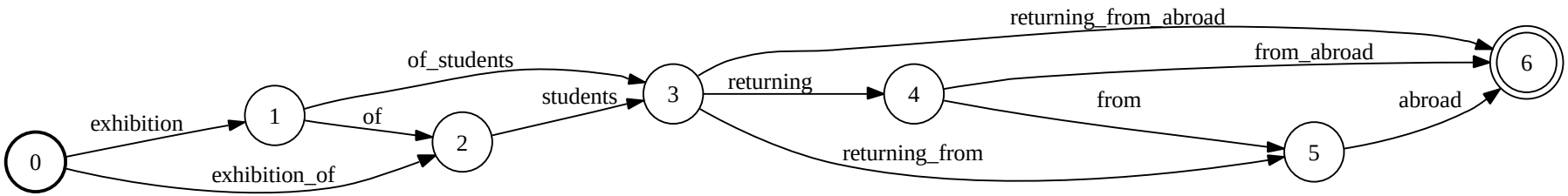
[2] C. Allauzen et al. 2003. Generalized Algorithms for Constructing Statistical Language Models. Proc. ACL

Phrase Segmentation Transducers



Sentence Acceptor

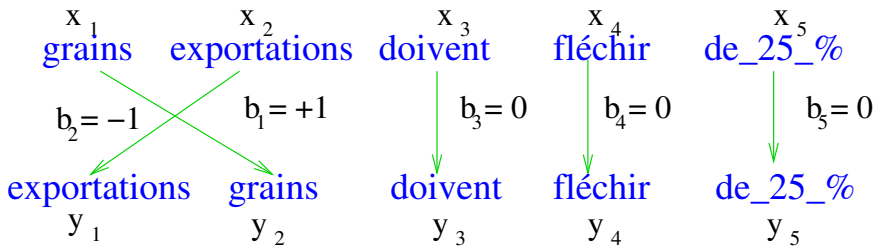
⇓ Phrase Segmentation Transducer W or Ω



Phrase Sequence Lattice

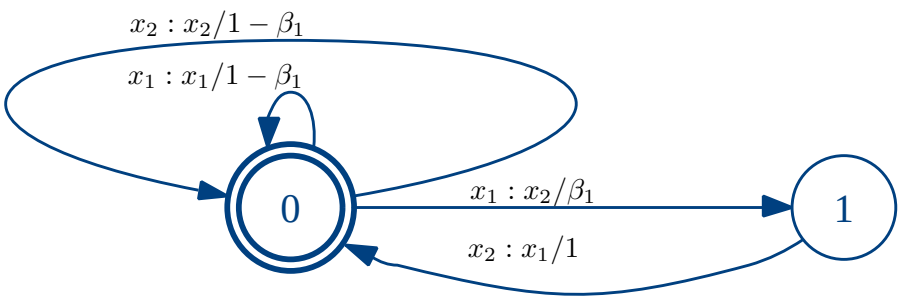
- Phrase Segmentation Transducers convert word sequences (or lattices) into phrase lattices according to Phrase Pair Inventory
- lattice is unweighted ⇒ all segmentations equally likely in first-pass decoding

Phrase Swapping by WFSTs [3]



Associate a jump sequence b_1^K with each sequence y_1^K

$$P(b_1^K|x_1^K, u_1^K, K, s_1^I) = \prod_{k=1}^{K} \underbrace{P(b_k|b_{k-1}, x_{k-1}, x_k, u_{k-1}, u_k)}_{\text{orientation prob., estimated from alignments}}$$



b_k specify relative offsets

MJ-1 : maximum jump of 1

$$b \in \{0, +1, -1\}$$

Extremely simple, but

→ Properly parameterized

→ Not degenerate

[3] Kumar , Byrne 2005. Local phrase reordering models for statistical machine translation. HLT-EMNLP.

Grammar constraints as LM acceptor

- often certain input word sequences are to be passed through the translation system intact

  此外，大约 三十 个 摊位 也 以 各类 行动 电视 手机 如 t-dmb (terrestrial digital media broadcasting)，s-dmb (satellite digital multimedia broadcasting) 及 dvb-h (digital video broadcasting-handhelds)，提供 杜林 冬 运 现场 实况 转播 的 画面，藉 以 吸引 参观 者 注意 .

- Separate translation of Foreign-language sequences is not ideal, as it prevents long-span translation, reordering and language models from looking accross boundaries

- Solution: Compose source language model with an additional constrained grammar
  - $G' = G \circ C$, where C accepts sequences $V^* \cdot u_1 \cdot V^* \cdot u_2 \cdot V^*$
    (V is the source language vocabulary)

- Useful to impose constraints on output and keep scores based on long-span models
  - parentheses or quotes properly matched
  - names correctly transliterated

Minimum Error Training (MET)

- MET is used to adjust the relevance of each component transducer to the translation metric (BLEU, TER, etc.)
- Effective in tuning systems to task-specific situations

Additional models for MET:

- Inverse phrase translation
- 3 phrase pair count features [4]

$\Rightarrow$ Total of 10 scaling factors are optimized on development set (typ. 3-4 iterations)

- weights are assigned to WFST likelihoods for test translation

[4] O. Bender, E. Matusov, S. Hahn, S. Hasan, S. Khadivi, and H. Ney (2007). The RWTH Arabic-to-English Spoken Language Translation System. ASRU.

Large Language Model Rescoring

Stupid backoff zero cut-off 5-gram language model [5]

Directly build sentence-specific LMs:

- Counts are extracted beforehand from all monolingual English data
- 5-grams are extracted from first-pass lattices
- All observed n-grams are kept and backoff weight α is fixed for all orders:

$$S(s_i|s_{i-n+1}^{i-1}) = \begin{cases} \frac{\#(s_{i-n+1}^{i})}{\#(s_{i-n+1}^{i-1})} & \text{if } \#(s_{i-n+1}^{i}) > 0 \\ \alpha\ S(s_i|s_{i-n+2}^{i-1}) & \text{otherwise} \end{cases}$$

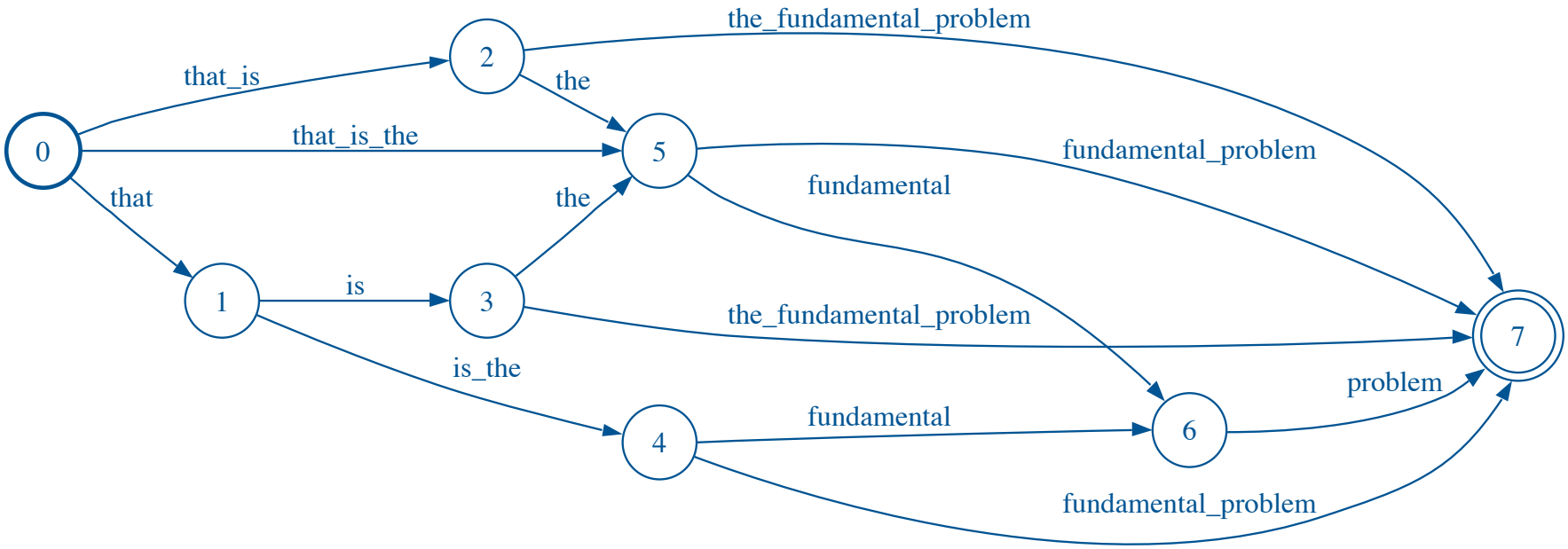
- equal weight interpolation with first-pass 4-gram LM
- exact search with OpenFST libraries in second-pass lattice rescoring

[5] T. Brants et al. 2007. Large Language Models in Machine Translation. EMNLP

Phrasal Segmentation Model Rescoring

Assign probability to sequences of English phrases

- phrases are translatable word sequences
- complement word-based N-grams



Source phrase segmentation transducer assigns 'bigram' probabilities:

$$P(u_1^K | s_1^I) = \prod_k P(u_k \mid u_{k-1}, s_1^K)$$

Model-1 Lattice-to-String Alignment Scores

IBM Model-1 is the simplest model of word alignment used in text alignment:

$$P_{M1}(t_1^J, a_1^J, J|s_1^I) = P_L(J|I)\frac{1}{I^J}\prod_{j=1}^{J} p_T(t_j|s_{a_j})$$

It is not powerful enough to guide translation, but is useful to rank competing translation hypotheses

- Finding the max. likelihood alignment is straighforward via dynamic programming
- For a fixed target sentence **T** and a translation lattice L, we can simultaneously find the best alignment of every lattice path to the target string
- This is done with non-WFST based lattice-to-string alignment procedures to keep compact lattice representation intact

Minimum Bayes Risk Decoding [6]

Taking the goal as BLEU maximization

- A baseline translation model to give the probabilities over translations: $P(\mathbf{S}|\mathbf{T})$
- A set $\mathcal{N}$ of N-Best Translations of T
- A Loss function $L(\mathbf{S}', \mathbf{S})$ that measures the the quality of $\mathbf{S}'$ relative to $\mathbf{S}$

MBR Decoder

$$\widehat{\mathbf{S}} = \underset{\mathbf{S} \in \mathcal{N}}{\operatorname{argmin}} \sum_{\mathbf{S}' \in \mathcal{N}} -\text{BLEU}(\mathbf{S}', \mathbf{S}) P(\mathbf{S}'|\mathbf{T})$$

$\widehat{\mathbf{S}}$ is sometimes called the 'consensus hypothesis'

- picks from the middle of the similar, relatively likely translation hypotheses
- must be done over an N-Best list

Rationale is to balance estimation criteria (e.g. MLE) with translation criteria (e.g. BLEU)

[6] S. Kumar W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. HLT-NAACL

Task Description

NIST 2008 Arabic-English MT task [7]

- development set (*mt02-05-tune*): odd-numbered sentences of the NIST MT02 through MT05 evaluation sets
- validation set (*mt02-05-test*): even-numbered sentences of the NIST MT02 through MT05 evaluation sets
- test sets: MT06 (with newswire and newsgroup subsets) and MT08

TTM system trained on all available Arabic-English data for NIST MT08

- $\sim$ 6M sentences, $\sim$ 150M words
- word-aligned using MTTK

English Language Model data includes:

- first-pass 4-gram: parallel corpus + subset from English GigaWord Third Edition ($\sim$ 965M words)
- second-pass zero-cutoff 5-gram: $\sim$ 4.7B words (newswire text)
- phrasal segmentation model: $\sim$ 1.8B word subset of the above text

[7] nist.gov/speech/tests/mt/2006 nist.gov/speech/tests/mt/2008

Results (1)

MET and successive lattice rescoring steps

- Reordering model is MJ1 with orientation probabilities

Lowercase BLEU/TER scores over five test sets from 2002 through 2008:

Method	mt02-05-tune	mt02-05-test	mt06-nist-nw	mt06-nist-ng	mt08-nist
TTM+MET	50.9 / 42.8	50.3 / 43.3	48.1 / 44.3	37.5 / 53.5	43.1 / 49.5
+5g	53.5 / 41.8	52.4 / 42.4	49.6 / 43.9	39.0 / 54.0	43.7 / 49.3
+PSM	53.9 / 42.1	53.3 / 42.7	50.1 / 44.3	39.0 / 54.7	44.3 / 49.3
+MBR	54.0 / 41.7	53.7 / 42.2	51.0 / 43.9	39.4 / 54.1	45.0 / 48.9

- Important gains from lattice rescoring (improved fluency)
- Phrasal segmentation model complements 5-gram rescoring with further (yet smaller) gains
- Minimum Bayes Risk on the 1000-best list produces consistent gains
- Ranks among the group of top single-system entries in NIST 2008 official results

Results (2)

Impact of orientation probabilities in reordering model

- Comparison with vs without orientation probs. (flat β_1 distribution)

Lowercase BLEU/TER scores over development and validation sets:

Method	with orientation probs.		without orientation probs.	
	mt02-05-tune	mt02-05-test	mt02-05-tune	mt02-05-test
TTM+MET	50.9 / 42.8	50.3 / 43.3	50.4 / 43.3	50.0 / 43.8
+5g	53.5 / 41.8	52.4 / 42.4	53.0 / 42.2	52.2 / 42.8
+PSM	53.9 / 42.1	53.3 / 42.7	53.4 / 42.5	53.1 / 43.1

- Degradation of around 0.4 BLEU in MET results, sustained after rescoring
- More informed phrase reordering model produces better MET lattice
- Further improvements in reordering model expected to benefit even more from large rescoring

Results (3)

Impact of phrase pair count features included in MET

- Comparison with vs without phrase pair counts feats. (10 vs 7 MET parameters)

Lowercase BLEU/TER scores over development and validation sets:

Method	with phrase pair cnt.		without phrase pair cnt.	
	mt02-05-tune	mt02-05-test	mt02-05-tune	mt02-05-test
TTM+MET	50.9 / 42.8	50.3 / 43.3	48.9 / 43.8	48.6 / 44.1
+5g	53.5 / 41.8	52.4 / 42.4	51.5 / 42.2	51.5 / 42.7
+PSM	53.9 / 42.1	53.3 / 42.7	52.6 / 42.3	52.6 / 42.7

- Significant contribution of phrase pair count features
- Simple features produce a very useful phrase selection effect
- Better quality MET lattice $\Rightarrow$ better quality after rescoring

Results (4)

Impact of Model-1 rescoring and order of rescoring steps

- Comparison of two different rescoring orders (without phrase pair counts feats.)

Lowercase BLEU/TER scores over development and validation sets:

Method	mt02-05-tune	mt02-05-test	Method	mt02-05-tune	mt02-05-test
TTM+MET	48.9 / 43.8	48.6 / 44.1	TTM+MET	48.9 / 43.8	48.6 / 44.1
+MOD1	50.5 / 42.5	50.4 / 43.0	+5g	51.5 / 42.2	51.5 / 42.7
+5g	52.2 / 41.6	52.1 / 42.3	+PSM	52.6 / 42.3	52.6 / 42.7
+PSM	52.9 / 41.9	52.6 / 42.6	+MOD1	53.0 / 41.8	52.6 / 42.5

- Model-1 is beneficial when directly rescoring MET lattice
- No real improvement if applying it after 5-gram and phrasal segmentation

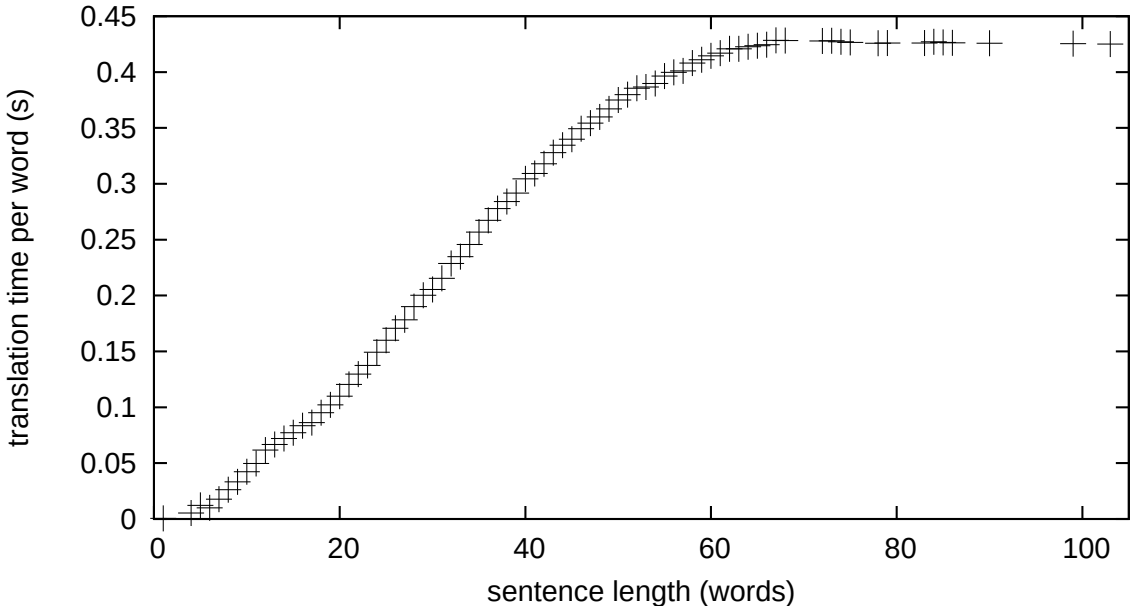
$\Rightarrow$ each of the rescoring techniques is a useful source of information

$\Rightarrow$ applying all techniques to the same MET lattice does not always provide gains

$\Rightarrow$ integration in MET prior to generating lattice would be helpful

Efficiency Considerations

- Efficiency/Parallelization via WFSTs with sentence-specific model parameters
- Memory requirements under ∼4Gb for most sentences
- Pruning for longest sentences (∼100 words) prior to LM composition is required
  - rather than via standard cost-based pruning, via minimum number of phrases
  - affects around 2% of the sentences
- *mt02-05-tune* (∼60k words) translated in 420 minutes (12s/sent on average, parallelizable)
- Translation time per word:



Conclusion and future work

Summary:

- CUED SMT system formulates translation as a series of transformations encoded in WFSTs
- Decodes using standard FST operations
- Able to efficiently handle very large quantities of data
- Good performance on 2008 NIST MT Arabic-English task
- Easy to apply to translation of ASR lattices

Future Work

- Integration of phrasal segmentation model and Model-1 into MET
- Focus on reordering model (current MJ1 model does not allow long-range reordering)

Thanks!
Questions and comments welcome.

Department of Engineering
University of Cambridge