Recent Improvements in the CMU Large Scale Chinese-English SMT System

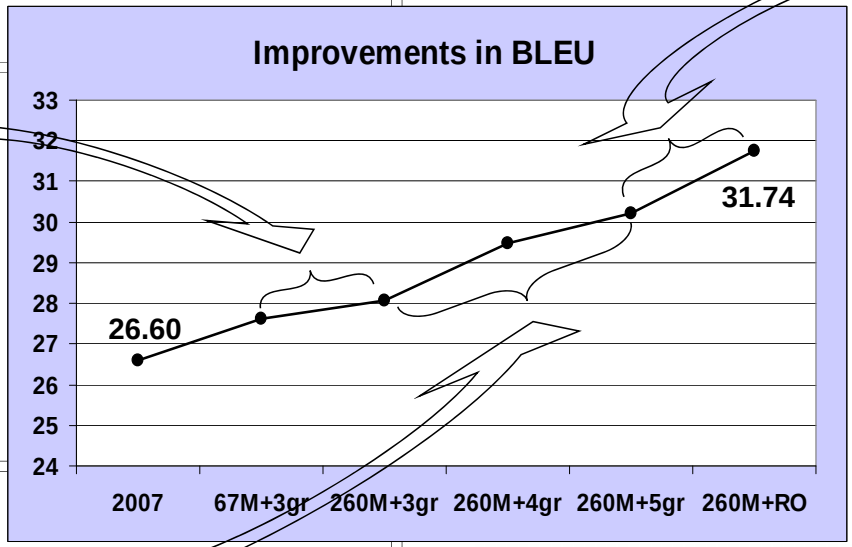
Almut Silja Hildebrand, Kay Rottmann, Mohamed Noamany,
Qin Gao, Sanjika Hewavitharana, Nguyen Bach, Stephan Vogel

Setup

- CMU SMT phrase based beam search decoder
- MT03 for dev test, MT06 NIST part (incl. audio) as unseen
- GALE parallel data: 10M sentences, 260M words
- LM data: GigaWordV3 + bilingual

Translation Model Training

- Parallelized GIZA++
- Speedup from 169 hours to 39 hours by using on average 11 CPUs in parallel opposed to 2 CPUs
- Allows for retraining of the full system on a regular basis



Language Model Training

- 2.7 Gigawords training data
- SriLM toolkit, 5-gram LM, interpolated Kneser Ney smoothing
- Interpolation of different sources

LM interpolation weights per source		
xin 0.30	cna 0.06	nyt 0.03
bil 0.26	un 0.07	ltw 0.01
afp 0.21	apw 0.05	

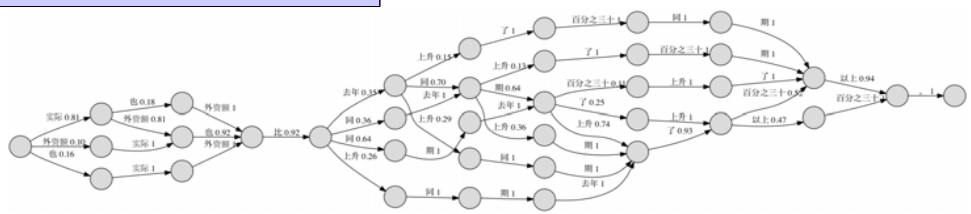
Improvements for longer LM history in BLEU/TER		
System	MT03	MT06
260M, 3gram	30.27 / 61.46	28.07 / 59.73
260M, 4gram	31.20 / 61.00	29.49 / 61.00
260M, 5gram	32.20 / 60.59	30.22 / 60.81

POS based Reordering

- Decode source lattice containing reordering alternatives
- Allows for long distance reordering without slowing down decoding

1. **POS tag training corpus**
2. **Automatically extract POS based reordering patterns**
3. **Use left and right context of patterns**
4. **Score patterns with relative frequencies**
5. **Apply patterns to source sentence to build lattice**

Improvements for reordering lattice decoding		
System	MT03	MT06
260M, standard	32.20 / 60.59	30.22 / 60.81
260M, lattice	33.53 / 59.74	31.74 / 59.59



Did not work: Number Tagging

- Tag and pretranslate testdata only
- Tag training data and test data
- Unbalanced number tags in training data

System	MT03	MT06
67M baseline	31.45 / 60.93	27.61 / 62.18
test data tagged	--	26.06 / 63.36
260M, 5gram	29.07 / 62.52	25.97 / 63.39

Carnegie Mellon