I²R Chinese-English Translation System for IWSLT-2007

Boxing Chen, Jun Sun, Hongfei Jiang, Min Zhang, Ai Ti Aw

Department of HLT
Institute for Infocomm Research (I²R), Singapore

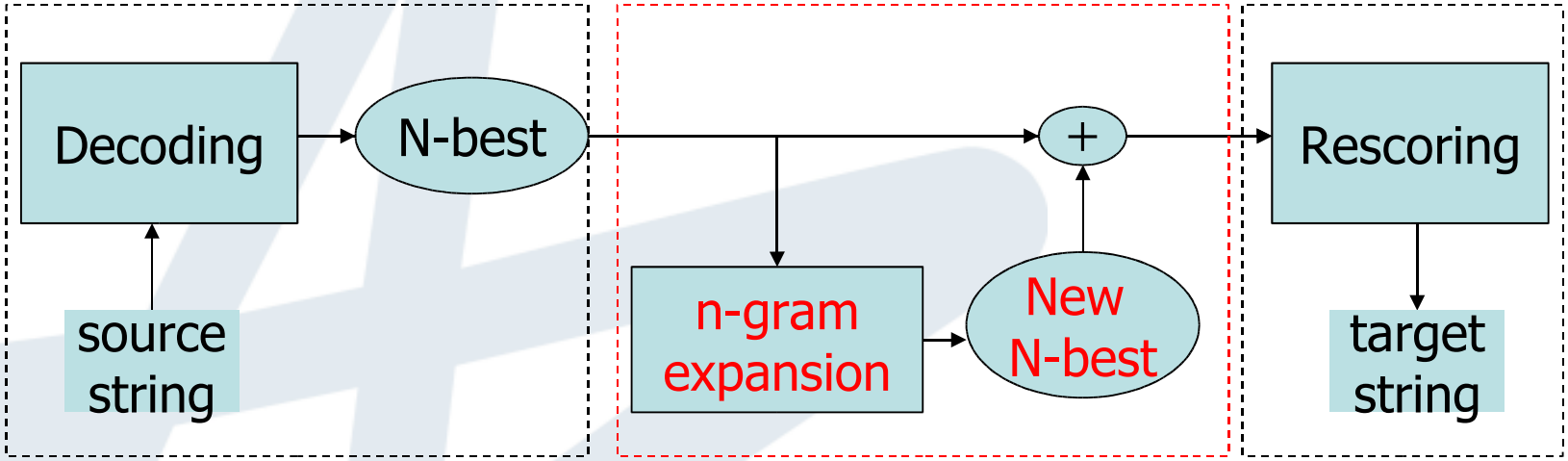
Outline

- Motivation
- Multi-pass approach
  - 1st pass: decoding
  - 2nd pass: regeneration
  - 3rd pass: rescoring
- Experiments and results
- Conclusion

Motivation

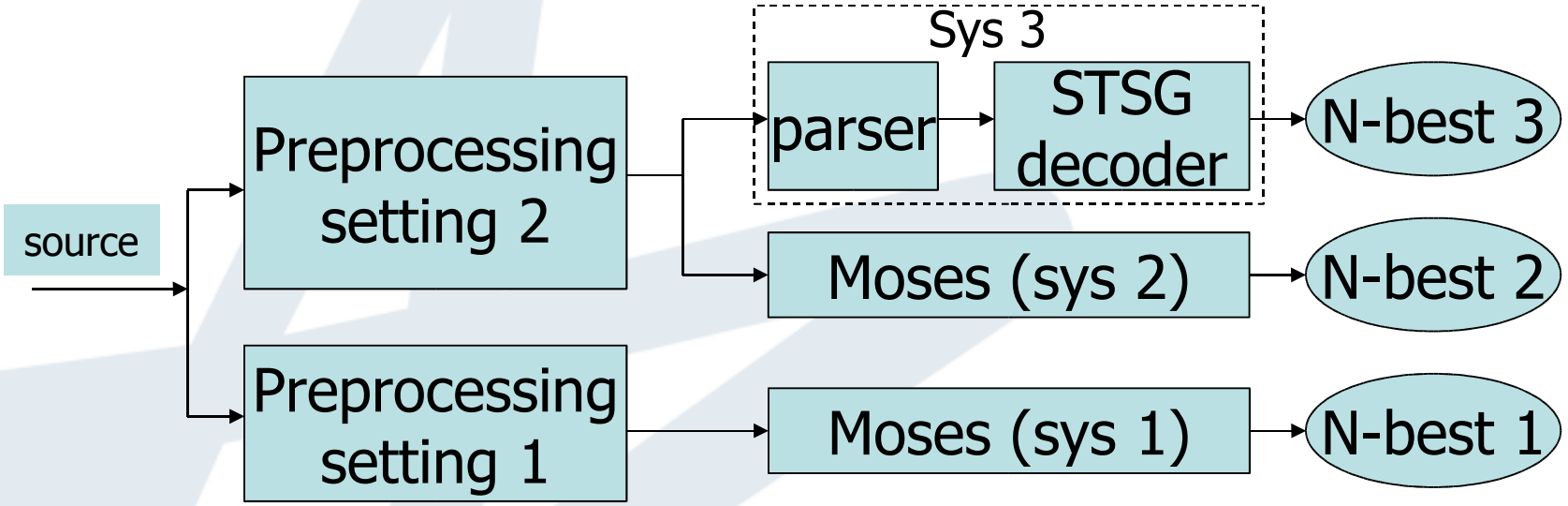
- A two-pass SMT system's performance could be improved from two aspects:
  - Scoring models
  - N-best Hypotheses
- Rescoring focus on improving the scoring models
- We try to improve the N-best hypotheses through an additional pass: regeneration and system combination

Multi-pass Approach



- 1st Pass
  - Decoding
  - Log-linear model
  - Multi decoders
- 2nd Pass
  - n-gram expansion
  - System combination
- 3rd Pass
  - Rescoring
  - Log-linear model
  - Additional features

1st Pass: Decoding

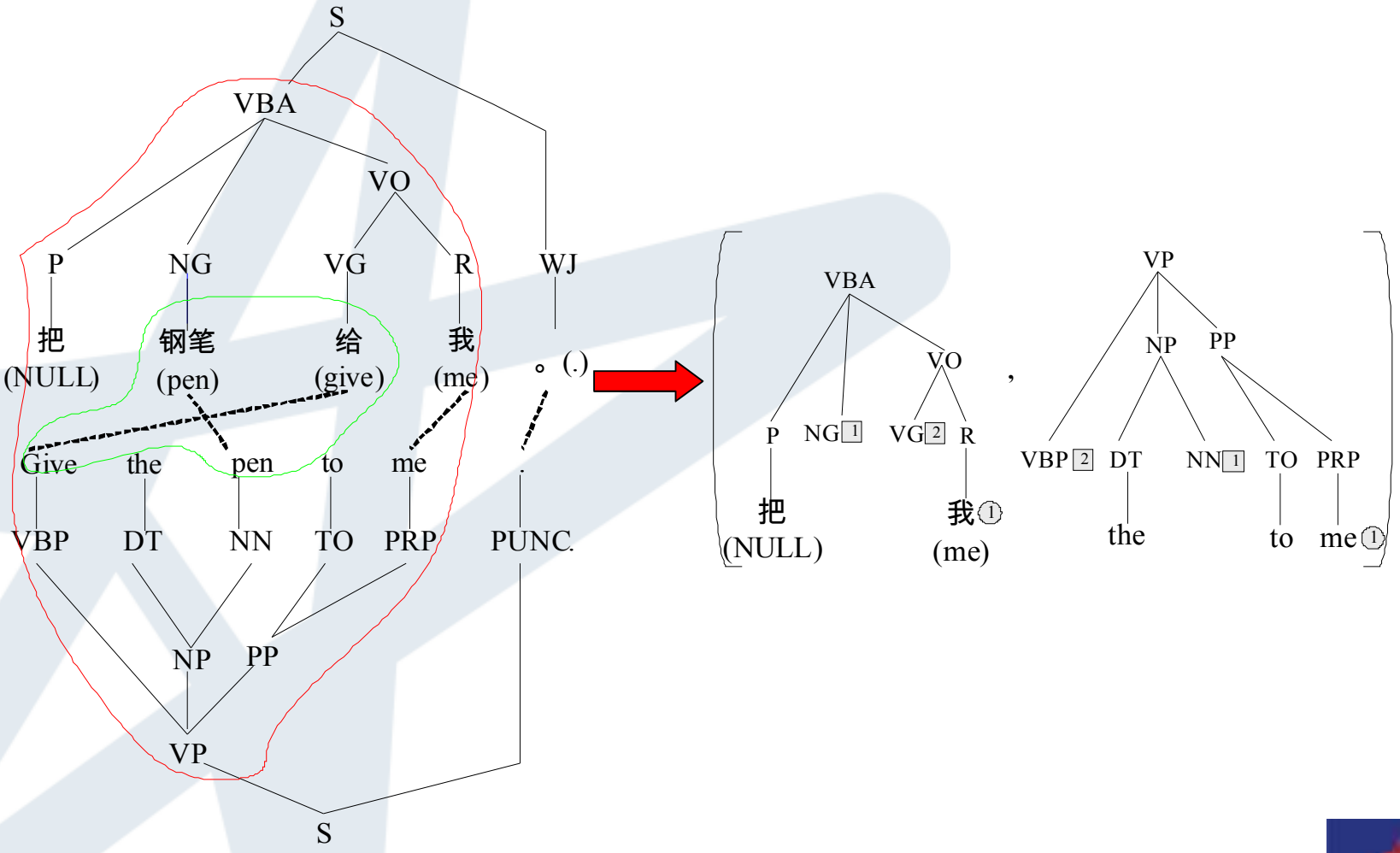


- 3 systems
  - Sys1: preprocessing setting 1 + Moses decoder
  - Sys2: preprocessing setting 2 + Moses decoder
  - Sys3: preprocessing setting 2 + STSG decoder

1st Pass: Syntax-based decoder

- STSG: Synchronous Tree Substitution Grammar
- A rule is a pair of elementary tree (PET) with alignment information.
  - PET is defined as a Triple $<\xi_s, \xi_t, A>$
    - ξ_s and ξ_t are source/target elementary tree
    - A is the alignments between leaf nodes of two elementary trees
- Two major benefits:
  - Possible to explicitly model the target syntax
  - Allow Multi-level global structure distortion

1st Pass: STSG Modelling



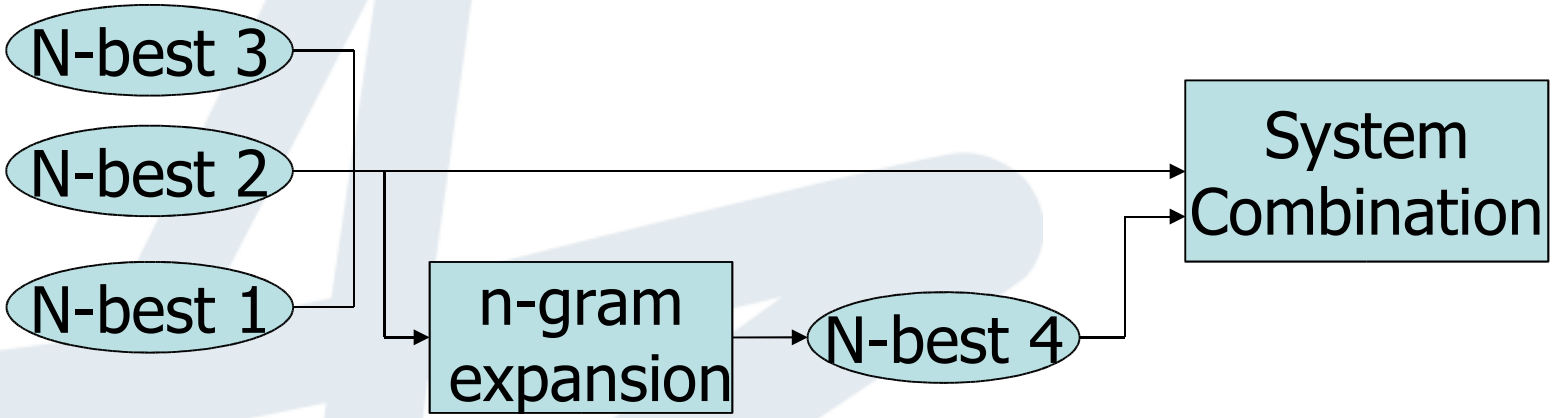
A*STAR I²R

2nd pass: n-gram expansion

- n-gram expansion generates new hypotheses
  - Collect all the n-grams from the original N-best
  - Continuously expand the partial hypothesis through the n-grams.

Reference:	my book is in the green basket .
Original entry:	my book is in the green case . my book is inside the green basket .
3-grams:	my book is, book is in, is in the, in the green, the green case, is inside the, the green basket …
n-gram expansion	Partial Hyp: my book is in n-gram: is in the New partial Hyp: my book is in the
New Hyp:	my book is in the green basket .

2nd Pass: System Combination



- System Combination
  - Hypotheses are simply joined
  - Duplicate hypotheses are removed

3rd Pass: Rescoring

- Rich additional feature functions (Chen et al., 2006)

Moses Features:

- Translation Model
- Reordering model
- Language Model
- Word penalty

Translation confidence

Rescoring Features:

1) Dir/Inv IBM model 1 and 3 score
2) CLA association score
3) lexicalized word/block reordering probabilities
4) 6-gram target LM
5) 8-gram target word-class based LM
6) source and target length ratio
7) question feature
8) frequency of n-grams in the N-best
9) n-gram post-probabilities
10) sentence length post-probabilities

Experiments: training data

- Task: Chinese-English Open data track
- Bilingual Training data: BTEC+HIT-corpus
  - Sys1 and Sys2:
    - 400K sentence-pairs
    - 4.5M target words
  - Sys3:
    - 90K sentence-pairs
    - 1.0M target words
- Additional target data: Tanaka corpus
  - 155K sentence-pairs, 1.4M target running words

Experiments: preprocessing

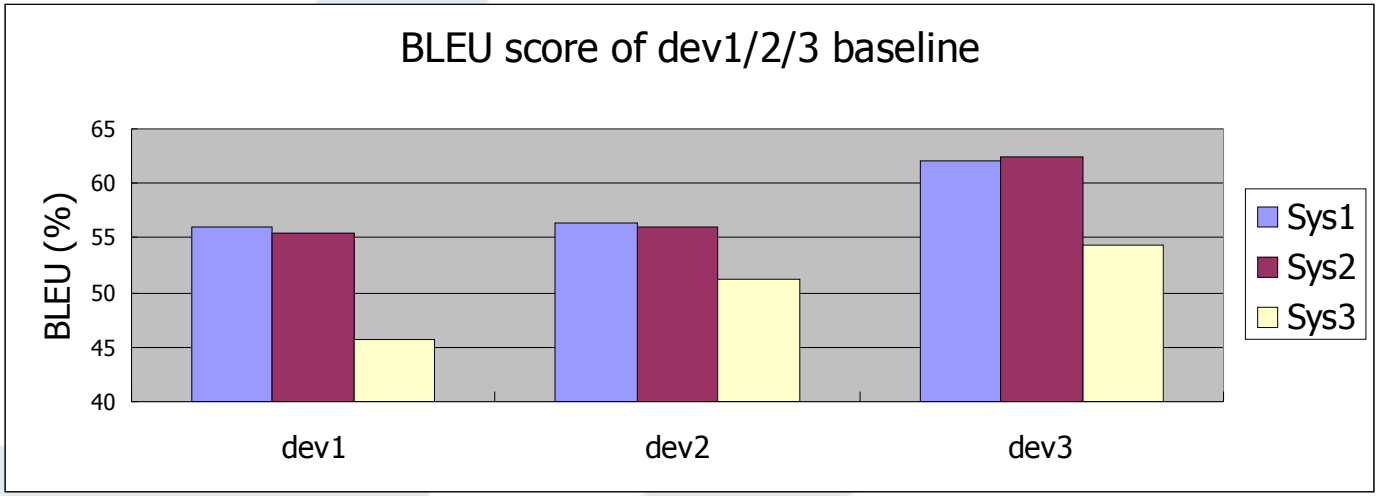
- Preprocessing
  - Tools: LDC-SEG (L) , ICTCLAS (I), Stanford parser

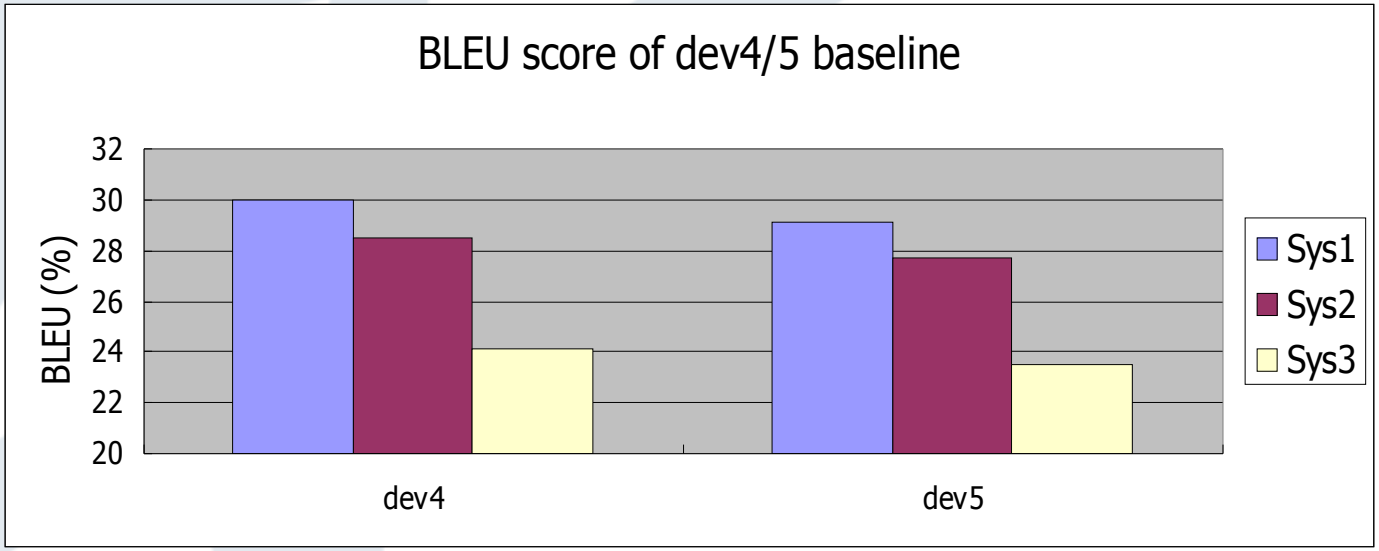
	Sys1		Sys2		Sys3	
	ch	en	ch	en	ch	en
Tokenization	L	x	I	x	I	x
Parsing					x	x
Txt-to-digit	x	x				
Lower-casing		x		x		x

Experiments: setting

- Two series of experiments:
  - DEV: dev1, TEST: dev2, dev3
  - DEV: dev4, TEST: dev5
- 6 types of MT outputs:
  - Sys1/2/3: 3 baselines
  - Resc1: rescoring on Sys1 N-best list
  - Resc2: rescoring on Sys1+Sys2 N-best lists
  - Comb: final translation output with n-gram expansion, system combination and rescoring incorporated

Results: Baseline



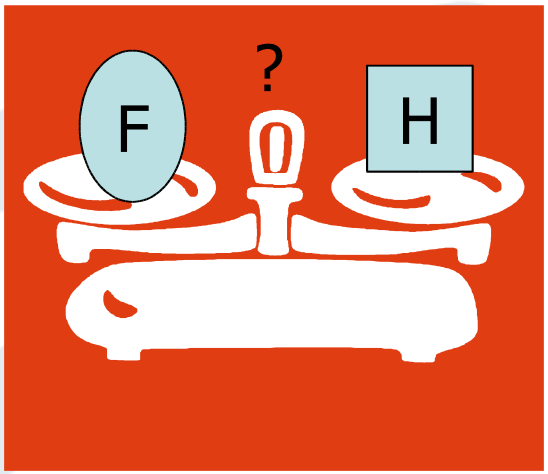


Results: Resc1/2 vs. Comb

Resc1/2:

Advantages:
More features (include local feat. used in decoding)

Disadvantages:
Less hypotheses

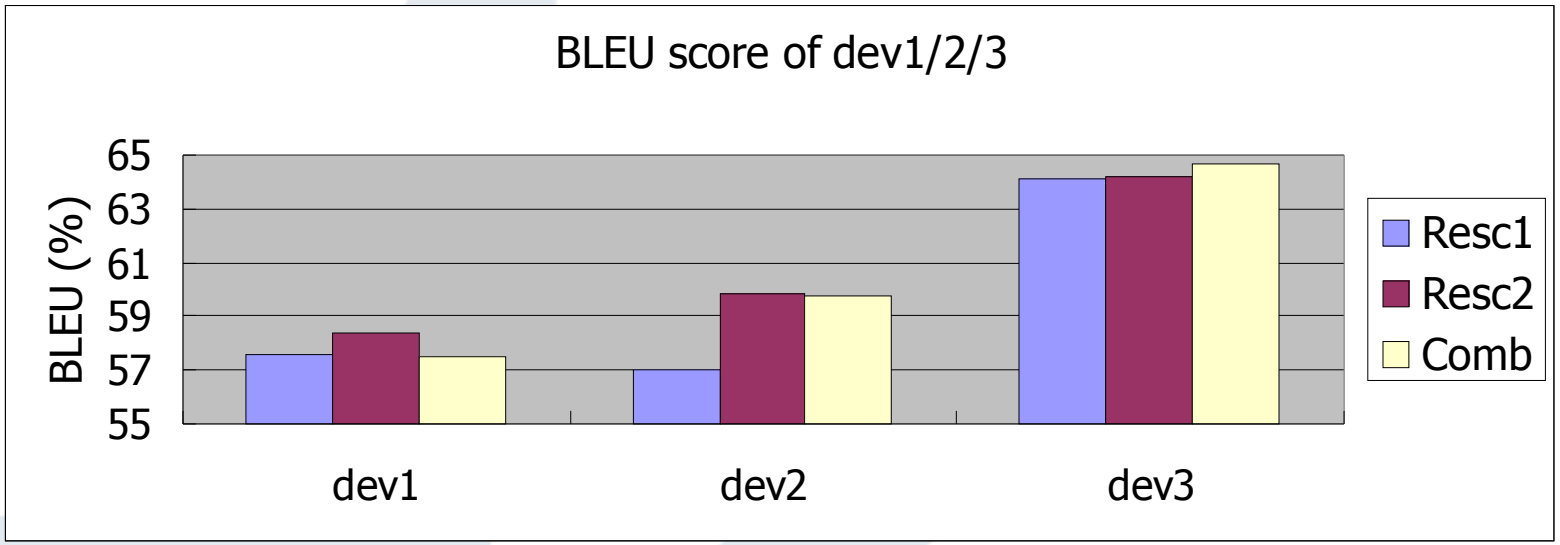


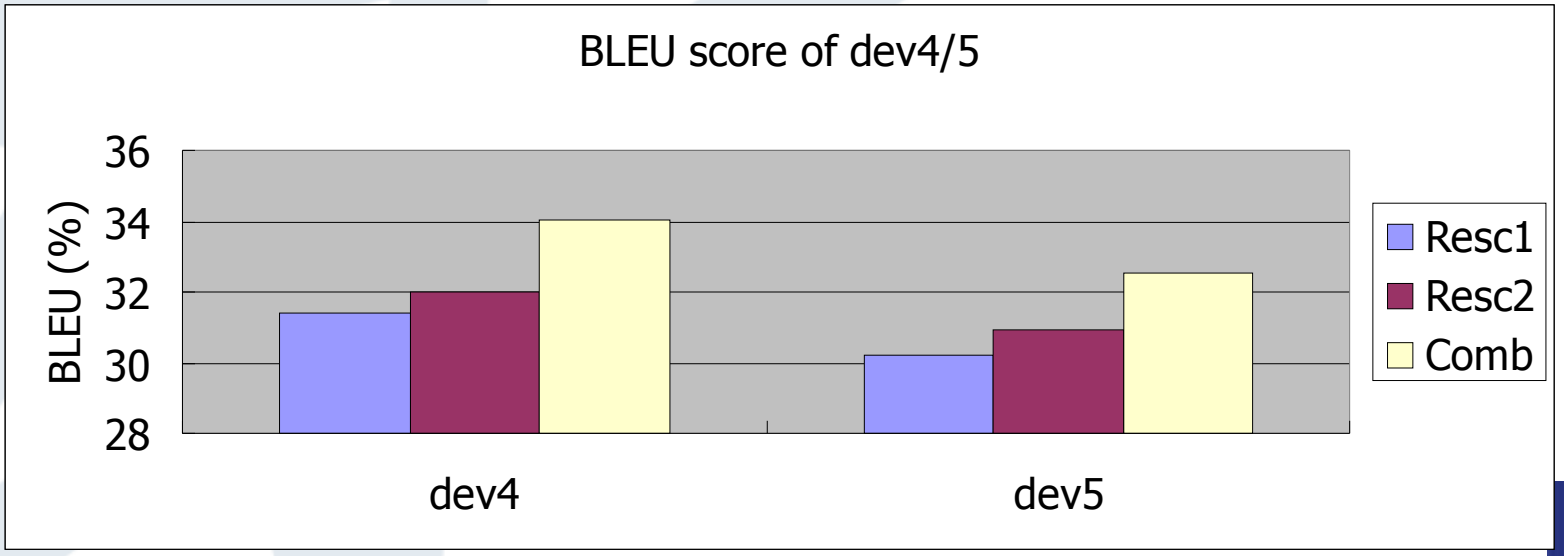
Comb:

Advantages:
More hypotheses

Disadvantages:
Less features (no local features)

Results: Resc1/2 vs. Comb





Results: Analysis

- Average length and relative improvements on BLEU (Resc2 vs. Comb)

	Dev1	Dev2	Dev3	Dev4	Dev5
Length	6.7	7.0	7.5	12.1	12.6
△	-1.5	-0.2	0.8	6.3	5.3

- Number of new generated hypotheses in Comb (about 500 sentences for each dev set.

	Dev1	Dev2	Dev3	Dev4	Dev5
#new hypo	29	18	12	59	74

- n-gram expansion benefits longer sentences more than short sentences. Because it permits long distance word movements through a low-order LM (e.g. a bi-gram LM).

Results: test set

- Test set are more similar to dev1 than other dev sets:
  - average length 6.5 (test) vs. 6.7(dev1)
- On dev1: "Resc2" produces better BLEU score than "Comb"

	Official submission		Only BTEC data	
	BLEU(%)	Rank	BLEU(%)	NIST
Run1 (Resc2)	40.77	1	38.67	6.740
Run2 (Comb)	39.42	2	37.04	6.756

Conclusion

- Multi-pass system
  - Multi-decoder to produce N-best lists
  - n-gram expansion to generate new hypotheses
  - Rich additional feature functions to do rescoring
- Rescoring gives significant improvements
- n-gram expansion and system combination give consistent improvement on longer sentences

Thanks for your attention!

Any questions?