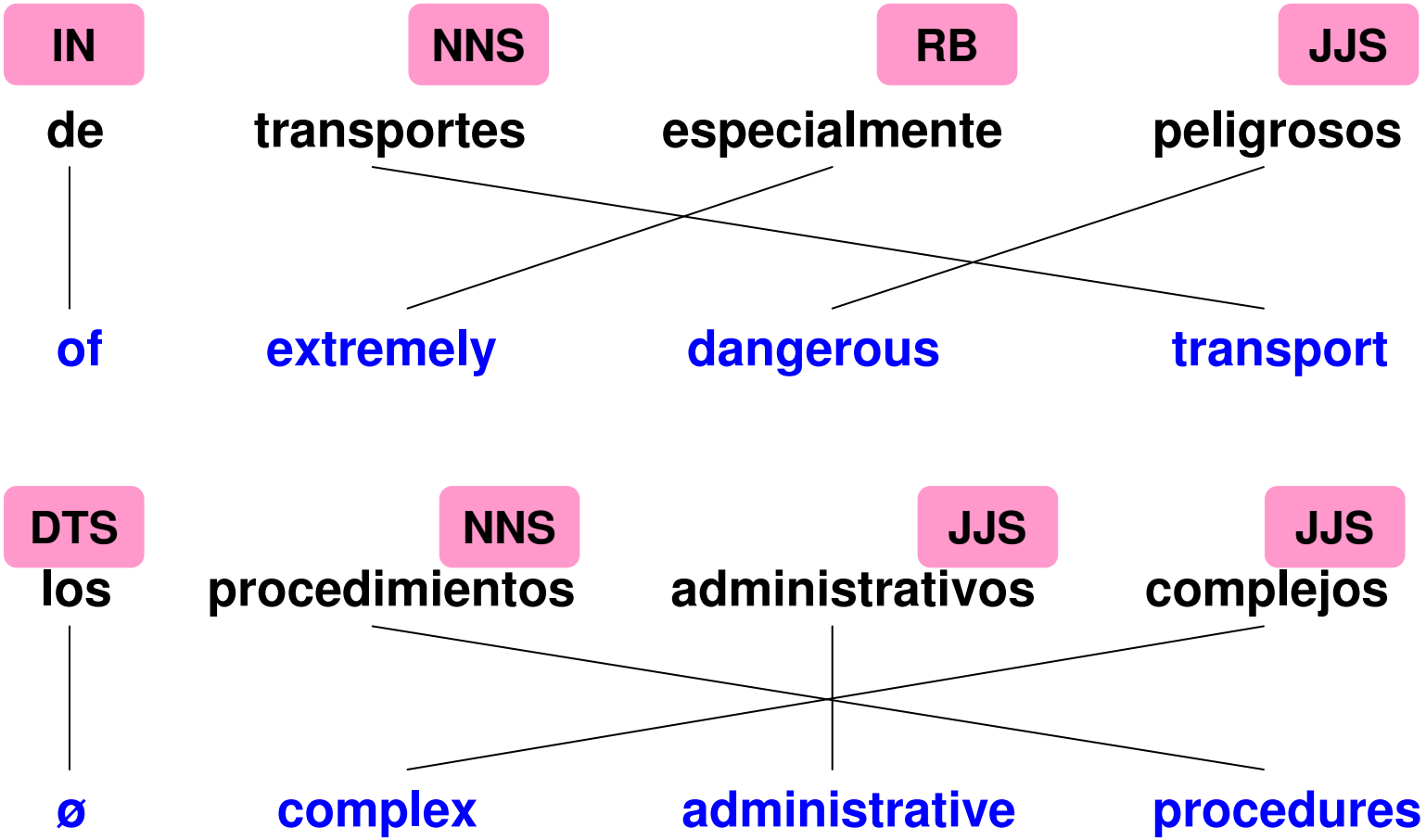
Morpho-Syntax in Statistical Machine Translation

Young-Suk Lee
IBM T. J. Watson Research Center

OpenLab 2006
March 30 – April 1, 2006

Reordering Rules: Motivations

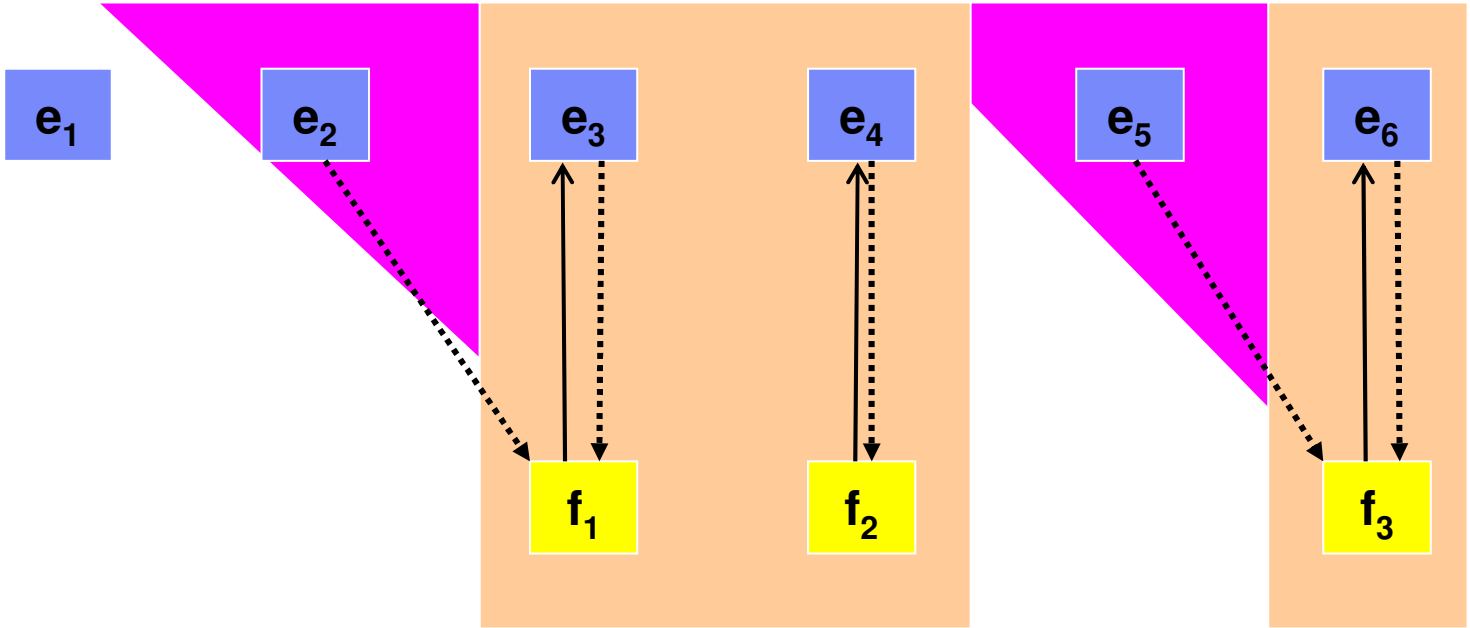


Outline

- **Baseline Phrase Translation System**
  - **Block Acquisition & Decoding**
- **Acquisition of Reordering Rules**
  - **Base Reordering Rules**
  - **Lexicalized Reordering Rules**
- **Experimental Results**
- **Related and Ongoing Work**

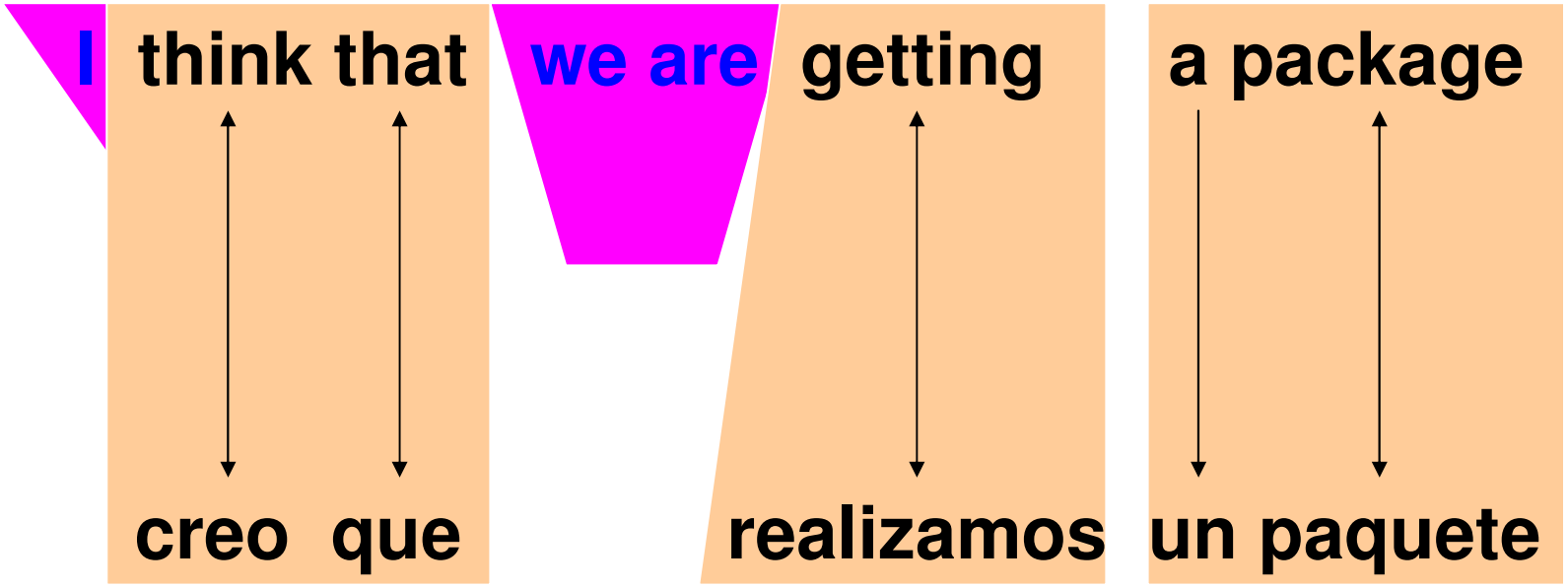
Baseline Block Acquisition

Block (b): a phrase translation pair consisting of source $\bar{f}$ & target $\bar{e}$ phrase



Tillmann 2003, EMNLP Proceedings

Extended Block Acquisition Algorithm



- Expansion word list: A list of target words typically aligned to null source words (e.g. *I, we, are*)
- Extend the target phrase to include an expansion word if it occurs in the neighborhood of a block

Decoding

- **Phrase translation models**
  - **Direct model:** $$p(\bar{e} \mid \bar{f}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{e}'} count(\bar{e}', \bar{f})}$$
  - **Source channel model:** $$p(\bar{f} \mid \bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}'} count(\bar{f}', \bar{e})}$$
  - **Block unigram model:** $$p(b) = \frac{count(b)}{\sum_{b'} count(b')}, b = (\bar{e}, \bar{f})$$

Decoding Cont'd ...

- **IBM Model 1 cost per phrase in both directions**

$$\sum_{j=1}^{m} -\log 10 \, max_i \; p(f_j \mid e_i), 1 \le i \le n$$

- **Word & part-of-speech tag trigram language models**
- **Word-level distortion models applied to blocks**
  - **Al-Onaizan 2004, DARPA MT Evaluation Workshop**
- **Word & block count penalty**
  - **Zens and Ney 2004, HLT Proceedings**

Acquisition of Base Reordering Rules

- Viterbi-align
  - Part-of-speech tagged source language corpus
  - Un-tagged target language corpus
- Identify the source language part-of-speech tag sequence (monotone increasing)
  - whose corresponding target word sequence is not monotone increasing
- Compute the reordering probabilities of each part-of-speech tag sequence

Reordering Probability Computation

$$p(reorder_i \mid \overline{tag}_k) = \frac{count(reorder_i, \overline{tag}_k)}{\sum_{reorder'} count(reorder', \overline{tag}_k)}$$

$DTS_1\ NNS_2\ JJS_3\ JJS_4$		$IN_1\ NNS_2\ RB_3\ JJS_4$	
reorder'	$p(reorder_i \mid \overline{tag}_k)$	***reorder'***	$p(reorder_i \mid \overline{tag}_k)$
1 2 3 4	0.107	1 2 3 4	0.179
1 2 4 3	0.059	1 2 4 3	0.047
1 3 2 4	0.239	1 3 2 4	0.094
1 3 4 2	0.109	1 3 4 2	0.560
1 4 2 3	0.111	1 4 2 3	0.072
1 4 3 2	0.375	1 4 3 2	0.048

One Best Reordering Rules

$$p(reorder_f \mid \overline{tag}) > p(reorder_s \mid \overline{tag}) + \alpha$$

IN_1 NNS_2 RB_3 JJS_4	IN_1 RB_3 JJS_4 NNS_2
DTS_1 NNS_2 JJS_3 JJS_4	DTS_1 JJS_4 JJS_3 NNS_2
DT_1 NN_2 JJ_3 IN_4	DT_1 JJ_3 NN_2 IN_4
NNS_1 JJS_2 CC_3 JJS_4	JJS_2 CC_3 JJS_4 NNS_1
DT_1 NN_2 CC_3 NN_4 JJ_5	DT_1 JJ_5 NN_2 CC_3 NN_4

Lexicalization of Exceptions

el apoyo operativo de la$_1$/DT Secretaría$_2$/NN General$_3$ /JJ del$_4$ /IN Consejo
the operational support of the$_1$ Secretary$_2$ General$_3$ of$_4$ the Council

$DT_1\ NN_2\ JJ_3$[~General] $IN_4 \rightarrow\ DT_1\ \ JJ_3\ \ NN_2\ \ IN_4$

El Fondo, por supuesto, debe continuar cumpliendo con su misión de investigación sobre la búsqueda de$_1$/IN variedades$_2$/NNS más$_3$/RB adaptadas$_4$ /JJS a la demanda y lo menos nocivas posible,
The Fund must of course continue to serve its purpose and pursue research into$_1$ varieties$_2$ more$_3$ suited$_4$ to demand and causing as little harm as possible .

$IN_1\ NNS_2\ \ RB_3\ \ JJS_4$[~adaptadas] $\rightarrow\ IN_1\ \ RB_3\ \ JJS_3\ \ NNS_2$

Lexicalized Reordering Rules

- Identify the key part-of-speech tag in the base reordering rules
- Replace the key part-of-speech tag with the corresponding word
  - DT NN JJ IN → DT NN General IN
- Compute reordering probabilities of lexicalized part-of-speech tag sequences
- Exception word list
  - If the reordering pattern with the highest probability is monotone increasing, select the word in the pattern as an exception

Lexicalized Reordering Probabilities

DT_1 NN_2 *General*$_3$ IN_4	
reorder´	$p(reorder_i \mid tag_k)$
1 2 3 4	0.454
1 2 4 3	0.021
1 3 2 4	0.201
1 3 4 2	0.012
1 4 2 3	0.194
1 4 3 2	0.098
4 1 2 3	0.007
4 1 3 2	0.013

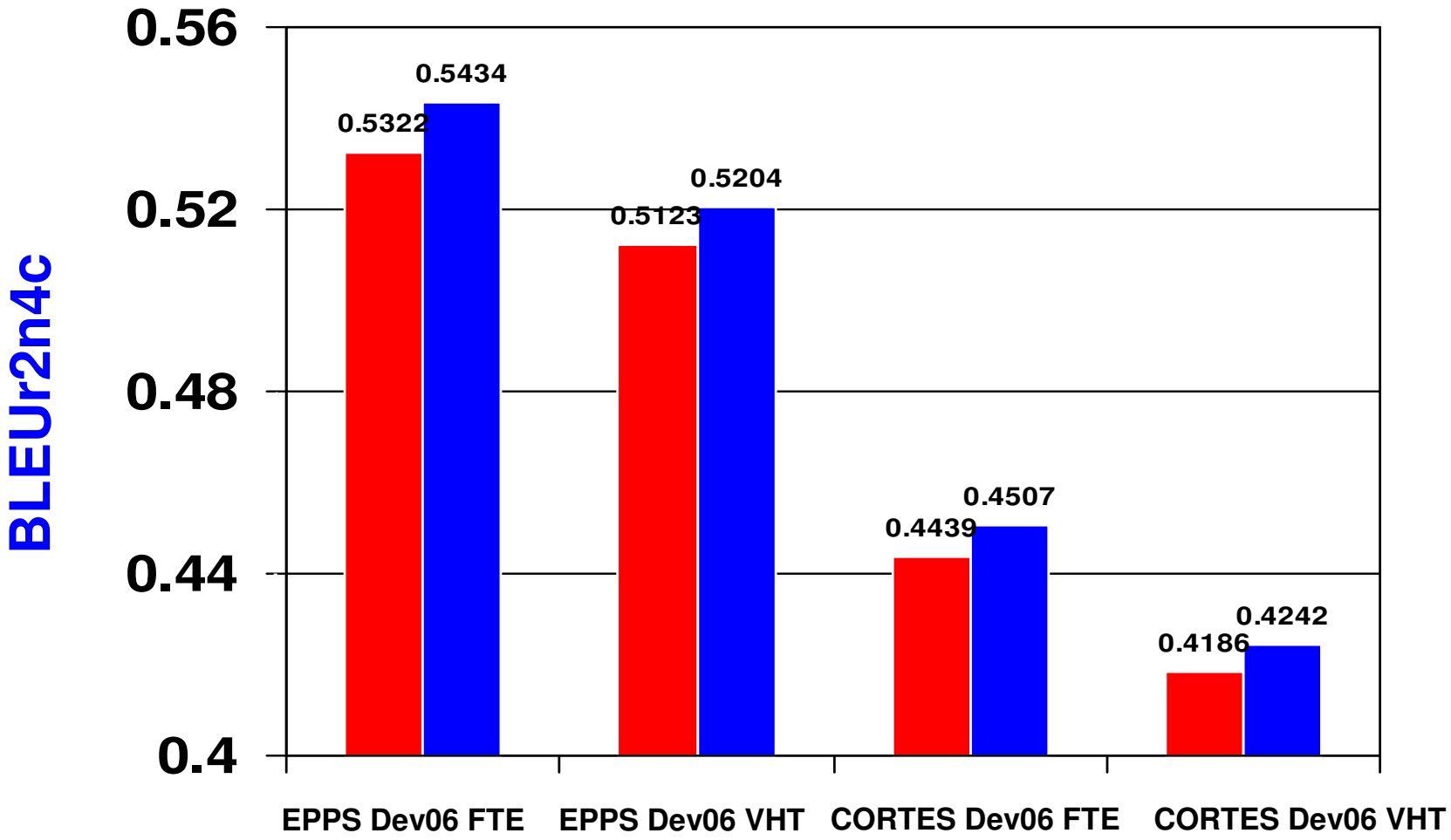
Performance Evaluations

- Translation model training corpus
  - ~1.3 M sentence pairs from EPPS distributed by RWTH
- Language model training corpus
  - EPPS English corpus: ~35 M words
  - UN parallel corpus English (LDC94T4A): ~45 M words
  - English gigaword second edition (LDC2005T12): ~2.5 B words

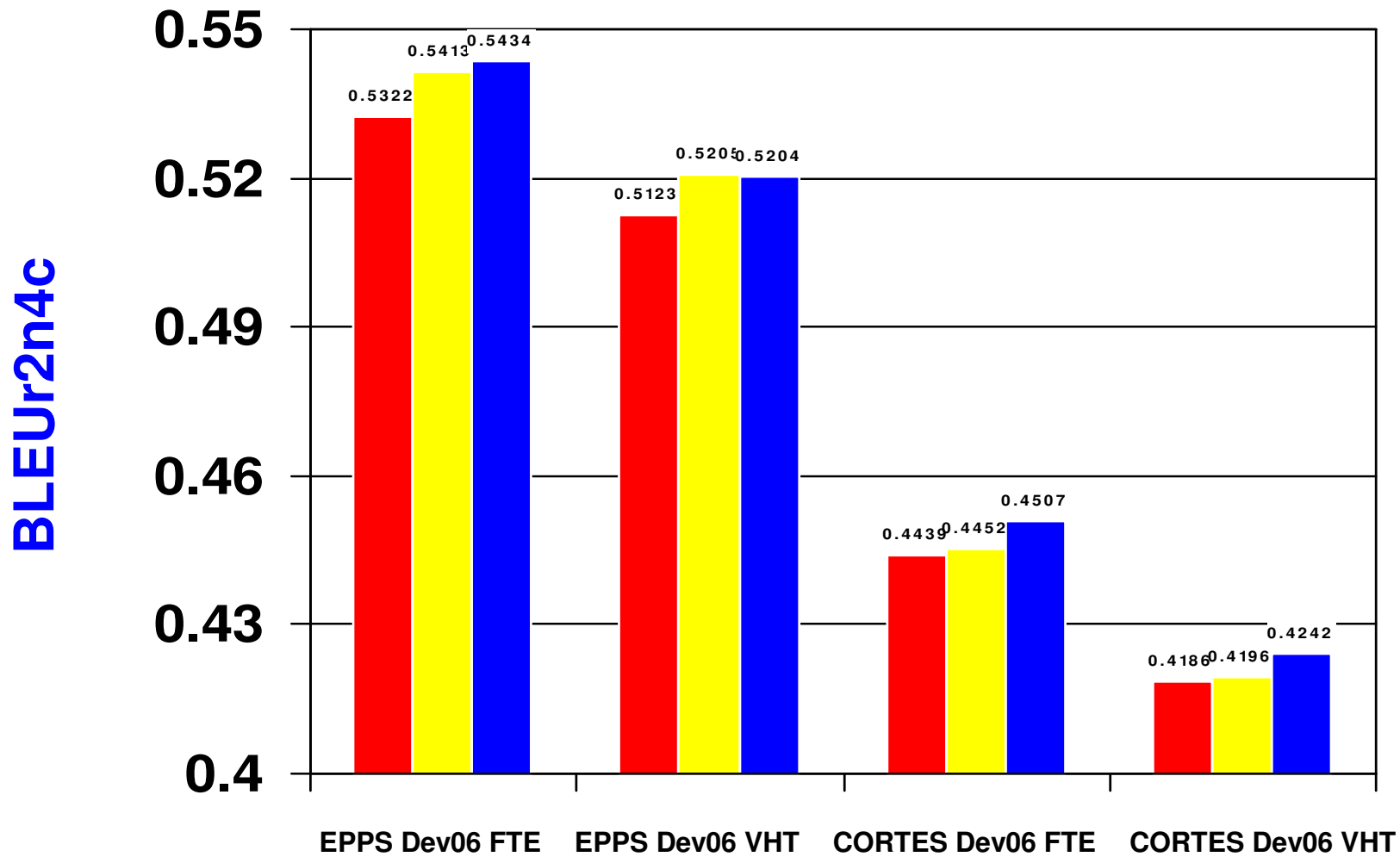
Evaluation Corpus Statistics

Data Sets	# of Segments	Avg. Segment Length
EPPS Dev06 FTE	699	35 words/segment
EPPS Dev06 VHT	792	31 words/segment
CORTES Dev06 FTE	753	37 words/segment
CORTES Dev06 VHT	920	31 words/segment

Lexicalized Reordering Rules: Impact



Base vs. Lexicalized Reordering Rules



Related Work

- **N-best Reordering in Arabic-to-English Translation**
  - **Statistically significant performance improvement by applying local reordering to noun phrase parsed Arabic**
  - ***IBM Site Report: DARPA MT Evaluation Workshop 2004***
- **Morphological Analysis for Statistical Machine Translation**
  - **Identify one to one word correspondences between Arabic and English to improve word to word translation qualities**
  - **Companion Volume of *HLT-NAACL 2004*, pages 57–60**
- **Local Reordering for Spanish-English Translations**
  - **Presentation at TC-STAR 2005 Evaluation Workshop**
  - **April 21-22, 2005, Trento, Italy**

Ongoing Work

- Non-local reordering models
  - [Se ha puesto a prueba]$_{VP}$ [su voluntad]$_{NP}$ →
    [Its will]$_{NP}$ [has been put to the test]$_{VP}$
  - Todas sus Señorías firmaron [con los electores]$_{PP}$ [un contrato]$_{NP}$ → All your ladies and gentlemen signed [a contract]$_{NP}$ [with the electors]$_{PP}$
- Integration of reordering models into the decoder