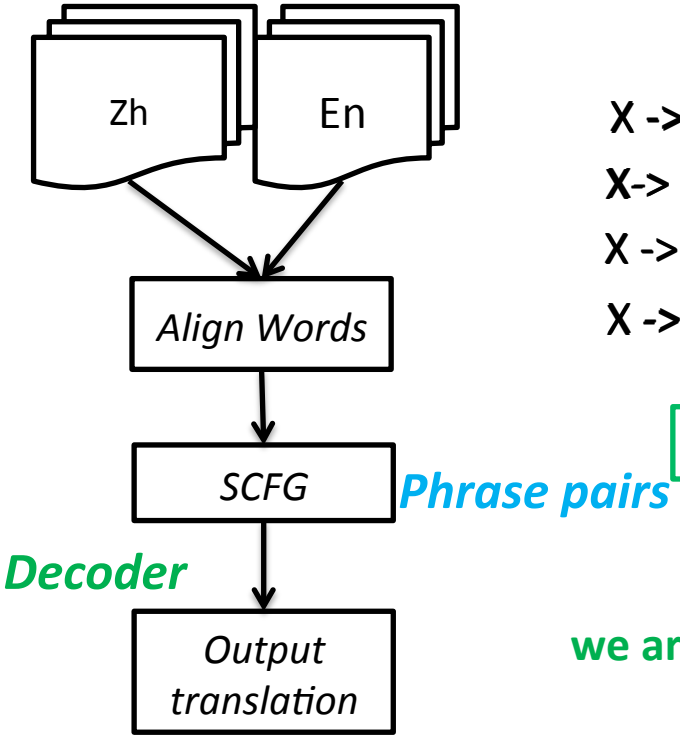
Simultaneous Translation for Hiero

Simon Fraser University

Maryam Siahbani, Anoop Sarkar

Hierarchical Phrase-based Translation (Hiero)



Synchronous Context-Free Grammar

X -> <我们十分X_1 / we are very much X_1>

X-> <关注 X_1 发生 的 X_2 /concerned with X_2 happens in X_1>

X -> <事情 / what >

X -> <非洲 地区 / African regions >

Phrase pairs

我们 十分	关注	非洲 地区	发生 的	事情

we are very much concerned with what happens in African regions

Find the correct translation for new input

Hiero Decoder (CKY)

we are very much concerned with what happens in African regions .

Bottom-up parsing algorithm

O(n^3)

Language Model (LM) computation

concerned with **LM**

happens in **LM**

we are very much concerned with **LM**

X -> <关注 X_1 发生 的 X_2 / concerned with X_2 happens in X_1>

X_1= African regions

X_1

X_2= what

X_2

我们 十分 关注 非洲 地区 发生 的 事情 。

Left-to-Right Decoding

X -> <我们十分 X_1 / we are very much [2,8]>

X -> <关注 X_1 / concerned with [3,8]>

X -> <X_1 发生 X_2 事情 / what happens [6,7] [8,1]>

X -> <的 / in >

X -> <非洲 地区 / African regions >

<s> [0,8]

<s>

LM

<s> we are very much

LM

<s> we are very much concerned with

[3,5]

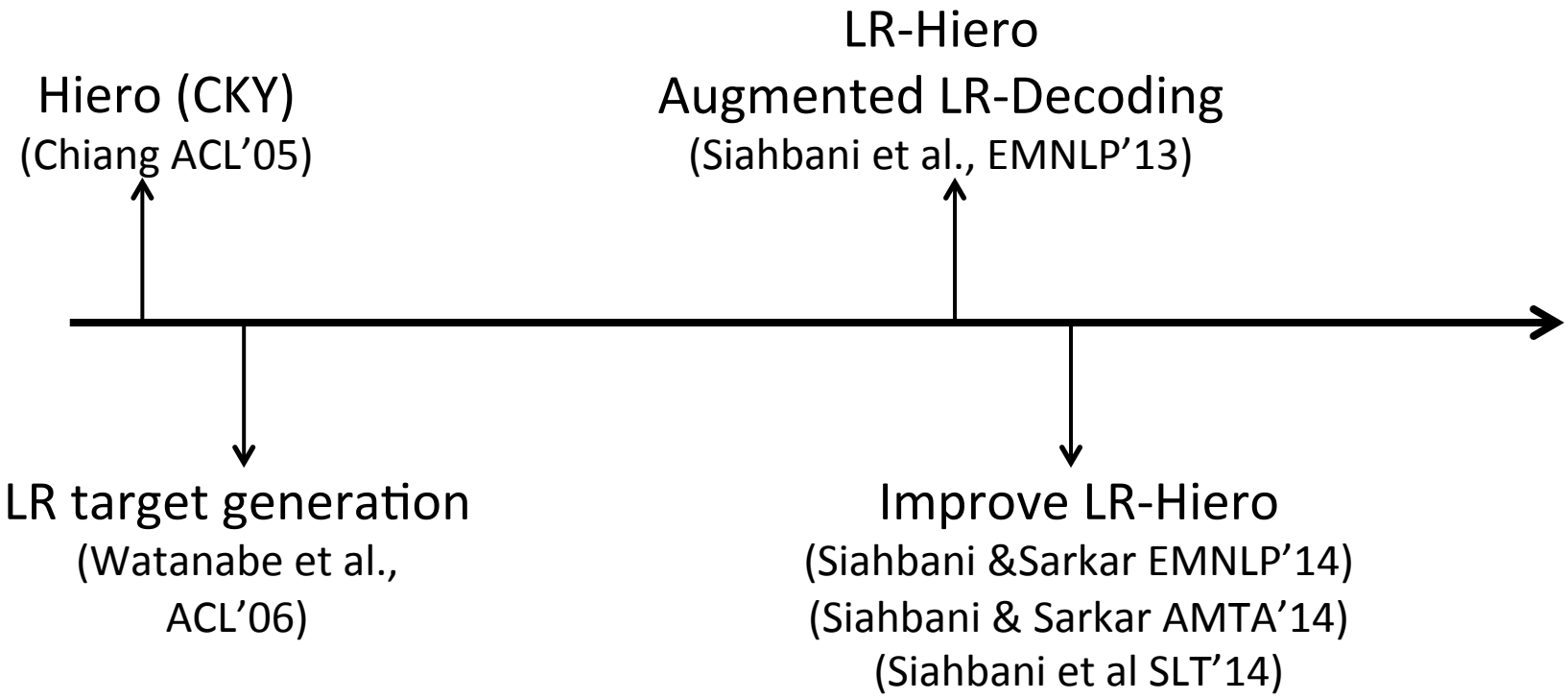
<s> we are very much concerned with what happens

<s> we are very much concerned with what happens in

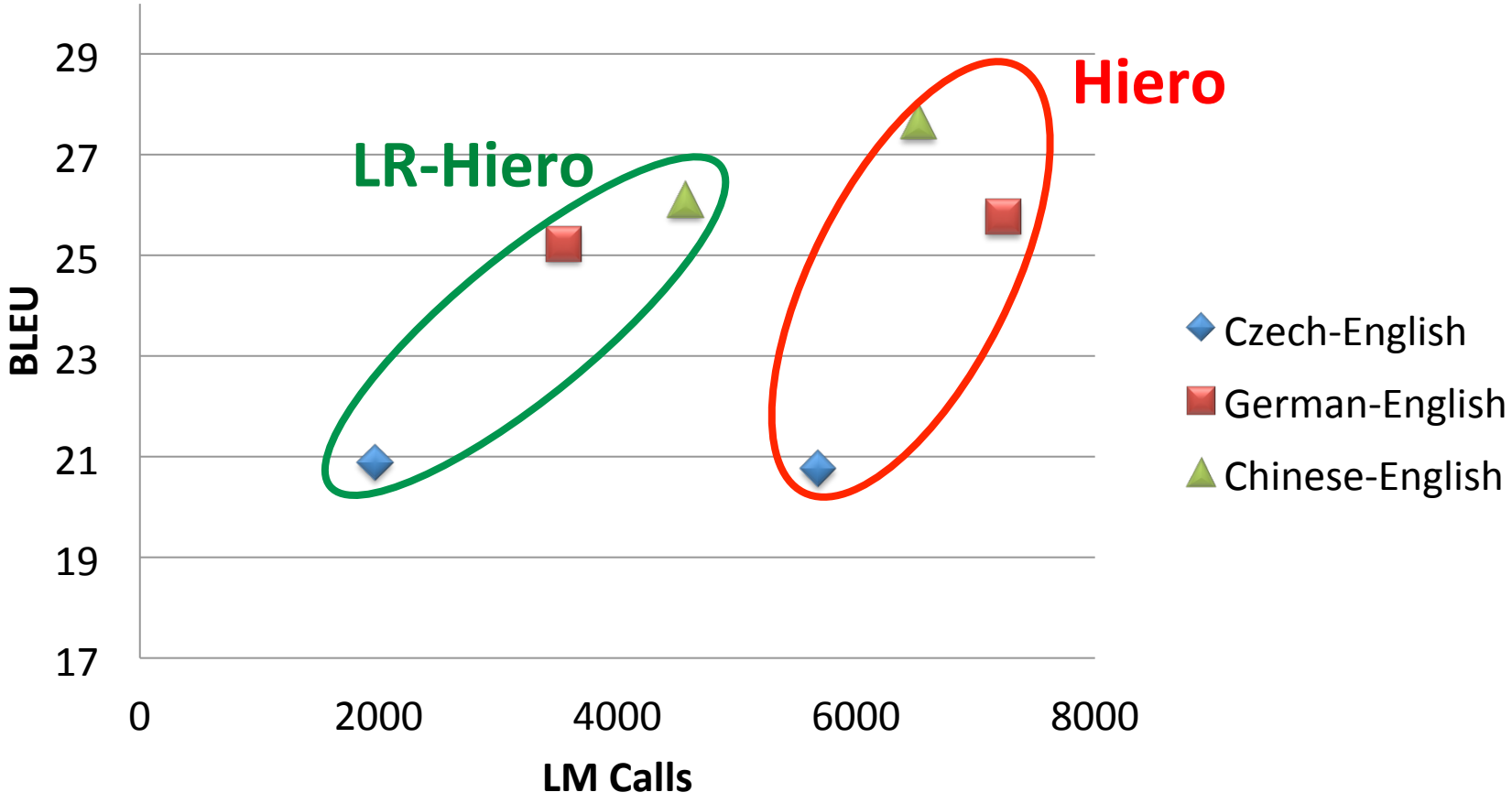
O(n^2)

Fewer LM calls

LR-Hiero


Hiero (CKY)
(Chiang ACL'05)
LR-Hiero
Augmented LR-Decoding
(Siahbani et al., EMNLP'13)
LR target generation
(Watanabe et al.,
ACL'06)
Improve LR-Hiero
(Siahbani &Sarkar EMNLP'14)
(Siahbani & Sarkar AMTA'14)
(Siahbani et al SLT'14)

Hiero vs LR-Hiero



Greibach Normal Form

X -> <我们十分 X_1 / we are very much X_1>

X -> < 关注 X_1 / concerned with X_1>

X -> <X_1 发生 X_2事情 / what happens X_2 X_1>

X -> <的 / in >

X -> <非洲 地区 / African regions >

$$X \rightarrow \langle \gamma, \bar{b}\, \beta \rangle$$

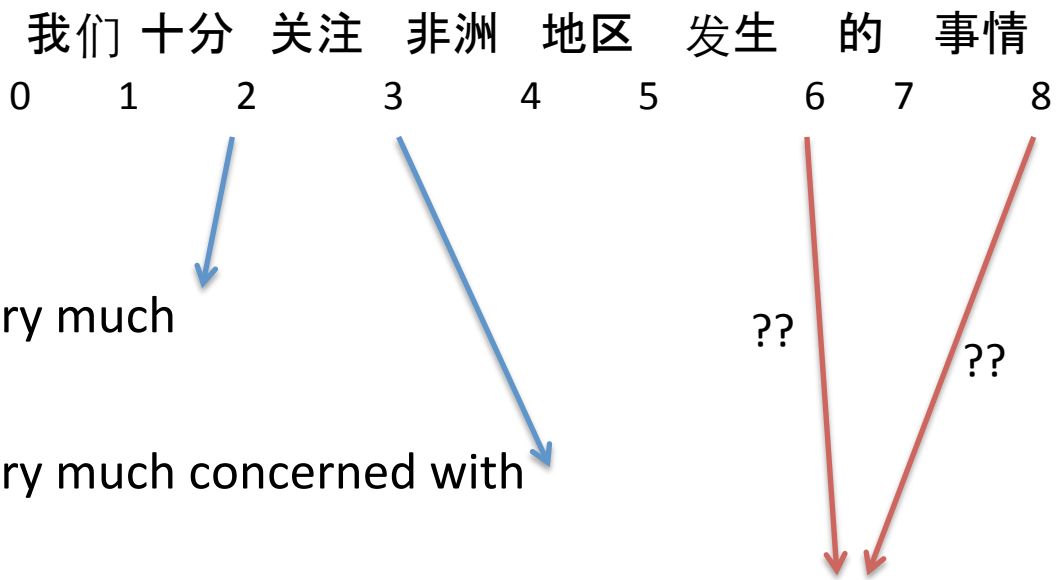
GNF

Non-GNF X -> <X_1 发生 的 X_2 / X_2 happens in X_1>

Simultaneous Translation

http://site.interpretereducationonline.com/interpreting-jokes/

Simultaneous Decoding



X -> <X_1 发生 X_2事情 / what happens X_2 X_1>

Wait till the end …

Good evening, I would like a taxi to the airport please → **4 sec** → buena noches, quiero un taxi al aeropuerto por favor

(higher fluency and latency)

Translate incrementally …

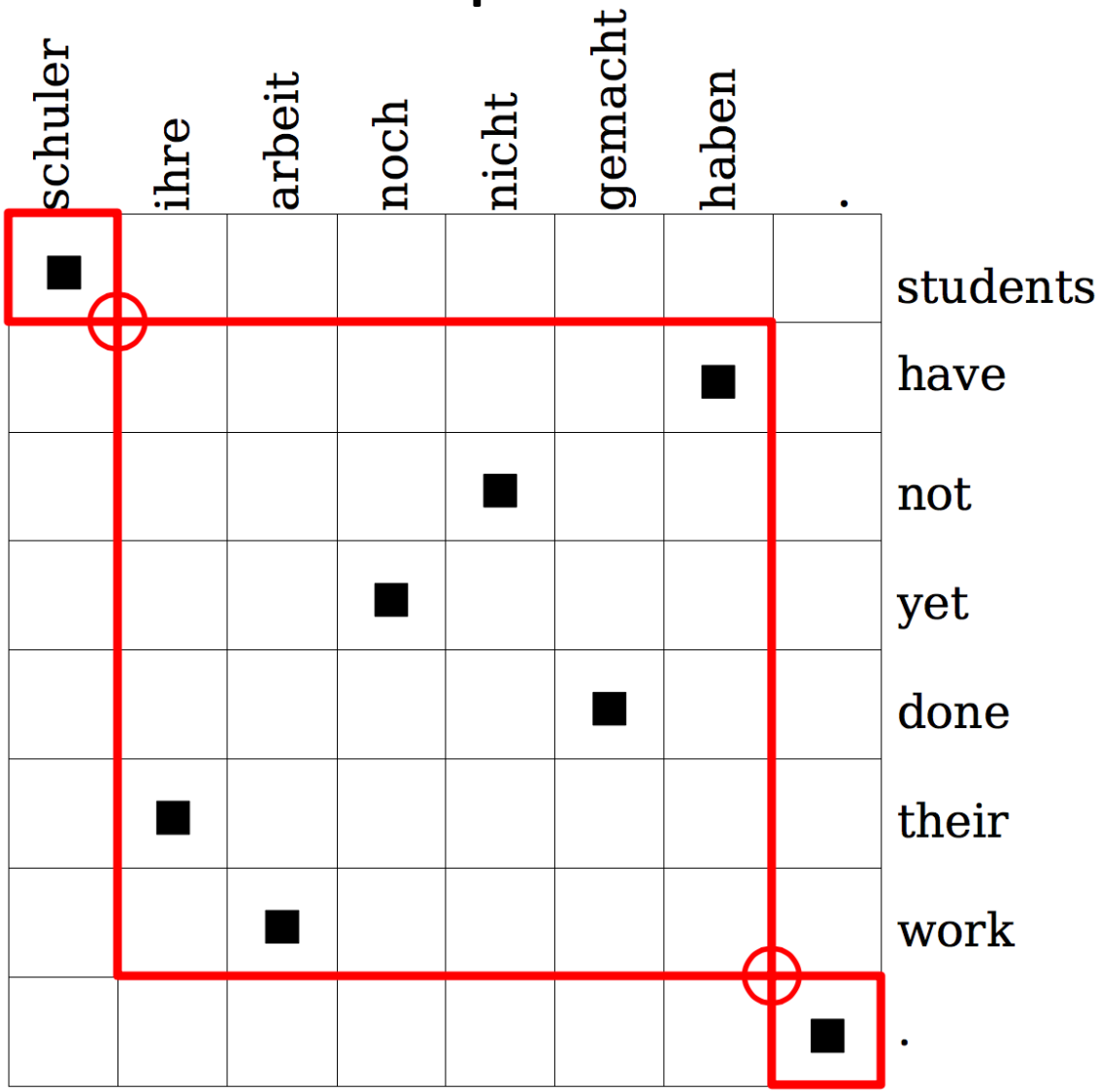
Good evening, → **0.2 sec** → buena noches

I would like → **0.2 sec** → me gustaría

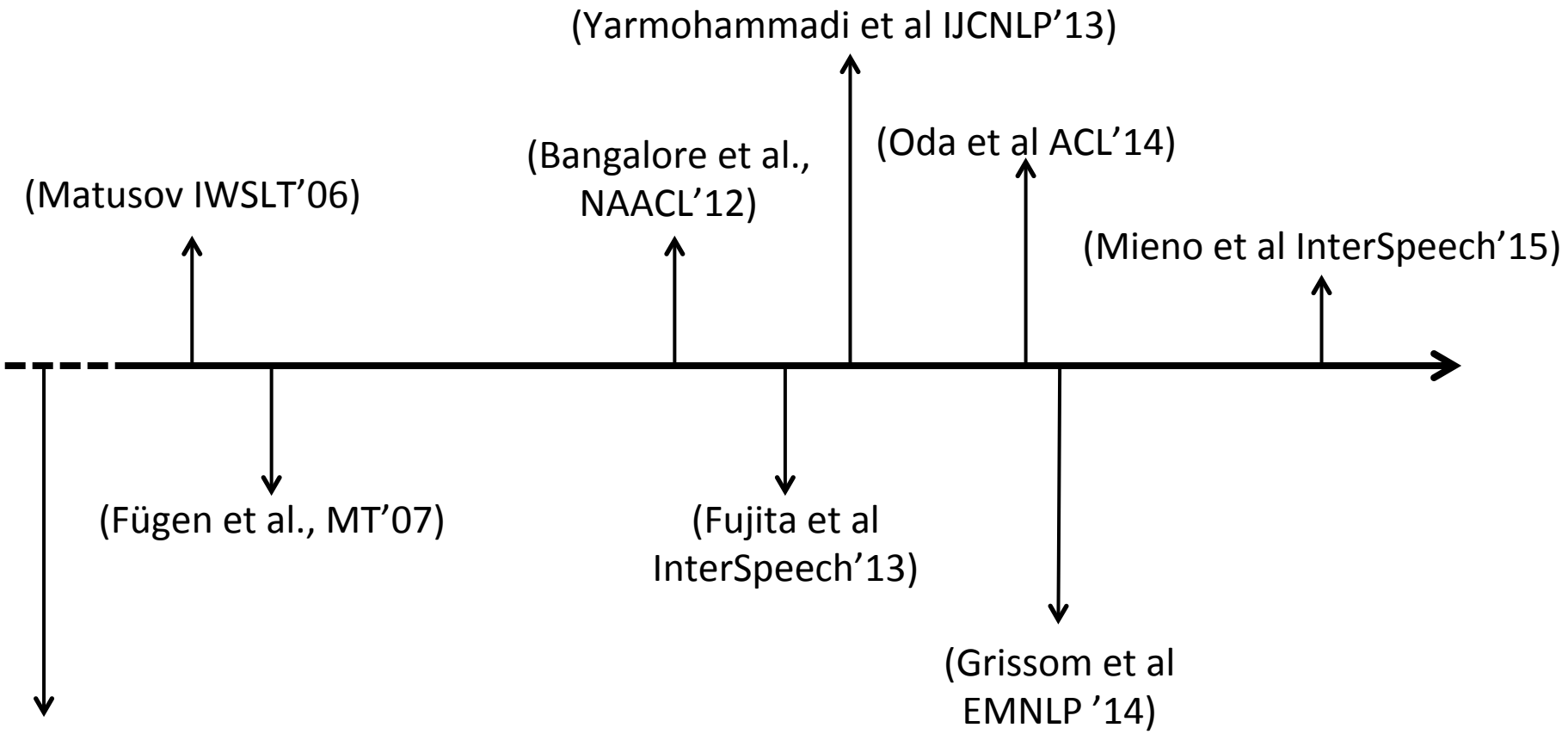
a taxi to the airport please → **0.8 sec** → un taxi al aeropuerto por favor

(lower fluency and latency)

Do not segment when reordering is required

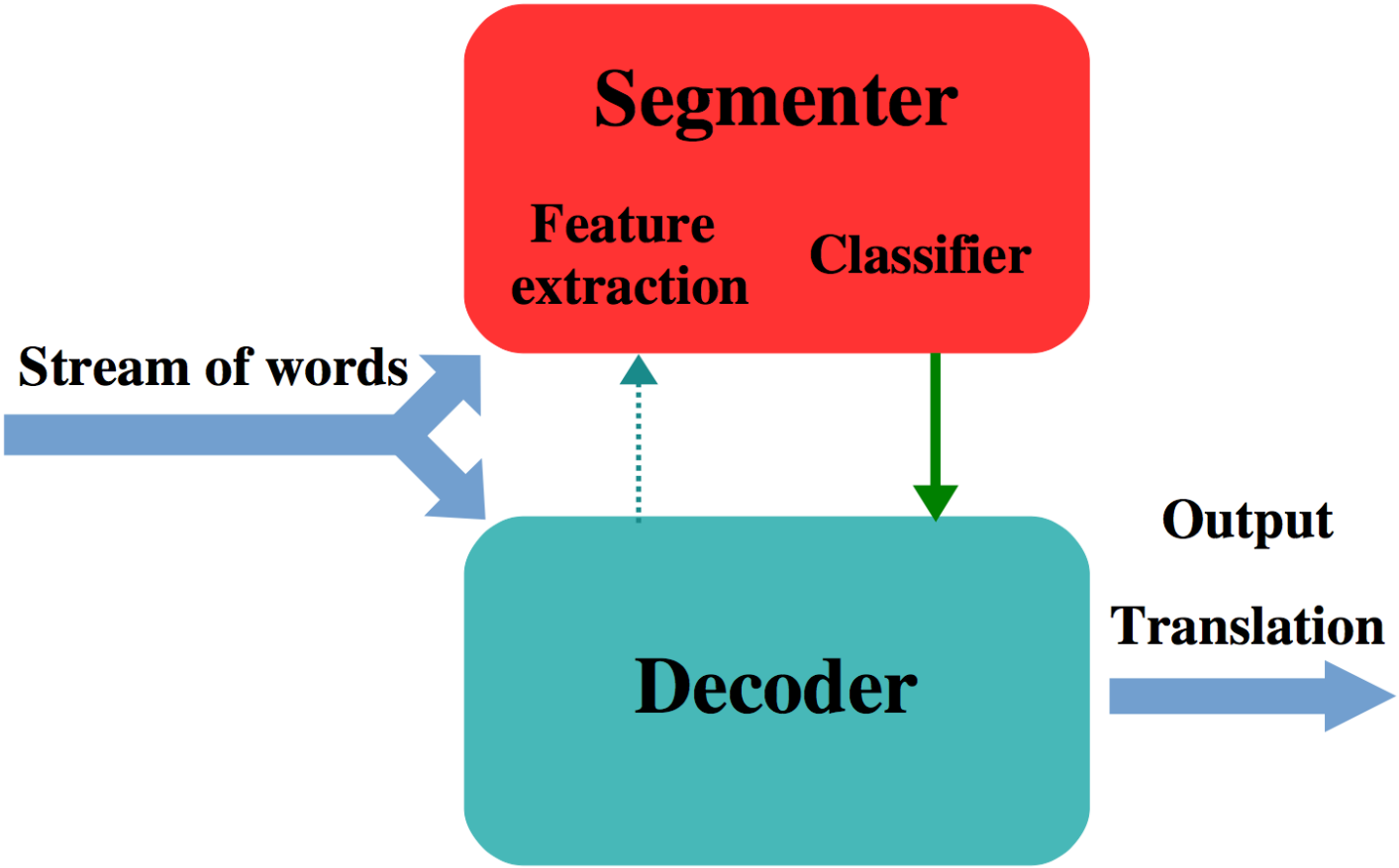


Sentence Segmentation



Simultaneous Translation using GNF SCFGs

(Siahbani et al, SLT 2014)



Hiero decoding made possible using LR-Hiero

Train the Segmenter

- Produce word alignment for training data (GIZA++)
- Find all monotone phrase pair boundaries
- Make sure phrase pairs are long enough (phrases of length 3 or more)
- Find a suitable set of features to classify segment boundaries
- Train a **classifier** to recognize segment boundaries

Datasets

- Train the segmenter:
  - IWSLT 2011 shared task: English-French TED talks
- Train the translation model:
  - EuroParl v7 plus IWSLT 2011 shared task data
- Train the language model:
  - WMT 2011 French data (EuroParl, News Commentary, UN)
- Tuning set and Test set
  - IWSLT 2010 shared task data (`dev2010`, `tst2010`)

Features for segmenter

- Basic
  - Word at segment boundary (punct, conj)
  - Position of boundary
  - Length of segment
- Part of speech
  - Trigram before segment
  - Bi/trigram at end of segment
- Decoder
  - Language model (lm)
  - P(e|f) phrase pair (tm_0)
  - P(f|e) phrase pair (tm_1)
  - Lex(e|f) lexical (tm_3)
  - Lex(f|e) lexical (tm_4)
  - Log-linear model score (c)
- Best performing segmenter F1 = 81.6%
  - Basic + POS + (lm,tm_0,c)

Results

	BLEU	Time (secs)
No segmentation	**25.72**	19.62
With segmentation	24.48	**0.84**

Our Current Work

- Pareto Optimality for balancing speed/latency versus fluency/accuracy
  - Take reordering into account
  - Let the decoder decide
  - The “least worst” BLEU score for different segment lengths may result in varying speed

Questions?

Extra Slides

Rule Extraction

Hiero Rule Extraction

- Search for *sub-phrases* within larger ones
  - Smaller phrases are replaced by non-terminal X

- Estimating rule frequency
  - Uniformly distribute the fractional count to *all* rules extracted from the phrase-pair

LR-Hiero: filtering non-GNF rules

X -> <我们十分 X_1 / we are very much X_1>

X -> <X_1 发生 的 X_2 / X_2 happens in X_1>

X -> <X_1 事情 / what X_1>　　X -> <非洲 地区 / African regions>　　X -> <发生 的 事情 / what happens in>

Constraints:

Length of phrase-pairs (usually 10)

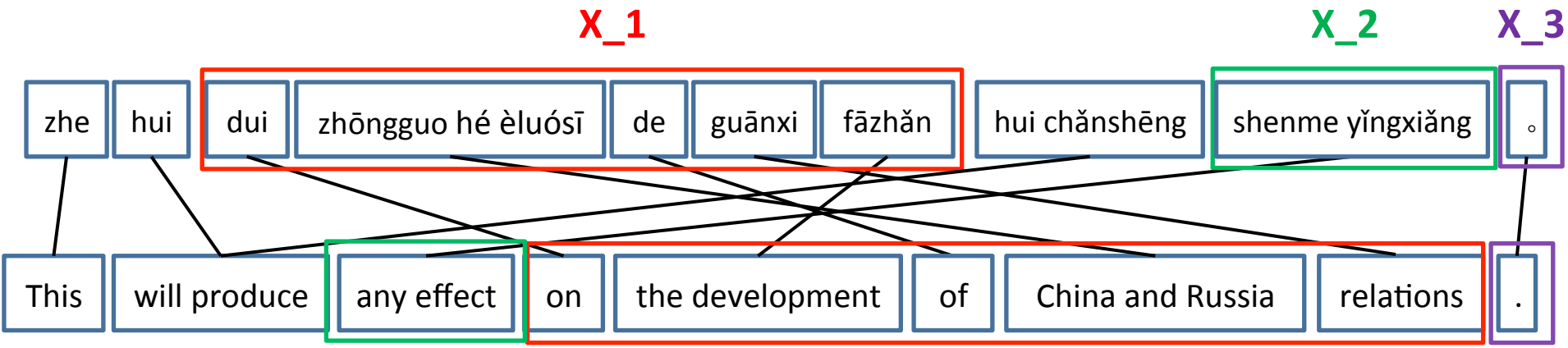
X -> <发生 的 X_1 / X_1 happens in>　　X -> <X_1 的 X_2 / X_2 X_1 in>

At most 2 non-terminals

Non-adjacent non-terminals on the source side

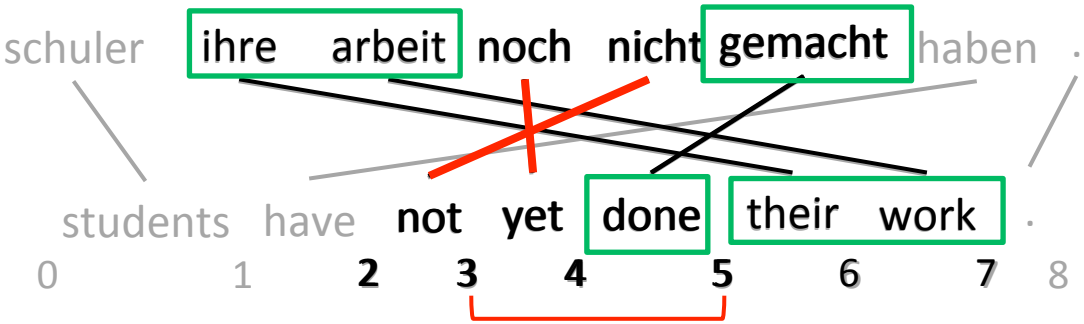
Hiero Rule Extraction: Issues

- Unable to capture all alignments



- Allowing more non-terminals in rules is not practical in CKY decoders LR-Decoding O(n^2)
- Relax constraints (initial phrase length & number of non-terminals)
  - increases the time complexity of rule extraction

New GNF Rule Extraction



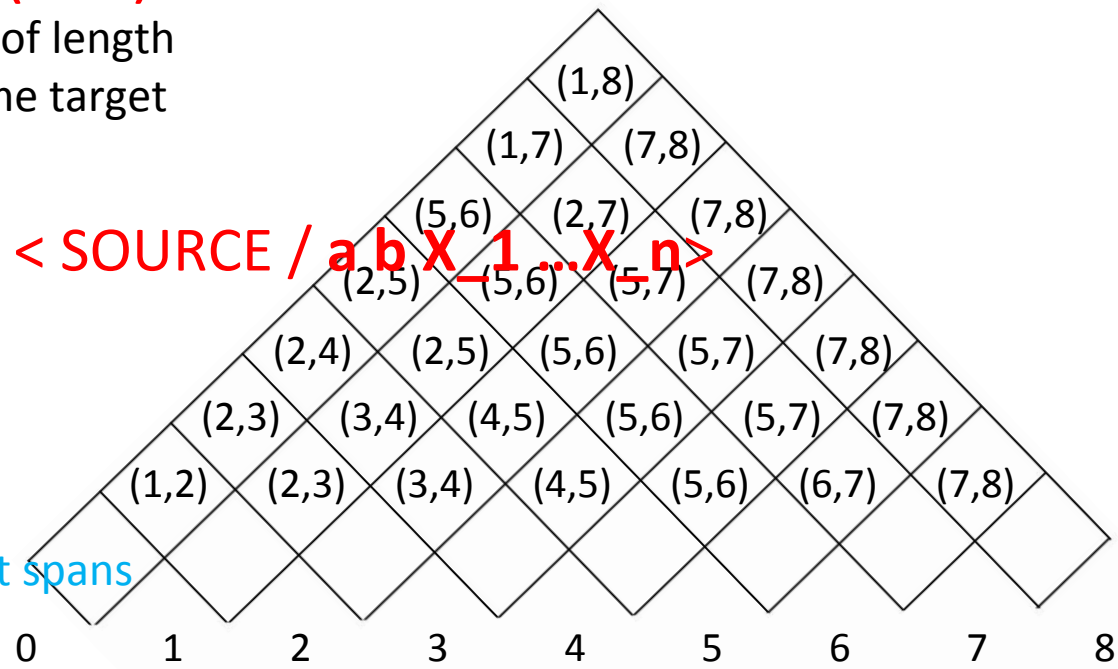
Largest Right Sub-phrase (LRS)

- the longest phrase pair (in terms of length of target side) which share the same target right boundary

LRS(2,5) = [4,5]
LRS(2,7) = [5,7]
LRS(5,7) = [6,7]

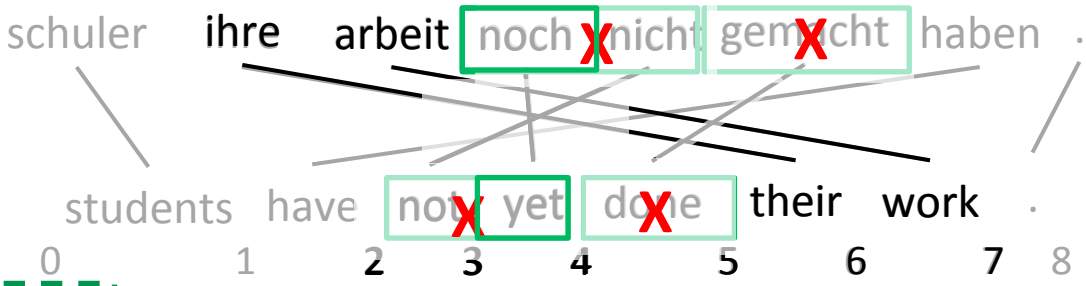
O(m^2)

X-> < SOURCE / a b X_1 ...X_n>



phrase pairs are identified to their target spans

New GNF Rule Extraction



[2,4]

X-> <noch nicht/not yet>

X-> <X_1 nicht/not X_1>

[5,7]

X-> <ihre arbeit/their work>

X-> <ihre X_1/their X_1>

[3,4]

X-> <noch / yet>

[6,7]

X-> <arbeit / work>

[2,5]

X-> <noch nicht gemacht/not yet done>

X-> <noch nicht X_1/not yet X_1>

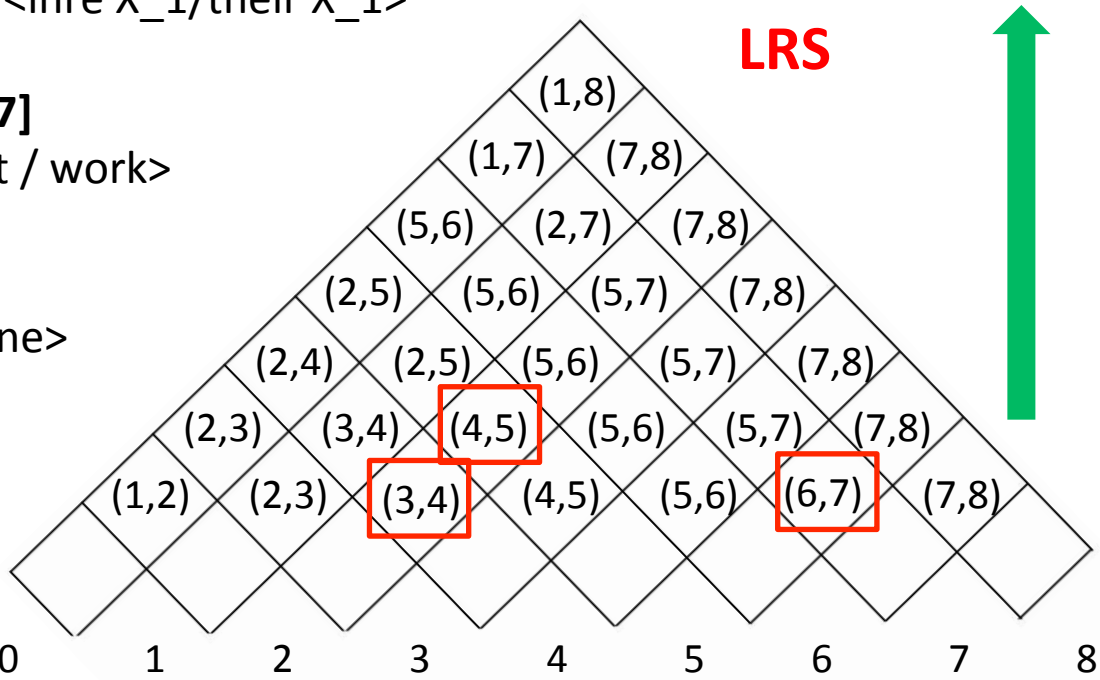
X-> <X_1 nicht X_2/ X_1 not X_2>

X-> <noch nicht X_1/not yet X_1>

X-> <X_1 X_2/ X_1 X_2>

[4,5]

X-> <gemacht / done>



Experiments

Kriya

	old	⌐old
new	306.3 M	74.6 M
⌐new	0	others

Cs-En

	old	⌐old
new	116.0 M	98.8 M
⌐new	0	others

De-En

Czech-English

German-English

Chinese-English

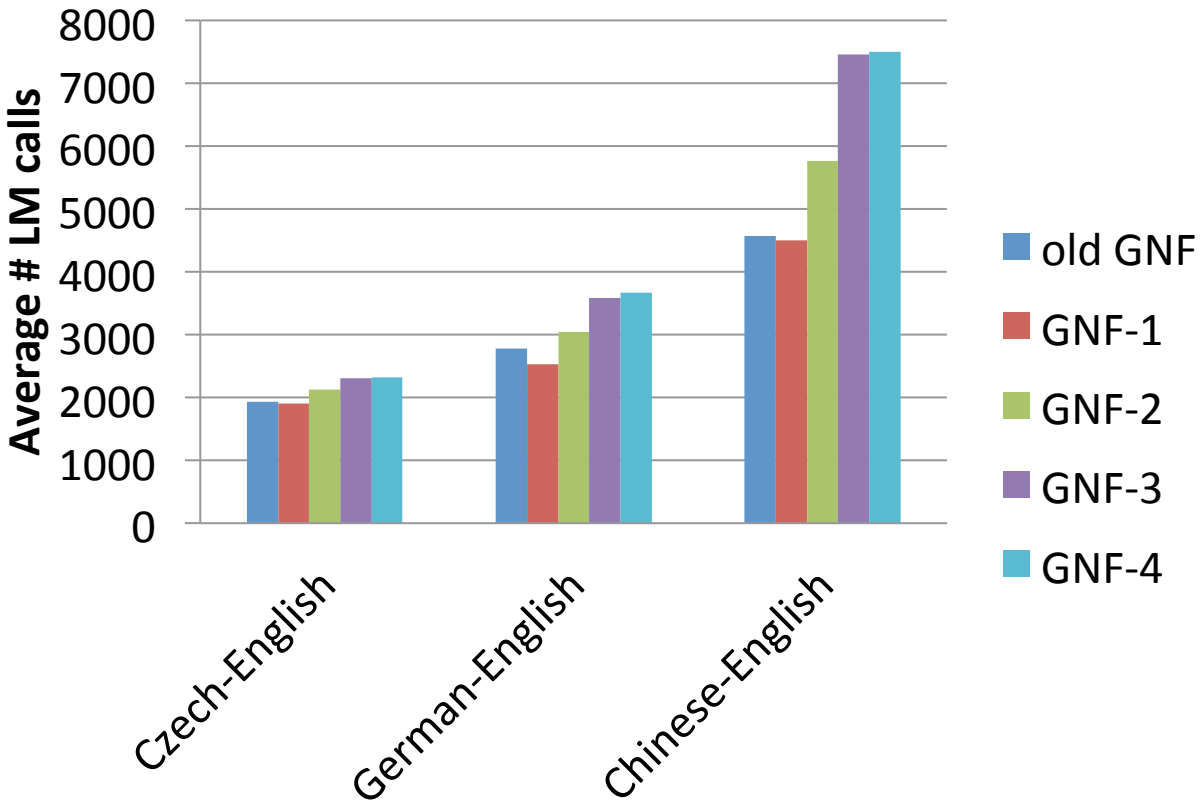
	old	⌐old
new	100.9 M	89.0 M
⌐new	0	others

Zh-En

SCFG & old GNF: initial phrase length 10, maximum source length 7, 2 non-terminals

GNF-4: all phrase pairs, maximum source length 10, 4 non-terminal

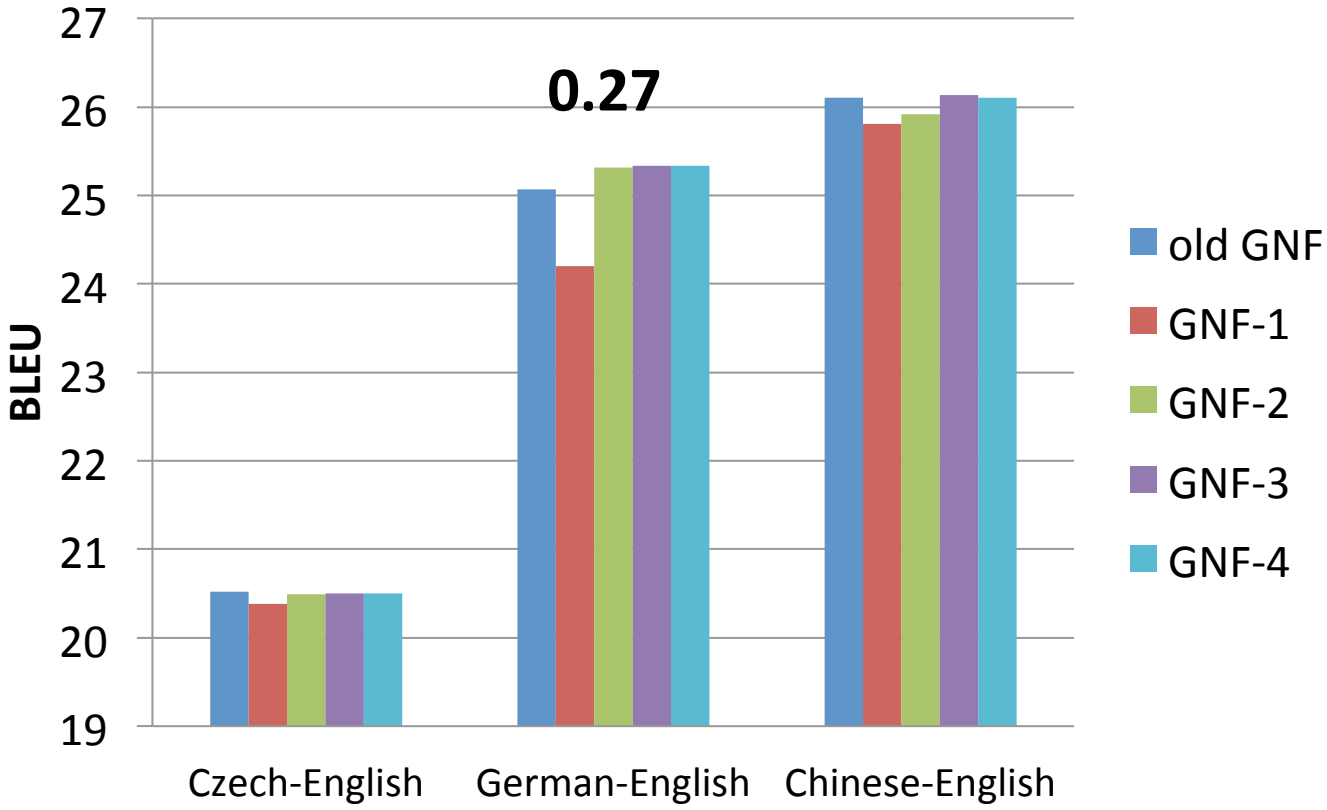
Results (LM calls)



Average number of language model calls on sample set of 50 sentences from testset.

GNF-x (new GNF rules): x non-terminal

Results (BLEU)



old GNF: maximum source length 7, 2 non-terminals

GNF-x (new GNF rules): maximum source length 10, x non-terminal

Alignment Coverage

Model	Czech-English	German-English	Chinese-English
SCFG	318	351	187
old GNF	278	300	132
GNF-4	306	**375**	163

Number of sentences (devset) covered in forced decoding mode

Conclusion and Future Directions

- A new algorithm for GNF rule extraction
- Sentence level GNF rules in LR-Hiero:
  - Improve alignment coverage
  - Marginally affects decoding speed
- Rules with more non-terminals are less frequent
- Elaborate features for rules with more than 2 non-terminals

Questions?

Experiments

- 3 language pairs:
  - Cs-En, De-En, Zh-En

	Corpus (train;dev;test)	
Cs-En	Europarl(v7)+CzEng(v0.9); News commentary(nc) 2008&2009; nc 2011	7.95M/3k/3k
De-En	Europarl(v7); WMT2006; WMT2006	1.5M/2k/2k
Zh-En	HK parallel-tex+GALE ph-1; MTC parts 1&3; MTC part 4	2.3M/1928/919

Experiments

- Hiero rule extraction; Kriya (Sankaran et al., 2012)
  - SCFG
  - GNF
- Configuration and settings:
  - Maximum 2 non-terminals
  - maximum source length 7
  - initial phrase length 10
- New GNF extraction:
  - Maximum 4 non-terminals
  - Maximum source length 10
  - All phrase pairs (including sentence level)

Left-to-Right Decoding

X -> <我们十分 X_1 / we are very much X_1>

X -> <关注 X_1 / concerned with X_1>

X -> <X_1 发生 X_2事情 / what happens X_2 X_1>

X -> <的 / in >

X -> <非洲 地区 / African regions >

<s> [0,8]

<s>

LM

<s> we are very much

LM

<s> we are very much concerned with

<s> we are very much concerned with what happens

<s> we are very much concerned with what happens in

[3,5]

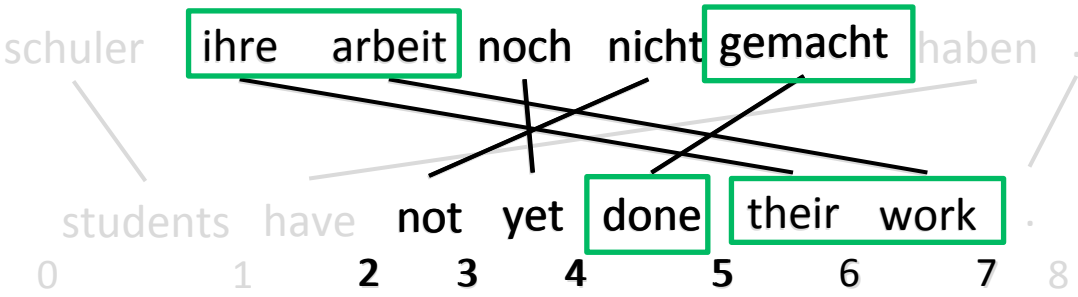
African regions

$$X \rightarrow \langle \gamma, \bar{b}\, \beta \rangle$$

GNF

Non-GNF X -> <X_1 发生 的 X_2 / X_2 happens in X_1>

New GNF Rule Extraction



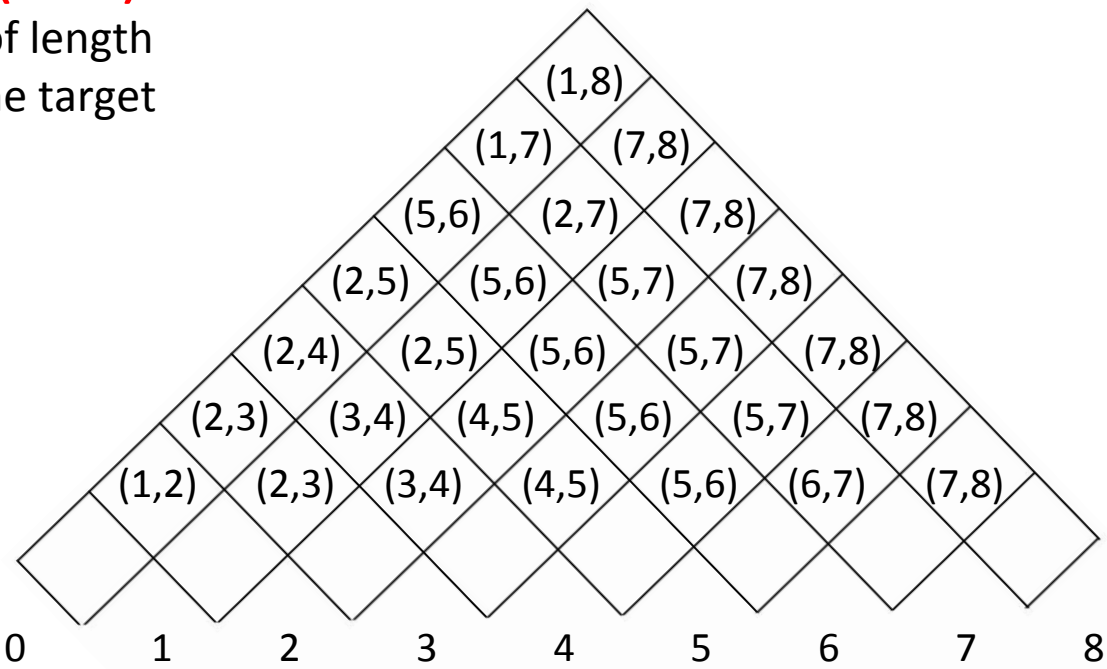
Largest Right Sub-phrase (LRS)

- the largest phrase pair (in terms of length of target side) which share the same target right boundary

LRS(2,5) = [4,5]
LRS(2,7) = [5,7]
LRS(5,7) = [6,7]

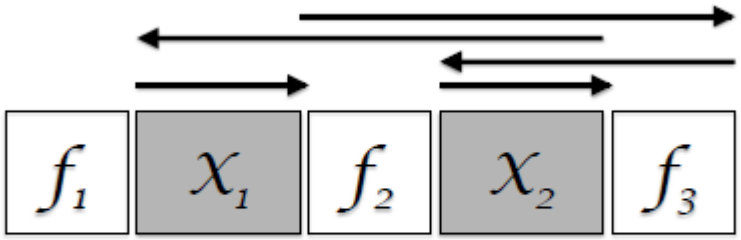
O(m^2)

Sub-problem



completed

Distortion Feature



$$r : \langle f_1 X_1 f_2 X_2 f_3, t X_2 X_1 \rangle \quad I = [\vdash, f_1, f_2, f_3, X_2, X_1, \dashv]$$

$$d(r) = \sum_{j=1}^{k} |I_j^{\mathcal{L}} - I_{j-1}^{\mathcal{R}}|$$

$$r : \langle X_1 \text{ noch nicht } X_2 / \text{not yet } X_2 X_1 \rangle \qquad I = [(1,1),(3,5),(5,6),(1,3),(6,6)]$$

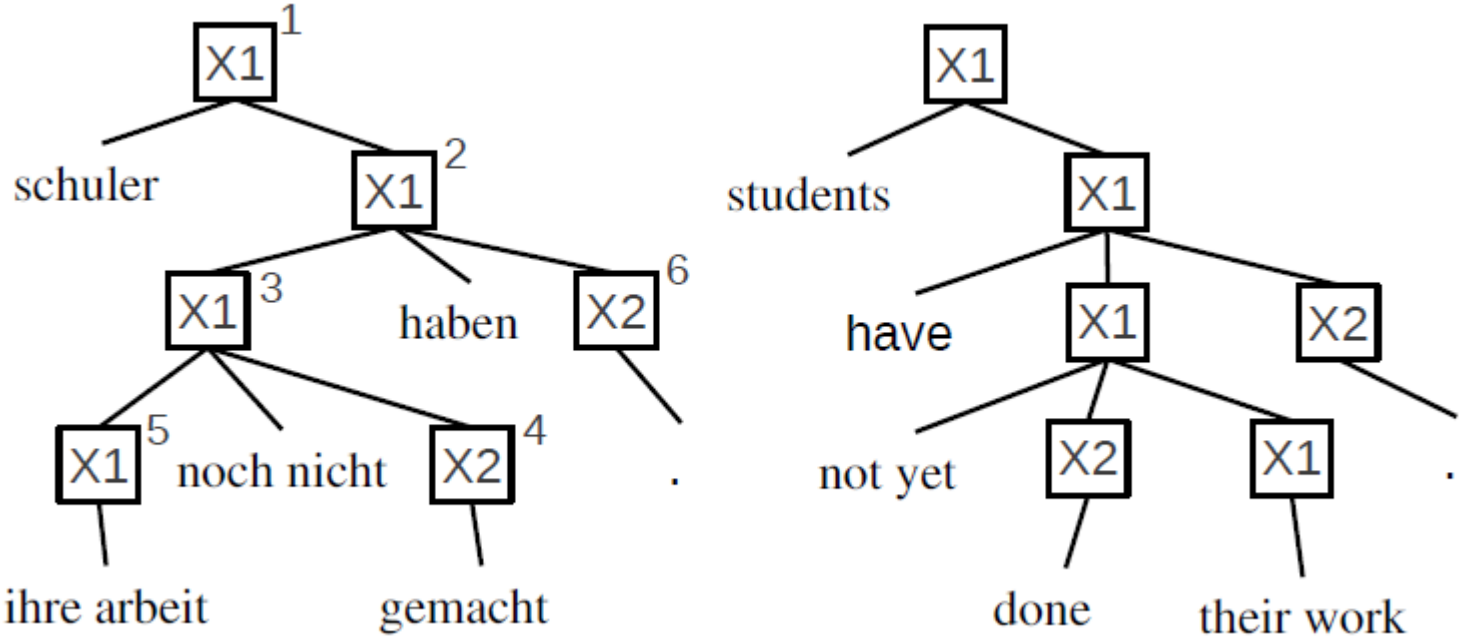
$$_1\text{ihre}_2\,\text{arbeit}_3\,\text{noch}_4\,\text{nicht}_5\,\text{gemacht}_6 \qquad d = |3-1| + |5-5| + |1-6| + |6-3|$$

completed

Reordering feature

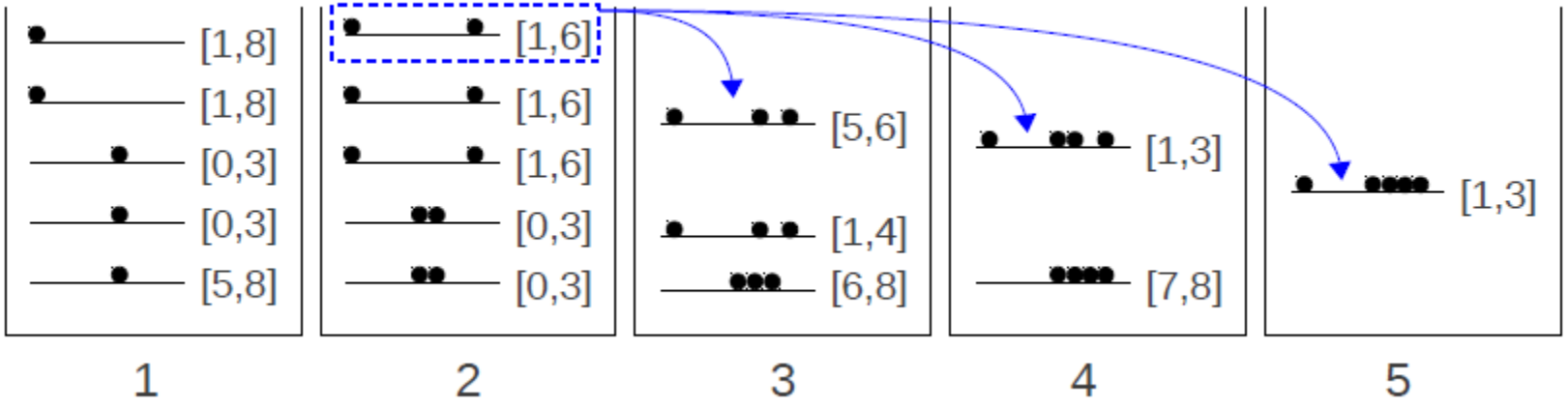
- Number of reordering rules (non-terminals on source and target side are reordered)

$$X \rightarrow \langle X_1 \, heban \, X_2 / have \, X_1 \, X_2 \rangle$$
$$X \rightarrow \langle X_1 \, noch \, nicht \, X_2 / not \, yet \, X_2 \, X_1 \rangle$$



LR-Decoding with Beam Search

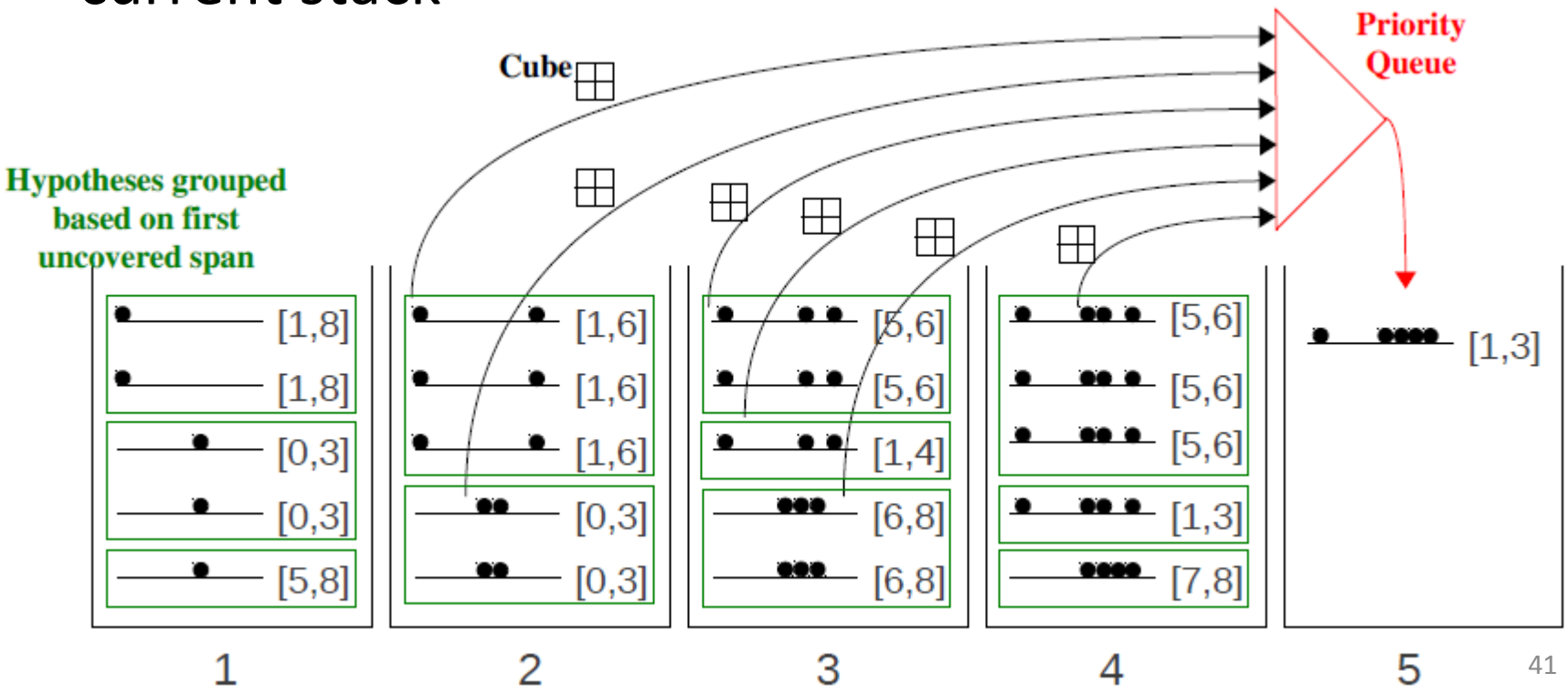
- Stacks: hypotheses with same number of source side words covered
- Exhaustively generating all possible partial hypotheses for a given stack



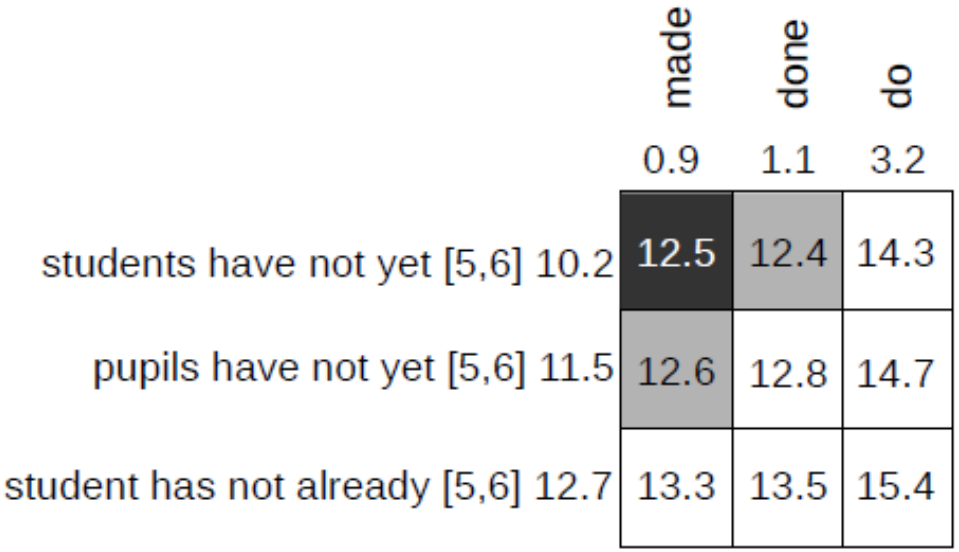
completed

Cube pruning

- each cube: a grouped of hypotheses and applicable rules
- Cubes are fed to a priority queue which fills the current stack

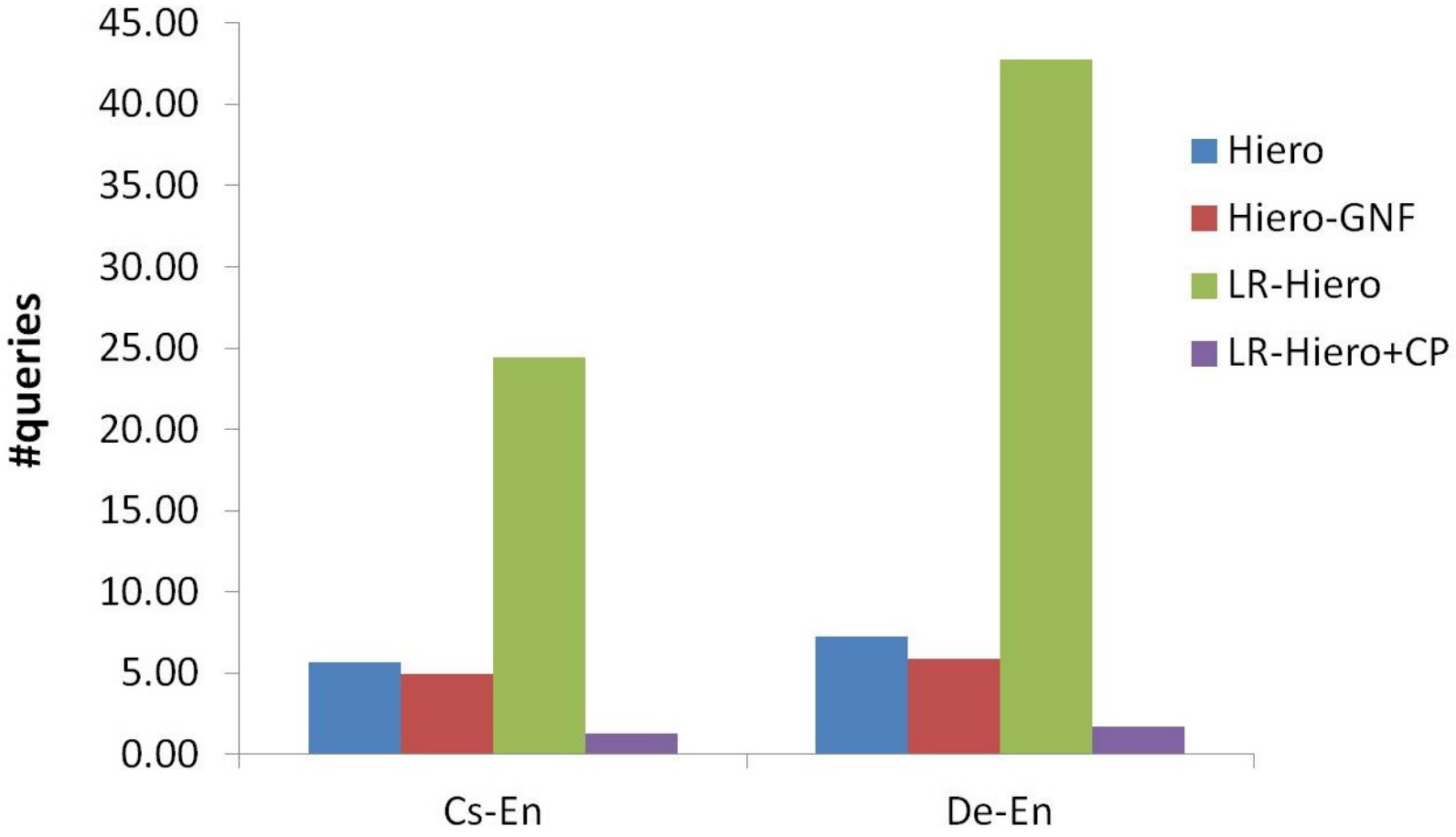


Cube pruning



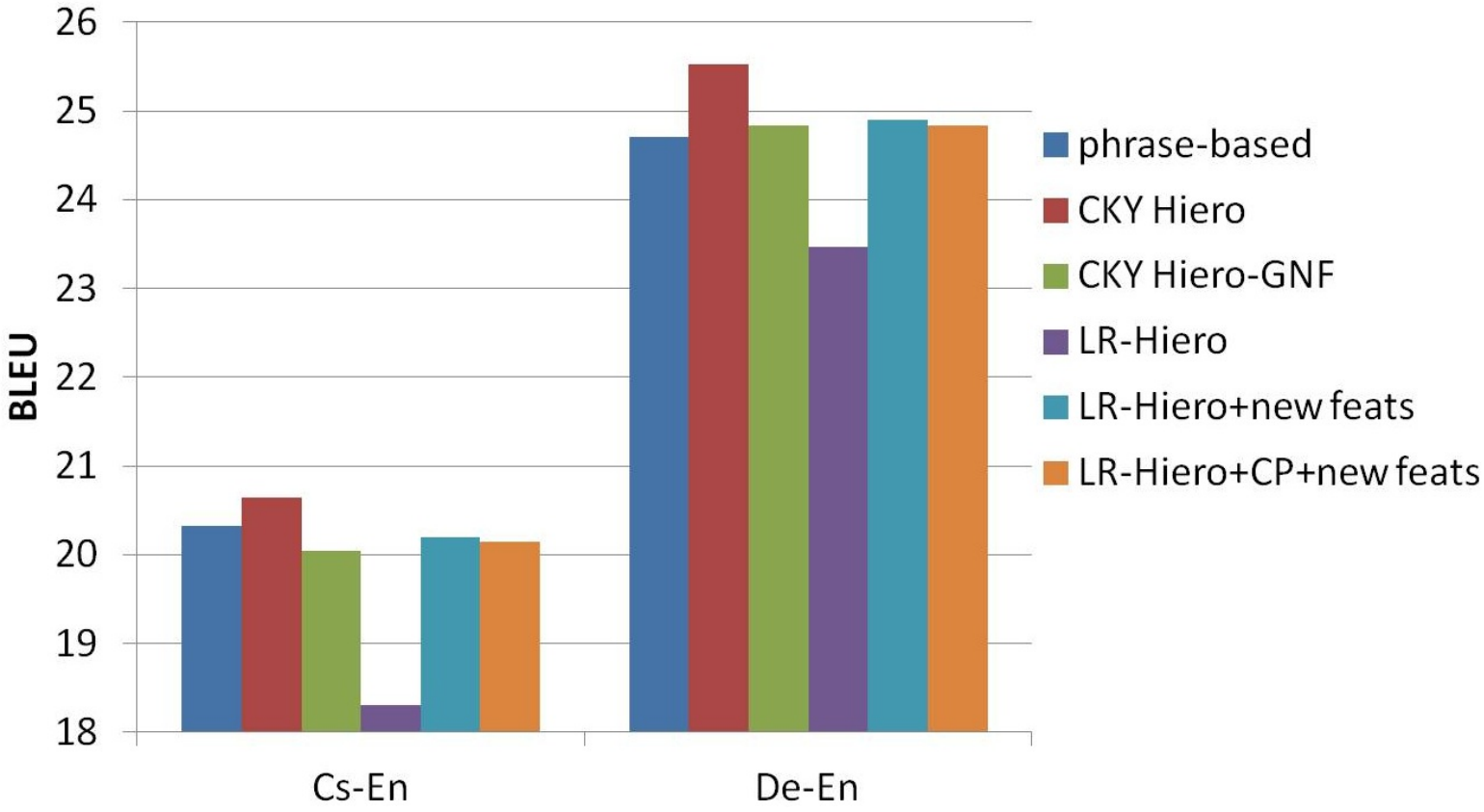
completed

Time efficiency: avg of LM queries



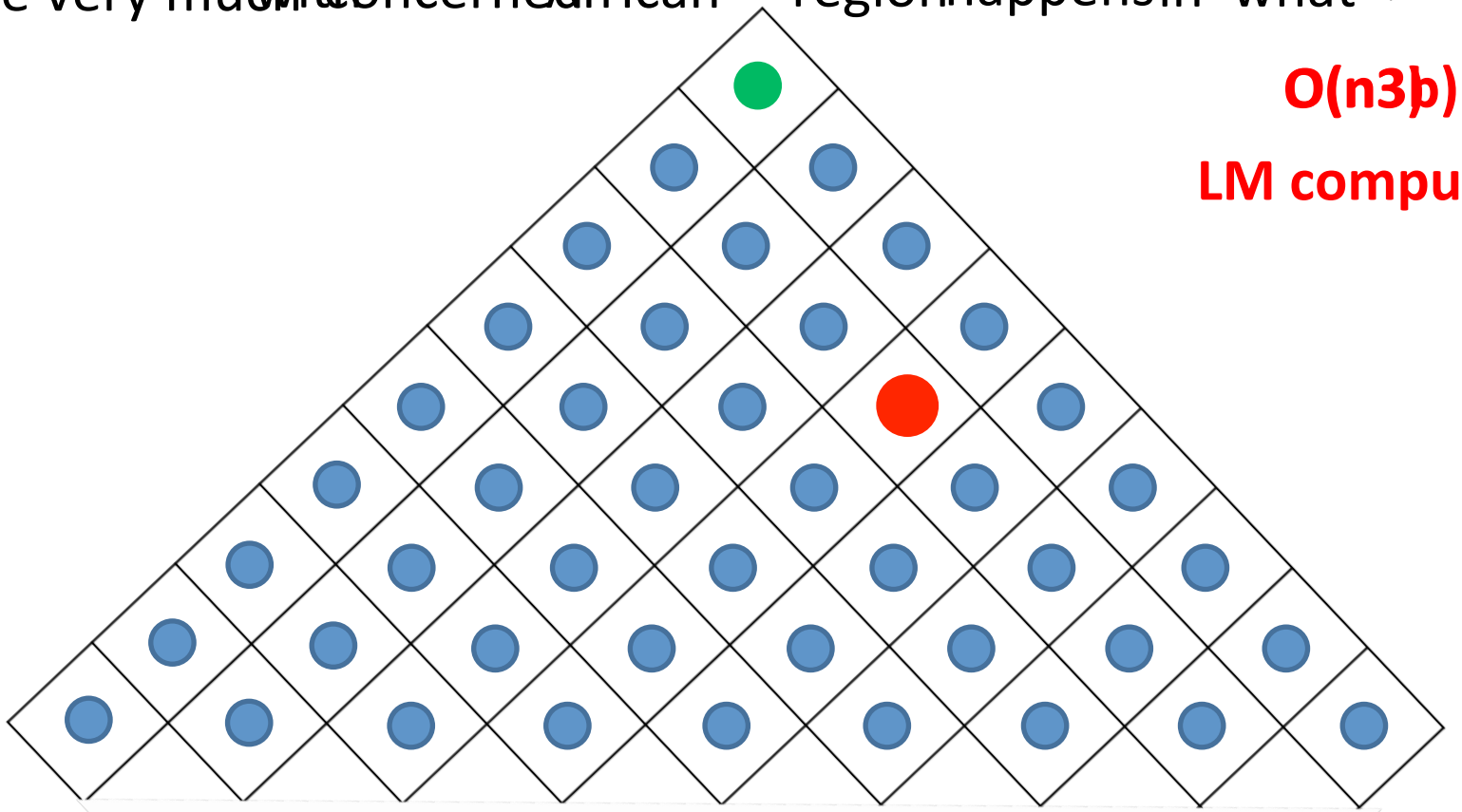
completed

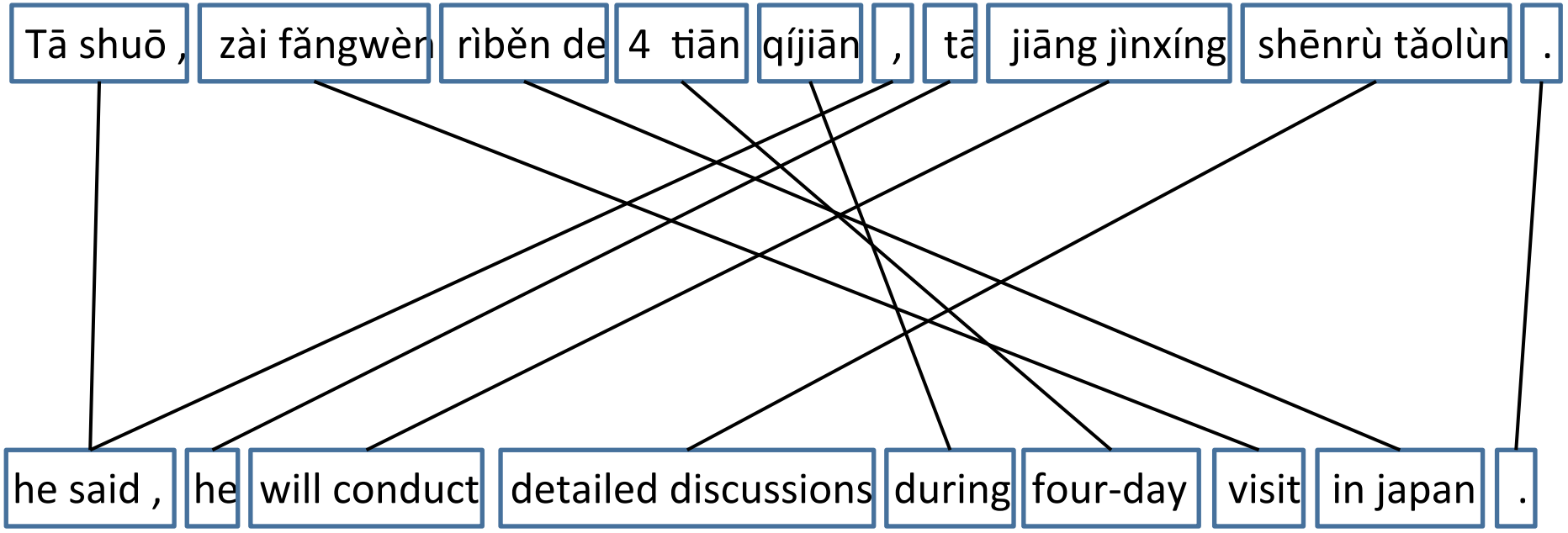
Translation quality



CKY decoding

we are very much concerned with what happens in African regionss .




Tā shuō ,
zài fǎngwèn
rìběn de
4 tiān
qíjiān
,
tā
jiāng jìnxíng
shēnrù tǎolùn
.
he said ,
he
will conduct
detailed discussions
during
four-day
visit
in japan
.