A Tutorial at AMTA 2016

Dependency-Based Statistical Machine Translation

Qun Liu **Liangyou Li**

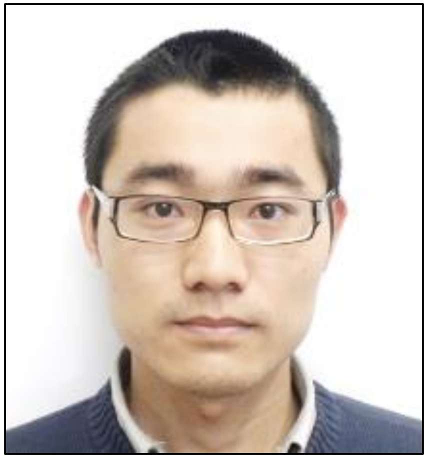
ADAPT Centre
Dublin City University
{qun.liu,liangyou.li}@adaptcentre.ie

28th October 2016, Austin TX USA

Speakers

Qun Liu
- Professor
- Dublin City University
- Chinese Academy of Science

http://computing.dcu.ie/~qliu/



Liangyou Li
- PhD Candidate
- Dublin City University

http://www.computing.dcu.ie/~liangyouli/

Outline

- Introduction
- Dependency-Based MT Evaluation
- Translation Models Based on Segmentation

 Coffee Break

- Translation Models Based on Synchronous Grammars
- Conclusion
- Lab Session

Statistical Machine Translation

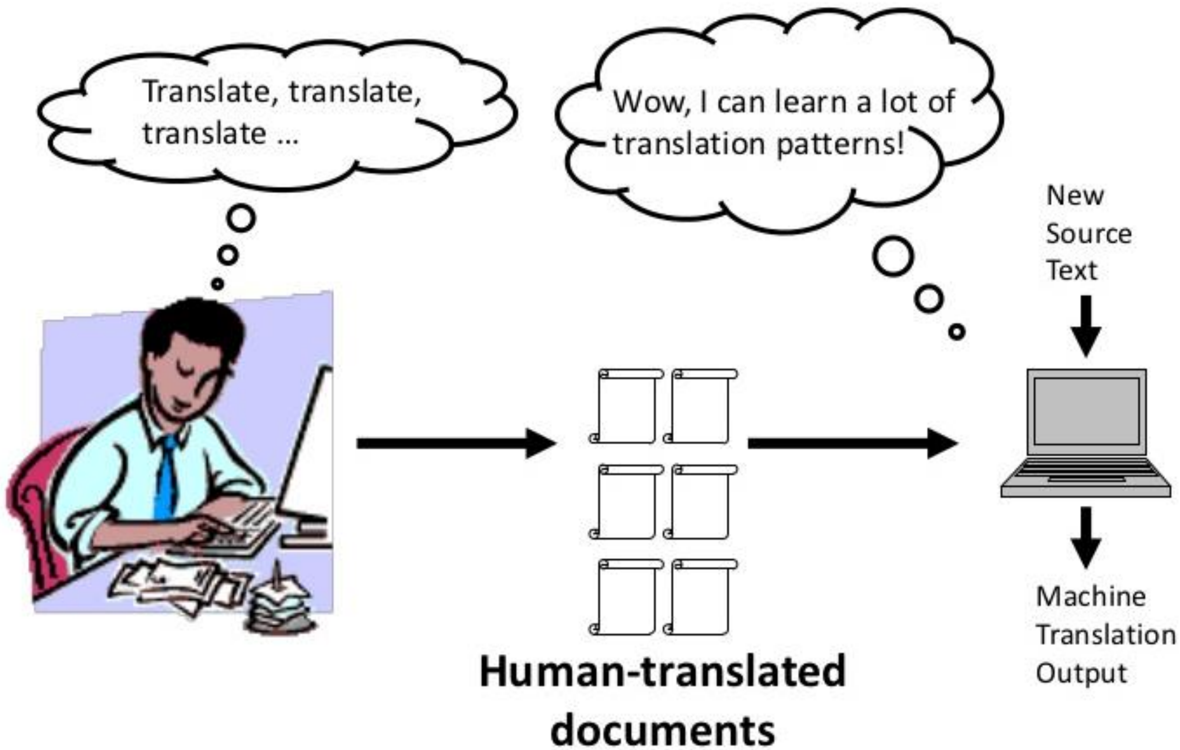
Dependency Structures

INTRODUCTION

Statistical Machine Translation

- What is SMT?
- Advantages of SMT
- Framework of SMT
- SMT Approaches

What is SMT?



SMT is a machine translation paradigm which relies on parallel corpora and machine learning techniques

[Image from TAUS USER CONFERENCE 2010]

Advantages of SMT

- Data driven
- Language independent
- Less dependent on language experts
- Fully automatic
- Fast prototype and deploy

Framework of SMT

- Noisy-Channel Model

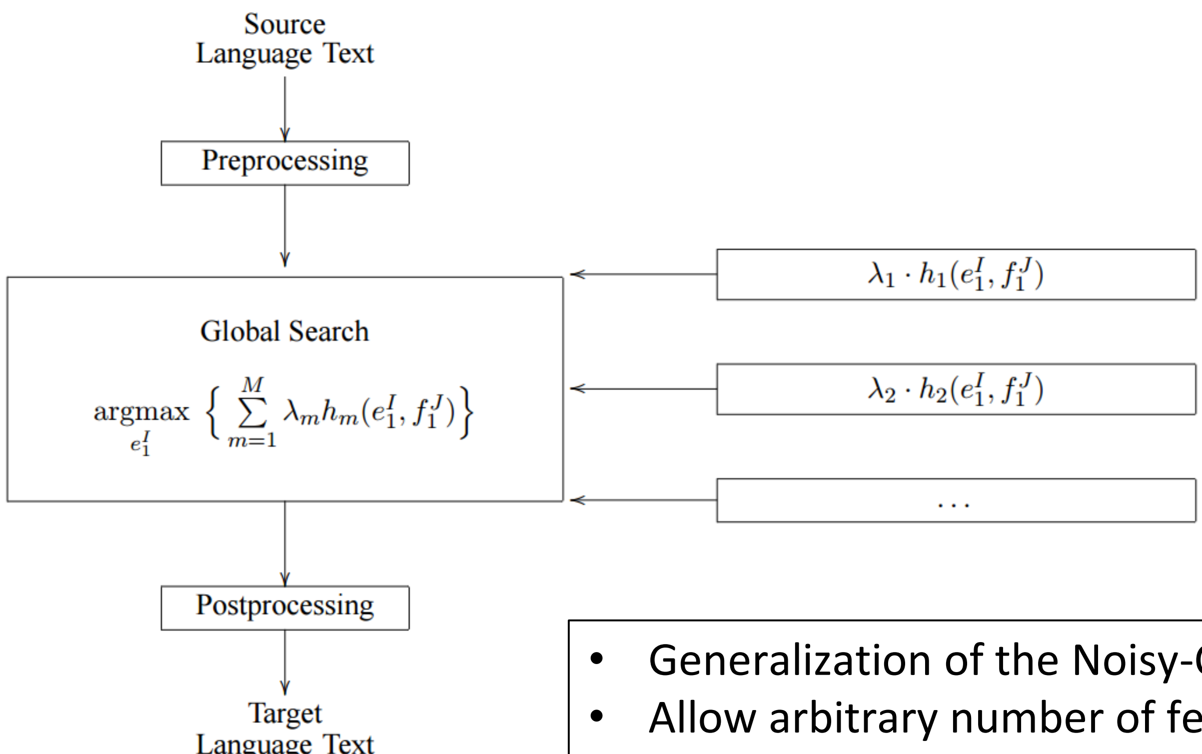
$$t \rightarrow p(t) \rightarrow p(s|t) \rightarrow s$$

$$\begin{aligned} t^* &= \text{argmax}\ p(t|s) \\ &= \text{argmax}\ \underbrace{p(t)}_{\text{Language Model}}\ \underbrace{p(s|t)}_{\text{Translation Model}} \end{aligned}$$

 [Brown 1988,1990,1993]

Framework of SMT

- Log-Linear Model



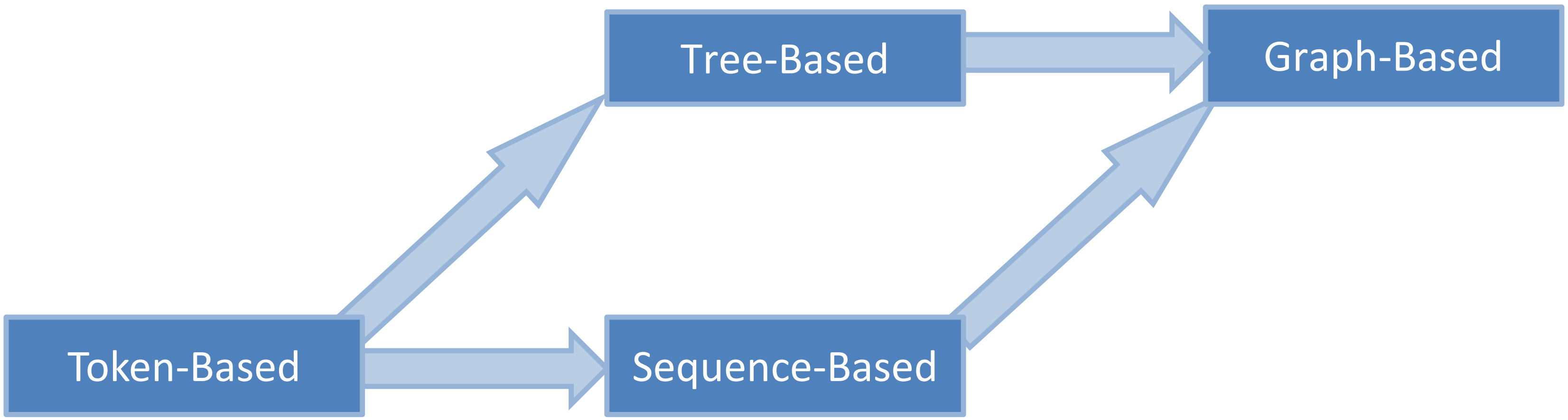
- Generalization of the Noisy-Channel Model
- Allow arbitrary number of features
- Features are tunable

[Och et al., 2002]

SMT Approaches

Dependency Edge
Dependency Path
Dependency Treelet
Hierarchical Phrase-based
Tree-to-String
String-to-Tree
Tree-to-Tree
Dependency-to-String
Tree-to-Dependency
Dependency-to-Dependency

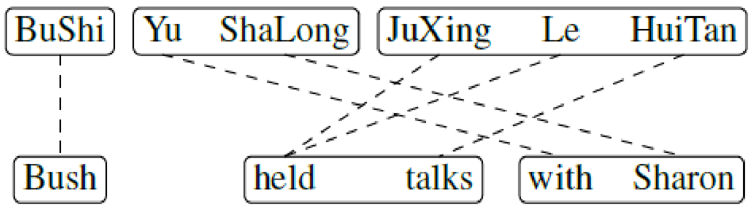
Dependency Graph Segmentation
Dependency Edge Replacement
Dependency Node Replacement



Word-based

Phrase-based

Phrase-Based SMT



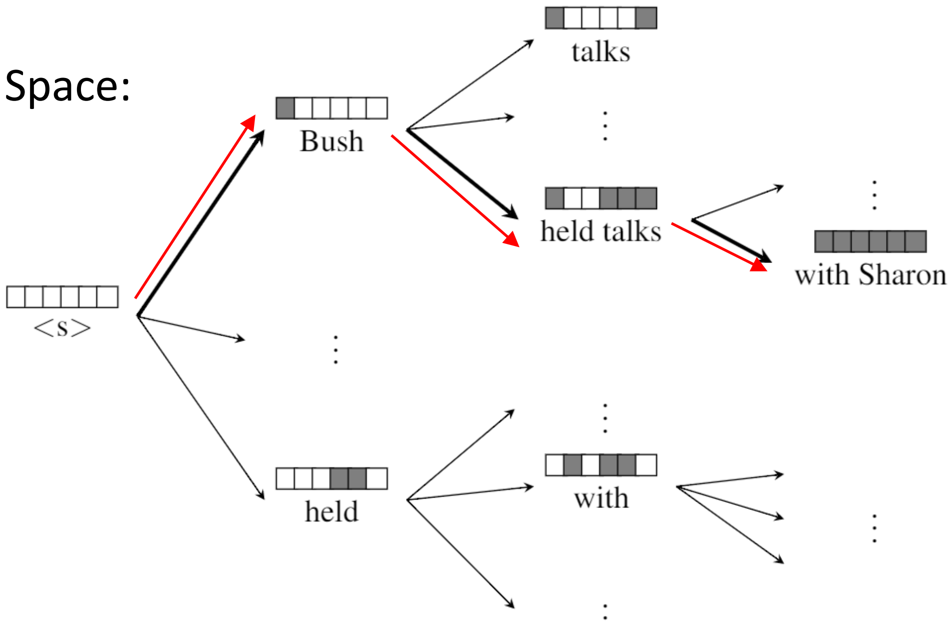
Source Phrase	Target Phrase	Probability
BuShi	Bush	0.5
	president Bush	0.3
	the US president	0.2
BuShi Yu	Bush and	0.7
	the president and	0.3

- Source sentences are segmented into phrases
- Source phrases are translated into target phrases
- Target phrases are reordered

[Koehn, 2003]

Phrase-Based SMT

Search Space:



Beam Search:

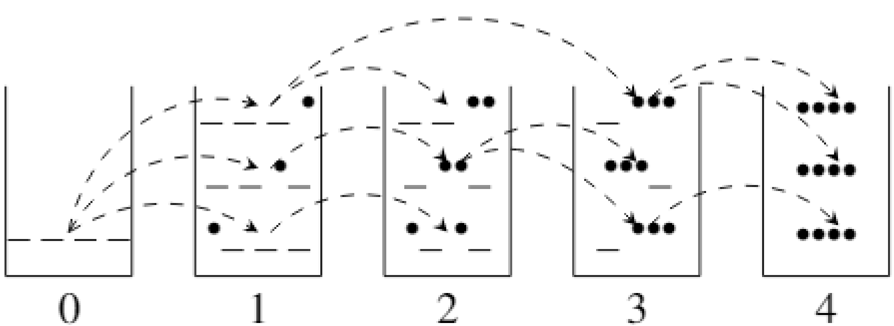


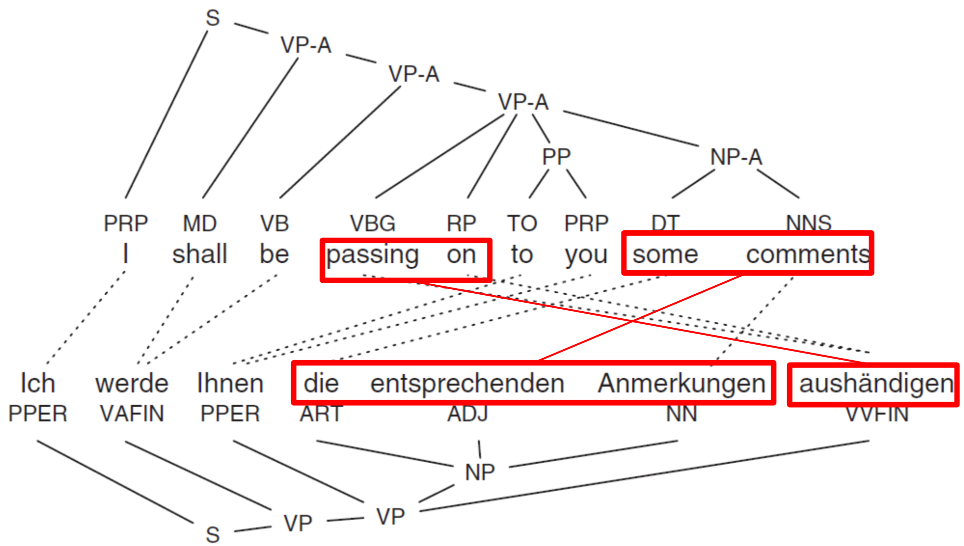
Illustration as in [Liu et al., 2014]

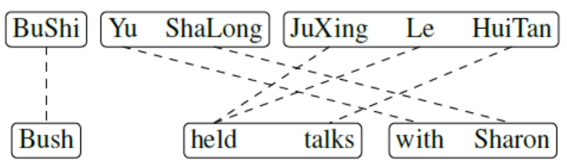
Tree-Based SMT

- Motivation
- Hierarchical Phrase-Based SMT
- String-to-Tree SMT
- Tree-to-String SMT
- Tree-to-Tree SMT
- Forest-Based SMT

Motivation

- Phrase reordering





- Reordering within phrases
- Only continuous phrases

- Generalizations
  - French *ne…pas* to English *not*
  - Chinese *Yu…WuGuan* to English *has nothing to do with*

Syntactic tree from [Koehn, 2010]

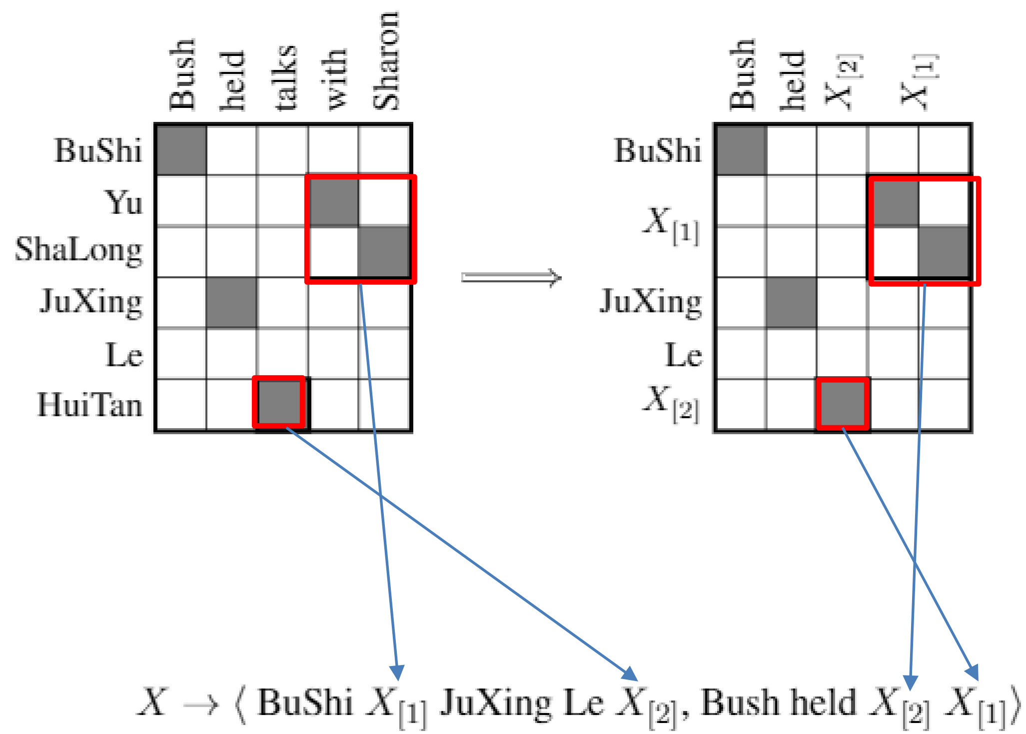
Hierarchical Phrase-Based SMT

- Rule Form

$$X \to \langle \gamma, \alpha, \sim \rangle$$

- Glue Rule

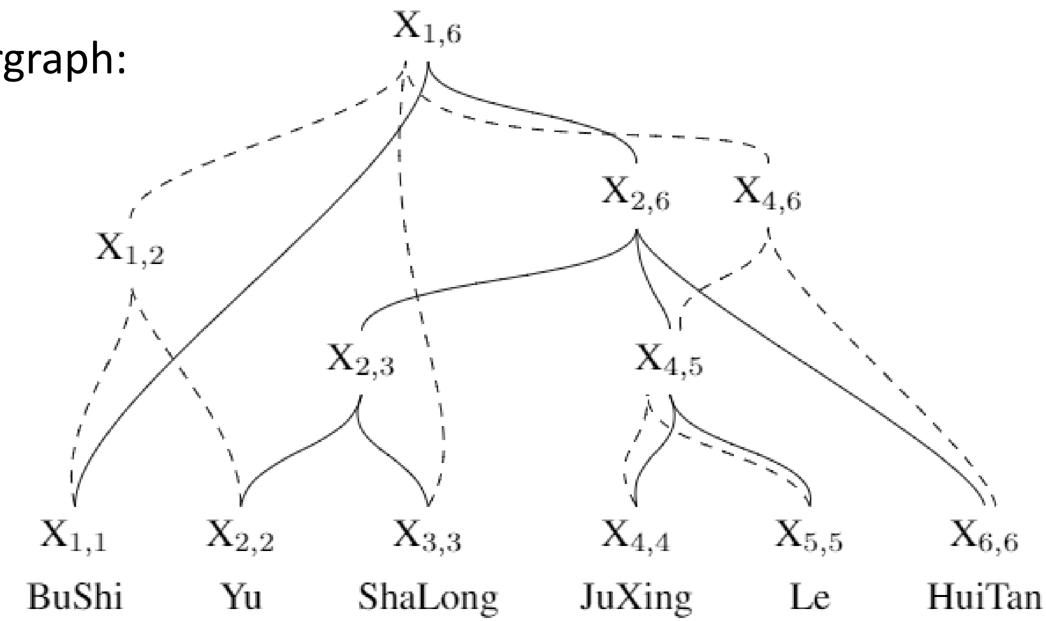
$$S \to \langle S_1 X_2, S_1 X_2 \rangle$$
$$S \to \langle X_1, X_1 \rangle$$



[Chiang, 2005, 2007]

Hierarchical Phrase-Based SMT

Search hypergraph:



Beam:

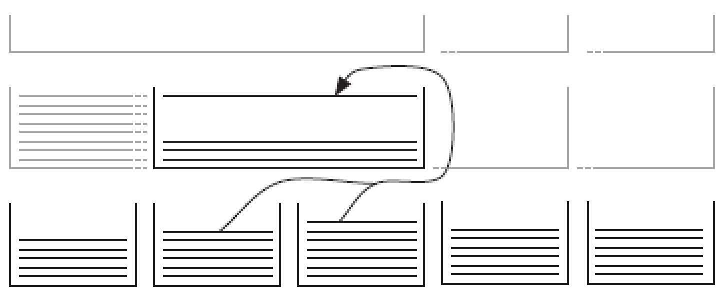
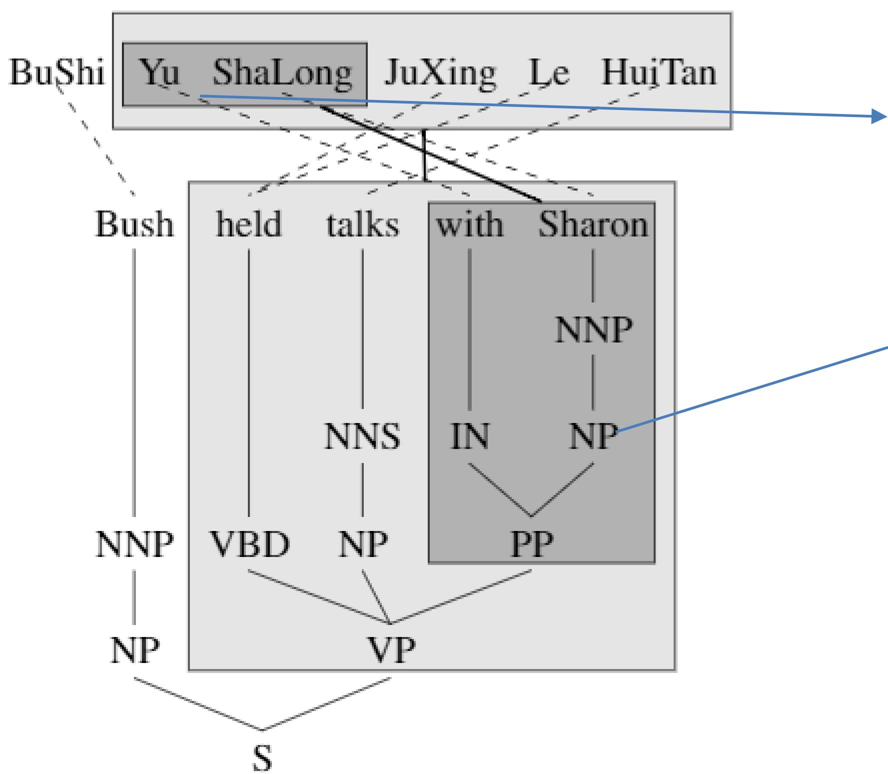
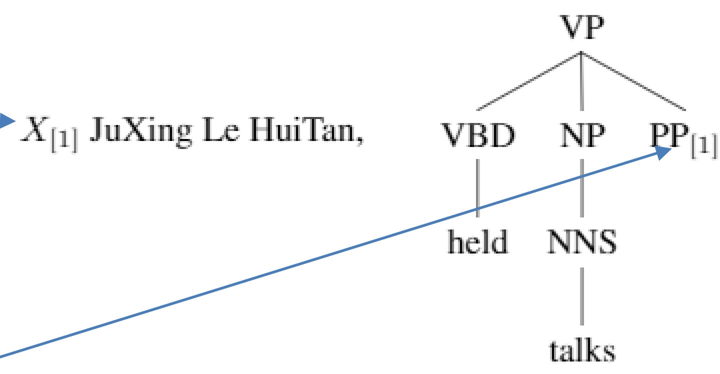


Image from [Koehn, 2010]

String-to-Tree SMT





Training

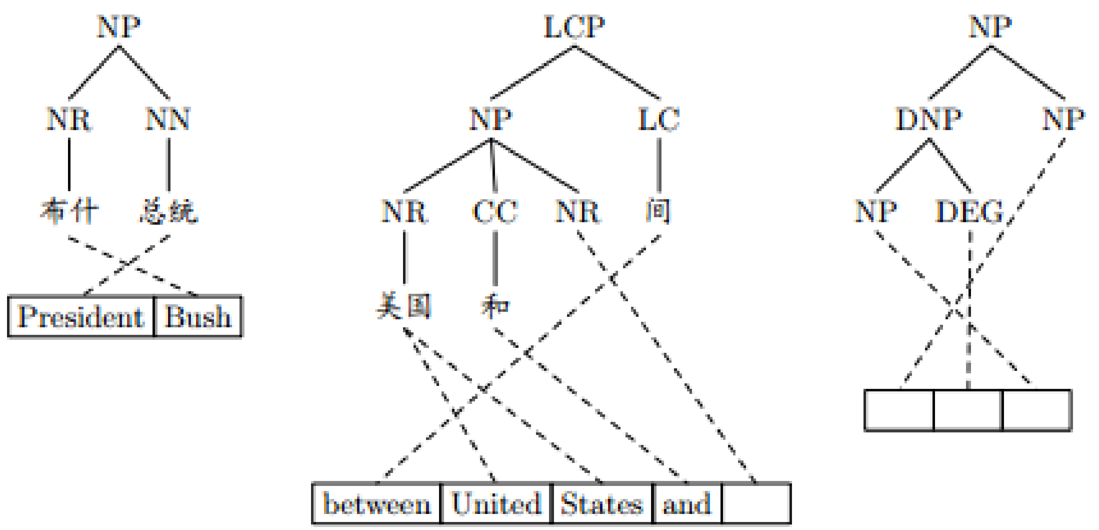
- Target sentences are parsed into trees
- Extract string—tree pairs

Decoding

- Parse source sentences using hierarchical phrases on the source side of rules
- Generate target trees using target subtrees in rules

[Galley et al., 2004, 2006]

Tree-to-String SMT

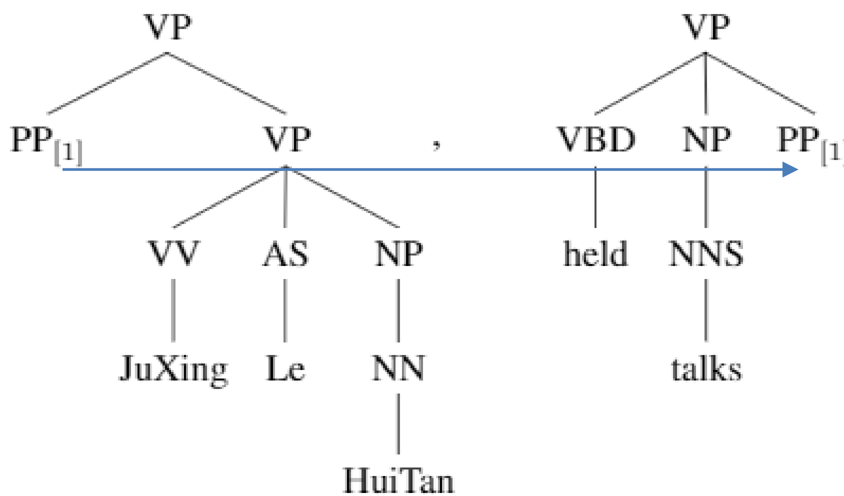


Training
- source sentences are parsed into trees
- Extract tree--string pairs

Decoding
- Parse source sentences beforehand
- Generate target words

[Liu et al., 2006]

Tree-to-Tree SMT



Training
- Source and target sentences are parsed into trees
- Extract tree--tree pairs

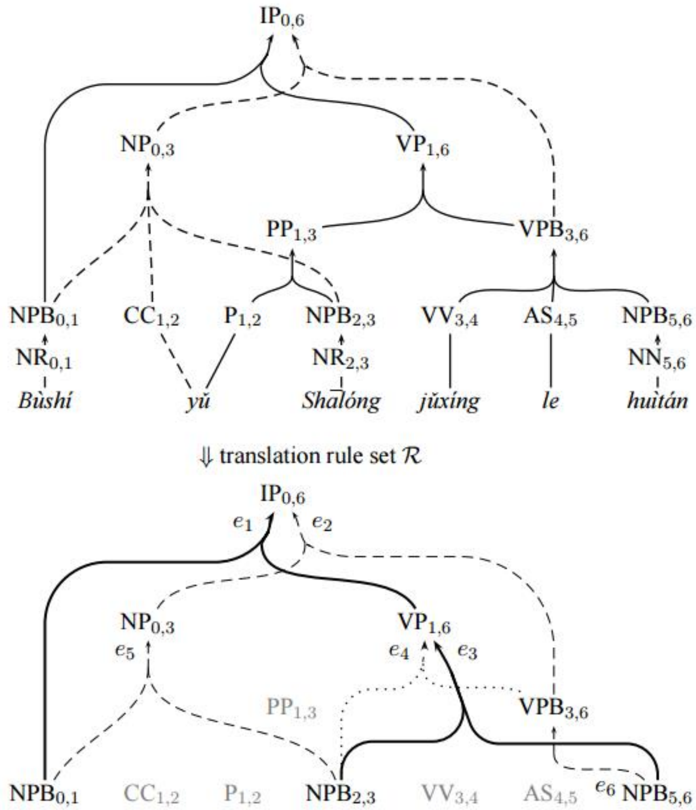
Decoding
- Parse source sentences
- Generate target trees using subtrees in rules

[Zhang et al., 2007]

Forest-Based SMT

(a)

(b)

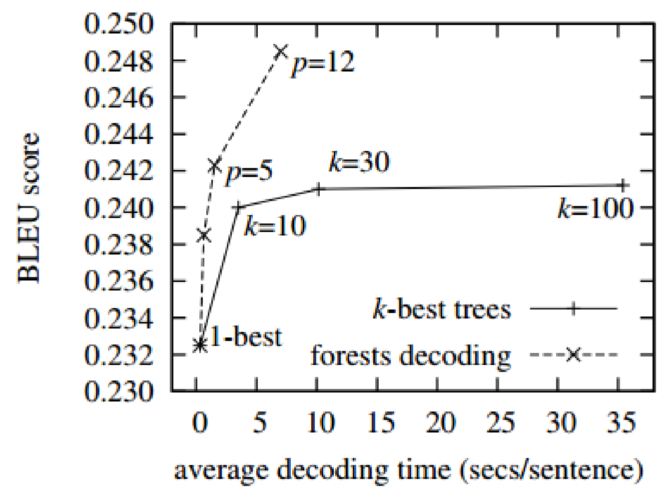


(c)

translation hyperedge		translation rule
e_1	r_1	IP(x_1:NPB x_2:VP) → x_1 x_2
e_2	r_6	IP(x_1:NP x_2:VPB) → x_1 x_2
e_3	r_3	VP(PP(P(*yǔ*) x_1:NPB) VPB(VV(*jǔxíng*) AS(*le*) x_2:NPB)) → held x_2 with x_1
e_4	r_7	VP(PP(P(*yǔ*) x_1:NPB) x_2:VPB) → x_2 with x_1
e_5	r_8	NP(x_1:NPB CC(*yǔ*) x_2:NPB) → x_1 and x_2
e_6	r_9	VPB(VV(*jǔxíng*) AS(*le*) x_1:NPB) → held x_1

approach \ ruleset	TR	TR+BP
1-best tree	0.2666	0.2939
30-best trees	0.2755	0.3084
forest ($p = 12$)	**0.2839**	**0.3149**

Table 1: BLEU score results from training on large data.



[Mi et al., 2008]

Graph-Based SMT

- Semantic Representation
- Semantic-Based SMT

Semantic Representation

- Abstract Meaning Representation (AMR)

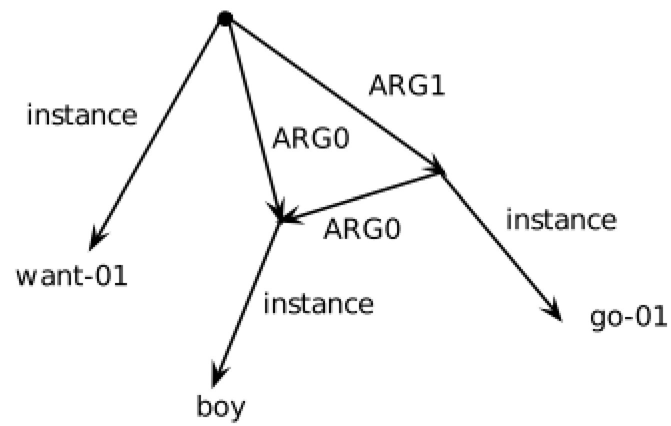
LOGIC format:

$\exists$ w, b, g:
instance(w, want-01) $\wedge$ instance(g, go-01) $\wedge$
instance(b, boy) $\wedge$ arg0(w, b) $\wedge$
arg1(w, g) $\wedge$ arg0(g, b)

AMR format (based on PENMAN):

```
(w / want-01
   :arg0 (b / boy)
   :arg1 (g / go-01
            :arg0 b))
```

GRAPH format:

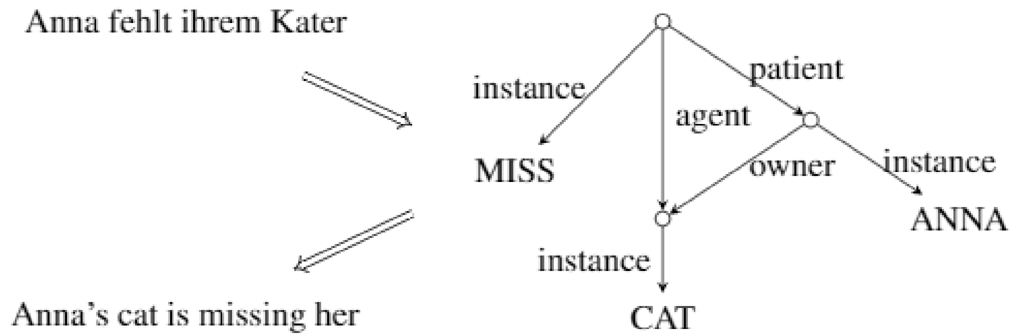


The boy wants to go

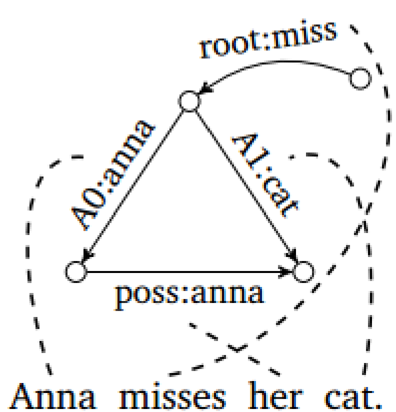
[Laura et al., 2013]

Semantic-Based SMT

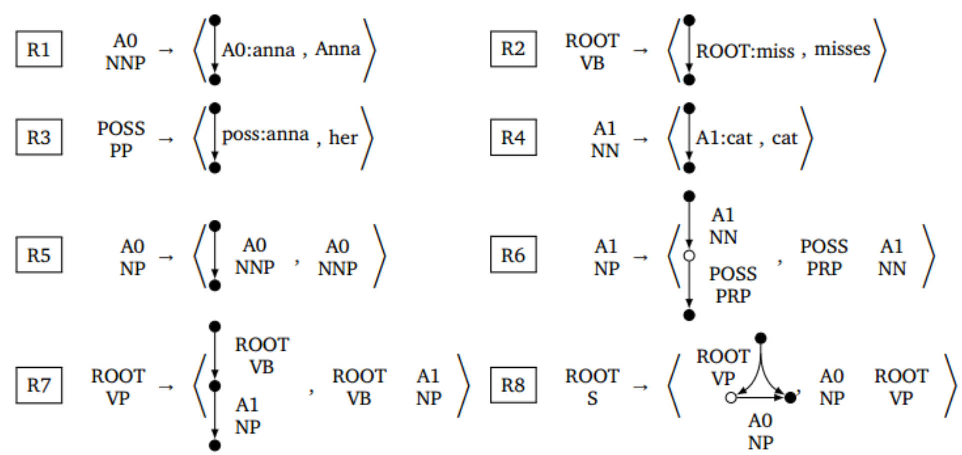
Translation process:



Edge-word alignments:



Rules:

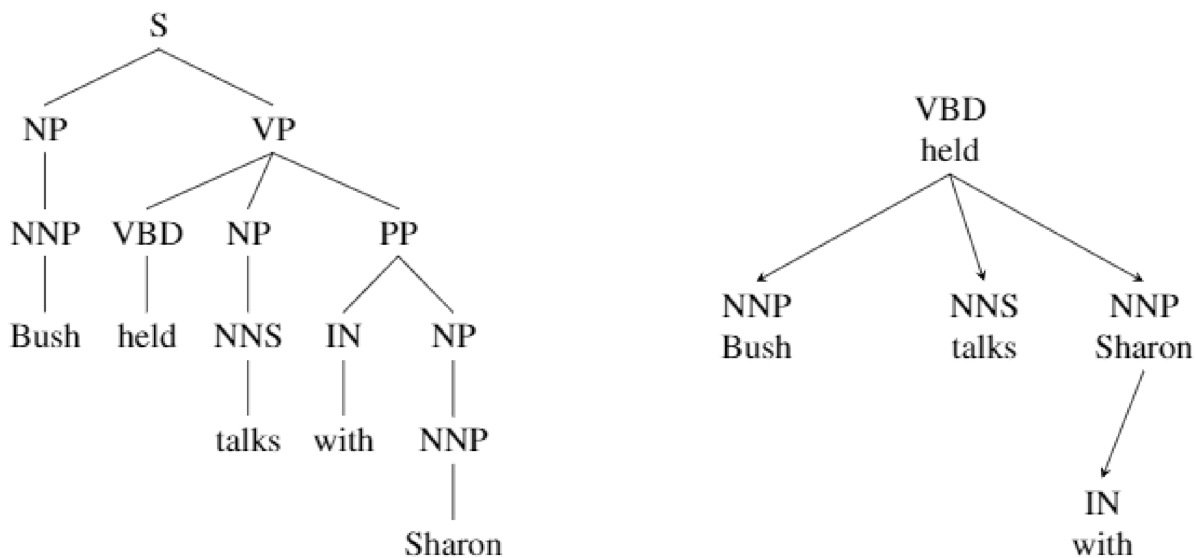


[Jones et al., 2012]

Dependency Structures

- Dependency Tree
- Why Dependency in SMT?

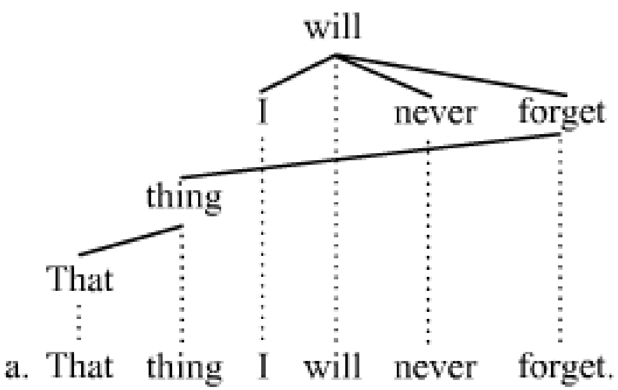
Dependency Tree



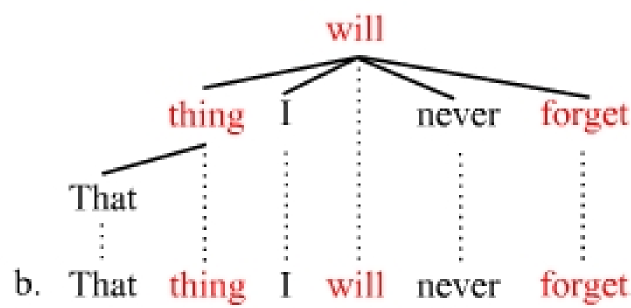
(a) Constituent Tree (b) Dependency Tree.

- Deep vs flat
- Word-node correspondence: one-to-one-or-many vs one-to-one
- Simple in formalism yet having CFG equivalent formal generative capacity [Ding et al., 2004]

Dependency Tree



Non-projective

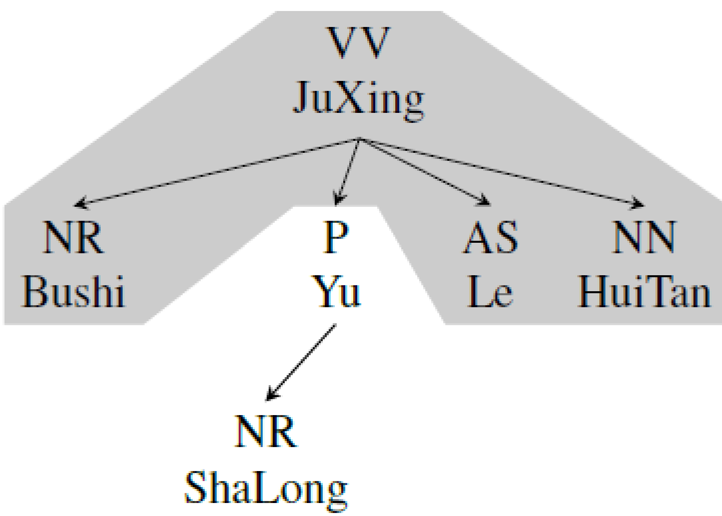


Projective

Image from Wikipedia

Why Dependency in SMT?

- Semantic relation between words
- Best inter-lingual phrase cohesion [Fox, 2002]
- Flexible translation units



Summary

- SMT models benefit from syntactic structures
  – HPB
  – T2S
  – S2T
  – T2T
- Dependency structures have the best inter-lingual phrasal cohesion property

References

- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul Rossin (1990). A Statistical Approach to Machine Translation. In: Computational Linguistics 16.2, pages 76–85.
- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul Roossin (1988). A Statistical Approach to Language Translation. In: Proceedings of the 12th Conference on Computational Linguistics - Volume 1. Budapest, Hungry, pages 71–76.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. In: Computational Linguistics 19.2, pages 263–311.
- David Chiang (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan, USA, pages 263–270.
- David Chiang (2007). Hierarchical Phrase-Based Translation. In: Computational Linguistics 33.2, pages 201–228.
- David Chiang (2012). Grammars for Language and Genes: Theoretical and Empirical Investigations. Springer.
- Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006). Scalable Inference and Training of Context-Rich Syntactic Translation Models. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, pages 961–968.
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu (2004). What's in a Translation Rule? In: Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, USA, pages 273–280.
- Liang Huang, Kevin Knight, and Aravind Joshi (2006a). A Syntax-Directed Translator with Extended Domain of Locality. In: Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing. New York City, New York, pages 1–8.
- Liang Huang, Kevin Knight, and Aravind Joshi (2006b). Statistical Syntax-Directed Translation with Extended Domain of Locality. In: Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts, USA, pages 66–73.
- Bevan Jones, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight (2012). Semantics-Based Machine Translation with Hyperedge Replacement Grammars. In: Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Technical Papers. Mumbai, India, pages 1359–1376.
- Philipp Koehn (2010). Statistical Machine Translation. 1st. New York, NY, USA: Cambridge University Press.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu (2003). Statistical Phrase-Based Translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. Edmonton, Canada, pages 48–54.
- Yang Liu and Qun Liu (2010). Joint Parsing and Translation. In: Proceedings of the 23rd International Conference on Computational Linguistics (Volume 2). Beijing, China, pages 707–715.
- Yang Liu, Qun Liu, and Shouxun Lin (2006). Tree-to-string Alignment Template for Statistical Machine Translation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, pages 609–616.
- Daniel Marcu,WeiWang, Abdessamad Echihabi, and Kevin Knight (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, pages 44–52.
- Haitao Mi, Liang Huang, and Qun Liu (2008). Forest-Based Translation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus, Ohio, USA, pages 192–199.
- Franz Josef Och and Hermann Ney (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA, pages 295–302.
- Chris Quirk and Simon Corston-Oliver (2006). The Impact of Parse Quality on Syntacticallyinformed Statistical Machine Translation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, pages 62–69.
- Kenji Yamada and Kevin Knight (2001). A Syntax-Based Statistical Translation Model. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Toulouse, France, pages 523–530.
- Kenji Yamada and Kevin Knight (2002). A Decoder for Syntax-Based Statistical MT. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA, pages 303–310.
- Min Zhang, Hongfei Jiang, Aiti Aw, Sun Jun, Sheng Li, and Chew Lim Tan (2007). A Tree-to-Tree Alignment-based Model for Statistical Machine Translation. In: Proceedings of Machine Translation Summit XI. Copenhagen, Denmark, pages 535–542.

Q&A

MT Evaluation Introduction

Human Evaluation

Automatic Evaluation

Dependency-Based Evaluation

DEPENDENCY-BASED MT EVALUATION

Introduction of MT Evaluation

Goal: evaluate translation performance of SMT systems

- Meaning preserved
- Grammatically correct

Difficulty: no single right answer

这个 机场 的 安全 工作 由 以色列 方面 负责 .
Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Image from [Koehn, 2010]

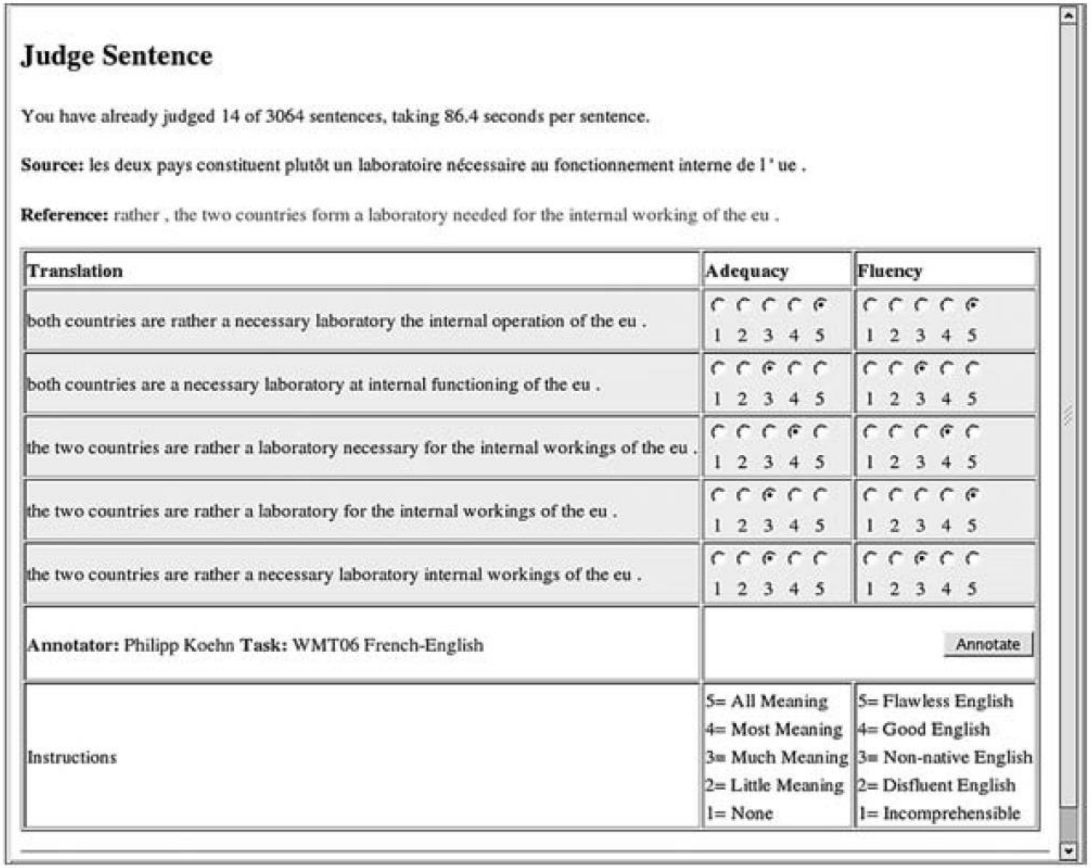
Direct Human Evaluation

Adequacy: same meaning?

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency: grammatically correct?

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible



[Graham et al., 2016]
Tables and images from [Koehn, 2010]

Rank-Based Human Evaluation

Human Evaluation

- Time-consuming
- expensive: e.g. professional translator?
- unrepeatable: precious human labor cannot be simply re-run
- low-agreement: both inter and intra judgement.
  – e.g. WMT11 EN-CZ task, multi-annotator agreement kappa value is very low; even the same strings produced by two systems were ranked differently each time by the same annotator [Callison-Burch, et al., 2011]

Automatic MT Evaluation

- Difficulty in automatic evaluation:
  - Language variability, language ambiguity
  - How to evaluate semantic and syntactic quality
- How to evaluate automatic evaluation metrics:
  - Usually calculate the correlation score with human judgements
- We expect:
  - Repeatable: can be re-used whenever we make some changes on SMT systems
  - Fast: minutes or seconds for evaluating 3k sentences vs hours of human labor
  - Cheap: compared with employment of human judges
  - Stable: each time of running, with same score for un-changed output
  - Reliable: give a higher score for better translation output
  - Further benefit: tune system parameters with automatic metrics

Automatic MT Evaluation

- Lexicon-based similarity metrics
  - BLEU [Papineni et al., 2002]
  - TER [Snover et al., 2006]
  - METEOR [Lavie et al., 2007; Denkowski et al., 2011]
- Semantic-based similarity metrics:
  - MEANT/HMEANT series [Lo et al., 2012, 2013]. Use semantic role labelling information, accuracy of labelling drops due to translation errors.
- Syntax-based metrics
  - Constituency structures
  - **Dependency structures**

BLEU

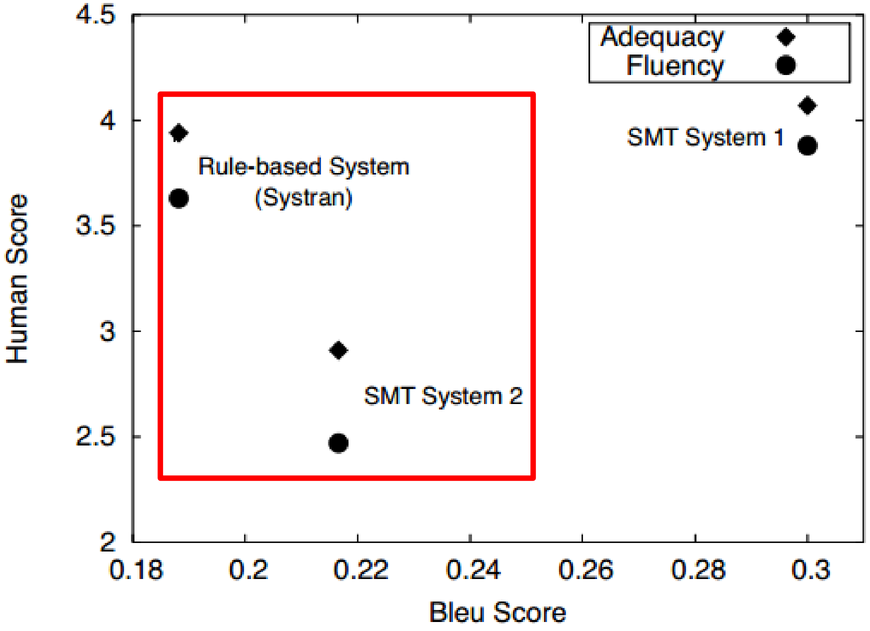
n-gram precision:

$$\text{B}\textsc{leu} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

length penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- Most widely used metric
- Language independent
- Multiple references
- No recall
- Geometric averaging
- Words are equally weighted
- Weak at semantic equivalents
- Document-level



[Papineni et al., 2002; Callison-Burch et al., 2006]

METEOR

- Precision, recall, F-measure
- Alignment and Word-order penalty
- Matching
  - Exact
  - Stem
  - WordNet
  - Paraphrase
- Function words, content words
- Tunable

Dependency-Based Evaluation

- Advantages of dependency structures
- Subtree and head-word chain matching
- Dependency relation matching
- RED metrics
- Parsing as Evaluation
- RNN-based MT evaluation

Advantages of Dependency Structures

- Syntactic equivalents
  - Structures and categories
- Better structures for languages with freer word-order
- Long-distance matching

Subtree And Head-Word Chain Matching

Subtree matching:

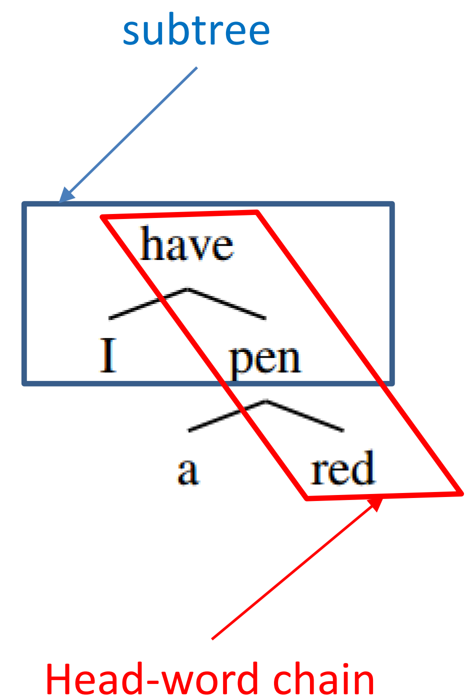
$$\mathrm{STM} = \frac{1}{D} \sum_{n=1}^{D} \frac{\sum_{t \in \mathrm{subtrees}_n(hyp)} \mathrm{count}_{\mathrm{clip}}(t)}{\sum_{t \in \mathrm{subtrees}_n(hyp)} \mathrm{count}(t)}$$

Head-word chain matching:

$$\mathrm{HWCM} = \frac{1}{D} \sum_{n=1}^{D} \frac{\sum_{g \in \mathrm{chain}_n(hyp)} \mathrm{count}_{\mathrm{clip}}(g)}{\sum_{g \in \mathrm{chain}_n(hyp)} \mathrm{count}(g)}$$

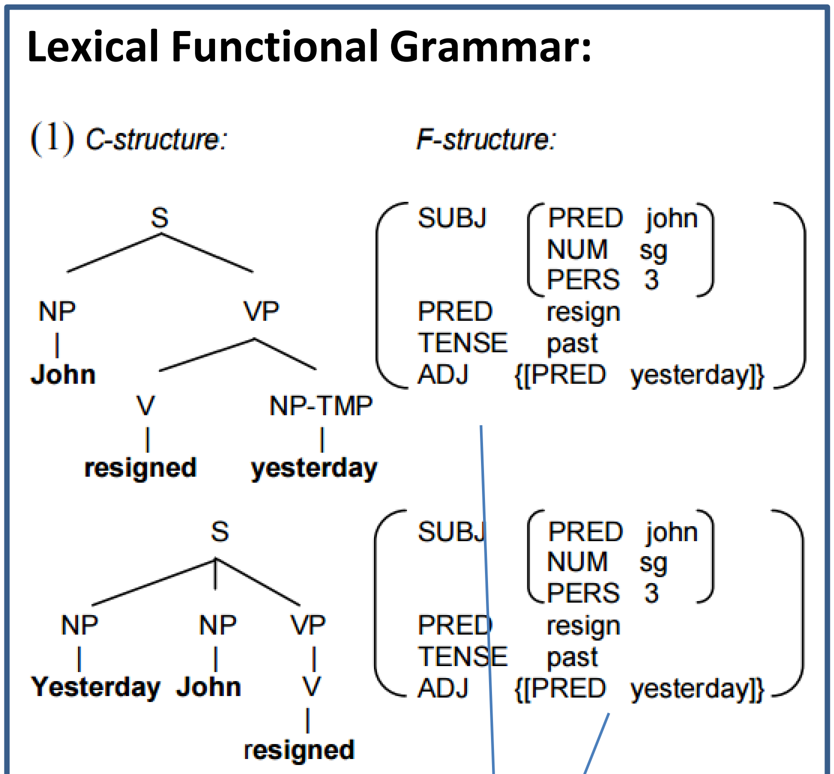
dependency constituent

Max Length/ Depth	BLEU	HWCM	STM	DSTM
1	0.126	0.130	—	—
2	0.132	0.142	0.142	0.159
3	0.117	0.157	0.147	0.150
4	0.093	0.153	0.136	0.121
kernel			0.065	0.090



[Liu et al., 2006]

Dependency Relation Matching



H_FL		H_AC		H_AVE	
d+WN	0.168	M+WN	0.294	M+WN	0.255
d	0.162	M	0.278	d+WN	0.244
d+WN_pr	0.162	NIST	0.273	M	0.242
BLEU	0.155	d+WN	0.266	NIST	0.238
d_pr	0.154	GTM	0.260	d	0.236
M+WN	0.153	d	0.257	GTM	0.230
M	0.149	d+WN_pr	0.232	d+WN_pr	0.220
NIST	0.146	d_pr	0.224	d_pr	0.212
GTM	0.146	BLEU	0.199	BLEU	0.197
TER	-0.133	TER	-0.192	TER	-0.182

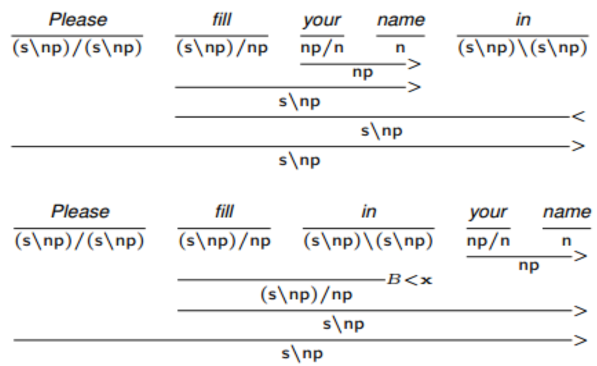
Table 5. Pearson's correlation between human scores and evaluation metrics. Legend: d = dependency f-score, _pr = predicate-only f-score, M = METEOR, WN = WordNet, H_FL = human fluency score, H_AC = human accuracy score, H_AVE = human average score.[9]

subj(resign, john), pers(john, 3), num(john, sg)
tense(resign, past), adj(resign, yesterday)
pers(yesterday, 3), num(yesterday, sg)

[Owczarzak et al., 2007]

Dependency Relation Matching

CCG



(det name_3 your_2)	(det name_4 your_3)
(dobj fill_1 name_3)	(dobj fill_1 name_4)
(ncmod _ fill_1 in_4)	(ncmod _ fill_1 in_2)
(xcomp _ please_0 fill_1)	(xcomp _ please_0 fill_1)

$\emptyset$	$\overleftarrow{left}$	**'Please'**	$\overrightarrow{right}$	{'fill'}
$\emptyset$	$\overleftarrow{left}$	**'fill'**	$\overrightarrow{right}$	{'in','name'}
{'your'}	$\overleftarrow{left}$	**'name'**	$\overrightarrow{right}$	$\emptyset$

Only parse references

Dependent ordering score (DOS):

- For each head word in the ref
  - For each left dependent
    - If the head appears in the MT output and the dependent is on the left, add value 1
  - Similar process for the right dependents

Final score:

recall in terms of DOS * length penalty

 [Mehay et al., 2007]

RED Metric

- RED: **RE**ference **D**ependency based MT evaluation metric
- Only use reference dependency tree
- Two kinds of reference dependency structures:
  – Head-word chains: capture the long-distance dependency information
  – Fixed and floating structures [Shen et al. 2010]: capture local continuous ngrams

 [Yu et al., 2014]

RED Metric

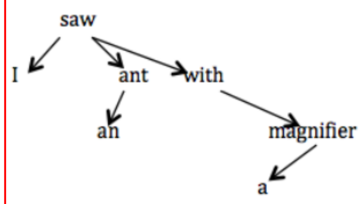


Figure 1: An example of dependency tree.

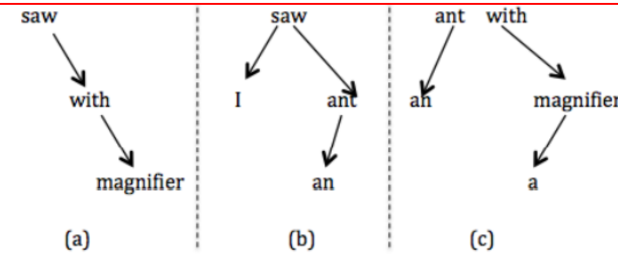


Figure 2: Different kinds of structures extracted from the dependency tree in Figure 1. (a): Head-word chain. (b): Fixed structure. (c): Floating structure.

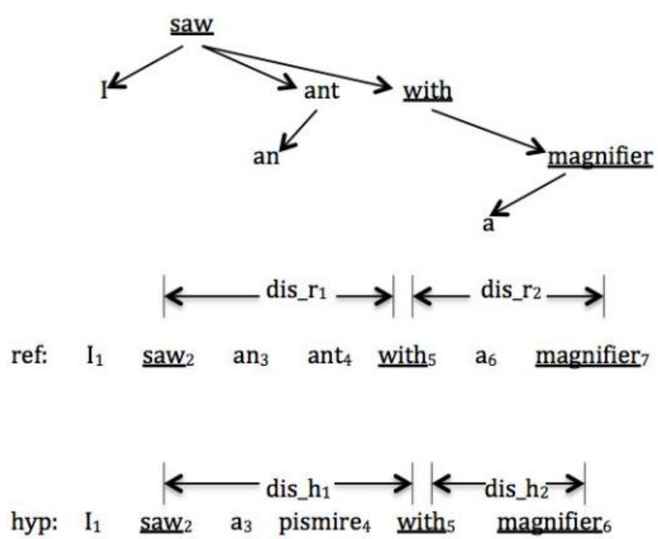


Figure 3: scoring head-word chain matching

Extra resources REDp (plus):

- stem and synonym
- paraphrase
- function word, content word

$$RED = \sum_{n=1}^{N} (w_{ngram} \times Fscore_n)$$

Evaluation

Tab 1: system-level correlation

data	WMT 2012					WMT 2013					
Metrics	cz-en	de-en	es-en	fr-en	ave	cz-en	de-en	es-en	fr-en	ru-en	ave
BLEU	.886	.671	.874	.811	.811	.936	.895	.888	.989	.670	.876
TER	.886	.624	.916	.821	.812	.800	.833	.825	.951	.581	.798
HWCM	.943	.762	.937	.818	.865	.902	.904	.886	.951	.756	.880
METEOR	.657	.885	.951	**.843**	.834	.964	.961	.979	.984	.789	.935
SEMPOS	.943	.924	.937	.804	.902	.955	.919	.930	.938	.823	.913
RED	**1.0**	.759	.951	.818	.882	.964	.951	.930	.989	.725	.912
REDp	.943	**.947**	**.965**	**.843**	**.925**	**.982**	**.973**	**.986**	**.995**	**.800**	**.947**

Tab 2: sentence-level correlation

data	WMT 2012					WMT 2013					
Metrics	cz-en	de-en	es-en	fr-en	ave	cz-en	de-en	es-en	fr-en	ru-en	ave
BLEU	.157	.191	.189	.210	.187	.199	.220	.259	.224	.162	.213
HWCM	.158	.207	.203	.204	.193	.187	.208	.247	.227	.175	.209
METEOR	**.212**	**.275**	**.249**	**.251**	**.247**	**.265**	**.293**	**.324**	**.264**	**.239**	**.277**
RED	.165	.218	.203	.221	.202	.210	.239	.292	.246	.196	.237
REDp	**.212**	.271	.234	.250	.242	.259	.290	.323	.260	.223	.271

HPB MT tuned on RED

Train \ Eval.		BLEU	METEOR	RED
MERT	BLEU	18.90	28.38	19.91
	METEOR	18.68	28.64	20.02
	RED	18.07	28.17	19.97
MIRA	BLEU	19.12	28.54	20.02
	METEOR	19.10	28.56	20.05
	RED	17.74	28.82	20.02

Table 1: Czech–English evaluation performance. In each column, the intensity of shades indicates the rank of values.

Train \ Eval.		BLEU	METEOR	RED
MERT	BLEU	11.25	17.36	14.95
	METEOR	10.44	17.00	14.86
	RED	9.51	16.81	14.58
MIRA	BLEU	11.52	17.54	15.14
	METEOR	11.43	17.56	15.26
	RED	11.29	17.67	15.25

Table 2: English–Czech evaluation performance. In each column, the intensity of shades indicates the rank of values.

System Name	TrueSkill Score		BLEU
	Tuning-Only	All	
BLEU-MIRA-DENSE	0.153	-0.182	12.28
ILLC-UvA	0.108	-0.189	12.05
BLEU-MERT-DENSE	0.087	-0.196	12.11
AFRL	0.070	-0.210	12.20
USAAR-Tuna	0.011	-0.220	12.16
DCU	-0.027	-0.263	11.44
METEOR-CMU	-0.101	-0.297	10.88
BLEU-MIRA-SPARSE	-0.150	-0.320	10.84
HKUST	-0.150	-0.320	10.99
HKUST-LATE	—	—	12.20

Table 4: Results on Czech-English tuning

System Name	TrueSkill Score		BLEU
	Tuning-Only	All	
DCU	0.320	-0.342	4.96
BLEU-MIRA-DENSE	0.303	-0.346	5.31
AFRL	0.303	-0.342	5.34
USAAR-Tuna	0.214	-0.373	5.26
BLEU-MERT-DENSE	0.123	-0.406	5.24
METEOR-CMU	-0.271	-0.563	4.37
BLEU-MIRA-SPARSE	-0.992	-0.808	3.79
USAAR-BASELINE-MIRA	—	—	5.31
USAAR-BASELINE-MERT	—	—	5.25

Table 5: Results on English-Czech tuning

[Li et al., 2015]

Parsing As Evaluation

- Train a maximum-entropy model-based dependency parser on references
  - References are parsed by the Stanford parser
- Parse hypotheses and use the normalized parsing probability as a score

$$DPM = \exp(\frac{Score(hyp)}{2n-1})$$

- Lexical score: unigram f-score
- Final score: $DPMF = DPM \times F\text{-}score$

[Yu et al., 2015]

Parsing As Evaluation

System-level

metrics	cs-en	de-en	es-en	fr-en	avg
TER	.886	.624	.916	.821	.812
BLEU	.886	.671	.874	.811	.811
METEOR	.657	.885	.951	.843	.834
•SEMPOS	.940	**.920**	.940	.800	.900
DPM	.943	.735	.888	.821	.847
DPMF	**.943**	.909	**.951**	**.850**	**.913**

(a) System level correlations on WMT2012.

metrics	cs-en	de-en	es-en	fr-en	ru-en	avg
TER	.800	.833	.825	.951	.581	.798
BLEU	.946	.851	.902	**.989**	.698	.877
•METEOR	.964	.961	.979	.984	.789	.935
DPM	.945	.880	.937	.951	.800	.903
DPMF	**.991**	**.975**	**.993**	.984	**.849**	**.958**

(b) System level correlations on WMT2013.

Sentence-level

Language	cs-en	de-en	es-en	fr-en	avg
BLEU	.157	.191	.189	.210	.187
METEOR	.212	.275	.249	.251	.247
•spede07_pP	.212	.278	.265	**.260**	.254
DPM	.146	.187	.211	.183	.182
DPMF	**.227**	**.279**	**.279**	.252	**.259**

(a) Sentence level correlations on WMT 2012.

Language	cs-en	de-en	es-en	fr-en	ru-en	avg
BLEU	.199	.220	.259	.224	.162	.213
METEOR	**.265**	.293	.324	.264	**.239**	.277
•SIMPBLEU-RECALL	.260	**.318**	**.387**	**.303**	.234	**.301**
DPM	.179	.204	.237	.194	.146	.192
DPMF	.258	.296	.316	.269	.227	.273

(b) Sentence level correlations on WMT 2013.

RNN-Based MT Evaluation

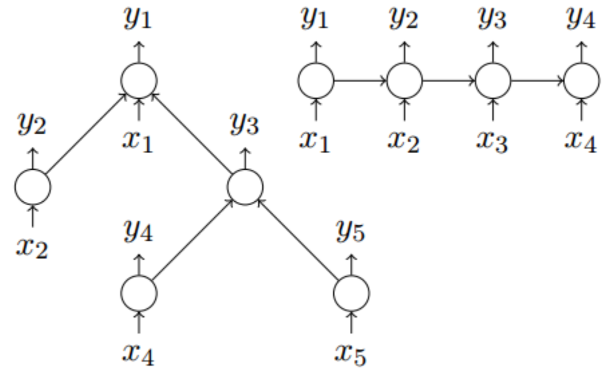


Figure 1: Tree-LSTM (left) and simple LSTM (right)

$$h_\times = h_{ref} \odot h_{tra}$$

$$h_+ = |h_{ref} - h_{tra}|$$

$$h_s = \sigma\left(W^{(\times)} h_\times + W^{(+)} h_+ + b^{(h)}\right)$$

$$\hat{p}_\theta = \text{softmax}\left(W^{(p)} h_s + b^{(p)}\right)$$

$$\boxed{\hat{y}} = r^T \hat{p}_\theta$$

Evaluation score

[Gupta et al., 2015, Tai et al., 2015]

Evaluation

Test	cs-en	de-en	fr-en	hi-en	ru-en	PAvg	SAvg
L+Sick(lstm)	.922 ± .051	.882 ± .028	.974 ± .009	.898 ± .011	.863 ± .023	.908 ± .024	.872 ± .060
LNF(50,150)	.972 ± .032	.900 ± .026	.974 ± .009	.900 ± .011	.882 ± .021	.925 ± .020	.913 ± .045
L(50,150)	.988 ± .022	.897 ± .027	.978 ± .008	.905 ± .010	.875 ± .022	.929 ± .018	.904 ± .042
L+Sick(50,150)	.993 ± .017	.904 ± .025	.978 ± .008	.908 ± .010	.881 ± .022	.933 ± .016	.915 ± .042
L+Sick(100,300)	.993 ± .018	.907 ± .025	.973 ± .009	.866 ± .012	**.890** ± .020	.926 ± .017	.902 ± .050
XL+Sick(100,300)	.913 ± .054	***.917*** ± .024	.978 ± .008	.904 ± .010	.884 ± .022	.919 ± .024	.889 ± .055
L+Sick(100,150)	**.994** ± .016	.911 ± .025	.975 ± .009	***.923*** ± .010	.870 ± .022	***.935*** ± .016	.904 ± .049
L+Sick(mix)	**.994** ± .017	.906 ± .025	***.979*** ± .008	.918 ± .010	.881 ± .022	***.935*** ± .016	**.919** ± .045
DiscoTK-party-tuned	.975 ± .031	**.943** ± .020	.977 ± .009	.956 ± .007	.870 ± .022	**.944** ± .018	.912 ± .043
LAYERED	.941 ± .045	.893 ± .026	.973 ± .009	**.976** ± .006	.854 ± .023	.927 ± .022	.894 ± .047
DiscoTK-party	.983 ± .025	.921 ± .024	.970 ± .010	.862 ± .015	.856 ± .023	.918 ± .019	.856 ± .046
REDSys	.989 ± .021	.898 ± .026	**.981** ± .008	.676 ± .022	.814 ± .026	.872 ± .021	.786 ± .047
REDSysSent	.993 ± .018	.910 ± .024	.980 ± .008	.644 ± .023	.807 ± .027	.867 ± .020	.771 ± .043
BLEU	.909 ± 0.54	.832 ± .034	.952 ± .012	.956 ± .007	.789 ± .027	.888 ± .027	.833 ± .058
Meteor	.980 ± .029	.927 ± .022	.975 ± .009	.457 ± .027	.805 ± .026	.829 ± .023	.788 ± .046

Table 3: Results: System-Level Correlations on WMT-14

Test	cs-en	de-en	fr-en	hi-en	ru-en	Average	Avg wmt12
L+Sick(lstm)	.204 ± .015	.232 ± .014	.289 ± .013	.319 ± .013	.236 ± .012	.256 ± .013	.254 ± .013
NFL(50,150)	.228 ± .015	***.288*** ± .014	.318 ± .014	.341 ± .014	.271 ± .012	.289 ± .014	.287 ± .014
L(50,150)	.225 ± .015	.272 ± .014	.328 ± .013	.346 ± .013	.280 ± .011	.290 ± .013	.287 ± .013
L+Sick(50,150)	.243 ± .016	.274 ± .013	.333 ± .013	.360 ± .014	.278 ± .011	.298 ± .013	.295 ± .014
L+Sick(100,300)	.233 ± .014	.286 ± .014	.343 ± .014	.358 ± .013	***.281*** ± .011	.300 ± .013	.297 ± .013
XL+Sick(100,300)	***.252*** ± .014	.279 ± .014	***.347*** ± .013	.367 ± .013	.274 ± .011	***.304*** ± .013	***.301*** ± .013
L+Sick(100,150)	.243 ± .016	.274 ± .014	.329 ± .013	***.368*** ± .012	.276 ± .011	.298 ± .013	.295 ± .013
L+Sick(mix)	.243 ± .016	.276 ± .013	.338 ± .013	.358 ± .013	.273 ± .011	.298 ± .013	.295 ± .013
DiscoTK-party-tuned	**.328** ± .014	**.380** ± .014	**.433** ± .013	.434 ± .013	**.355** ± .010	**.386** ± .013	**.386** ± .013
BEER	.284 ± .015	.337 ± .014	.417 ± .013	**.438** ± .014	.333 ± .011	.362 ± .013	.358 ± .013
REDcombSent	.284 ± .015	.338 ± .013	.406 ± .012	.417 ± .014	.336 ± .011	.356 ± .013	.346 ± .013
Meteor	.282 ± .015	.334 ± .014	.406 ± .012	..420 ± .013	.329 ± .010	.354 ± .013	.341 ± .013
BLEU_NRC	.226 ± .014	.272 ± .014	.382 ± .013	.322 ± .013	.269 ± .011	.294 ± .013	.267 ± .013
sentBLEU	.213 ± .016	.271 ± .014	.378 ± .013	.300 ± .013	.263 ± .011	.285 ± .013	.258 ± .014

Table 4: Results: Segment-Level Correlations on WMT-14

Summary

- Dependency structures are helpful on MT evaluation
  - Subtrees
  - Head-word chains
  - Fixed/floating structures
  - Dependency relations
  - RNN
- Extra resources are important to evaluation performance but language-dependent.

Thanks Lifeng Han for his help on this section.

References

- Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 85–91. Association for Computational Linguistics.
- Jesu ́s Gime ́nez and Llu ́ıs Ma`rquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 256–264. Association for Computational Linguistics.
- Yvette Graham and Qun Liu. 2016. Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL),* San Diego, CA
- Rohit Gupta, Constantin Orasan, Josef van Genabith (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In:Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1066–1072, Lisbon, Portugal.
- Yifan He and Andy Way. 2009. Learning Labelled Dependencies in Machine Translation Evaluation. Proceedings of the 13th Annual Conference of the EAMT, pages 44–51, Barcelona, May 2009
- Philipp Koehn. 2010. Statistical Machine. Translation. Cambridge University Press.
- Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Liangyou Li, Hui Yu, Qun Liu. 2015. MT Tuning on RED: A Dependency-Based Evaluation Metric. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 428-433, Lisboa, Portugal, 17-18 September 2015.
- Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 25–32.
- Chi-kiu Lo and Dekai Wu. 2013. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 422–428, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic mt evaluation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 243–252, Montre ́al, Canada, June. Association for Computational Linguistics.
- Dennis Mehay and Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In Proceedings of the MT summit.
- Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Sta- tistical Translation, SSST '07, pages 80–87, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Karolina Owczarzak, Josef van Genabith and Andy Way. 2007. Evaluating Machine Translation with LFG Dependencies. J. Machine Translation. Vol. 21, No. 2 (Jun., 2007), pp. 95-119.
- Karolina Owczarzak. 2008. A Novel Dependency-Based Evaluation Metric for Machine Translation. PhD thesis. DCU.
- K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of trans- lation edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas, pages 223–231.
- Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 259–268. Association for Computational Linguistics.
- Milos Stanojevic and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In Proceedings of the Ninth Workshop on Statistical Machine Translation.
- H Yu, X Wu, J Xie, W Jiang, Q Liu, S Lin. 2014. RED: A Reference Dependency Based MT Evaluation Metric. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2042–2051, Dublin, Ireland, August 23-29 2014.
- H Yu, X Wu, W Jiang, Q Liu, SX Lin. 2015. An Automatic Machine Translation Evaluation Metric Based on Dependency Parsing Model.arXiv preprint arXiv:1508.01996

Q&A

- Introduction
- Dependency-Based MT Evaluation
- Translation Models Based on Segmentation
- Translation Models Based on Synchronous Grammars
- Conclusion
- Lab Session

Structure Segmentation

Why Segmentation?

Dependency Tree Segmentation

Dependency Graph Segmentation

TRANSLATION MODELS BASED ON SEGMENTATION

Structure segmentation

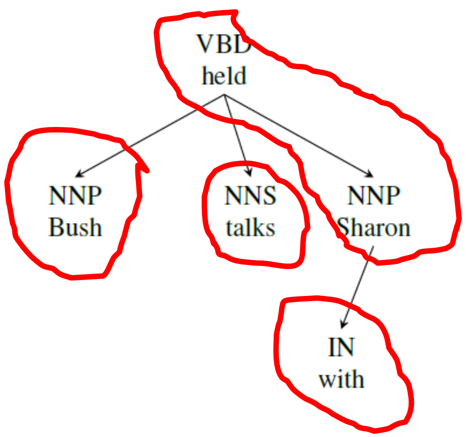
Segmentation divides structures into units.

Sentence -> phrases
Phrase-based models

Sentence Segmentation

`Morgen` `fliege` `ich` `nach Kanada` `zur Konferenz`

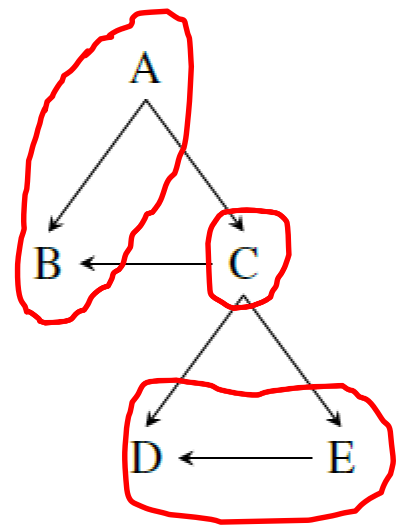
Tree Segmentation



tree -> treelets
treelet-based models

graph -> subgraphs
graph-based models

Graph Segmentation

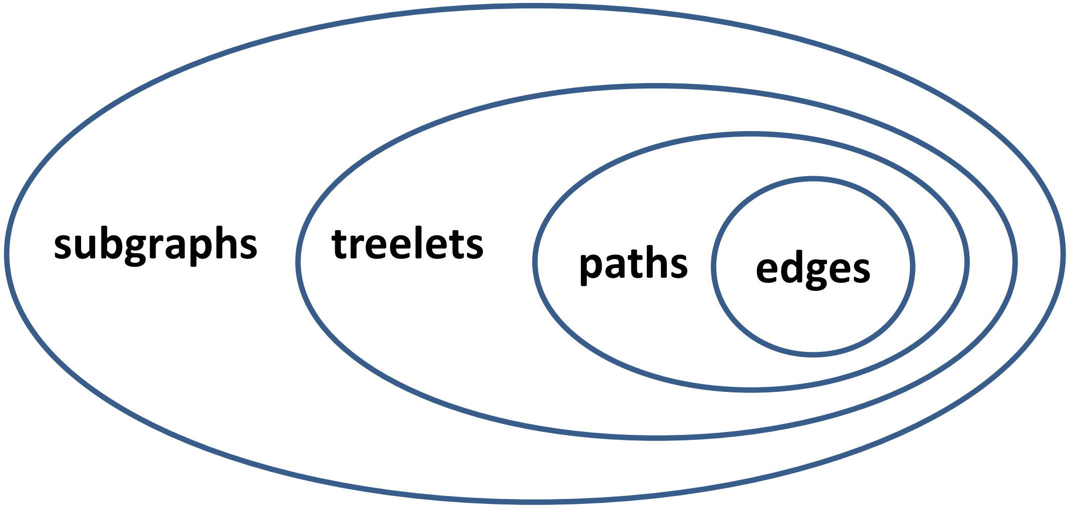


Why Segmentation?

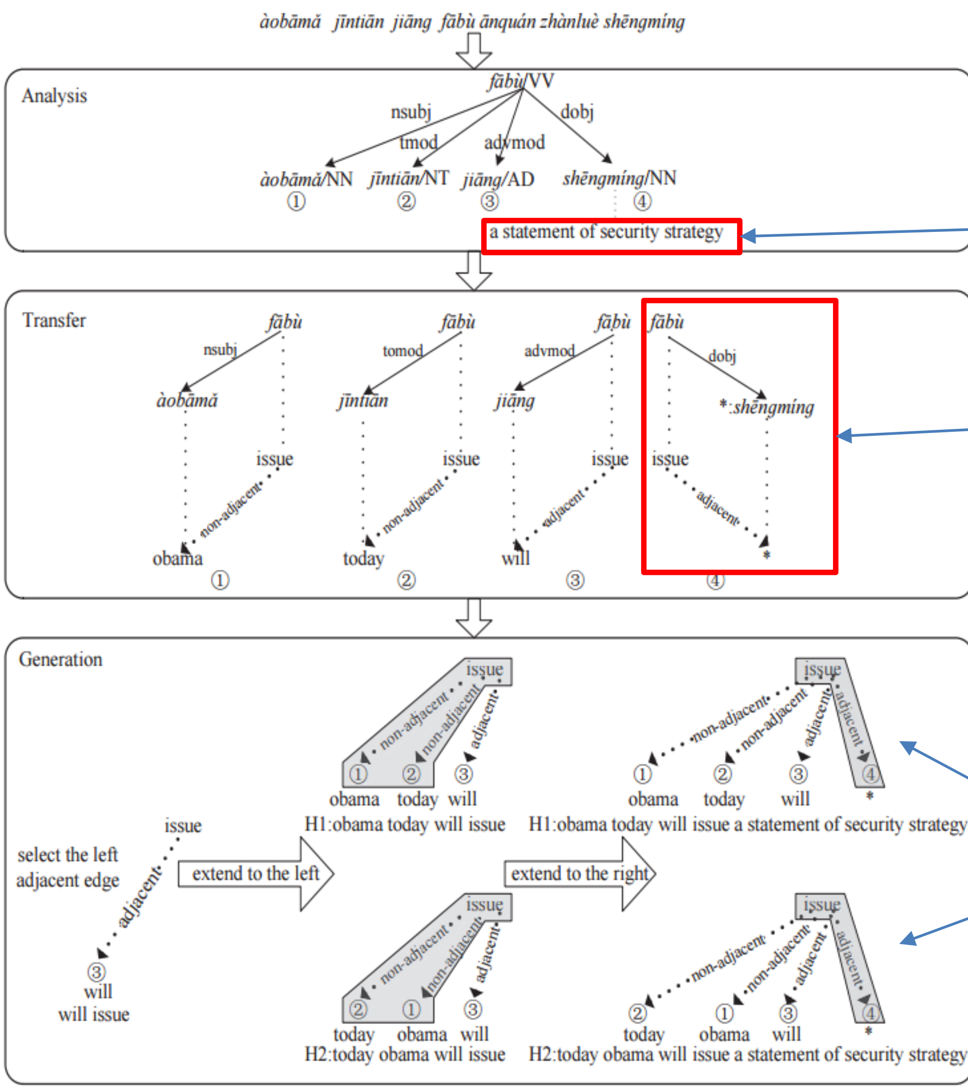
- Intuitive
  - Instead of translating a whole sentence at a time, translating parts and then combing them
- Small model
  - Not reply on recursive rules
- Flexible translation units
  - Such as treelets and subgraphs covering discontinuous spans.
- Fast decoding in practice
  - Phrase-based model vs hierarchical phrase-based model

Dependency Segmentation

- Dependency Tree Segmentation
  - edges, paths, treelets
- Dependency Graph Segmentation
  - subgraphs



Dependency Edge Model



Bottom-up translate

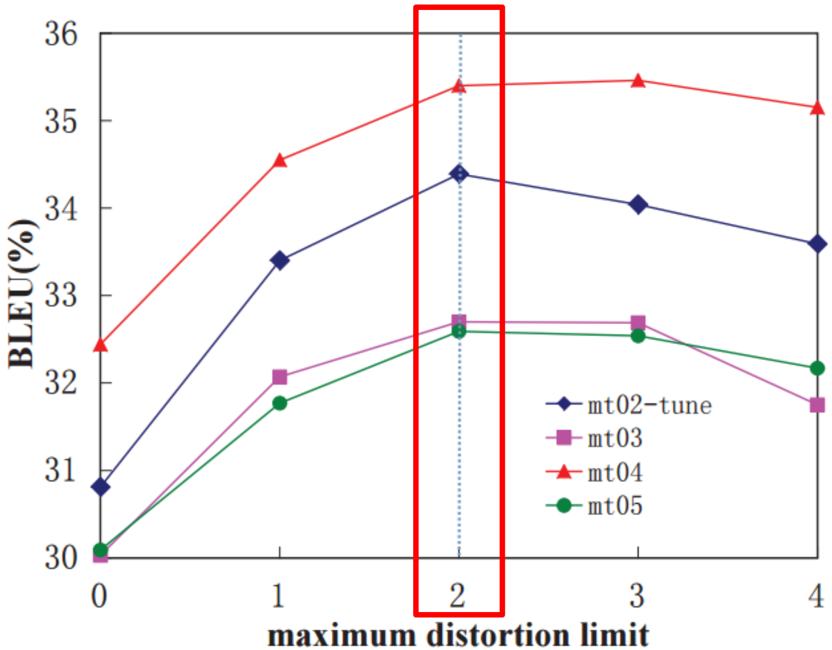
Subtree translations

Non-terminal nodes for integrating subtree translations

Different word order
For efficiency:
distortion limitation

[Chen et al., 2012]

Dependency Edge Model



Low distortion limit:

- less reordering is allowed.
- Target words are in the similar order with source words.
- Fast decoding

High distortion limit:

- Allow too much reordering
- Introduce many bad translations
- Low efficiency

Tab 1: BLEU scores

System	Rule #	MT03	MT04	MT05	Average
Moses	44.49M	32.03	32.83	31.81	32.22
DEBT	30.7M	**32.7***	**35.4***	**32.59***	**33.56**

Incorporating phrasal rules

[Chen et al., 2012]

Dependency Path Model

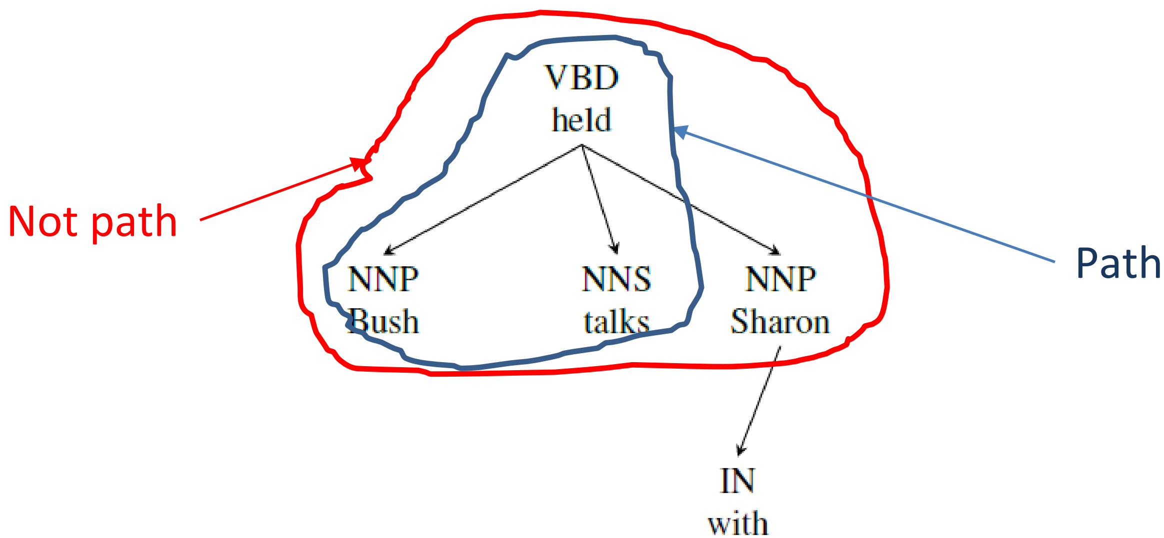
A sequence of nodes $n_1, \ldots, n_k, \ldots n_m$ and the dependency links between them form a **path** if the following conditions hold:

a. $\forall\, i\ (1 \le i < k)$, there is a link from n_{i+1} to n_i.

b. $\forall\, i\ (k \le i < m)$, there is a link from n_i to n_{i+1}.

n_k is a head word

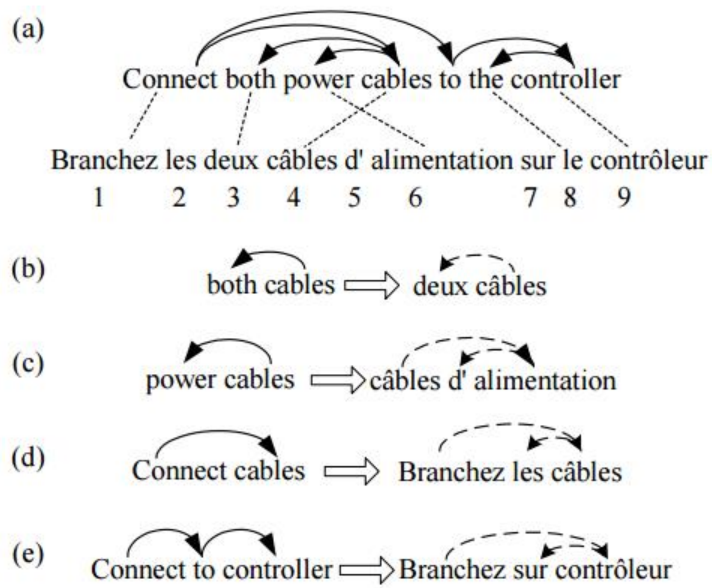
monotonic



[Lin, 2004]

Dependency Path Model

Rules:



Decoding:

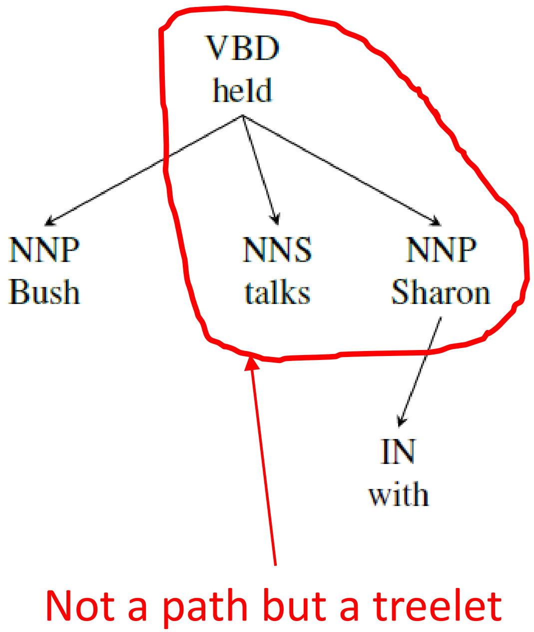
- A source sentence is parsed into a dependency tree
- Extract all paths and find transfer rules
- Find a sequence of transfer rules which
  - cover the source tree
  - generate a target tree
  - Have the highest probability
- Obtain a target sequence from the target tree

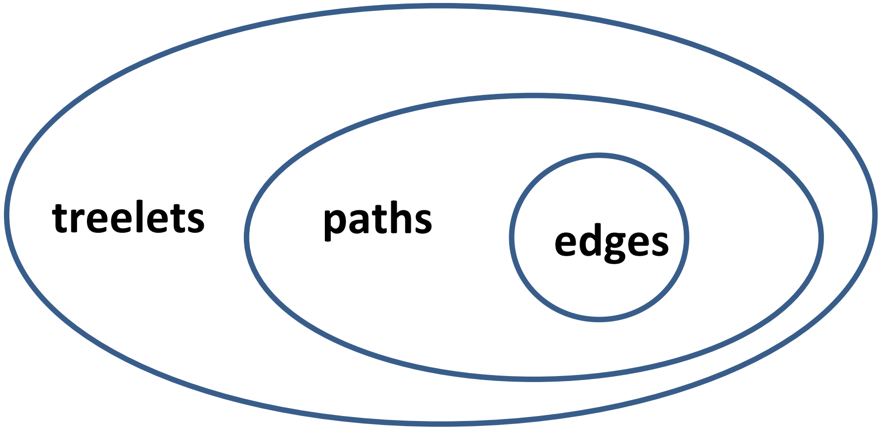
Worse than the phrase-based model

[Lin, 2004]

Dependency Treelet Model

A treelet is defined to be an arbitrary connected subgraph of a dependency tree.



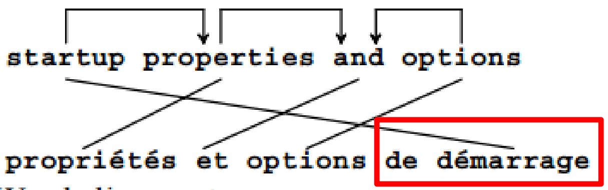


[Quirk et al., 2005]

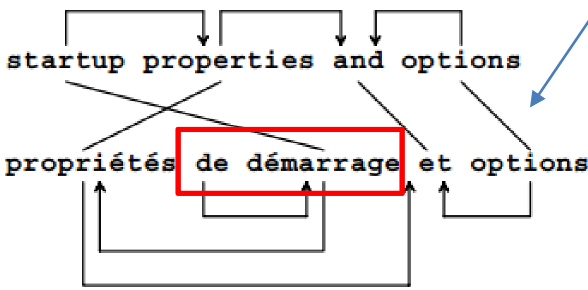
Dependency Treelet Model

Projection based on word alignments

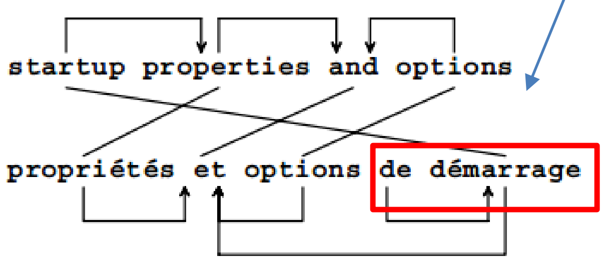
Reattachment to keep target word order



(a) Word alignment.

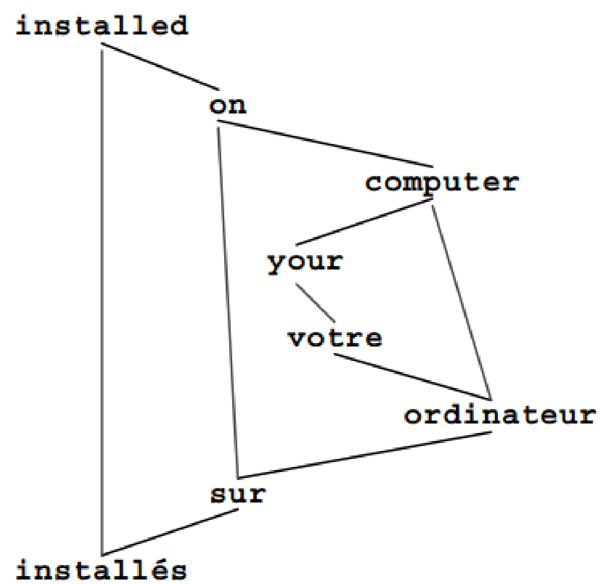


(b) Dependencies after initial projection.



(c) Dependencies after reattachment step.

Example translation rule

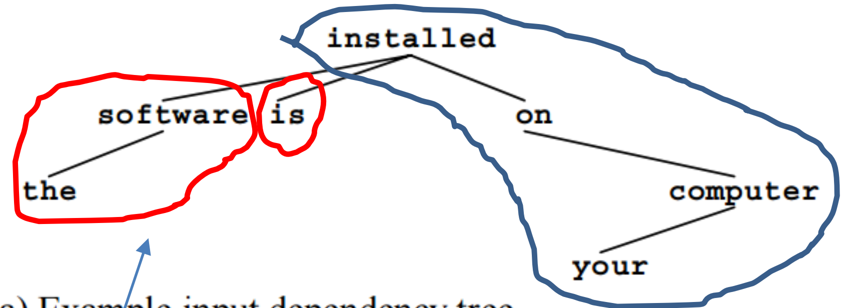


[Quirk et al., 2005]

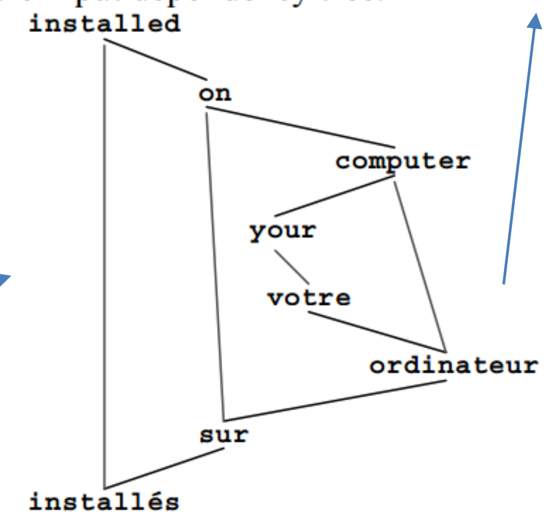
Dependency Treelet Models

- Bottom-up decoding
- Translations of treelets are attached together to form a complete translation
- Attachment during decoding: combinatory problem

Attach target trees to the head word
=Insert translations into *installes sur*
3*4=12 possibilities!



(a) Example input dependency tree.



(b) Example treelet translation pair.

[Quirk et al., 2005]

Evaluation

Tab 1: System comparison

	BLEU Score	Sents/min
Pharaoh monotone	37.06	4286
Pharaoh	38.83	162
MSR-MT	35.26	453
Treelet	40.66	10.1

Tab 3: Influence of treelet or phrase size

Max. size	Treelet BLEU	Pharaoh BLEU
1	37.50	23.18
2	39.84	32.07
3	40.36	37.09
4 (default)	40.66	38.83
5	40.71	39.41
6	40.74	39.72

Tab 2: Influence of reordering

Ordering strategy	BLEU	Sents/min
No order model (monotone)	35.35	39.7
Greedy ordering	38.85	13.1
Exhaustive (default)	40.66	10.1

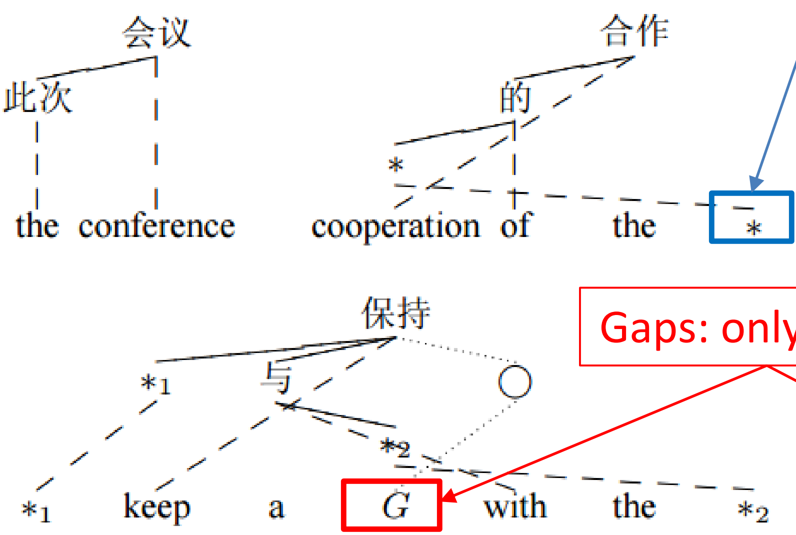
Tab 4: Continuity vs Discontinuity

	BLEU Score	Sents/min
Contiguous only	40.08	11.0
Allow discontiguous	40.66	10.1

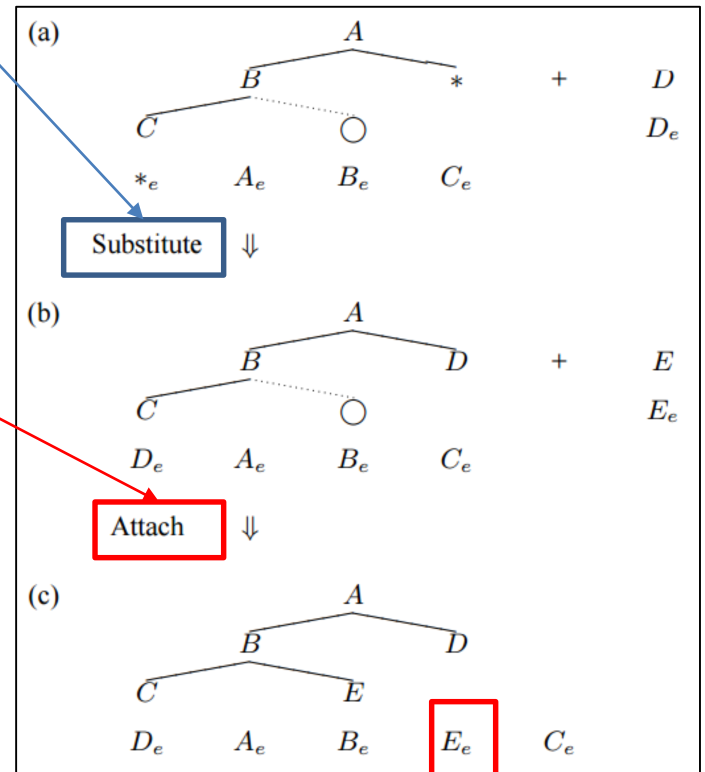
[Quirk et al., 2005]

Allowing Variables and Gaps

Variables: source -- target



Gaps: only target



Systems	BLEU-4
PB	20.88 ± 0.87
DTSC	20.20 ± 0.81
DTSC + phrases	21.46 ± 0.83

Weak at reordering

 [Xiong et al., 2007]

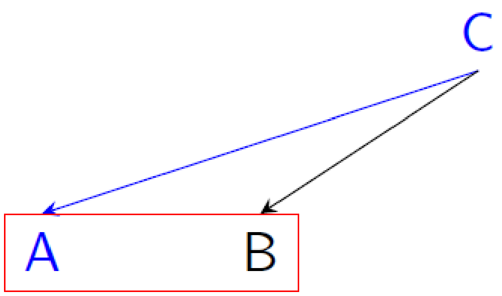
Dependency Graph Segmentation

- Why Graph Segmentation?
- How to Construct Graphs?
- Segmentational Graph-Based Model
- Context-Aware Segmentation

Why Graph Segmentation?

Treelet-Based Models (Menezes and Quirk, 2005; Quirk et al., 2005; Xiong et al., 2007)

- **tree-based**, translate a dependency tree by segmenting it into treelets
- Treelets are any connected subgaphs in the tree structure
- Treelet may cover discontinuous phrases which are linguistically-motivated and thus more reliable
- weakness: lower phrase coverage, only consider phrases connected in the tree

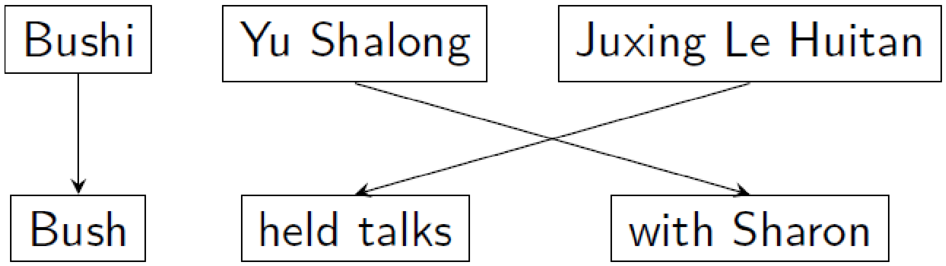


Sentence: **A B C**

Why Graph Segmentation?

Phrase-Based Models (Koehn et al., 2003)

- **sequence-based**, translate a sentence by segmenting it into phrases

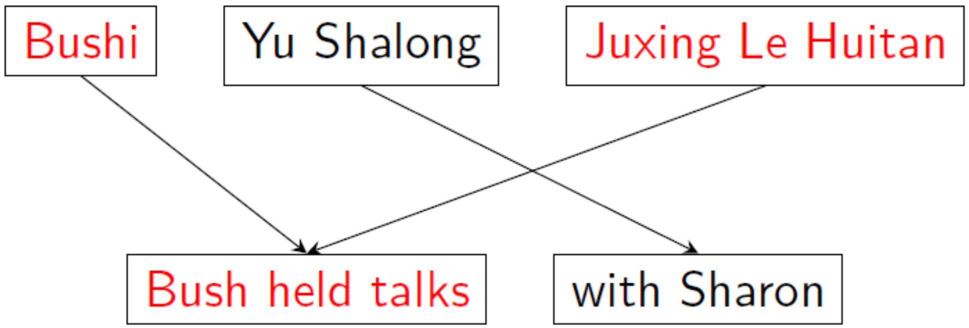


- make full use of continuous phrases, have higher phrase coverage
- weakness: cannot learn generalizations (discontinuous phrases)
  such as French *ne . . . pas* → English *not*

Why Graph Segmentation?

Allow discontinuous phrases + higher phrase coverage ?

DTU model achieves this by directly extracting both continuous and discontinuous phrases from sentence pairs (Galley and Manning, 2010)



Without linguistic structures to restrict the discontinuity:

- Extract plenty of discontinuous phrases which may be unreliable
- Learn a huge model

Why Graph Segmentation?

⟹ **graph-based model** which takes subgraphs as the basic translation units:

- Graphs combine dependency relations and bigram relations
- So both continuous phrases and linguistically-informed discontinuous phrases are connected.

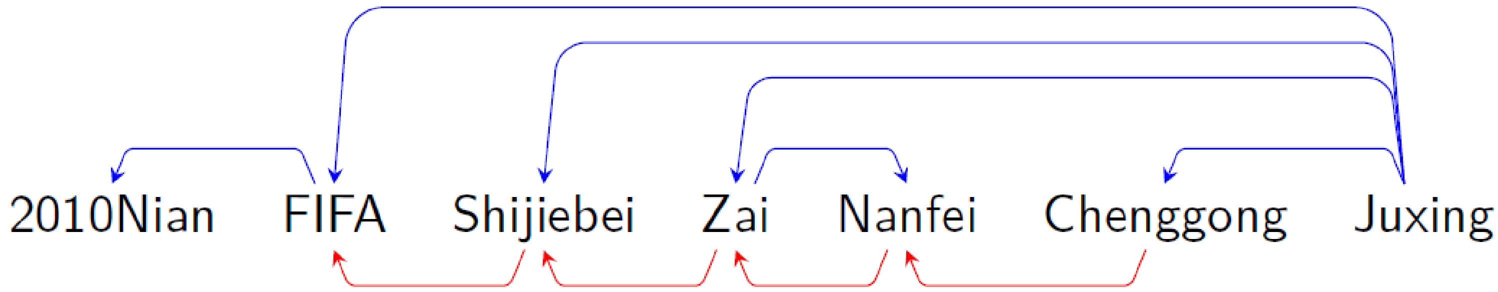
Model	Coverage	Discontinuity	Structure
Phrase-Based	•		sequence
Treelet-Based		•	tree
DTU	•	•	sequence
This work	•	•	**graph**

[Li et al., 2016]

How to Construct Graphs?

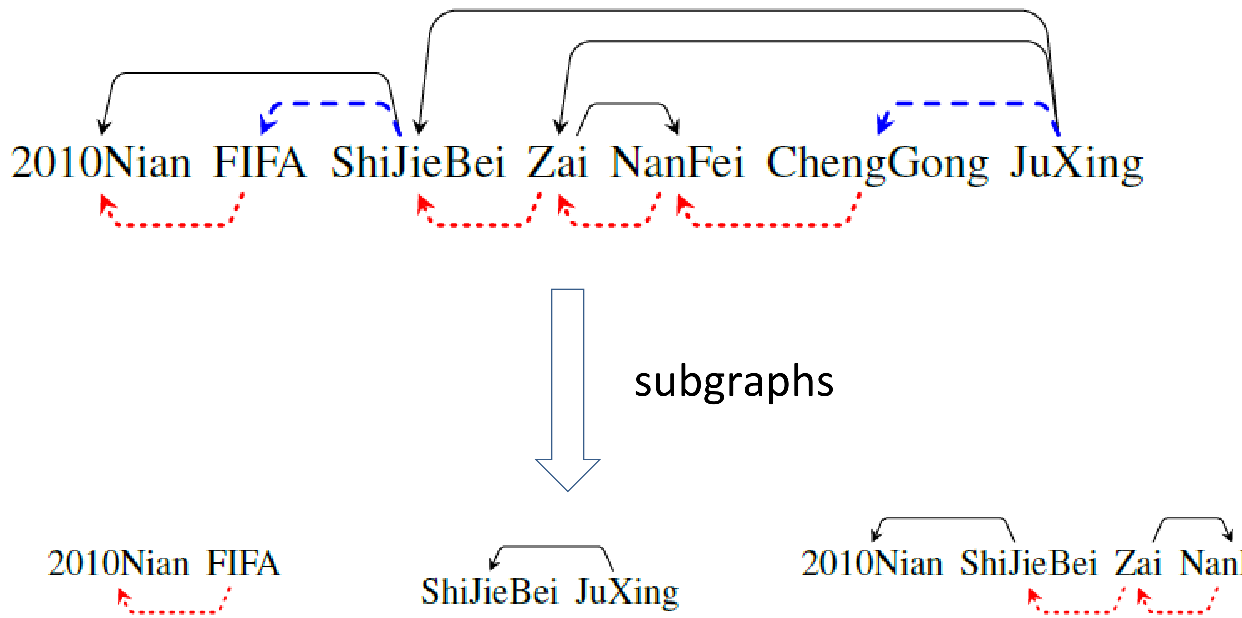
Dependency Relations:
encourage linguistically-informed discontinuous phrases

Bigram Relations:
encourage continuous phrases to improve phrase coverage



[Li et al., 2016]

How to Construct Graphs?



(a) Subgraph connected by bigram links (b) Subgraph connected by dependency links (c) Subgraph connected by both dependency and bigram links

Segmentational Graph-Based Models

- Training

Given a graph-string pairs, we extract subgraph-phrase pairs which are consistent with word alignment

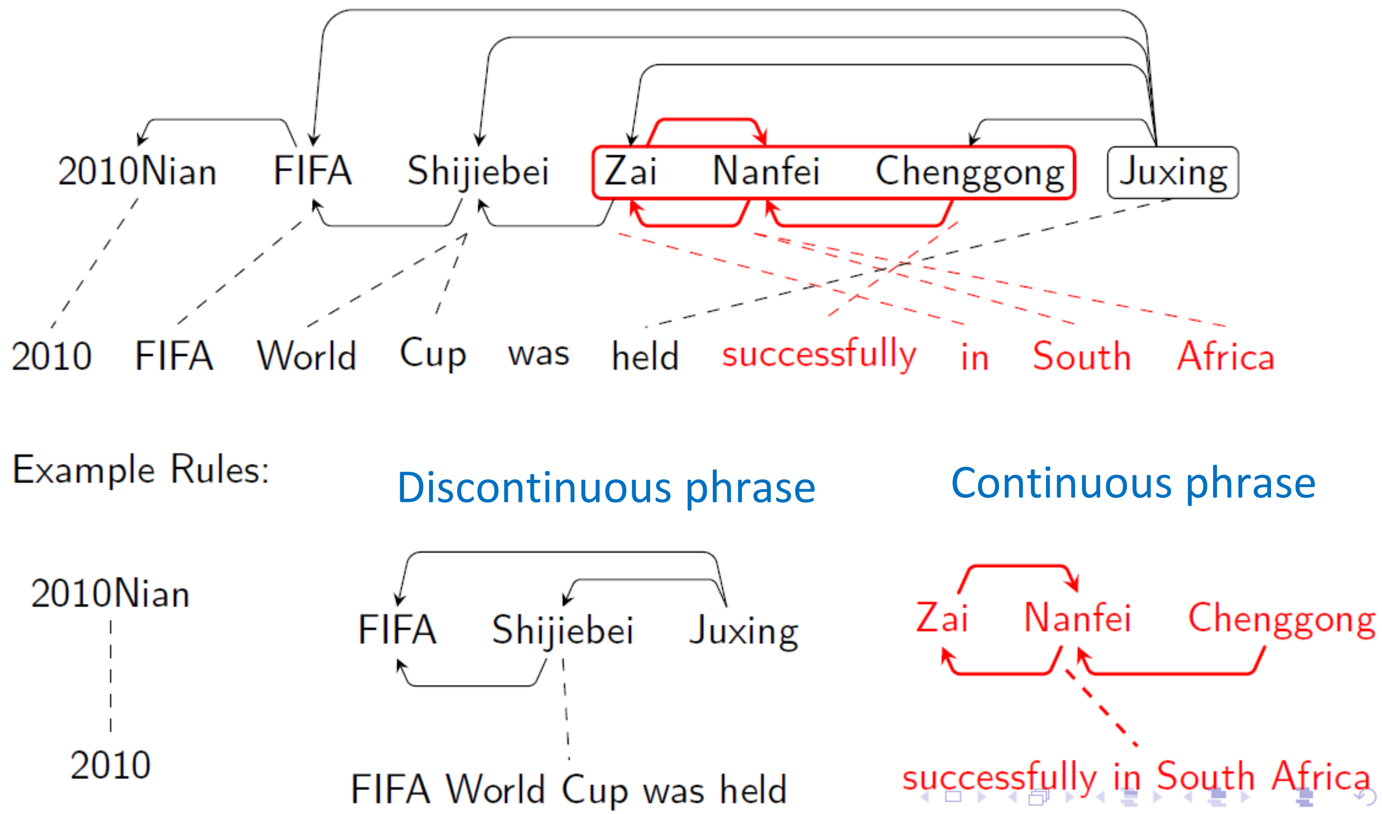
For each target phrase:

1. find a set of source words which are aligned to the phrase
2. if source words are connected, output a subgraph-phrase pair
3. extend with unaligned source words
4. go back to Step 2 until no more unaligned words are added.

 [Li et al., 2016]

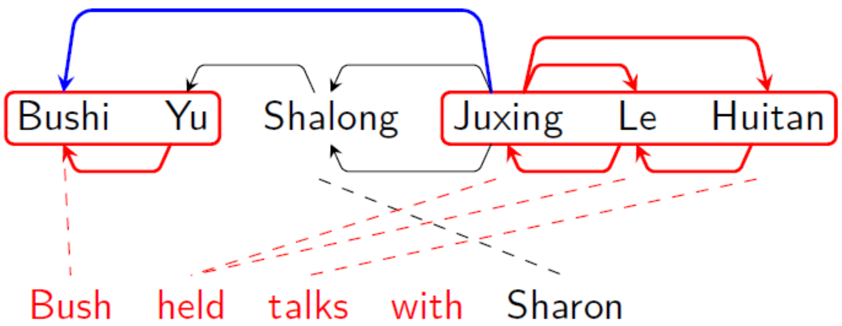
Segmentational Graph-Based Models

- Training

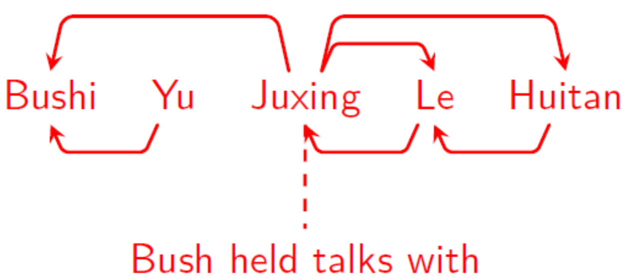


Segmentational Graph-Based Models

- Training

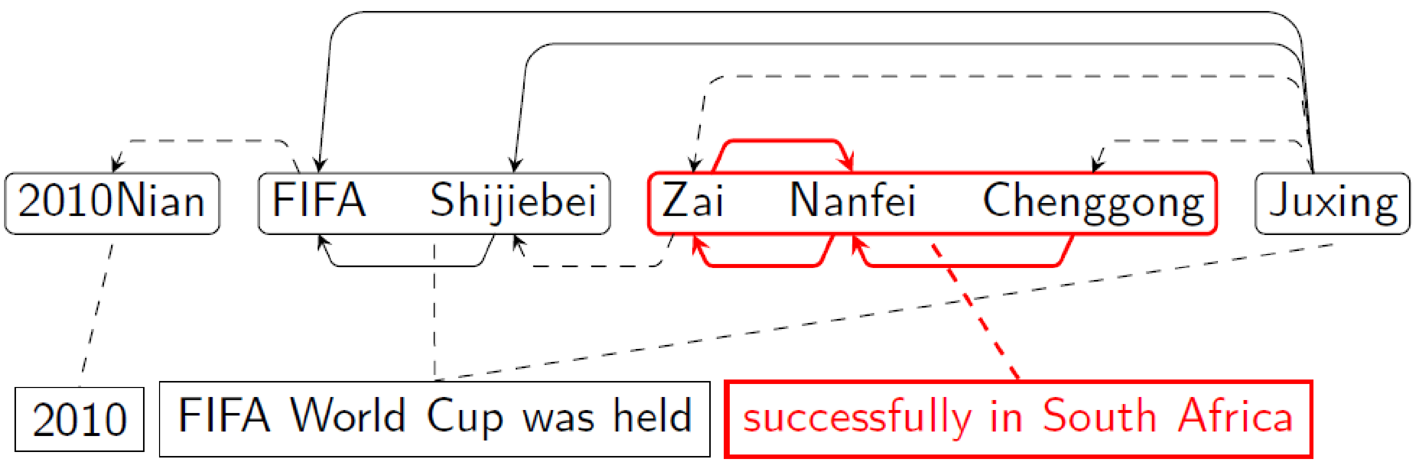


A special rule in the graph-based model



Segmentational Graph-Based Models

- Decoding
  - It generates translations from left to right
  - Beam search is used to find a complete translation



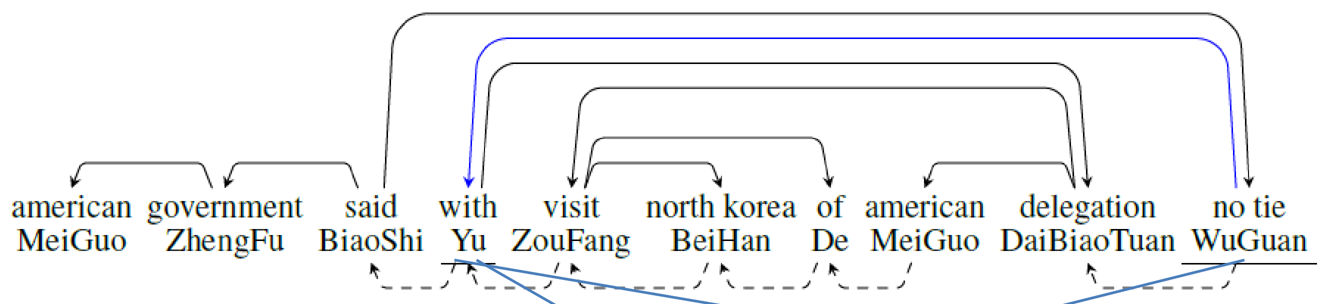
Evaluation

Tab 1: BLEU scores

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
PBMT	33.2	31.8^{+}	19.5	21.9
Treelet	33.8^{*}	31.4	19.6	22.2^{+}
DTU	34.7^{*+}	32.6^{*+}	19.7^{*}	22.4^{*}
SegGBMT	34.7^{*+}	32.4^{*+}	$20.1^{*+\ddagger}$	$22.9^{*+\ddagger}$

Tab 2: system rule number

System	# Rules	
	ZH–EN	DE–EN
DTU	224M+	352M+
SegGBMT	99M+	153M+

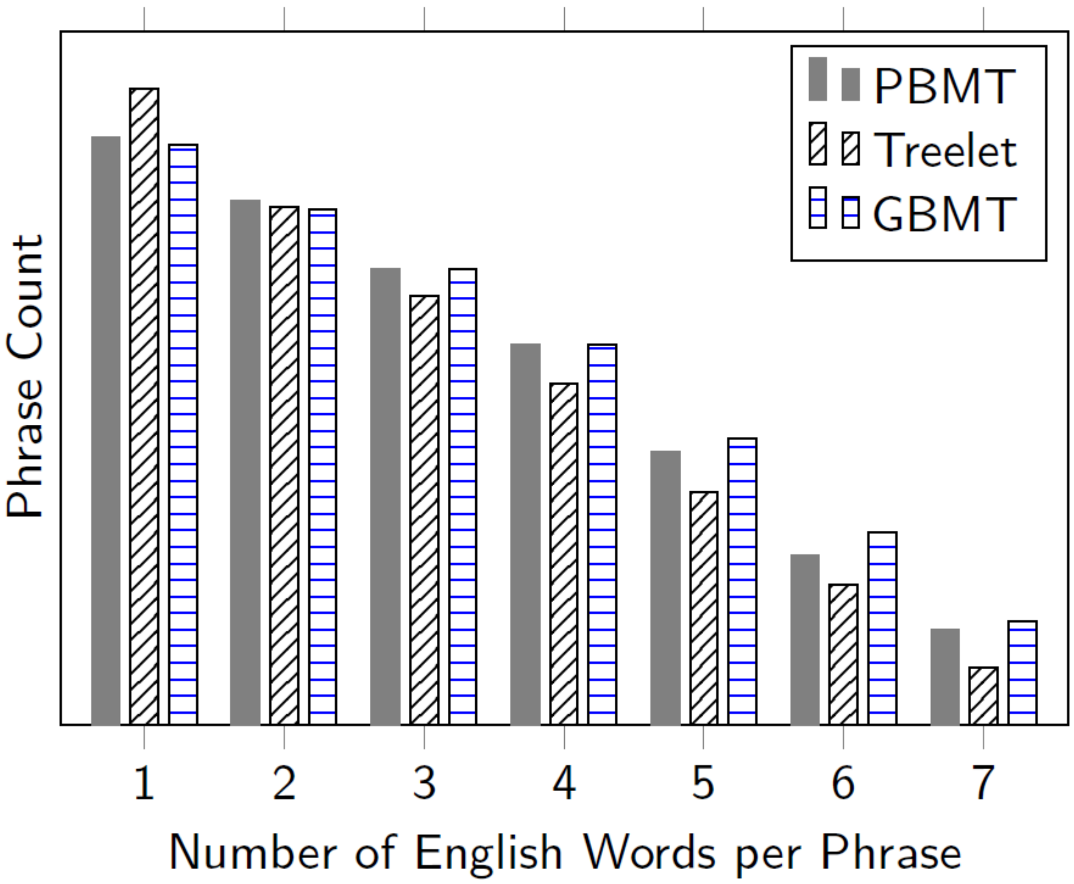


Ref: The american government said that it has nothing to do with the american delegation to visit north korea

PBMT: The government has said that the united states and north korea delegation has visited the united states

SegGBMT: The united states has indicated that it has nothing to do with the us delegation visited north korea

Evaluation



Higher phrase coverage leads to larger phases to be used

- Treelet tends to use smaller phrases. (only dependency relations, low coverage)
- GBMT uses more larger phrase pairs. (+bigram relations)

Evaluation

Tab 1: rule number according to their types

Rule Set	# Rules ZH–EN	# Rules DE–EN
PhrRule	70M+	107M+
TreeRule	42M+	73M+
PhrRule+TreeRule	82M+	129M+
SpecRule	16M+	23M+
All	99M+	153M+

70%

42%--48%

Share >30%

15%--17%

Tab 2: BLEU scores

Rule Set	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
PhrRule	34.4	32.3	19.6	22.0
TreeRule	33.8	32.0	19.8+	22.4+
+PhrRule	34.6*	32.2	20.1+*	22.9+*
+SpecRule	34.7	32.4	20.1+	22.9+

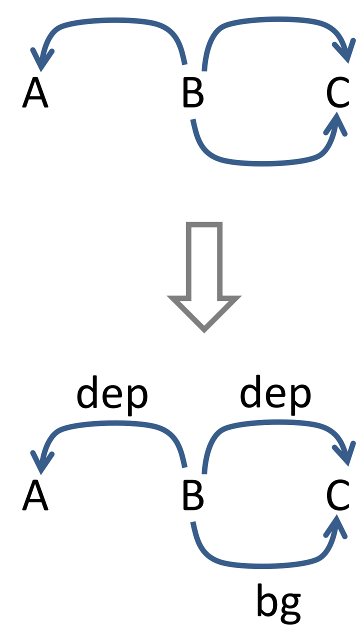
Inconsistency: more TreeRules are extracted and used?

small contribution but the best

Evaluation

Tab 1: Influence of edge types

Metric	System	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
BLEU ↑	SegGBMT	34.7	32.4	20.1	22.9
	+ET	34.7	32.7*	20.1	22.9
METEOR ↑	SegGBMT	32.4*	32.4*	28.4	29.7
	+ET	32.2	32.3	28.4	29.7
TER ↓	SegGBMT	60.1	61.6	63.1	59.3*
	+ET	59.0*	60.3*	63.2	59.4

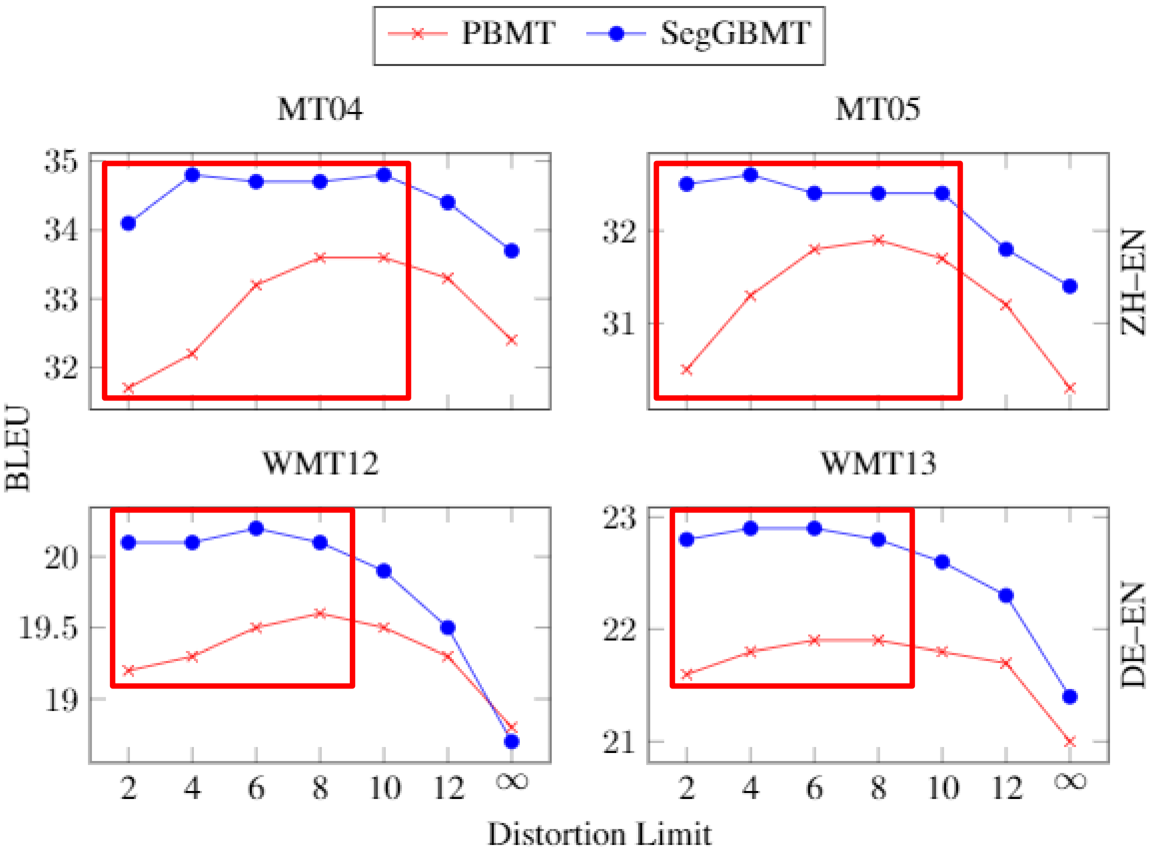


Tab 2: rule number

System	# Rules ZH–EN	# Rules DE–EN
SegGBMT	99.2M+	153.4M+
+ET	99.7M+	153.8M+

Less ambiguity

Evaluation

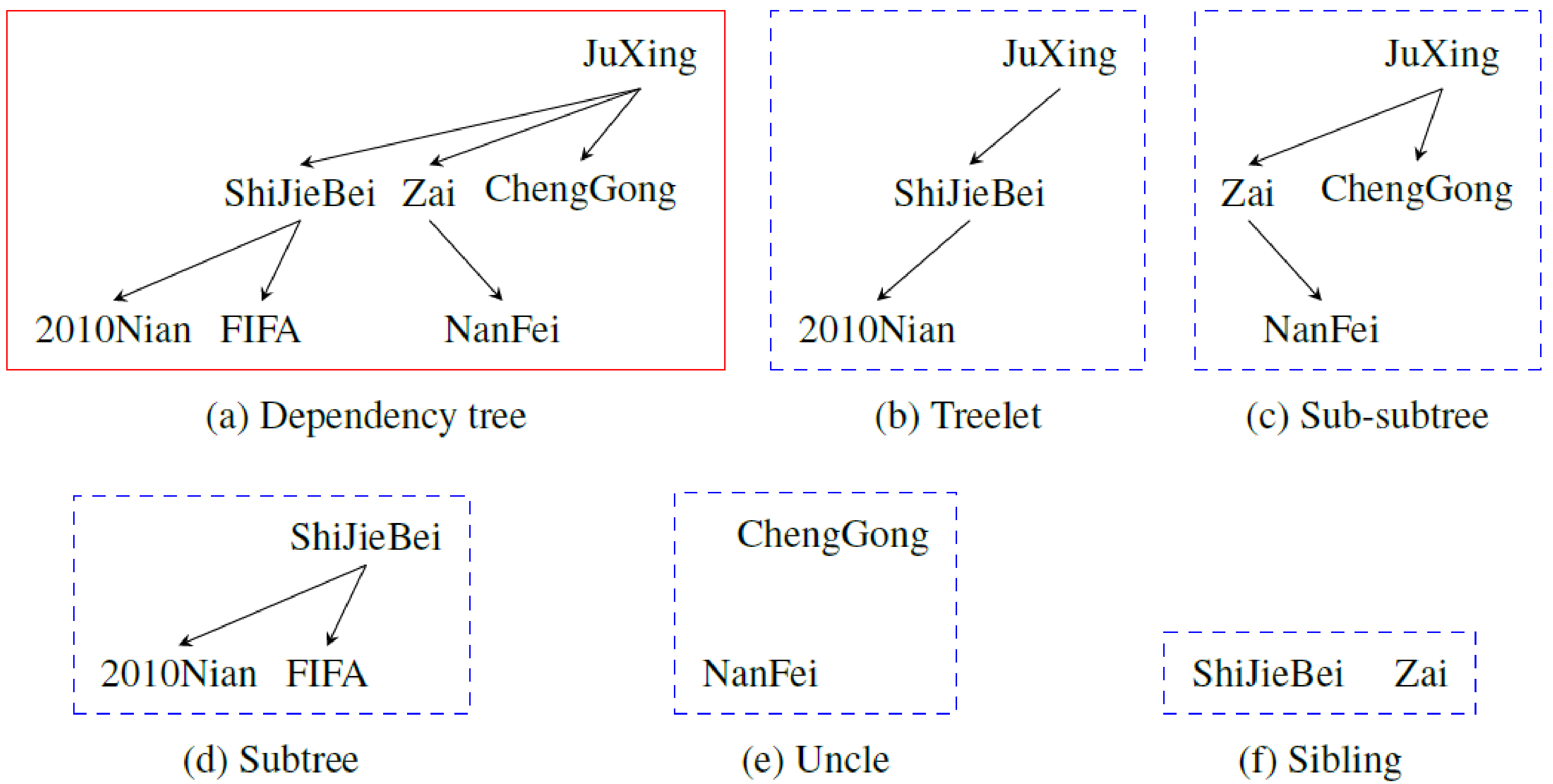


distortion limit:

- disallows long-distance phrase reordering
- speed up the decoder
- often improve translation performance.

Less sensitive:
Even though the distortion limit is small, subgraphs can cover long-distance discontinuous phrases.

Evaluation



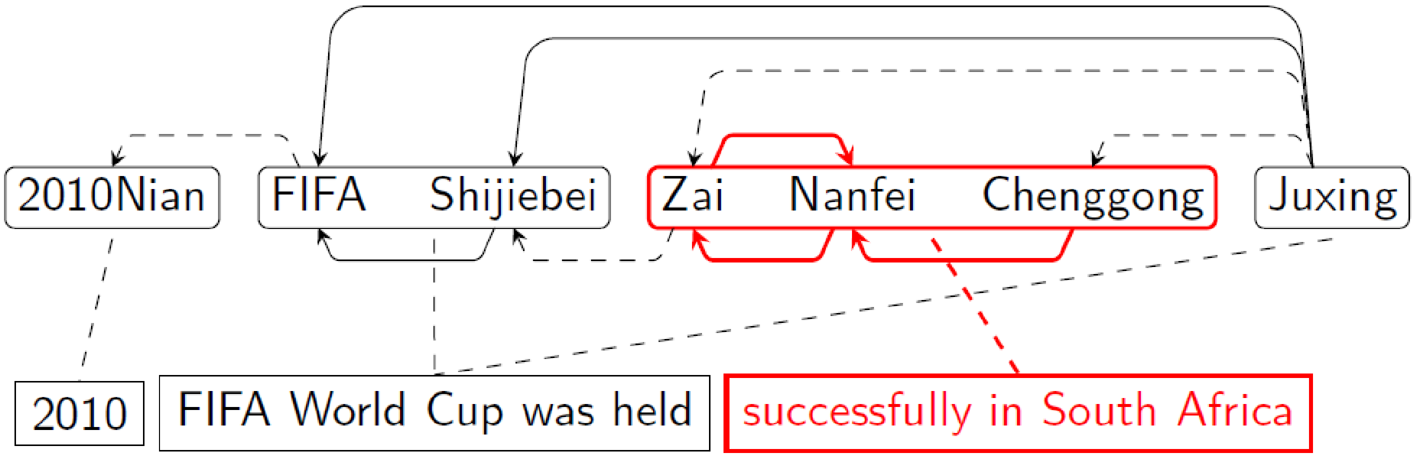
Given the dependency tree in (a), SegGBMT can cover dependency configurations (b)–(f).

Context-Aware Segmentation

- Why context-awareness?
- Graph segmentation model
- Context-aware rules

Why Need Context-Awareness?

- Better subgraph selection
- Better rule selection



Graph Segmentation Model

Basic Assumption:

$$p(G(\tilde{s}_1)\cdots G(\tilde{s}_I)) = \prod_{i=1}^{I} P(G(\tilde{s}_i)|G(\tilde{s}_1)\cdots G(\tilde{s}_{i-1}))$$

Sparse Features:

$$\left\{\begin{matrix} n.w \\ n.c \end{matrix}\right\} \times \left\{\begin{matrix} n'.w \\ n'.c \end{matrix}\right\} \times \left\{\begin{matrix} C \\ P \\ H \end{matrix}\right\} \times \left\{\begin{matrix} in \\ out \end{matrix}\right\}$$

Current node n

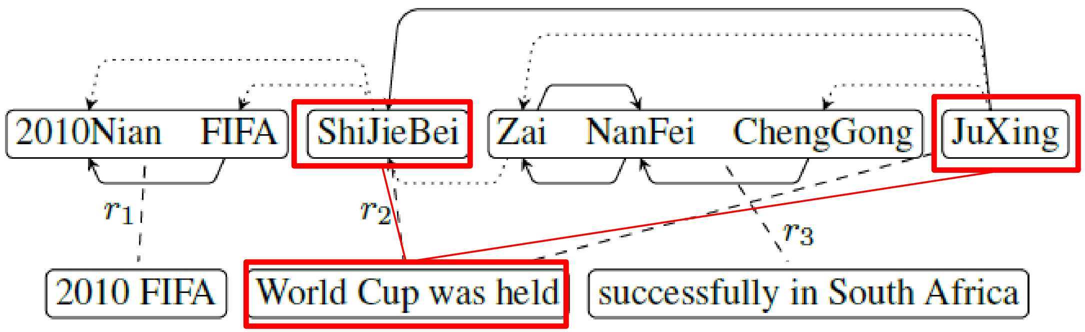
A node n’ connected to the current node

Where the node n’ comes from

Connection directions

[Li et al., 2016]

Graph Segmentation Model



Sparse Features for r_2:

Extract for each node

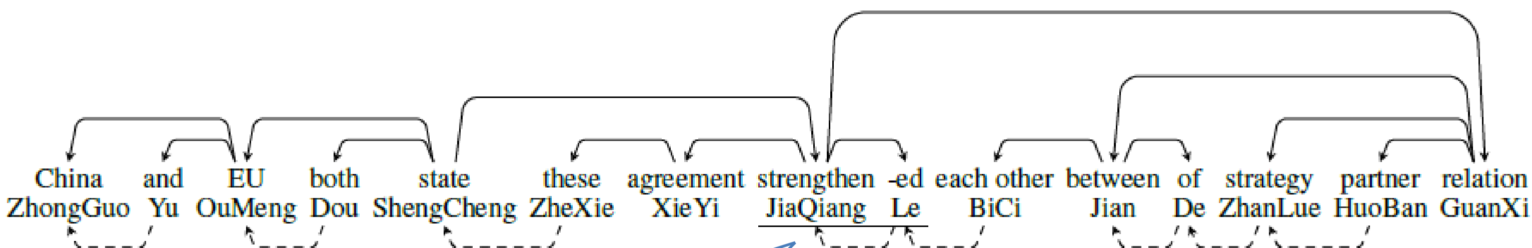
w=ShiJieBei@w=JuXing@p=C@d=in
w=ShiJieBei@c=1@p=C@d=in
w=ShiJieBei@w=2010Nian@p=P@d=out
w=ShiJieBei@c=2@p=C@d=out
w=ShiJieBei@w=FIFA@p=P@d=out
w=ShiJieBei@c=3@p=C@d=out
c=4@w=JuXing@p=C@d=in
c=4@c=1@p=C@d=in

c=4@w=2010Nian@p=P@d=out
c=4@c=2@p=C@d=out
c=4@w=FIFA@p=P@d=out
c=4@c=3@p=C@d=out
w=JuXing@w=ShiJieBei@p=C@d=out
w=JuXing@c=4@p=C@d=out
c=1@w=ShiJieBei@p=C@d=out
c=1@c=4@p=C@d=out

Full generalization

Evaluation

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
SegGBMT	34.7	32.4	20.1	22.9
SegGBMT+GSM	35.1*	32.6	20.4*	23.2*



Ref: Both China and the EU claimed that these agreements have strengthened their strategic partnership

SegGBMT: China and the EU have claimed that these agreements to strengthen their mutual strategic partnership

SegGBMT+GSM: China and the EU have claimed that these agreements have strengthened their strategic partnership

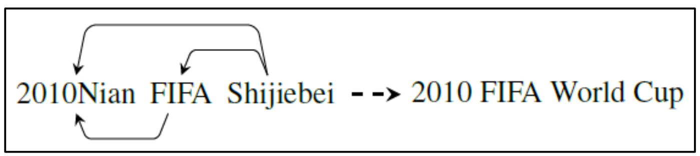
Context-Aware Rules

Rule form: $\langle g, t \rangle \longrightarrow \langle g, c, t \rangle$

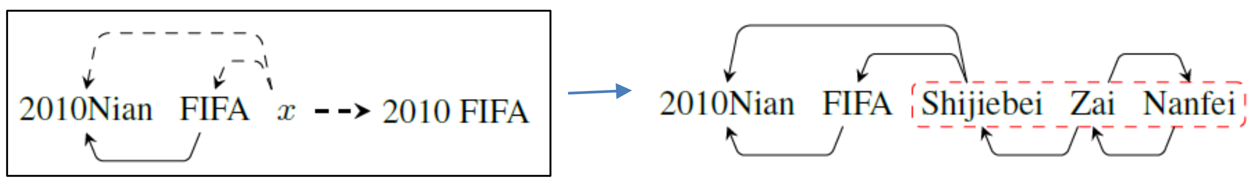
Rule Types:

Segmenting rules and selecting rules are extensions of basic rules by adding context information so that basic rules are split into different groups according to their contexts.

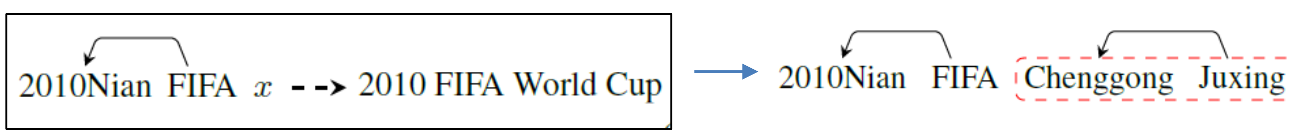
Basic Rule



Segmenting Rule



Selecting Rule



[Li et al., 2016]

Evaluation

Tab 1: BLEU scores

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
PBMT	33.2	31.8	19.5	21.9
Treelet	33.8*	31.7	19.6	22.1*
DTU	**34.5***	32.3*	**19.8***	**22.3***
GBMT$_{ctx}$	**35.4**$^{*+}$	**33.7**$^{*+}$	**20.1**$^{*+}$	**22.8**$^{*+}$

Tab 3: number of rules

Rule Type	# Rules	
	ZH–EN	DE–EN
Basic Rule	84.7M+	115.7M+
Segmenting Rule	128.4M+	167.3M+
Selecting Rule	30.2M+	35.7M+
Total	243.5M+	318.9M+

Tab 2: Influence of context

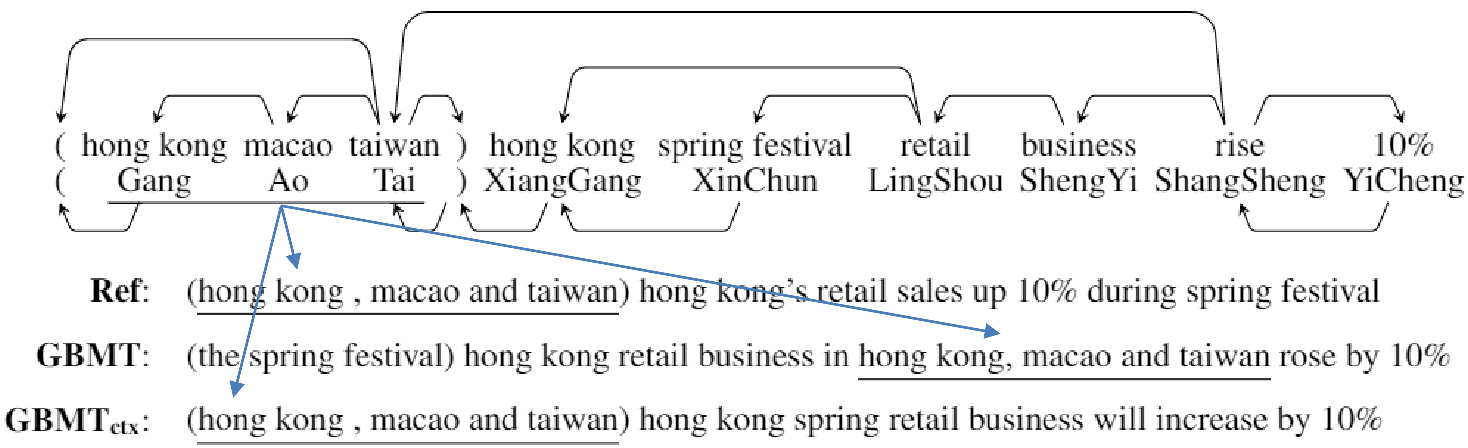
System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
GBMT	34.7	32.4	19.8	22.4
GBMT$_{ctx}$	**35.4**	**33.7**	**20.1**	**22.8**

Selecting rules are less often used?

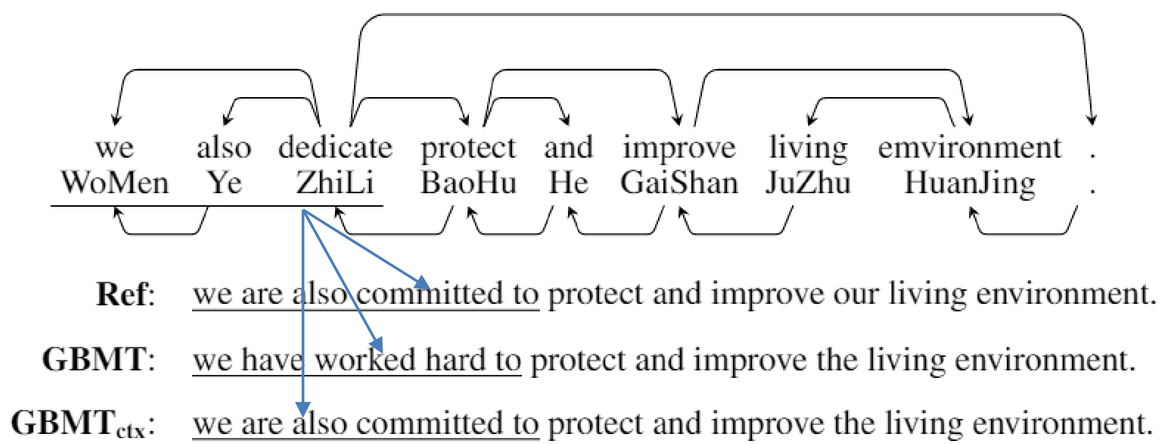
Tab 4: influence of rules

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
Basic Rule	34.7	32.4	19.8	22.4
+Seg. Rule	**34.9**	**33.0**	**20.2**	**23.0**
+Sel. Rule	34.8	32.5	**20.0**	**22.7**
All	**35.4**	**33.7**	**20.1**	**22.8**

Evaluation



(a) subgraph selection



(b) target-phrase selection

Summary

Segmentation-based models are flexible to use translation units. However, they are weak at phrase reordering.

Main research lines:

- Segmenting Dependency Tree
  - Edge
  - Path
  - Treelet
- Segmenting Dependency Graph
  - Subgraph
  - Contexts are helpful

References

- Hongshen Chen, Jun Xie, Fandong Meng, Wenbin Jiang, and Qun Liu (2014). A Dependency Edge-based Transfer Model for Statistical Machine Translation. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, pages 1103–1113.
- Michel Galley and Christopher D. Manning (2010). Accurate Non-hierarchical Phrase-Based Translation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, USA, pages 966–974.
- Philipp Koehn (2010). Statistical Machine Translation. 1st. New York, NY, USA: Cambridge University Press.
- Dekang Lin (2004). A Path-based Transfer Model for Machine Translation. In: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, pages 625–630.
- Arul Menezes and Chris Quirk (2005). Dependency Treelet Translation: The Convergence of Statistical and Example-Based Machine-Translation? In: Proceedings of the Workshop on Example-Based Machine Translation. Phuket, Thailand.
- Chris Quirk, Arul Menezes, and Colin Cherry (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan, USA, pages 271–279.
- Chris Quirk and Simon Corston-Oliver (2006). The Impact of Parse Quality on Syntacticallyinformed Statistical Machine Translation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, pages 62–69.
- Deyi Xiong, Qun Liu, and Shouxun Lin (2007). A Dependency Treelet String Correspondence Model for Statistical Machine Translation. In: Proceedings of the Second Workshop on Statistical Machine Translation. Prague, pages 40–47.
- Liangyou Li, AndyWay, Qun Liu (2016). Graph-Based Translation Via Graph Segmentation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, pages 97–107,
- Liangyou Li, AndyWay, Qun Liu (2016). Context-Aware Graph Segmentation for Graph-Based Translation. In Proceedings of EACL. (Submitted)

Q&A

- Introduction
- Dependency-Based MT Evaluation
- Translation Models Based on Segmentation
- Translation Models Based on Synchronous Grammars
- Conclusion
- Lab Session

Synchronous Grammars

String-to-Dependency Models

Dependency-to-String Models

Dependency Graph-to-String Models

TRANSLATION MODELS BASED ON SYNCHRONOUS GRAMMARS

Synchronous Grammars

- Synchronous context free grammar (SCFG)
  - Hierarchical phrase-based models
- Synchronous tree substitution grammar (STSG)
  - Tree-to-string models
  - String-to-tree models
  - Tree-to-tree models

SCFG

An SCFG is a tuple $\langle N, T, T', P, S \rangle$, where

- N is a finite set of non-terminal symbols.
- T and T' are finite sets of terminal symbols.
- $S \in N$ is the start symbol.
- P is a finite set of productions of the form $\langle A \rightarrow R, A' \rightarrow R', \sim \rangle$, where $A, A' \in N$, R is a $\boxed{\text{sequence}}$ over $N \bigcup T$ and R' is a $\boxed{\text{sequence}}$ over $N \bigcup T'$. $\sim$ is a one-to-one mapping between non-terminal symbols in R and R'.

SCFG

$\langle S_{\boxed{1}}, S_{\boxed{1}} \rangle$

$\overset{(14)}{\Rightarrow} \langle S_{\boxed{2}} X_{\boxed{3}}, S_{\boxed{2}} X_{\boxed{3}} \rangle$

$\overset{(14)}{\Rightarrow} \langle S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}}, S_{\boxed{4}} X_{\boxed{5}} X_{\boxed{3}} \rangle$

$\overset{(15)}{\Rightarrow} \langle X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}}, X_{\boxed{6}} X_{\boxed{5}} X_{\boxed{3}} \rangle$

$\overset{(9)}{\Rightarrow} \langle \text{Aozhou } X_{\boxed{5}} X_{\boxed{3}}, \text{Australia } X_{\boxed{5}} X_{\boxed{3}} \rangle$

$\overset{(11)}{\Rightarrow} \langle \text{Aozhou shi } X_{\boxed{3}}, \text{Australia is } X_{\boxed{3}} \rangle$

$\overset{(8)}{\Rightarrow} \langle \text{Aozhou shi } X_{\boxed{7}} \text{ zhiyi, Australia is one of } X_{\boxed{7}} \rangle$

$\overset{(7)}{\Rightarrow} \langle \text{Aozhou shi } X_{\boxed{8}} \text{ de } X_{\boxed{9}} \text{ zhiyi, Australia is one of the } X_{\boxed{9}} \text{ that } X_{\boxed{8}} \rangle$

$\overset{(6)}{\Rightarrow} \langle \text{Aozhou shi yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}} \text{ de } X_{\boxed{9}} \text{ zhiyi,}$
$\text{Australia is one of the } X_{\boxed{9}} \text{ that have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle$

$\overset{(10)}{\Rightarrow} \langle \text{Aozhou shi yu Beihan you } X_{\boxed{2}} \text{ de } X_{\boxed{9}} \text{ zhiyi,}$
$\text{Australia is one of the } X_{\boxed{9}} \text{ that have } X_{\boxed{2}} \text{ with North Korea} \rangle$

$\overset{(12)}{\Rightarrow} \langle \text{Aozhou shi yu Beihan you bangjiao de } X_{\boxed{9}} \text{ zhiyi,}$
$\text{Australia is one of the } X_{\boxed{9}} \text{ that have diplomatic relations with North Korea} \rangle$

$\overset{(13)}{\Rightarrow} \langle \text{Aozhou shi yu Beihan you bangjiao de shaoshu guojia zhiyi,}$
$\text{Australia is one of the few countries that have diplomatic relations with North Korea} \rangle$

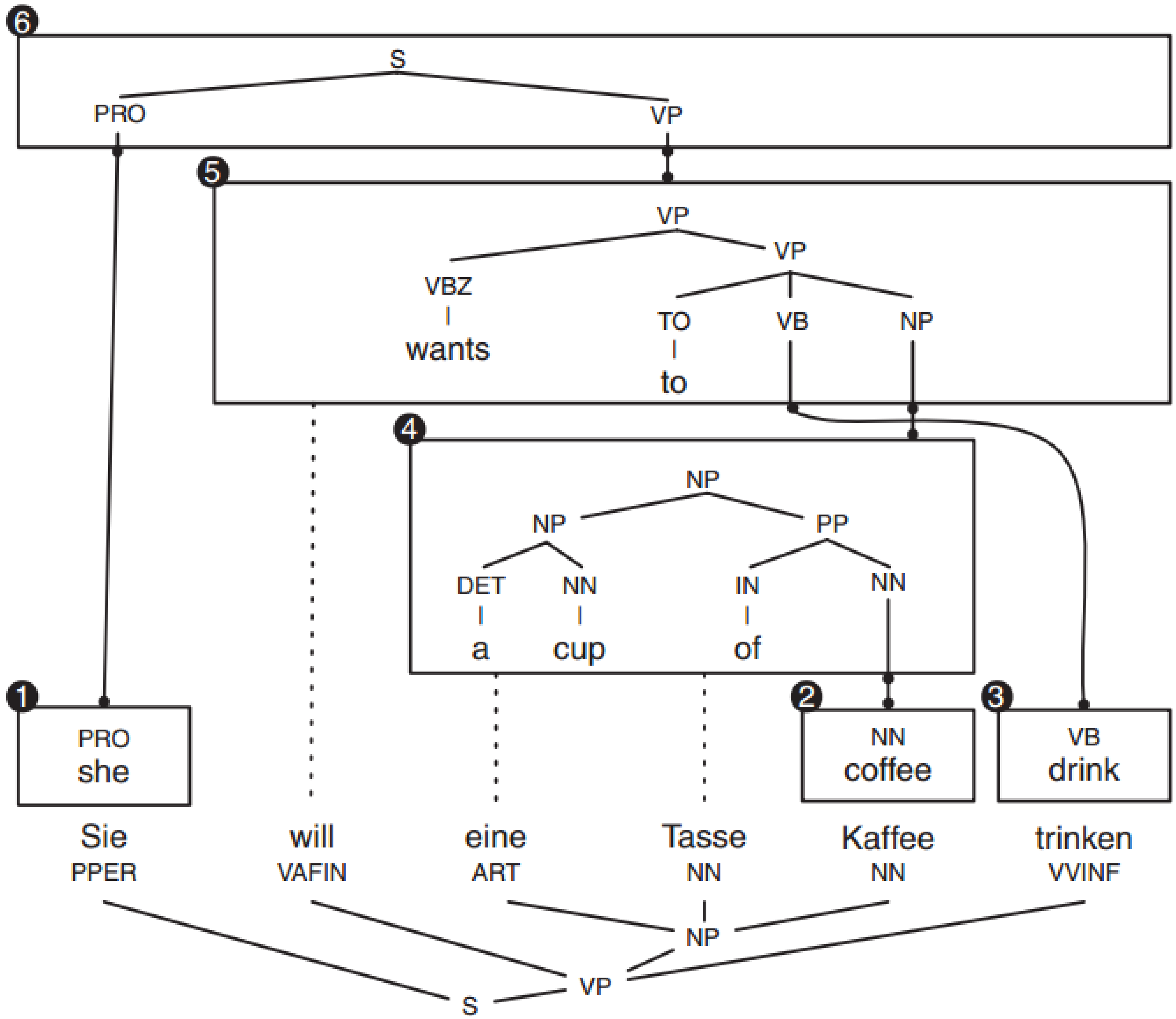
[Chiang, 2007]

STSG

An STSG is a tuple $\langle N, T, T', P, S \rangle$, where

- N is a finite set of non-terminal symbols.
- T and T' are finite sets of terminal symbols.
- $S \in N$ is the start symbol.
- P is a finite set of productions of the form $\langle A \rightarrow R, A' \rightarrow R', \sim \rangle$, where $A, A' \in N$, R is a $\boxed{\text{tree}}$ over $N \bigcup T$ and R' is a $\boxed{\text{tree}}$ over $N \bigcup T'$. $\sim$ is a one-to-one mapping between non-terminal symbols in R and R'.

STSG



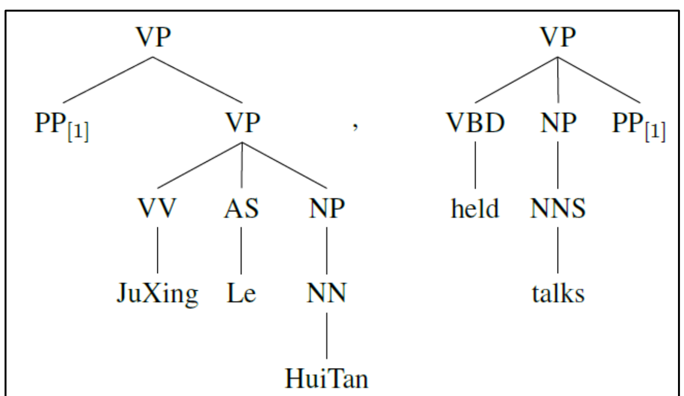
[Koehn, 2010]

Why Synchronous Grammars?

- Target phrase reordering
  - Recursive rules

$$X \rightarrow \langle \text{ BuShi } X_{[1]} \text{ JuXing Le } X_{[2]}, \text{Bush held } X_{[2]}\ X_{[1]} \rangle,$$

- Linguistic theory
  - Syntax annotations



String-to-Dependency Model

- Extension of hierarchical phrase-based model
- Well-formed dependency structures
- Dependency tree on the target side
- Dependency language model

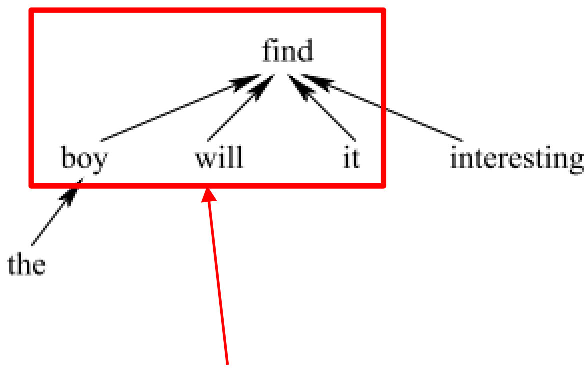
[Shen et al., 2008, 2010]

Well-Formed Dependency Structures

A dependency structure $d_i d_{i+1} \ldots d_j$, or $d_{i..j}$ for short, is **fixed on head** h, where $h \in [i, j]$, or **fixed** for short, if and only if it meets the following conditions

1. $d_h \notin [i, j]$
2. $\forall k \in [i, j]$ and $k \neq h$, $d_k \in [i, j]$
3. $\forall k \notin [i, j]$, $d_k = h$ or $d_k \notin [i, j]$

Head node + full subtrees
Continuous span



Ill-formed

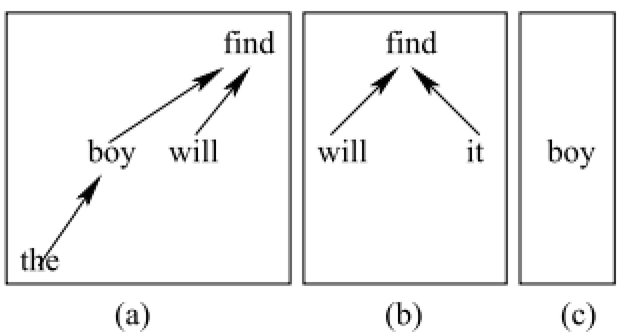


Figure 2
Fixed dependency structures.

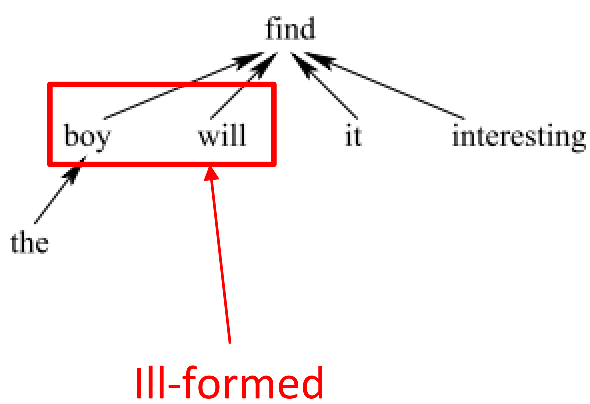
 [Shen et al., 2008, 2010]

Well-Formed Dependency Structures

A dependency structure $d_i \ldots d_j$ is **floating with children** C, for a non-empty set $C \subseteq \{i, \ldots, j\}$, or **floating** for short, if and only if it meets the following conditions

1. $\exists h \notin [i,j], s.t. \forall k \in C, d_k = h$
2. $\forall k \in [i,j]$ and $k \notin C, d_k \in [i,j]$
3. $\forall k \notin [i,j], d_k \notin [i,j]$

Sibling subtrees
Continuous span



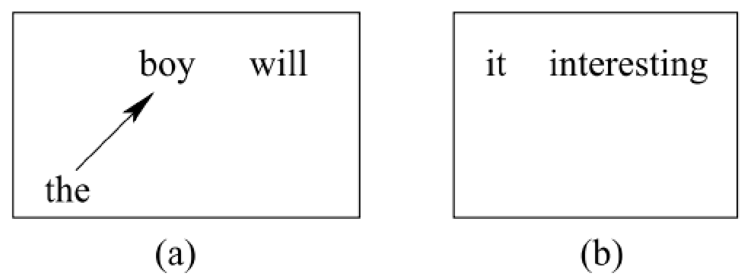


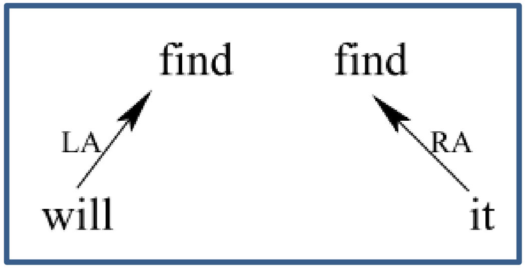
Figure 3
Floating dependency structures.

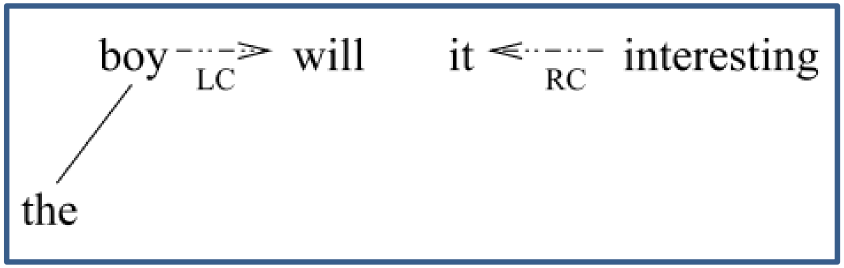
[Shen et al., 2008, 2010]

Construct Target Dependency Tree

- Four operations:

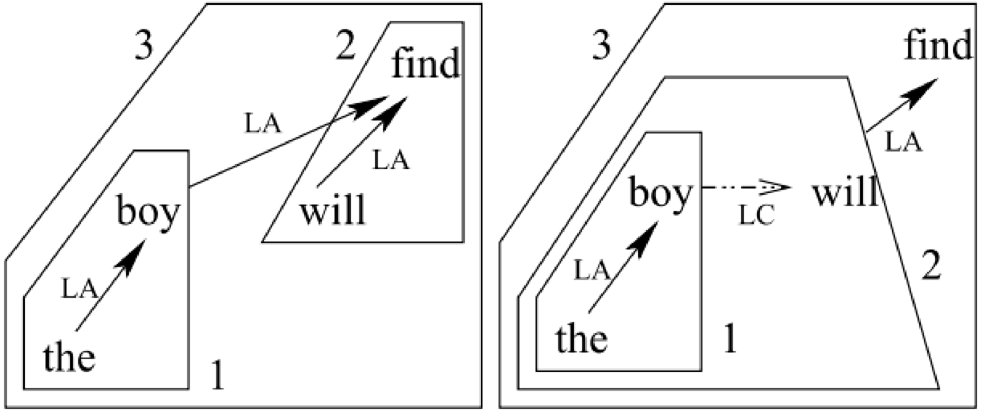
fixed





floating

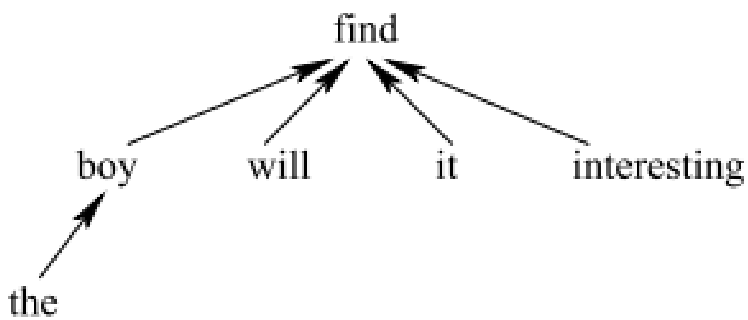
- Examples



(a) (b)

[Shen et al., 2008, 2010]

Dependency Language Model



$$
\begin{aligned}
P = \; & P_T(\text{find}) \\
& \times P_L(\text{will} \mid \text{find-as-head}) \\
& \times P_L(\text{boy} \mid \text{will, find-as-head}) \\
& \times P_L(\text{the} \mid \text{boy-as-head}) \\
& \times P_R(\text{it} \mid \text{find-as-head}) \\
& \times P_R(\text{interesting} \mid \text{it, find-as-head})
\end{aligned}
$$

root word

Left dependents: from right to left

Recursive on subtrees

Right dependents: from left to right

[Shen et al., 2008, 2010]

Training and Decoding

- Training
  - Similar to [Chiang, 2007]
  - Keep target dependency structures
  - Only extract well-formed dependency structures
- Decoding
  - Similar to [Chiang, 2007]
  - Build target dependency trees
- Non-terminal
  - POS of the head in fixed structures
  - X for floating structures

Evaluation

Tab 1: The number of rules

Model	Arabic-to-English	Chinese-to-English
baseline	337,542,137	193,922,173
filtered	32,057,337	39,005,696
str-dep	35,801,341	41,013,346
labeled	41,201,100	43,705,510

Only phrases covered by well-formed structures

POS-based non-terminals

Tab 2: Evaluation results

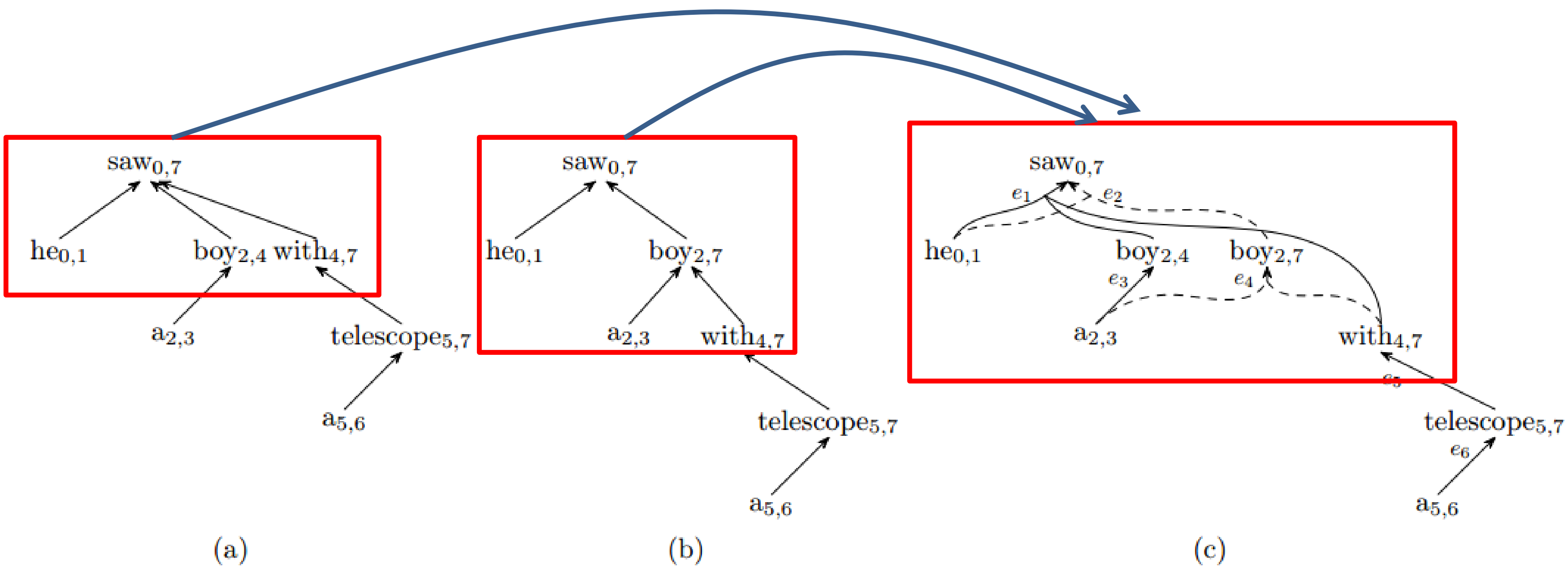
Model	BLEU		TER		METEOR
	lower	mixed	lower	mixed	
		Decoding (3-gram LM)			
baseline	36.40	34.79	54.98	56.53	57.25
filtered	36.02 (∗)	34.23 (∗)	55.29 (∗)	57.03 (∗)	57.60 (+)
str-dep	37.44 (+)	35.62 (+)	54.64 (∗)	56.47 (∗)	57.42 (+)
labeled	38.37 (+)	36.53 (+)	54.14 (+)	55.99 (∗)	58.42 (+)

Worse but use fewer translation rules

Dependency language model is useful

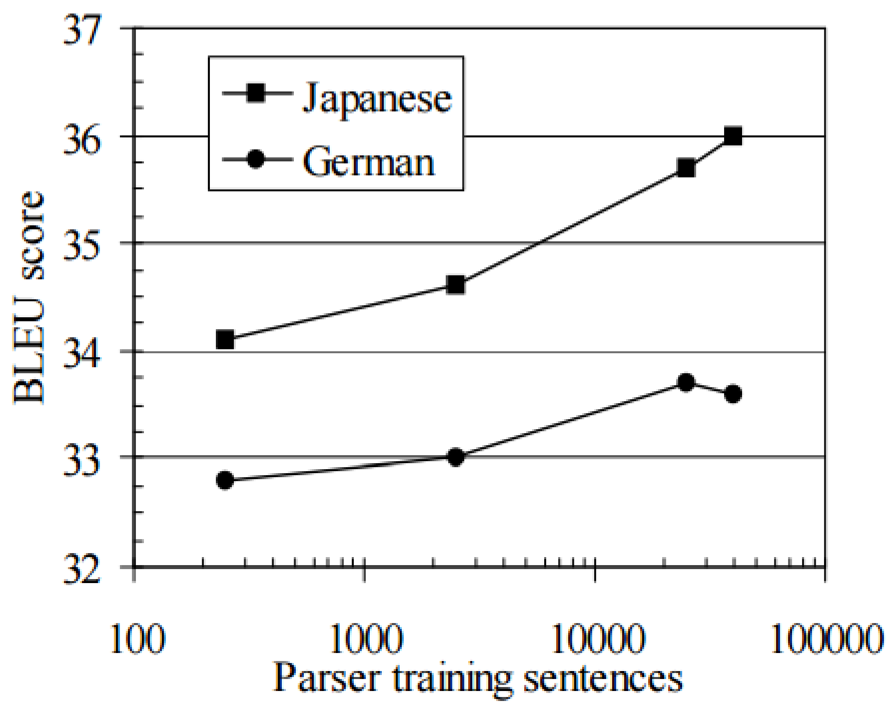
Syntactic non-terminals are helpful

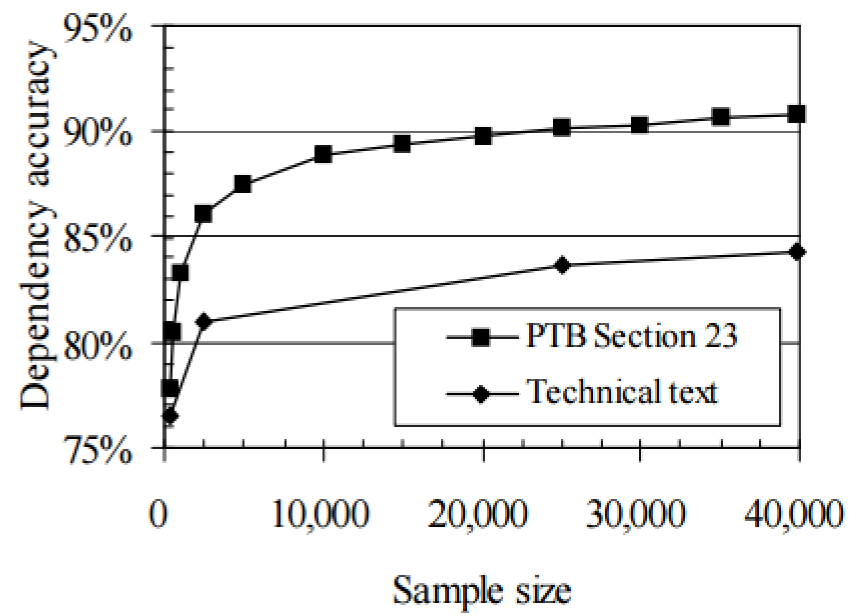
Dependency Forest



[Tu et al., 2010]

Why Dependency Forest?



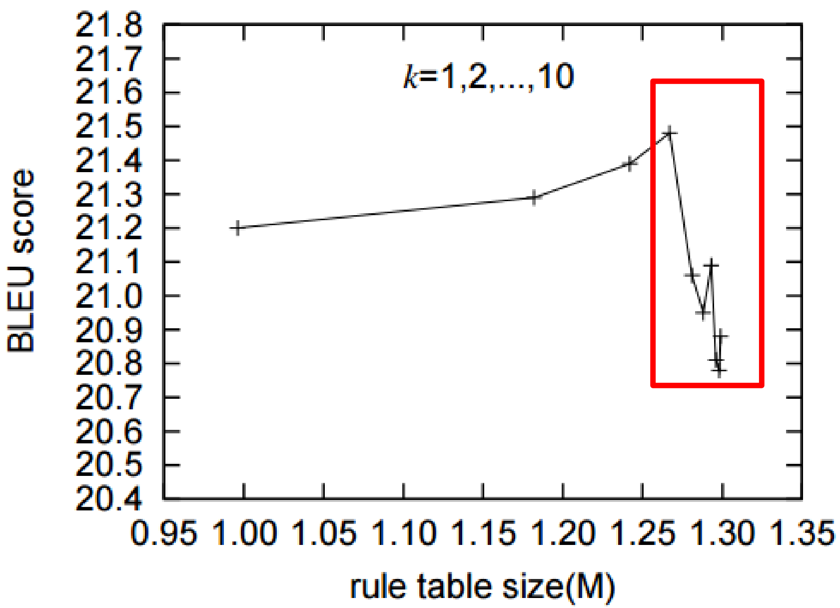


[Quirk et al., 2006]

String-to-Dependency Models

Tab1: Evaluation Result

Rule	DepLM	NIST 2004	NIST 2005	NIST 2006	time
tree	tree	33.97	30.21	30.73	19.6
tree	forest	34.42*	31.06*	31.37*	24.1
forest	tree	34.60*	31.16*	31.45*	21.7
forest	forest	**35.33****	**31.57****	**32.19****	28.5



Korean--Chinese

Tab2: model size

Rules	Size	New Rules
tree	7.2M	-
forest	7.6M	16.86%

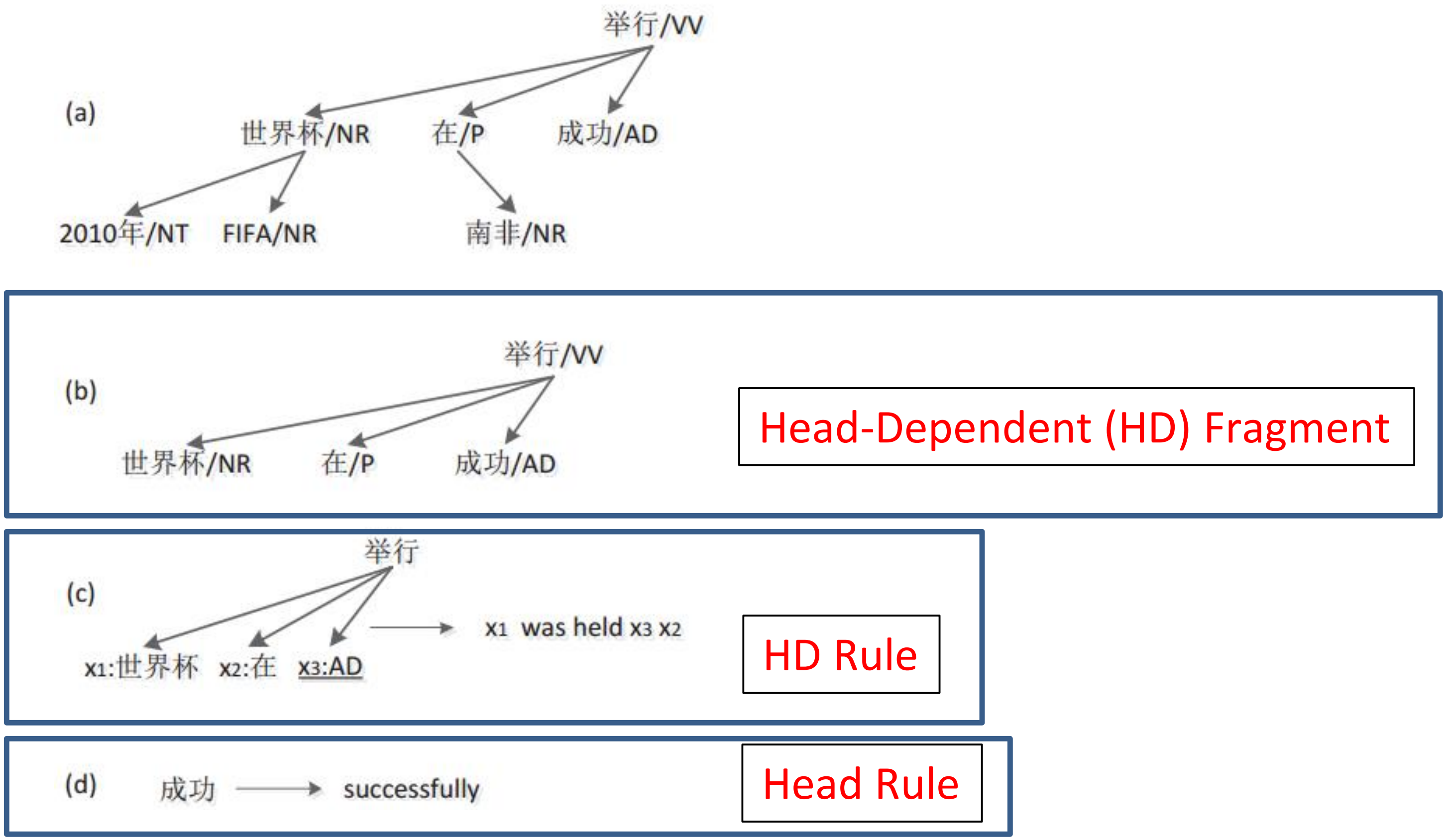
[Tu et al., 2010]

Dependency-to-String Model

- Fast decoding
  - Linear in practice [Huang et al., 2008]

- Dependency-to-string model
- Handling non-syntactic phrases

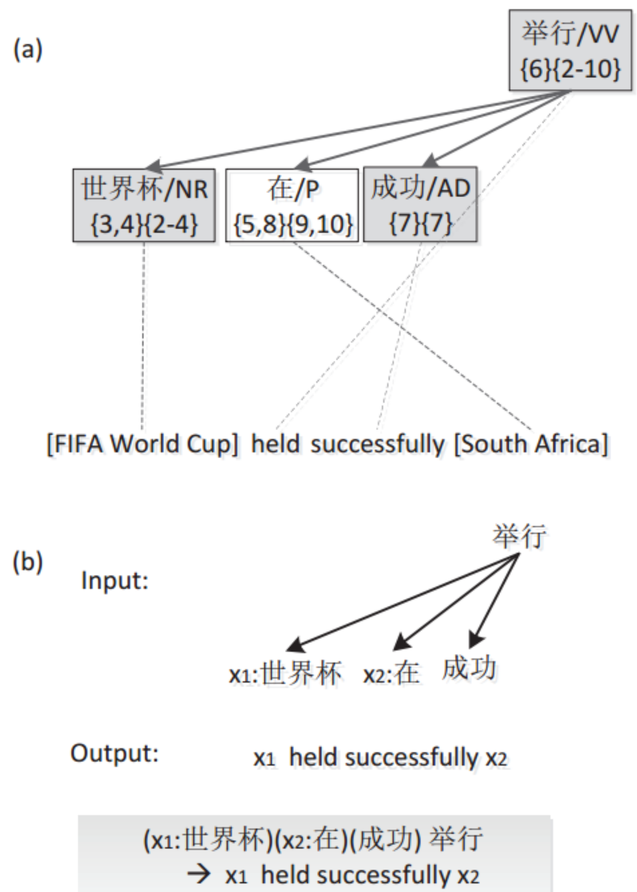
Dependency-to-String Model



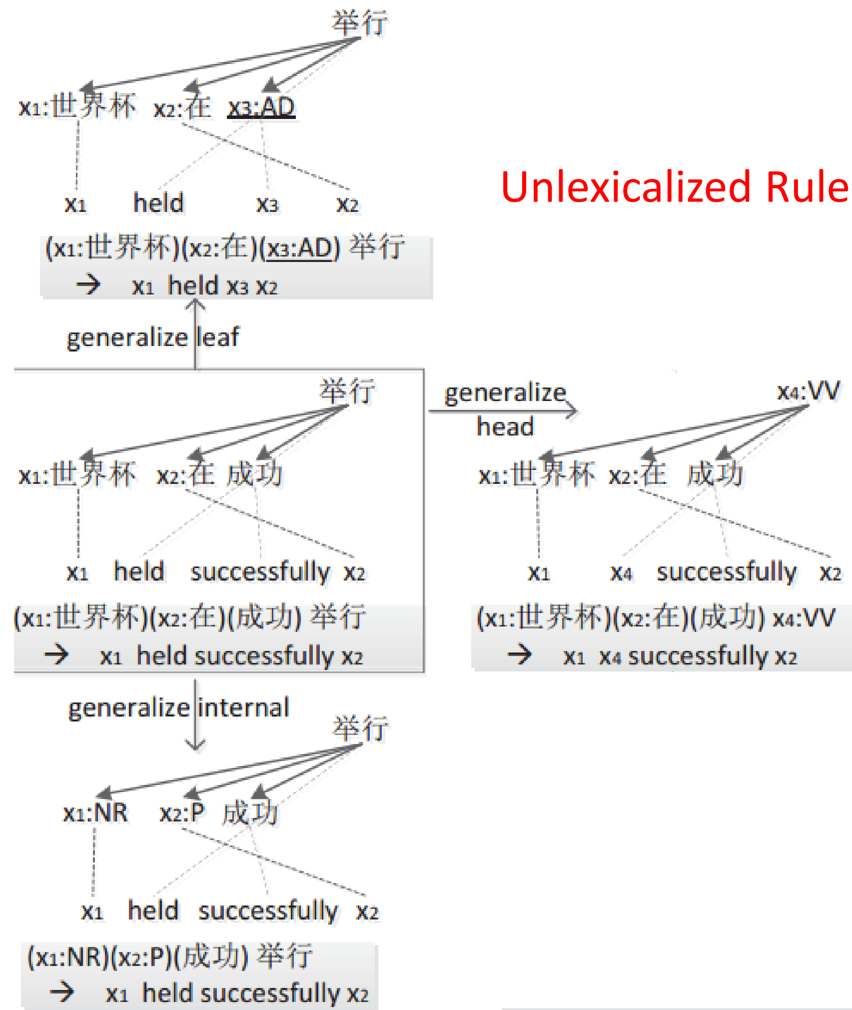
[Xie et al., 2011]

Dependency-to-String Model

Lexicalized HD Rule:



Unlexicalized Rule



 [Xie et al., 2011]

Dependency-to-String Model

- Decoding
  - CYK algorithm
  - Post-order traverse

Tab: Evaluation Results

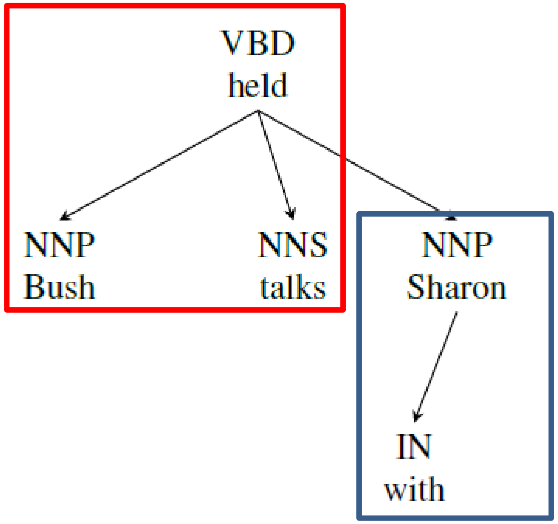
System	Rule #	MT04(%)	MT05(%)
cons2str	30M	34.55	31.94
hiero-re	148M	35.29	33.22
dep2str	56M	**35.82**$^{+}$	**33.62**$^{+}$

[Xie et al., 2011]

Handling Non-syntactic Phrases

Dependency structures are flat.

Non-syntactic phrases:
- Large number
- Local reordering
- Important to phrase coverage
- Improve systems performance



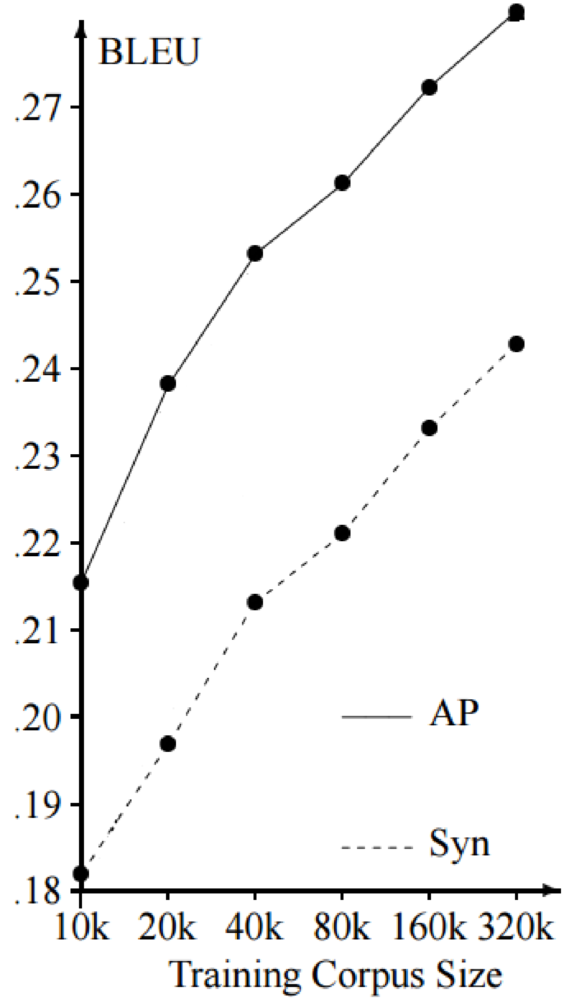
Syntactic phrases:
- Smaller amount
- Reliable
- Long-distance reordering
- Easy to use in models

Handling Non-syntactic Phrases

Important to phrase coverage and systems performance

Method	Training corpus size					
	10k	20k	40k	80k	160k	320k
AP	84k	176k	370k	736k	1536k	3152k
Syn	19k	24k	67k	105k	217k	373k

Table 1: Size of the phrase translation table in terms of distinct phrase pairs (maximum phrase length 4)



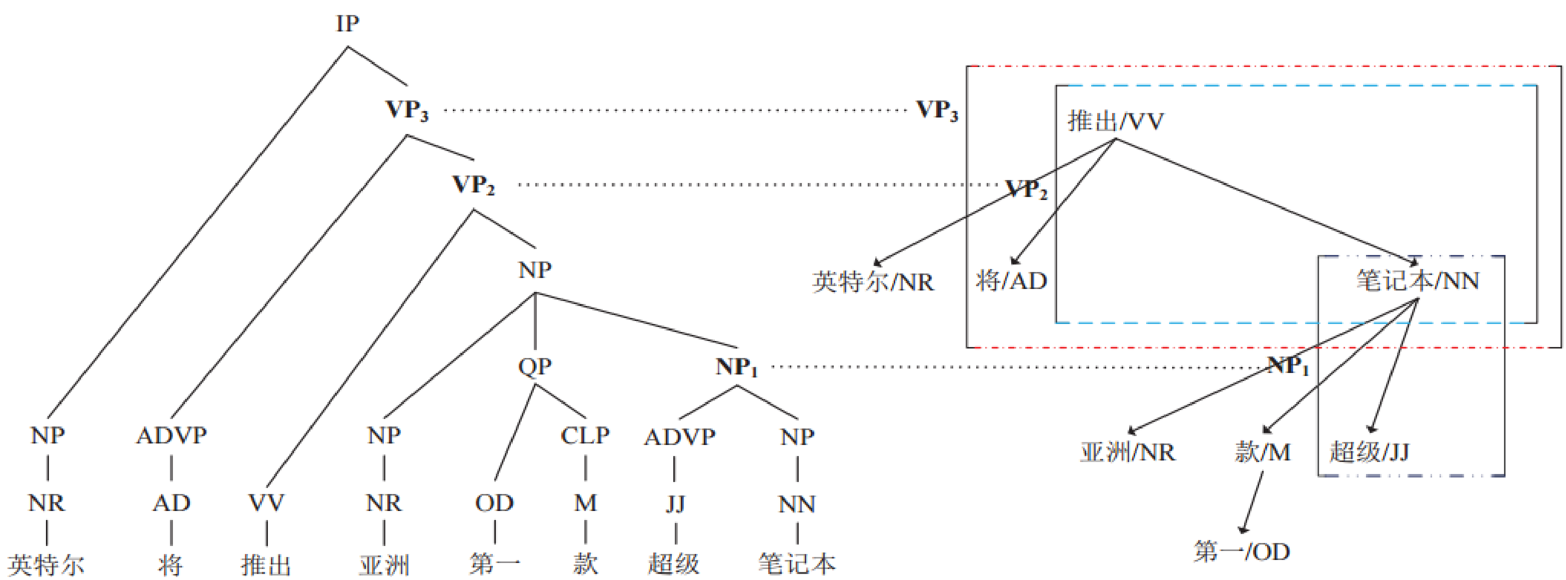
[Koehn et al., 2003]

Handling Non-syntactic Phrases

- Methods:
  - Using constituent trees
  - Integrating fixed/floating structures
  - Decomposing dependency structures

Using Constituent Tree

Phrases that cannot be captured by a dependency tree can be captured by a constituency tree

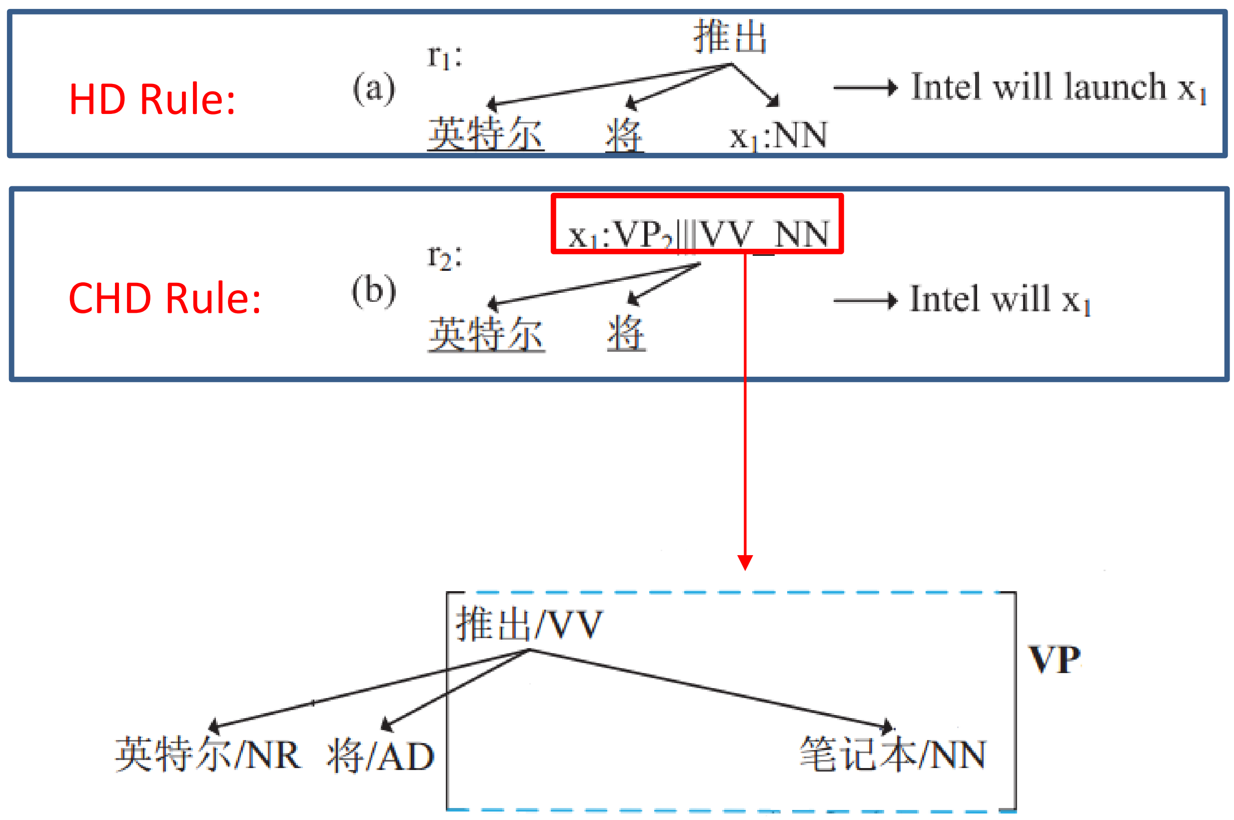


(a) (b)

	Intel	will	launch	Asia	first		super	laptop
Chinese:	英特尔	将	推出	亚洲	第一	款	超级	笔记本
English:	Intel	will	launch	the	first	Ultrabook	in	Asia

 [Meng et al., 2013]

Using Constituent Tree



[Meng et al., 2013]

Evaluation

Tab 1: Evaluation results. (+phrase pairs)

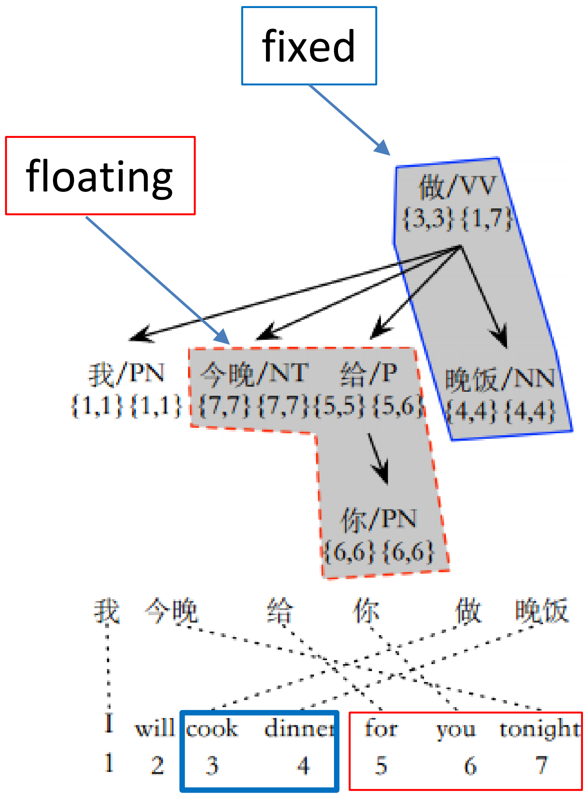
System	Rule #	MT03	MT04	MT05	Average
Moses-chart	116.4M	34.65	36.47	34.39	35.17
cons2str	25.4M+32.5M	33.14	35.12	33.27	33.84
dep2str	19.6M+32.5M	34.85	36.57	34.72	35.38
consdep2str	23.3M+32.5M	**35.57***	**37.68***	**35.62***	**36.29**

Tab 2: The proportion (%) of 1-best translations that employ CHDR-phrasal rules (CHDR-phrasal Sent.) and the proportion (%) of CHDR-phrasal rules in all CHDR rules in these translations (CHDR-phrasal Rule)

System	MT03	MT04	MT05
CHDR-phrasal Sent.	50.71	61.80	56.19
CHDR-phrasal Rule	10.53	13.55	10.83

[Meng et al., 2013]

Integrating Fixed/Floating Structures



System	Rule#	MT03	MT04	MT05	Average
Moses-Chart	116.4M	34.65	36.47	34.39	35.17
dep2str	37M+32.5M	34.92	36.82	34.71	35.48
dep2str-aug	37M+32.5M	**35.66*** (**+0.74**)	**37.61*** (**+0.79**)	**35.74*** (**+1.03**)	**36.33** (**+0.85**)

The same number of rules:
- Use bilingual phases during decoding
- But focus on phrases covered by fixed/floating structures

[Xie et al., 2014]

Dependency Decomposition

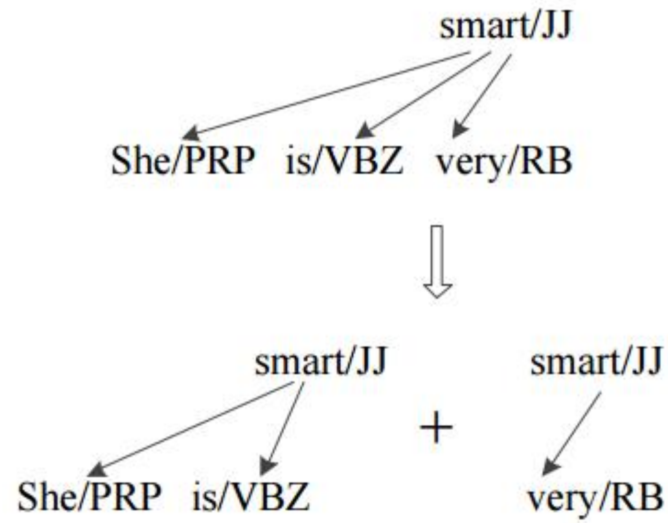
Formal definition:

$$
\begin{aligned}
L_i \cdots L_1 H R_1 \cdots R_j \\
&= L_m \cdots L_1 H R_1 \cdots R_n \\
&+ L_i \cdots L_{m+1} H R_{n+1} \cdots R_j
\end{aligned}
$$

subject to

$$
\begin{aligned}
& i \geq 0, j \geq 0 \\
& i \geq m \geq 0, j \geq n \geq 0 \\
& i + j > m + n > 0
\end{aligned}
$$

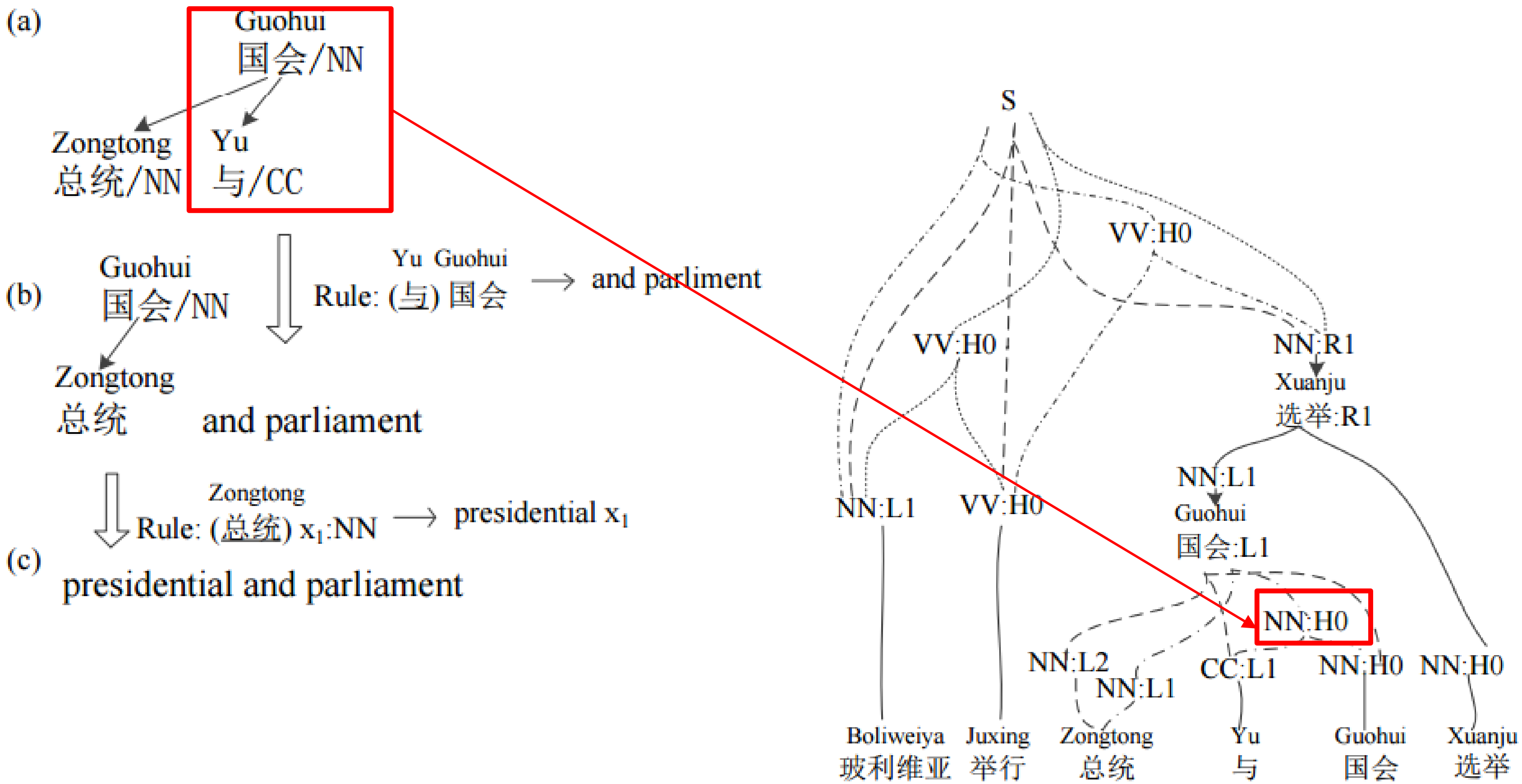
Example:



During training: extract more rules

During decoding: translate an HD fragment in two steps

Decomposition During Decoding



Evaluation

Tab 1: Influence of decomposition

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
HPBMT	36.5	34.3	**20.5**	**23.0**
D2S	35.1	33.1	20.0	22.3
+Decomp	**36.6***	**34.9***	**20.4***	22.7*

Tab 2: Influence of phrase pairs

System	ZH–EN		DE–EN	
	MT04	MT05	WMT12	WMT13
HPBMT	36.5	34.3	20.5	23.0
D2S+Decomp	36.6	34.9	20.4	22.7
+Phrase	**37.7***	**35.5***	**20.8***	**23.4***

Tab 3: Rule number

System	# Rules	
	ZH–EN	DE–EN
HPBMT	388M	684M
D2S	27M	41M
+Decomp	84M	92M
+Phrase	161M	206M

 [Li et al., 2014]

Revisit Non-syntactic Phrases

- Non-syntactic phrases exist in linguistically syntax-based models
  - STSG (over SCFG)
  - Focus on subtrees
  - Same generative capability on string pairs
  - Stronger generative capability on tree pairs
- Add patches to tree-based models [previous slides]

Revisit Non-syntactic Phrases

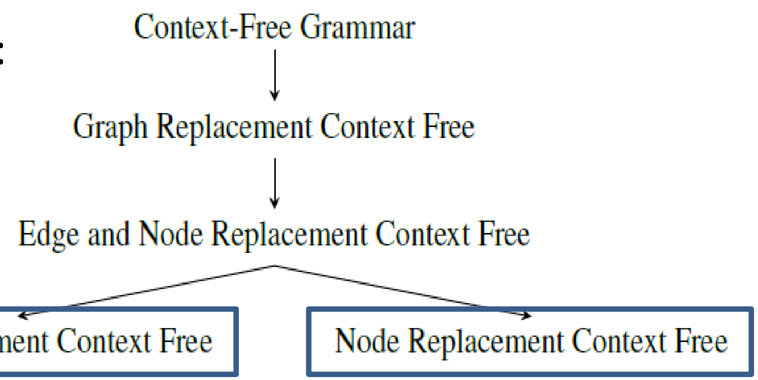
- Graphs vs Trees
  - More complex structures
  - More powerful to model sentences
    - AMR for semantic, graphs for feature structures
  - Graph grammars
  - Non-syntactic phrases could be connected
  - Subgraphs, without the definitions of syntactic and non-syntactic phrases

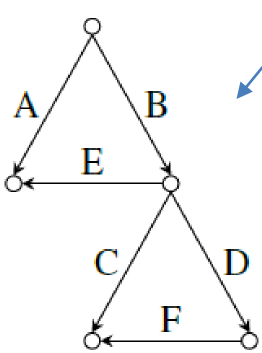
Dependency Graph-to-String Models

- Graph grammars
  - Edge replacement grammar (ERG)
  - Node replacement grammar (NRG)
- Models based on graph grammars
  - ERG-based model
  - NRG-based model

Graph Grammars

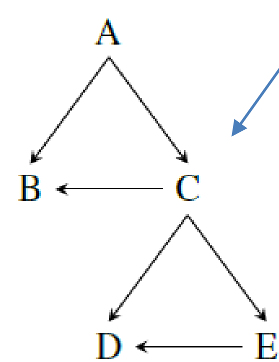
Hierarchy of graph grammars:





Ignore node label in this tutorial

(a) Edge-labeled graph



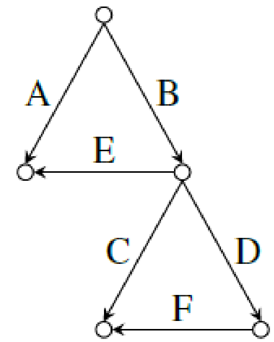
Ignore edge label in this tutorial

(b) Node-labeled graph

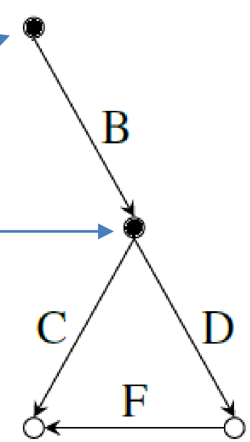
[Kukluket al., 2008]

Edge Replacement Grammar

- Graph
  - Edge-labeled
  - Directed



- Graph fragment definition
  - Basic deviation units
  - Graph
  - External nodes
  - Prevent hyperedges



[Li et al., 2015]

Edge Replacement Grammar

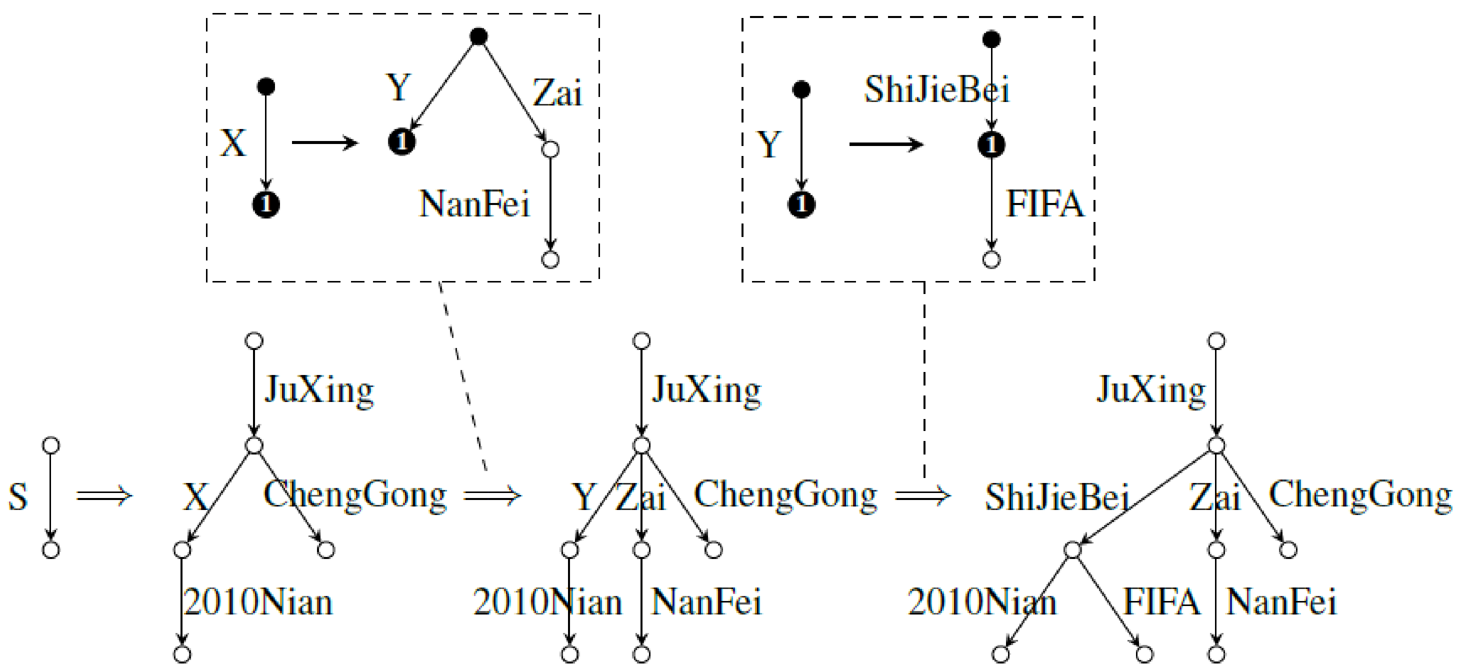
An *edge replacement grammar* is a tuple $\langle N, T, P, S \rangle$, where

- N and T are disjoint finite sets of non-terminal symbols and terminal symbols, respectively.
- P is a finite set of productions of the form $A \to R$, where $A \in N$ and R is a graph fragment, where edge-labels are from $N \bigcup T$.
- $S \in N$ is the start symbol.

[Li et al., 2015]

Edge Replacement Grammar

- Derivation



Synchronous Edge Replacement Grammar

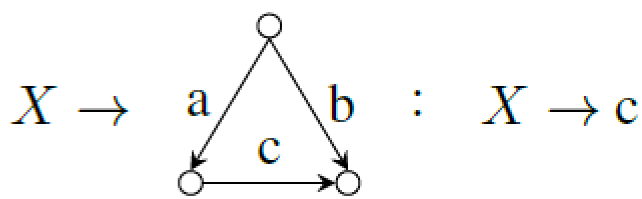
A *synchronous ERG* (SERG) is a tuple $\langle N, T, T', P, S \rangle$, where

- N is a finite set of non-terminal symbols.
- T and T' are finite sets of terminal symbols.
- $S \in N$ is the start symbol.
- P is a finite set of productions of the form $\langle A \rightarrow R, A \rightarrow R', \sim \rangle$, where $A \in N$, R is a graph fragment over $N \bigcup T$ and R' is a graph fragment over $N \bigcup T'$. $\sim$ is a one-to-one mapping between non-terminal symbols in R and R'.

[Li et al., 2015]

Synchronous Edge Replacement Grammar

- SERG has a stronger generative capacity over structure pairs than both SCFG and STSG
  - STSG has a stronger generative capacity over structures than SCFG [Chiang, 2012]
  - Any STSG can easily be converted into an SERG by labeling edges in tree structures
  - The following SERG generates a trivial example of a graph pair, which no STSG can generate

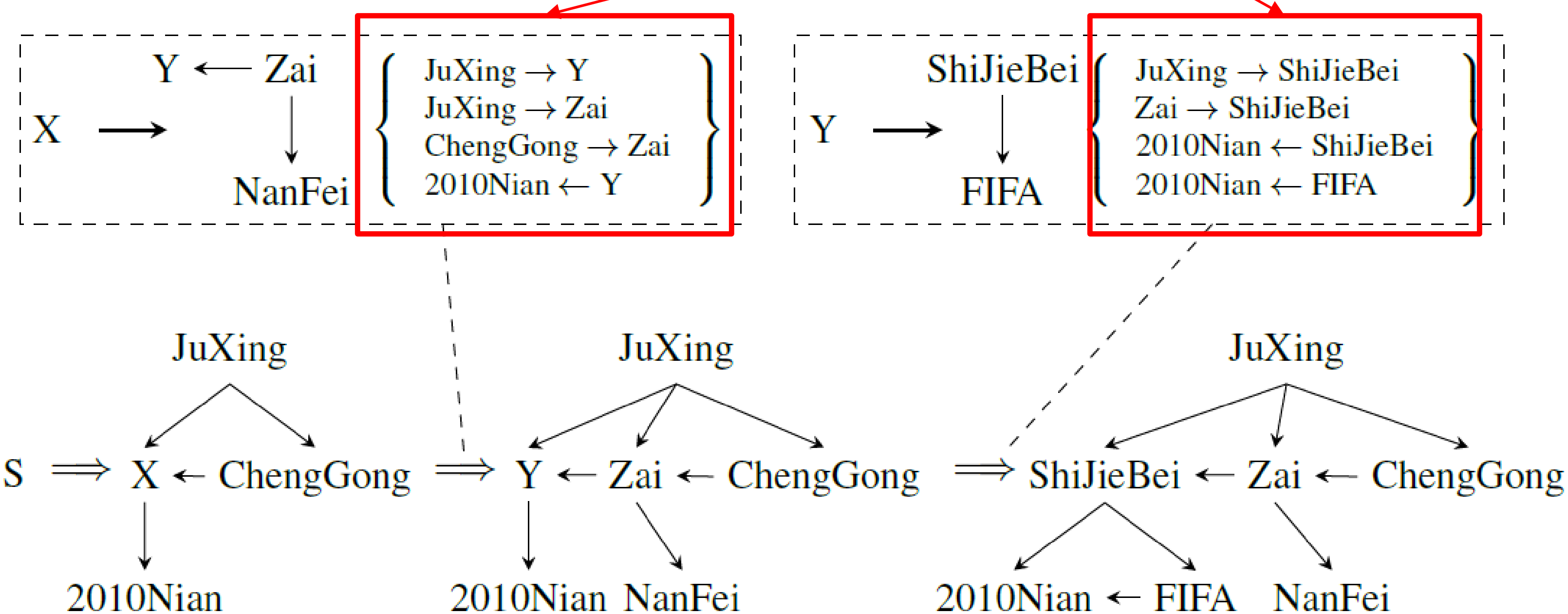


[Li et al., 2015]

Node Replacement Grammar

- Derivation

Embedding mechanism which can be ignored during parsing [Kukluket al., 2008]



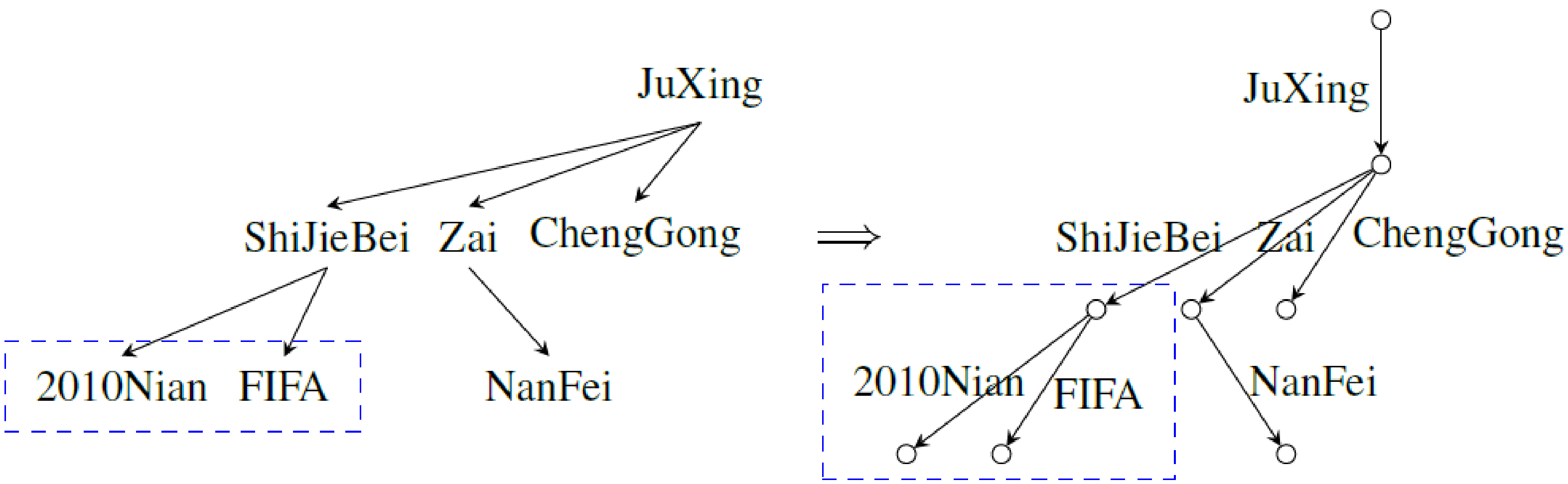
Synchronous Node Replacement Grammar

- For machine translation
- SNRG has a stronger generative capacity over structure pairs than both SCFG and STSG

ERG-Based Model

- Create edge-labeled graphs
- Practical restrictions
- Training
- Decoding

Create Edge-Labeled Graphs



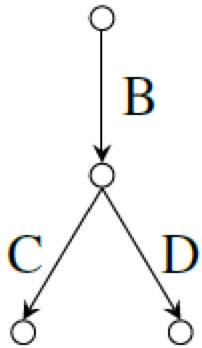
[Li et al., 2015]

Practical Restrictions

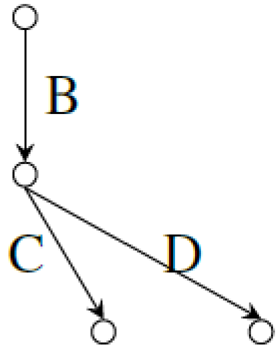
- Word-order restriction
- Continuity restriction
- Non-terminal restriction

Word-Order Restriction

- Keep word order



(a) C B D



(b) B C D

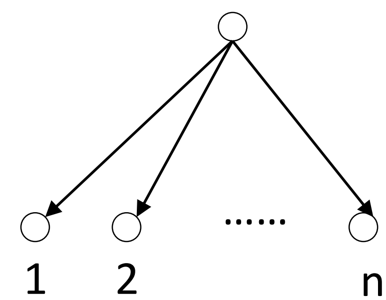
Continuity Restriction

- Subgraphs cover continuous phrase (from exponential to polynomial)

Decoding Process

Data: Input graph G of a sentence s
Result: Translation t

```
for span length l = 1 to l_s do
    for all subgraph g of size l do
        for all rule r do
            if r can be applied to g then
                create new hypothesis h ;
                add h to chart ;
            end
        end
    end
end
```

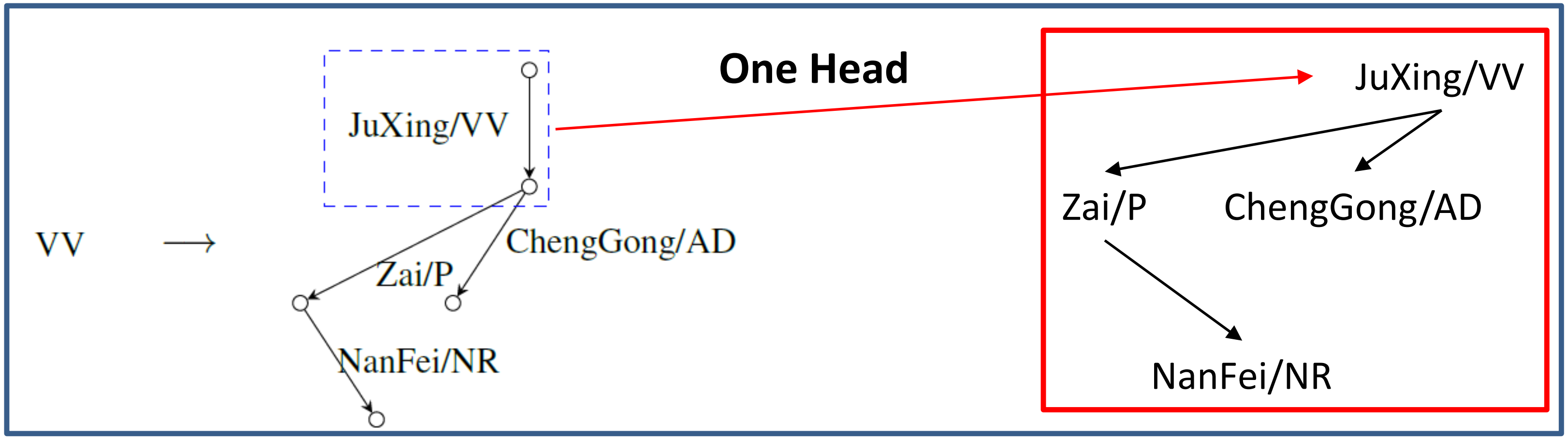


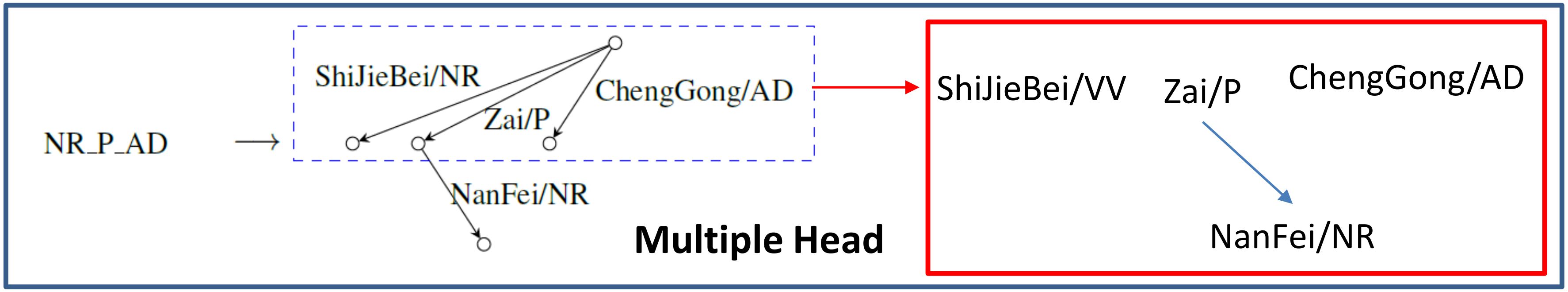
$O(2^n)$ subgraphs

$\Downarrow$ continuity

$O(n^2)$ subgraphs

Non-terminal Restriction

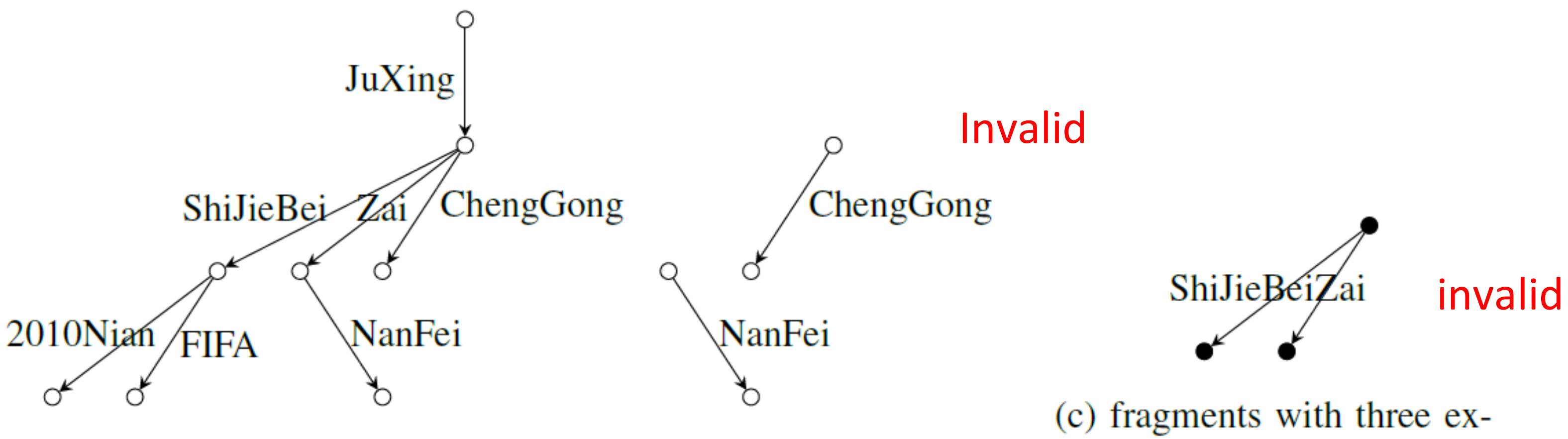




Training

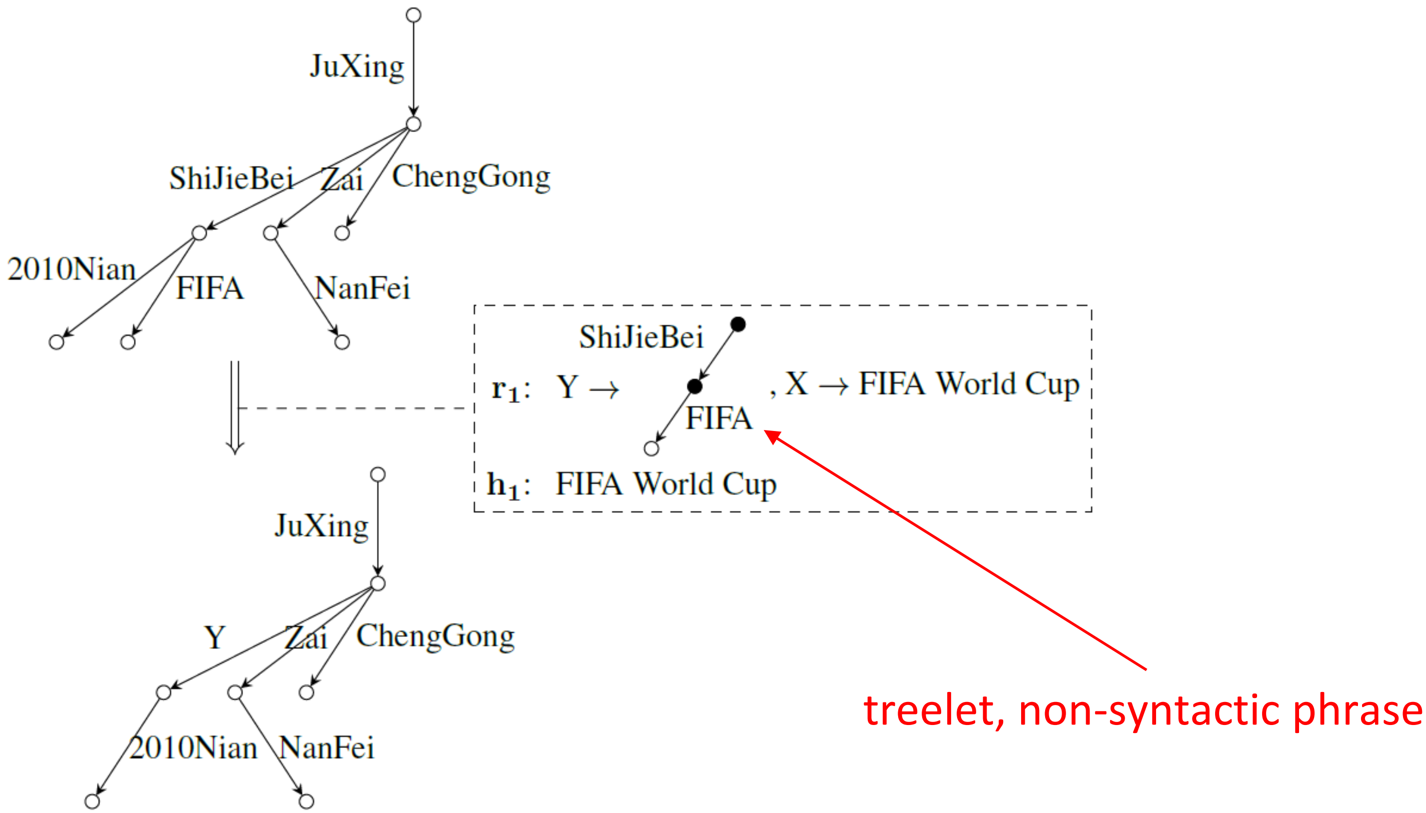
Similar to [Chiang, 2007], but:

- Check if the source side is a valid graph
- Keep dependency structures in rules
- Induce non-terminals for the source side



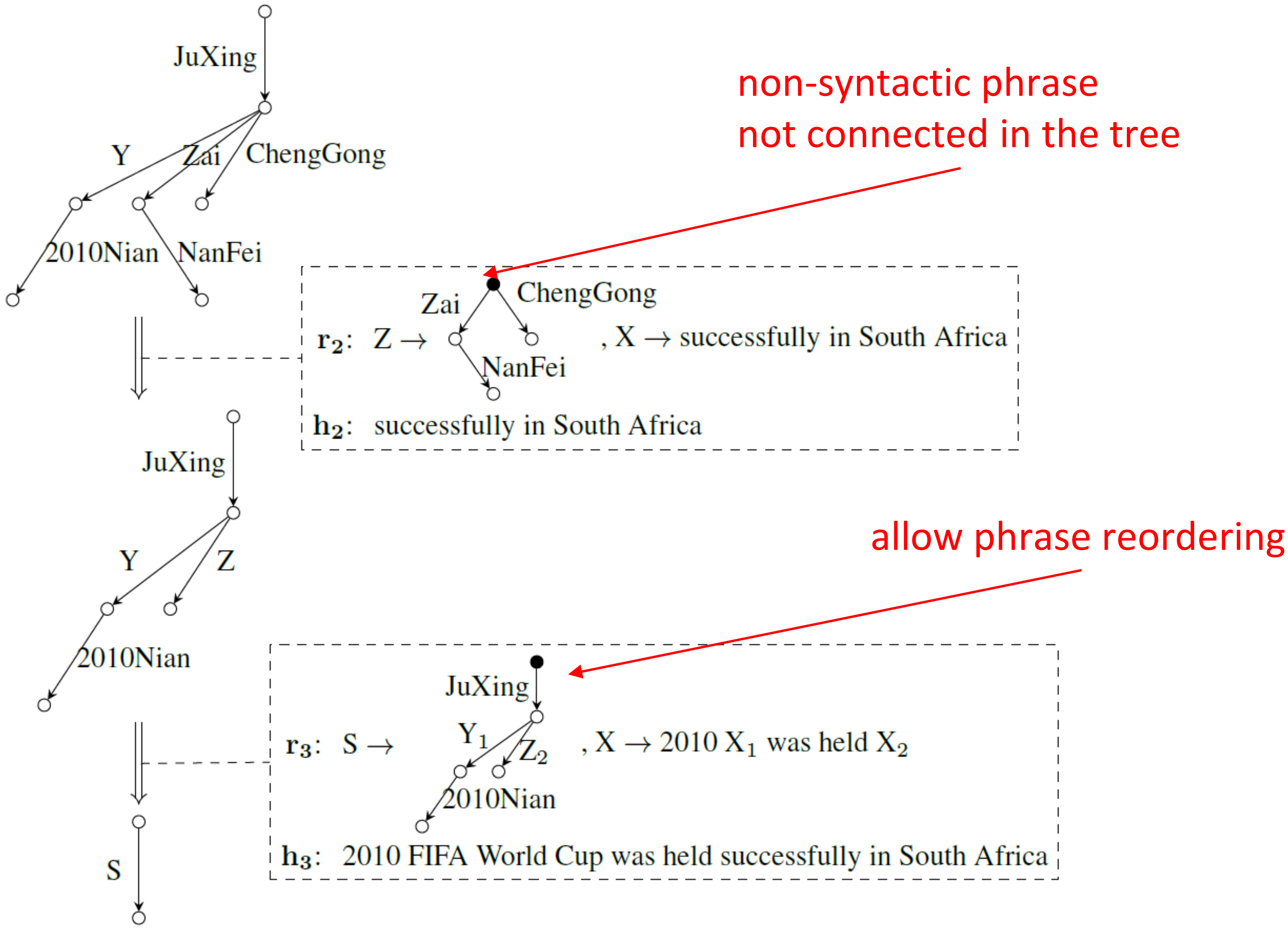
(a) DEG (b) unconnected fragment (c) fragments with three external nodes

Decoding



JuXing
ShiJieBei
Zai
ChengGong
2010Nian
FIFA
NanFei
ShiJieBei
$\mathbf{r_1}$: Y →
, X → FIFA World Cup
FIFA
$\mathbf{h_1}$: FIFA World Cup
JuXing
Y
Zai
ChengGong
2010Nian
NanFei
treelet, non-syntactic phrase

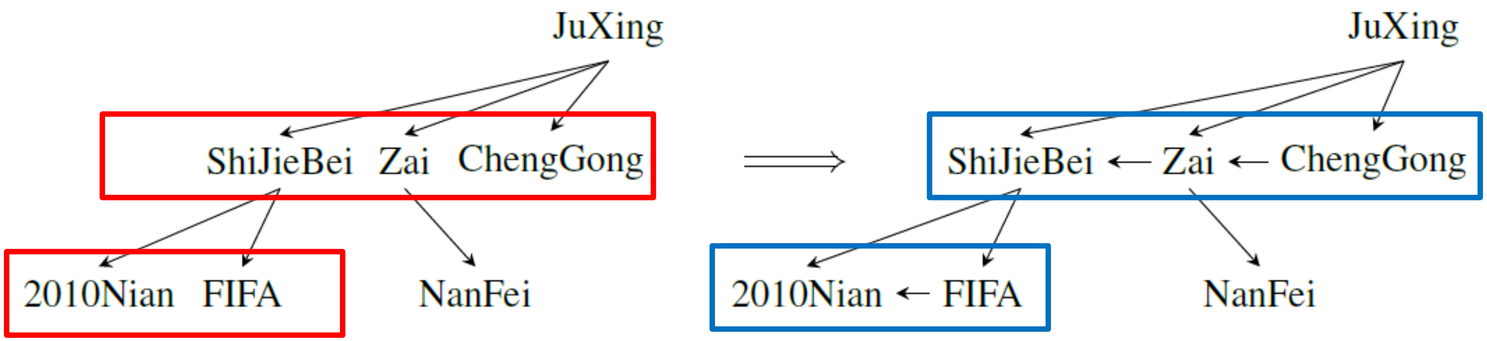
Decoding



JuXing
Y
Zai
ChengGong
2010Nian
NanFei
non-syntactic phrase
not connected in the tree
Zai
ChengGong
$\mathbf{r_2}$: Z →
, X → successfully in South Africa
NanFei
$\mathbf{h_2}$: successfully in South Africa
JuXing
Y
Z
allow phrase reordering
2010Nian
JuXing
$\mathbf{r_3}$: S →
Y_1
Z_2
, X → 2010 X_1 was held X_2
2010Nian
$\mathbf{h_3}$: 2010 FIFA World Cup was held successfully in South Africa
S

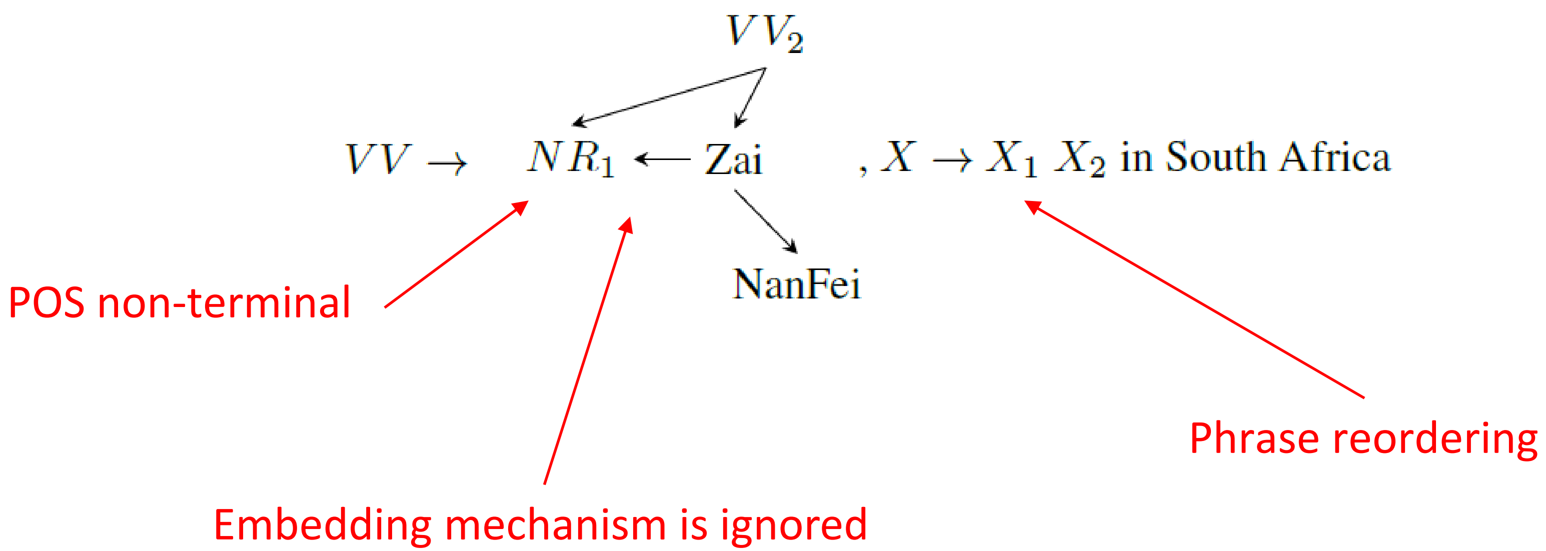
NRG-Based Model

- Node-labeled graphs

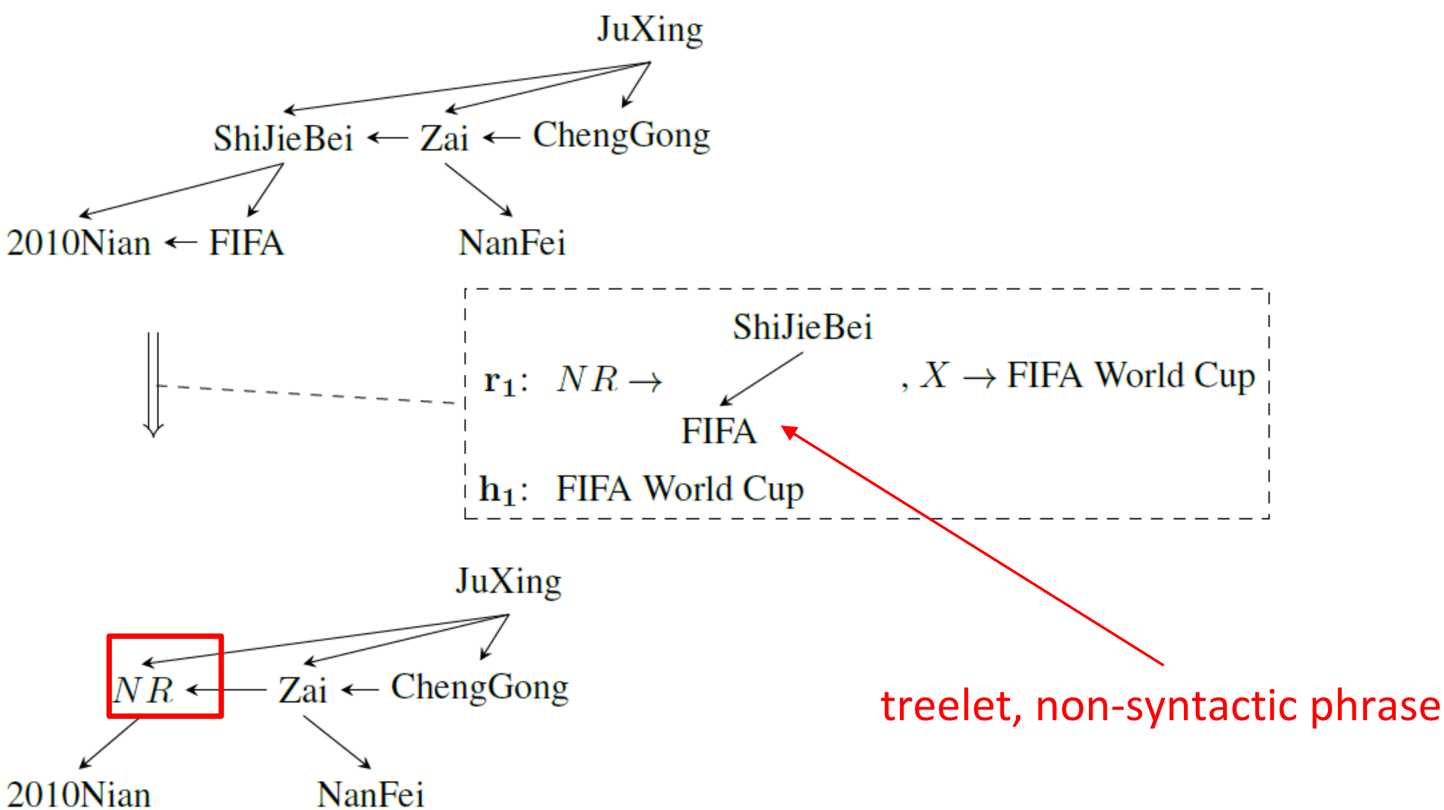


NRG-Based Model

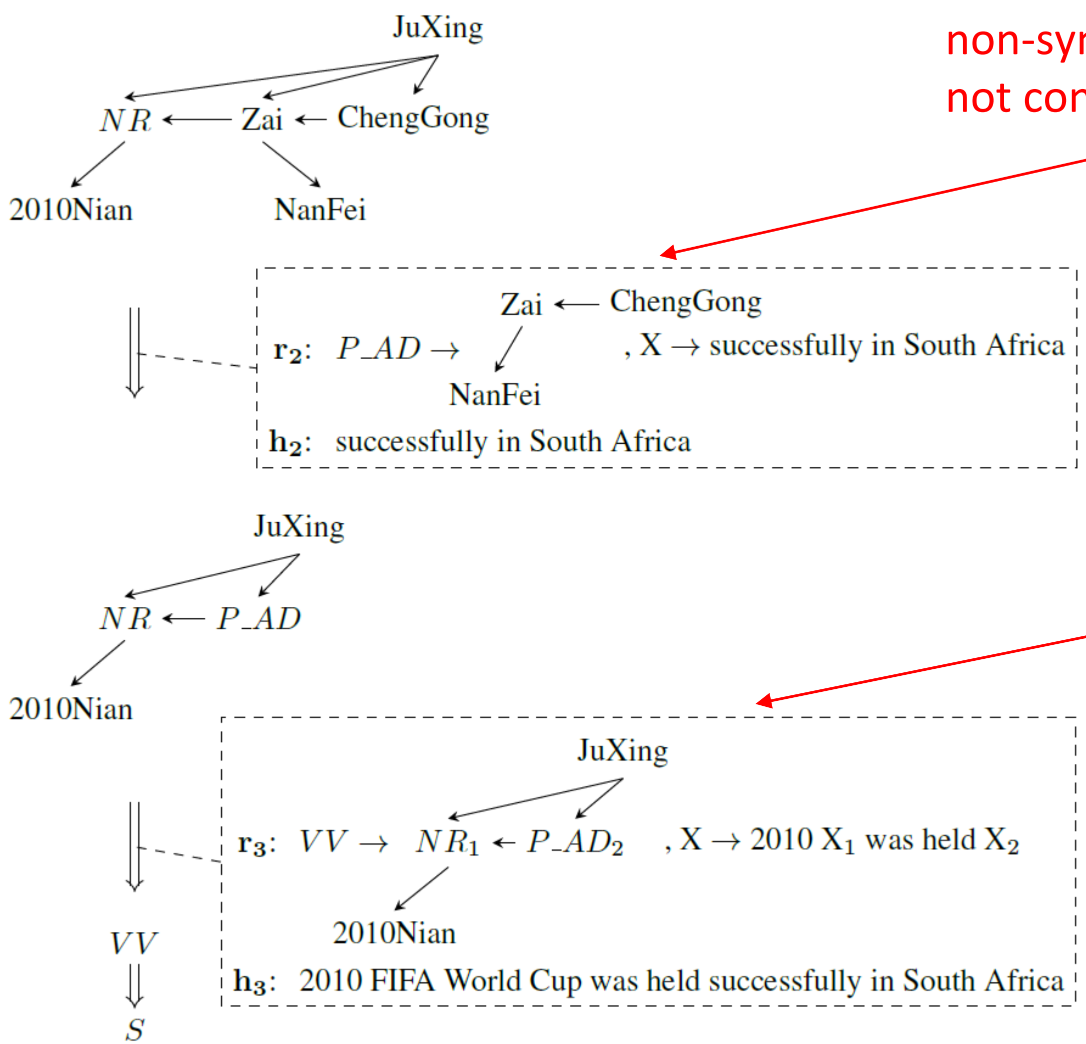
- The same practical restrictions
- Similar training and decoding processes
- Rule example:



NRG-Based Model



NRG-Based Model



non-syntactic phrase
not connected in the tree

allow phrase reordering

Evaluation

Tab 1: BLEU scores

System	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
HPBMT	36.5	34.3	20.5	23.0
SERG	**37.7**	**35.8**	20.6	**23.2**
SNRG	**37.7**	**35.8**	20.7	**23.4**

Tab 3: influence of sibling edges

System	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
SNRG	**37.7**	**35.8**	**20.7**	**23.4**
-Sib	33.7	32.0	19.8	22.3

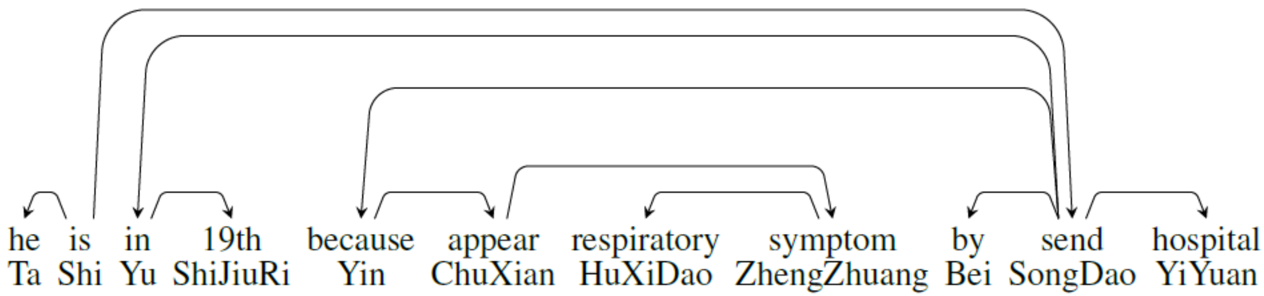
Tab 2: Influence of POS non-terminals

System	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
SERG	**37.7**	**35.8**	**20.6**	**23.2**
-NT	37.0	34.9	20.1	22.8
SNRG	**37.7**	**35.8**	20.7	23.4
-NT	37.2	34.7	20.7	**23.6**

Tab 4: Influence of edge types

System	ZH–EN MT04	ZH–EN MT05	DE–EN WMT12	DE–EN WMT13
SNRG	37.7	**35.8**	20.7	23.4
+ET	37.6	35.4	20.8	23.5

Evaluation



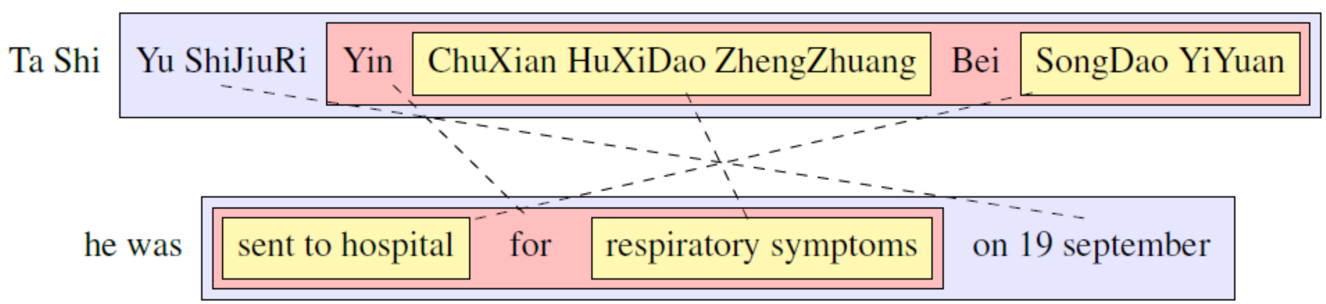
Ref: he was sent to the hospital for respiratory symptoms on the 19th

HPBMT: he is in 19 due to respiratory symptoms were sent to the hospital

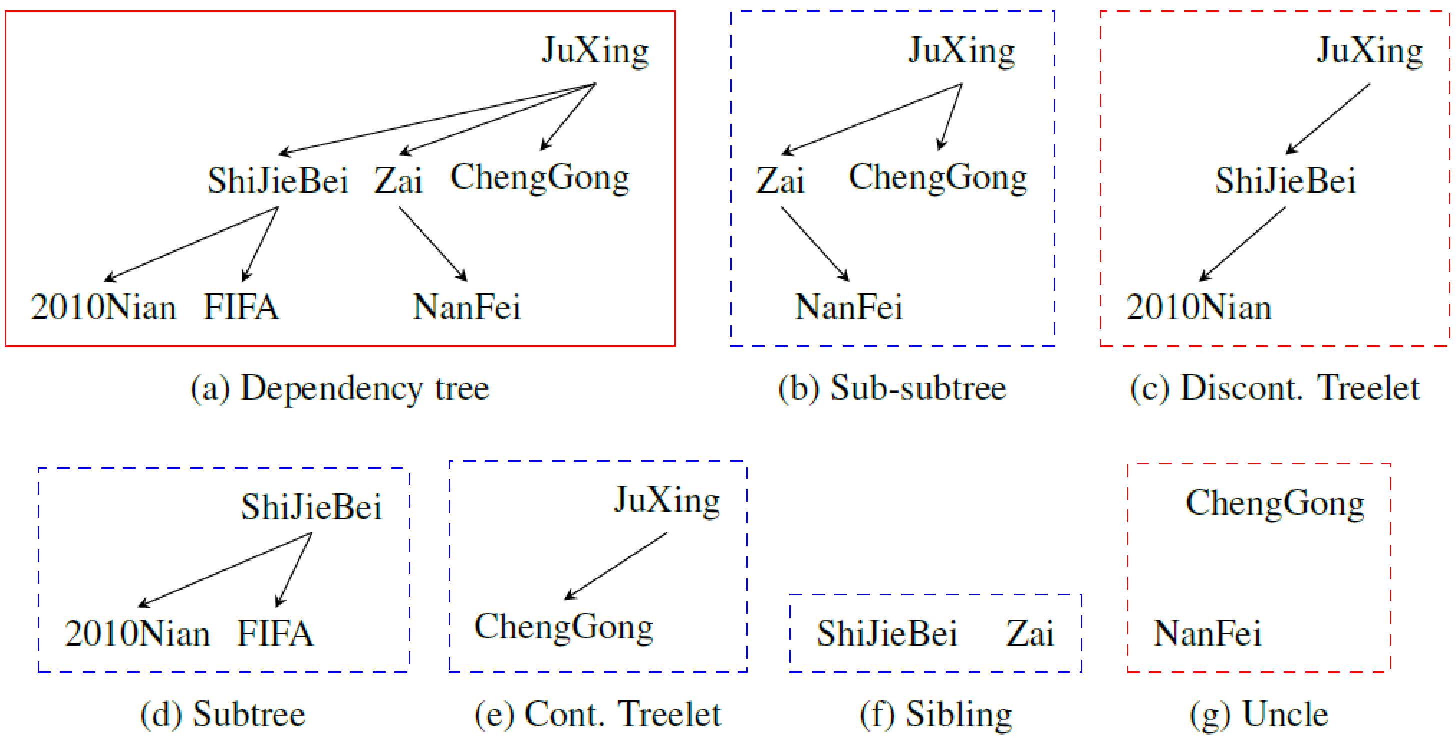
SERG: he was sent to hospital for respiratory symptoms on 19 september

SNRG: he was sent to hospital for respiratory symptoms on 19 september

⇩ Correct reordering



Evaluation



Given the dependency tree in (a), SERG and SNRG can cover dependency configurations (b), (d), (e), and (f). *Discont. Treelet* denotes a treelet covering a discontinuous phrase while *Cont. Treelet* means a treelet covering a continuous phrase.

Summary

- Models based on synchronous grammars can learn recursive rules.
- Non-terminals in recursive rules are used for target-phrase reordering
- Graph grammars
  - SERG
  - SNRG

References

- David Chiang (2007). Hierarchical Phrase-Based Translation. In: *Computational Linguistics* 33.2, pages 201–228.
- David Chiang (2012). Grammars for Language and Genes: Theoretical and Empirical Investigations. Springer
- Liang Huang and Haitao Mi (2010). Efficient Incremental Decoding for Tree-to-string Translation. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, pages 273–283
- Jacek Kukluk (2007). Inference of Node and Edge Replacement Graph Grammars. PhD thesis. University of Texas at Arlington.
- Fandong Meng, Jun Xie, Linfeng Song, Yajuan Lu, and Qun Liu (2013). Translation with Source Constituency and Dependency Trees. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, pages 1066–1076
- Chris Quirk and Simon Corston-Oliver (2006). The Impact of Parse Quality on Syntactically informed Statistical Machine Translation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, pages 62–69.
- Libin Shen, Jinxi Xu, and Ralph Weischedel (2010). String-to-Dependency Statistical Machine Translation. In: *Computational Linguistics* 36.4, pages 649–671.
- Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin (2010). Dependency Forest for Statistical Machine Translation. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Volume 2)*. Beijing, China, pages 1092–1100
- Jun Xie, Haitao Mi, and Qun Liu (2011). A Novel Dependency-to-string Model for Statistical Machine Translation. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, United Kingdom, pages 216–226.
- Jun Xie, Jinan Xu, and Qun Liu (2014). Augment Dependency-to-String Translation with Fixed and Floating Structures. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, pages 2217–2226
- Liangyou Li, Andy Way, Qun Liu. (2015). Dependency Graph-to-String Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, Lisbon, Portug

Q&A

- Introduction
- Dependency-Based MT Evaluation
- Translation Models Based on Segmentation
- Translation Models Based on Synchronous Grammars
- Conclusion
- Lab Session

CONCLUSION

SMT Benefits From Structures

- Sequence-based
  - Phrase-based
- Tree-based
  - Hierarchical phrase-based
  - Tree-to-string
  - String-to-tree
  - Tree-to-tree
  - Forest-based
  - **Dependency-based**
- Graph-based
  - Semantic-based
  - **Dependency graph-based**

Dependency-Based Evaluation

- Automatic evaluation is important
  - Lexical
  - Semantic
  - Syntactic
- Dependency structures and relations provide rich information for evaluation
  - Subtree, head-word chain, fixed/float structures
  - Dependency relations
  - RNN

Segmentational Dependency-Based Models

- Segmenting dependency structures provide various translation units
  - Edge
  - Path
  - Treelet
- Dependency graphs provide subgraphs as the basic translation units.

Recursive Dependency-Based Models

- Synchronous grammars provide theoretical foundation for SMT
- Recursive rules provide information on how to perform phrase reordering
- SMT systems also benefit from linguistic non-terminals
- Tree-based models are weak at translating non-syntactic phrases
- Dependency graphs naturally take various phrases into consideration

Thank you very much !

Q&A

- Introduction
- Dependency-Based MT Evaluation
- Translation Models Based on Segmentation
- Translation Models Based on Synchronous Grammars
- Conclusion
- Lab Session

Dependency-Based Models

Dependency Format

Download and Try

LAB SESSION

Dependency-Based Models

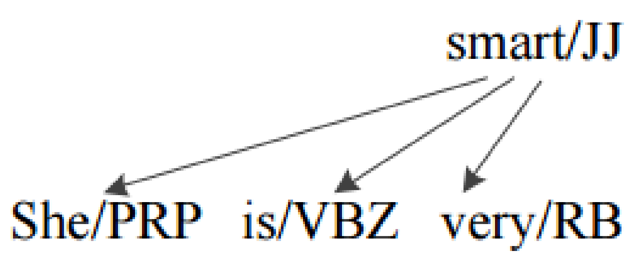
- **Dependency tree-to-string model**
  - Liangyou Li, Jun Xie, Andy Way, Qun Liu. (2014). Transformation and Decomposition for Efficiently Implementing and Improving Dependency-to-String Model In Moses. In *Proceedings of SSST-8*.
- **Segmentational graph-based model**
  - Liangyou Li, Andy Way, Qun Liu. (2016). Graph-Based Translation Via Graph Segmentation. In *Proceedings of ACL*.
- **Context-ware segmentational graph-based model**
  - Liangyou Li, Andy Way, Qun Liu. (2016). Context-Aware Segmentation for Graph-Based Translation. Submitted to *EACL 2017*.
- **SERG-based dependency graph-to-string model**
  - Liangyou Li, Andy Way, Qun Liu. (2015). Dependency Graph-to-String Translation. In *Proceedings of EMNLP*.
- **SNRG-based dependency graph-to-string model**
  - Paper in preparation

Dependency Format

- Using factors
  - Word | POS | fid | relation

She|PRP|3|nsubj is|VBZ|3|cop very|RB|3|advmod smart|JJ|-1|ROOT

⇕



Dependency Format

- moses-graph/scripts/training/stanford-dep-2-factor.perl

```
nsubj(smart-4, She-1)
cop(smart-4, is-2)
advmod(smart-4, very-3)
root(ROOT-0, smart-4)
```

⇩

She|PRP|3|ROOT is|VBZ|3|cop very|RB|3|advmod smart|JJ|-1|ROOT

Download and Try

- Binaries, sample data, and lab instructions
  - https://drive.google.com/drive/folders/0BzwIbrtQHxlLZ2hlTjVKWnNqWkk?usp=sharing
- Source codes
  - git clone https://llysuda@bitbucket.org/llysuda/moses-graph.git

Or download from my webpage:

http://www.computing.dcu.ie/~liangyouli

Please follow the instructions to build your models ☺