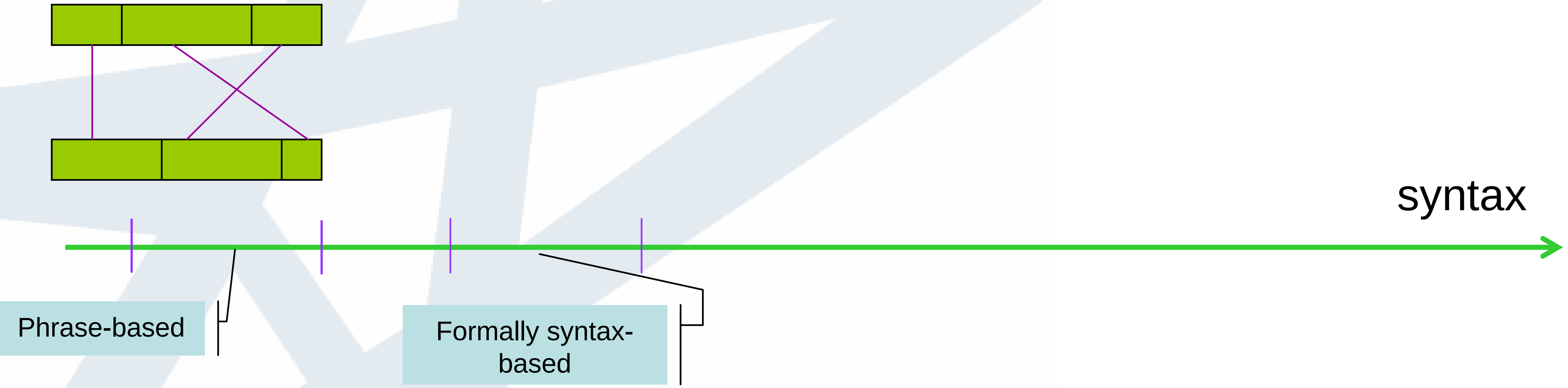
Chime Seminar

Linguistically Annotated BTG for Statistical Machine Translation

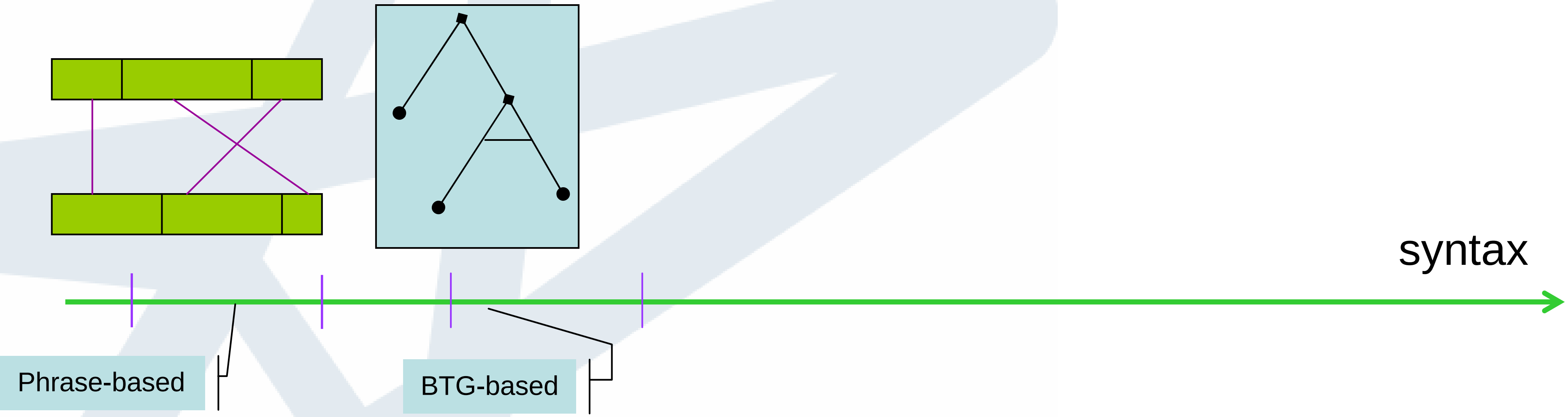
Deyi Xiong, Min Zhang, Aiti Aw, Haizhou Li
Human Language Technology
Institute for Infocomm Research
dyxiong@i2r.a-star.edu.sg

SMT: from phrase to syntax

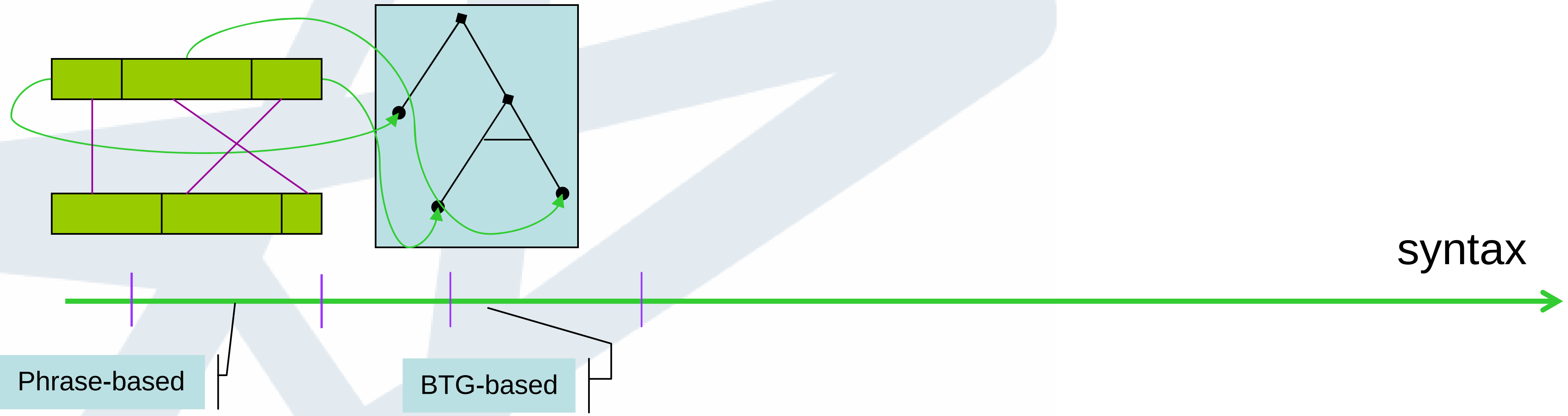
SMT: from phrase to syntax



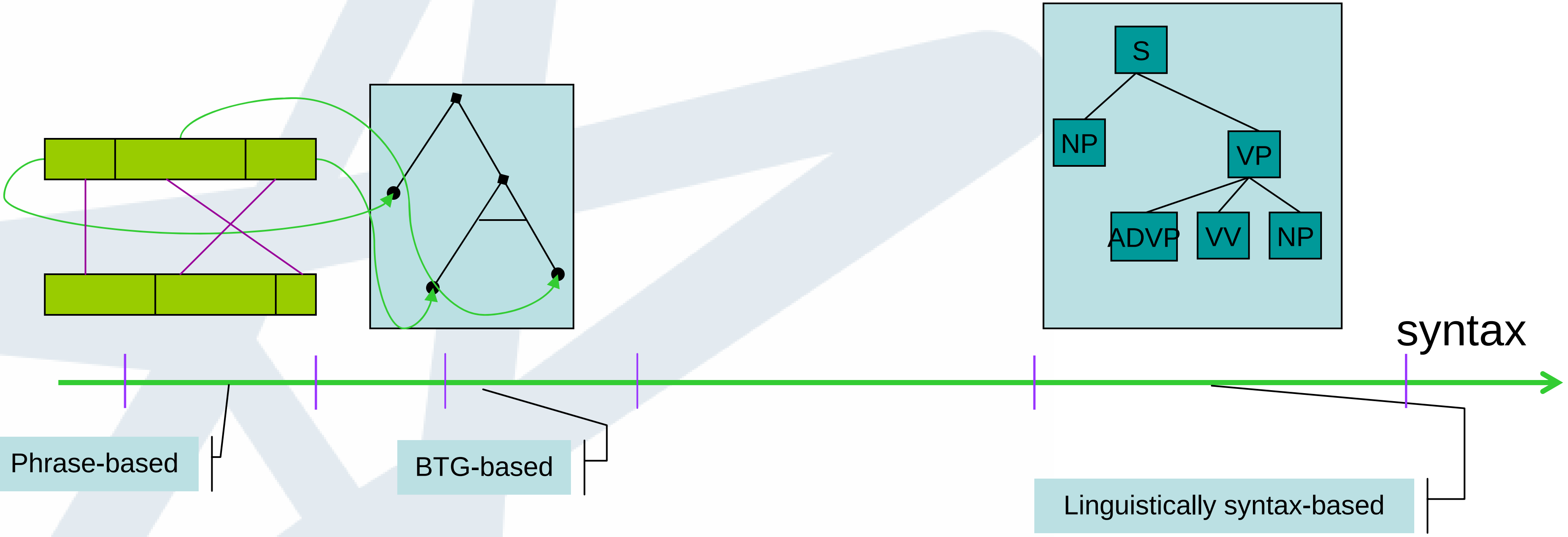
SMT: from phrase to syntax



SMT: from phrase to syntax



SMT: from phrase to syntax



Efforts to combine phrase and syntax: two directions

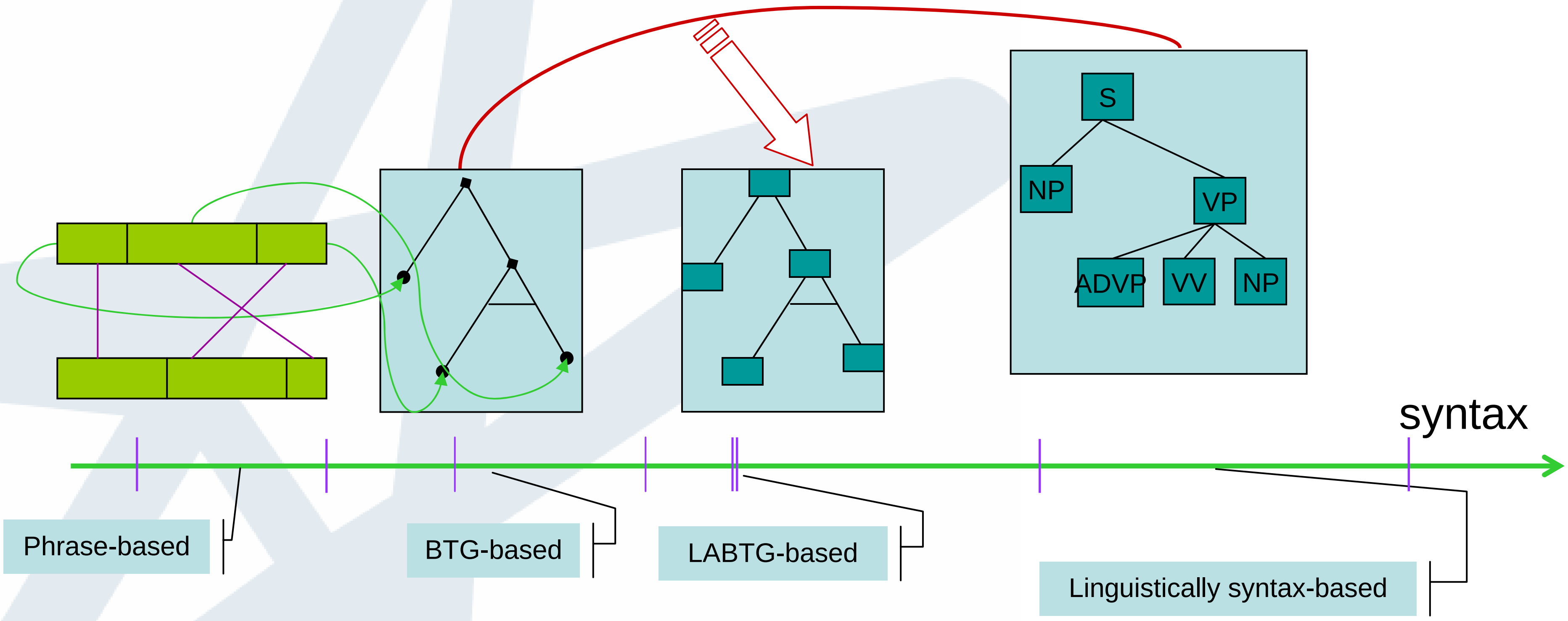
- Right to left: starting with linguistic representations, then inserting non-linguistic phrases (Marcu et al. 2006; Liu et al. 2007)

- Left to right: starting with a pure phrase-based system, then integrating linguistic syntax (Hassan et al. 2007)

The third direction

- Middle to right: starting with a formally syntax-based representations, then adding linguistic syntax
- Natural choice for combining phrases and syntax
  - Being syntactic although in the formal sense
  - Keeping the strength of phrase-based representation

The 3rd direction



Outline

- BTG-based SMT
- Linguistically Annotated BTG for SMT
  - LABTG rules
  - LABTG model
- Experiments
- Conclusions

BTG-based SMT: rules

- Lexical rules (r^l)
  - A → x/y
- Merging rules (r^m)
  - Straight order: A→[A,A]
  - Inverted order: A →<A, A>

BTG-based SMT: model

- Given a source sentence, we obtain the final translation *t* by using a series of lexical and merging rules
  - Derivation (D): a sequence of applications of lexical and merging rules
- Model P(D): translation model, reordering model, language model

$$P(D) = P_T(r^l_{1..n_l})^{\lambda_T} \cdot P_R(r^m_{1..n_m})^{\lambda_R} \cdot P_L(e)^{\lambda_L} \cdot exp(|e|)^{\lambda_w}$$

- Reordering model: use boundary words as features (Xiong et al. 2006)

BTG-based SMT: decoding

- CKY algorithm with beam search
  - Bottom-up
  - Dynamic programming
    - Store leftmost/rightmost boundary words to update the language model

Outline

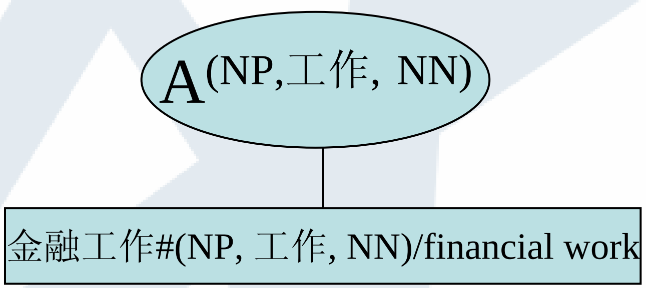
- BTG-based SMT
- Linguistically Annotated BTG for SMT
  - LABTG rules
  - LABTG model
- Experiments
- Conclusions

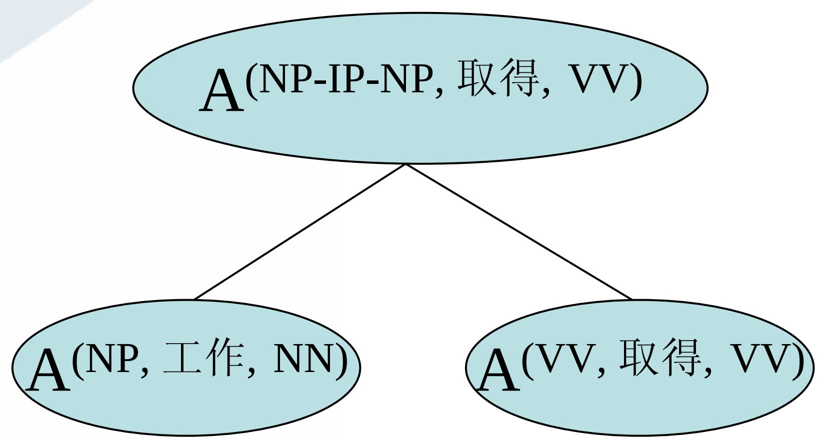
LABTG rules

- Annotated lexical rules
  - $A^a \rightarrow x\#a/y$
- Annotated merging rules
  - Straight order: $A^a \rightarrow [A^{a^l}, A^{a^r}]$
  - Inverted order: $A^a \rightarrow \langle A^{a^l}, A^{a^r} \rangle$

LABTG rules: linguistic annotation and examples

- Linguistic annotation
  - Syntactic label
  - Head word
  - POS tag of head word
- examples





Extract LABTG rules

- Obtain many-to-many word alignments
- Extract annotated lexical rules
  - Extract bilingual phrases
  - Annotate bilingual phrases according to source parse trees
- Extract annotated merging rules
  - Extract reordering examples: (A, A^l, A^r, o) (Xiong et al. 2006)
  - Annotate A, A^l, A^r according to source parse trees

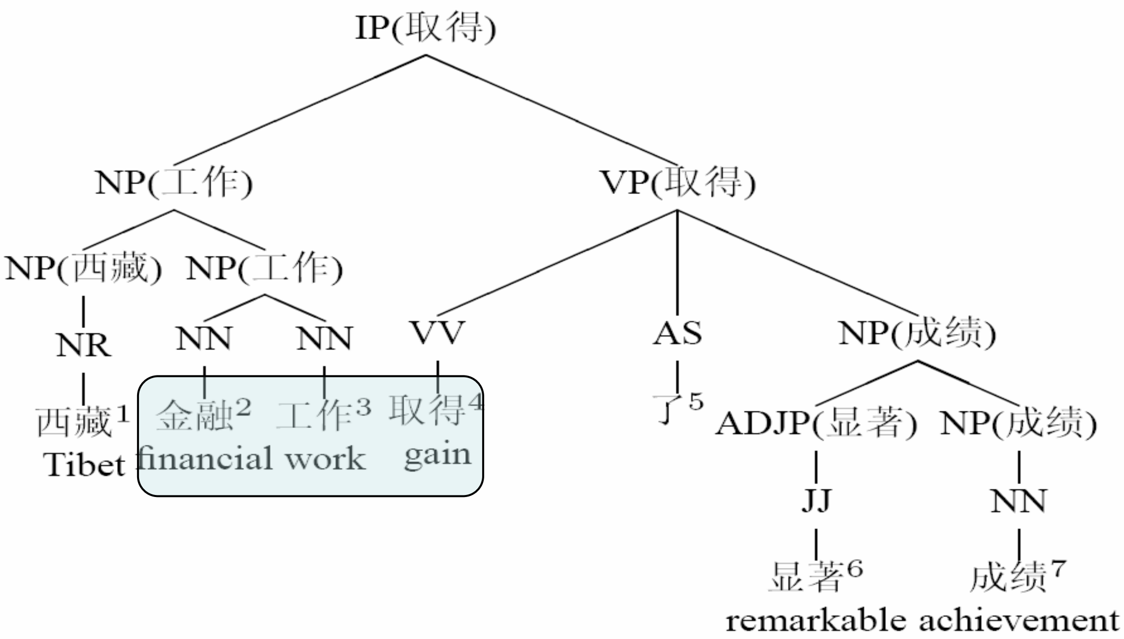
How to annotate phrases

- Syntactic phrases
  - Linguistic annotations directly come from corresponding sub-trees
- Non-syntactic phrases
  - Locate the smallest sub-tree covering the phrase
  - Determine the head word of the sub-tree
  - Construct the composite label
    - L-C-R
    - L/R: label of the left/right context node under the sub-tree
    - C: label of the sub-tree root node

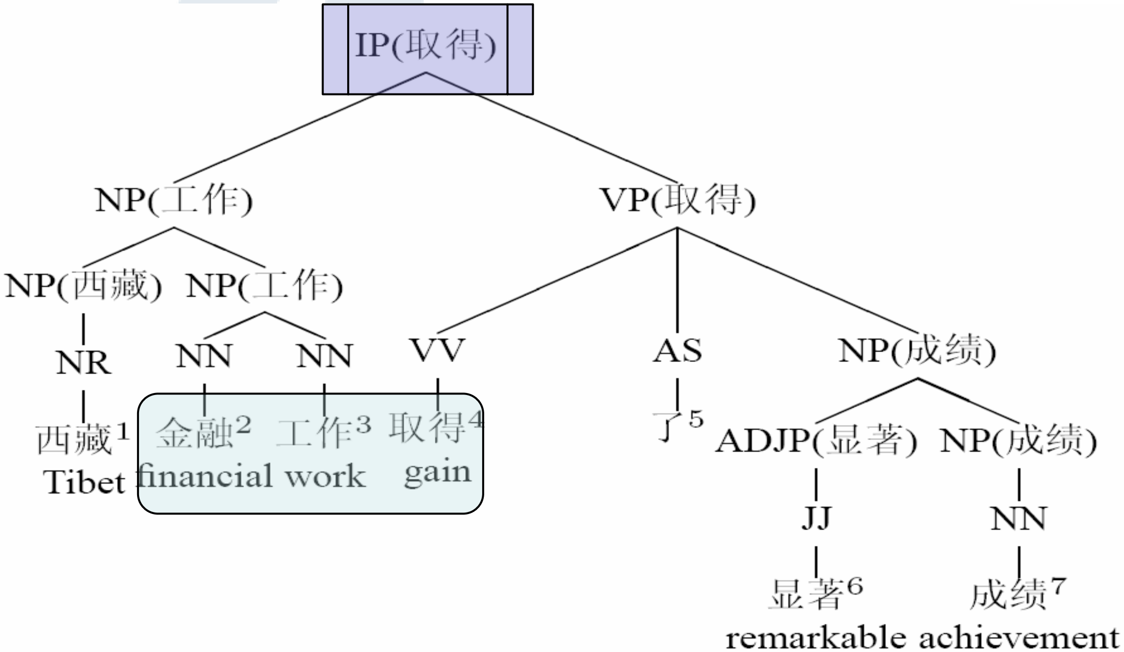
Context node

- First definition: Boundary node
  - The highest leftmost/rightmost sub-node of the sub-tree: not overlapping the phrase
- Second definition: neighboring node

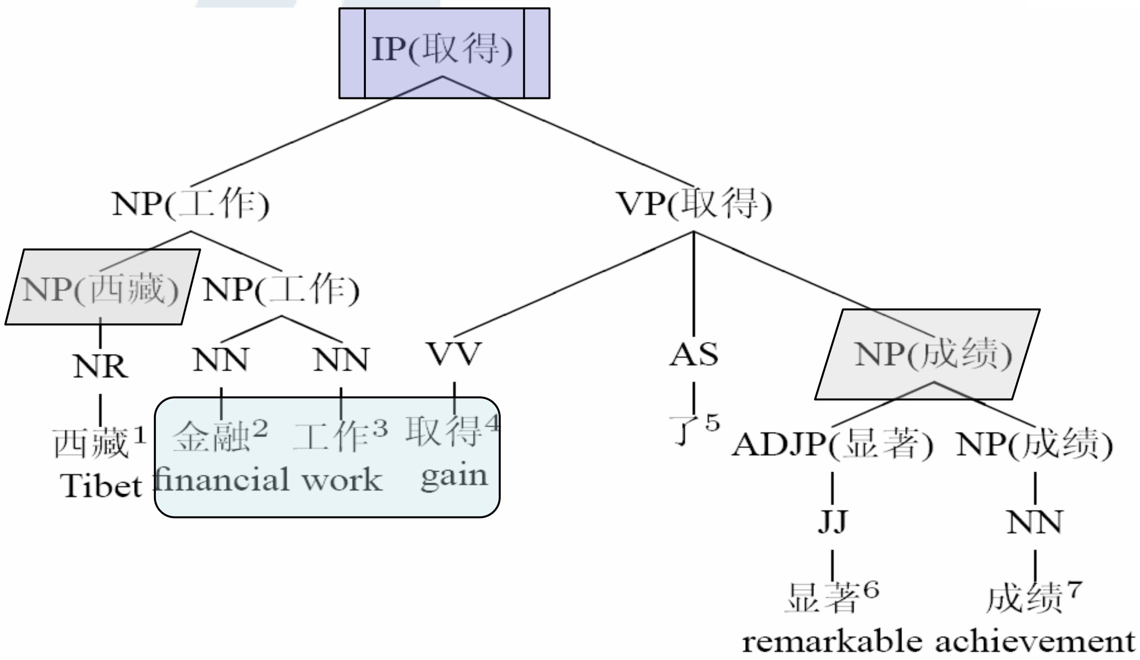
Annotation: An example



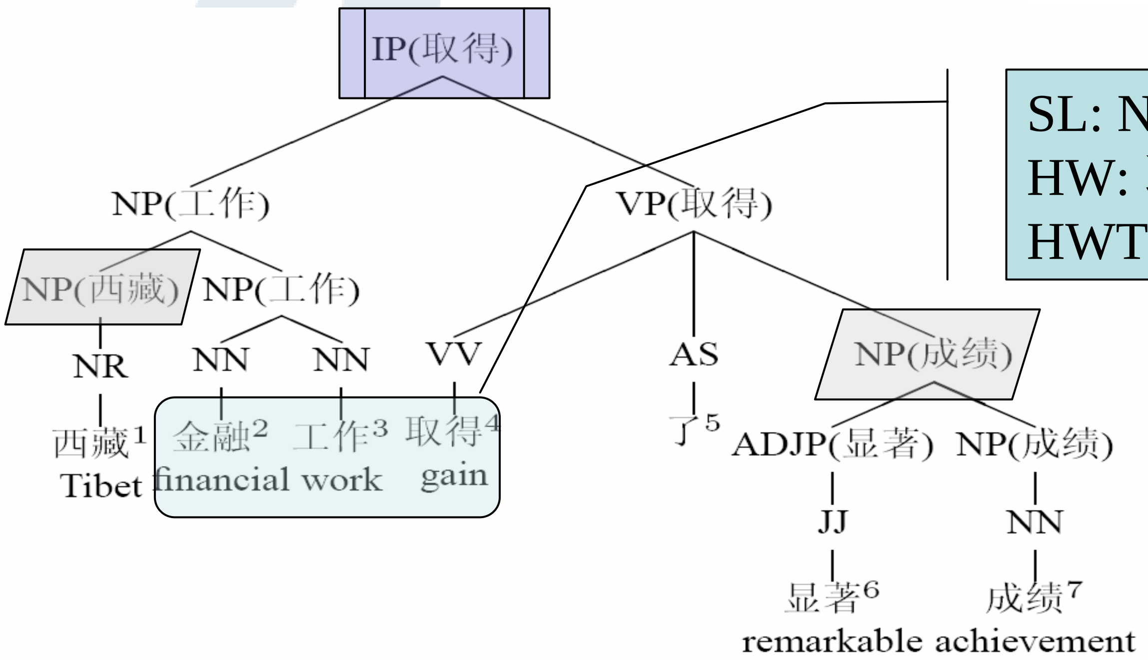
Annotation: An example



Annotation: An example



Annotation: An example



SL: NP-IP-NP
HW: 取得
HWT: VV

Outline

- BTG-based SMT
- Linguistically Annotated BTG for SMT
  - LABTG rules
  - LABTG model
- Experiments
- Conclusions

LABTG model

- The new log-linear model

$$P(D) = \boxed{P_{T_a}(ar^l_{1..n_l})}^{\lambda_{T_a}} \cdot P_{R_b}(r^m_{1..n_m})^{\lambda_{R_b}} \cdot \boxed{P_{R_a}(ar^m_{1..n_m})}^{\lambda_{R_a}} \cdot P_L(e)^{\lambda_L} \cdot exp(|e|)^{\lambda_w}$$

- The annotated translation model

- The annotated reordering model

Annotated translation model

- Model p(x#a|y) and p(y|x#a)
- Calculate these probabilities by relative counts
  - p(x#a|y) = count(x#a, y) / count(y)
  - p(y|x#a) = count(x#a, y) / count(x#a)
- No need to smooth
  - Not so many different annotations *a* for a same source phrase *x:* 1.14 annotations per source phrase
  - but for unseen pair (x#a, y), backoff to its un-annotated version (x, y)

Annotated reordering model

- Maximum entropy based model

$$
\begin{aligned}
P_{R_a}(ar^m) &= p_\theta(o|A^a, A_l^{a_l}, A_r^{a_r}) \\
&= \frac{exp(\sum_i \theta_i h_i(o, A^a, A_l^{a_l}, A_r^{a_r}))}{\sum_o exp(\sum_i \theta_i h_i(o, A^a, A_l^{a_l}, A_r^{a_r}))}
\end{aligned}
$$

- Binary features with linguistic annotations

$$
h_i(o, A^a, A_l^{a_l}, A_r^{a_r}) = \begin{cases} 1, & A_l^{a_l}.hw = w, o = straight \\ 0, & otherwise \end{cases}
$$

Outline

- BTG-based SMT
- Linguistically Annotated BTG for SMT
  - LABTG rules
  - LABTG model
- Experiments
- Conclusions

Experiments setup

- Chinese-to-English translation task
  - NIST MT-05 as test set
  - NIST MT-02 as dev set
- Bilingual training corpus: FBIS (The Foreign Broadcast Information Service)
  - Only used 218.9k sentence pairs due to the Chinese parser failure (14.9K sentences failed)
- Source side parsed by a Chinese parser: (Xiong et al. 2005)
- 4-gram language model: 181.1M words (Xinhua section of the English Gigaword corpus)
- Three systems: Moses, BTG, LABTG

Statistics on training data

Item	Number
Bilingual phrase	4.55M
Annotated lexical rules	4.65M
Reordering examples	2.8M
Boundary word features	115K
Linguistic annotation features	85K

LABTG vs. Moses and BTG

System	BLEU
Moses	0.2386
MEBTG	0.2498
LABTG	0.2667

2.8

1.7

Annotation schemes

- Try different annotation schemes
  - C: only use the root node label of the covering sub-tree for non-syntactic phrases
  - N-C-N: neighboring node as context node for non-syntactic phrases
  - B-C-B: boundary node as context node for non-syntactic phrases
  - Annotating syntactic phrases with composite labels

Result

Annotation scheme	BLEU
C	0.2626
N-C-N	0.2591
B-C-B	0.2667
Annotating syntactic nodes with composite label	0.2464

Conclusion:

Neither too specific nor too general annotations are the best choice for LABTG

Effect of annotated translation model

Translation model	BLEU
P_T	0.2498
P_{T_a}	0.2581
P_{T_a} (-NULL)	0.2548

- Source-side linguistic annotation provide two kinds of information for phrase selection:
  - Syntactic category: phrase with different categories has different translations
  - context

Effect of annotated reordering model

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

Effect of annotated reordering model

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

132 features

Effect of annotated reordering model

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

132 features

6.7K features

Effect of annotated reordering model

1.55 BLEU points

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

132 features

6.7K features

85K features

Effect of annotated reordering model

1.55 BLEU points

Reordering Configuration	BLEU (%)
BWR	24.97 ± 0.90
BWR + LAR (SL)	25.88 ± 0.95
BWR + LAR (+BNL)	26.27 ± 0.98
BWR + LAR (+BNL+HWT)	26.52 ± 0.96
Only allowed SPs reordering	25.12 ± 0.87

132 features

6.7K features

85K features

Outline

- BTG-based SMT
- Linguistically Annotated BTG for SMT
  - LABTG rules
  - LABTG model
- Experiments
- Conclusions

Conclusions

- Source-side syntactic knowledge can be helpful for phrase selection
- Linguistic annotations according to source-side parse trees significantly improve phrase reordering
- Incorporation linguistic knowledge into SMT model should be carefully conducted.

The end