

Learning Translation boundaries for Phrase-Based Decoding

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

Segmentation/bracketing problem in machine translation

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Segmentation/bracketing problem in machine translation

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Possible segmentation

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Possible segmentation

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Possible segmentation

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]



Wrong segmentations

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Wrong segmentations

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]

[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]

Wrong segmentations

~~[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]~~

~~[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]~~

~~[佛塞特] [完成] [单人] [热气球] [绕地球一周] [壮举]~~

~~[Fossett] [Completes] [his Epic Voyage] [of Solo] [Circumnavigation of Earth] [in a Hot Air Balloon]~~

Three fundamental problems for phrase-based MT

- Segmentation
 - Splits the input sentence into cohesive segments
- Translation
 - Translates source segments into target segments
- Reordering
 - Reorder target segments

Segmentation: less studied

- Assume a uniform distribution over segmentations
 - Of course not uniformly distributed
- Syntactic constraints: Segmentations which violate constituent boundary on the source side will be penalized.
 - Why do we think a constituent can be considered as a segment unit which is translated coherently given syntactic divergences?

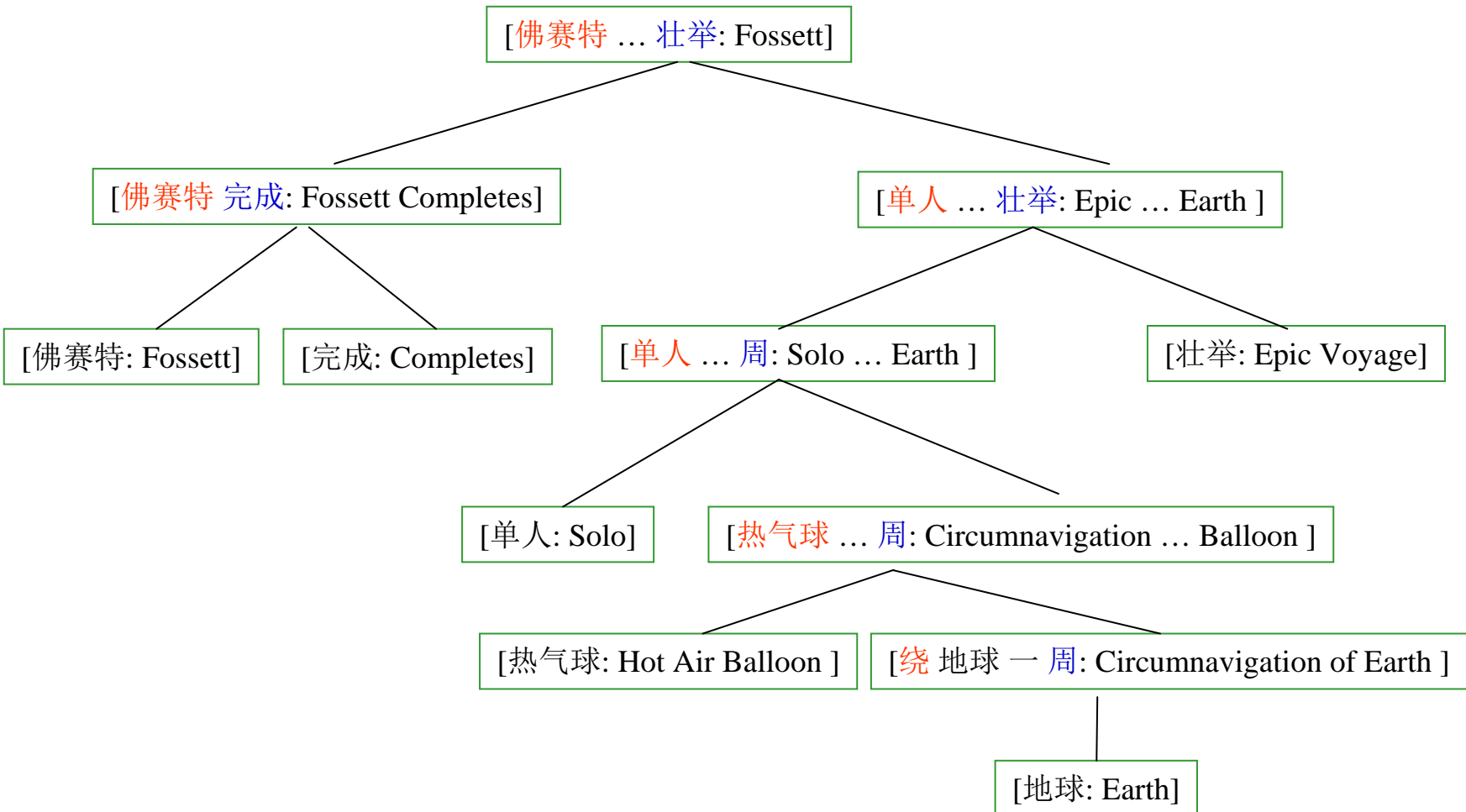
Outline

- Define translation boundary for segmentation
- Learn translation boundaries
- Integrate translation boundaries into phrase-based MT
- Experiments

Translation boundary: Definition

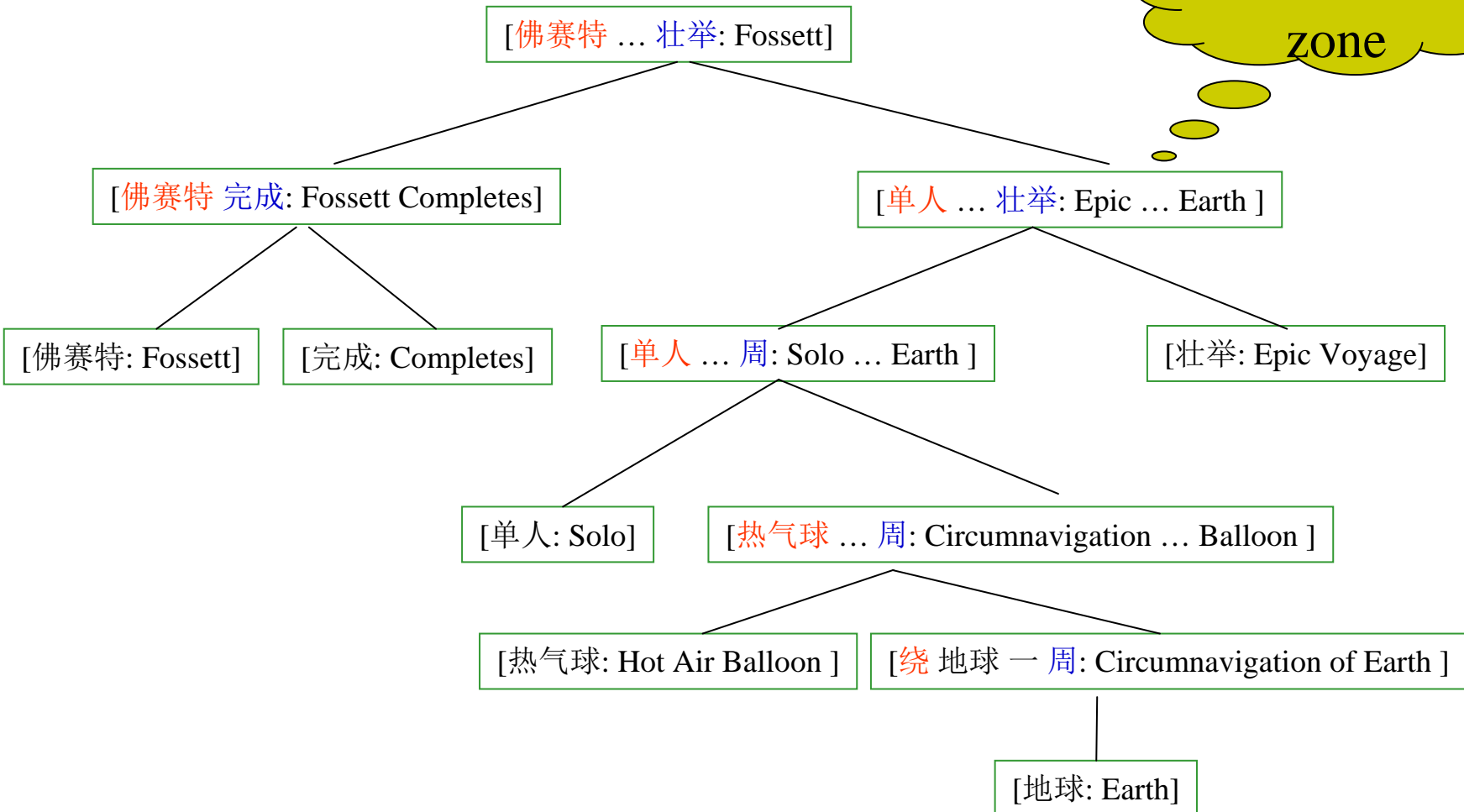
- Translation zone:
 - A pair of aligned source and target phrase
 - No words outside the source phrase are aligned to words inside the target phrase
 - No words inside the source phrase are aligned to words outside the target phrase
- Beginning boundary: B_y
 - The leftmost source word in a translation zone
- Ending boundary: E_y
 - The rightmost source word in a translation zone

Translation boundary: Examples

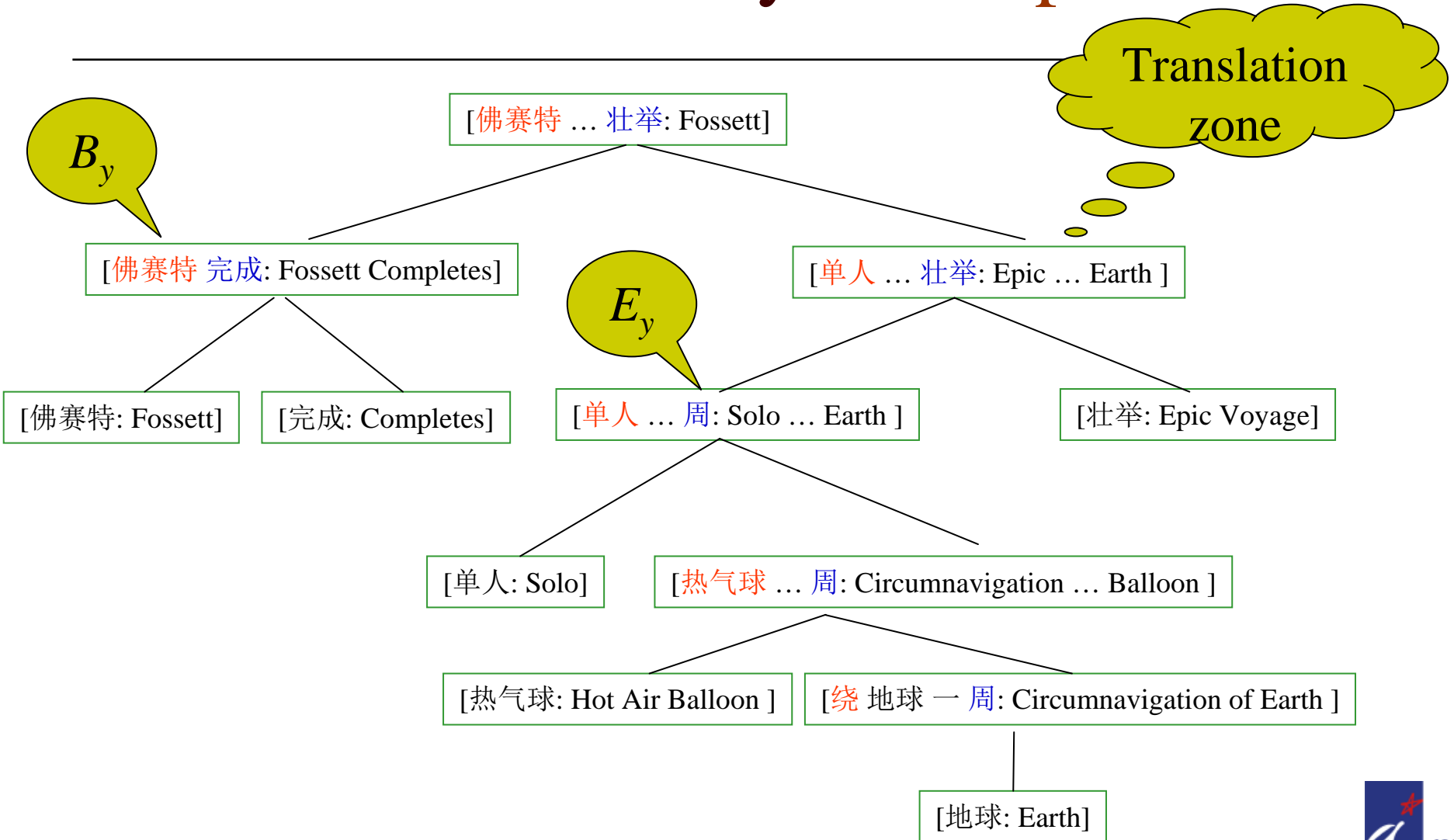


Translation boundary: Examples

Translation
zone



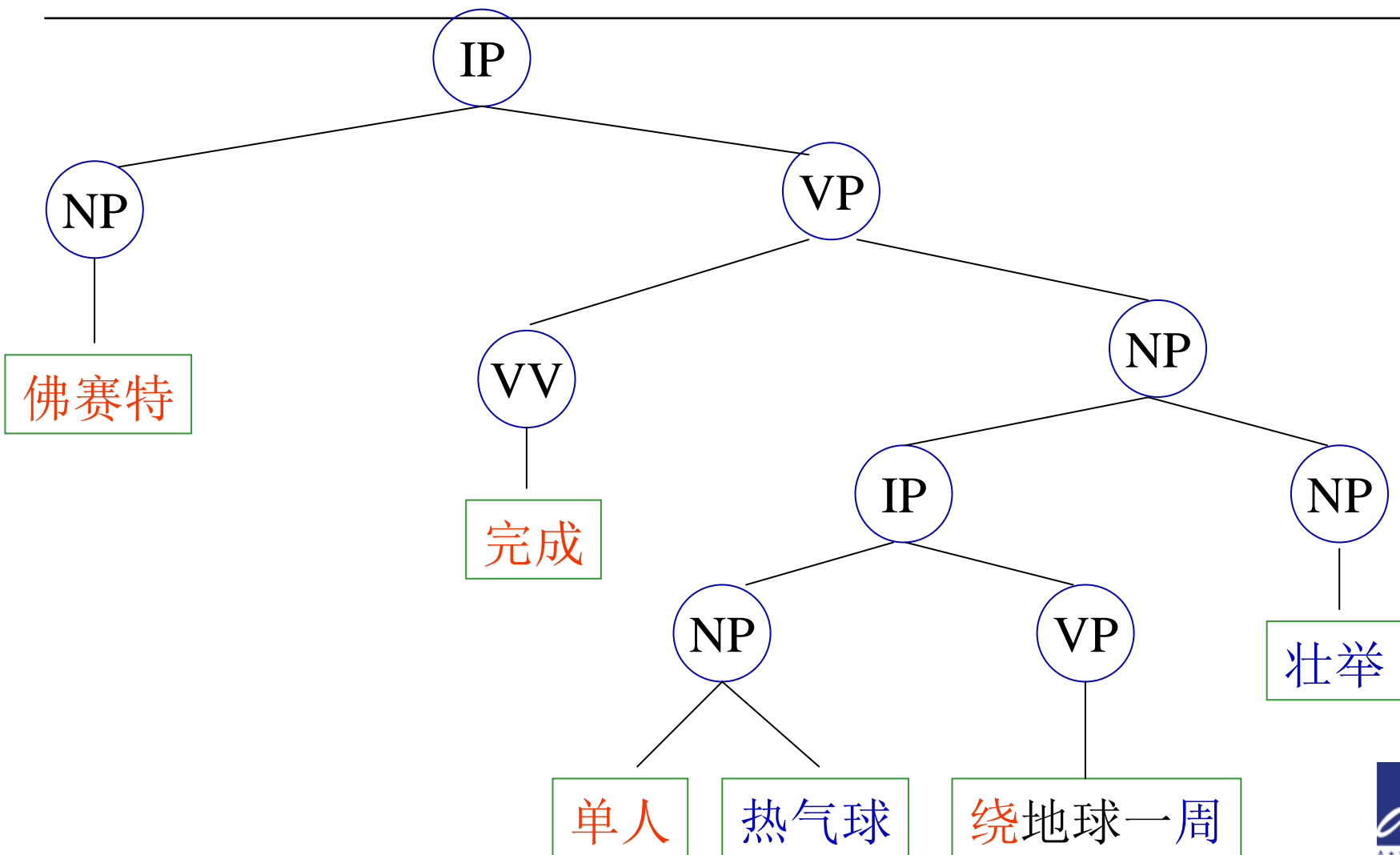
Translation boundary: Examples



Translation boundary: Distribution

Item	Count (M)	P (%)
Sentences	3.8	–
Words	96.9	–
Words $\in B_y$	22.7	23.4
Words $\in E_y$	41.0	42.3
Words $\notin B_y$ and $\notin E_y$	33.2	34.3

Constituent boundary



Constituent boundary vs. translation boundary

- Gold standard: translation boundaries derived from alignments between source sentences and reference translations
- Classifier: constituent boundary deducer using source side parse trees

Classification Task	Avg. Accuracy (%)
B_y / B_n	46.9
E_y / E_n	52.2

Constituent boundary vs. translation boundary

- Gold standard: translation boundaries derived from alignments between source and reference translations
- Classifier: constituent boundaries from source side parse tree

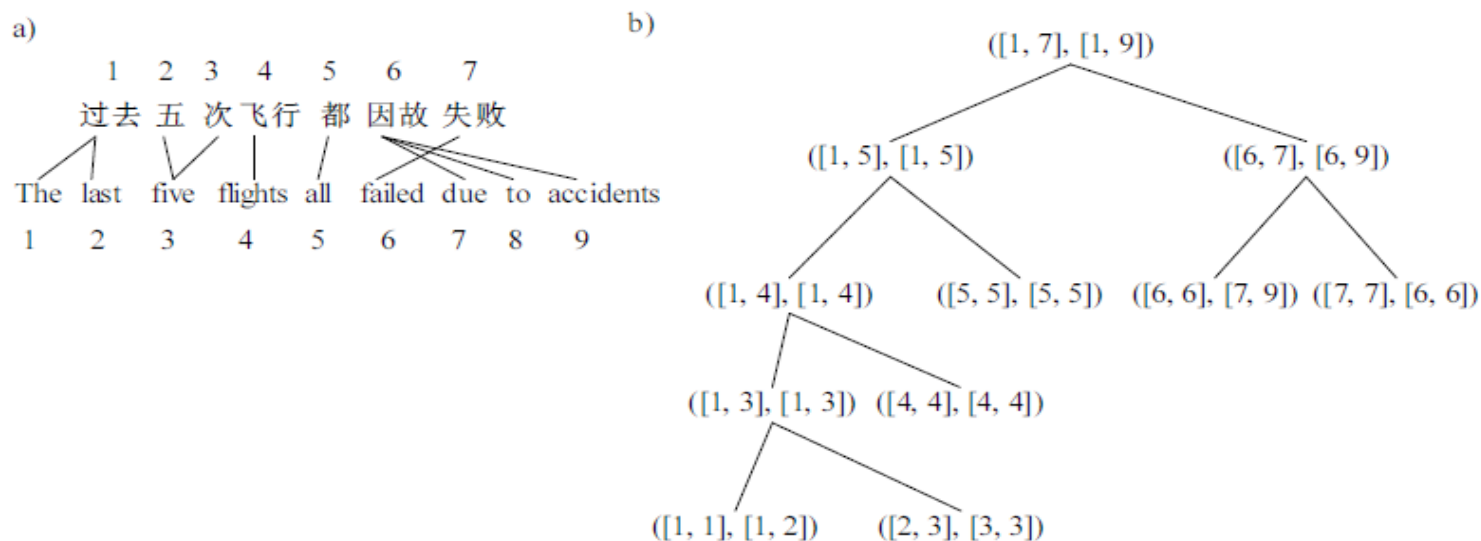
Only 50% constituent boundaries are really translation boundaries

Classification Task	Avg. Accuracy (%)
B_y/B_n	46.9
E_y/E_n	52.2

Outline

- Define translation boundary for segmentation
- Learn translation boundaries
- Integrate translation boundaries into phrase-based MT
- Experiments

Automatically extract translation boundaries from word alignments



- Word alignments → hierarchical segmentation structure by (Zhang et al., 2008)'s shift-reduce algorithm
- Each node is a translation zone
- Visit each multi-source-word node to extract boundary words

Build translation boundary classifiers

- Two classifiers: one for predicting beginning boundaries, the other for predicting ending boundaries
- Based on Maximum Entropy Markov Models (MEMM)
- Features:
 - Word features: two words before and after the current word position respectively
 - Class features: the previous class (Markov order 1, MEMM1), the previous two classes (Markov order 2, MeMM2)

Translation boundary classifiers:

Evaluation

- 4 gold standards: derived from 4 different reference translations
- 2 classifiers

Classification Task	Avg. Accuracy (%)	
	MEMM 1	MEMM 2
B_y/B_n	71.7	70.2
E_y/E_n	59.2	58.8

Translation boundary classifiers: Evaluation

- 4 gold standards: derived from 4 different reference translations
- 2 classifiers

Significantly higher than accuracies of constituent boundary deducer

Classification Task	Avg. Accuracy (%)	
	MEMM 1	MEMM 2
B_y/B_n	71.7	70.2
E_y/E_n	59.2	58.8

Divergences among 4 reference translations

- Classifier: translation boundaries derived from one reference translation
- 3 gold standards: translation boundaries derived from the other 3 reference translations

Classification Task	Avg. Accuracy (%)
B_y/B_n	80.6
E_y/E_n	75.7

Divergences among 4 reference translations

- Classifier: translation boundary derived from one reference
the accuracy of translation boundary classifier is not that low when
- 3 go considering vast divergences of reference translations

Classification Task	Avg. Accuracy (%)
B_y/B_n	80.6
E_y/E_n	75.7

Outline

- Define translation boundary for segmentation
- Learn translation boundaries
- Integrate translation boundaries into phrase-based MT
- Experiments

A segmentation model

- Integrate translation boundaries into phrase-based decoding as soft constraints
- Segmentation model: translation boundary violation counting feature
 - Accumulates whenever a partial translation covers $c_i \dots c_j$, where c_i is not in B_y or c_j is not in E_y

Outline

- Define translation boundary for segmentation
- Learn translation boundaries
- Integrate translation boundaries into phrase-based MT
- Experiments

Experimental setup

- Training data: about 4M sentence pairs from LDC corpus
- Language model: trained on 181M-word Giga Xinhua corpus
- Development set: NIST03
- Test set: NIST05
- Base system: BTG-based phrasal SMT system with state-of-the art reordering model (Xiong et al., 2006)

Using “right” translation boundaries derived from reference translations

- 4 set of translation boundaries derived from 4 different reference translations

System	BLEU-4 (%)
Base	33.05
Ref1	33.99*
Ref2	34.17*
Ref3	33.93*
Ref4	34.21*

Using translation boundaries from our trained classifiers

Constituent boundary based constraints

	BLEU-4 (%)
Baseline	33.05
Condeducer	33.18
XP+	33.58*
BestRef	34.21*+
MEMM 1	33.70*
MEMM 2	34.04*+

Using translation boundaries from our trained classifiers

Constituent boundary based constraints

	BLEU-4 (%)
Baseline	33.05
Conceducer	33.18
XP+	33.58*
BestRef	34.21*+
MEMM 1	33.70*
MEMM 2	34.04*+

translation boundary based constraints

Conclusions

- Automatically learn translation boundaries from word alignments without any additional resources
- Build a segmentation model based on translation boundaries predicted by classifiers
- Translation boundary is better than constituent boundary