Parser Self-Training for Syntax-Based Machine Translation

Nara Institute of Science and Technology
Augmented Human Communication Laboratory

Makoto Morishita, Koichi Akabe, Yuto Hatakoshi, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura

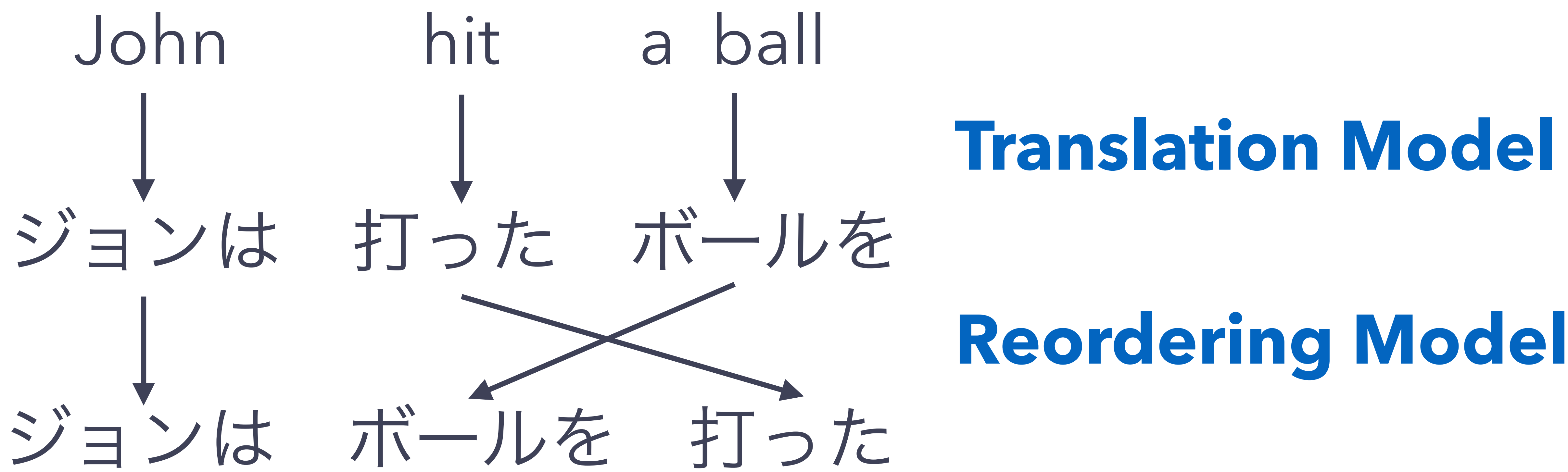
2015/12/03
IWSLT 2015

NAIST®

Background

Phrase-Based Machine Translation

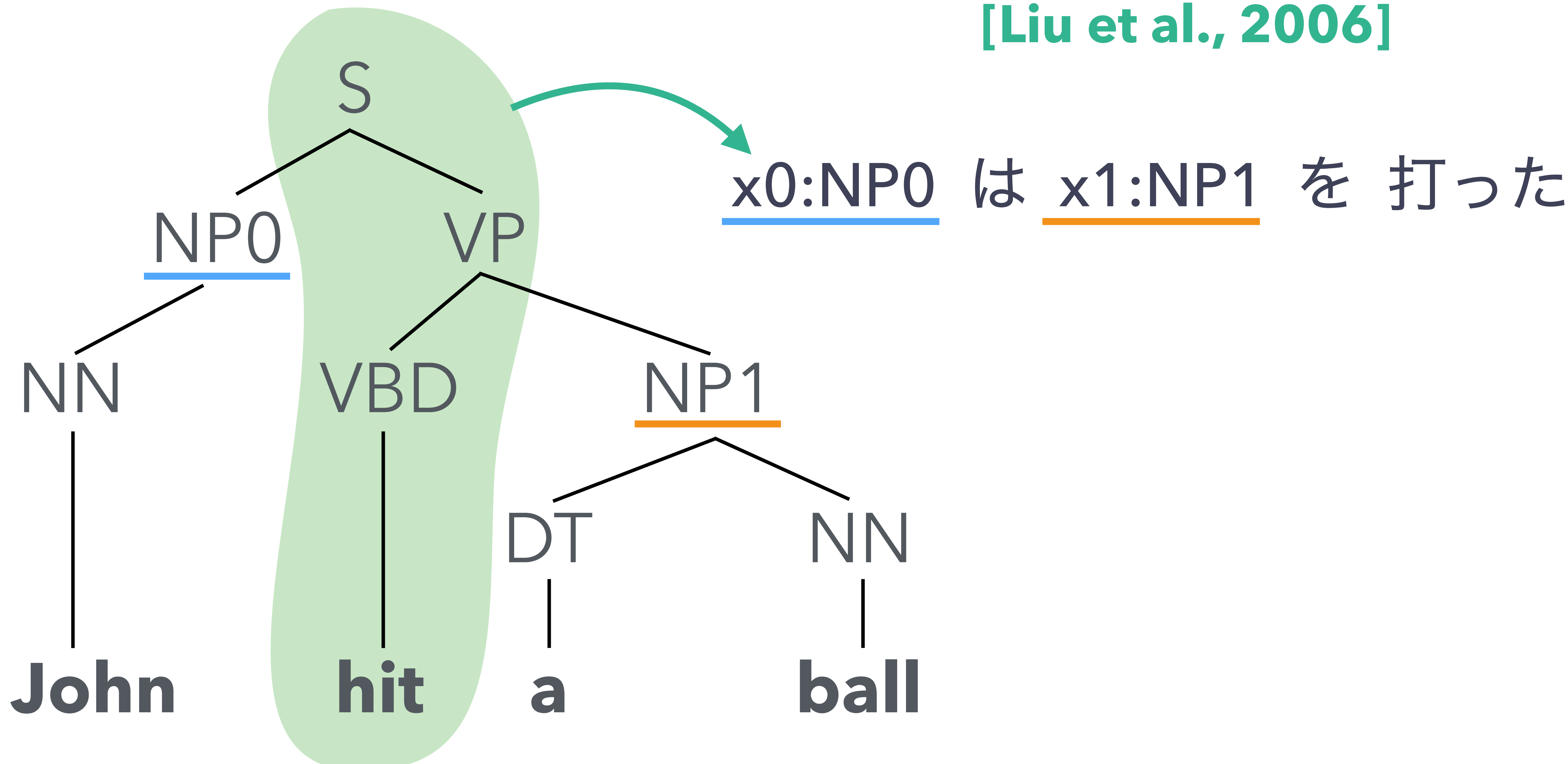
[Koehn et al., 2003]



- Translate and reorder by phrases.
  - Easy to learn translation model.
  - Low translation accuracy on language pairs with different word order.

Tree-to-String Machine Translation

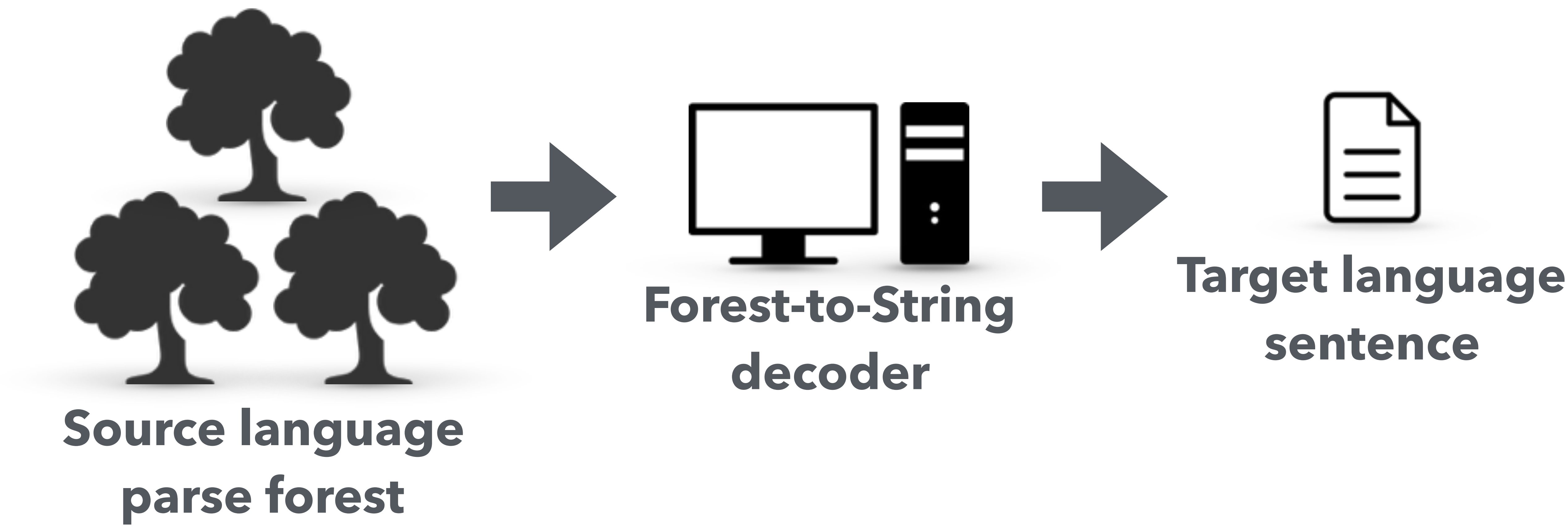
[Liu et al., 2006]



- Use the source language parse tree in translation
  - High translation accuracy on language pairs with different word order.
  - Translation accuracy is affected greatly by the parser accuracy.

Forest-to-String Machine Translation

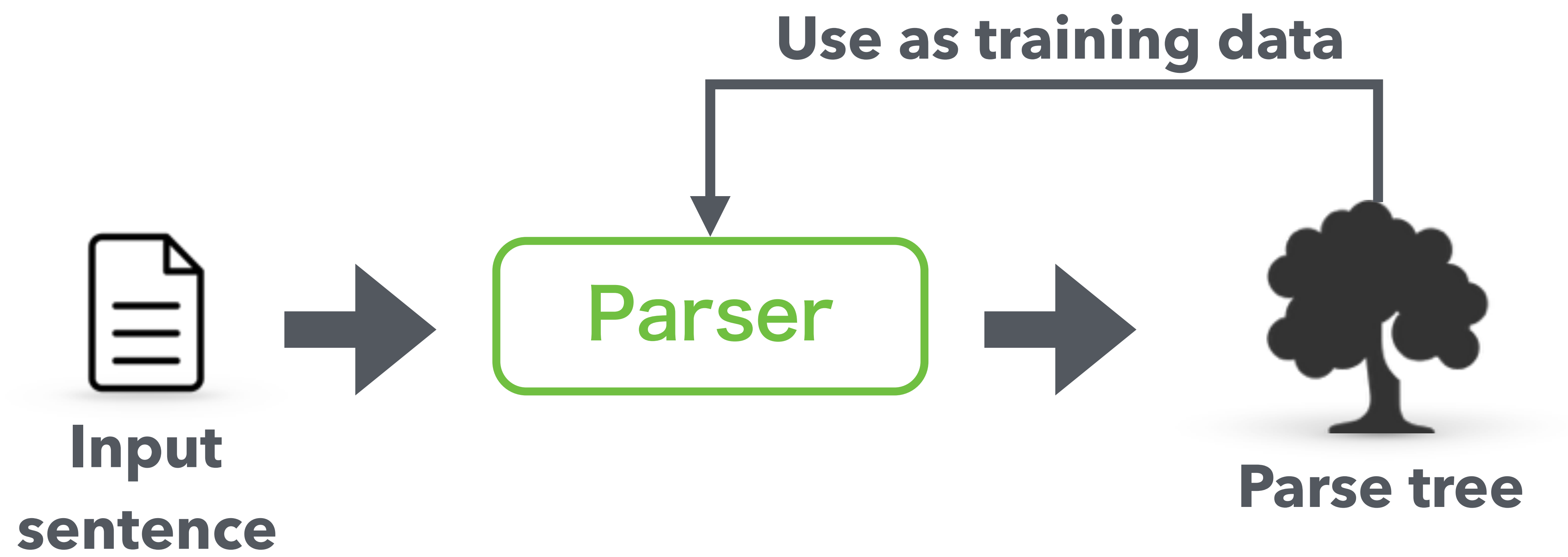
[Mi et al., 2008]



- Use the source language parse forest in translation
  - Decoder can choose the parse tree that has high translation probability from the parse tree candidates

[Zhang et al., 2012]

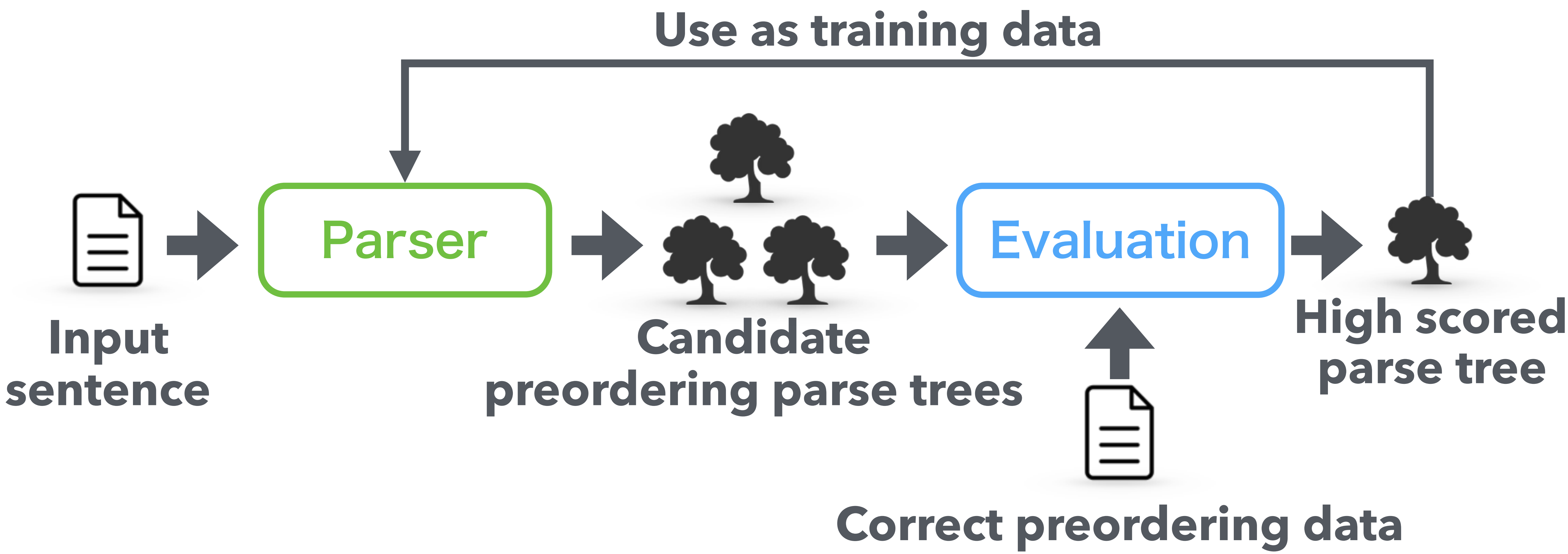
Parser Self-Training [McClosky et al., 2006]



- Use the parser output as training data.
- Improve the parser accuracy.
  - Parser is adapted to the target domain.

Self-Training for Preordering

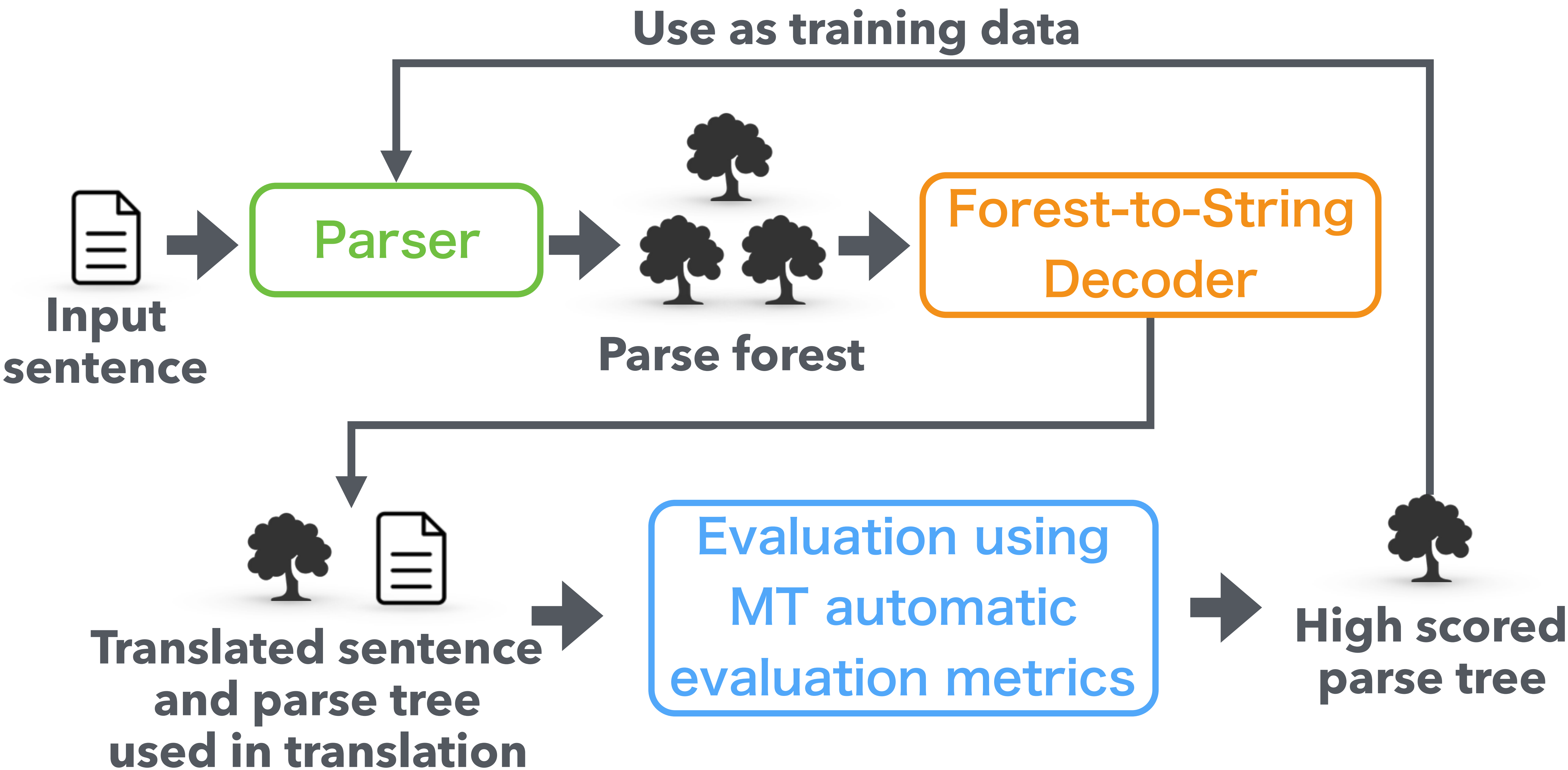
[Katz-Brown et al., 2011]



- By selecting the parse trees, more effective self-training (Targeted Self-Training).
  - Use only high scored parse trees.
  - However, in this method, we need hand-aligned data.
  - It is costly to make hand-aligned data.

Proposed Method

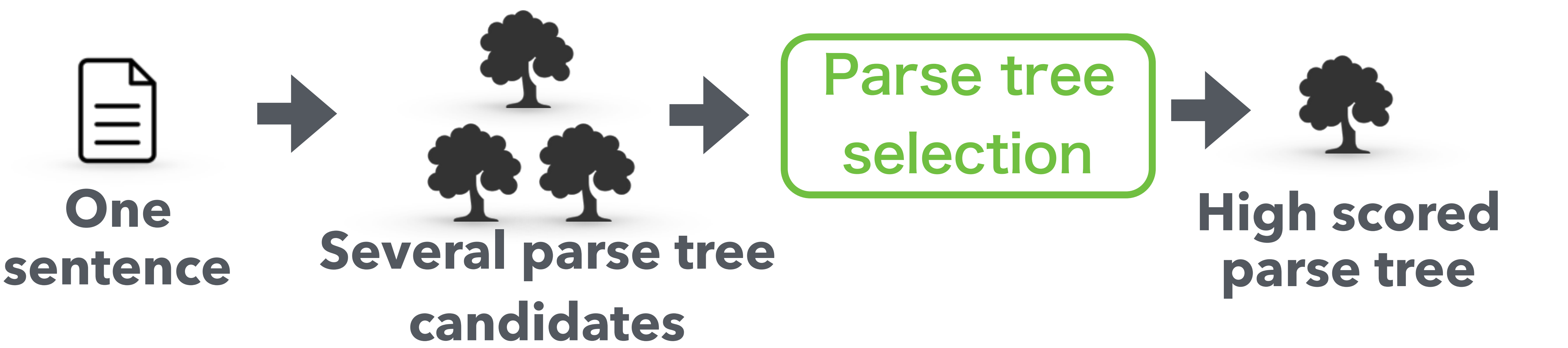
Proposed Method



- Targeted Self-Training using MT automatic evaluation metrics
  - low cost and accurate evaluation

Selection Methods

- Parse tree selection
  - Select a parse tree to use from a single sentence

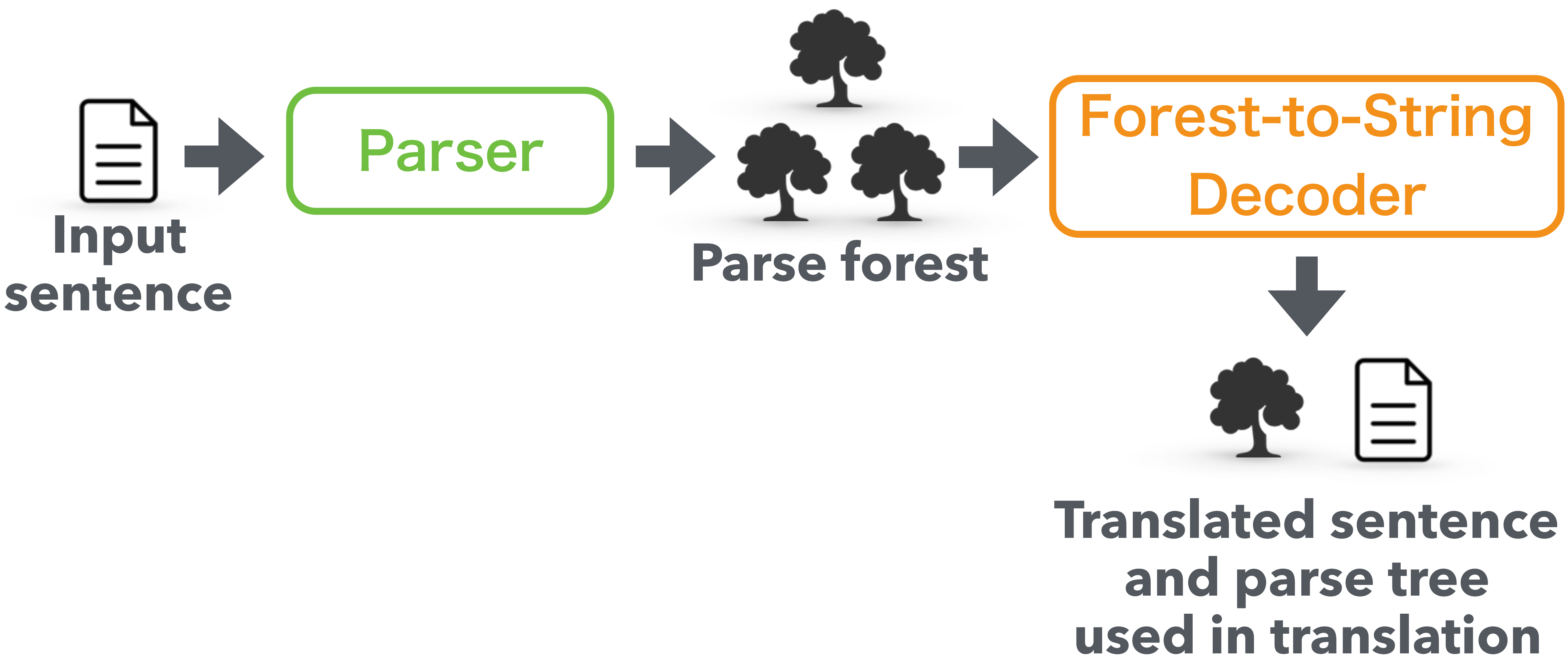


- Sentence selection
  - Select the sentences to use from the entire corpus

Parse Tree Selection

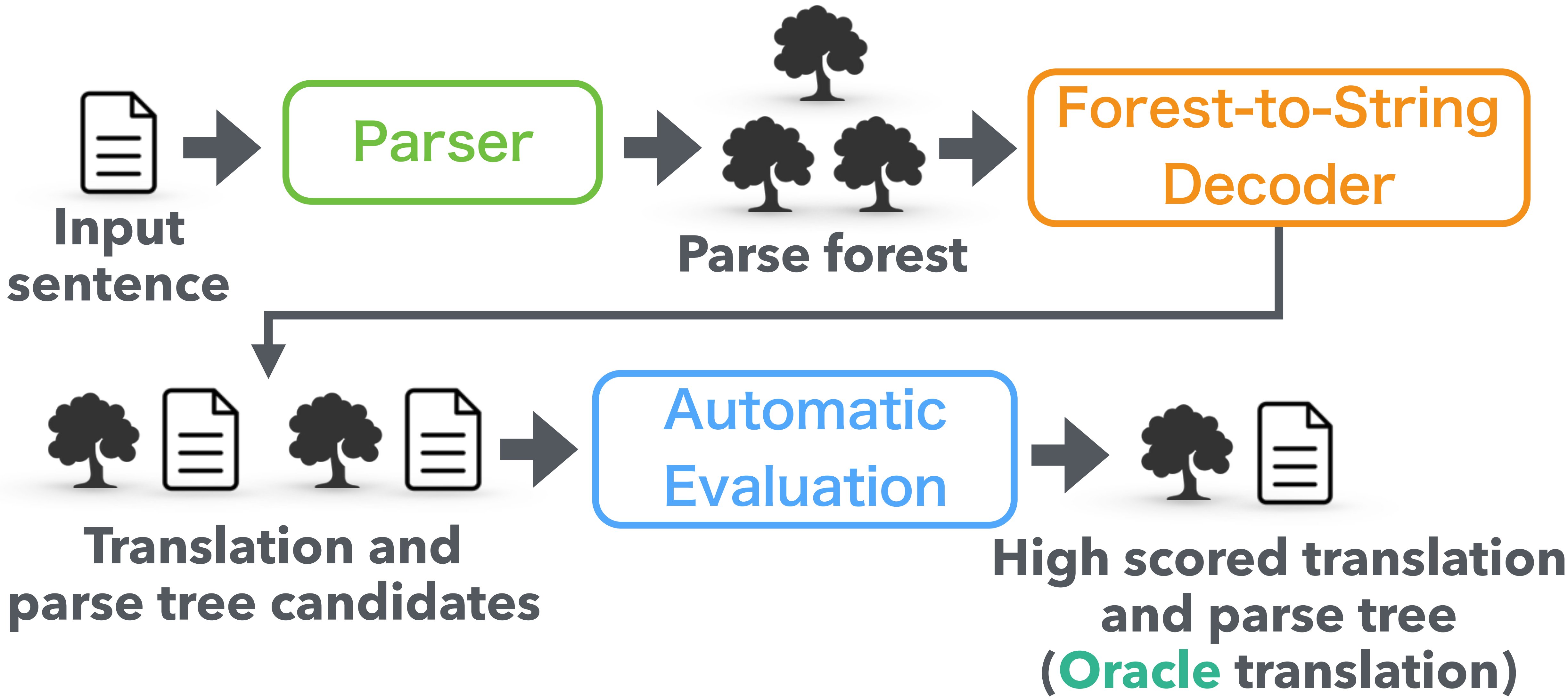
- Parser 1-best
  - Use the parser 1-best tree.
  - Traditional self-training [McClosky et al. 2006].
- Decoder 1-best
  - Use the parse tree used in translation.
- Evaluation 1-best
  - Among the translation candidates, use the parse tree used in highest scored translation.

Decoder 1-best



- Decoder 1-best
  - Use the parse tree used in translation.

Evaluation 1-best



- Evaluation 1-best
  - Among the translation candidate, use the parse tree used in highest scored translation.
  - This highest scored translation is called Oracle translation.

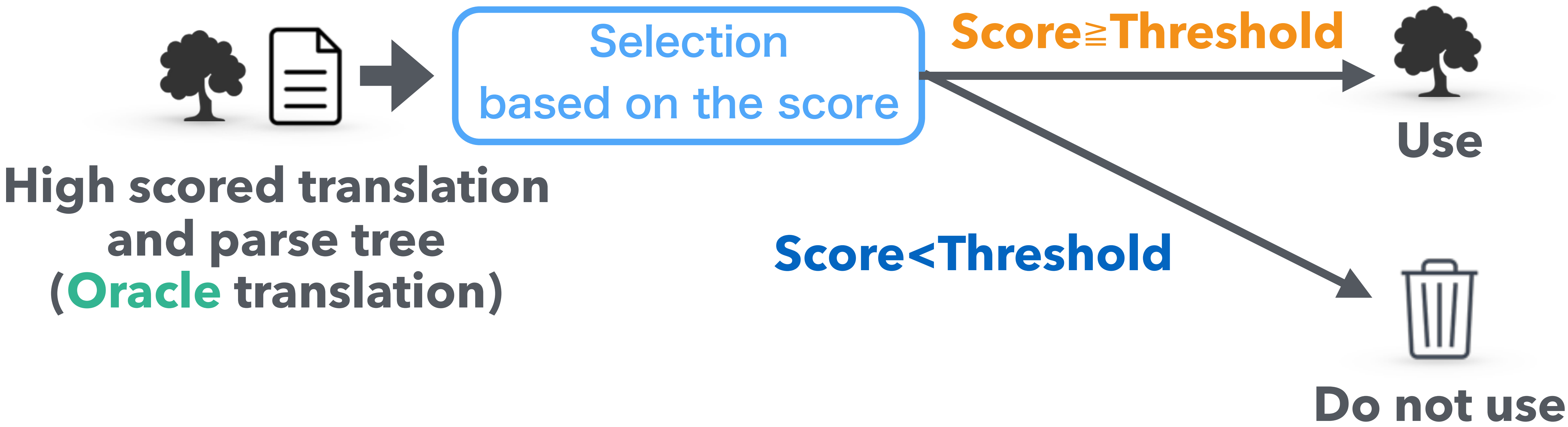
Sentence Selection

- Random
  - Select sentences randomly from the corpus.
  - Traditional self-training.

- Threshold of the evaluation score
  - Use sentences that score over the threshold.

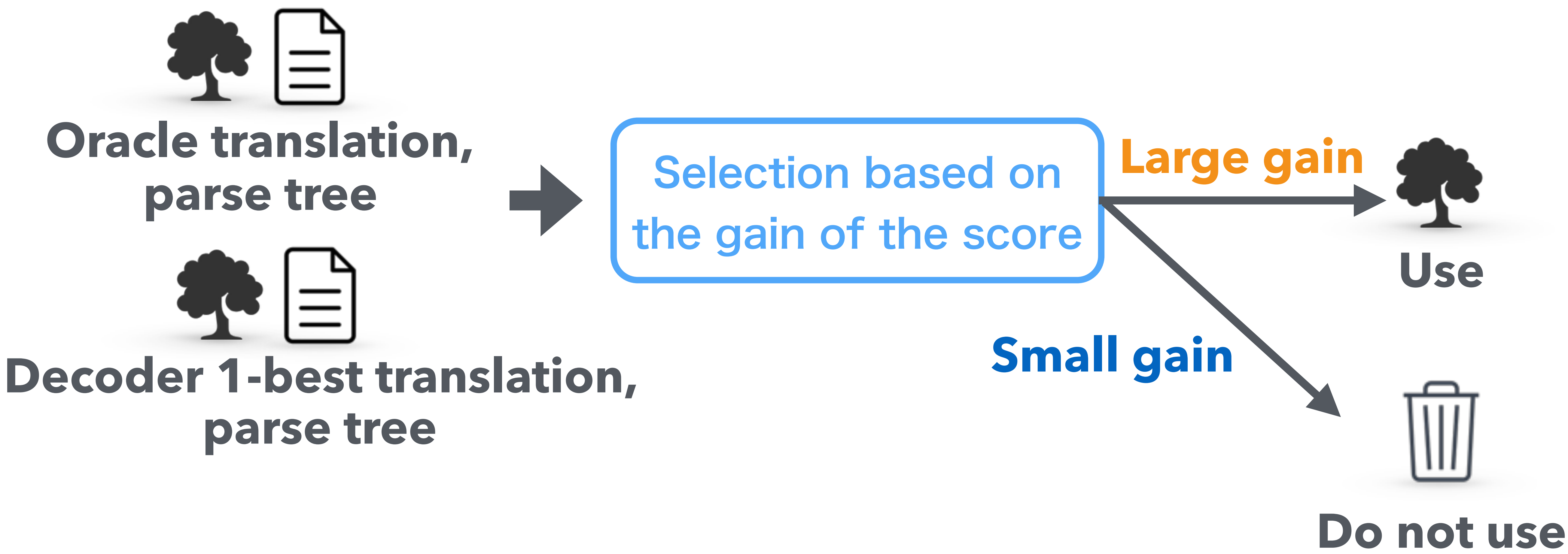
- Gain of the evaluation score
  - Use sentences that have a large gain in score between decoder 1-best and oracle translation.

Threshold of the Evaluation Score



- Threshold of the evaluation score
  - Use sentences that score over threshold.

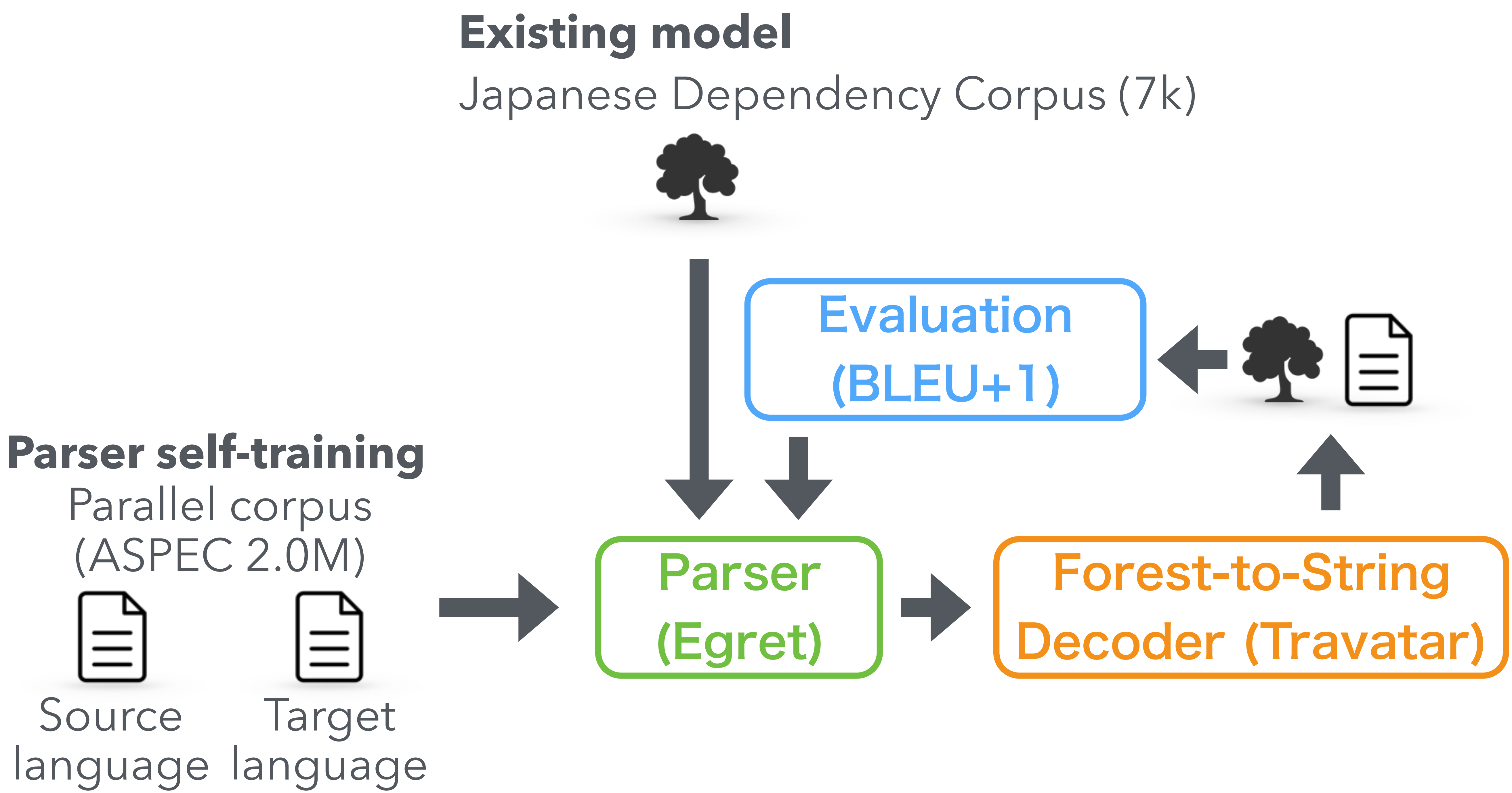
Gain of the Evaluation Score



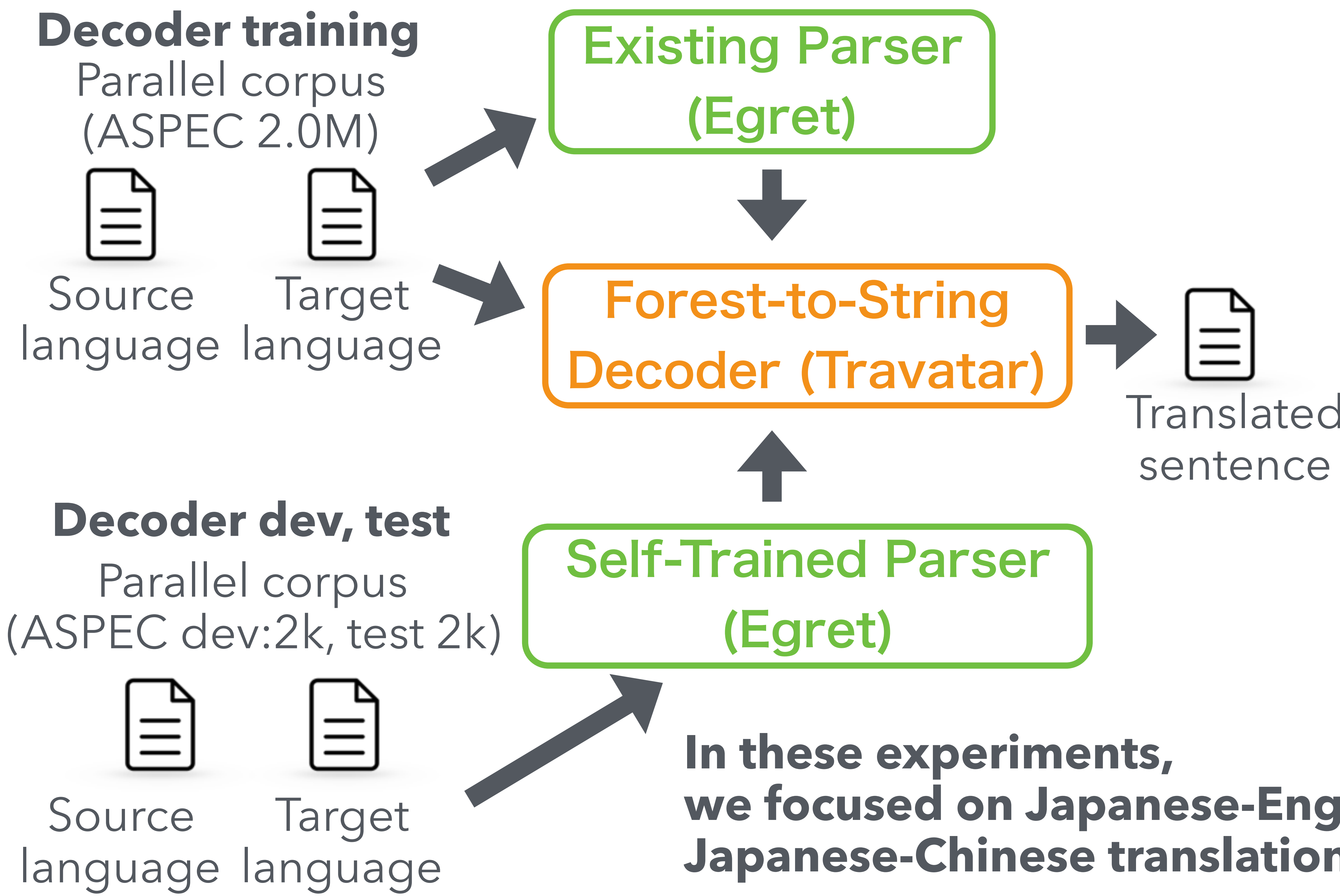
- Gain of evaluation score
  - Use sentences that have a large gain in score between decoder 1-best and oracle translation.

Experiments

Experimental Setup (for Self-Training)



Experimental Setup (for Evaluation)

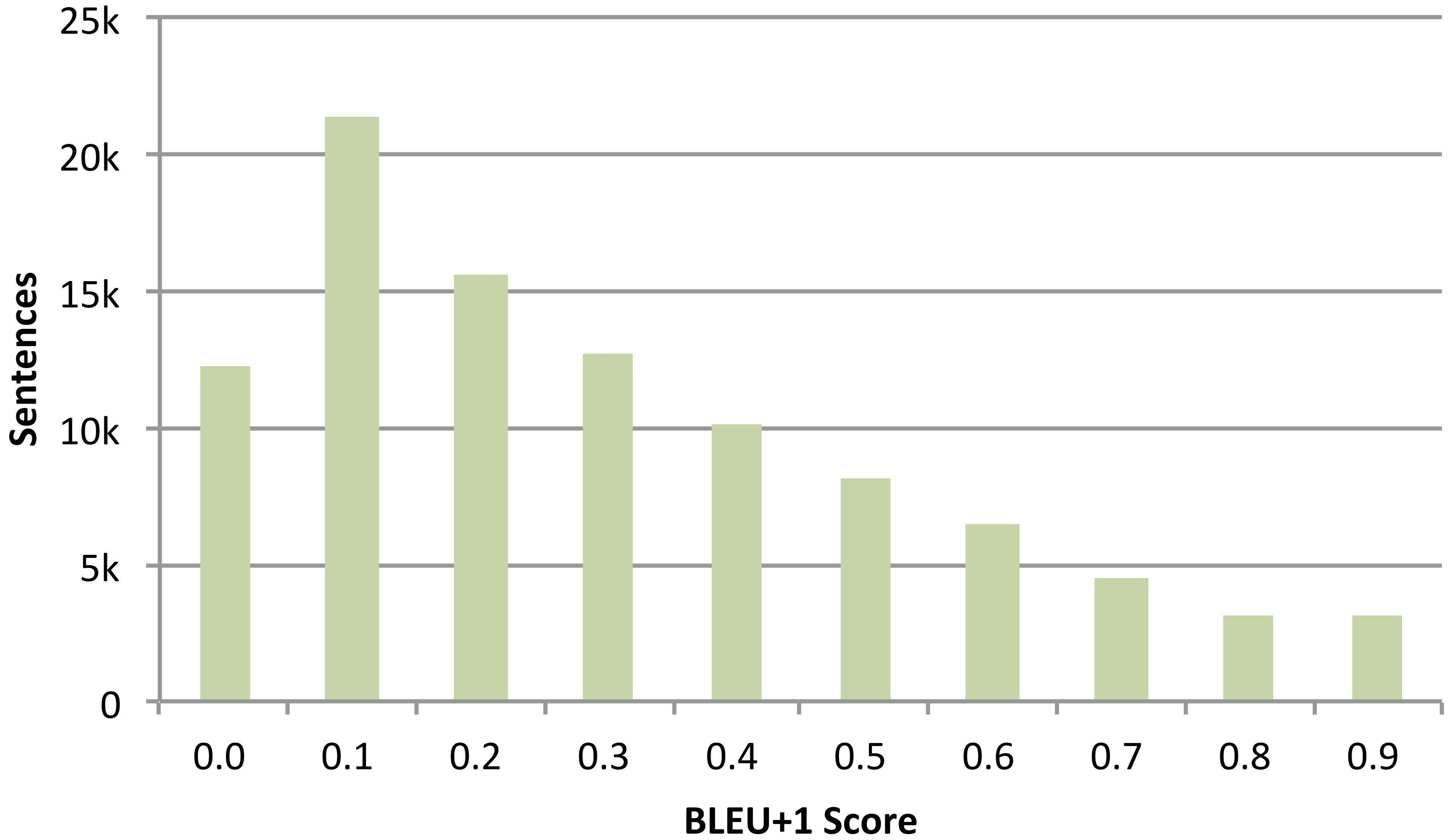


In these experiments, we focused on Japanese-English, Japanese-Chinese translation

Experiment Results (Japanese-English Translation)

Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	23.83	72.27
Parser 1-best	**Random**	96	23.66	71.77
Decoder 1-best	**Random**	97	23.81	72.04
Oracle	**Random**	97	23.93	72.09

Oracle Translation Score Distribution



- It contains a lot of noisy sentences.

Experiment Results (Japanese-English Translation)

Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	23.83	72.27
Parser 1-best	**Random**	96	23.66	71.77
Decoder 1-best	**Random**	97	23.81	72.04
Oracle	**Random**	97	23.93	72.09
Oracle	**BLEU+1 Threshold**	120	** 24.26	72.38
Oracle	**BLEU+1 Gain**	100	* 24.22	72.32

*** : $p < 0.05$ ** : $p < 0.01$**

- By self-training, the accuracy significantly improved

Manual Evaluation

Tree selection	Sentence selection	Score	Significance between Baseline	Significance between Parser -best
Baseline		2.38	—	—
Parser 1-best	**Random**	2.42	No	—
Oracle	**BLEU+1 Threshold**	2.50	Yes (99% level)	Yes (90% level)

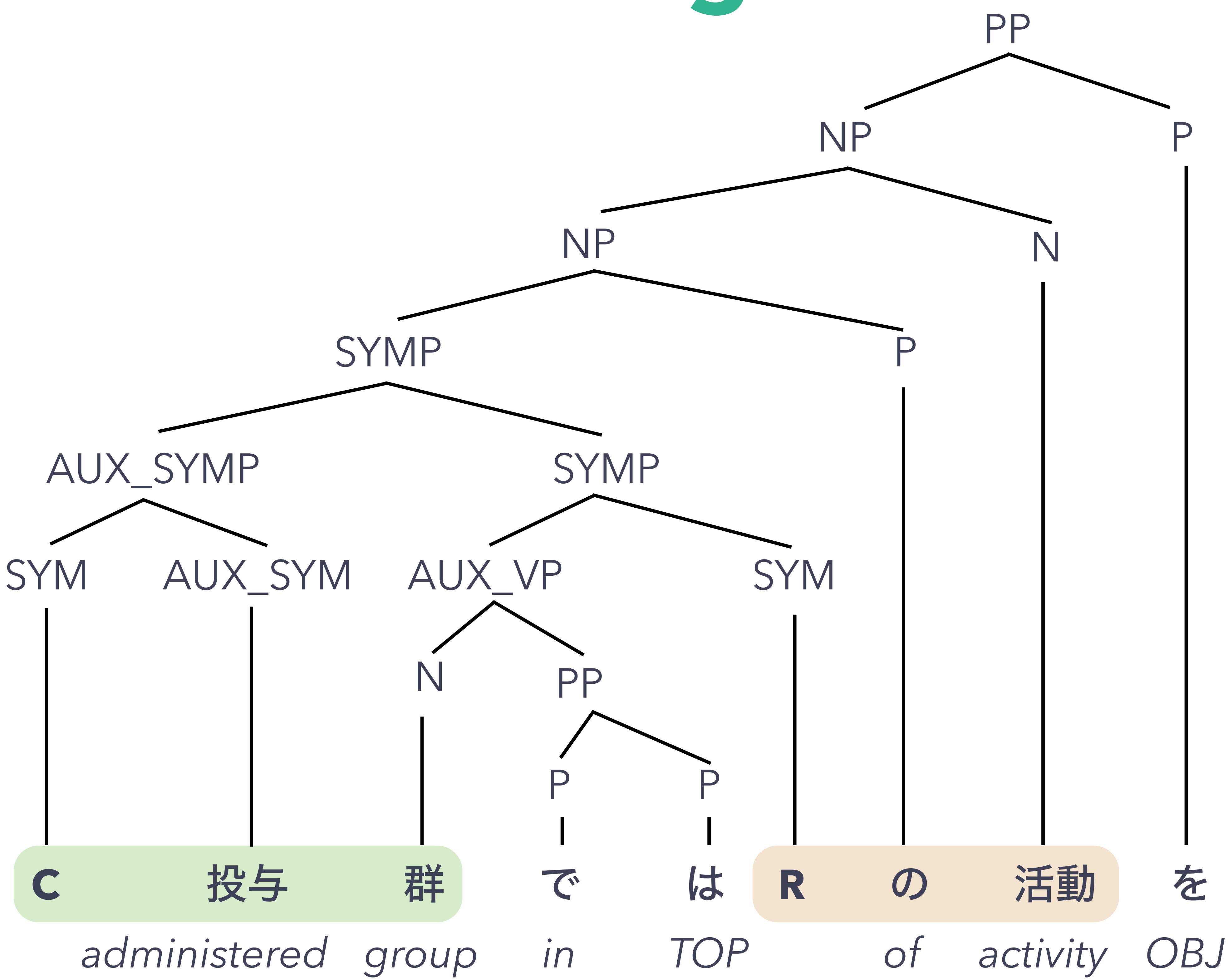
Score range is 1 to 5

- We could verify that our method is effective.

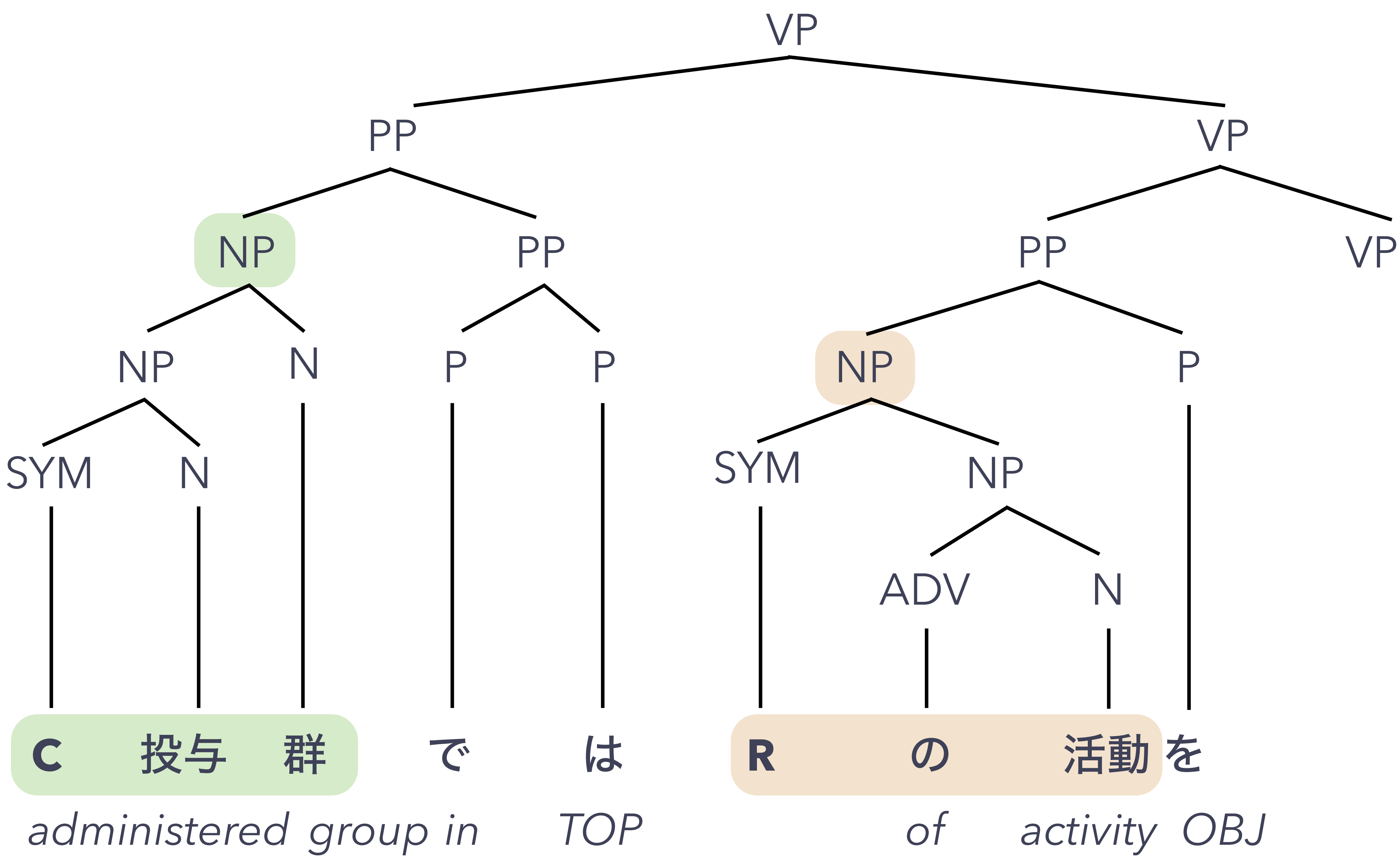
Example of an improvement

Source	C投与群ではRの活動を240分にわたって明らかに増強した
Reference	in the C - administered group, thermal reaction clearly increased the activity of R for 240 minutes.
Baseline	for 240 minutes clearly enhanced the activity of C administration group R.
Self-Trained	for 240 minutes clearly enhanced the activity of R in the C - administration group.

Before Self-Training



After Self-Training



Experiment Results (Japanese-Chinese Translation)

Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	29.60	81.32
Parser 1-best	**Random**	129	29.75	** 81.55
Decoder 1-best	**Random**	130	29.76	* 81.53
Oracle	**Random**	130	** 29.89	** 81.66
Oracle	**BLEU+1 Threshold**	82	* 29.86	** 81.60
Oracle	**BLEU+1 Gain**	100	* 29.85	** 81.59
Oracle (ja-en)	**BLEU+1 Threshold**	120	* 29.87	* 81.58

*** : $p < 0.05$ ** : $p < 0.01$**

- By self-training, the accuracy significantly improved
- By using ja-en self-trained model, it also improved the accuracy.

Experiment Results (Japanese-Chinese Translation)

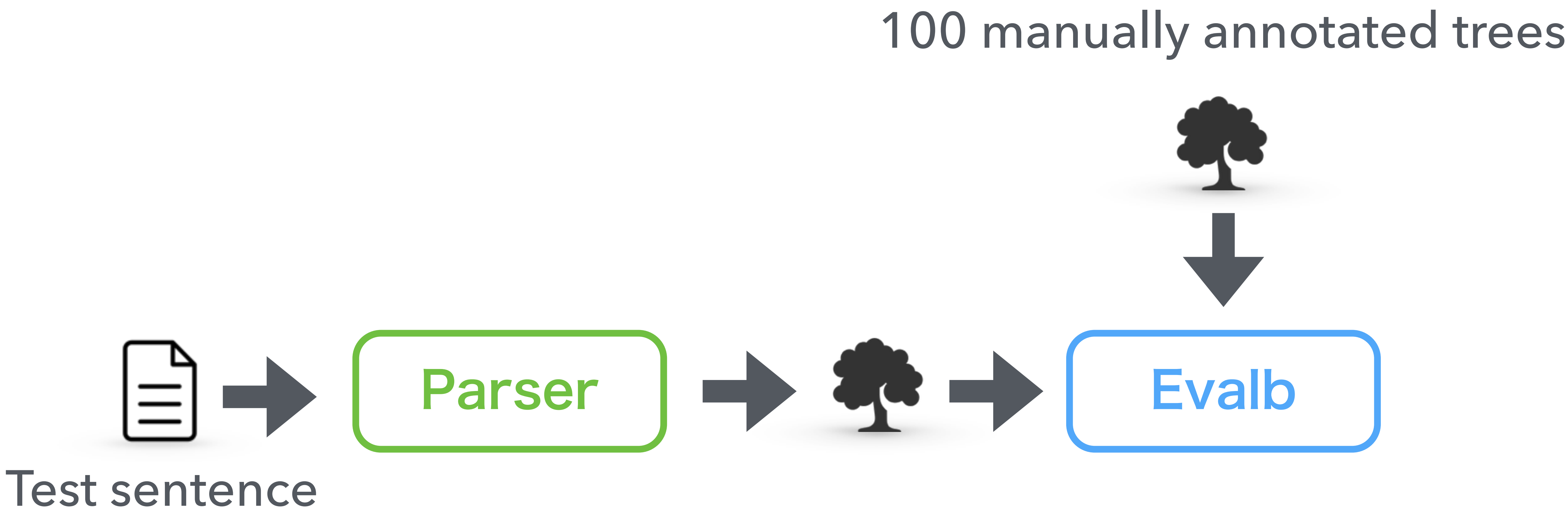
Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	29.60	81.32
Parser 1-best	**Random**	129	29.75	** 81.55
Decoder 1-best	**Random**	130	29.76	* 81.53
Oracle	**Random**	130	** 29.89	** 81.66
Oracle	**BLEU+1 Threshold**	82	* 29.86	** 81.60
Oracle	**BLEU+1 Gain**	100	* 29.85	** 81.59
Oracle (ja-en)	**BLEU+1 Threshold**	120	* 29.87	* 81.58

*** : $p < 0.05$ ** : $p < 0.01$**

- By self-training, the accuracy significantly improved
- By using ja-en self-trained model, it also improved the accuracy.

Parser Accuracy

Experimental Setup



- Evalb: tool of scoring parsing accuracy based on Collins, 1997.
- We test Ja-En parsers.

Experiment Results

Tree selection	Sentence selection	Recall	Precision	F-Measure
Baseline		84.88	84.77	84.83
Parser 1-best	**Random**	86.52	86.41	* 86.46
Oracle	**BLEU+1 Threshold**	88.13	88.01	** 88.07

*** : $p < 0.05$ ** : $p < 0.01$**

- Our method improves not only MT results, but also parser accuracy itself.

Conclusion

Conclusion

- By our proposed self-training method, translation and parser accuracy improved.
- Self-Training does not rely on target language
  - By using Ja-En self-trained model, Ja-Zh translation accuracy improved.
- Future work
  - Verify this method is applicable in other languages.
  - Self-training using several target languages data.
  - Test the effect when performing the parser self-training repeatedly.

END

Experiment Results (Japanese-English Translation)

Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	23.83	72.27
Parser 1-best	**Random**	96	23.66	71.77
Decoder 1-best	**Random**	97	23.81	72.04
Oracle	**Random**	97	23.93	72.09
Oracle	**BLEU+1 ≧ 0.7**	206	** 24.27	72.38
Oracle	**BLEU+1 ≧ 0.8**	120	** 24.26	72.38
Oracle	**BLEU+1 ≧ 0.9**	58	** 24.26	72.49
Oracle	**BLEU+1 Gain**	100	* 24.22	72.32

*** : $p < 0.05$**
**** : $p < 0.01$**

Experiment Results (Japanese-Chinese Translation)

Tree Selection	Sentence Selection	Sentences (k)	BLEU	RIBES
Baseline		-	29.60	81.32
Parser 1-best	**Random**	129	29.75	** 81.55
Decoder 1-best	**Random**	130	29.76	* 81.53
Oracle	**Random**	130	** 29.89	** 81.66
Oracle	**BLEU+1 ≧ 0.7**	240	** 29.86	** 81.60
Oracle	**BLEU+1 ≧ 0.8**	150	** 29.91	81.47
Oracle	**BLEU+1 ≧ 0.9**	82	* 29.86	** 81.60
Oracle	**BLEU+1 Gain**	100	* 29.85	** 81.59
Oracle (ja-en)	**BLEU+1 ≧ 0.8**	120	* 29.87	* 81.58

*** : $p < 0.05$**
**** : $p < 0.01$**

Why decoder 1-best parse tree is better than parser 1-bset?

- Probability considered in Forest-to-String translation
  - Parse tree probability
  - Translation model
  - Language model
- The rule that use correct tree have high probability on translation model.
  - The rule that use incorrect tree have low probability.
- By using language model, the correct parse tree tends to be chosen.
  - The correct tree have high probability on language model.