Monolingual Phrase Alignment as Parse Forest Mapping

Sora Kadotani*, Yuki Arase**

***NTT Human Informatics Laboratories, **Osaka University**

*This work was completed at Osaka University

Outline

- Tackle the problem of monolingual phrase alignment conforming to syntactic structures

- The existing method formulates phrase alignment as the unordered tree mapping
  → the alignment quality is affected by syntactic ambiguities

- The proposed method aligns parse forests rather than 1-best trees

- The experimental results indicated that our method improves the syntactic phrase alignment quality of the state-of-the-art method

Monolingual Phrase Alignment

- Identify semantically corresponding phrase pairs in sentences
- Fundamental technique for many tasks
  - Paraphrase recognition
  - Textual entailment recognition
  - Semantic textual similarity estimation

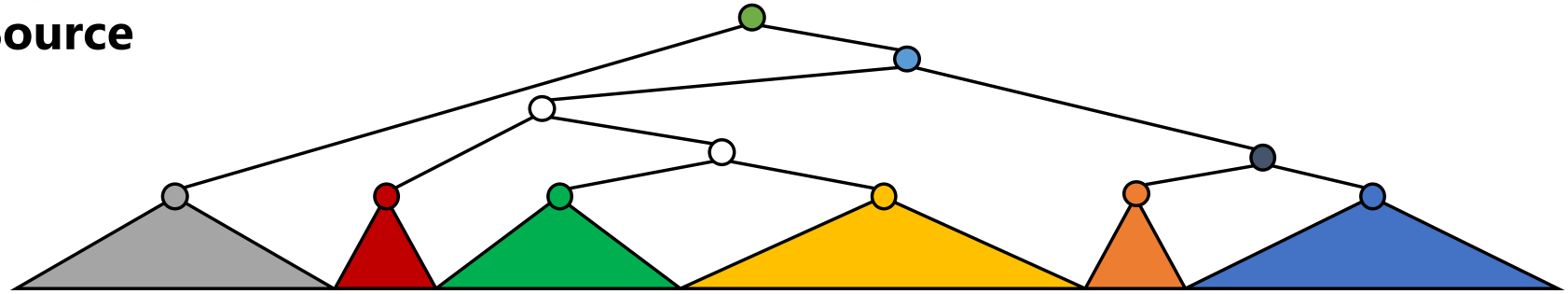
Source:
Yousaf Raza Gilani made this statement at the Karachi Shipyard when talking with reporters

Target:
Yousaf Raza Gilani gave this statement while speaking to media persons at Karachi Shipyard

TreeAligner [Arase and Tsujii, EMNLP'20]

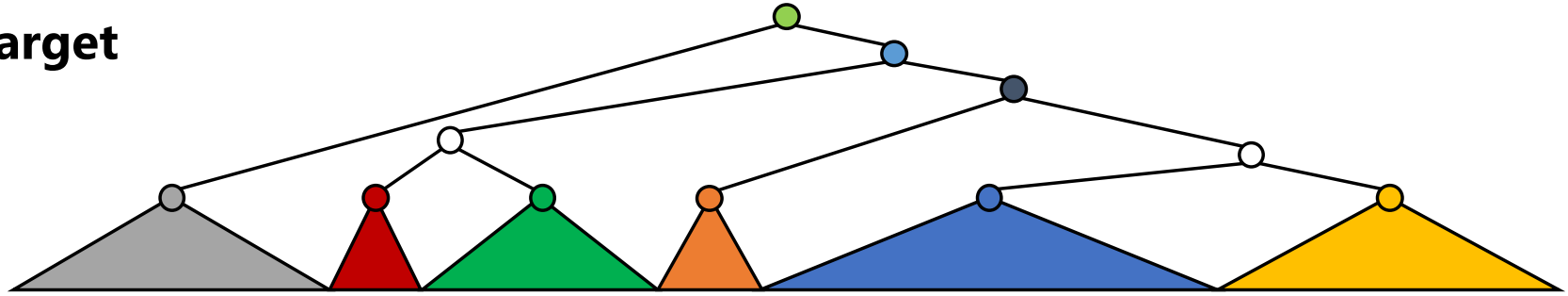
- Formulate phrase alignment as the unordered tree mapping between source and target constituency trees
- Based on dynamic programming
  - Recursively align phrases from leaves to root nodes

Source



Yousaf Raza Gilani made this statement at the Karachi Shipyard when talking with reporters

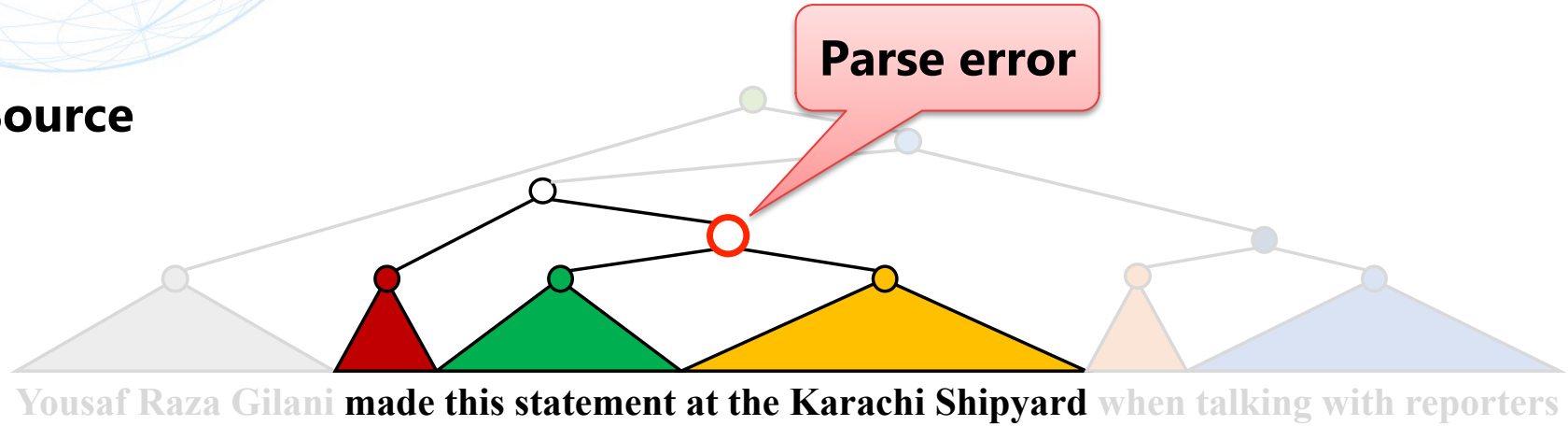
Target

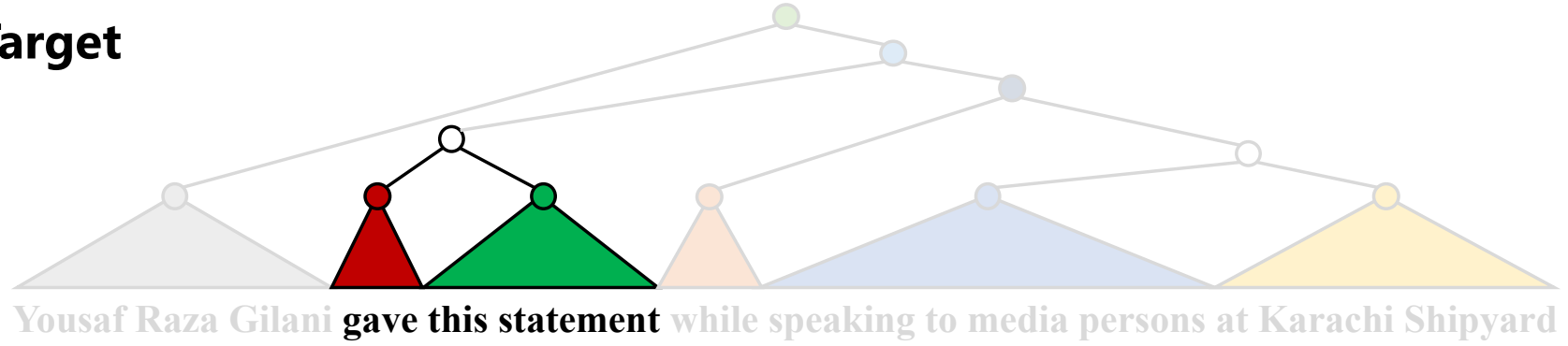


Yousaf Raza Gilani gave this statement while speaking to media persons at Karachi Shipyard

TreeAligner [Arase and Tsujii, EMNLP'20]

- Syntactic ambiguities cause parse errors in practical parsers
- Parse errors cannot be corrected in *TreeAligner* and degrade the performance of phrase alignment

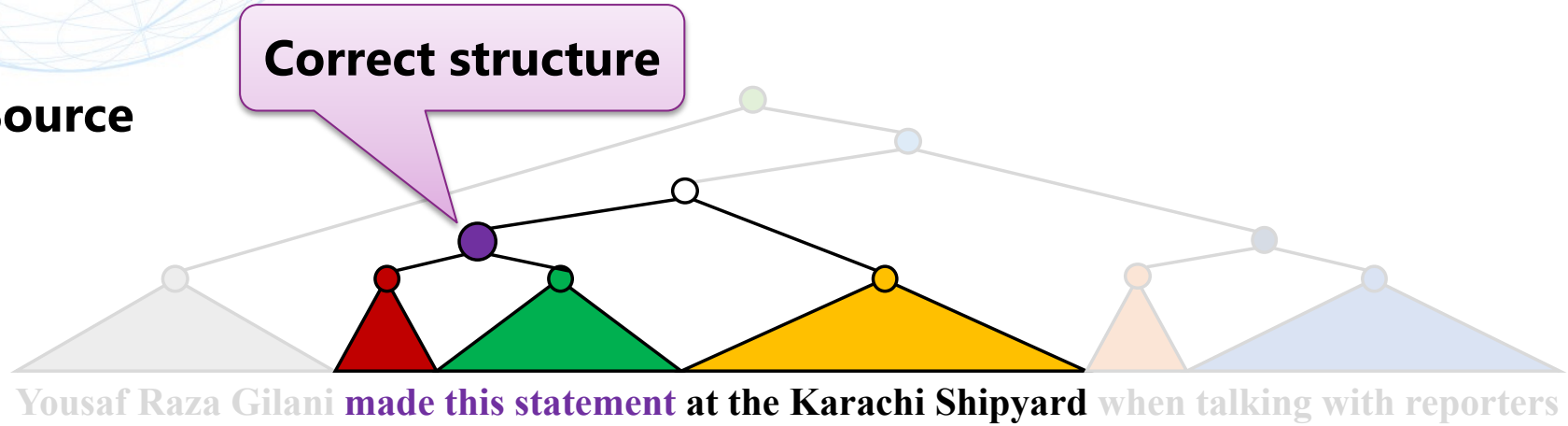




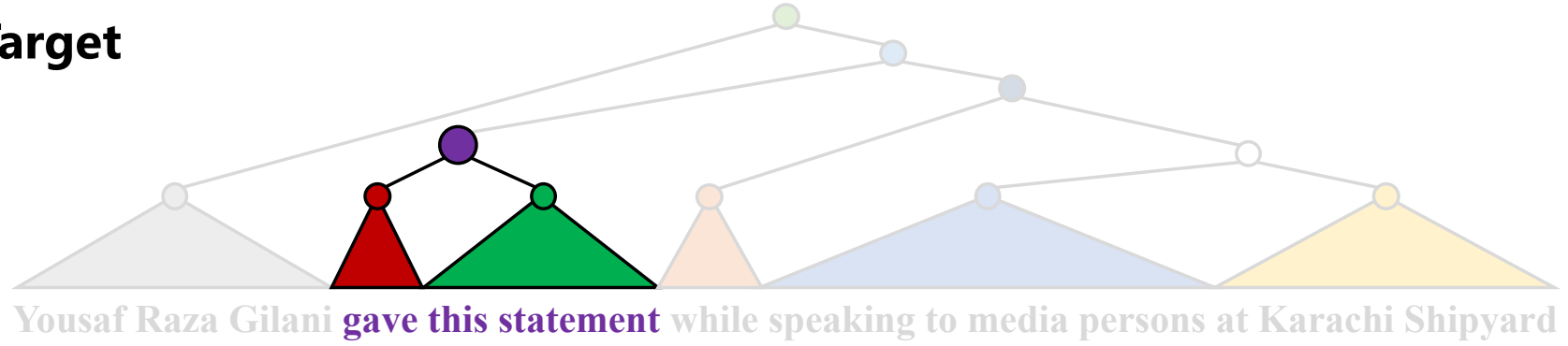
TreeAligner [Arase and Tsujii, EMNLP'20]

- Syntactic ambiguities cause parse errors in practical parsers
- Parse errors cannot be corrected in *TreeAligner* and degrade the performance of phrase alignment

Source



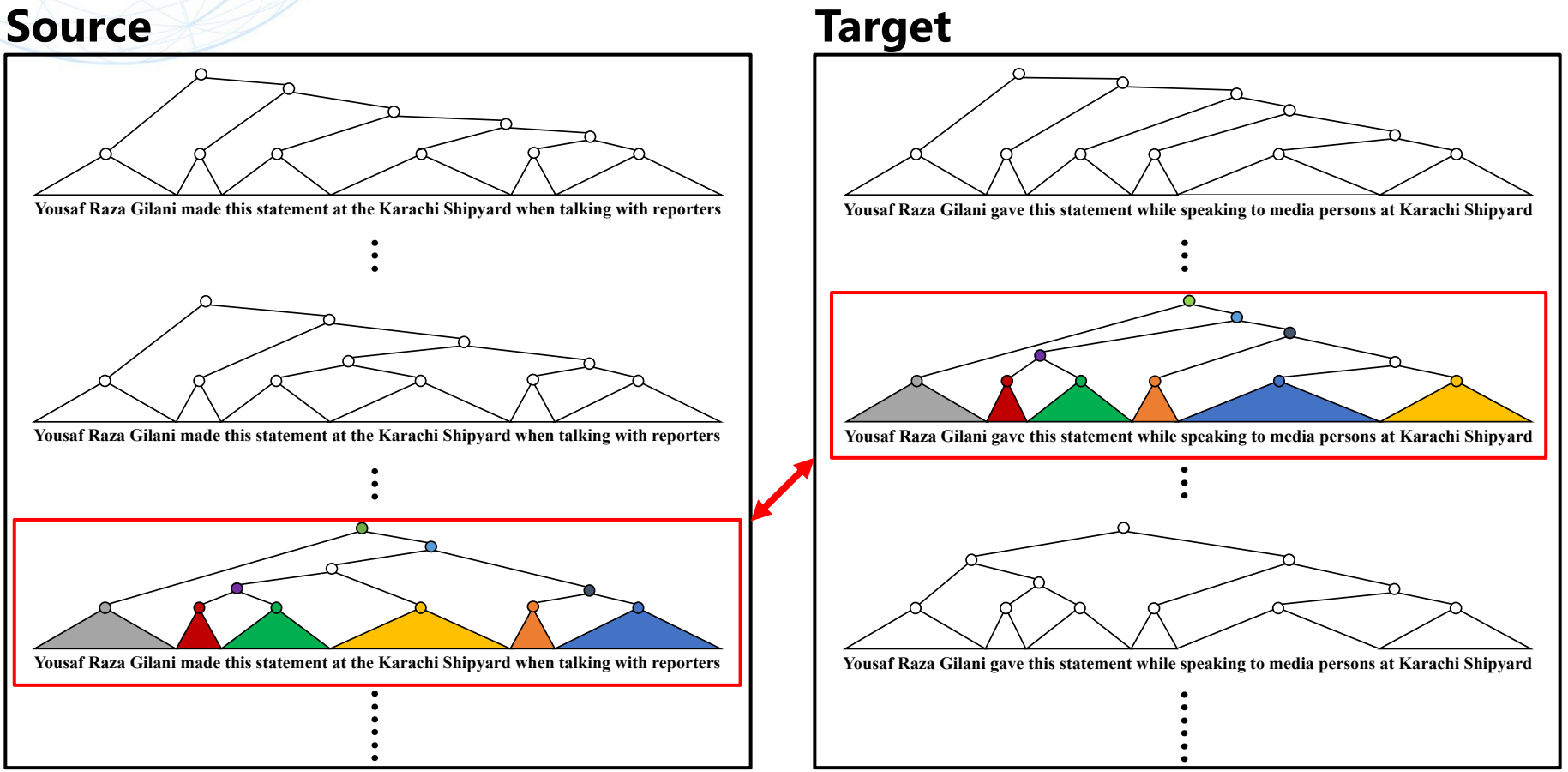
Target



Proposed Method: *ForestAligner*

Set of possible constituency trees

- Expand *TreeAligner* to align parse forests instead of trees
- Syntactic ambiguities can be resolved by referring to the other sentence



Naive Approach to Align Forests

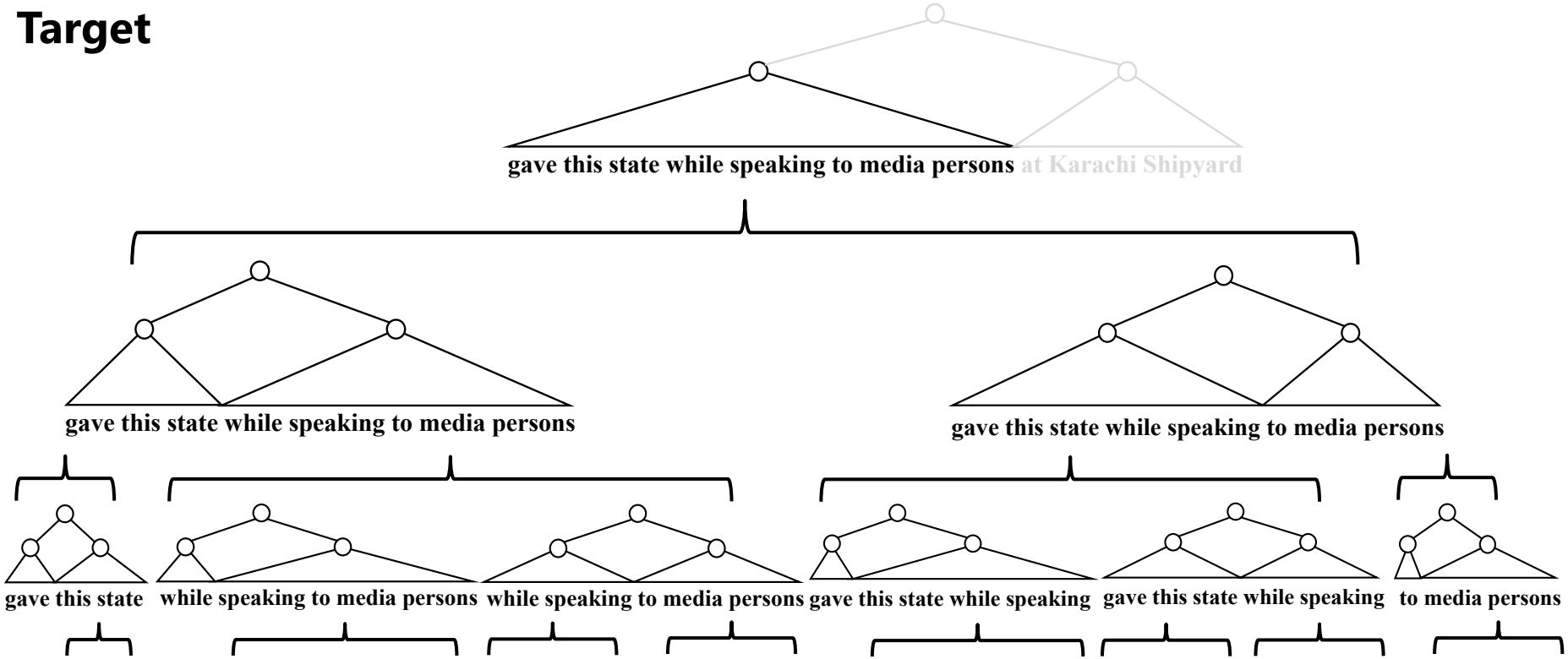
- Align the all combinations of possible trees and find the best pair

- Computationally expensive
  - If there are n and m source and target possible constituency trees, we need to compute phrase alignment for $n \times m$ combinations
  - The number of possible constituency trees of a sentence can be hundreds

Packed Forest Structure

- *ForestAligner* achieves efficient alignment by mapping forests using a packed structure
- Packed forest structure compactly stores possible syntactic structures under the same nodes

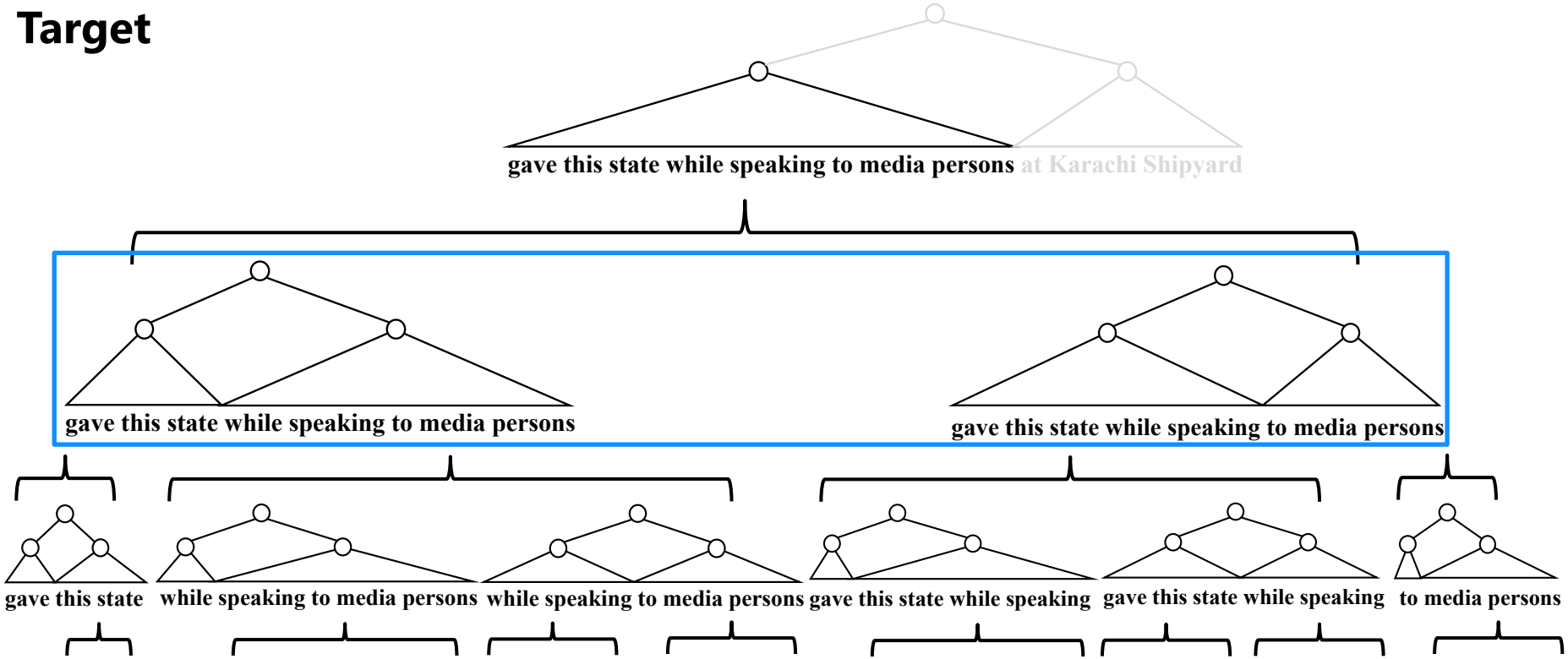
Target



Packed Forest Structure

- *ForestAligner* achieves efficient alignment by mapping forests using a packed structure
- Packed forest structure compactly stores possible syntactic structures under the same nodes

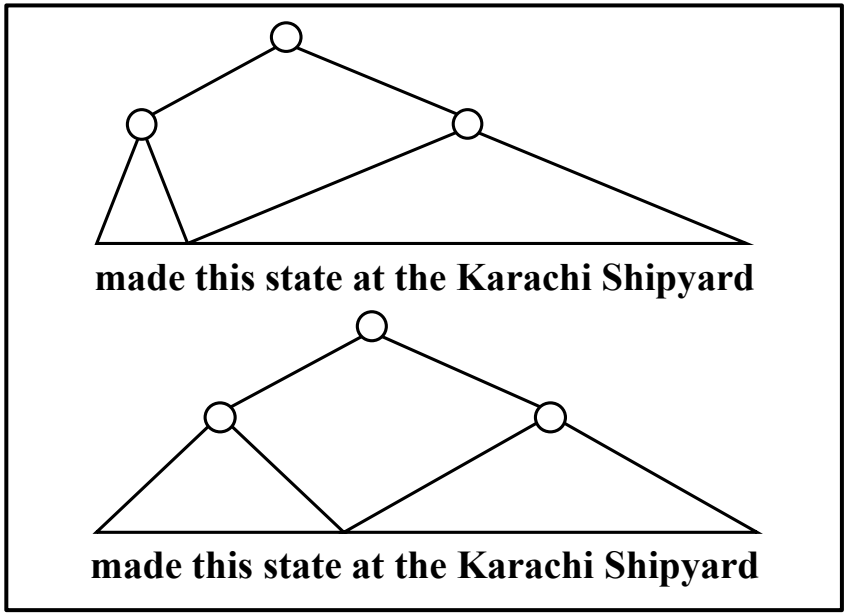
Target



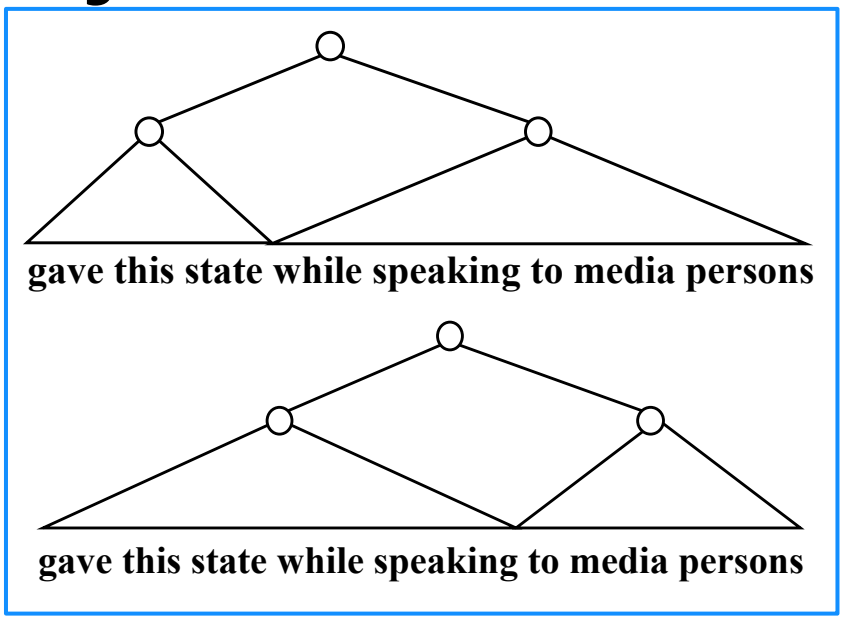
Alignment on Packed Forests

- Recursively execute the following procedures from leaves to root nodes

Source

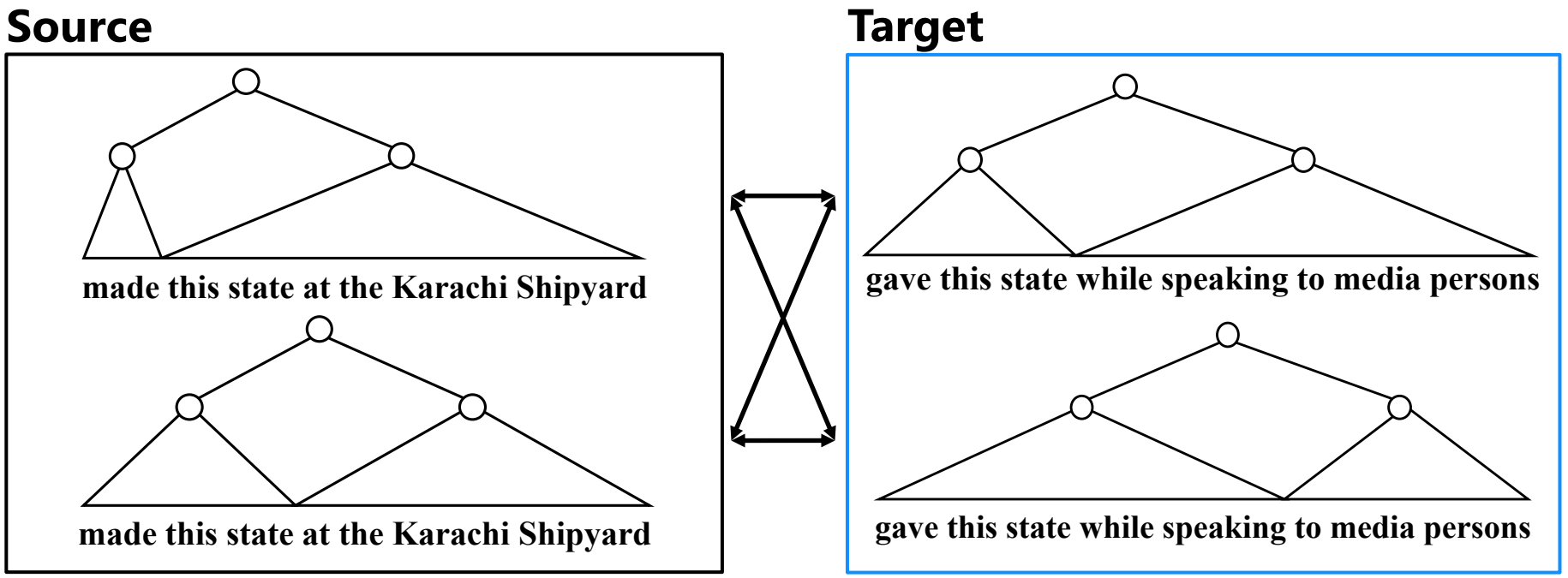


Target



Alignment on Packed Forests

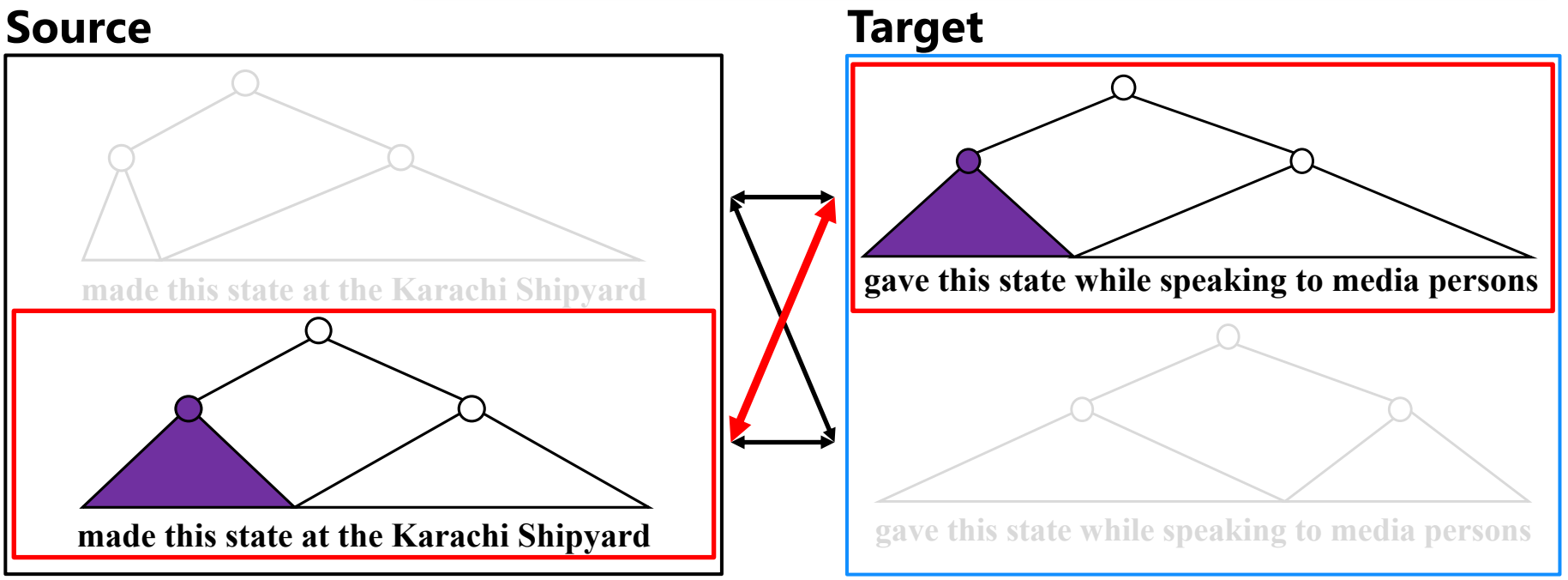
- Recursively execute the following procedures from leaves to root nodes
  1. Compute the cost to align all combination of possible subtrees



Alignment on Packed Forests

- Recursively execute the following procedures from leaves to root nodes
  1. Compute the cost to align all combination of possible subtrees
  2. Memorise only the pair with the minimum cost

Will be considered in the alignment of the upper nodes



Syntactic Plausibility

- *ForestAligner* considers the likelihood of not only phrase alignment but also syntactic structure
- Define the cost to align forests as follows

$$\widehat{D}\left(F_i^s, F_j^t\right) = D\left(F_i^s, F_j^t\right) - \lambda_{\text{syn}} \frac{S(T_i^s) + S(T_j^t)}{2}$$

T_k: the subtree rooted at node k
F_k: the forest under node k
$D(\cdot,\cdot)$: the tree edit distance between forests
$S(\cdot)$: the likelihood of a subtree obtained from a syntactic parser
λ_{syn} : the hyperparameter that balances both terms

Experiment: Setup

- **Dataset: SPADE** **[Arase and Tsujii, LREC'18]**
  - Consist of the gold constituency trees of English paraphrase sentence pairs with phrase alignment annotations
  - Dev: 50 sentence pairs (8,708 phrase pairs)
    Test: 151 sentence pairs (25,709 phrase pairs)
- **Evaluation Metrics**
  - Alignment recall (**ALIR**), precision (**ALIP**), F-measure (**ALIF**)
  - Phrase span matching ratio (**PSMR**) against the gold trees
- **Parser: Enju** **[Miyao and Tsujii, CL'08]**
  - Based on a wide-coverage probabilistic HPSG grammar
- **Comparison Method: *TreeAligner***
  - The previous state-of-the-art method
  - Use 1-best trees

Experiment: Results

- Phrase alignment quality (ALIR, ALIP, ALIF)
  - The scores of *TreeAligner* significantly dropped when aligning 1-best trees compared to the case of aligning the gold trees
  - *ForestAligner* improved ALIR by **1.3%**, ALIP by **2.6%**, ALIF by **2.0%** compared to *TreeAligner* with 1-best trees
  - Confirm the effectiveness of forest alignment
- Correctness of phrase structures (PSMR)
  - *ForestAligner* moderately improved *TreeAligner* by **0.3%**
  - Investigate what kind of parse errors in 1-best trees were fixed and newly introduced by *ForestAligner*

	Structure	ALIR (%)	ALIP (%)	ALIF (%)	PSMR (%)
TreeAligner	Gold tree	88.2	86.6	87.4	100.0
TreeAligner	1-best tree	79.8	76.7	79.3	93.1
ForestAligner	Forest	**81.1**	**79.3**	**80.2**	**93.4**

Experiment: Analysis

- How to analyze
  1. Randomly sample 40 sentences where the PSMR score increased (20 sentences) and decreased (20 sentences) compared to *TreeAligner*
  2. Manually categorize the sampled sentences into error types
- *ForestAligner* tends to fix NP attachment errors while increases modifier attachment errors

Error type	Improved	Deteriorated
PP attachment	8	8
NP attachment	5	1
Modifier attachment	1	4
Coordination	2	2
Other	4	5

Summary

- Tackled the problem of monolingual syntactic phrase alignment

- Existing method: *TreeAligner*
  - Formulate phrase alignment as the unordered tree mapping
  - Have the problem that the alignment quality is easily affected by syntactic ambiguities

- Proposed Method: *ForestAligner*
  - Expand *TreeAligner* to align parse forests rather than 1-best trees
  - Achieve efficient alignment by mapping forests on a packed structure

- Confirmed the effectiveness of forest alignment through the experiment