対訳コーパスを利用した構文解析器の自己学習

奈良先端科学技術大学院大学
知能コミュニケーション研究室
森下 睦・赤部 晃一・波多腰 優斗
Graham Neubig・吉野 幸一郎・中村 哲

言語処理学会 第22回年次大会
2016/3/10

NAIST®

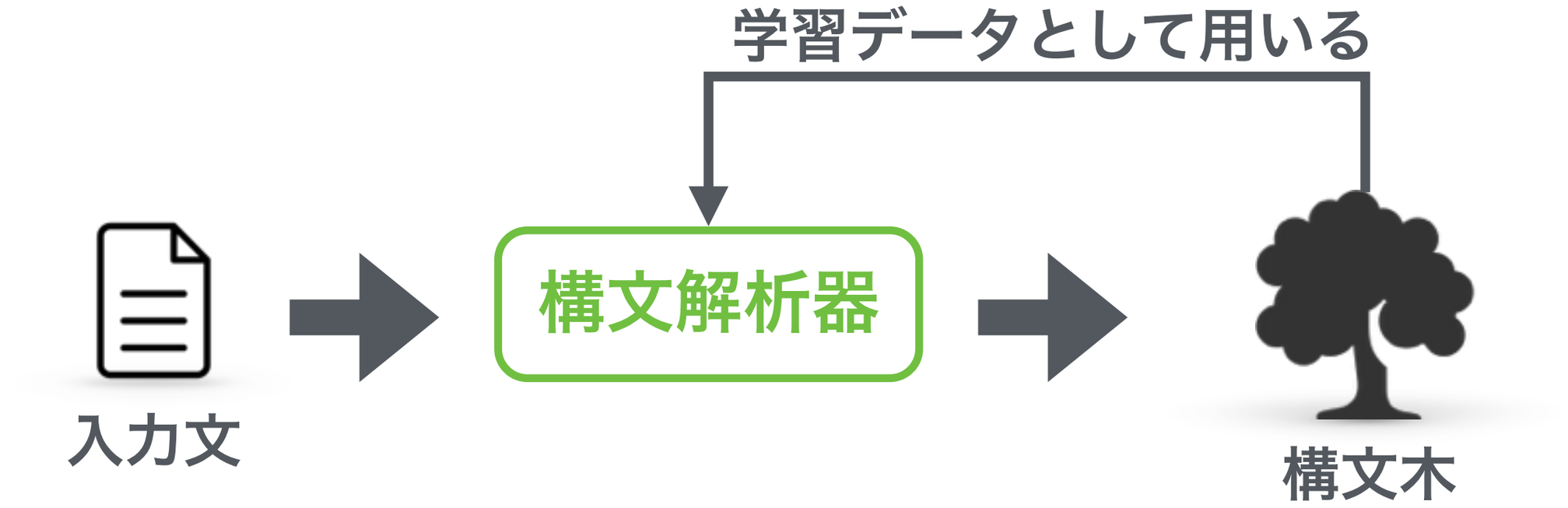
背景

構文解析器の学習には

- 人手でアノテーションされた構文木を学習データとする
- アノテーション作業にはコストがかかる
  - コストをかけずに正しい構文木を作成したい

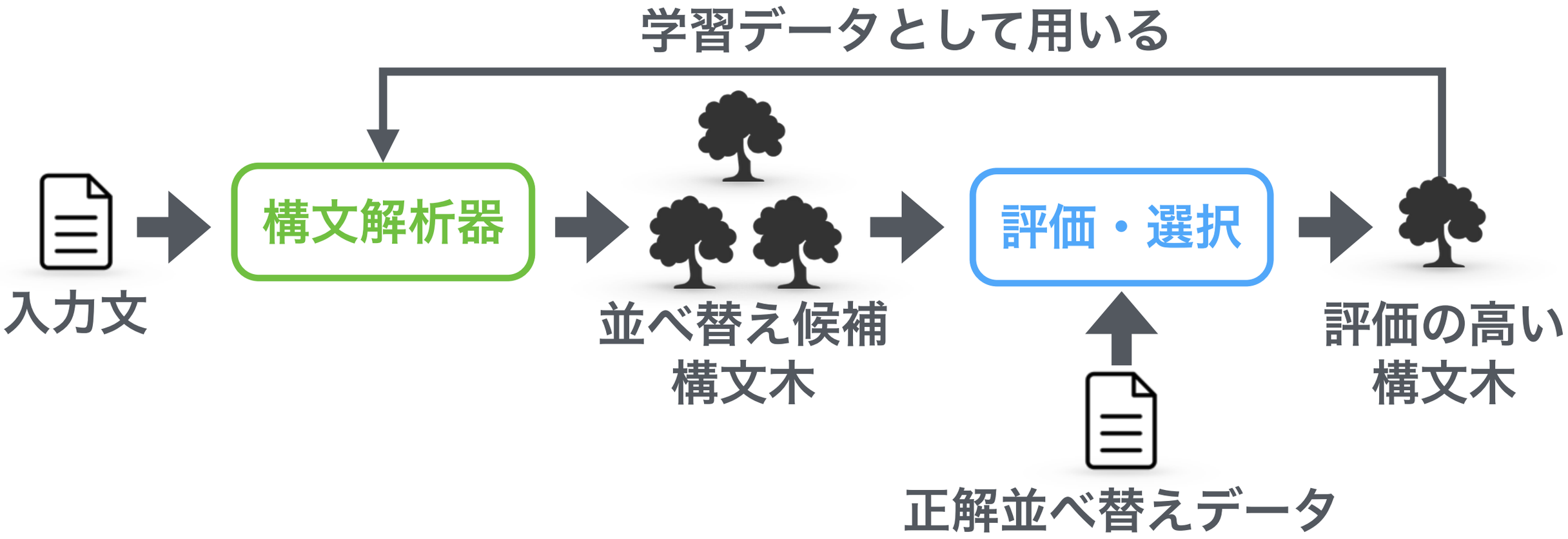
構文解析器の自己学習

[McClosky et al., 2006]



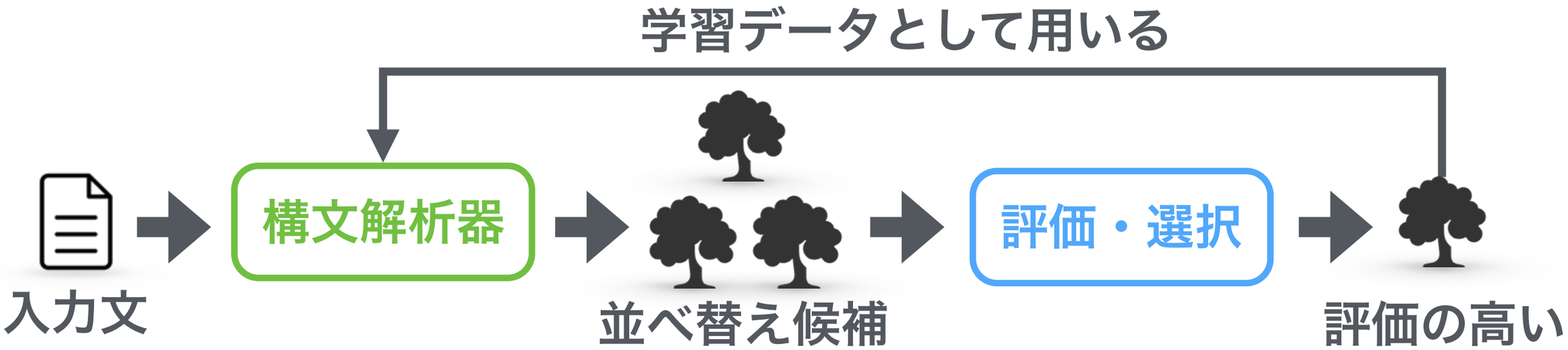
- 構文解析器の出力を学習データとする
- 構文解析器の精度を向上
  - 入力文のドメインへ適応する効果がみられる
- 学習データには誤ったデータが混入している可能性

事前並べ替えにおける構文解析器の自己学習 [Katz-Brown et al., 2011]



- 学習するデータを選択する標的自己学習による効果的な自己学習
  - 改善に役立つデータのみを学習に用いる
  - ただしこの手法では正解並べ替えデータが必要
  - 正解データを作るためには大きなコストがかかる

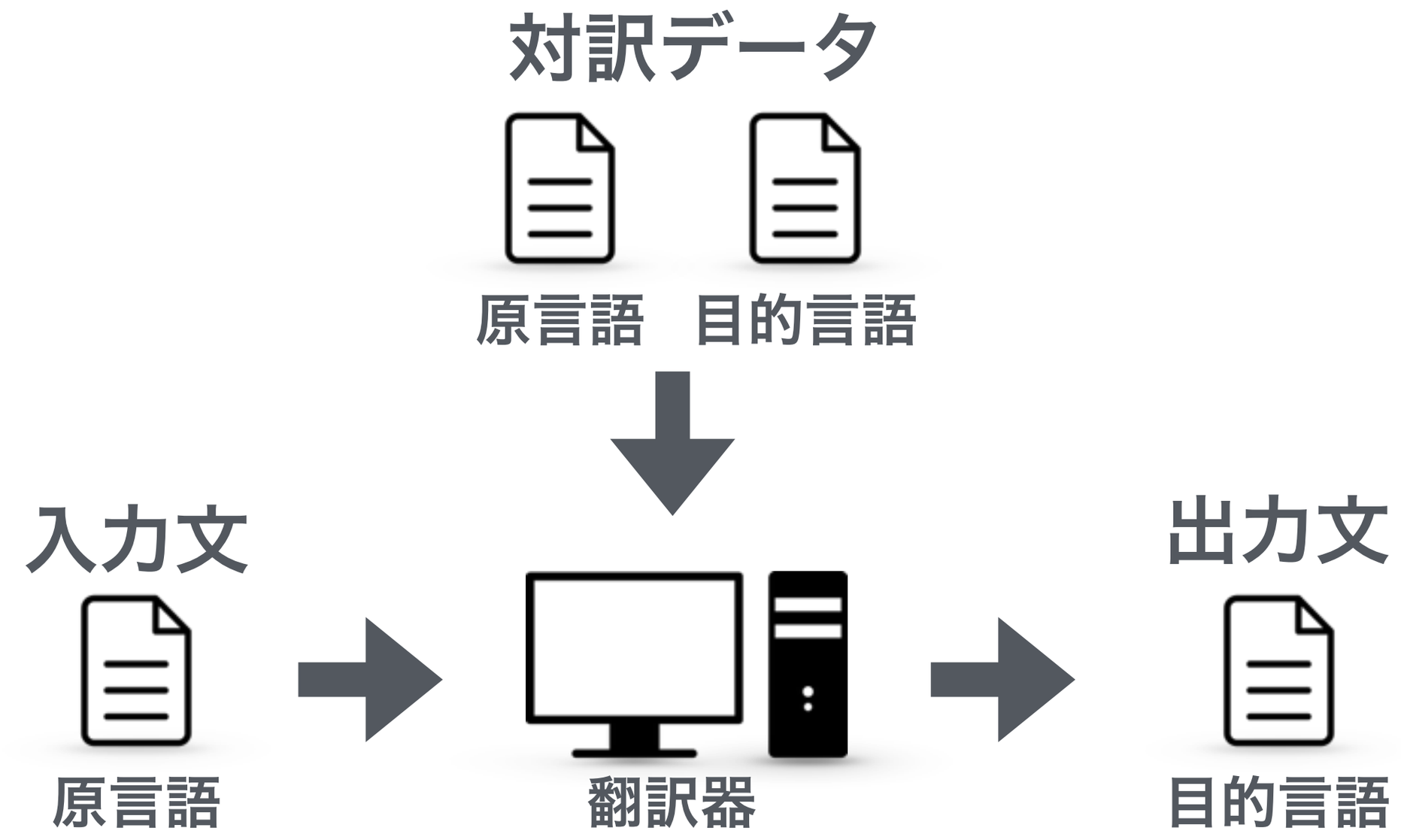
モチベーション



- 評価・選択をコストをかけずに行いたい
  - 構文木を用いた機械翻訳を使用
  - 正しい構文木の翻訳精度は高い
  - 機械翻訳の自動評価尺度で構文木の精度を計測
    - 対訳コーパスのみで評価可能

機械翻訳

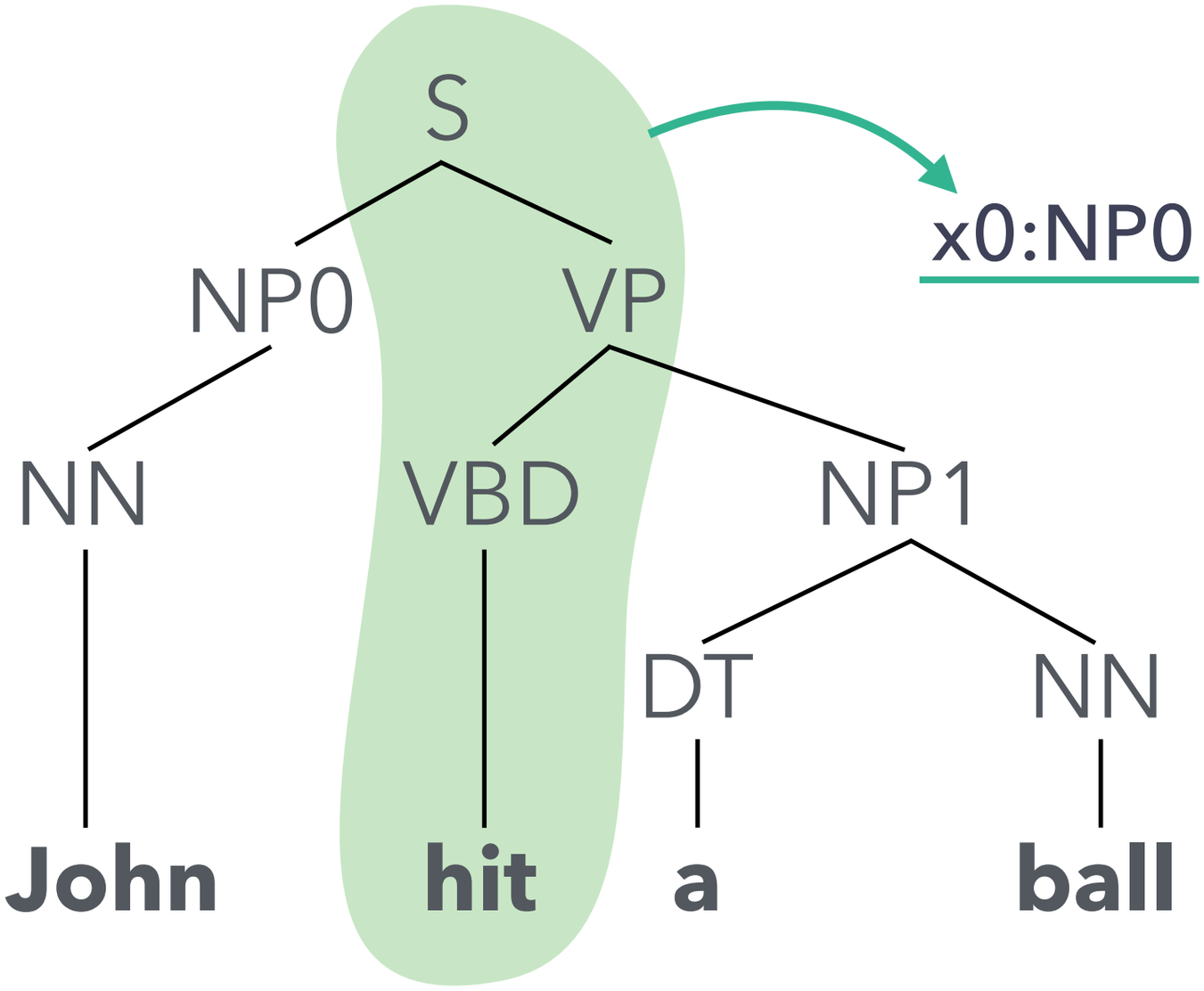
統計的機械翻訳



- 対訳データから翻訳規則を自動的に学習
- 翻訳規則に基づいて原言語文を目的言語に変換

Tree-to-String翻訳

[Liu et al., 2006]



x0:NP0 は x1:NP1 を 打った

- 原言語の構文木を翻訳に利用
  - 構文解析の誤りが翻訳結果に悪影響を及ぼす
  - 構文木が正しい場合，正しい訳になりやすい

Forest-to-String翻訳 [Mi et al., 2008]

原言語構文森

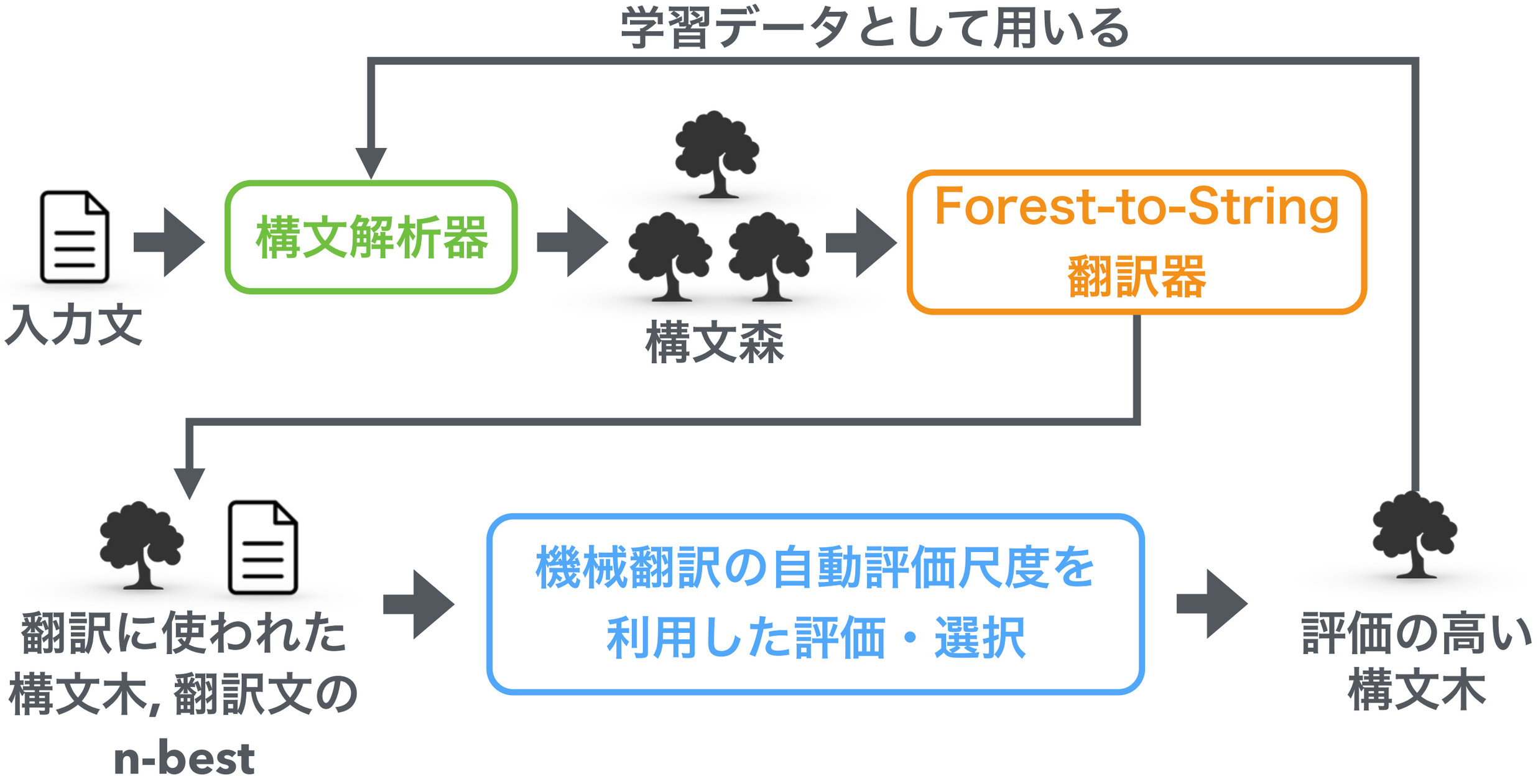
Forest-to-String
翻訳器

目的言語文

- 原言語の構文森を翻訳に利用
  - 構文木の候補から, 翻訳モデルのスコアが高くなる構文木を選択でき, 翻訳精度の改善につながる [Zhang et al., 2012]

提案手法

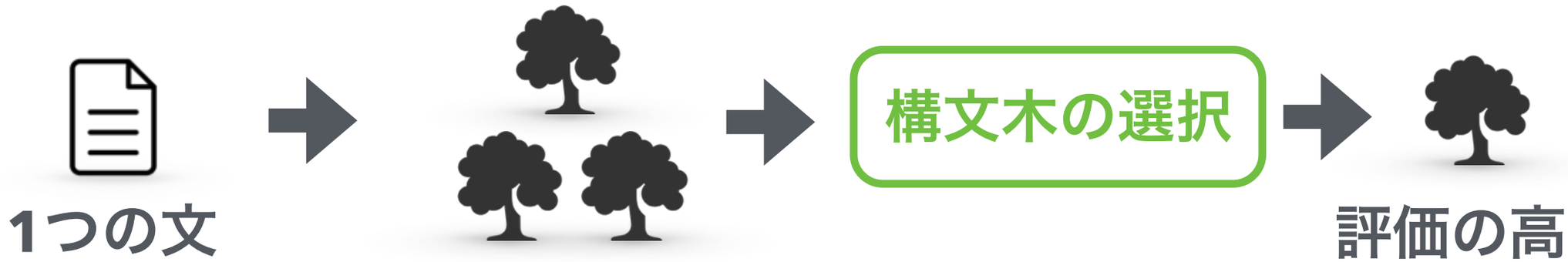
提案手法



- 機械翻訳の自動評価尺度を利用した標的自己学習
  - 低コストかつ適切な選択が可能

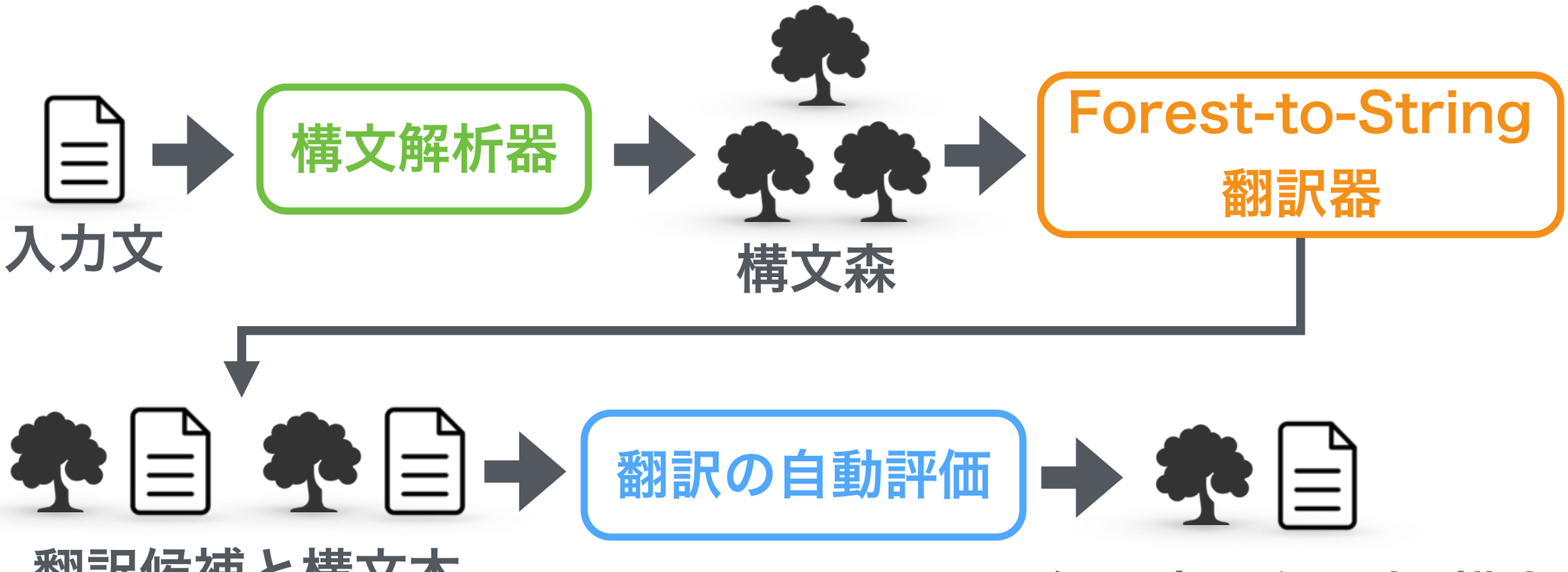
選択方法

- 構文木の選択
  - 文の構文木の候補からどれを学習するか選択



- 文の選択
  - どの文を自己学習に使用するか選択

構文木の選択法

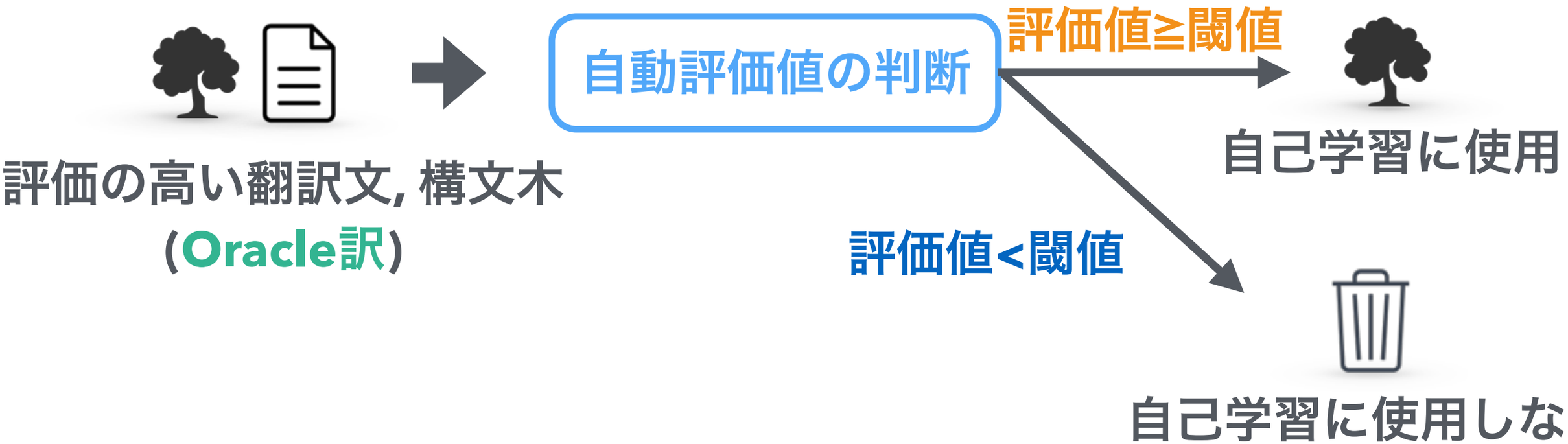


翻訳候補と構文木

評価の高い翻訳文, 構文木
(Oracle訳)

- 訳の候補の中から, 機械翻訳の自動評価値が最も高い訳に使われた構文木を使用
  - 自動評価尺度が最も高い訳のことをOracle訳という

文の選択法



- 自動評価値が一定以上の訳の構文木のみ使用

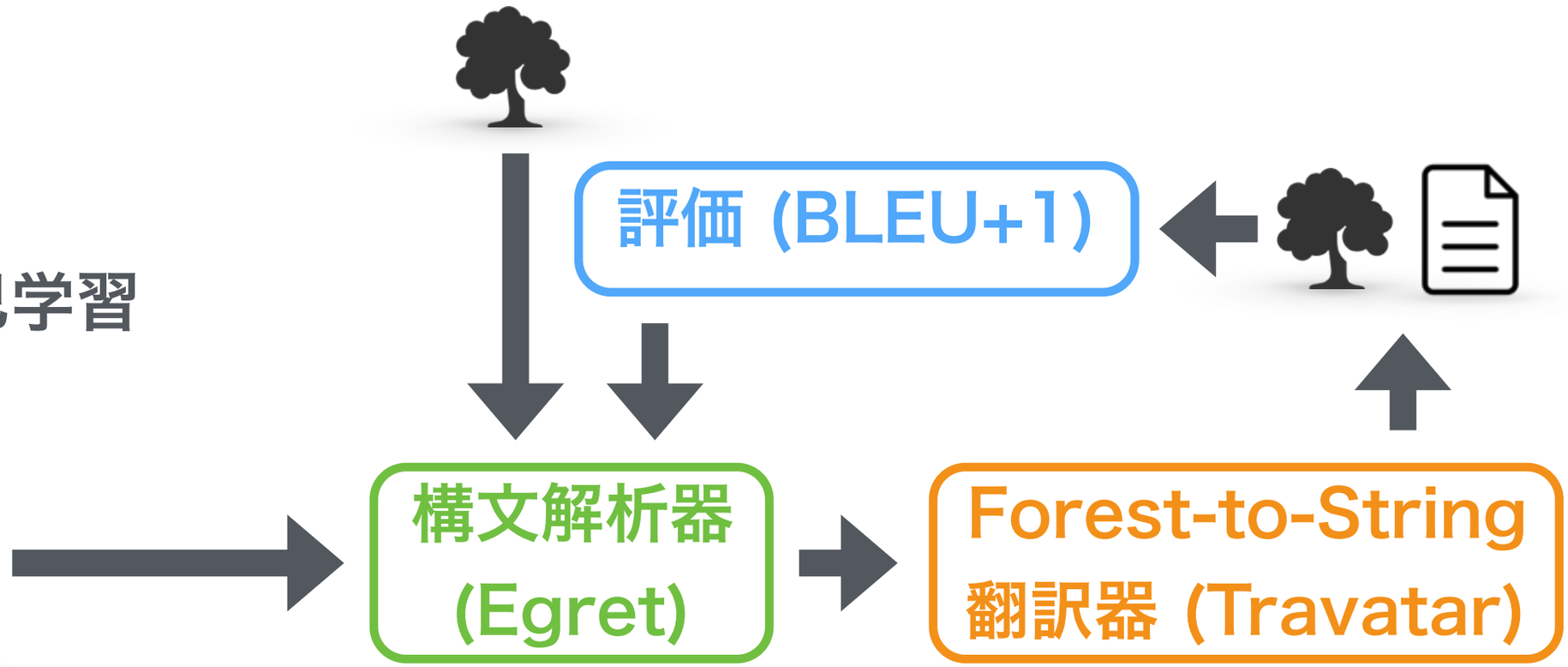
実験的評価

実験設定 (自己学習時)

既存モデル
日本語係り受けコーパス (7k)

構文解析器 自己学習
対訳コーパス
(ASPEC 2.0M)

原言語 目的言語



実験設定 (精度評価時)

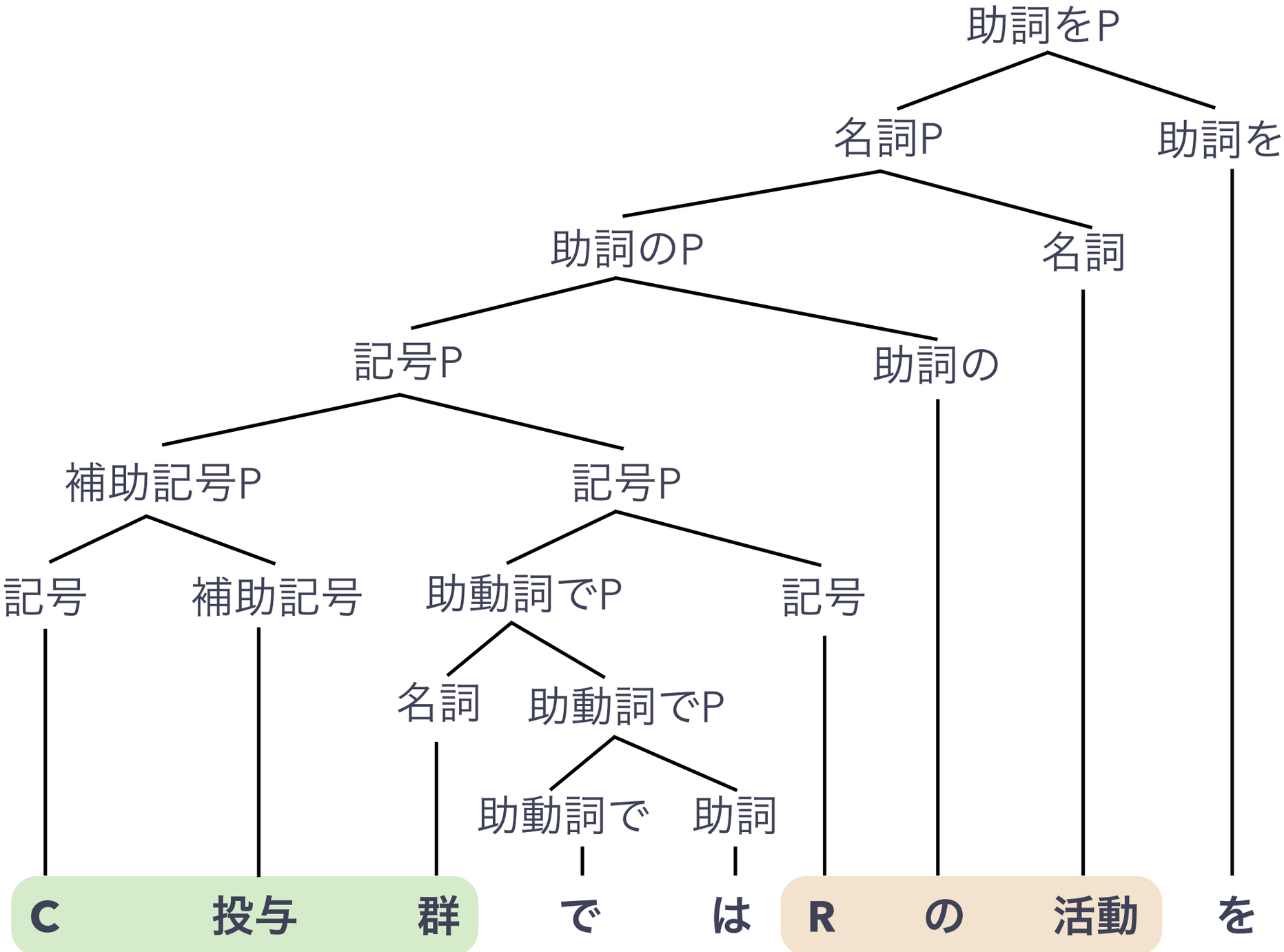
ASPEC テストセット
100文

- 手動で100文の正解構文木を作成
- Evalb: 構文解析精度計測ツール
  - F値を計測

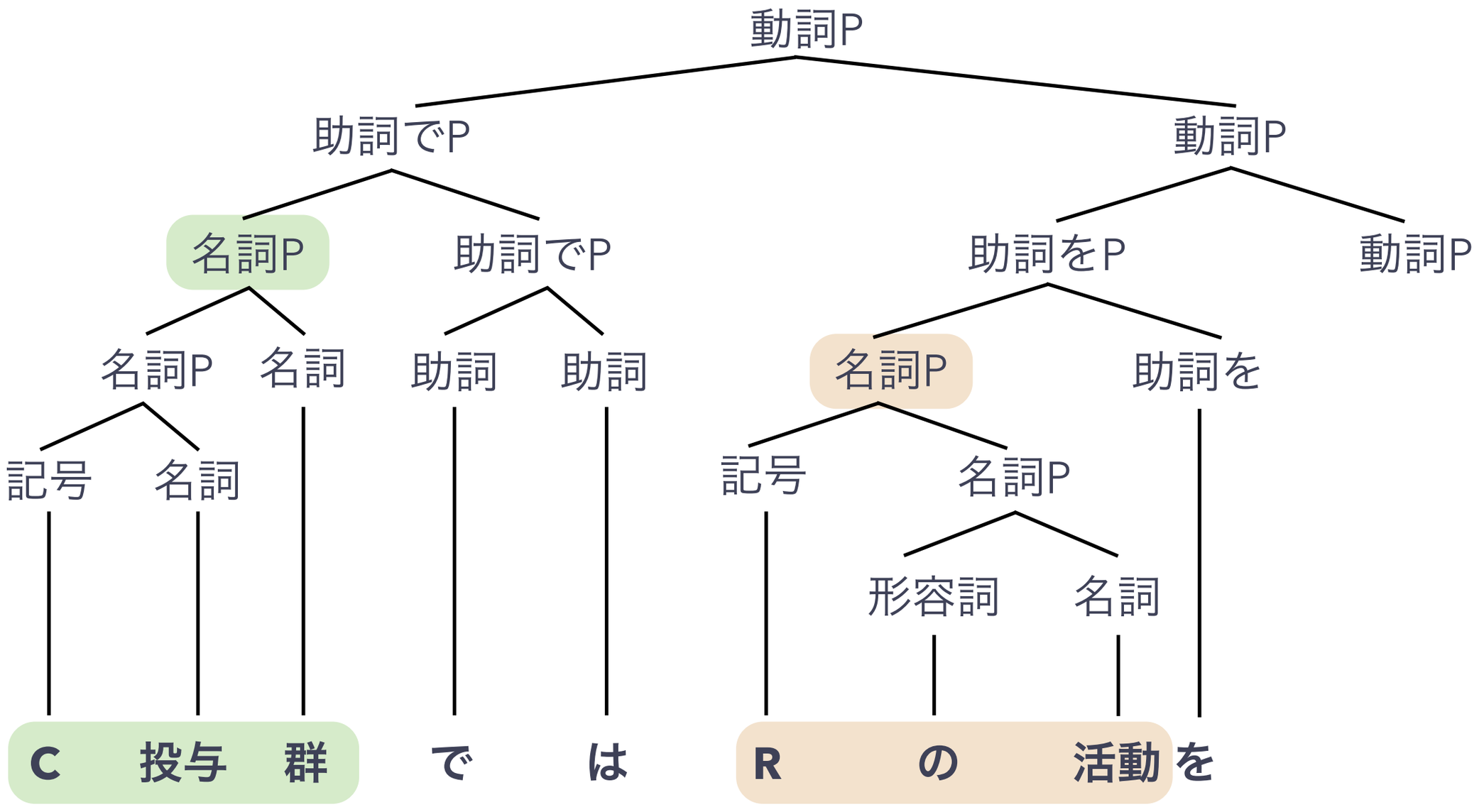
実験結果

	自己学習手法	使用文数	F値	(a)との有意差	(b)との有意差
(a)	ベースライン (自己学習無し)	—	84.83	—	—
(b)	既存自己学習手法 [McClosky et al., 2006]	96k	86.46	あり (95%)	—
(c)	BLEU+1 ≧ 0.7	206k	88.07	あり (99%)	あり (95%)
(d)	BLEU+1 ≧ 0.8	120k	88.07	あり (99%)	あり (95%)
(e)	BLEU+1 ≧ 0.9	58k	87.23	あり (99%)	なし

自己学習前の構文木


助詞をP
名詞P
助詞を
助詞のP
名詞
記号P
助詞の
補助記号P
記号P
記号
補助記号
助動詞でP
記号
名詞
助動詞でP
助動詞で
助詞
C
投与
群
で
は
R
の
活動
を

自己学習により改善された構文木



まとめ

まとめ

- 対訳コーパスを用いた
  構文解析器の自己学習により解析精度が向上
  - 既存の対訳コーパスを構文解析器の学習に活用可能
  - 既存自己学習手法と比較しても有意に精度向上

- 今後の課題
  - さらに幅広い分野で手法が有効であることを確認
  - 新たな構文木，文選択手法について検討

END