様々な分野における対訳コーパスを用いた構文解析器の自己学習効果の検討

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・小田悠介・Graham Neubig・吉野 幸一郎・中村 哲

第226回自然言語処理研究会
2016/05/16

NAIST®

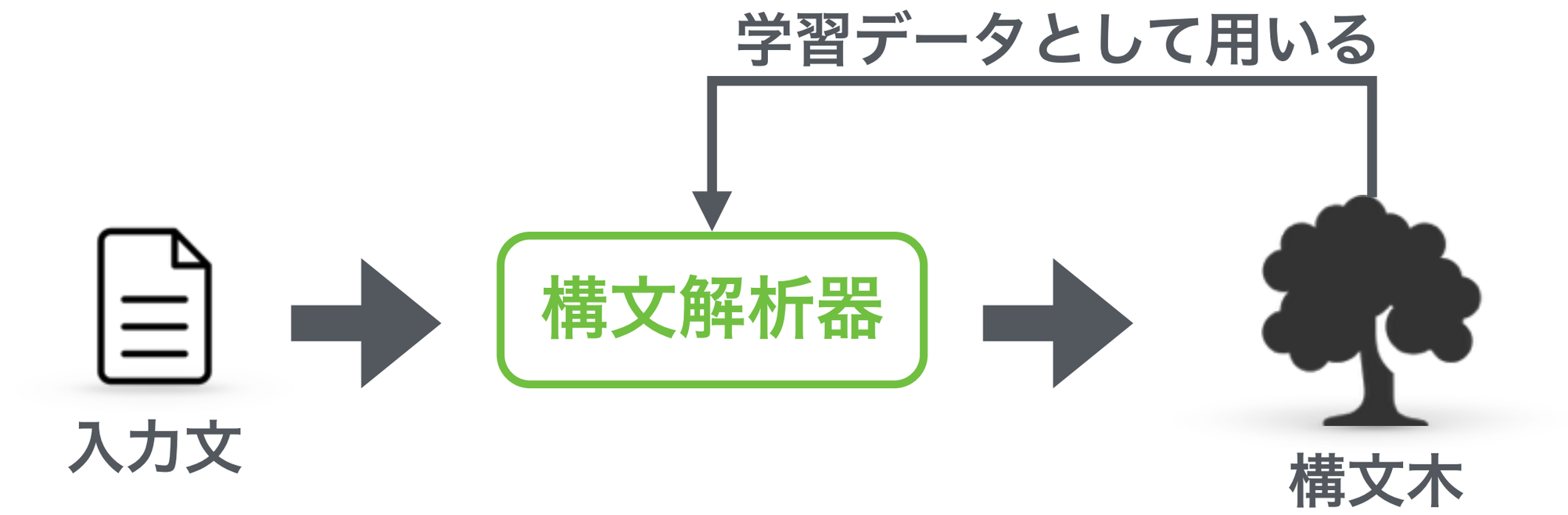
背景

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

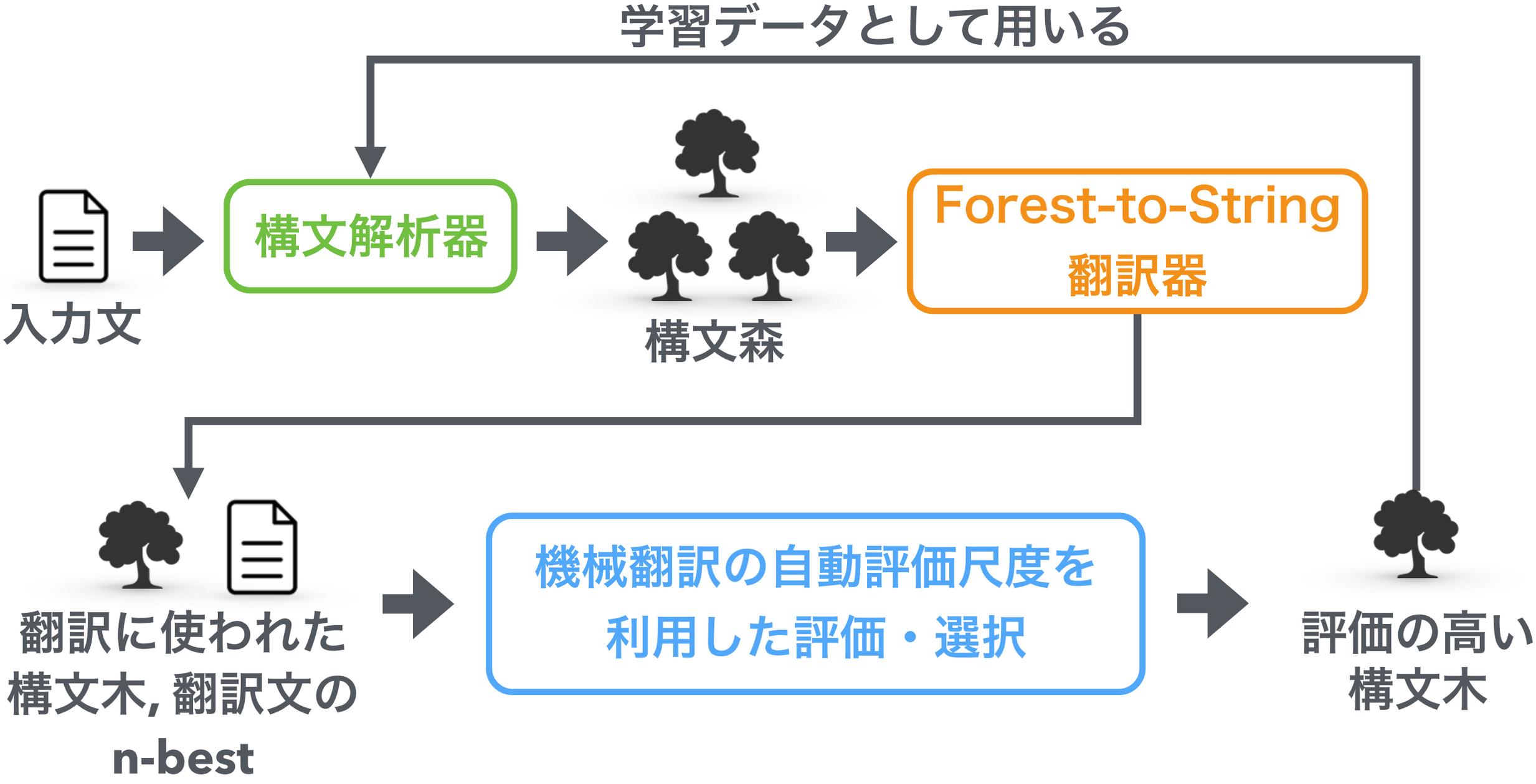
[McClosky et al., 2006]



- 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
  - 入力文の分野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

対訳コーパスを用いた構文解析器の自己学習

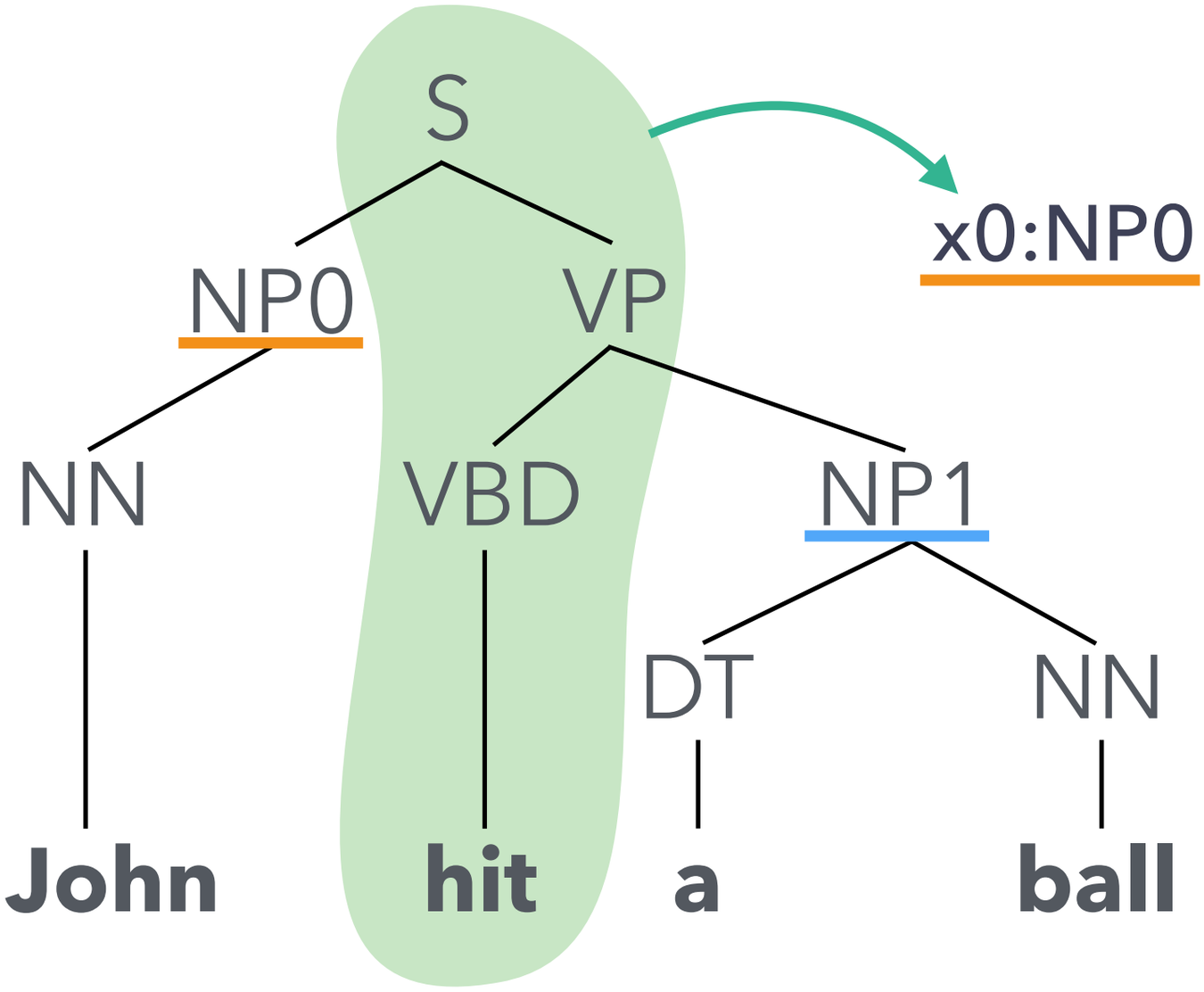
[Morishita et al., 2015]



- 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

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

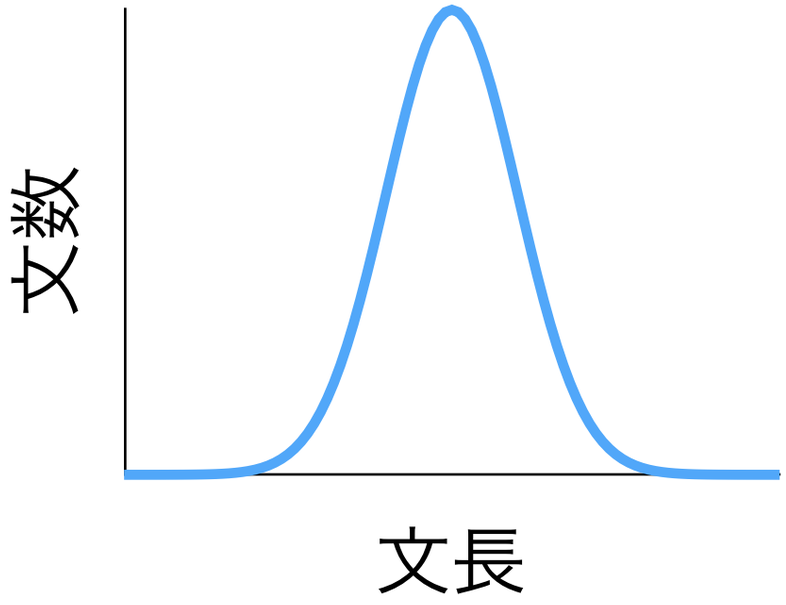
Forest-to-String
翻訳器

目的言語文

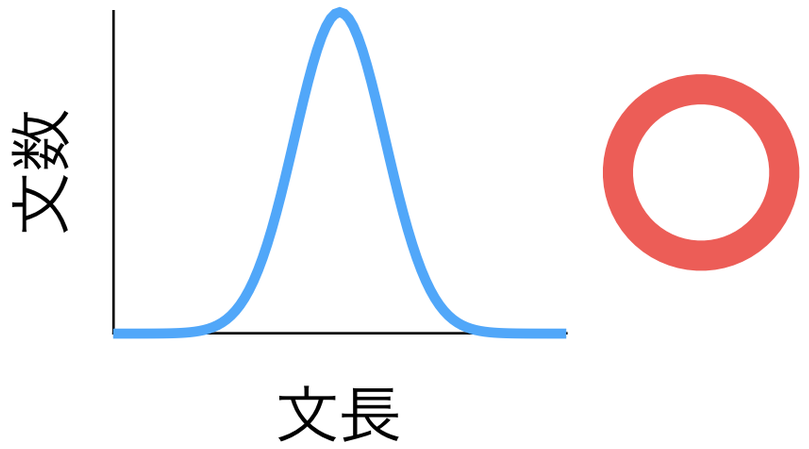
- 原言語の構文森を翻訳に利用
  - 構文木の候補から, 翻訳モデルのスコアが高くなる構文木を選択でき, 翻訳精度の改善につながる [Zhang et al., 2012]

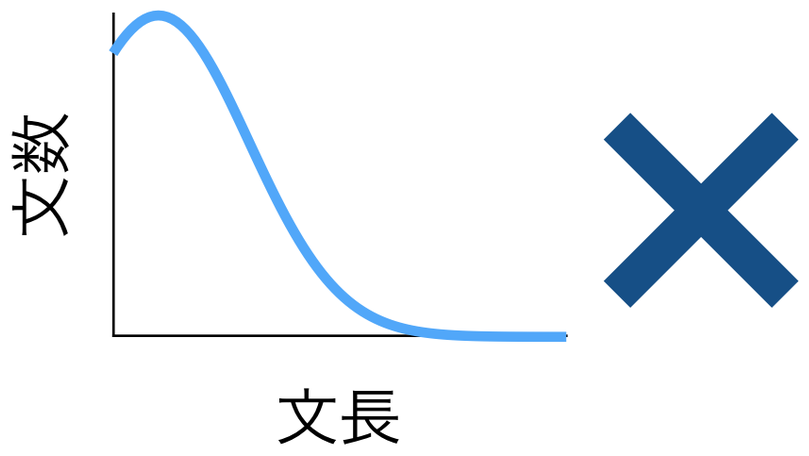
文長の分布の保持

コーパスの文長分布



選択された文長分布





- 選択された文の文長分布をコーパス全体と一致させる
  - 短い文ばかりが選ばれることを防ぐ

今までわかっていること

- 単一分野に対して適用した場合,
  構文解析精度が向上する

  - ASPEC (科学技術論文抜粋) に対して適用
  - 複数の分野に対して適用した事例はない

本研究で明らかにすること

- 対訳コーパスを用いた構文解析器の自己学習は,
  様々な分野に対して適用可能か
  - 既存の自己学習手法を超えられるか

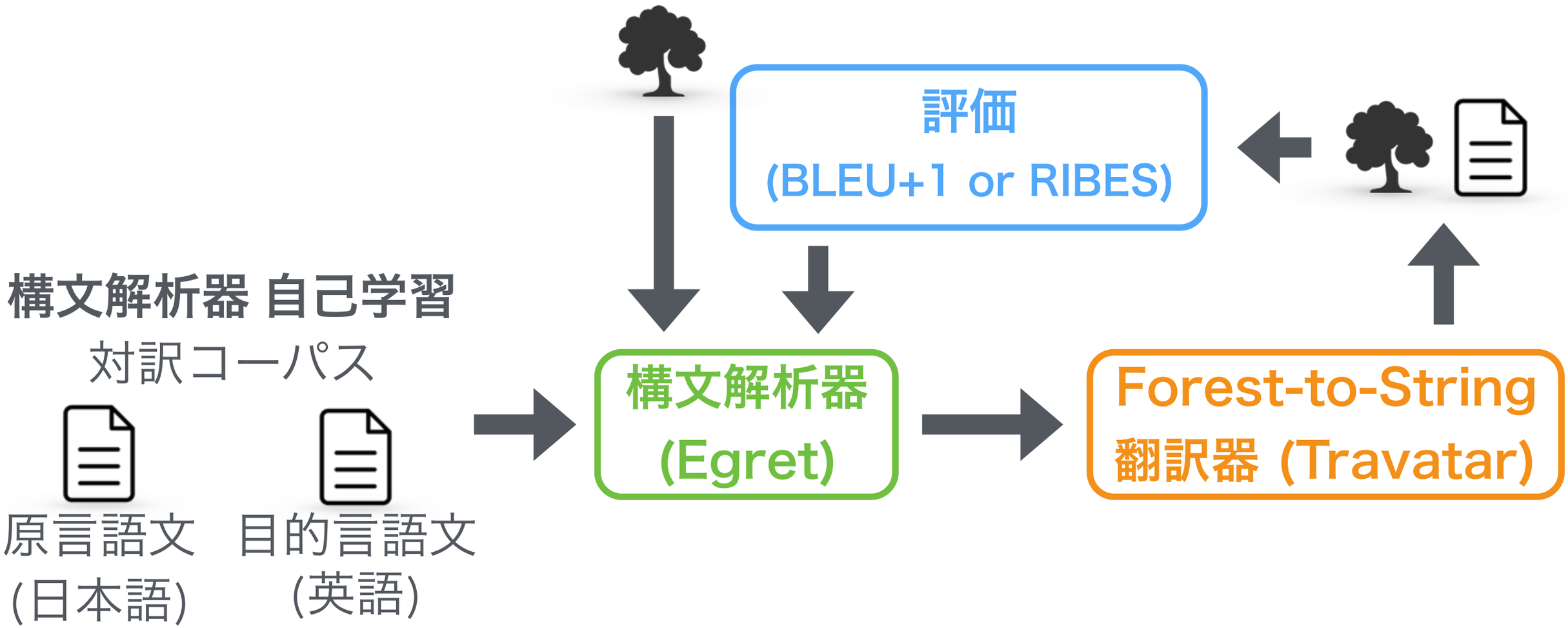
- どのような特徴を持った分野において,
  自己学習による効果が得られやすいか

- 単一分野を対象に自己学習を行った場合と,
  複数分野を対象とした場合の精度比較
  - 自己学習は複数分野を対象にした方が良い？

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 (34k)



使用した対訳コーパス

コーパス名	文数	翻訳器の学習に使用	自己学習に使用
青空文庫	108k	○	○
BTEC	465k	○	○
KFTT	440k	○	○
法律文書	260k	○	○
例辞郎	424k	○	○
田中コーパス	150k	○	○
TED	97k	○	○
英辞郎	1969k	○	×
WWWJDIC	394k	○	×
Wikipedia	403k	○	×

JDCに含まれる分野

分野	train文数	test文数
Yahoo!知恵袋	1579	491
白書	1158	340
Yahoo!ブログ	1788	491
書籍	2181	485
雑誌	2439	395
新聞	2446	471
日常会話例文	11700	1300
日本経済新聞	8747	979
レシピ	661	62
論文抄録	286	30
特許	1494	208

様々な分野が含まれている

実験設定 (精度評価時)

- Evalb: 構文解析精度測定ツール
  - F値を計測
- ブートストラップ・リサンプリング法により統計的有意差を検証

実験結果

20万文を用いた自己学習結果

	構文木の選択法	文の選択法	文長の分布の保持	F値
(a)	—	—	—	82.95
(b)	構文解析器 1-best [McClosky et al. 2006]	ランダム	なし	82.34
(c)	自動評価尺度 1-best (BLEU+1)	自動評価値上位	なし	81.13
(d)	自動評価尺度 1-best (BLEU+1)	自動評価値上位	あり	83.23
(e)	自動評価尺度 1-best (RIBES)	自動評価値上位	あり	* 83.26

*: $p < 0.05$

- 有意に高い構文解析精度の向上

各分野での精度測定結果

分野	ベースライン	提案手法
Yahoo!知恵袋	84.09	83.32
白書	71.89	** 74.41
Yahoo!ブログ	79.38	80.25
書籍	74.46	** 75.90
雑誌	78.66	* 80.01
新聞	79.04	* 80.18
日常会話例文	92.74	** 91.95
日本経済新聞	86.33	85.92
レシピ	84.02	82.53
論文抄録	83.99	82.12
特許	86.65	86.71
全体	82.95	* 83.26

特定の分野で
有意な向上

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

精度向上が期待できるドメインの特徴

- 4つの特徴について検討
  1. 既存モデルの解析精度
  2. 自己学習に使用した文と各分野の類似度
  3. 文の平均文長
  4. 自己学習後の未知bigramの減少率

- 各分野のF値上がり幅と特徴との相関を求める

既存モデルの解析精度

- 既存モデルでの解析精度が低いほど,
  自己学習による効果が大きい？

- 結果

特徴	F値上がり幅との相関係数	p値
既存モデル精度	-0.79	0.0040

- 相関がある

自己学習に使用した文と各分野の類似度

- 自己学習に使用した文と各分野の類似度が近い場合，自己学習効果が高い？

- 検証方法

- 自己学習に使用した文を基に言語モデルを作成
- 各テストセットとのPerplexityを求め，相関を見る

自己学習に使用した文と各分野の類似度

- 結果

特徴	F値上がり幅との相関係数	p値
自己学習文とのPerplexity	0.26	0.4360

- 相関はない

文の平均文長

- McCloskyらは自己学習が有効な文として，
  文長が20~50単語の文を挙げている
  - 本手法においても同様のことが言える？

- 結果

特徴	F値上がり幅との相関係数	p値
平均文長	-0.14	0.6875

- 相関はない

自己学習後の未知bigram減少率

- McCloskyらは自己学習が有効な文として,
  既存モデルで既知の単語が,
  未知のbigramで現れた文と報告している
  - 本手法においても同様のことが言える？

- 既存モデルのbigram, 自己学習後のbigramを求め,
  どの程度未知bigramが減少したか確認

自己学習後の未知bigram減少率

- 結果

特徴	F値上がり幅との相関係数	p値
未知bigram減少率	0.69	0.0181

- 相関がある
  - 未知bigramが減少すると精度が向上する

精度向上が期待できるドメインの特徴

	特徴	F値上がり幅との相関係数	p値
1.	既存モデル精度	-0.79	0.0040
2.	自己学習文とのPerplexity	0.26	0.4360
3.	平均文長	-0.14	0.6875
4.	未知bigram減少率	0.69	0.0181

- 既存モデル精度および未知bigram減少率との相関が見られた

単一分野に対して自己学習を行った場合との比較

単一分野に対しての自己学習

- 今回の実験では,
  - 複数の分野に対して自己学習を行い,
  - 複数の分野で精度計測
- 単一分野に対して自己学習を行った場合との差は？

ASPECを用いた比較

- ASPEC: 科学技術論文抜粋対訳コーパス
  - 100文について人手で正解構文木を作成
  - これをもとに精度評価
- ASPECに対して自己学習
  - JDC, ASPECのテストセットで計測

実験結果

	テストセット JDC	ASPEC
ベースライン	82.95	84.53
構文解析器 1-best	82.34	86.40
様々な分野で自己学習	83.26	86.36
ASPECについて自己学習	79.41	88.07

- ASPECで自己学習するとJDCでは精度低下

実験結果

	テストセット JDC	テストセット ASPEC
ベースライン	82.95	84.53
構文解析器 1-best	82.34	86.40
様々な分野で自己学習	83.26	86.36
ASPECについて自己学習	79.41	88.07

- ASPECで自己学習するとJDCでは精度低下
- ただし，ASPECについては最高精度が出せる

実験結果

	テストセット JDC	テストセット ASPEC
ベースライン	82.95	84.53
構文解析器 1-best	82.34	86.40
様々な分野で自己学習	83.26	86.36
ASPECについて自己学習	79.41	88.07

- ASPECで自己学習するとJDCでは精度低下
- ただし，ASPECについては最高精度が出せる
- 様々な分野で自己学習すると両テストセットで精度が向上

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
  - 4種類のドメインにおいて精度向上が見られた

- 精度向上が期待できるドメインの特徴について調査
  - 既存モデル精度が低く，未知bigramが減少した場合に
    精度が向上する可能性が高い

- 一つのドメインに絞って学習を行うと,
  最も解析精度が高くなると思われる

- 今後の課題
  - 精度が低下した原因の調査
  - 解析対象を絞った場合の文選択手法の検討
    - 未知bigramを減らすように学習する？
    - 構文解析器のモデルをより考慮した学習方法を検討？

モデルの公開

- 今回自己学習を行ったモデルを公開します
  - http://www.otofu.org
- 何かあれば森下までご連絡ください

morishita.makoto.mb1@is.naist.jp

END