

# 構文木の生成確率を考慮した 対訳コーパスを用いた構文解析器の自己学習



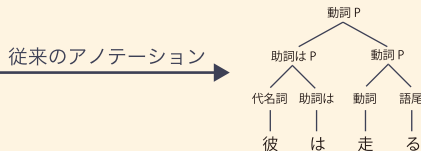
森下 睦, Graham Neubig, 吉野 幸一郎, 中村 哲  
[morishita.makoto.mb1, neubig, koichiro, s-nakamura]@is.naist.jp  
奈良先端科学技術大学院大学 (NAIST)

## 1 研究背景・目的

- 構文解析器の学習には大量の構文木が必要
- 現在構文木が存在する分野には偏りがあり、分野外の文に対しては解析精度が低い
- また、新たに構文木を手でアノテーションするには、大きなコストがかかり現実的ではない

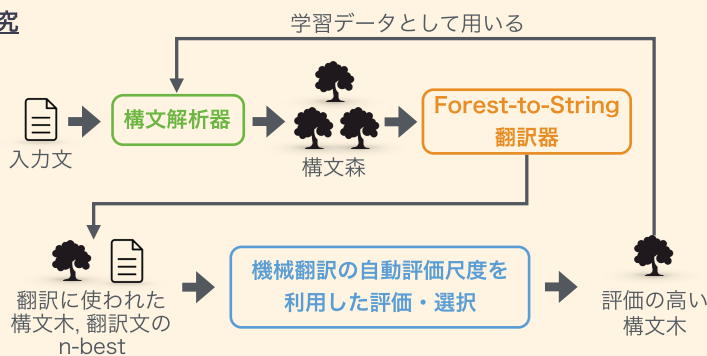
### 目的

低コストで構文解析器の精度を多分野で向上させる



## 2 対訳コーパスを用いた構文解析器の自己学習

### 先行研究



対訳コーパスのみを用いて、構文解析器の精度向上が見込める

Morishita, M., Akabe, K., Hatakoshi, Y., Neubig, G., Yoshino, K. and Nakamura, S.: Parser Self-Training for Syntax-Based Machine Translation, Proc. IWSLT, pp. 232–239 (2015).

## 3 Tree Entropy

### 先行研究

- 構文解析器の能動学習の際に有効な文選択法
- コーパスの中から、Tree Entropy が最も高い文を、次のアノテーション対象とする

$$H(V) = - \sum_{v \in \mathcal{V}} p(v) \log_2(p(v))$$

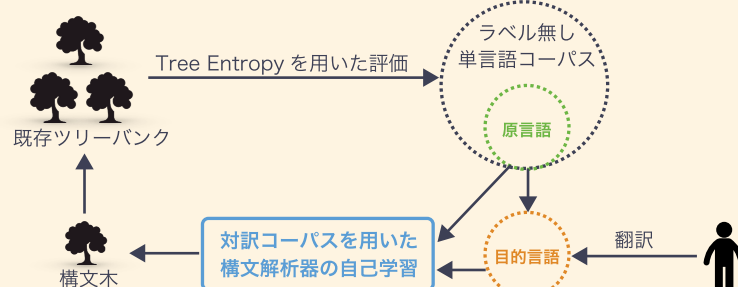
$v$ : 構文木  $\mathcal{V}$ : 生成可能な構文木の集合

- Tree Entropy は文が長くなると  $\mathcal{V}$  の数が大きくなり、値が大きくなるため正規化する

$$H_n(V) = \frac{H(V)}{\log_2(|\mathcal{V}|)}$$

Hwa, R.: Sample Selection for Statistical Parsing, Computational Linguistics, Vol. 30, No. 3, pp. 253–276 (2004).

## 4 提案手法



Tree Entropy を用いて、次にアノテーションする文を選択  
→ 選択された文のみを翻訳、自己学習に使用する

学習に必要な翻訳文を最小限にして、  
対訳コーパスを用いた構文解析器の自己学習が可能

## 5 実験設定

### 比較手法

	構文木選択法	文選択法
MT-Eval	機械翻訳の自動評価尺度	機械翻訳の自動評価尺度
MT-Entropy	機械翻訳の自動評価尺度	Tree Entropy
Parser-Entropy	構文解析器 1-best	Tree Entropy

### 実験設定

原言語 (構文解析言語): 英語

目的言語: 中国語

Forest-to-String 翻訳器:  
Travatar

構文解析器: Egret

既存ツリーバンク: WSJ

使用コーパス: UM-Corpus

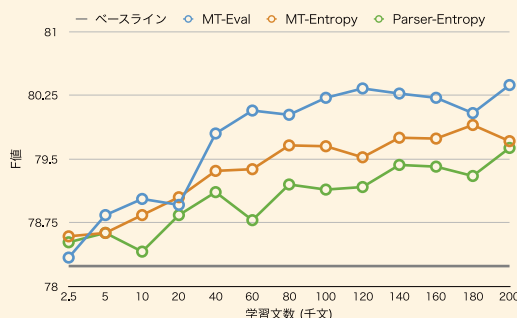
翻訳器学習用: 約 145 万文

自己学習用: 約 72 万文

構文解析器テストセット:

Google Web Treebank  
(Answers, Email, Reviews,  
Newsgroup, Weblog)

## 6 実験結果



MT-Eval は最も高い精度が得られているものの、コーパス全体を事前に翻訳する必要がある一方 MT-Entropy は Tree Entropy を基に選択された文のみを翻訳するため、必要な翻訳文が少ない

Tree Entropy を用いることで、学習に必要な翻訳文を  
最小限に抑えつつ、構文解析器の精度向上が可能になった