日本語言い換えデータベースの改善と評価

水上雅博，Graham Neubig，Sakriani Sakri，戸田智基，中村哲（NAIST）

目的

目的：**フリーで汎用的な高品質の言い換えデータベースを自動構築する。**

関連研究

言い換えデータに関して
- 敬語表現 [大竹+, 2001]
- 文末表現 [中川+, 2005]
- 意味関係表現 [Saeger+, 2009]
- 動詞含意関係 [橋本+, 2011]

…他多数

➡ ドメインや表現を限定しない

言い換え抽出手法に関して
- 対訳コーパスとピボット [Bannard+, 2001]
- 構造変換 [Edmonds+, 1999]

➡ 日本語に適した頑健な手法

従来法

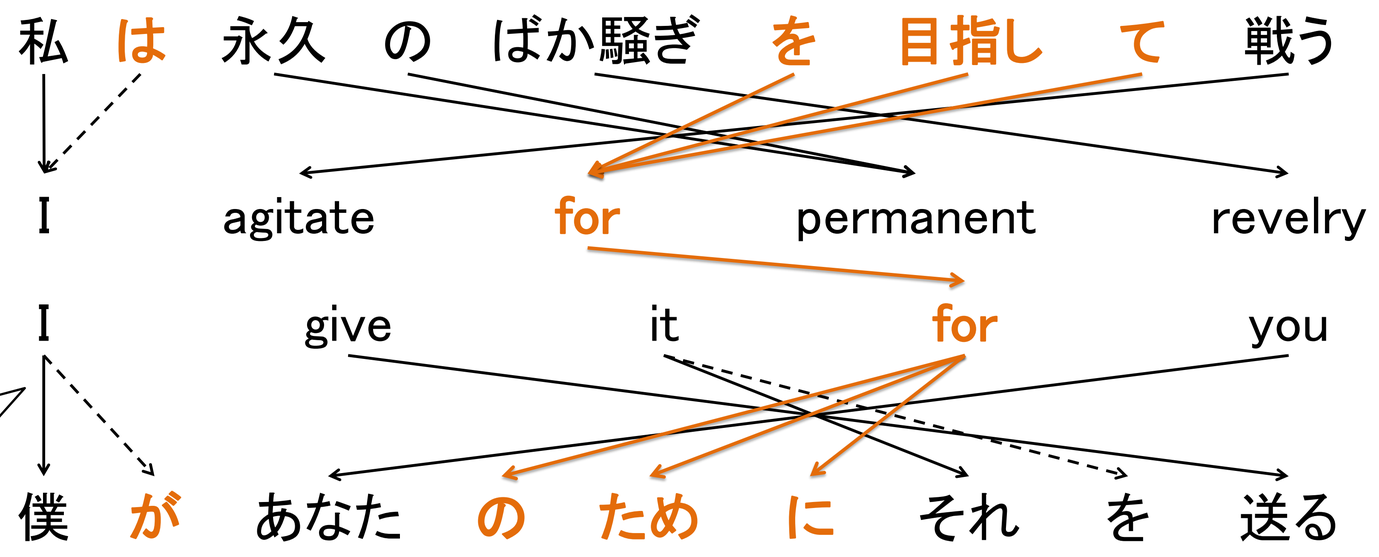
日英対訳コーパスとピボットによる言い換えデータ抽出 [水上+, 2014]

➡ 旅行会話コーパスを対象に、ピボット法で言い換えを抽出

従来法の問題

- 旅行会話に**ドメインが限定**
- 助詞等のヌルアライメント

- アライメントが交差しており、**誤アライメント**が生じる
- 助詞（は、が…etc）に対応する英単語がなく、**アライメントが不足**

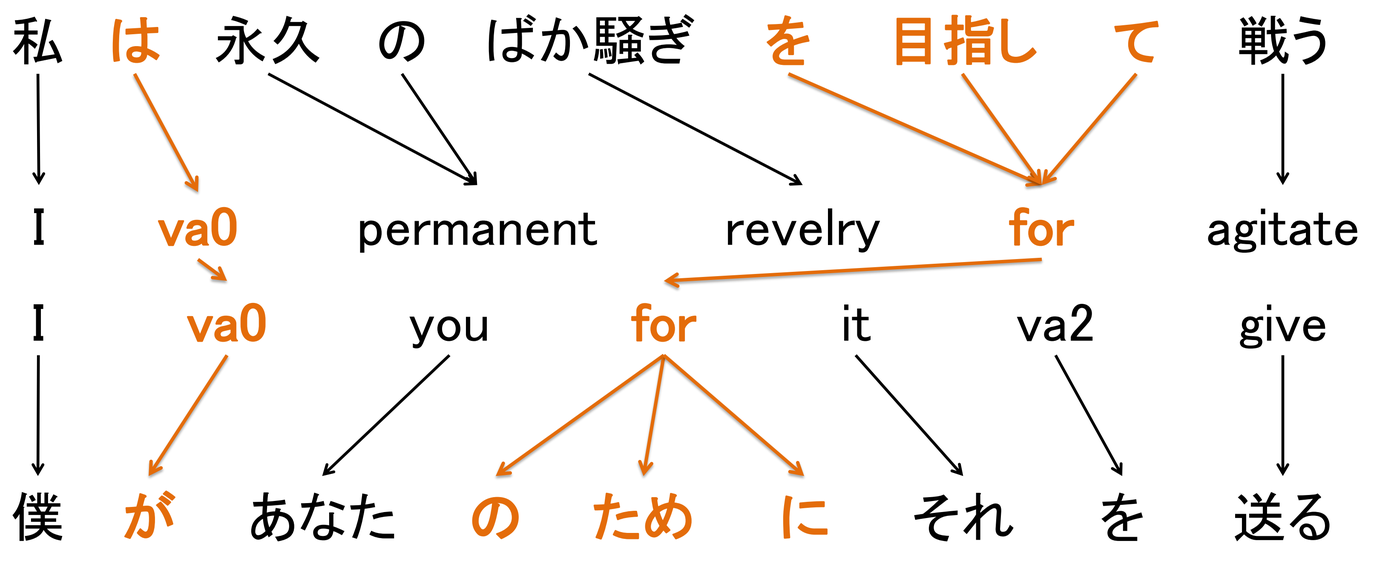


提案法

1. フリーの日英対訳コーパスを混合一般日英対訳コーパスを構築

Corpus	
Tanaka	150k
AOZORA	108k
Common Crawl	821k
Kyoto Wiki	440k
WWWJDIC	373k
Total	1.9M

2. **事前並べ替え(HF; [磯崎+, 2010])**を用いてアライメントを改善



評価実験と考察

設定	
アライメント	GIZA++ [Och+, 2003]
トークナイズ(en)	Stanford Parser [Socher+, 2013]
トークナイズ(ja)	Kytea [Neubig+, 2013]
フレーズ最大長	7 words
フレーズ数	67.1M
原文	24文
テスト文	85文
評価者	日本人20代男女3名

提案法	正解率[%]
意味	84.7
文法	55.3
両方	**45.8**
従来法	正解率[%]
意味	**58.9**
[Bannard+,2005]	正解率[%]
意味	64.5
両方	**48.9**

評価の一致率 κ=0.60 [Fleiss+, 1971]
“そこそこ”の一致性 [Lindis+, 1977]

考察と課題

- （特に助詞による）**文法的な不一致**の発生
- 言語モデルを用いて**文法的な等価性**を評価、精度向上を期待
- アプリケーションへ利用