文献紹介

長岡技術科学大学 電気系 自然言語処理研究室
修士１年　松本　宏

文献

Title: Enriching SMT Training Data via Paraphrasing

Authors: Wei He, Shiqi Zhao, Haifeng Wang , Ting Liu

Conf. : IJCNLP. 2011

Page: 803-810

概要

- 統計的機械翻訳(SMT)のコーパスの換言を行う
- 主な目的としては
  - 統計機械翻訳の精度向上
  - Coverage率を上げる

背景

背景 #1

- SMTにおいてコーパスサイズは重要
- 様々な構築はたくさん行われてきた
  - マニュアル
  - Webからのコーパス構築
  - など
- コーパスの換言を行い増築する手法も研究されている

背景 #2

- 主な換言手法は以下
  - 入力文の換言
  - 学習コーパスの換言
- 既存の換言手法はSMTの向上を示している

背景 #3

- しかし、
- コーパス換言手法において既存手法の問題点
  - 処理の非効率
  - 文脈の非考慮

処理の非効率

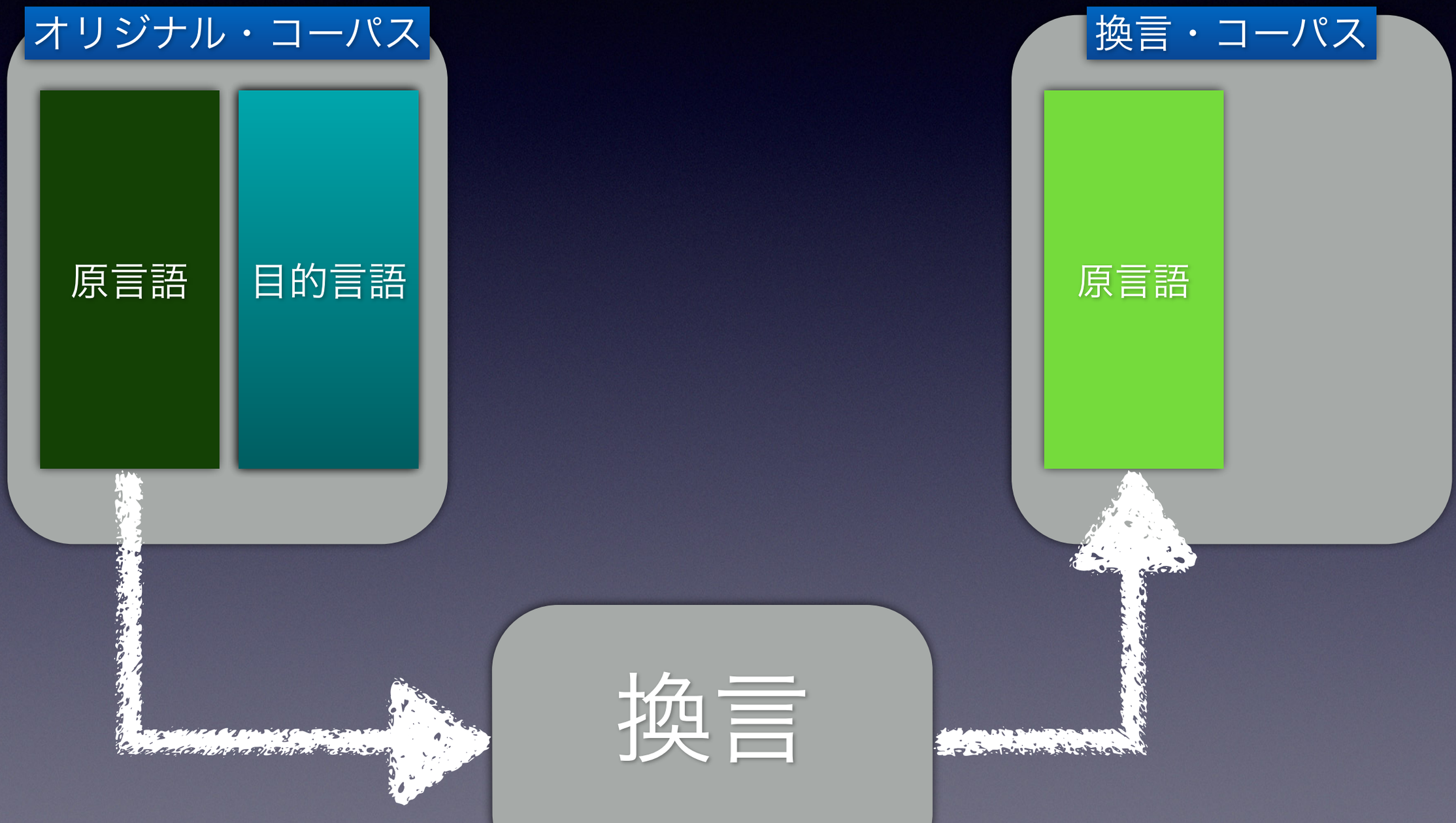
- 既存手法は：
  - 2段ディコーディング・処理
    1. 換言処理
    2. 翻訳処理

文脈の非考慮

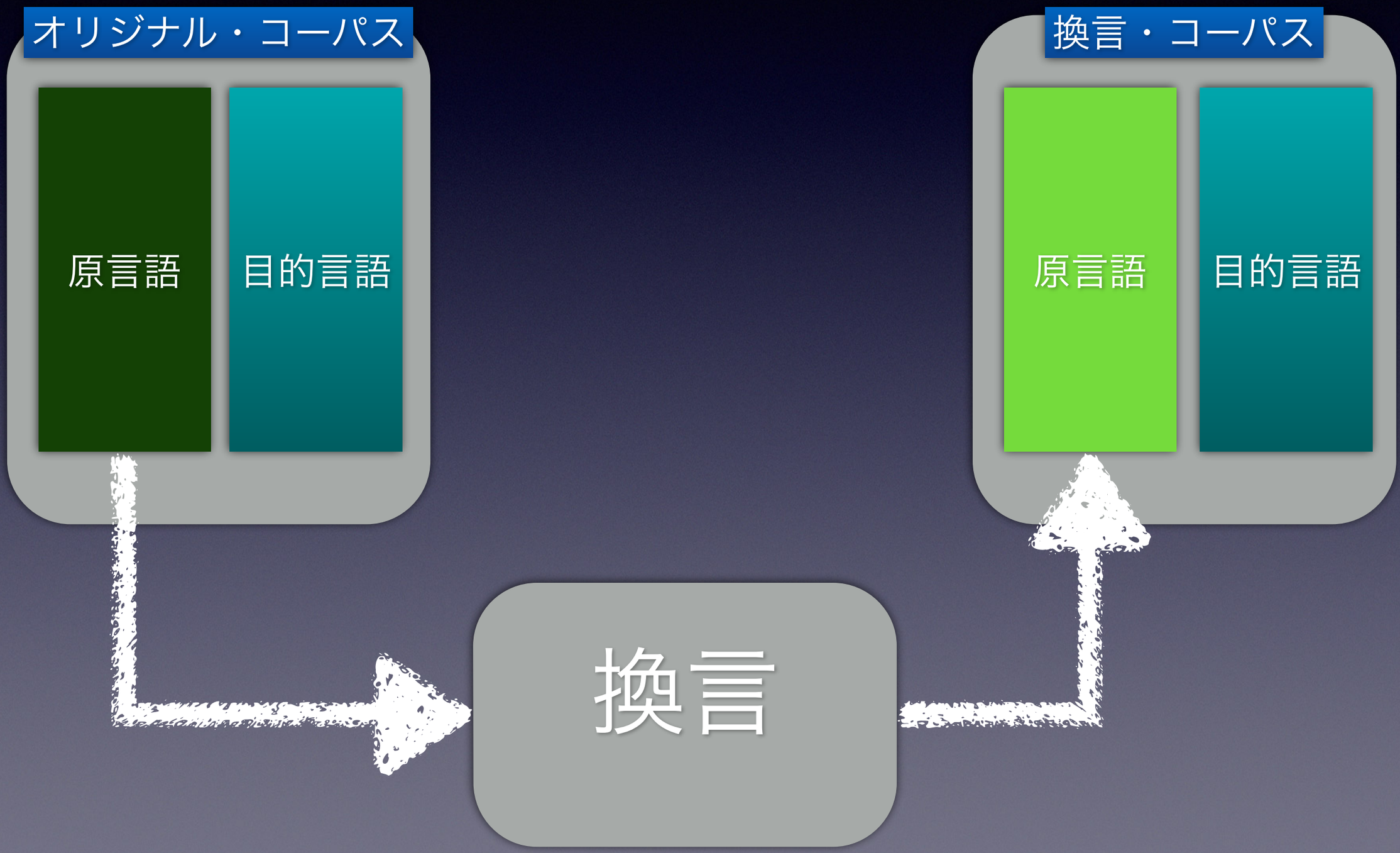
- bank(川岸、銀行) と shore(岸)は換言可能
- しかし、文脈に依存する
  - bankとshoreは川に関する文脈のみ換言可能

手法

手法 #1



手法 #1

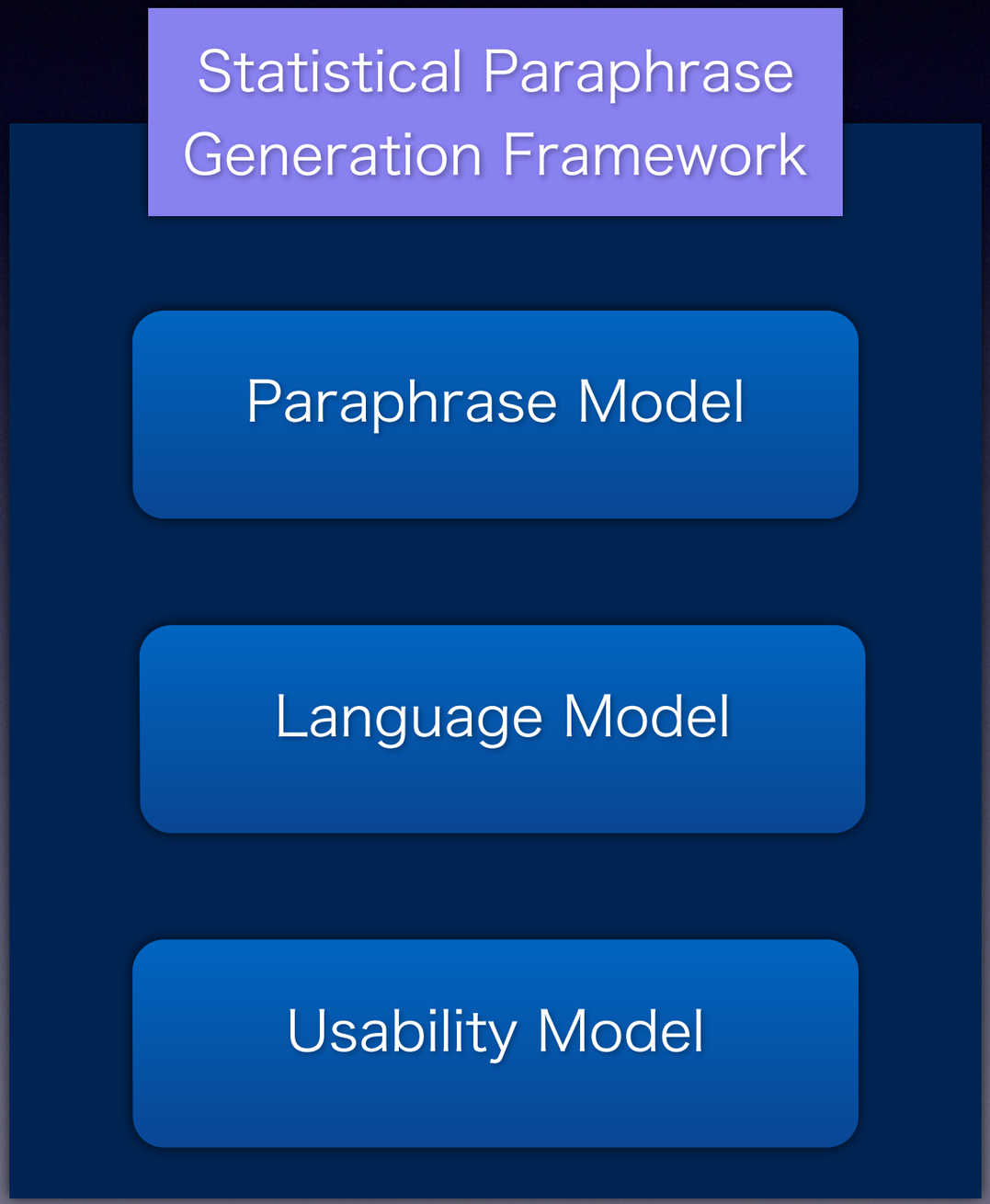


手法 #2

換言

- 統計換言生成フレームワークの利用
  - 統計的手法
  - 言語非依存

統計換言生成フレームワーク



Paraphrase Model

- 換言モデルは妥当性の制御

- 入力文SをI個に区切った単位s
- (s, t)を換言対とした時の尤度を ϕ (s, t)

- S,Tの換言スコアは以下のように表現できる

$$p_{pm}(\bar{s}_1^I, \bar{t}_1^I) = \prod_{i=1}^{I} \varphi_{pm}(\bar{s}_1, \bar{t}_1)^{\lambda_{pm}}$$

Language Model

- 言語モデルは流暢性の制御

- 4-gramの言語モデル
  - 流暢性の確保
  - 換言の曖昧性を解消
- λは重み

$$p_{lm}(T) = \prod_{j=1}^{J} p(t_j|t_{j-3}t_{j-2}t_{j-1})^{\lambda_{lm}}$$

Usability Model

- ユーザビリティ・モデルはより適した換言の制御
  - 新しいngramを選ぶことと定義

- J単語からなる換言文Tがあるとき、
- Novel(TM, t, n, j)は翻訳モデルにとって新規ngramかの判断をする
  - 新規なら1, そうでないなら0

$$p_{nm}(t) = exp(\sum_{j=1}^{J}\sum_{n=1}^{N} Novel(TM, t, n, j))^{\lambda_{nm}}$$

まとめると

換言生成フレームワークは以下のようになる

$$
\begin{aligned}
I \\
p(T|S) &= \lambda_{pm} \sum_{i=1} log \varphi_{pm}(\overline{s}_i, \overline{t}_i) \\
&+ \lambda_{lm} \sum_{j=1}^{J} \log p(t_j | t_{j-3} t_{j-2} t_{j-1}) \\
&+ \lambda_{nm} \sum_{j=1}^{J} \sum_{n=1}^{N} Novel(TM, T, n, j)
\end{aligned}
$$

実験

Experimental Setup

- 対象は英中翻訳
- データはSinorama and FBIS corpora
  - 319,694行
- 少資源言語を模擬するために
  - ランダムに29,000行抽出
- 評価セット: English-Chinese NIST MT 2008

Experiment Setup

３種類のコーパスを用意する

1. オリジナル・コーパス (Ori.)

2. 最良換言コーパス (1-Para)

3. 複良換言コーパス (M-Para)

Results #1

	29k	Full
Ori.	324k	3603k
+1-PARA	507k (+56%)	4514k (+25%)
+1-PARA+M-PARA	878k (+171%)	8359k (+132%)

Results #2

	Model	BLEU-4	TER
29k	Baseline	17.91	66.83
	Ori. +1-PARA	19.26	65.98
	Ori. +1-PARA +M-PA	19.57	65.88
Full	Baseline	25.46	62.36
	Ori. +1-PARA	26.52	61.36
	Ori. +1-PARA +M-PA	26.33	61.47