Improving Statistical Machine Translation with Word Class Models

Joern Wuebker, Stephan Peitz, Felix Rietig and Hermann Ney

紹介文献

Title: Improving Statistical Machine Translation with Word Class Models

Author: Wuebker, Joern and Peitz, Stephan and Rietig, Felix and Ney, Hermann

Booktitle: EMNLP

pages: 1377—1381

Year: 2013

概要

- 統計機械翻訳にクラス分類の取り入れ
  - スパースネス問題解消のために語彙を減らす

- 実験:
  - 仏独翻訳

データスパースネスはSMTの問題

問題のモデル推定のために語彙サイズの削減

- 手法として分類(Word Class)の利用

目的:

学習されて得るWord Classと本来の単語が持つ単語アイデンティティ(Word Identity)の比較

手法

今回、SMTで利用される以下のモデル

- LM: 言語モデル
- TM: 翻訳モデル
- HRM:階層的並び替えモデル

- を、WordClassモデルに差し替え、それぞれの変化を調べる

デコーダ

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, s_1^K, f_1^J) \right\}$$

1つのモデルを上記の式で表現できる

- f: 原言語
- e: 目的言語
- s: 隠れ・アライメント
- 先のモデルを追加モデルとして追加可能

実験

- French → German
  - Data from WMT 2012 shared task
    - Broadcast news and broadcast conversations

		French	German
train	Sentences	1.9M	
	Running Words	57M	50M

使われるクラス数:

- 100, 200, 500 クラス

結果:仏独翻訳

	dev		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
-TM +wcTM	21.2	64.2	24.7	59.5
-LM +wcLM	22.2	62.9	25.9	58.9
-HRM +wcHRM	24.6	61.9	27.5	58.1
phrase-based	24.6	61.8	27.8	57.6
+ wcTM	24.7	61.4	28.1	57.1
+ wcLM	24.9	61.2	28.4	57.1
+ wcHRM	25.4‡	60.9‡	28.9‡	56.9‡
+ wcLM[7]	25.5‡	60.7‡	29.2‡	56.6‡
+ $wcModels_{200}$	25.5‡	60.8‡	29.3‡	56.4‡
+ $wcModels_{500}$	25.2†	60.8‡	29.0‡	56.6‡