文献紹介ゼミ

電気M1　宮西　由貴

文献情報

- タイトル
  Improving Word Alignment using Word Similarity
- 著者
  - Theerawat Songot
  - David Chiang

概要

- 単語アライメントの手法の提案
  - 単語類似度を使用した手法
  - 単一言語のリソースを使用
  - ニューラルネットワークで学習
- 提案手法の評価

単語アライメントについて

- 単語アライメント
  - 2言語間の単語の照応関係
  - ソース文:$\boldsymbol{e} = e_1, e_2, \ldots e_l$
  - ターゲット文:$\boldsymbol{f} = f_1, f_2, \ldots, f_m$
  - eとfのアライメント: $\mathbf{a} = a_1, a_2, \ldots, a_m$(null / e_{a_i})

- 機械翻訳と単語アライメント
  - 単語の機械翻訳を行っているのと同様

単語類似度のモデル

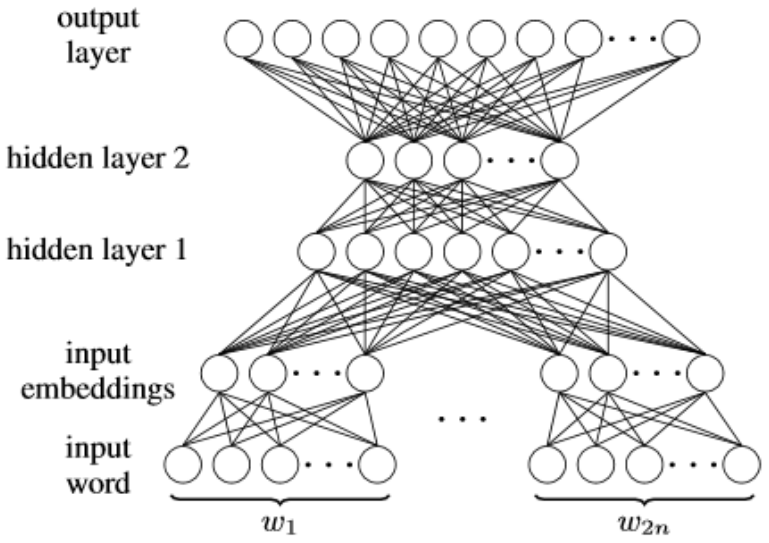
- モデルの定式化
  - ある単語wとw'の類似度:p(w'|w)
  - w'∈V　(ボキャブラリ)

$$p(w' \mid w) = \sum_{c} p(c \mid w)\, p(w' \mid c)$$

  - c:単語wの周辺2nのトークン(右側n，左側n)
  - p(w'|c)は過学習する可能性あり

単語類似度のモデル

- p(w’|c)へのフィードフォワードNNの使用
  - 入力:
    文脈c(one-hot)
    タグ付き(<s>,</s>,<unk>)
  - 出力:
    それぞれのw’∈V
- 今回の設定
  - n(cの右側or左側の長さ):5
  - 隠れレイヤー×2



単語類似度の学習方法

- 学習データ(単一言語データから取得)
  - ターゲット単語$w \in V$
  - 文脈c
- p(c|w)とp(w'|c)を別々に学習
  - p(c|w)…最尤推定
  - p(w'|c)…noise-contrastive estimation
    (シンプル&高速)

単語類似度のモデルを変形

- 今までの式と学習法からの改善点
  - False positive(偽陽性)は起こらないようにしたい
  - 式を変形:

$$p(w' \mid w) = \frac{1}{Z(w)} \exp \sum_c p(c \mid w) \log p(w' \mid c)$$

  - Z(w)は標準化定数

$p(w' \mid \text{country})$	
country	0.8363
region	0.0558
nation	0.0522
world	0.0282
city	0.0273

$p(w' \mid \text{region})$	
region	0.8338
area	0.0760
country	0.0524
province	0.0195
city	0.0181

$p(w' \mid \text{area})$	
area	0.8551
region	0.0524
zone	0.0338
city	0.0326
areas	0.0258

上位k件を
抽出
(k=5と設定)

単語アライメントのモデル

- スタンダードIBMモデルを拡張
  - いろんな関係のモデルに楽に応用可

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \propto \prod_{i=1}^{m} p(f_i \mid e_{a_i}) = \prod_{i=1}^{m} t(f_i \mid e_{a_i})$$

  - 単語類似度を用いる

$$p(f \mid e) = \sum_{e', f'} p(e' \mid e)\, t(f' \mid e')\, p(f \mid f')$$

$$p(f \mid f') = \frac{p(f' \mid f)\, p(f)}{p(f')}$$

  - 今回はp(f|f')≈p(f|f')とみなす

実験

- 中国語–英語，アラビア語–英語で実験
  - 中→英：9.5M+12.3Mの学習コーパス
    39.6k+50.9kの実験データ
  - ア→英：4.2M+5.4Mの学習コーパス
    10.7k+15.1kの実験データ

実験結果

Model	Precision	Recall	F1	BLEU		METEOR	
				Test 1	Test 2	Test 1	Test 2
Chinese-English							
Baseline	65.2	76.9	70.6	29.4	26.7	29.7	28.5
Our model	71.4	79.7	75.3	29.9	27.0	30.0	28.8
Baseline (resource-limited)	56.1	68.1	61.5	23.6	20.3	26.0	24.4
Our model (resource-limited)	66.5	74.4	70.2	24.7	21.6	26.8	25.6
Arabic-English							
Baseline	56.1	79.0	65.6	37.7	36.2	31.1	30.9
Our model	60.0	82.4	69.5	38.2	36.8	31.6	31.4
Baseline (resource-limited)	56.7	76.1	65.0	34.1	33.0	27.9	27.7
Our model (resource-limited)	59.4	80.7	68.4	35.0	33.8	28.7	28.6