単語分散表現のアライメントに基づく文間類似度を用いたテキスト平易化のための単言語パラレルコーパスの構築

首都大学東京　梶原智之　小町守　kajiwara@jnlp.org

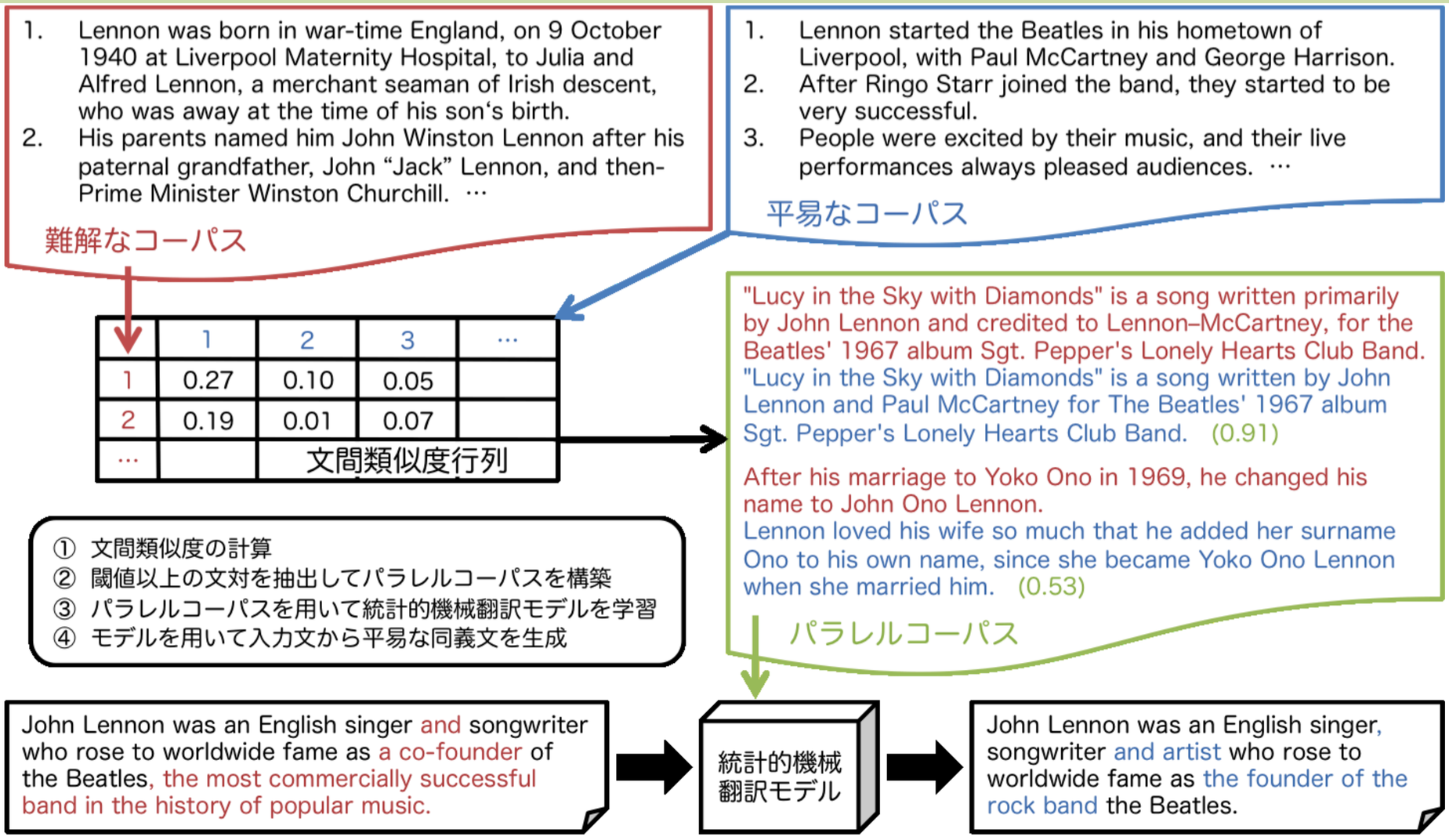
テキスト平易化

難解なテキストの意味を保持したまま平易に書き換える

文圧縮 + （平易な）言い換え

従来のコーパス

- Zhu et al. (2010)
  文をTF-IDFベクトルで表現
  cos類似度の閾値で切る
- Coster and Kauchak (2011)
  Zhu et al. (2010) の拡張
  文の出現順序も考慮する
- Hwang et al. (2015)
  外部知識（Wiktionary）を用いて単語間類似度を考慮



本研究：単語分散表現のアライメントに基づく文間類似度を用いたコーパス構築

単語分散表現を用いることでラベル付きデータや辞書などの外部知識に頼らずに、異なる単語間の類似度を考慮した文間類似度 [1] を計算する

1. Average Alignment （多対多の単語アライメント）
2. Maximum Alignment （多対一の単語アライメント）
3. Hungarian Alignment （一対一の単語アライメント）

[1] Song and Roth (2015) Unsupervised Sparse Vecor Densification for Short Text Similarity

1. Average Alignment

全ての単語対の単語間類似度Φを平均

$$S_{ave}(x,y)=\frac{1}{|x||y|}\sum_{i=1}^{|x|}\sum_{j=1}^{|y|}\phi(x_i,y_j)$$

2. Maximum Alignment

上の手法はnoisy → 各単語x_iに対して最も類似度が高い単語y_jのみを考慮

$$S_{\max}(x,y)=\frac{1}{|x|}\sum_{i=1}^{|x|}\max_j\phi(x_i,y_j)$$

3. Hungarian Alignment

単語をノード、単語間類似度をエッジとする重み付き完全２部グラフを考え、このグラフの最大マッチングを求める（この問題はHungarian法で解ける）

$$S_{hun}(x,y)=\frac{1}{\min(|x|,|y|)}\sum_{i=1}^{\min(|x|,|y|)}\phi(x_i,h(x_i))$$

内的評価：文間類似度を用いたパラレルとノンパラレルの２値分類

	両方向含意 vs. その他		片方向含意 vs. その他	
	MaxF1	AUC	MaxF1	AUC
Zhu et al. (2010)	0.550	0.509	0.431	0.391
Coster and Kauchak (2011)	0.564	0.495	0.415	0.387
Hwang et al. (2015)	0.712	0.694	0.607	0.529
Additive Embeddings	0.691	0.695	0.518	0.487
1. Average Alignment	0.419	0.312	0.391	0.297
2. Maximum Alignment	0.717	0.730	0.638	0.618
3. Hungarian Alignment	0.524	0.414	0.354	0.275

※ Additive Embeddings: 単語アライメントを使用しない比較手法
単語分散表現を足した文ベクトルのCOS類似度

① English Wikipedia と Simple English Wikipedia からタイトルが一致する 126,725 文書対を収集

② Maximum Alignment の文間類似度で 492,993 文対を収集
単語アライメントの閾値：単語間類似度 > 0.49
文アライメントの閾値：　文間類似度 > 0.53

外的評価：統計的機械翻訳の枠組みでのテキスト平易化

	文対数	平均文長		BLEU	
		難解	平易	両含意	片含意
Baseline (None)				42.1	22.3
Zhu et al. (2010)	107,516	21.2	17.4	42.0	22.1
Coster and Kauchak (2011)	136,862	23.6	21.1	44.3	23.8
Hwang et al. (2015)	284,238	26.0	19.8	43.9	23.1
Ours	492,493	25.3	17.9	47.5	26.3

※ English Wikipedia 全体の平均文長：25.1
※ Simple English Wikipedia 全体の平均文長：16.9

梶原智之, 小町守. 単語分散表現のアライメントに基づく文間類似度を用いたテキスト平易化のための単言語パラレルコーパスの構築. NLP若手の会 第11回シンポジウム, P31, August 29th, 2016.