文献紹介ゼミ

林 秀治

紹介する文献

- Better evaluation for grammatical error correction
- Daniel Dahlmeier, Hwee Tou Ng
- NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
- pp.568-572

概要

- 文法誤り訂正を評価する新たな手法の提案
- MaxMatch(M2)という、原文間とのフレーズレベルの編集を効率的に計算する手法
- Helping Our Own(HOO) shared task dataでテストを行いよりよい結果が得られている

背景

- 誤り訂正の評価は主に、システムによる訂正と人手による訂正間のF値で行われる
- しかし、テストセットの訂正の曖昧性のため評価は難しい
  - 1つの文字列を別の文字列に変える訂正でも必ずしもユニークでない
  - 訂正箇所が曖昧性を持つより長いフレーズの場合がある

曖昧性による影響

- 以下のセットのとき

  誤：Our baseline system feeds word into PB-SMT pipeline.

  正：Our baseline system feeds **a** word into PB-SMT pipeline.
- HOOでは(ε→a)のようにaの挿入である
- しかし、実際は(word→{a word, words})の2パターン考えられる
- そこでMaxMatch(M2)を提案し、この問題を解決する

手法

- 誤り訂正システムが作ったSource sentences S={s1,...,sn}とhypotheses H={h1,...,hn}のセットを考える。
- 同じ文に対してのGold standard G={g1,...gn}
- 各アノテーションのgi={gi1,...gir}は訂正のセット
- 編集は(a,b,C)からなる
  - a:始点, b:終点, C:訂正(のセット)

手法

- 評価は以下の2stepで行う
  - 各source-hypothesis pair(si,hi)に対して system edits eiを抽出
  - Gold standard Gと比較して評価

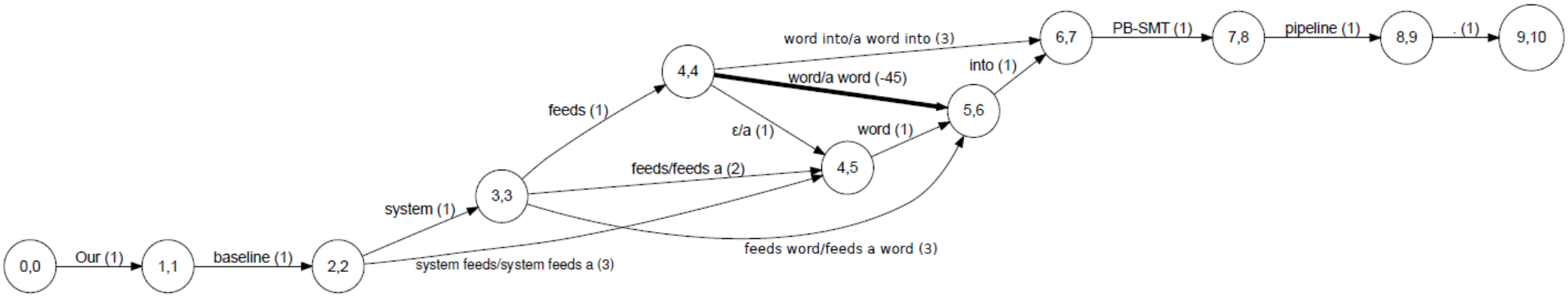
Edit lattice

- Levenshtein距離を使う

		Our	baseline	system	feeds	a	word	into	PB-SMT	pipeline	.
	0	1	2	3	4	5	6	7	8	9	10
Our	1	0	1	2	3	4	5	6	7	8	9
baseline	2	1	0	1	2	3	4	5	6	7	8
system	3	2	1	0	1	2	3	4	5	6	7
feeds	4	3	2	1	0	1	2	3	4	5	6
word	5	4	3	2	1	1	1	2	3	4	5
into	6	5	4	3	2	2	2	1	2	3	4
PB-SMT	7	6	5	4	3	3	3	2	1	2	3
pipeline	8	7	6	5	4	4	4	3	2	1	2
.	9	8	7	6	5	5	5	4	3	2	1

Edit lattice

- Levenshtein距離を基にlatticeを作成
- 長いフレーズに対応するために隣接したものを接続する



Evaluating edits

- System edits{e1,...,en}とgold edits{g1,...,gn}のprecision, recall, F1で評価する

$$
\begin{aligned}
P &= \frac{\sum_{i=1}^{n} |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|} \\
R &= \frac{\sum_{i=1}^{n} |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^{n} |\mathbf{g}_i|} \\
F_1 &= 2 \times \frac{P \times R}{P + R},
\end{aligned}
$$

評価実験

- HOO shared taskのデータを使用
  - NLPの論文と人手によるgold-standardからなる
- HOOの他の結果と比較

Team	HOO scorer			M^2 scorer		
	P	R	F_1	P	R	F_1
JU (0)	10.39	3.78	5.54	12.30	4.45	6.53
LI (8)	20.86	3.22	5.57	21.12	3.22	5.58
NU (0)	29.10	7.38	11.77	31.09	7.85	12.54
UI (1)	50.72	13.34	21.12	54.61	14.57	23.00
UT (1)	5.01	4.07	4.49	5.72	4.45	5.01

出力の例

- HOOでは人手と一致しなかったものがいくつか一致した

M^2 scorer	. . . should basic translational unit be (word → a word) . . .
HOO scorer	. . . should basic translational unit be *(ϵ → a) word . . .
M^2 scorer	. . . development set similar (with → to) (ϵ → the) test set . . .
HOO scorer	. . . development set similar *(with → to the) test set . . .
M^2 scorer	(ϵ → The) *(Xinhua portion of → xinhua portion of) the English Gigaword3 . . .
HOO scorer	*(Xinhua → The xinhua) portion of the English Gigaword3 . . .