P4-25 文法誤り訂正の自動評価のための原文・参照文・訂正文間の n-gram F-score

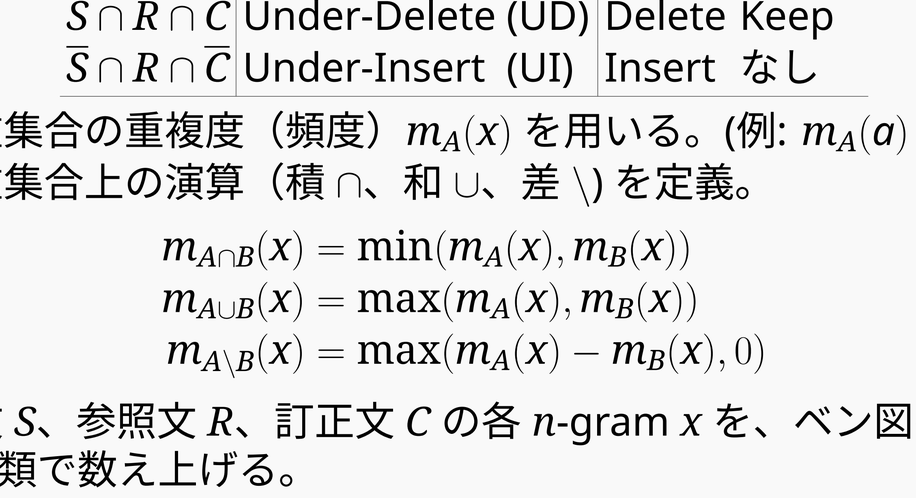
古山翔太（東工大/産総研, 岡崎研 D2） 永田亮 (甲南大/産総研) 高村大也（産総研） 岡崎直観（東工大/産総研）

概要

- 文法誤り訂正とは？文書中の誤りを訂正するタスク。
  - 例: *"What is you?"* → *"Who are you?"*
- システムの出力への良い自動評価手法がほしい。
  - システムの質を人手で評価するのは避けたい。
- M^2 とその派生が広く使われている。
  - 文間のフレーズアライメントを取り、訂正の個数を数え、F-score を計算。
- M^2 系の問題点は？
  - アライメントの計算に時間がかかる。また、アノテーションされたデータや、その他のツールが必要となる。
- 提案手法 GREEN
  - 文を n-gram の多重集合とみなし、文間の編集を集合操作で抽出する。
    - アライメントの計算が不要。
  - 原文から参照文の編集、原文から訂正文の編集を比較し、F-score を計算。
  - 既存の手法より人手評価と高い相関を示した。

手法

- 提案手法 GREEN は、文を最大長 N までの n-gram の多重集合として扱う。(例: 文 A ="*a a b*" は、$N=2$ で、$\{a, a, b, a\ a, a\ b\}$。)
- 原文 S, 参照文 R, 訂正文 C を多重集合で表し、ベン図を用いて文中の n-gram に対して 7 分類を行う。



原文 S:	*What*	*is*	*you*	*?*
参照文 R:	*Who*	*are*	*you*	*?*
訂正文 C:	*Who*	*is*	*you*	*!*

領域	名前		$S \to R$	$S \to C$
$S \cap \overline{R} \cap \overline{C}$	True Delete	(TD)	Delete	Delete
$\overline{S} \cap R \cap C$	True Insert	(TI)	Insert	Insert
$S \cap R \cap C$	True Keep	(TK)	Keep	Keep
$S \cap R \cap \overline{C}$	Over-Delete	(OD)	Keep	Delete
$\overline{S} \cap \overline{R} \cap C$	Over-Insert	(OI)	なし	Insert
$S \cap \overline{R} \cap C$	Under-Delete	(UD)	Delete	Keep
$\overline{S} \cap R \cap \overline{C}$	Under-Insert	(UI)	Insert	なし

- 多重集合の重複度（頻度）$m_A(x)$ を用いる。(例: $m_A(a)=2$)
- 多重集合上の演算（積 $\cap$、和 $\cup$、差 $\backslash$) を定義。

$$m_{A\cap B}(x) = \min(m_A(x), m_B(x))$$
$$m_{A\cup B}(x) = \max(m_A(x), m_B(x))$$
$$m_{A\backslash B}(x) = \max(m_A(x) - m_B(x), 0)$$

- 原文 S、参照文 R、訂正文 C の各 n-gram x を、ベン図を用いた 7 分類で数え上げる。

$$\begin{aligned}
\mathrm{TD}_{S,R,C}(x) &= m_{S\cap\overline{R}\cap\overline{C}}(x) = m_{S\backslash(R\cup C)}(x) \\
&= \max\{m_S(x) - \max(m_R(x), m_C(x)), 0\} \\
\mathrm{TI}_{S,R,C}(x) &= m_{\overline{S}\cap R\cap C}(x) = m_{(R\cap C)\backslash S}(x) \\
&= \max\{\min(m_R(x), m_C(x)) - m_S(x), 0\} \\
\mathrm{TK}_{S,R,C}(x) &= m_{S\cap R\cap C}(x) = \min(m_S(x), m_R(x), m_C(x)) \\
\mathrm{OD}_{S,R,C}(x) &= m_{S\cap R\cap\overline{C}}(x) = m_{(S\cap R)\backslash C}(x) \\
&= \max\{\min(m_S(x), m_R(x)) - m_C(x), 0\} \\
\mathrm{OI}_{S,R,C}(x) &= m_{\overline{S}\cap\overline{R}\cap C}(x) = m_{C\backslash(S\cup R)}(x) \\
&= \max\{m_C(x) - \max(m_S(x), m_R(x)), 0\} \\
\mathrm{UD}_{S,R,C}(x) &= m_{S\cap\overline{R}\cap C}(x) = m_{(S\cap C)\backslash R}(x) \\
&= \max\{\min(m_S(x), m_C(x)) - m_R(x), 0\} \\
\mathrm{UI}_{S,R,C}(x) &= m_{\overline{S}\cap R\cap\overline{C}}(x) = m_{R\backslash(S\cup C)}(x) \\
&= \max\{m_R(x) - \max(m_S(x), m_C(x)), 0\}
\end{aligned}$$

- 各 n で S, R, C ごとの True Positive (TP), False Positive (FP), False Negative (FN) を計算する。

$$\mathrm{TP}_{n,S,R,C} = \sum_{\forall n\text{-gram } x} (\mathrm{TD}_{S,R,C}(x) + \mathrm{TI}_{S,R,C}(x) + \mathrm{TK}_{S,R,C}(x))$$
$$\mathrm{FP}_{n,S,R,C} = \sum_{\forall n\text{-gram } x} (\mathrm{OD}_{S,R,C}(x) + \mathrm{OI}_{S,R,C}(x))$$
$$\mathrm{FN}_{n,S,R,C} = \sum_{\forall n\text{-gram } x} (\mathrm{UD}_{S,R,C}(x) + \mathrm{UI}_{S,R,C}(x))$$

- 各 n でコーパス全体 ($\mathbb{S} = (S_1, \ldots), \mathbb{R} = (R_1, \ldots), \mathbb{C} = (C_1, \ldots)$) での precision, recall の幾何平均を取り、GREEN の F_β 値を求める。

$$\mathrm{precision}(N, \mathbb{S}, \mathbb{R}, \mathbb{C}) = \left(\prod_{n=1}^{N} \frac{\sum_i \mathrm{TP}_{n,S_i,R_i,C_i}}{\sum_i (\mathrm{TP}_{n,S_i,R_i,C_i} + \mathrm{FP}_{n,S_i,R_i,C_i})}\right)^{\frac{1}{N}}$$
$$\mathrm{recall}(N, \mathbb{S}, \mathbb{R}, \mathbb{C}) = \left(\prod_{n=1}^{N} \frac{\sum_i \mathrm{TP}_{n,S_i,R_i,C_i}}{\sum_i (\mathrm{TP}_{n,S_i,R_i,C_i} + \mathrm{FN}_{n,S_i,R_i,C_i})}\right)^{\frac{1}{N}}$$
$$\mathrm{GREEN}_\beta(N, \mathbb{S}, \mathbb{R}, \mathbb{C}) = \frac{(1+\beta^2)\mathrm{precision}(N, \mathbb{S}, \mathbb{R}, \mathbb{C})\mathrm{recall}(N, \mathbb{S}, \mathbb{R}, \mathbb{C})}{\beta^2\mathrm{precision}(N, \mathbb{S}, \mathbb{R}, \mathbb{C}) + \mathrm{recall}(N, \mathbb{S}, \mathbb{R}, \mathbb{C})}$$

- 複数の参照文 ($R_{i_1}, \ldots, R_{i_m}$) がある場合は、文単位でスコアを最大にする $\hat{R}_i$ を選ぶ。

$$\hat{R}_i = \underset{R\in\{R_{i_1},\ldots,R_{i_m}\}}{\mathrm{argmax}}\, \mathrm{GREEN}_\beta(N, (S_i), (R), (C_i))$$

実験

- CoNLL-2014 評価タスクでの評価
  - Corpus-level 評価を行い、Expected Wins 法の人手評価値との相関係数、1 システムに対する1 回の評価の実行時間の平均を計測した。
  - アライメントによる F-score は、precision-oriented な評価を行うため、$\beta = 0.5$ を用いる。n-gram による手法は、参照データ数のバイアスを緩和するため、recall-oriented な評価を行い、$\beta = 2.0$ を用いる。
  - n-gram の計測の分割は、単語単位、文字単位で行い、それぞれ、word+ 手法名、char+ 手法名と表記する。
  - charGREEN は、既存手法より高い相関を示す。
  - GREEN は、既存手法よりも高速に評価を行える。
  - M^2 は人手でアノテーションしたアライメントを必要とする。ERRANT、CLEME は品詞タグ付けによるアライメントを行う。PT-M^2 は BERT を用いる。GREEN は追加のデータやツールが不要かつ高速で、高性能である。

	r	ρ	時間 (秒)
アライメントによる F-score			
M^2	0.623	0.687	4.34
ERRANT	0.644	0.687	12.35
PT-M^2	0.686	0.786	109.82
CLEME	0.648	0.709	10.15
n-gram による precision			
wordGLEU	0.696	0.445	2.69
charGLEU	0.606	0.593	5.00
n-gram による F-score			
wordGREEN	0.741	0.698	**0.55**
charGREEN	**0.786**	**0.813**	1.94

- 評価手法ごとの違いは？
  - AMU : CoNLL-2014 タスクで最も人手評価値が高い出力
  - INPUT : CoNLL-2014 タスクの入力文をそのまま出力とする
  - IPN : CoNLL-2014 タスクで最も人手評価値が低い出力
  - NULL : すべての入力に対して、空の文を出力とする
- アライメントによる F-score は、悪い出力を高く評価してしまうことがある。

	AMU	INPUT	IPN	NULL
アライメントによる F-score				
M^2	35.01	0.00	7.09	28.01
ERRANT	31.97	0.00	5.95	0.20
PT-M^2	35.94	0.00	5.72	2.44
CLEME	25.14	0.00	4.41	33.44
n-gram による precision				
wordGLEU	58.08	56.34	55.08	0.00
n-gram による F-score				
wordGREEN	79.26	76.93	76.31	43.46
人手評価	62.84	45.64	29.99	-