



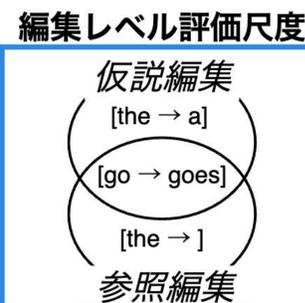
編集レベル評価: 仮説文と参照文を編集の一致度で定量化

- 参照あり評価では仮説文の品質を参照文と比較して評価
 - **編集レベルでの評価** が代表的
- **編集**: 入力文と訂正文を比較し, 単語やフレーズ単位の書き換えとして抽出したもの
 - 入力文に対するスパンと訂正後の文字列を含む

入力文:
He go to the school .

仮説文:
He **goes** to **a** school .

参照文:
He **goes** to **the** school .



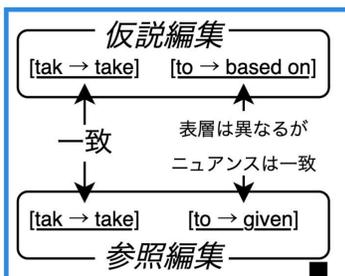
既存の編集レベル評価の問題: 表層の一致に依存

- **スパン・訂正後文字列の相違**がわずかでもあれば不正解
ハードな編集の一致判定
- 表層は異なるがニュアンスは類似する
“惜しい”仮説編集を評価できない
- 編集が多様化するFluency editドメインでは顕著に

入力文:
... and **take** an appropriate action **to** the results .

仮説文:
... and **take** an appropriate action **based on** the results .

参照文:
... and **take** an appropriate action **given** the results .

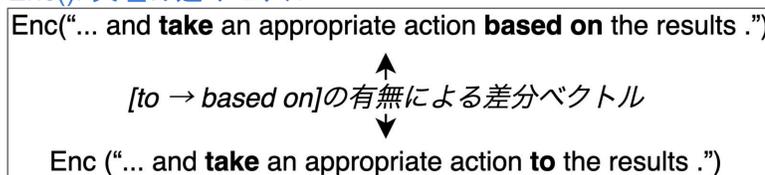


提案法: 編集の類似度を編集ベクトルの類似度で定量化

編集をベクトル化

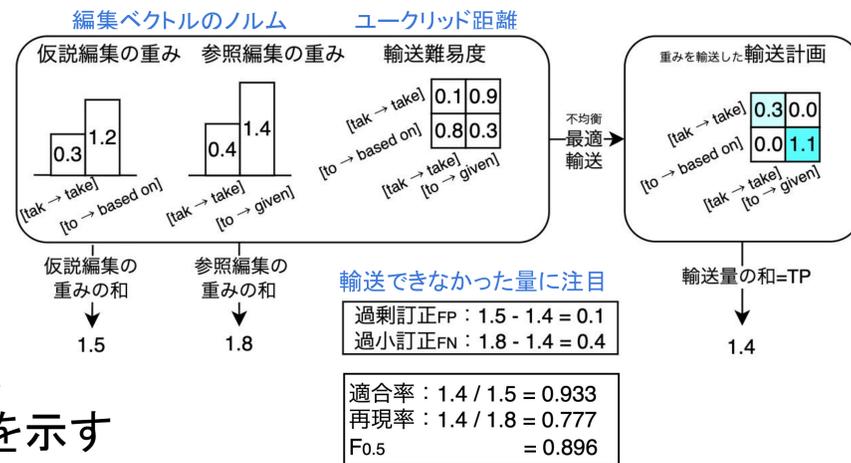
- 編集の適用前後における文埋め込みの差分 = 編集が文埋め込みに与える影響
- 編集の一致度をベクトル類似度で定量化できる
ソフトな編集の一致判定が可能に

参照編集における[to -> based on]の編集ベクトルを計算する場合
Enc(): 文埋め込みモデル



仮説編集から参照編集に編集ベクトルを最適輸送

- **不均衡最適輸送** (Unbalanced Optimal Transport, UOT)
 - 輸送の過不足を許容 → 過剰・過小訂正を評価
- 入力: 仮説・参照編集の重み, 輸送難易度
- 出力: 編集間の輸送量 (編集アライメント)



輸送計画からの適合率・再現率・Fβ計算

- 輸送できた量は仮説・参照の一致度を示す = True Positive
- 輸送できなかった量は過剰訂正・過小訂正の度合いを示す
False Positive False Negative

実験: 編集レベル評価の中では人手評価と高く相関・説明性も高い

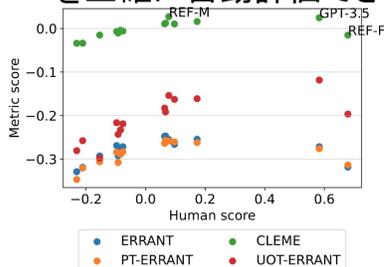
- **SEEDAメタ評価データセット**
 - 13システムを性能順に並び替え
 - 人手評価とのPearson, Spearman相関で評価
 - Base / +Fluency: +Fluencyの方が積極的に訂正
- **提案法は編集レベル評価では最高相関**
 - 既存尺度のERRANT, PT-ERRANT, CLEMEよりも高相関
 - どの編集が対応するかも可視化できる→説明性
- 既存評価よりシステムの順位づけ, 分析の両方で有用
人手評価との相関が大事 説明性が大事

	SEEDA-E Base				SEEDA-E +Fluency			
	Official		10 Refs		E-Fluency		NE-FLuency	
	r	ρ	r	ρ	r	ρ	r	ρ
n-gram レベル評価尺度								
GLEU	.909	<u>.965</u>	.949	.958	.278	.600	.781	.921
GREEN	.921	<u>.965</u>	.910	.979	.547	.802	<u>.745</u>	<u>.908</u>
編集レベル評価尺度								
ERRANT	.881	.895	<u>.952</u>	.951	-.005	.424	.114	.508
PT-ERRANT	<u>.924</u>	.951	.957	<u>.986</u>	.005	.310	.276	.578
CLEME	.910	.930	.932	.937	-.043	.297	.473	.653
UOT-ERRANT	.950	.979	.942	.993	<u>.445</u>	<u>.684</u>	.705	.851

(太字は1位, 下線は2位)

分析

- 人手評価で上位のシステムを正確に自動評価できた



- 編集ベクトルノルムは誤りタイプごとに異なる
- UOTにおける有用な重み付けに

下位3件		上位3件	
誤りタイプ	ノルム	誤りタイプ	ノルム
正書法	0.170 ± 0.406	接続詞	1.247 ± 0.556
句読点	0.835 ± 0.597	その他	1.318 ± 0.685
副詞	0.944 ± 0.485	綴り	1.409 ± 0.610

- 編集ベクトルは綴り, 主語動詞の一致などの種類でクラス形成
 - 類似の編集は類似のベクトルに

- 編集ベクトル類似度には意味があり UOTにおける輸送難易度に有用

