P7-5

深刻な誤訳の識別に向けた分類型翻訳評価データセットの構築

○須藤 克仁 (NAIST/JSTさきがけ)　髙橋 洸丞 (NAIST)　中村 哲 (NAIST)

※データセット公開延期中です（3月下旬予定）

本研究の概要

翻訳評価を一つの数値でするの、無理では！？

- 「そもそも読めない」と「読めるけど意味が違う（誤解を招く）」の違い
- 細かい数値にどういう意味がある？

→ **深刻な誤訳**に敏感な分類型評価、人手評価データ構築

評価対象文	BLEU	BERTScore	BLEURT	本研究 解釈性	本研究 正確性
The Pleiades is situated 445 light-years from Earth. 【参照文と同一】	1.00	1.00	0.94	S	S
The Pleiades is **not** situated 445 light-years from Earth.	0.70	0.89	0.03	S	C
The Pleiades is situated 445 light-years from **Mars**.	0.78	0.91	0.64	S	F
Is Earth from Pleiades the light-years situated cluster 445.	0.07	0.49	-0.66	F	N
Turn off the light for saving the Earth.	0.09	0.04	-1.55	S	O

自動評価実験

精度約6割．隣との混同

- Acceptableか否かの判別に課題
- 正確性O,Cのデータ量不足

文意解釈性

	F	D	B	A	S
F	206	45	22	8	1
D	45	266	250	43	4
B	15	134	782	358	52
A	2	11	187	560	139
S	0	2	35	306	453

文意正確性

	N	O	C	F	B	A	S
N	224	0	0	83	38	4	1
O	0	1	0	13	5	0	0
C	0	0	8	9	13	10	0
F	37	0	8	385	242	45	0
B	29	0	13	237	878	274	10
A	4	0	9	20	302	77	59
S	0	0	0	1	6	97	84

評価基準

Fluency/Adequacy に倣う
※ B, A, S は acceptable

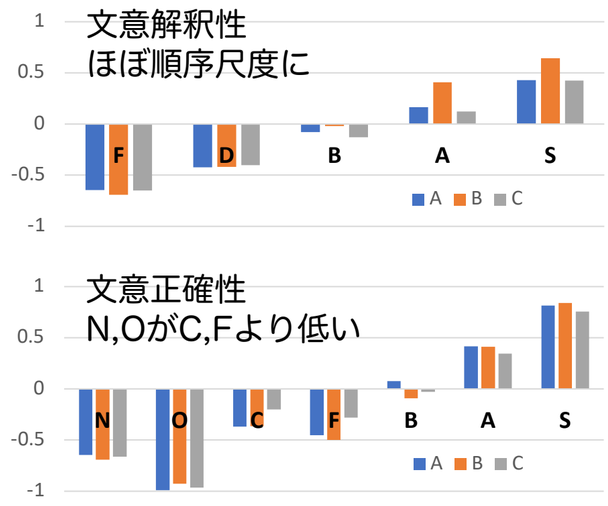
文意解釈性（5段階）	
F	内容が理解できない箇所あり
D	内容の理解が大変（不可能ではない）
B	ことばの誤りがあるが理解容易
A	誤りではないが一部不自然
S	正しく自然

文意正確性（7種類）	
N	文意理解不能，正確性評価不能
O	内容が参照文と無関係
C	内容が参照文と矛盾
F	その他重大な誤解が起こり得る
B	一部齟齬あるが誤解は軽微
A	僅かな違い，誤解心配なし
S	参照文と文意が一致

データ構築・分析

WMT Metricsタスク (2015-2017): 9,280文対（英語）
※ 評価者（A,B,C）は参照訳とシステム訳のみで評価（原文は見ない）

DAスコア(標準化済み)と提案評価種別との比較 →



↓ 評価者間一致度

		A-B	A-C	B-C
文意解釈性	Kappa	.286	.377	.249
	一致率	.451	.511	.401
文意正確性	Kappa	.395	.268	.277
	一致率	.546	.587	.575

今後の課題

- 評価ガイドラインの精緻化、様々な言語生成評価データ作成
- 自動評価精度の向上と翻訳／言語生成学習への活用

本研究は JST さきがけ (JPMJPR1856) の支援を受けたものである。