発表番号
P4-7

単語単位で評価が可能な機械翻訳向け自動評価

高橋洸丞, 須藤克仁, 中村哲(NAIST)

本研究の概要

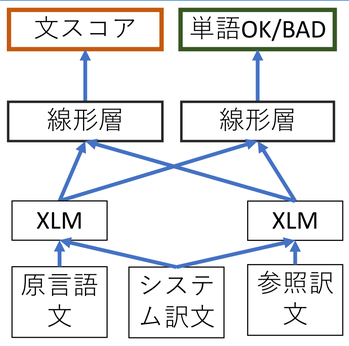
- 機械翻訳(MT)の自動評価として広く用いられているBLEUやBERT$_{SCORE}$、COMETなどは、文単位で評価スコアの出力する
- **文単位の評価は文内部の誤り箇所や誤り度合いを明示できない**
- **本研究では、単語単位での評価を行い、その評価結果を基に文単位での評価スコアを算出する自動評価モデルを提案する**
- 2021年のWMT metricsタスクによる実験において、提案手法は従来の文単位でのみ評価を行うモデルと同等、言語対によっては上回る人手評価との相関が得られた

参考モデル

COMET-22 [Ricard Rei + 2022]

- 2つのcross-encoders
- 出力
  - **単語単位 OK/BAD (エラーの有無)**
  - **文単位スコア**

文スコアと単語評価は個別に出力
独自に追加された[露語-英語]のMQMデータで訓練



MQMデータ

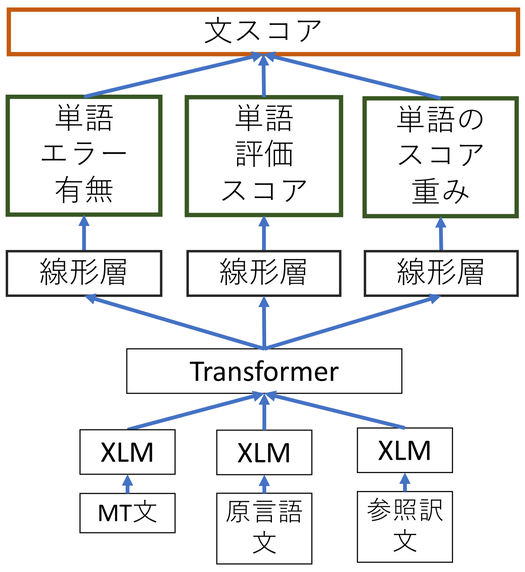
Multidimensional Quality Metrics (MQM:人手評価)
文単位の評価時は、文単位のスコアに変換が必要

文スコア =
「誤りカテゴリ」と
「誤り度合い」の組み合わせ

原言語文	MT文	誤りカテゴリ	誤り度合い	文スコア
自治区总工会相关负责人表示。	A spokesman for the <v>regional trade union</v> said.	Accuracy / Mistranslation	Major	5.0

提案手法

- 3つのbinary-encoders
- 出力
  - **単語エラー有無 [0:エラーなし, 1:エラー有り]**
  - **単語スコア (エラー度合い)**
  - **単語重み (文スコア計算時に対象単語をどれだけ重要視するか)**
  - **文単位スコア (単語単位の出力から計算)**

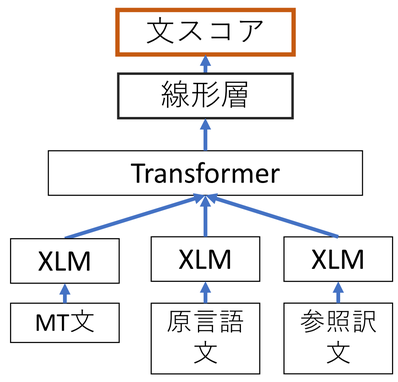


単語単位の出力と一貫性のある文スコア

$$\text{システム訳文長} = \sum_{k=1} (\text{エラー有無}_k * \text{単語スコア}_k * \text{重み}_k)$$

実験設定

- 訓練データ (英-独, 中-英)
  - WMT20 Metrics Shared Task MQM データ
  - TED Talks MQMデータ
- ベースライン
  - 文スコアのみを直接予測するモデル
- 参考モデル
  - COMET-22 segment (文スコア)
  - COMET-22 tagging (単語エラー有無)
- メタ評価
  - 単語単位: 適合率, 再現率, F値, エラースパンヒット率(HSH, TSH)
  - 文単位: ピアソン, ケンドールの相関係数
- テストデータ(英-独, 中-英)
  - WMT21 Metrics Shared Task MQMデータ



ベースライン:
文スコアのみを
直接予測するモデル

- エラースパンヒット率
  - 正解ラベルと予測結果のエラースパンが接触していれば1ヒット
  - HSH $= \frac{\text{正解ラベルと予測結果のエラースパンのヒット数}}{\text{予測エラースパン数}}$
  - TSH $= \frac{\text{正解ラベルと予測結果のエラースパンのヒット数}}{\text{正解エラースパン数}}$

単語単位の実験結果

評価モデル	英-独			中-英			英-独 + 中-英		
	適合率	再現率	F値	適合率	再現率	F値	適合率	再現率	F値
COMET-22 tagging	**0.270**	0.331	**0.297**	**0.329**	0.596	**0.424**	**0.311**	0.495	**0.382**
提案モデル	0.220	**0.397**	0.283	0.258	**0.726**	0.381	0.248	**0.601**	0.351

評価モデル	英-独		中-英		英-独 + 中-英	
	HSH	TSH	HSH	TSH	HSH	TSH
COMET-22 tagging	**0.209**	**0.519**	0.253	0.583	0.237	0.563
提案モデル	0.199	0.500	**0.283**	**0.664**	**0.252**	**0.613**

提案モデルは
- COMET22よりもrecallが高い
- COMET22よりもエラースパンのヒット比率が高い

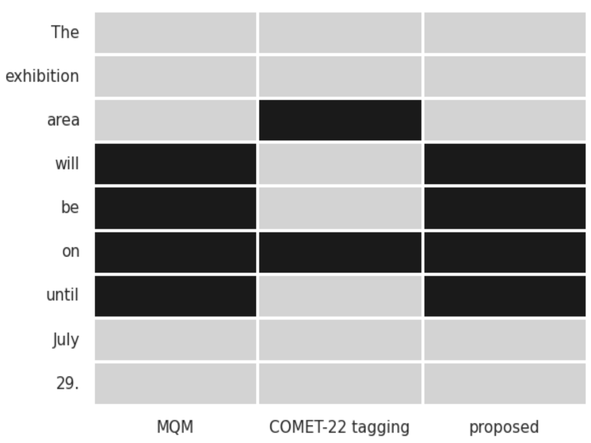
文単位の実験結果

評価モデル	英-独		中-英		英-独 + 中-英	
	ピアソン	ケンドール	ピアソン	ケンドール	ピアソン	ケンドール
COMET-22 tagging	**0.361**	**0.266**	**0.509**	**0.387**	0.476	0.316
COMET-22 segment	0.232	0.220	0.252	0.241	0.344	0.313
文スコアのみ	0.258	0.187	0.469	0.335	**0.534**	**0.367**
提案モデル	0.289	0.194	0.484	0.347	**0.539**	**0.369**

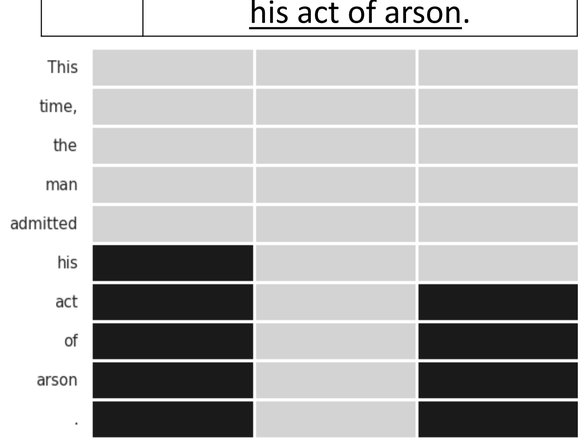
提案モデルは
- en-de+zh-enでは文スコアと同等
- en-de, zh-enの個別では、文スコアを超える相関係数値

事例分析 (黒色:エラー有, 灰色:エラー無)

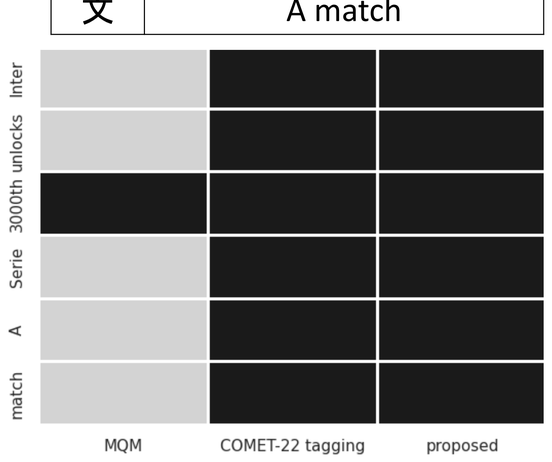
原文	展销区将展至7月29日。
参照訳文	The exhibition area will be open until July 29.
MT文	The exhibition area will be on until July 29.



原文	这一次，男子承认了自己纵火的行为。
参照訳文	This time, the man admitted to his arson.
MT文	This time, the man admitted his act of arson.



原文	国际米兰解锁第3000场意甲比赛-中新网
参照訳文	Inter Milan Has Its 3000th Serie A Game - China News
MT文	Inter unlocks 3000th Serie A match



原文	波兰将退出欧洲防止对女子施暴的条约
参照訳文	Poland to Withdraw from the European Treaty on Violence against Women
MT文	Poland to withdraw from European treaty to prevent violence against women

提案モデル(proposed)は
・(成功):スパン完全一致、小さなずれ　　・(失敗):広すぎるエラースパン、大きなずれ