Pxx　RUSE: 文の分散表現を用いた回帰モデルによる機械翻訳の自動評価

嶋中宏希 †　梶原智之 †‡　小町守 †　shimanaka-hiroki@ed.tmu.ac.jp
†: 首都大学東京　‡：大阪大学

1. 研究の背景と概要

- WMTのMetricsタスクで優秀な成績を収めたほとんどの機械翻訳自動評価手法が表層に基づく素性を利用している.
  → 意味が似ていても表層が異なる文に対して誤った評価をしてしまう.
- 本研究では，汎用的な文の分散表現を利用し，表層ではなく意味的な情報を考慮した自動評価手法を提案する.
- 実験の結果，我々の提案手法は**文の分散表現のみを素性として用いた回帰モデル**で**最高性能を達成**した.

例:

出力文: This is not a major issue.

参照文: It is nothing major.

評価手法	スコア	ランキング
Human	0.892	32/560
Blend	- 0.0734	423/560
Our metric	**0.554**	**60/560**

2. 提案手法　RUSE: Regressor Using Sentence Embeddings

- 本研究では，事前学習された汎用的な**文の分散表現を用いた回帰モデル**で人手評価値の学習を行い機械翻訳の自動評価を行う.

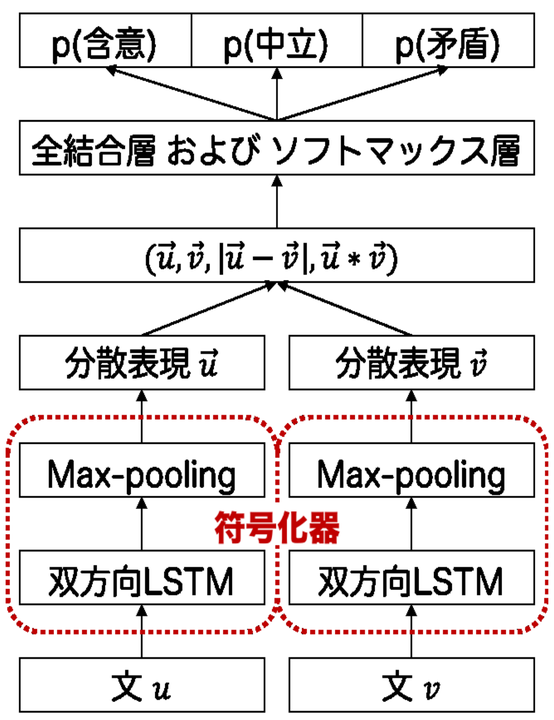

図 1: InferSentの概要

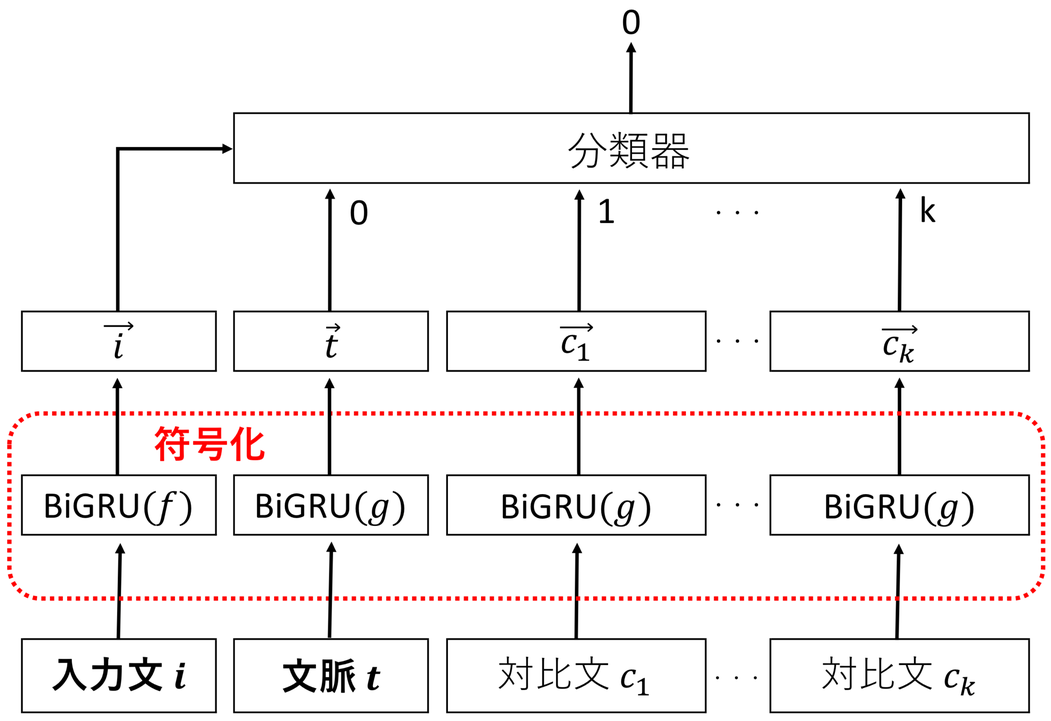

図 2: Quick-Thoughtの概要

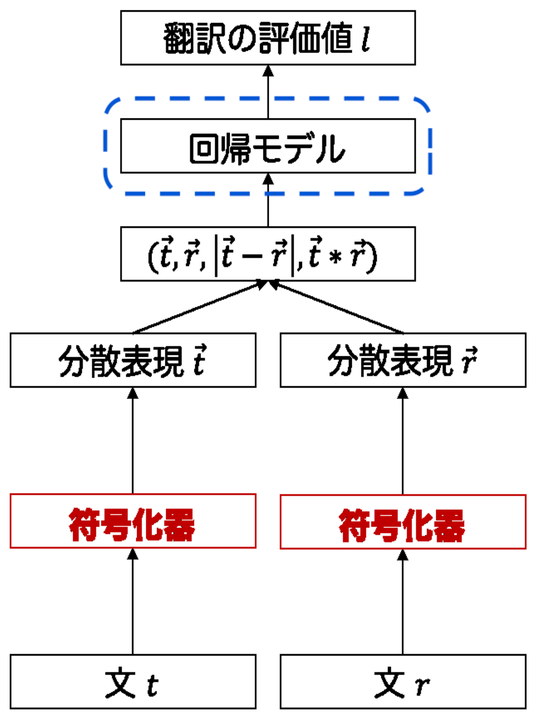

図 3: 提案手法（RUSE）の概要

3. 実験設定

◆ 汎用的な文の分散表現

- InferSent (IS) [Conneau et al., 2017]
  - 学習データ: SNLI Corpus, MultiNLI Corpus
  - 次元: 4,096
- Quick-Thought (QT) [Logeswaran and Lee, 2018]
  - 学習データ: Toronto-Books Corpus, UMBC corpus
  - 次元: 4,800
- Universal Sentence Encoder (USE) [Cer et al., 2018]
  - 学習データ: 様々なWebデータ, SNLI Corpus
  - 次元: 512

◆ 人手評価値の学習データセット

- WMTの人手評価値データセットを用いた.

表 1: WMTの人手評価値付きの文対数（英語方向）

	cs	de	fi	lv	ro	ru	tr	zh
WMT15	500	500	500	-	-	500	-	-
WMT16	560	560	560	-	560	560	560	-
WMT17	560	560	560	560	-	560	560	560

◆ 機械翻訳自動評価のための回帰モデル

MLP と SVR を用いてそれぞれ実験を行った.

- 多層パーセプトロン (MLP) (Chainer を使用)
  - Number of Layers ∈ {1, 2, 3} (ReLU関数)
  - Number of Units ∈ {512, 1024, 2048, 4096}
  - Batch size ∈ {64, 128, 256, 512, 1024}
  - Dropout rate ∈ {0.1, 0.3, 0.5}
  - Optimizer ∈ {Adam}
- サポートベクター回帰 (SVR) (scikit-learnを使用)
  - RBF カーネル
  - C ∈ {0.1, 1.0, 10}
  - ε ∈ {0.01, 0.1, 1.0}
  - γ ∈ {0.001, 0.01, 0.1}

4. 実験結果

- 学習&開発データ: WMT15とWMT16のデータセット (5,360文対)
- 評価データ: WMT17 のデータセット (各言語対 560文対)
- MLPでは10分の1を開発データとして用い，SVRでは10分割交差検定を行った.

表2 : **文単位**での人手評価とのピアソンの相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg.
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
Blend [Ma et al., 2017]	0.594	0.571	0.733	0.577	0.633	0.671	0.661	0.633
MEANT2.0 [Lo, 2017]	0.578	0.565	0.687	0.586	0.607	0.596	0.639	0.608
chrF++ [Popović, 2017]	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
RUSE (MLP) with IS	0.556	0.568	0.706	0.650	0.626	0.649	0.634	0.627
RUSE (MLP) with QT	0.601	0.587	0.737	0.685	0.661	0.692	0.647	0.658
RUSE (MLP) with USE	0.592	0.596	0.681	0.621	0.598	0.645	0.620	0.622
RUSE (MLP) with IS + QT + USE	0.614	0.637	**0.756**	**0.705**	**0.680**	0.704	0.677	0.682
RUSE (SVR) with IS + QT + USE	**0.624**	**0.644**	0.750	0.697	0.673	**0.716**	**0.691**	**0.685**

表3 : **システム単位**での人手評価とのピアソンの相関係数

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg.
BLEU	0.971	0.923	0.903	0.979	0.912	0.976	0.864	0.933
Blend [Ma et al., 2017]	0.968	**0.976**	0.958	0.979	**0.964**	0.984	0.894	0.960
BEER [Stanojević et al., 2015]	0.972	0.960	0.955	0.978	0.936	0.972	0.902	0.954
UHH_TSKM [Duma et al., 2017]	**0.996**	0.937	0.921	0.990	0.914	0.987	0.902	0.950
RUSE (MLP) with IS + QT + USE	0.995	0.964	**0.985**	**0.996**	0.956	**0.993**	**0.937**	**0.975**
RUSE (SVR) with IS + QT + USE	**0.996**	0.964	0.983	0.988	0.951	**0.993**	0.930	0.972