エントレインメント尺度および戦略が対話システムの評価に与える影響の調査

金崎 翔大 [1,2]　河野 誠也 [2]　湯口 彰重 [3,2]　桂井 麻里衣 [1]　吉野 幸一郎 [2]
[1] 同志社大学大学院理工学研究科　[2] 理化学研究所ガーディアンロボットプロジェクト
[3] 東京理科大学先進工学部
{kanezaki21,katsurai}@mm.doshisha.ac.jp
{seiya.kawano,koichiro.yoshino,akishige.yuguchi}@riken.jp

概要

人間同士の会話では，やりとりを重ねるうちに話者のふるまいが同調するエントレインメント現象がしばしば発生する．こうした現象を扱う対話システムを構築しようとする場合，どのようなエントレインメント尺度を活用するか，どのようなタイミングでどのようなエントレインメント度合いを用いるかの 2 点を明らかにする必要がある．本研究ではニューラル雑談対話モデルをエントレインメント度合いに応じてリランキングするシステムを用い，複数の語彙的エントレインメント尺度，および複数のエントレインメント戦略を網羅的に組み合わせた主観評価実験を行った．

1　はじめに

対話が進行するにつれて，語彙，構文構造，文体，韻律など，様々な要素において話者間のふるまいが類似する現象をエントレインメントとよぶ．エントレインメントは対話タスクの成功率や自然さと相関することが報告されており，その追跡を通じて対話システムの性能を評価しようとする試みがある [1, 2, 3, 4]．

我々はこれまでに，対話文脈においてエントレインメントを柔軟に考慮するように応答文を選択する手法を構築してきた [5, 6, 7]．この中で，Word Mover's Distance （WMD）に基づく尺度 [2, 3] と，Bidirectional Encoder Representations from Transformers （BERT）に基づく尺度 [4] の二種類のエントレインメント尺度（スコア）を定義して用いてきた．また，どのような対話文脈でどのようなエントレインメント尺度を持つ応答を選択するか決定するため，人間による対話コーパスを用いてエントレインメント予測モデルを学習し，目標とする同調度合いを決定するエントレインメント戦略を構築してきた．これらの手法に対してコーパス内の対話履歴を正解データとみなした定量評価実験を行ってきたものの，対話システム構築にあたっては実際のユーザによる評価が重要である．そこで本研究では，各エントレインメント尺度と，エントレインメント予測モデルにその他の戦略を織り込んだ複数のエントレインメント戦略の組み合わせを用い，エントレインメントを行う対話モデルにおける最適な設定を網羅的に探索する．具体的には，各組み合わせによってリランキングされた対話応答に対してクラウドソーシングによる主観評価実験を行う．

2　エントレインメントスコア

エントレインメントスコアは，発話者 S_1，応答者 S_2 による発話ペア (U_{S_1}, U_{S_2}) に関し，ターゲット発話 U_{S_2} に与えられる．言語的エントレインメントを測定するフレームワークとして Local Interpersonal Distance （LID） [2] が提案されており，このスコア計算に異なる 2 つの指標を用いる．

2.1　WMD に基づくスコア

二つの文の類似度を計算するために提案された WMD [8] を用いて，発話 U_{S_1}, U_{S_2} 間の単語分散表現空間上での意味的距離をスコアとする [2, 3]．

$$\mathrm{LID}_{\mathrm{WMD},U_{S_2}} = \mathrm{WMD}(U_{S_1}, U_{S_2}) \tag{1}$$

WMD は単語埋め込みを用いた手法であるため，文脈埋め込みと比較して単語の構成性を考慮でき，よりスタイルに焦点を置いたスコアになっていることが期待される．

2.2 BERT に基づくスコア

BERT 文脈埋め込み空間におけるコサイン類似度により二つの文の類似度を計算する手法として提案された BERTScore [9] を用いて，発話 U_{S_1}, U_{S_2} 間の埋め込み空間上での意味的距離をエントレインメントスコアとする．

$$\mathrm{LID}_{\mathrm{BERT},U_{S_2}} = 1 - \mathrm{BERTScore}(U_{S_1}, U_{S_2}) \quad (2)$$

こちらは発話文全体の内容を考慮したエントレインメントスコアとなることが期待される．

3 エントレインメント戦略

エントレインメント戦略の一つとして，人間が行っているエントレインメント戦略の模倣を考える．このため，エントレインメントスコアの予測値 $\hat{\mathrm{LID}}_{U_{S_2}}$ を予測するモデルを構築する（4.2 節）．また，実際に人間が行っているエントレインメント戦略が最適であるかどうかを確認するため，以下の 2 種類の戦略を試行する．

$$\mathrm{LID}_{\mathrm{oracle}} = \mathrm{LID}_{U_{S_2}} \quad (3)$$

$$\mathrm{LID}_{\mathrm{min}} = \min_{R_i \in \boldsymbol{R}} \left(\mathrm{LID}_{R_i}\right) \quad (4)$$

ここで R_i は応答候補集合 $\boldsymbol{R}$ から選択された応答候補の一つである．$\mathrm{LID}_{\mathrm{oracle}}$ は実際に人間同士の対話データに存在するデータから計算したエントレインメントスコアであり，人間のエントレインメント戦略を模倣した場合に相当する．また $\mathrm{LID}_{\mathrm{min}}$ は常に最大限ユーザ発話に同調する発話を選択することに相当する．

4 エントレインメントを考慮した応答リランキング

各エントレインメントスコアおよびエントレインメント戦略の効果を検証するため，著者らが過去に提案したリランキング手法 [6] を適用する．手法の概要を図 1 に示す．以下で各モジュールを説明する．

図 1 リランキング手法の概要

4.1 ニューラル雑談対話モデル

ニューラル雑談対話モデルは，入力として発話文 U_{S_1} を受け取り，出力として入力発話に対する n 個の n-best 応答候補集合 $\boldsymbol{R} = \{R_1, R_2 \ldots, R_n\}$ および応答候補の尤度集合 $\boldsymbol{l_R} = \{l_{R_1}, l_{R_2} \ldots, l_{R_n}\}$ を出力する．ただし，応答候補集合 $\boldsymbol{R}$ は尤度の降順になっており，$l_{R_1} \geq l_{R_2} \geq \ldots, \geq l_{R_n}$ を満たす．

4.2 エントレインメント予測モデル

エントレインメント予測モデルは，発話文 U_{S_1} が入力されたとき，その応答文 U_{S_2} が持つべきエントレインメントスコア $\mathrm{LID}_{\mathrm{WMD/BERT},U_{S_2}}$ を予測するように学習する[1)]．図 2 に示すように，予測モデルは Gated Reccurent Unit （GRU） [10] を用いた階層的エンコーダモデルとバイアス項を含む全結合層（Linear）を用いたエントレインメントデコーダモデルで構成される．

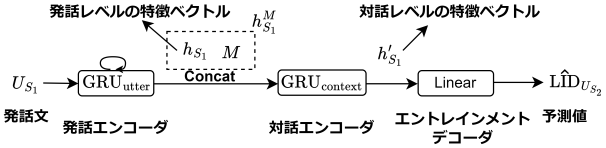

図 2 エントレイメント予測モデルの概要

発話文中の各単語 $u_{S_1,i} \in U_{S_1}$ を，式（5）で定義するエンコーダで発話レベルの特徴ベクトル $\boldsymbol{h_{S_1}} = \boldsymbol{h_{S_1,|U_{S_1}|}}$ に変換する[2)]．

$$\boldsymbol{h_{S_1,i}} = \mathrm{GRU}_{\mathrm{utter}}(\boldsymbol{h_{S_1,i-1}}, \mathrm{Embedding}(u_{S_1,i})) \quad (5)$$

ここで，Embedding$(\cdot)$ は，U_{S_1} における各単語 $u_{S_1,i} \in U_{S_1}$ を固定長の密ベクトル表現に変換する単語埋め込み層である．

本研究では発話文だけではなく，発話に付随する属性情報を対話エンコーダの追加の入力として用いることでエントレインメントの予測性能の向上を期待する [11, 12]．具体的には，属性情報 $\boldsymbol{M}$ を発話レベルの特徴ベクトル $\boldsymbol{h_{S_1}}$ の一部と見なして次式のように結合する．

$$\boldsymbol{h^M_{S_1}} = \mathrm{Concat}(\boldsymbol{h_{S_1,i}}, \boldsymbol{M}) \quad (6)$$

ここで，属性情報 $\boldsymbol{M}$ は 2 人の話者 S_1，S_2 の社会的関係ベクトル v_{uS}，v_{uT} を用いた．

次に，式（7）で定義される対話エンコーダを用いて，各ターンまでに得られた発話レベルの特徴ベクトル $\boldsymbol{h_{S_1}}$ の系列を対話レベルの特徴ベクトル $\boldsymbol{h'_{S_1}}$ に統合する[3)]．

1） “$\mathrm{LID}_{\mathrm{WMD/BERT}}$” は $\mathrm{LID}_{\mathrm{WMD}}$ または $\mathrm{LID}_{\mathrm{BERT}}$ を意味する．これ以降，LID と略す．

2） $\boldsymbol{h_{S_1}}$ は発話エンコーダ $\mathrm{GRU}_{\mathrm{utter}}$ に対する最後の単語 $u_{S_1,|U_{S_1}|}$ の入力に対応する，隠れベクトル $\boldsymbol{h_{S_1,|U_{S_1}|}}$ である．

3） $\boldsymbol{h'}$ は発話 S_1 の 1 ターン前の対話レベルの特徴ベクトルを

$$\boldsymbol{h}'_{S_1} = \mathrm{GRU}_{\mathrm{context}}(\boldsymbol{h}^{\boldsymbol{M}}_{S_1}, \boldsymbol{h}') \tag{7}$$

さらに，デコーダ Linear(·) を用いて，各ターンの特徴ベクトル $\boldsymbol{h}'_{S_1}$ から予測値 $\hat{\mathrm{LID}}_{U_{S_2}}$ を得る．

$$\hat{\mathrm{LID}}_{U_{S_2}} = \mathrm{Linear}(\boldsymbol{h}'_{S_1}) \tag{8}$$

4.3 リランキング機構

リランキング機構では，応答候補集合 $\boldsymbol{R}$ からエントレインメント戦略に従った候補を選択して利用する．各エントレインメント戦略から得られた目標エントレインメントスコア $\mathrm{LID}_{\mathrm{target}}$ と，応答候補 $R_i \in \boldsymbol{R}$ のエントレインメントスコア LID_{R_i} との差の絶対値を，エントレインメントの実現距離 $d_{R_i,\mathrm{target}}$ とする．

$$d_{R_i,\mathrm{target}} = |\mathrm{LID}_{R_i} - \mathrm{LID}_{\mathrm{target}}|. \tag{9}$$

この値が小さくなるように応答を選択する．具体的には，$R_i \in \boldsymbol{R}$ に対し，正規化した尤度 l_{R_i} と正規化した実現距離 $d_{R_i,\mathrm{target}}$ の重み付き調和平均を算出し，最大値をとる応答 $R_{\text{F-beta}}(\beta)$ を選択する．

$$R_{\text{F-beta}}(\beta) = \underset{R_i \in \boldsymbol{R}}{arg\,max} \left[(1+\beta^2) \frac{\tilde{\frac{1}{|l_{R_i}|}} \times \tilde{\frac{1}{d_{R_i,\mathrm{target}}}}}{\beta^2 \tilde{\frac{1}{|l_{R_i}|}} + \tilde{\frac{1}{d_{R_i,\mathrm{target}}}}} \right]. \tag{10}$$

$\tilde{\cdot}$ はそれぞれの応答候補リストを基準とした Min-Max 正規化を表し，β は 0 以上の実数となる重み係数である．本研究では，$\beta = 1$ と設定した．

5 評価実験

実験では，WMD と BERT に基づく 2 種類のエントレイメントスコア及び各エントレインメント戦略の組み合わせと，クラウドソーシングによる対話システムの主観評価値との関係を調べる．エントレインメント戦略の設定には次の 5 種類がある．

- $\mathrm{LID}^{\cdot}_{\mathrm{predict}}$: 予測されたエントレインメントスコアを利用
- $\mathrm{LID}^{\cdot}_{\mathrm{predict}} w/ v_{S_2}$: 予測されたエントレインメントスコアを利用（当該対話コンテキストにおいてシステム側の発話者が持つ属性情報を追加）
- $\mathrm{LID}^{\cdot}_{\mathrm{predict}} w/ v_{S_1}, v_{S_2}$: 予測されたエントレインメントスコアを利用（当該対話コンテキストにおいて両方の発話者が持つ属性情報を追加）
- $\mathrm{LID}^{\cdot}_{\mathrm{oracle}}$: 対話コーパス上の正解から計算されたエントレインメントスコアを利用（人間のエントレインメント戦略の模倣）
- $\mathrm{LID}^{\cdot}_{\mathrm{min}}$: 常に最もユーザ発話に同調した発話候補を利用

表す．ただし，本研究では対話履歴は 1 ターンのみであるためゼロベクトルを用いた．

5.1 データセット

実験には Twitter のデータを使用した [13]．Twitter から，ユーザ間のフォロー・フォロワー（FF）関係，返信ペア，すべてのユーザとツイートに関連するメタデータを得るために，まず 5 人のシードユーザを選んだ．シードユーザには，Twitter 上で多くのフォロワーを持ち，情報工学に関する有名な日本人研究者を用いた．次に，彼らの友人とフォロワーのリストを取得した．これらのリストに含まれるユーザから，2 次の FF 関係であるフォロー及びフォロワー関係のリストを取得する．最初のシードユーザと上記の関係リストを用いて，ユーザーをノード，“フォロー” 関係を有向エッジとするソーシャルグラフを構築した．その後，グラフ内のノード・ユーザーから全てのツイート（2018 年 3 月から 2022 年 3 月まで）を取得した．ツイートデータの分割は，50,000 ペア（2018 年 3 月から 2022 年 1 月まで）を学習データ，5,000 ペア（2022 年 2 月）を開発データ，300 ペア（2022 年 3 月）をテストデータとした．社会的属性情報を表すユーザベクトル v_{S_1}, v_{S_2} は先行研究 [14] に従い，ユーザのフォロワーネットワークを node2vec アルゴリズムに入力して計算した．

5.2 実験設定

応答候補集合を生成するニューラル雑談対話モデルには，NTT から公開されている Twitter のデータで事前学習済み Transformer Encoder-Decoder モデルを用いた [15]．リランキング対象の応答候補は 40 個とした．

単語分散表現モデルについては，Akama らが提案したスタイルの類似性と統語的・語義的な類似性の両方を考慮した単語分散表現モデル [16] を用いた．単語分散表現モデルの学習には学習データセット内に含まれる発話を用いた．BERT モデルについては，学習済み多言語 BERTScore を用いた [9]．

5.3 クラウドソーシングにおける主観評価

ワーカに発話文 U_{S_1} を提示し，応答について 5 つの項目に関する 5 段階評価の質問を行った．

- 自然性：人間らしい自然な応答ができているか
- 楽しさ：応答文は面白かったか，会話を続けた

表 1　主観評価と自動評価の実験結果

$\mathrm{LID}_{\mathrm{target}}$, 予測モデル	主観評価					自動評価	
	自然性	楽しさ	話題追随性	共感性	スタイル類似性	sMAPE	BLEU （×100）
Human	3.65**	3.13	3.72*	3.46	3.39	-	-
R_1（baseline）	3.27	3.00	3.47	3.34	3.29	24.89%（$\mathrm{LID}^{\mathrm{WMD}}$） 24.78%（$\mathrm{LID}^{\mathrm{BERT}}$）	0.86
$\mathrm{LID}^{\mathrm{WMD}}_{\mathrm{predict}}$	3.34	3.12	3.47	3.39	3.42	11.08%	0.49
$\mathrm{LID}^{\mathrm{WMD}}_{\mathrm{predict}}$（w/ v_{S_2}）	3.30	3.00	3.49	3.33	3.31	11.38%	0.50
$\mathrm{LID}^{\mathrm{WMD}}_{\mathrm{predict}}$（w/ v_{S_1}, v_{S_2}）	3.25	2.96	3.30	3.34	3.20	11.30%	0.70
$\mathrm{LID}^{\mathrm{WMD}}_{\mathrm{oracle}}$	3.24	3.00	3.37	3.22	3.37	8.24%	0.93
$\mathrm{LID}^{\mathrm{WMD}}_{\mathrm{min}}$	3.32	3.13	3.75*	**3.59***	3.64*	40.87%	1.22
$\mathrm{LID}^{\mathrm{BERT}}_{\mathrm{predict}}$	3.34	3.14	3.56	3.46	3.38	10.35%	0.54
$\mathrm{LID}^{\mathrm{BERT}}_{\mathrm{predict}}$（w/ v_{S_2}）	3.30	3.02	3.53	3.46	3.46	9.70%	0.32
$\mathrm{LID}^{\mathrm{BERT}}_{\mathrm{predict}}$（w/ v_{S_1}, v_{S_2}）	3.32	3.03	3.46	3.39	3.31	9.29%	0.52
$\mathrm{LID}^{\mathrm{BERT}}_{\mathrm{oracle}}$	3.30	3.11	3.47	3.31	3.34	7.00%	0.92
$\mathrm{LID}^{\mathrm{BERT}}_{\mathrm{min}}$	**3.46**	**3.18**	**3.89***	3.50	**3.86***	36.00%	1.03

各項目ごとに R_1（baseline）との両側 t 検定を行い有意差がある場合: * : $p \leq 0.05$, ** : $p \leq 0.01$

いか
- 話題追従性：人間の話題に沿っていたか
- 共感性：システムは人間と共感・協調できているか
- スタイル類似性：システムは人間と似た口調で応答ができているか

すべての評価項目において「5」が最も良い評価，「1」が最も悪い評価であり段階的な評価とした．

5.4　自動評価

応答の評価指標としてテストデータ中に存在する参照発話との BLEU[17] を用いた．また，選択された応答の理想的なエントレインメントスコアとの誤差の計算には式（11) で表される symmetric mean absolute percentage error （sMAPE）を用いた．n は評価データの数である．

$$\begin{aligned}&\mathrm{sMAPE}(\mathbf{LID}_{\mathrm{R_{F\text{-}beta}}}, \mathbf{LID}_{\mathrm{oracle}})\\ &= \frac{200}{n}\sum_{j=1}^{n}\frac{|\mathrm{LID}_{\mathrm{R_{F\text{-}beta}},j} - \mathrm{LID}_{\mathrm{oracle},j}|}{|\mathrm{LID}_{\mathrm{R_{F\text{-}beta}},j}| + |\mathrm{LID}_{\mathrm{oracle},j}|}\end{aligned} \tag{11}$$

5.5　実験結果

表 1 に主観評価及び自動評価の結果を示す．また，比較として Twitter における人間の応答とリランキング前の応答についても評価を行った．sMAPE については，それぞれのリランキング機構に用いたエントレインメントスコアによって計算した．

主観評価について注目すると，oracle 戦略を用いたシステムを比較しても，min 戦略は各主観評価において高い評価を得られることが示された．また，話題追随性，スタイル類似性については BERT 尺度に基づくリランキングが，共感性については WMD 尺度に基づくリランキングがより大きく作用することが示唆された．これは我々の，WMD 尺度がよりスタイル類似性に作用し，BERT 尺度はより話題追随性に作用するという予想とは異なる結果となった．また意外なことに，いくつかの主観評価項目においては，人間の応答よりも BERT スコアによるリランキング応答がより高い評価を得ており，これは必ずしも人間のエントレインメント戦略が対話相手にとって最適ではない可能性を示している．

次に，自動評価について注目する．エントレインメント誤差（sMAPE）はベースラインであるリランキング前の応答と比較して，$\mathrm{LID}_{\mathrm{min}}$ を除いて誤差が小さくなった．一方で，最も同調するようにリランキングをする場合の $\mathrm{LID}_{\mathrm{min}}$ では，エントレインメント尺度にかかわらず，エントレインメント誤差が増加している．これは，人間が対話文脈に合わせて柔軟に同調度合いを調整していることを示唆している．しかし主観評価で示唆されたように，このような人間の対話戦略を模倣することが対話システムにとって常に良い戦略とは限らない．

6　おわりに

本研究では，エントレインメント現象を考慮する対話システムを構築し，その評価実験をした．特に，異なるエントレインメントスコア及びエントレインメント戦略を用いた場合に，各主観評価項目にどのような影響が出るかを調査した．

クラウドソーシングによる主観評価の結果，当初の予想と異なり，話題追随性，スタイル類似性については BERT 尺度に基づくリランキングが，共感性については WMD 尺度に基づくリランキングが大きく作用することが示唆された．また，人からの印象を良くするためには常にエントレインメント度合を最大化するような戦略が望ましいことが示された．

謝辞

本研究は JSPS 科研費 22K17958，22H04873，20H04484 の助成を受けた．

参考文献

[1] Ani Nenkova, Agustin Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In **Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers**, pp. 169–172. Association for Computational Linguistics, 2008.

[2] Md. Nasir, Sandeep Nallan Chakravarthula, Brian R.W. Baucom, David C. Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. Modeling Interpersonal Linguistic Coordination in Conversations Using Word Mover' s Distance. In **Proc. INTERSPEECH 2019**, pp. 1423–1427, 2019.

[3] Seiya Kawano, Masahiro Mizukami, Koichiro Yoshino, and Satoshi Nakamura. Entrainable neural conversation model based on reinforcement learning. **IEEE Access**, Vol. 8, pp. 178283–178294, 2020.

[4] Yuning Liu, Aijun Li, Jianwu Dang, and Di Zhou. Semantic and acoustic-prosodic entrainment of dialogues in service scenarios. In **Companion Publication of the 2021 International Conference on Multimodal Interaction**, ICMI '21 Companion, p. 71–74, New York, NY, USA, 2021. Association for Computing Machinery.

[5] 金崎翔大, 河野誠也, 湯口彰重, 桂井麻里衣, 吉野幸一郎. 対話における後続発話のエントレインメント予測. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 50–55, 2021.

[6] 金崎翔大, 河野誠也, 湯口彰重, 桂井麻里衣, 吉野幸一郎. エントレインメント予測に基づいたニューラル雑談対話モデルの応答リランキング. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 3Yin248–3Yin248, 2022.

[7] 金崎翔大, 河野誠也, 湯口彰重, 桂井麻里衣, 吉野幸一郎. エントレインメントスコアを用いた応答リランキングとその自動評価. 言語処理学会 第 29 回年次大会 発表論文集, pp. 1963–1968, 2023.

[8] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In **International Conference on Machine Learning**, pp. 957–966, 2015.

[9] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[10] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In **Proc. of EMNLP**, pp. 1724–1734, 2014.

[11] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. **Politeness: Some universals in language usage**, Vol. 4. Cambridge university press, 1987.

[12] Cindy Gallois, Tania Ogay, and Howard Giles. Communication accommodation theory: A look back and a look ahead. In **Theorizing about intercultural communication**, pp. 121–148. Thousand Oaks: Sage, 2005.

[13] Seiya Kawano, Shota Kanezaki, Angel Fernando Garcia Contreras, Akishige Yuguchi, Marie Katsurai, and Koichiro Yoshino. Analysis of style-shifting on social media: Using neural language model conditioned by social meanings. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

[14] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 855–864, 2016.

[15] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[16] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised learning of style-sensitive word vectors. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 572–578, 2018.

[17] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.