音声言語処理のための注意機構を用いた音声認識仮説統合

叶高朋[1]　小川厚徳[1]　マーク・デルクロア[1]　渡部晋治[2]
[1] 日本電信電話株式会社　[2] カーネギーメロン大学
{takatomo.kanou.xe, atsunori.ogawa.gx, marc.delcroix.hc}@hco.ntt.co.jp
shinjiw@cmu.edu

概要

音声要約や音声翻訳などの音声言語処理（Spoken language processing：SLP）の多くは, 音声認識（Automatic speech recognition：ASR）モデルとテキスト翻訳・要約などの自然言語処理（Natural language processing：NLP）モデルの直列接続で実現されている．このような直接接続のシステム構成では, 音声認識誤りが NLP モデルの性能に悪影響を与えることが知られている．本研究では, 音声認識誤りの悪影響を低減させるため, 複数の ASR モデルの出力を統合しながら, 翻訳・要約を行う方法を提案する．提案手法は要約・翻訳にとどまらず, 様々な直列接続の SLP システムに適用可能である．本研究では, まず音声要約と音声翻訳の評価実験を通して提案手法の有効性を確認した．

1　はじめに

音声要約や音声翻訳など, 多くの音声言語処理（Spoken language processing：SLP）は, 音声認識（Automatic speech recognition：ASR）モデルと自然言語処理モデル（Natural language processing：NLP）の直列接続で実現されている．このような直列接続のシステムの学習には, 音声と出力テキストの大規模なペアデータを必要とせず, 個々の ASR モデルと NLP モデルに最先端のモデルを採用することができる．一方, 2 つの異なるモデルの学習と推論が独立して行われるため, 音声認識誤りが後段の NLP モデルの推論に悪影響を与え, SLP システム全体の精度が低下する．

このような認識誤り伝搬の問題に対し, 先行研究では認識誤りに対して頑健になるような拡張を NLP モデルに行い, 再学習することで認識誤り伝搬の影響を低減した [1–5]．これらの先行研究では, 1 つの ASR から複数の認識仮説を出力させ, ASR 結果の信頼度などの補助情報を抽出し, 信頼度を基に複数の認識仮説の統合を行う音声翻訳を提案している．また我々も先行研究において, Bidirectional Encoder Representations from Transformers（BERT）[6] の分散表現に基づいた音声認識仮説の統合に関する手法 [7] を提案し, 複数の認識仮説が音声要約において, 音声認識誤りに対する頑健性を向上させることを確認した．

しかし, これらの先行研究では, 単一の ASR モデルを使い複数仮説を生成しているため, 認識仮説のバリエーションが乏しい．これは, N-best デコーディングが多様な結果を出力しないことに起因する．さらに, [3–5] では, ASR モデルから補助情報を抽出して, NLP モデルで再学習するため, ASR モデルと NLP モデルで同じ単語辞書を使用しなければならないという制約がある．これは, 従来の直列接続のメリットであった, ASR と NLP モデルの独立性を損なうもので, ASR モデルまた NLP モデル, もしくは両方の性能を低下させる．そのため, 最もよい性能の ASR モデルと言語処理モデルを利用した直列接続でシステムを構成できず, SLP モデルの再学習を行う際に, 初期状態が最適なものではないという問題がある．

次に, 音声認識誤りの影響を低減させる方法として, 音声認識精度を改善し認識誤りそのものを減らす方法がある．先行研究では, 複数の ASR システムを統合することで音声認識精度を改善できることが報告されている [8–11]．これらシステム統合法では, 異なる ASR モデルから出力された仮説を比較することで, 認識誤りの位置を推定し修正することが出来る．本研究では, それらシステム統合法の中で, 最もよく使われる Recognizer output 投票 error reduction (ROVER) [10] に着目する．ROVER は ASR の後処理として広く利用され, アラインメントと投票の 2 段階の処理を通して認識仮説を統合する [12–16]．ROVER では, 音声認識結果を単語単位で比較・修正するため, ASR モデルと NLP モデルに対して特別な

拡張を必要とせず, 直列接続の独立性を損なわない.

しかし, ROVER にもいくつかの問題がある. まず, アラインメントの処理は, 単語単位の動的計画法で行われるため単語同士の類似性や文脈情報を考慮できない. 次に, 投票の処理では, 各システムの認識結果を平等に扱うため, 多くの認識器が間違えた中, 1 つの認識器が正解を出力するような場合は, 正しく修正できない. さらに直列接続の SLP システムでは, 入力音声を一字一句正しく認識した結果が必ずしも良い出力につながらないという問題ある [17]. 具体的には, ASR モデルはフィラーや言い間違い, 言いよどみなども正しく認識しようとする. しかし, これらは翻訳や要約には不要で, NLP モデルの精度を下げる一因となる. そのため SLP システムでは, 最終的な要約・翻訳精度を最大化するような認識結果を用いる必要があるが, ROVER は学習可能な重みを持たないため, 特定のタスクに対して最適化することができない.

本研究では, ROVER に代わる方法として, 複数の ASR モデルの出力結果を NLP モデル内部で統合する方法を提案する. 提案手法では, 注意機構を用いて, 複数の ASR モデルの出力結果をアラインメントし統合する処理を NLP モデルの分散表現に基づいて行う. 注意機構は学習可能な重みをもち, 特定のタスクに最適なシステム統合を学習することが出来る. 提案法は, 直近の我々の研究 [7] から派生し, 階層的注意機構 [18, 19] やマルチストリームコンビネーションの研究 [20], マルチモーダル処理の研究 [21] とも関連する. 提案手法では, 複数の ASR モデルの出力結果をマルチストリームとみなし, 注意機構を用いて統合する.

本研究では, 音声翻訳と音声要約において提案手法を ROVER などの先行研究 [3, 4, 10, 22] と比較した. 翻訳と要約では解く問題や, 扱うコンテキスト情報の長さが異なる. 例えば, 翻訳では, 多言語間の複雑な単語の対応関係を学習する必要があるが, 一発話単位で翻訳が行われるため, 複数文にまたがる依存関係を考慮しない. 一方, 要約は同一言語間で変換を行うため, 単語の対応関係の学習は容易であるが, 文章全体の要点を見つけるために, 複数文にまたがる依存関係を学習する必要がある. 異なる 2 つのタスクに対して, 本実験では両タスクに共通の Conformer ASR モデル [23] を, 翻訳タスクには Transformer モデル [24] を, 要約タスクには BERTSum モデル [25] をそれぞれ採用して SLP システムを構築した. 提案手法は, このような異なるタスクとモデル構造の SLP システムに対して, 最適なシステム統合を行い, 両タスクにおいて性能を改善することができた.

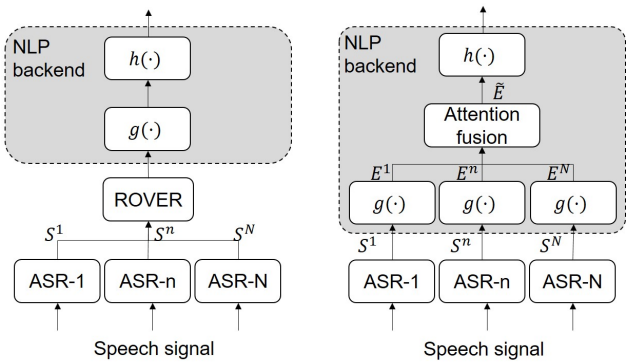

図 1 ROVER と 注意機構を用いた仮説統合の比較. 図内において, 注意機構を用いた仮説統合モジュールを Attention Fusion と表記する.

2 ASR 仮説の統合

2.1 従来の ASR 仮説統合

複数の ASR 仮説の統合し認識精度を改善する研究としては, Lattice を使った方法や, Minimum Bayes risk decoding を使った方法などが提案されている [8, 9, 11]. 本研究では, その中で最も一般的な手法である ROVER に着目した. ROVER [10] はアラインメントと投票の 2 段階の処理を通して, 異なる長さの認識仮説を統合する. まずアラインメントステップでは, 動的計画法を用いて挿入, 削除, 置換を追加した Word Transition Network を作成し, 投票ステップでは, 各時間において最も出現頻度も高い認識仮説を正解として採用する.

ROVER は ASR の後処理として行われ, 認識結果の単語列を直接比較するため, 様々な SLP システムに取り入れることが出来る. しかし, 単語同士の類似度や文脈情報などの有益な情報を考慮できず, また学習可能な重みも持たない. そのため, 各タスクにおいて, 単語の重要度を考慮して ASR システム統合を行い, SLP システムの性能を最大化するというような最適化ができない. 本研究では, これらの ROVER の持つ問題を解決するため, 注意機構を用いた ASR システムの統合法を紹介する.

2.2 ASR 仮説の分散表現

ASR 仮説の統合は, 単語の類似度, 文脈情報などを考慮するため, NLP の分散表現に基づいた統合を行う．ここで入力単語列 $S^n \in \mathbb{R}^L$ に対する NLP モデルの最初の数層の出力結果を $E^n \in \mathbb{R}^{L\times D}$ とする．

$$E^n = g(S^n), \tag{1}$$

ここで, $g(\cdot)$ は NLP モデルの最初の数層の写像処理, D は隠れ層の次元数を表す．ASR システムの総数は N であり, S^n は n 個目の ASR が出力した認識結果を表す．各 S^n の長さが異なるため, 事前に最後尾に Mask トークンで Padding を行い各系列を同じ長さ L に揃える．提案手法と ROVER 法の違いを図 1 に示す．図 1 中の $h(\cdot)$ は $g(\cdot)$ 以降の隠れ層の処理を表す．ROVER は ASR モデルと NLP モデルのデータ受け渡し間に仮説の統合を行うのに対して, 提案手法では, NLP モデルの Encoder 内部で ASR 仮説の統合を行う．

2.3 注意機構による仮説アライメント

まず, ROVER のアラインメントに相当する処理として, 本研究では注意機構を用いて入力された各 ASR 出力系列の位置合わせを行う．まず, SLP システムの再学習時に, 検証データを元に各 ASR モデルの性能を計測し, 最も性能が高い出力 ASR モデルを同定し, その出力系列を参照仮説 E^r とする．注意機構の Query には参照仮説 E^r を Key と Value には任意の ASR システムの出力仮説 E^n を用いる．この時は, 参照仮説 E^r は任意の認識仮説 E^n にも含まれる．

注意機構を用いたアラインメント結果 $\tilde{E}^n$ は以下のように得られ,

$$\tilde{E}^n = \mathrm{softmax}\left((E^r W^Q)(E^n W^K)^{\mathrm{T}}\right) E^n W^V, \tag{2}$$

ここで, T は転置記号を表す．提案手法は ROVER と比べた際, ソフトアライメントを実現し, Encoder で学習された文脈情報や単語の類似度を考慮できるため, 単語のみでなくフレーズなどより長い単位を考慮したアライメントを実現可能である．

注意機構には Multi-head-attention を採用し, 注意機構は各 Query, Key, Value に対し学習可能な重み $W^Q, W^K, W^V \in \mathbb{R}^{D\times D'}$ を持つ．本研究では, 注意機構の入力次元数 D' と出力次元数 D は同じであり, 注意機構の重みは正方行列となる．これらの重みを単位行列で初期化する．その理由は, 複数の ASR システム統合を行う際に, ランダムな写像を行うと BERT などの事前に学習された分散表現を十分に考慮できないために, 学習開始時点では, 注意機構において恒常写像が行われ, 参照仮説が選択されるよう意図したためである．これにより, 事前学習モデルの性能を大きく劣化させずに仮説統合の学習を行うことが出来る．

2.4 注意機構による仮説の統合

次に, ROVER の投票に相当するステップでは, 階層的注意機構 [18, 26, 27] と同様にアラインメントされた各系列の位置 l ごとに異なる N 個の ASR モデルから出力された仮説を統合する．$\tilde{E} \in \mathbb{R}^{N\times L\times D}$ はアラインメント済みの N 個の長さ L の音声認識結果の系列である．この時, 位置 l の $E_l \in \mathbb{R}^{N\times D}$ について, 注意機構を用いた統合を以下のようにあらわす．

$$\alpha_l = \mathrm{softmax}\left((e_l^r)^{\mathrm{T}} W^Q \tilde{E}_l\right), \tag{3}$$

$$e_l^{\mathrm{att}} = \alpha_l \tilde{E}_l^{\mathrm{T}} \tag{4}$$

ここで, $\alpha_l \in \mathbb{R}^{1\times N}$ は各 ASR システムの出力に対する注意重みである．提案手法と ROVER との違いは, 提案手法では正解単語を出現頻度に基づいて仮説単語から選択するのではなく, 特徴空間上で参照仮説と他の仮説の類似度を計算し重みとして, すべての仮説を重ね合わせて正解単語を表現する点である．これにより, 同時に複数の単語の情報を考慮することが可能となる．

3 実験

本研究では, 音声要約・翻訳タスクについて関連する先行研究と提案手法の比較を行った．データセットには, TED Talk から作成した要約コーパスである TEDSumarry [7] と Youtube のビデオから作成した HOW2 [28] を採用し, ROUGE [29]・BLEU [30] スコアでの客観評価を実施した．本実験における ASR モデルは ESPnet[1] で公開されている Tedlium/HOW2 のレシピに従い構築した．Tedliium は TEDTalk から作成した音声認識コーパスであり, TEDSumarry 用の ASR モデルの学習に用いられる．要約・翻訳モデルは OpenNMT[2] をもとに公開されているレシピ・実装[3] を用いて構築した．

本実験では, Topline として正解の音声認識結果を翻訳・要約した結果 **(0) ASR-GT** を採用し, Baseline

1） https://github.com/espnet
2） https://opennmt.net/
3） https://github.com/nlpyang/BertSum

として, 評価セットで最も性能の良い ASR システムを用いた結果を **(1) ASR w/ASR BPE**, BERT と同じ辞書を使用した ASR システムを用いた結果 **(3) ASR w/BERT BPE**, ROVER ですべての ASR システムの結果を統合し, 翻訳・要約した結果を **(3) ROVER** に示す．また, 追加で再学習を行う手法として, 音声認識結果で NLP を再学習したシステムである **(4) Retrain** と, 先行研究である **(5) Confidence [22]** と我々の先行研究である **(6) Nbest fusion [7]** を用意し比較した．

表 1 ROUGE スコア (R1, R2, RL) での各システムの比較結果: システム (7) が提案手法

システム	TED			HOW2		
	R1	R2	RL	R1	R2	RL
(0) ASR-GT	32.1	6.2	19.0	56.5	37.8	59.3
(1) ASR w/ASR BPE	29.9	**6.9**	18.3	47.4	27.1	46.1
(2) ASR w/BERT BPE	28.9	6.2	17.8	45.3	26.8	45.0
(3) ROVER	29.9	5.8	19.2	48.0	27.3	47.1
(4) Retrain	31.5	5.6	**20.4**	47.2	27.0	45.6
(5) Confidence [22]	30.1	6.8	**20.4**	48.4	**29.0**	47.3
(6) Nbest fusion [7]	**31.9**	6.0	19.3	49.3	28.8	48.2
(7) System fusion	**31.9**	6.1	19.0	**50.1**	**29.0**	**48.3**

表 1 の結果より, 提案手法は先行研究の **(5) Confidence** および, 我々の先行研究 **(6) Nbest fusion** よりも HOW2 データにおいて良い要約性能を示した．一方で TEDSumarry データにおいては, HOW2 データほど顕著な優位性を確認できていない．これは, TEDSumarry では正解の音声認識を用いる場合でも, 要約が困難で Topline の性能が低く, Baseline との差が少ない．そのため, 性能改善の余地が少ないことが一つの原因だと考えられる．Topline と Baseline の差が少ない理由として, 表 2 に示すように, TEDTalk での ASR 性能が高く認識誤りが少ないことが考えられる．

次に, 音声翻訳の性能を同様のデータを用いて評価した結果を表 3 に示す．ここで, 翻訳モデルと同じ辞書を使用した ASR システムを用いた結果 **(3) ASR w/MT BPE** を示す．表 3 の結果において, 両

表 2 各 BPE サイズと音声認識性能の単語誤り率による評価． 30.5k* は BERT モデルの単語辞書のサイズ. ROVER は, 各 ASR システムを ROVER で統合した際の認識性能

	BPE サイズ						
	best		2-nd	3-th	4-th	5-th	
	500	5k	10k	20k	30k	30.5k*	ROVER
HOW2	n/a	13.0	13.6	14.1	14.3	14.6	12.2
TED	8.5	n/a	8.7	9.5	10.0	10.4	8.3

表 3 BLUE スコアによる各音声翻訳システムの評価. システム (7) が提案手法.

	システム	TED (En-De)	HOW2 (En-Pt)
(0)	ASR-GT	27.2	55.2
(1)	ASR w/ASR BPE	24.3	44.6
(2)	ASR w/MT BPE	24.1	44.8
(3)	ROVER	24.4	45.1
(4)	Retrain	24.0	45.5
(5)	Posterior [4]	25.1	45.8
(6)	Nbest fusion [7]	25.1	45.6
(7)	System fusion	**25.4**	**46.0**

データに対して提案手法の優位性が確認できた．翻訳においても, TED よりも HOW2 データでの Topline と Baseline の差が大きく, 全体的にスコアの差が大きい．これは前述の音声認識性能の差と, 英語（En）からポルトガル語（Pt）への翻訳が, 英語からドイツ語（De）へのよりも容易であるためだと考えられる．よってこれらの結果から提案手法について次のようなことが言える．

提案手法は, 複数の ASR システムを NLP 内部で統合することで, 様々なタスク・データにおいて音声認識誤りで劣化する翻訳・要約性能を回復できる．しかし, ASR モデルの性能が高く, NLP モデルの性能が低いケースでは, 音声認識誤りが SLP システムに及ぼす影響が少なく, 精度改善の余地が少ない．加えて提案法は NLP モデルの分散表現に基づいて仮説統合を行うため, NLP モデルの性能が低い場合は, NLP モデルの分散表現が十分に学習されておらず, NLP の分散表現を基に仮説を統合する提案法は効果が小さく, 他の手法との優位な差が出にくいこと考えられる．

4 まとめ

本研究では, ASR と NLP を組み合わせて実現される SLP システムにおいて, 音声認識誤りによる性能劣化を改善する手法を提案した．本提案手法は, ROVER を参考に複数の ASR の出力結果を, NLP の分散表現に基づいて統合する．これにより, 従来 ROVER では難しかった, 単語の類似度や文脈情報を考慮した仮説統合を実現し, 音声要約・翻訳, 双方のタスクで既存手法を上回る音声認識誤りへの頑健性を示した．今後は, 注意機構による Alimento と統合方法の改善．他の音声言語処理タスクへの適応などを検討していく．

参考文献

[1] Nicola Bertoldi, Richard Zens, and Marcello Federico. Speech translation by confusion network decoding. In **ICASSP**, pp. 1297–1300, 2007.

[2] Masaya Ohgushi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. An empirical comparison of joint optimization techniques for speech translation. In **INTERSPEECH**, pp. 2619–2623, 2013.

[3] Kaho Osamura, Takatomo Kano, Sakti Sakriani, Katsuhito Sudoh, and Satoshi Nakamura. Using spoken word posterior features in neural machine translation. In **IWSLT**, 2018.

[4] Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. Tight integrated end-to-end training for cascaded speech translation. In **SLT**, pp. 950–957, 2021.

[5] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Neural lattice-to-sequence models for uncertain inputs. In **IWSLT**, pp. 1380–1389, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186, 2019.

[7] Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. Attention-based multi-hypothesis fusion for speech summarization. In **ASRU**, pp. 487–494, 2021.

[8] Gunnar Evermann and Woodland Philip. Posterior probability decoding, confidence estimation and system combination. In **Speech Transcription Workshop**, Vol. 27, pp. 78–81, 2000.

[9] Björn Hoffmeister, Dustin Hillard, Stefan Hahn, Ralf Schlüter, Mari Ostendorf, and Hermann Ney. Cross-site and intra-site ASR system combination: Comparisons on lattice and 1-best methods. In **ICASSP**, pp. 1145–1148, 2007.

[10] Jonathan Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In **ASRU**, pp. 347–354, 1997.

[11] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu. Minimum bayes risk decoding and system combination based on a recursion for edit distance. **Comput. Speech Lang.**, Vol. 25, No. 4, pp. 802–828, 2011.

[12] Olivier Siohan, Bhuvana Ramabhadran, and Brian Kingsbury. Constructing ensembles of asr systems using randomized decision trees. In **ICASSP**, Vol. 1, pp. 197–200, 2005.

[13] Venkata Ramana Rao Gadde, Andreas Stolcke, Dimitra Vergyri, Jing Zheng, M. Kemal Sönmez, and Anand Venkataraman. Building an ASR system for noisy environments: Sri's 2001 SPINE evaluation system. In **INTERSPEECH**, 2002.

[14] Yulia Tsvetkov, Florian Metze, and Chris Dyer. Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In **EACL**, pp. 616–625, 2014.

[15] Arne Mauser, Richard Zens, Evgeny Matusov, Sasa Hasan, and Hermann Ney. The RWTH statistical machine translation system for the IWSLT 2006 evaluation. In **IWSLT**, pp. 103–110, 2006.

[16] Christian Fügen, Alex Waibel, and Muntsin Kolss. Simultaneous translation of lectures and speeches. **Mach. Transl.**, pp. 209–252, 2007.

[17] Xiaodong He, Li Deng, and Alex Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In **ICASSP**, pp. 5632–5635, 2011.

[18] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In **NAACL-HLT**, pp. 1480–1489, 2016.

[19] Jindrich Libovický and Jindrich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In **ACL**, pp. 196–202, 2017.

[20] Xiaofei Wang, Ruizhi Li, Sri Harish Mallidi, Takaaki Hori, Shinji Watanabe, and Hynek Hermansky. Stream attention-based multi-array end-to-end speech recognition. In **ICASSP**, pp. 7105–7109, 2019.

[21] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In **ICCV**, pp. 4203–4212, 2017.

[22] Shi-Yan Weng, Tien-Hong Lo, and Berlin Chen. An effective contextual language modeling framework for speech summarization with augmented features. In **EUSIPCO**, pp. 316–320, 2020.

[23] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In **Interspeech**, pp. 5036–5040. ISCA, 2020.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, pp. 5998–6008, 2017.

[25] Alexandra Savelieva, Bryan Au-Yeung, and Vasanth Ramani. Abstractive summarization of spoken and written instructions with BERT. In **SIGKDD**, Vol. 2666 of **CEUR Workshop Proceedings**, 2020.

[26] Potsawee Manakul, Mark J. F. Gales, and Linlin Wang. Abstractive spoken document summarization using hierarchical model with multi-stage attention diversity optimization. In **INTERSPEECH**, pp. 4248–4252, 2020.

[27] Tzu-En Liu, Shih-Hung Liu, and Berlin Chen. A hierarchical neural summarization framework for spoken documents. In **ICASSP**, pp. 7185–7189, 2019.

[28] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. **CoRR**, Vol. abs/1811.00347, , 2018.

[29] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **ACL**, pp. 74–81, 2004.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.