5Q-07

クラウドソーシングを用いて作成した環境音キャプションコーパスの評価

岩月道生 †, 糸山克寿 †, 西田健次 †, 中臺一博 †,‡
† 東京工業大学工学院システム制御系
‡ ホンダ・リサーチ・インスティチュート・ジャパン

1 はじめに

システムが周囲の環境を理解する「環境理解」研究において，音響信号は，映像信号と並び重要な入力情報である．映像信号は主に空間的な関係性を抽出する際に用いられるのに対し，音響信号は時系列信号であることから，時間的な関係性を抽出する上で特に重要な情報である．我々は，このことを考慮して，音イベントの順序性に着目して，音響シーンに対する音響キャプション生成に取り組んでいる [1,2]．また，こうした研究を進めるにあたり，適切な学習用コーパスが存在しないことも問題であり，この解決のために，音イベントの順序性に着目した音響キャプション用コーパスの設計も行った [2]．本稿では，設計指針に則って構築したコーパスの評価について報告する．

2 音響キャプション用コーパス

既存の音響キャプション研究向けコーパスでは，ある程度まとまったシーン全体をキャプショニングすることを想定しているため，数十秒の音響信号全体に対する，キャプションが一文で記述されていることが多く，シーン内の音イベントの時間的な関係性は記述されていない [3,4]．また，特定のシーンに偏ったコーパスとなっているものも存在する [3]．一方，車や楽器といった音イベントごと切り出された音源識別用のコーパスも複数存在している [5]．これらから選択した音イベントを適宜つなぎ合わせてコーパスとして使用する報告もある [1] が，こうしたデータセットでは，実環境評価が難しいという問題がある．そこで，音イベントの順序性に着目した音響キャプション用コーパスを以下の指針に基づいて構築した [2]．

- 音イベント単位のアノテーション：
  一つの音響信号に対し，音イベント単位で時系列順にアノテーションを実施．得られたキャプションを以降では「キャプション列」と呼ぶものとする．
- 多様なシーンをカバー：
  アノテーション対象の音響信号クリップには MSR-VTT [6] に含まれる動画の音響信号を使用し，特定のシーンに限定しない，多様な状況で収録された音を使用．
- アノテーション精度の向上：
  アノテーターに音響信号に対応した映像信号を同時に提示．
- 構築コストの低減：
  クラウドソーシング (Amazon Mechanical Turk) を用いたアノテーションの実施

結果として， 798 個の音響信号クリップに対して，6,322 個の英語のキャプション列が含まれるコーパスが構築された．

3 コーパスの評価

構築したコーパスを評価するために，研究を進めている音響信号キャプション生成モデル [1] を構築したコーパスで学習し，性能評価を行った．比較のために，同様に音響キャプション研究を行っている Wu らの作成したコーパス [3] でも同じ音響信号キャプション生成モデルを用いて，性能評価を行った．なお，Wu らのコーパスは病院内の音響シーンのみを対象としたもので，中国語で行ったアノテーションを英語に機械翻訳で変換している．3,707 個のクリップに対して，11,107 個のキャプション（列）からなっている．以降，本稿では今回岩月らが構築した評価対象のコーパスを IC，Wu らの構築した比較対象のコーパスを WC と呼ぶ．

3.1 本コーパスの分割

1 つのコーパスで学習と評価を行うために，IC をトレーニングデータセット，評価データセット，テストデータセットに分割したデータセットを作成した．コーパスの分割は，音響信号クリップに対して行い，異なる 2 つのデータセット間には，同じクリップに対するキャプションが含まれないようにした．また，学習時や評価時にデータの偏りによる評価結果の偏りを防ぐ

Evaluation of Environmental Sound Caption Corpus using Crowdsourcing
Michio Iwatsuki†, Katsutoshi Itoyoma†, Kenji Nishida†, Kazuhiro Nakadai†,‡
† Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
‡ Honda Research Institute Japan Co., Ltd.

Table 1　出力されたキャプションの例

IC	man speaking . foot steps on tile . birds chirping . crowd cheering . ball being hit . shoes squeaking on court .
WC	The doctor is talking to the patient . Voice The doctor is talking to the family members of the patient and his family .

Table 2　評価セットとテストセットでの BLEU

	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$
評価セット	0.363	0.284	0.228	0.197
テストセット	0.364	0.282	0.224	0.193

Table 3　出力されたキャプションの種類数

	音響信号クリップ数	出力キャプションの全種類数
IC	158	93
WC	371	24

ため，3 つのデータセットに含まれる単語の割合が同等になるよう分割した．WC についてはすでに 3 つのセットに分割されているので，そのセットをそのまま使用した．

3.2　評価に使用した深層学習モデル

使用した深層学習モデルは，エンコーダ･デコーダモデルに基づいており，エンコーダ部，デコーダ部には，それぞれ 1 つずつ Long-Short Term Memory (LSTM) [7] を用いている．また，エンコーダとデコーダの隠れ状態の間にアテンション [8] を加えたモデルとなっている．IC，および WC のそれぞれで学習を行い，2 つのモデルを得た．IC を用いる場合は，キャプション列をピリオドでつなげて 1 文にすることによって，順序を表現した．

3.3　結果と考察

Table 2 に構築したコーパスで学習したモデルに対し，評価データセットとテストデータセットを用いた場合の BLEU [9] 値を示す．BLEU は機械翻訳などの分野でモデルの評価に一般的に用いられる指標であり，値が大きいほど良いモデルであるといえる．2 つのコーパスの BLEU 値が十分近いことから，IC は過学習していない，つまり評価に十分な規模を持っているといえる．Table 1 に学習した 2 つのモデルのそれぞれのテストデータセットでの出力結果サンプルを示す．一つの行が一つの出力サンプルに対応する．また，Table 3 にそれぞれのテストデータセットの音響信号クリップ数と，そのセットから出力されたキャプション文の全種類数を示した．WC で学習したモデルの出力は，音響信号全体の要約的なキャプションを生成しているが，出力されるキャプションの種類が少ないことから，音イベントの順序など音響シーンの差異をある程度吸収したキャプションとなってしまっているといえる．一方で IC で学習したモデルの出力は音の順序が正しく表現されておりバリエーションが多い．

4　まとめ

本稿では，音イベントの順序に着目して構築した音響キャプション生成用のコーパスを，研究開発中のエンコーダ・デコーダベースの音響キャプション生成モデルも用いて，評価を行った．先行研究である Wu らの作成したコーパスで学習したモデルと比較することにより，構築したコーパスが音イベントの順序により注目したキャプション生成に適していることを示した．

今後は，生成されるキャプションのクオリティを上げるため，構築したコーパスの拡充，ならびに音響キャプション生成モデルの改良を行う予定である．

謝辞　本研究は JSPS 科研費 16H02884, 17K00365 および 19K12017 の助成を受けた．

参考文献

[1] 岩月 ほか，“Listen and Tell: 深層学習を用いた音響シーンのキャプション生成”，IPSJ2019，6T-3，2019.

[2] 岩月 ほか，“音環境説明ロボットの実現に向けた環境音キャプションコーパスの構築”，RSJ2019, 2I1-05, 2019.

[3] M. Wu *et al.*，“Audio Caption: Listen and Tell.”, ICASSP2019, 830–834, 2019.

[4] C. D. Kim *et al.*，“AudioCaps: Generating Captions for Audios in the Wild.”, NAACL-HLT 2019, 119–132, 2019.

[5] E. Fonseca *et al.*，“General-purpose Audio with AudioSet Labels: Task Description, Dataset, and Baseline.”, DCASE2018, 69–73, 2018.

[6] J. Xu *et al.*，“MSR-VTT: A Large Video Description Dataset for Bridging Video and Language.”, CVPR2016, 5288–5296, 2016.

[7] F. A. Gers *et al.*，“Learning to Forget: Continual Prediction with LSTM.”, *Neural Comput*, 12(10), 2451–2471, 2000.

[8] D. Bahdanau *et al.*，“Neural Machine Translation by Jointly Learning to Align and Translate.”, ICLR2015.

[9] K. Papineni *et al.*，“BLEU: A Method for Automatic Evaluation of Machine Translation.”, ACL2002, 311-318, 2002.