動画からの動作キャプション生成における書き換え技術の適用

平川 幸司　　小林 哲則　　林 良彦
早稲田大学理工学術院
hirakawa@pcl.cs.waseda.ac.jp

1 はじめに

動画中で描写される**人間の動作を簡潔に表す動作キャプション**を生成する手法を提案する．具体的には，動作に限らない一般的な状況を描写するキャプション生成器の存在を仮定し，生成されたキャプション[1]を動作キャプションへと変換する書き換え器を学習する．

動画の特定の区間に対してキャプションを付与する研究として Dense-captioning events [1] が知られている．これにより区間に対して生成されるキャプションは，一般的に一文で表されるが，動作以外の状況描写を含み，複文の構造を持つことが多い．これに対し，我々が想定する動作キャプションは，人間の動作を単文により簡潔に表すものである．以上から，一般的なキャプションから動作キャプションへの書き換えは，動作に関係ない部分を排除しつつ，一連の動作のそれぞれを単文として表すタスクとなる．

本研究では，Seq2Seq モデルによる書き換えの学習を行うが，得られる学習データが当面は少量であることに鑑み，類似したタスクである Split and Rephrase [2] におけるデータを援用する．ただし，このデータは本研究から見ればドメイン外であるため，ドメイン適応 [3] を行うことが必要となる．また，書き換えにおいては，用いられる単語を変更する必要は少ないことを考慮し，コピー機構 [4] を利用する．本論文では，評価実験の結果からこれらの手法の有効性を確認する．

2 動作キャプションとその生成

2.1 動作キャプション

動画に対して動作キャプションを適切に付与することができれば，TRECVID Ad-hoc Video Search [5] で多く要求されるような，人間の動作を含む動画の自然言語クエリによる検索が可能となる．

[1]単にキャプションと書くときは一般的なキャプションを意味し，動作キャプションと区別する．

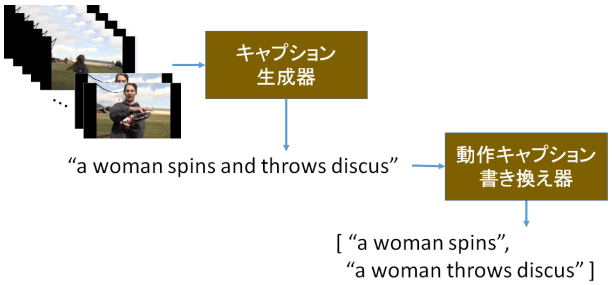

図 1: 書き換えによる動作キャプションの生成

本研究では図 1 に示すように，一般的なキャプションを適切に生成できるキャプション生成器の存在を仮定し，その出力を Seq2Seq モデルにより動作キャプションへ書き換えるアプローチをとる．本研究で想定する動作キャプションとは，人間が動作を行っている動画中の区間に対して，その動作を簡潔に記述するものであり，次のような言語特徴を持つものとして規定される．

- 単文の系列
- 各単文は人間が主語で，動詞は現在形もしくは現在進行形
- 単文はそれぞれが独立

ここで「それぞれが独立」ということは，生成される単文同士に前後関係はなく書き換えたキャプションに代名詞を用いず，繰り返される単語の省略を行なわない，ということを意味する．具体的には”he”や”she”などの代名詞を ”man” や ”woman”などの一般的な表現に置き換える．

本研究では，ActivityNet Captions [1] のキャプションデータをこの基準に合致するように書き換えることで書き換えの学習データを作成する．具体的には，分割された単文から動作を含まないキャプション (例: ”We see a title card after a shot.”) を除き，時制，主語をそろえる，という方法をとった．また，書き換えの際には原文に出現する表現をなるべく残すようにし

た．ActivityNet Captions のデータには動作記述を含まないキャプションも存在し，その場合は文の除去を表す記号”None”を書き換え後の表現とした．現在までに 8,172 件のキャプションに対するデータを作成した．

2.2 Seq2Seq による動作キャプション生成

基本的には，上記の方法で作成した学習データを用いて，Seq2Seq モデルによる動作キャプションへの書き換えを学習するが，(1) 学習データの量が十分ではないこと，(2) これと関係して，学習データに含まれる多くの単語がテスト時に出現しない Out-of-vocabulary (OOV) の問題が生じること，(3) タスクの性質上，書き換え前の単語の多くは書き換えにおいて保存されるべきであること，などを考慮する必要がある．

3 提案手法

学習データ量の問題に関しては，類似したタスクである Split and Rephrase [2] におけるデータを援用することで，学習データの増量を図る．ただし，このデータは本研究から見ればドメイン外であるため，ドメイン適応 [3] を行う．さらに，OOV の問題に対処し，また，本書き換えタスクでは用いられる単語を変更する必要が少ないことを考慮し，コピー機構 [4] を利用する．これらの手法を導入することにより，一般キャプションから動作キャプションへの書き換えの精度向上を狙う．

3.1 Split and Rephrase データの利用

Split and Rephrase [2] とは，情報の欠落がないように一つの複文を複数の単文に分割にするタスクである．本タスクのためのデータセット benchmark-v1.0[2]が公開されている．このデータセットは WebNLG data [3]に含まれる複文を単文に分割することで作成されており，データの規模は 886,857 件に達する．本研究における動作キャプションへの書き換えと Split and Rephrase を比較すると，後者では情報の欠落がないことを前提としているが，前者では入力となる一般キャプションから動作記述以外を取り除くことが必要となる．この点で，本研究の書き換えタスクは，Split and Rephrase タスクに類似しながらも，文圧縮 [6] の側面を持ち合わせている．

[2] https://github.com/shashiongithub/Split-and-Rephrase

[3] https://gitlab.com/shimorina/webnlg-dataset

3.2 ドメイン適応

機械翻訳の分野では，特定のドメインにおける機械翻訳の精度を向上させるために，対象とは異なるドメイン (out domain) における大量の対訳データを援用して翻訳モデルを学習し，これを対象ドメイン (in domain) における相対的に少量の対訳データによってドメイン適用させる研究が行われている [3]．本研究においては，まず in domain の動作キャプションのデータと out domain の Split and Rephrase のデータを合わせた全体で初期の学習を行い，次にこの学習で得られたパラメータを初期値として用い，in domain の動作キャプションの書換データのみで再学習 (ファインチューニング) を行う．

動作キャプションの生成においては，入力となるキャプションに含まれる動作記述の部分のみを単文として表現する必要があるため，書き換え結果として得られる単文の数は，Split and Rephrase の場合よりも多くなることはない．大量データによって単文系列への書き換えを学習し，動作キャプションのデータを用いた再学習によって，このような動作記述でない文のフィルタリングが達成されることを期待する．

3.3 コピー機構の採用

機械翻訳や要約生成において固有表現は OOV となりやすく，また，出現頻度が低い現言語の単語が全く異なる単語に置き換えられる場合がしばしば生じる．このような事態を避けるため，入力の特定の単語をそのまま出力へと引き渡すコピー機構 [4] の研究が行われてきた．本研究の書き換えタスクは同一言語 (英語) 内の変換であり，単語を置き換える必要性が低いこと，また，利用可能な in domain の学習データが当面は少量であることから，コピー機構が導入された Seq2Seq の要約生成モデルである Pointer-generator networks [7] を用いる．

4 評価実験

Out domain である Split and Rephrase タスクの学習データの援用，ドメイン適応，コピー機構の有効性を評価した．

表 1: 生成された動作キャプションの評価結果

S&R の文数	コピー機構	ドメイン適応	BLEU	ROUGE	総合評価値	WMD	#S/C
0	なし	なし	0.467	0.674	0.551	1.94	1.82
	あり	なし	0.565	0.774	0.653	1.79	1.85
6k	なし	なし	0.457	0.665	0.542	1.98	1.86
	あり	なし	0.576	0.804	0.671	1.95	1.68
	なし	あり	0.460	0.669	0.546	3.24	1.84
	あり	あり	0.592	0.801	0.681	1.74	**1.75**
60k	なし	なし	0.431	0.601	0.502	2.54	1.73
	あり	なし	0.252	0.612	0.357	**0.899**	2.82
	なし	あり	0.431	0.645	0.517	2.01	1.87
	あり	あり	**0.612**	**0.826**	**0.703**	1.84	1.60
300k	なし	なし	0.468	0.214	0.294	1.60	2.41
	あり	なし	0.329	0.480	0.391	2.31	2.15
	なし	あり	0.457	0.661	0.540	1.98	1.88
	あり	あり	0.547	0.776	0.642	1.76	1.50
880k (all)	なし	なし	0.453	0.195	0.272	1.671	3.31
	あり	なし	0.335	0.585	0.426	0.972	2.42
	なし	あり	0.450	0.655	0.534	3.18	1.74
	あり	あり	0.332	0.653	0.440	2.02	1.01
60k (test)	あり	あり	0.559	0.805	0.660	1.884	1.38

4.1 実験設定

これまでに準備した動作キャプションのデータ約 8,000 件を train:6,158 件，validation:1,014 件，test:1,000 件に分割し，学習・テストを行った．Split and Rephrase の学習データを援用することの有効性や，その際の適切なデータ量を調べるため，N 件 (N=0, 6k, 60k, 300k, 880k) のデータをランダムサンプリングし，動作キャプションの学習データに追加した．また，ドメイン適応を行う場合と行わない場合，コピー機構を用いる場合と用いない場合を比較する．

4.2 評価指標

評価指標には，BLEU (Precision 相当)，ROUGE (Recall 相当)，および，これらの調和平均による総合評価値 (F1 相当) を用いる．また，書き換え前後での意味的距離を Word Mover's Distance (WMD) [8] により評価する．WMD は，対応する単語同士の分散表現の L2 距離に基づいて文書間の意味的距離を定量化するもので，値が 0 に近いほど意味が近いとみなす．

さらに，1 キャプションを書き換えた後の動作キャプションの数 (#S/C) も調査する．test データの正解から求めた#S/C は 1.75 であり，生成される動作キャプションの数もこれに近いことが望ましい．なお，Split and Rephrase のデータにおける#S/C は，2.52 と報告されている．

4.3 主要な実験結果と考察

表 1 に validation データ，および，最良と考えられる設定で test データ (最下行) に対して生成した動作キャプションに対する各評価指標を示す．

コピー機構を用い，ドメイン適応を行うことにより，Split and Rephrase のデータを 60,000 件程度まで増やしても BLEU，ROUGE，総合評価の各指標は向上する．しかし，それ以上増やすと各指標は低下する．これより，out domain のデータ量を適切に定めることが必要であること，また，コピー機構の果たす役割が大きいことが分かる．今回，学習に用いた動作キャプションのデータ数は約 6,000 件であるので，in domain データに対する適切な out domain のデータ量は 10 倍程度であると推定できる．なお，動作キャプションでは書き換え結果が存在しない，すなわち，入力キャプションが動作記述を含まない場合が存在する．これに対して，Split and Rephrase ではそのようなことがないため，このデータ量が増えると相対的に，動作記述でない文を排除する能力が低下すると考えられる．

書き換えによる意味の保存を評価するために WMD 指標を導入したが，表 1 の結果からは一定の傾向を見出すことは難しい．文に対する分散表現を利用して意味的類似度を求めるなどの改善が必要と考えられる．

生成された文数 (#S/C) を見ると，ドメイン適応を行うことにより，学習データに追加する Split and Rephrase の文数を増加させても，この指標を一定に

表 2: 出力される動作キャプションの例 (”あり/なし” などはコピー機構，ドメイン適応の有無を表す)

設定	出力結果
正解	a basketball player is holding a ball . a basketball player is running up . a basketball player is shooting a basket .
なし/なし	a basketball player is holding a ball . a basketball player is running up . a basketball is drinking a basket .
あり/なし	a basketball player is holding a ball . a basketball player is shooting a basket .
なし/あり	a basketball player is holding a ball . a basketball player is running up . a basketball is drinking a basket .
あり/あり	a basketball player are holding a ball . a basketball player is running up . a basketball player is shooting a basket .

抑えることができていることが分かる．これは，動作キャプションへの書き換えにおいて一定数現れる非動作記述の排除が狙いどおりに行えている可能性を示すが，詳細についてはさらに調査を行う必要がある．

表 2 に”an intro leads into a basketball player holding a ball and then running up and shooting a basket .” というキャプションに対して生成された動作キャプション例を示す．この例では，コピー機構の利用により，不適当な動詞 (”drinking”) を排除できた．

5 おわりに

本稿では，動作キャプション生成における書き換え技術の適用法を提案した．評価実験の結果から，(1) out domain であっても形式的に類似した書き換えタスク (Split and Rephrase) のデータを適切な量だけ利用し，(2) ドメイン適応とコピー機構の導入を行うことにより，生成する動作キャプションの精度向上に有効であることを確認した．今回は，out domain のデータからランダムにサンプリングしたデータを利用したが，適切なコーパスフィルタリングを行う [9, 10] ことにより，より効果的な学習が可能となると考えられる．このためには，動作キャプションで必要な文は人間の動作記述であることに着目し，学習データに加えるべき文を選択することが必要となる．

謝辞

本研究は JSPS 科研費 (17H01831) の助成を受けた．

参考文献

[1] R. Krishna *et al.*, Dense-captioning events in videos., in: ICCV, 2017, pp. 706–715.

[2] S. Narayan *et al.*, Split and rephrase, arXiv preprint arXiv:1707.06971.

[3] C. Chu *et al.*, An empirical comparison of domain adaptation methods for neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2017, pp. 385–391.

[4] J. Gu *et al.*, Incorporating copying mechanism in sequence-to-sequence learning, CoRR abs/1603.06393.

[5] G. Awad *et al.*, Trecvid 2017: Evaluating adhoc and instance video search, events detection, video captioning and hyperlinking, in: Proceedings of TRECVID 2017, NIST, USA, 2017.

[6] K. Filippova *et al.*, Sentence compression by deletion with lstms, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 360–368.

[7] A. See *et al.*, Get to the point: Summarization with pointer-generator networks, arXiv preprint arXiv:1704.04368.

[8] M. Kusner *et al.*, From word embeddings to document distances, in: International Conference on Machine Learning, 2015, pp. 957–966.

[9] A. Axelrod *et al.*, Domain adaptation via pseudo in-domain data selection, in: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, 2011, pp. 355–362.

[10] M. van der Wees *et al.*, Dynamic data selection for neural machine translation, CoRR abs/1708.00712.