VideoCLIP を用いた実験動画からのプロトコル生成

山本航輝 [1]　西村太一 [1]　亀甲博貴 [2]　森信介 [2]
[1] 京都大学大学院　情報学研究科　[2] 京都大学 学術情報メディアセンター
{yamamoto.koki.76n,nishimura.taichi.43x}@st.kyoto-u.ac.jp
{kameko,forest}@i.kyoto-u.ac.jp

概要

本研究では，研究における再現性向上の目的の下，生化学分野における一人称視点の実験映像からの，実験手順を表すプロトコルの自動生成に取り組んだ．プロトコルの生成において，フレームに映っている物体の名称に加え，VideoCLIP により推定した実験者の動作を動詞として利用することで，物体の名称のみを用いた場合よりも多くの情報を用いたプロトコル生成を行なった．その結果，正しい動詞を利用することでより適切なプロトコルを生成することができた．

1　はじめに

心理学や化学など様々な研究領域において，研究の再現性の低さが問題視されている．Baker の調査 [1] によれば，研究者約 1,500 人の内 70 %以上が他の研究者の実験結果を再現できなかった経験があると述べている．科学的研究結果を普遍的なものにするために，再現性を向上させることが求められている．

研究領域の中でもとりわけ，薬品や実験器具を扱う生化学実験の再現における重要な要素のひとつとしてプロトコルが挙げられる．プロトコルとは必要な器具や薬品の名称とともに実験の手順を時系列順に記した文書であり，実験を再現する研究者はプロトコルを参照することで, 実験手順や試薬の量といった実験の再現に必要な情報を得ることができる．プロトコルに記述漏れや誤りがある場合，研究者は誤った情報をもとに実験を再現することに繋がり，プロトコルを正しく記述することは実験の再現性に関わる大きな要因のひとつだと言える．

このようなプロトコルの記述漏れや誤りを防ぐ手段のひとつとして，自動的なプロトコルの作成が挙げられる．人間の手を介さずにプロトコルを記述することにより，ヒューマンエラーによる記述誤りを防ぐと同時に，人手によるプロトコル作成のコストを削減することができる．

本研究では自動的なプロトコル作成の一例として，視覚モデルと言語モデルを用いた，実験映像からのプロトコル生成に取り組む．本研究で提案する手法は西村らが発表したプロトコル生成手法 [2] を拡張したものである．西村らは生化学実験の実験映像中に現れる試薬や実験器具の名称を言語モデルに与え，プロトコルの生成を行なっていた．しかし，加える試薬の量や実験者の動作等といった視覚情報が与えられておらず，それらの情報だけでは十分とは言えない．本研究では，実験映像から実験者の動作を推定し，映像中の試薬や実験器具に加え，動作も加味することでより多くの情報を用いたプロトコル生成に取り組んだ．その結果，言語モデルに正しい動詞を与えることでより適切なプロトコルが生成できることを確認した．

2　BioVL2 データセット

本研究では，生化学分野における一人称視点の実験映像データセットである BioVL2 データセット [2] を用いてプロトコルの生成に取り組む．BioVL2 データセットは，生化学分野における基礎的な 4 種類の実験の様子を撮影した一人称視点の動画像データセットである．

BioVL2 データセットには，(1) 実験映像に対応するプロトコル，及び (2) 映像フレーム中に現れる試薬，実験器具のバウンディングボックスの 2 種類のアノテーションが与えられている．(1) に関して，プロトコルは各実験映像の実験手順を時系列順に記したものであり，それぞれの手順と，実験映像中の各手順に対応する開始時刻と終了時刻が付与されている．(2) に関して，実験映像から 4 秒ごとにフレームを切り出し，フレーム中に現れる試薬及び実験器具のバウンディングボックス及びそれらの名称が 4 秒毎のフレームに対して付与されている．

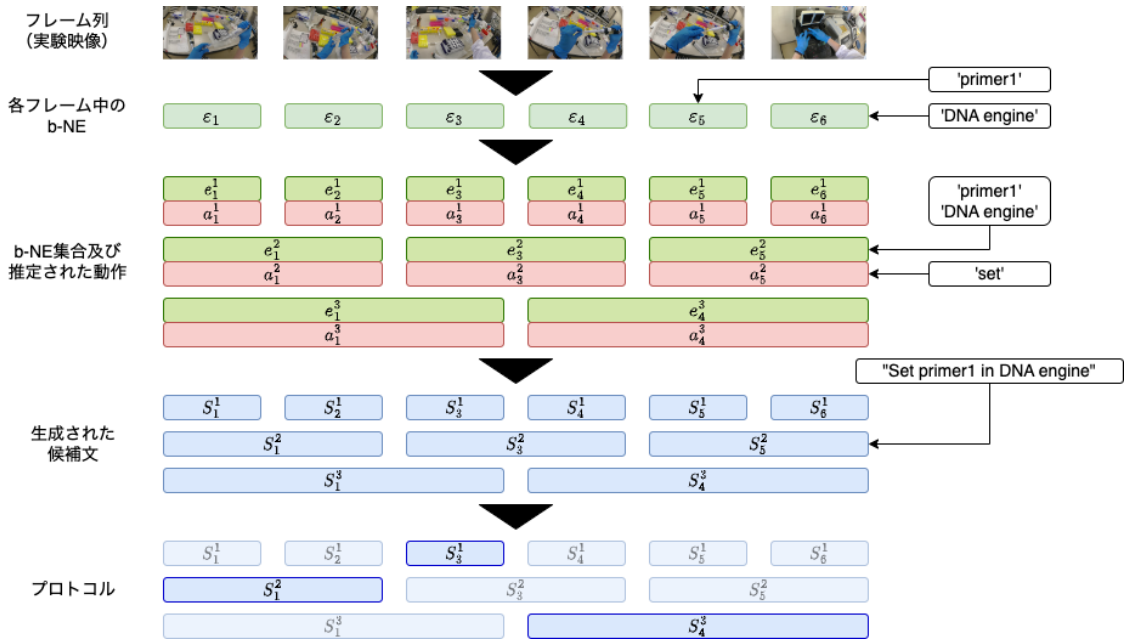

図 1　プロトコル生成の手順

3　実験映像からのプロトコル生成

本研究で扱う BioVL2 データセットは合計 3 時間ほどの動画像データセットであり，学習データ量の観点から，映像データを入力として直接プロトコルを生成するような深層学習モデルを構築することは難しい．そこで本研究ではプロトコルの生成手法として西村らの手法 [2] を拡張し，適用する．図 1 に手法の概要を示す．西村らの研究では, 牛久らの手法 [3] を基に，BioVL2 データセットを用いて以下の 4 つの手順により実験映像からプロトコルを生成する．一つの実験映像に対するプロトコル生成の手順を以下に示す．

1. 4 秒に 1 枚フレームを抽出し，あらかじめアノテーションされた物体名（bio-Named Entity, 以下 b-NE とする）を抽出する．
2. 連続するフレームの b-NE を統合し，b-NE 列を得る．図 1 の例では，5 番目のフレームと 6 番目のフレームから’primer1’ と’DNA engine’ という b-NE が得られ，b-NE 列 {’primer1’}，{’DNA engine’}，{’primer1’，’DNA engine’} を得る．
3. 得られた b-NE 列を，事前学習済みの文生成モデルに与え，手順の候補文とその尤もらしさを表すスコアを得る．
4. 手順 3 により得られた候補文の中から，候補文のスコアに基づいたビタビアルゴリズムにより，最も尤もらしい文の組み合わせを探索し，プロトコルとする．

西村らの手法の問題点は，手順 3 において実験映像中に現れる b-NE のみを文生成モデルに与えることで候補文の生成を行なっており，映像の視覚的情報を無視してしまっている点である．そこで本研究では上記の手順 3 を，「得られた b-NE 列，及び連続するフレーム間の実験映像から推定される動作を文生成モデルに与え，手順の候補文を生成する」という手順に変更した．

3.1　b-NE の抽出と統合

BioVL2 の実験映像では動画の 4 秒毎に，フレーム中に現れる b-NE がバウンディングボックスによりアノテーションされている．これらのアノテーションから，各フレームにおける b-NE 及び連続するフレーム列における b-NE 列を以下のようにして得る．i 番目のフレームを f_i，i 番目のフレームから連続する l 個のフレーム列を $\boldsymbol{f}_n^l = (f_n, f_{n+1}, ..., f_{n+(l-1)})$ とし，フレーム f_i が b-NE 集合 ε_i を持つとする．この時フレーム列 $\boldsymbol{f}_i^l$ が持つ b-NE 集合は $e_n^l \in \varepsilon_i \times \varepsilon_{i+1} \times ... \times \varepsilon_{i+(l-1)}$ と書ける（$\times$ は集合のデカルト積を表す）．西村らの手法と同様に，フレーム列の長さ l について $l = 1, 2, 3$ に対応する b-NE 列を獲得する．例として，図 1 のフレーム f_5, f_6 が b-NE としてそれぞれ’primer1’, ’DNA engine’ を持つ場合，フレーム列 $\boldsymbol{f}_5^2$ が持つ b-NE 集合は’primer1’, ’DNA engine’ となる．

3.2 実験映像からの動作推定

続いて，連続するフレーム列の間の実験映像から，実験者の動作を推定する．生化学分野における実験を対象とした大規模動画像データセットは存在せず，実験映像を入力とした動作推定を行うモデルを学習させることは困難である．よって本研究では動作推定を行うモデルとして，タスクに応じたモデルの再学習を必要とせず，ゼロショットで下流タスクを解くことのできる VideoCLIP [4] を用いる．

3.2.1 VideoCLIP

VideoCLIP は動画と言語の対応関係を学習した Vision-Language モデルである．VideoCLIP は HowTo100M [5] データセットを用いた対照的学習により事前学習されている．VideoCLIP は事前学習に用いていない様々な動画像データセットを対象とした Vision-Language タスクにおいて高い精度を記録している．

VideoCLIP や CLIP 等といった，大規模なデータセットで事前学習を行うことによりゼロショットでの下流タスク解決が可能な Vision-Language モデルにおいては，タスクに応じてプロンプトと呼ばれる入力テキストの形式を変えることが精度向上につながることが知られている [6]．タスクに応じたプロンプト最適化の具体例として，VideoCLIP を用いて動画中の動物が犬・猫のどちらであるかを分類するという 2 クラス分類タスクについて考える．この時，テンプレート文として “a video of {class}”という文をあらかじめ用意しておき，テンプレート文中の’{class}’ の部分を’dog’, ’cat’ という単語で置き換えることにより，それぞれのクラスに対応したプロンプトを得る．このようなプロンプトをテキスト入力として用いることで，’dog’ や’cat’ といったクラス名を直接テキスト入力に使用した場合よりも高い精度を記録できることが知られている．本研究で用いるプロンプトについては 4.2 で述べる．

3.2.2 VideoCLIP を用いた動作の推定

連続するフレーム列 $\boldsymbol{f}_n^l$ 間に対応する部分実験映像を $\boldsymbol{v}_n^l$ とし，i 番目の動作の候補となる単語を，あらかじめ用意したテンプレート文に埋め込んだプロンプトを $\boldsymbol{t}_i$ とする．$\boldsymbol{v}_n^l$ と $\boldsymbol{t}_i$ の類似度を VideoCLIP を用いて算出し，類似度の高い動詞の組 a_n^l を動作の候補として選出する．

3.3 b-NE 列及び動作からの候補文生成

続いて，選出された動詞の組 a_n^l と得られた b-NE 列 e_n^l を用いて実験手順の候補となる文を生成する．西村らの手法と同様，文生成モデルには WLP データセット [7] で事前学習した Transformer にコピー機構を加えたモデルを用いる．コピー機構をモデルに組み込むことで，入力として渡される b-NE 及び動詞を正しく出力に反映させることができる．b-NE 列 e_n^l と，VideoCLIP を用いて推定された動詞を Transformer に与え，候補文を生成する．

次の手順にて最適な候補文を探索するために，生成された候補文毎に，文の尤もらしさを以下の式によりスコアとして計算しておく．

$$\mathrm{Score}(e_n^l) = \prod_{i=1}^{N} p(d_i | d_1, d_2, ..., d_{k-1}; e_i^l)$$

ここで d_i は出力文の i 番目の単語，N は単語列の長さを表す．

3.4 候補文を用いたプロトコルの選定

最後に，生成された候補文の中からスコアが最も高い文の系列，すなわちプロトコルを構成する文章として最も尤もらしい文の系列をプロトコルとして出力する．候補文からプロトコルの選定を行うための探索は，西村らの手法に倣い，ビタビアルゴリズムにより行う．

4 実験と結果

前節で説明したプロトコル生成手法により，BioVL2 データセットの実験映像からプロトコル生成を行うタスクに取り組んだ．

4.1 事前学習

文生成モデルの事前学習には WLP データセット [7] を用いた．WLP データセットは生物学分野における実験プロトコルを収集したテキストデータセットであり，各プロトコルの文には，単語レベルでの b-NE のアノテーションが付与されている．b-NE のアノテーションは，試薬を表す Reagent や実験器具を表す Device などいくつかのタグに分けられているが，BioVL2 において付与されているアノテーションが Reagent, Device, Location であることから，これら 3 種類のタグ及び実験者の動作を表す Action を文生成モデルの入力として与えた．

動詞の候補	テンプレート	BLEU1	BLEU2	BLEU3	BLEU4	METEOUR	ROUGE-L
b-NE のみ	-	41.8	30.4	22.6	**17.1**	20.4	28.2
WLP 中の動詞 100 位まで	動詞のみ	25.4	14.8	8.8	5.1	19.5	22.6
	動詞 + b-NE	28.1	17.3	10.8	6.6	17.5	25.3
WLP 中の動詞 30 位まで	動詞のみ	27.1	17.4	12.1	8.9	15.2	26.0
	動詞 + b-NE	34.3	23.4	16.3	11.6	17.2	20.6
BioVL2 中の動詞 14 個	動詞のみ	36.9	25.3	17.3	12.2	20.3	23.2
	動詞 + b-NE	41.2	26.0	16.9	10.5	18.9	25.9
	正解の動詞	**53.3**	**36.1**	**24.0**	16.0	**24.4**	**41.1**

表 1　自動評価指標によるスコア．太字は各評価指標において最も高いスコアを表す

4.2　プロンプト

VideoCLIP を用いた動作推定において，3.2.1 で説明した，動詞の候補を埋め込むテンプレート文を利用した．テンプレート文として，動詞のみを埋め込むテンプレート文と，動詞と b-NE を埋め込むテンプレート文の 2 種類を用意した．動詞のみを埋め込むテンプレート文は “a video of {verb} in the laboratory, a type of actions”，実験映像中の b-NE と動詞を埋め込むテンプレート文は”a video of {verb} using {b-NE} in the laboratory, a type of actions”とし，例えば図 1 では，’{verb}’ に’set’ を，’{b-NE}’ に’primer1 and DNA engine’ を埋め込むことでプロンプト “a video of set using primer1 and DNA engine in the laboratory, a type of actions”を形成する．また，’{verb}’ に埋め込む動詞の候補として，WLP データセット中の動詞の中で出現頻度が上位 30 位までの動詞，100 位までの動詞，及び BioVL2 データセット中の 14 の動詞を用いた．

4.3　生成プロトコルの評価

提案手法に加え，ベースラインとして，推定した動詞を用いず b-NE のみを文生成モデルに与えた場合，及びあらかじめアノテーションされた正解となる動詞を b-NE と共に文生成モデルに与えた場合についてプロトコルを生成し，比較を行なった．生成されたプロトコル例を付録に示す．生成されたプロトコルの自動評価尺度として，BLEU [8]，METEOR [9]，ROUGE-L [10] を用い，BLEU の N の値は $N = 1, 2, 3, 4$ とし評価を行なった．各動詞の候補及び b-NE を用いて生成したプロトコルの自動評価尺度による結果を表 1 に示す．この結果から以下の 3 つのことがいえる．

第一に，推定した動詞の情報を文生成モデルに与える場合，正しく動詞を推定しなければ逆に生成結果に悪影響を与えてしまうということである．BioVL2 データセットにおいてあらかじめアノテーションされた正解となる動詞を b-NE に加えてプロトコル生成を行なった結果がほとんどの評価指標において最も高いスコアを記録しているが，動詞を用いず b-NE のみで生成を行なった結果がそれに続くスコアとなっている．このことから，プロトコル生成に動作の情報を用いることは有用であるが，適切な動作の情報を用いなければそれらがノイズとなってしまうといえる．

第二に，テンプレート文に動詞のみを埋め込む場合よりも，動詞と b-NE の両方を埋め込んで動詞の推定を行う方がよいということである．この原因として’Centrifuge’ など，専門的な実験器具を操作する動詞の推定には対象となる実験器具の名称が必要であることが挙げられる．

第三に，候補となる動詞の集合と，BioVL2 データセットに現れる動詞の集合との乖離が小さいほど良いプロトコルが得られるということである．これは候補となる動詞の集合が大きくなることで，推定した動詞が BioVL2 データセットに含まれる動詞と異なる確率が高くなり，結果として正解のプロトコルと生成されたプロトコルの差異が大きくなることが原因であると考えられる．

5　おわりに

本研究では，実験映像からの実験者の動作を加味したプロトコル生成を行なった．本研究におけるプロトコルの評価結果は，実験者の動作の情報を与えることは有用であるが，与える動作の情報が正しいものでなければ，プロトコルの生成において悪影響を及ぼすということを示している．今後の方針として，より正確な動詞の推定手法を用いたプロトコルの生成を行うことが考えられる．

謝辞

本研究は JSPS 科研費 JSPS 21J20250 の助成を受けたものです.

参考文献

[1] Baker M. 1,500 scientists list the lid on reproducibility. **Nature**, Vol. 533, pp. 452–454, 2016.

[2] Atsushi Ushiku Atsushi Hashimoto Natsuko Okuda Fumihito Ono Hirotaka Kameko Taichi Nishimura, Kojiro Sakoda and Shinsuke Mori. Biovl2 dataset: Egocentric biochemical video-and-language dataset. **Journal of Natural Language Processing**, 2022.

[3] Atsushi Ushiku, Hayato Hashimoto, Atsushi Hashimoto, and Shinsuke Mori. Procedural text generation from an execution video. In **International Joint Conference on Natural Language Processing**, pp. 326–335, 2017.

[4] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for
zero-shot video-text understanding. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Online, November 2021. Association for Computational Linguistics.

[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In **International Conference on Computer Vision**, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **Proceedings of the International Conference on Machine Learning**, pp. 8748–8763, 2021.

[7] Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. An annotated corpus for machine reading of instructions in wet lab protocols. In **North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 97–106, 2018.

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

[9] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Association for Computational Linguistics**, pp. 65–72, 2005.

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

正解のプロトコル		b-NE のみ		BioVL2 の動詞		正解の動詞	
手順 1	Add phenol chloroform.	手順 1	Add 200 l chloroform and mix by inverting the phenol.	手順 1	Incubate at room temperature for 5 minutes.	手順 1	Add phenol l of chloroform to the chloroform and mix.
手順 2	Invert violently.	手順 2	Add 200 l chloroform and mix by inverting the tube several times.	手順 2	Add 200 l of phenol to the tube and incubate at room temperature for 10 minutes.	手順 2	Centrifuge the chloroform at room temperature for 5 minutes.
手順 3	Centrifuge for 2 minutes at room temperature.	手順 3	Place the tube in the magnet for 5 minutes.	手順 3	Incubate at room temperature for 2 minutes.	手順 3	Transfer the new tube.
手順 4	Transfer the DNA layer to a new tube.	-		-		-	

表 2　フェノールクロロホルム法の映像に対するプロトコル生成結果の例

A　BioVL2 データセット

BioVL2 データセットは，生化学分野における実験映像を収集した一人称視点の映像データセットである．生化学分野での基礎的な実験である PCR，ミニプレップ法，アガロースゲル作成，DNA 抽出の 4 種類の実験をそれぞれ 8 動画ずつ撮影している．実験映像の撮影は，実験者の頭にヘッドマウントカメラを取り付け行われており（図 2），計 32 動画，約 178 分の動画像データセットとなっている．また BioVL2 データセットには，(1) 実験映像に対応するプロトコル，及び (2) 実験映像中に現れる試薬，実験器具の映像中のバウンディングボックスの 2 種類のアノテーションが与えられている．

A.1　実験映像とプロトコルの対応関係のアノテーション

実験映像に対応するプロトコルは，実験者が実験内容を口頭で説明したものが文書化され与えられている．プロトコルは実験者の動作ごとに手順として区切られており，例えば “Add phenol chloroform and invert violently.”という内容は “Add phenol chloroform”と “Invert violently”という手順に分けられる．また各手順に対して，実験映像における開始時刻と終了時刻が付与されている．

A.2　実験映像中の物体に対するアノテーション

実験映像中の物体に対して付与されたバウンディングボックスの例を図 3 に示す．このアノテーションは，映像から 4 秒ごとにフレームを抽出し，(1) 手が物体と触れている，かつ (2) 触れている物体がプロトコルに現れる場合に対して，その物体をバウンディングボックスで囲い，物体名を付与することで為されている．

B　生成プロトコル例

表 2 に，実験映像からのプロトコル生成例を示す．ここでは，正解のプロトコル，b-NE のみを入力とした場合，BioVL2 の動詞を動詞推定の候補とした場合，及び VideoCLIP による推定ではなく正解となる動詞を与えた場合について例を示している．なお，動詞推定の際のテンプレート文は動詞と b-NE を埋め込むテンプレート文を利用している．

生成されたプロトコルを見ると，正解の動詞を与えることで，正解のプロトコルに近い動詞を出力できていることがわかる．一方で，VideoCLIP により推定した動詞を与えた場合，誤った動詞を出力してしまっている例が多く見られた．b-NE のみを用いて生成を行うモデルが出力する動詞は，WLP データセットで事前学習した際の b-NE との共起関係から得られるものであり，そのようにして得られる動詞を用いる場合よりも，誤って推定した動詞を与える方が，モデルへの悪影響が大きいと考えられる．

また，生成されたプロトコル中の “for 5 minutes”や “200 l”といった量や時間に関する数値は，WLP データセット中の共起関係に依るものであり，実際の映像からこれらの数値を読み取っているわけではない．プロトコル中のこのような数値が誤っている場合，実験再現の失敗に繋がると考えられるため，実験中の詳細な数値を生成プロトコルに反映する方法について検討する必要がある．

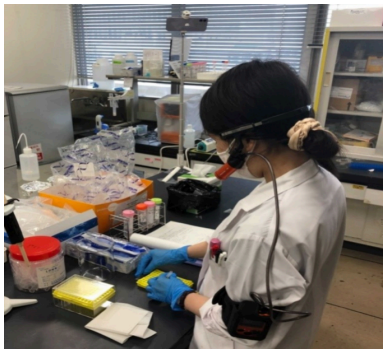

図 2　実験映像撮影の様子

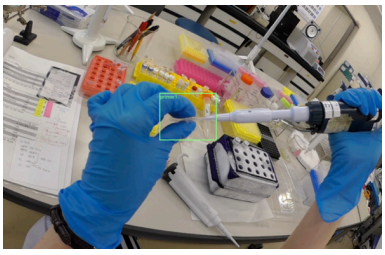

図 3　映像中の物体に付与されたバウンディングボックスの例