3P1-2in2

人の挙動を表現するテキスト生成の一考察

A Study on Text Generation describing Human Behavior

樺山 絵里 *1 Eri Kabayama
小林 一郎 *1 Ichiro Kobayashi
麻生 英樹 *2 Hideki Asoh
持橋 大地 *3 Daichi Mochihashi
アッタミミ ムハンマド *4 Attamimi Muhammad
中村 友昭 *4 Tomoaki Nakamura
長井 隆行 *4 Takayuki Nagai

*1お茶の水女子大学 Ochanomizu University
*2産業技術総合研究所 National Institute of Advanced Industrial Science and Technology
*3統計数理研究所 The Institute of Statistical Mathematics
*4電気通信大学 The University of Electro-Communications

This paper discusses a method to generate sentences describing human behaviors observed by Kinect camera, and the ability of transferring the existing linguistic resources to the categories which have not yet been described by words. We have set 60 human behavior categories, and collected natural language descriptions to make bi-gram language for generating sentences. Through experiments, we have shown that our method expands the ability of our former studies in terms of both generating sentences and transferring linguistic resources.

1. はじめに

非言語情報を言葉で説明するテキスト生成の研究が盛んになってきている．Usiku ら [1] は静止画に対する説明文を n-gram 言語モデルを用いて生成している．また，小林ら [2] も同様に n-gram 言語モデルを用いて動画中の人の動作を説明するテキスト生成を行っている．観察対象を説明するテキストを生成するためには，十分な言語資源が必要となるが，説明対象ドメイン毎に充実した言語資源があることは期待し難い．このことから，Asoh ら [3] は，観測した視覚情報である人の動作を過去に説明したことが無い場合に，これまでに見た動作を説明するのに用いた言語資源を zero-shot 学習 [4] により転移させることにより，その動作を説明することを試みている．本研究では，先行研究 [2] [5] で提案された言語生成手法および言語資源の転移手法手法をより汎用的にするため，対象動作の種類を増やし，その上で時系列データからの言語生成およびテキスト生成のための言語資源の転移を行った．

2. 時系列データからの言語生成

本研究では，人の動作の時系列データを，Kinect センサを使用して，9 関節に関しての x, y, z 座標の形で取得する．今回対象とするのは，図 1 に示す意味的な構成からなる簡単な人の動作である．先行研究 [5] における動作の意味的な構成に「高く」「低く」を加えたものに相当する．

対象とする動作は,「右/左/両」「手/足」を「速く/遅く」「高く/低く」「上げる/下げる」の組み合わせの 60 の動作から構成されると定義する．動作カテゴリは，例えば「両」「手」「速く」「高く」「上げる」の組み合わせにより「両手を速く高く上げる」というカテゴリが構成される．それぞれの動作カテゴリに対して，自然言語文を収集し，それを元にバイグラムモデルを構築する．このバイグラムモデルを用いて，尤度が高い単語の組み合わせを動的計画法を用いて抽出することにより，テキストを生成する．また，その際，文長が長い文は尤度が低くなってしまうため，図 2 のように文長を揃えるために仮想の単語 null を導入し，文長に左右されないテキスト生成を実現する．この言語資源から構築されるバイグラムを用いて，テキスト生成を行う.「両手を速やかに高く上げる」という動作に対する生成結果を表 1 の full に示す.

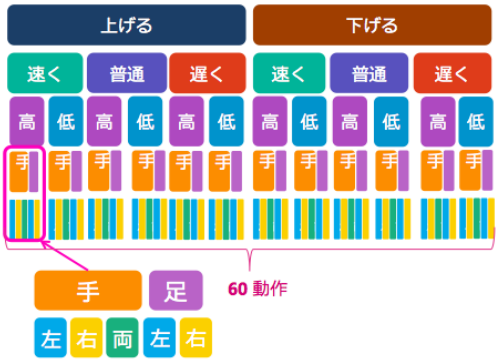

図 1: 動作カテゴリの構成

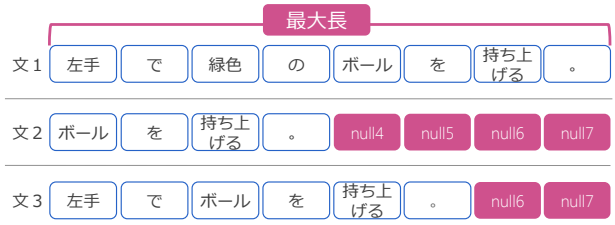

図 2: 仮想の単語 null の導入

3. zero-shot 学習に基づく言語資源推定

特定の意味を説明するための言語資源が存在しないことの影響を評価するために，先行研究 [2] で提案した手法を用いて一部の動作に対する言語資源を取り除き，最小二乗推定による

表 1: 「左手を速く低く上げる 」という動作に対する削減された言語資源の下での生成文

言語資源	生成文
full	• 左手 を 素早く 上げる 。 null6 null7 null8 null9 null10 null11 EOS • 左手 を 速く 低く 上げる 。 null6 null7 null8 null9 null10 null11 • 左手 を 速く 低く 上げる 。 null5 null6 null7 null8 null9 null10
three-quarters	• 左手 を 素早く 上げる 。 null6 null7 null8 null9 null10 null11 EOS • 左手 を 速く 低く 上げる 。 null6 null7 null8 null9 null10 null11 • 左手 を 速く 低く 上げる 。 null5 null6 null7 null8 null9 null10
half	• 左手 を 素早く 上げる 。 null6 null7 null8 null9 null10 null11 EOS • 左手 を 速く 低く 上げる 。 null6 null7 null8 null9 null10 null11 • 左手 を 速く 低く 上げる 。 null5 null6 null7 null8 null9 null10
min	• 左足 を 素早く 上 に 挙げる 。 null7 null8 null9 null10 null11 • 人 が 左足 を 素早く 上 に 挙げる 。 null7 null8 null9 • 左足 を 素早く 上 に 挙げる 。 null6 null7 null8 null9 null10

表 2: BLEU スコアおよび生成文の対数尤度に基づく評価結果

	推定された言語資源	full	three-quarters	half	min
BLEU	(データ全動作)	1.0	0.7949	0.6566	0.3528
	欠損動作	(1.0)	0.3940	0.3763	0.2739
対数尤度	min,half,three-quarters 共通欠損動作	-1.9464	-2.0115	-4.3354	0.2053
	half, three-quarters 共通欠損動作	-2.2846	-2.4807	-4.8463	—

zero-shot 学習を行う．これにより，他の動作に対する言語資源用いて，欠損している動作に対する言語モデルを推定する．その後，推定された言語モデルを用いて説明文の生成を行い，得られた説明文の品質を評価する．zero-shot 学習は，動作カテゴリに対して，バイグラムを表した行列 Ψ を動作カテゴリに対して，それが含んでいる動作要素を表した行列 A およびそれぞれの動作要素に対して使用される言語資源の確率を表す行列 Φ とに分解できるということを前提に，特定の動作カテゴリのバイグラムが欠損している場合，行列の一般逆行列を使って欠損部分の要素を推定する．上述した内容は以下の式で表せる．ここで A^+ は A の一般逆行列を示す．

$$\hat{\Phi} = \min_{\Phi} ||\Psi - A\Phi||^2 = A^+\Psi \quad (1)$$

4. 実験

4.1 実験設定

先行研究 [5] と同様に，zero-shot 学習により，データ欠損動作の言語モデルをどの程度正確に推定可能であるかを検証するために，動作カテゴリの構成において出現していない構成要素が存在しないようにバランスを考慮しつつ，その一部を取り除いた言語資源のデータを用意した．full は全データ，three-quarters，half は全データから 1/4 および 1/2 のデータをそれぞれ取り除き，min は全構成要素が現れる最小限の言語資源とする．生成された文の定量的な評価手法として，BLEU スコアによる評価と生成文の対数尤度評価を用いた．

4.2 実験結果および評価

言語モデル，full，three-quarters，half，min に対して，それら全てに共通して推定された言語モデルである動作「左手を速く低く上げる」に関するテキストの生成結果を表 1 に，評価結果を表 2 に示す．

4.3 考察

今回，人の動作が 5 つのカテゴリ要素によって説明されるとして，先行研究 [5] よりも複雑な意味的な構成から成る言語資源の転移を行ったが，先行研究 [5] 同様, 言語資源が減少するにつれて, 評価が 下がり, 文章として説明ができていないものが増える ことが確認された．

5. おわりに

人の動作を表す時系列データから動作の振る舞いの大きさを考慮したパターン抽出を行い，そのパターンに基づいて速さを含んだ動作に対する説明文生成を行った．また，先行研究 [5] よりも複雑な動作の意味的な構成において対応する言語資源を推定する zero-shot 学習を行い，言語資源の量と生成文の精度は比例するという結果を得た．今後は，より一般的な人の動作を対象に提案手法を改良していくつもりである．

謝辞

本研究は，JSPS 科研費 26280096 の助成を受けて実施した．

参考文献

[1] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. A Understanding Images with Natural Sentences. the 19th Annual ACM International Conference on Multimedia (ACMMM 2011), pp.679-682, 2011.

[2] 小林瑞季, 麻生英樹, 小林一郎, 人の動作を対象にした確率的言語生成への取り組み，言語処理学会第 20 回年次大会，pp.920-923, 北海道大学，2014.

[3] Hideki Asoh and Ichiro Kobayashi, Zero-Shot Learning of Language Models for Describing Human Actions Based on Semantic Compositionality of Actions, The 28th Pacific Asia Conference on Language, Information and Computing, Dec. 12-14,; Phuket, Thailand, 2014.

[4] Larochelle, H., Erhan, D., & Bengio, Y. Zero- data learning of new tasks. AAAI Conference on Arti- ficial Intelligence, 2008.

[5] 樺山絵里，麻生英樹，小林一郎，持橋大地，アッタミミムハンマド，中村友昭，長井隆行，言語モデルを用いた人の動作を説明するテキスト生成およびその評価，言語処理学会第 22 回年次大会，pp.14-15，東北大学，2016.