Zero-shot 学習した言語モデルによるテキスト生成結果の評価

樺山 絵里[†]　麻生 英樹[‡]　小林 一郎[†]　持橋 大地[¶]　Muhammad Attamimi[§]
中村友昭[§]　長井隆行[§]
[†]お茶の水女子大学　[‡]産業総合技術研究所　[¶]統計数理研究所　[§]電気通信大学大学院

[†]{g1120513,koba}@is.ocha.ac.jp,[‡]h.asoh@aist.go.jp,[¶]daichi@ism.ac.jp,
[§]m_att@apple.ee.uec.ac.jp,[§]tnakamura@uec.ac.jp,[§]tnagai@ee.uec.ac.jp

1　はじめに

動画や時系列データなどの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．このような研究はロボット工学の分野で古くから研究が進められており，たとえば，Ushiku ら [1] は静止画に対する説明文を n-gram モデルを用いて生成している．観察対象を説明するための n-gram などを学習するためには，言語資源＝学習用データが必要となるが，説明対象ごとに十分な言語資源があることは期待し難い．この問題に対して，われわれは，人の動作の説明を対象として，一部の動作に対する学習用データが存在しない場合に，他の動作に対する学習用データを用いて言語モデルを zero-shot 学習する方法を提案した [2]．本発表では，その手法において，与えられる学習用データの量を変化させた時に生成される文に対する評価を行った結果について報告する．

2　言語モデルの学習とテキスト生成

2.1　概要

我々は，人の動作を説明する文章を生成することを目指して研究を進めている．これまでに，動作の認識と認識結果から文章生成を行うシステムを試作してきた [3]．ここでは，認識結果から文章生成を行うための手法として，統計的言語モデル（バイグラム）を用いている．認識結果である動作ごとの言語モデルを構築するために，各動作に対する説明文を学習用データ（言語資源）として収集して，言語モデルを学習している．しかし，全ての動作について学習用データの説明文を収集することは，動作の数が増えると現実的ではない．そこで，いくつかの動作について言語モデルを学習するための説明文が得られない場合でも，他の動作についての説明文データを用いて言語モデルの学習を可能にする方法となる言語モデルの zero-shot 学習の方法を提案した [2]．本発表では，提案した zero-shot 学習法の性能評価のために，学習用データが与えられる動作の数を減らした場合に，学習された言語モデルを使って生成される文の妥当性がどのように低下してゆくかを調べた結果について報告する．

2.2　言語モデル構築

本研究では，収集したテキストから構築したバイグラムモデルを用いて，尤度が高くなるような単語の組合わせを見つけることにより文の生成を行うとする．一般に，観測対象が同じ現象であったとしても，人によってその対象の説明の仕方は様々であり，選択する語彙や説明文の長さが異なる．構築したバイグラムモデルから尤度の高い単語の組合わせを抽出することによってテキスト生成を行う場合，単語数が少ない文のほうが尤度が高くなってしまう．このことから，本研究では，文長に左右されないテキスト生成を行うために，小林ら [3] が用いた，疑似単語（番号付き null）をバイグラムモデルに導入することにより文長に関わらず尤度の次数を同じにする手法を適用する．

3　Zero-shot 学習に基づく言語資源推定

本節では，我々が提案している zero-shot 学習の方法について，今回の実験に即して簡単に述べる．Zero-shot 学習は，マルチタスク学習の一種であり，対象とする学習課題に関する学習用データ無しで学習を行う手法である．近年，一般物体認識のようなカテゴリ数が非常に多いパターン識別の課題などに関して研究が盛んになっている．そうした問題では，正解ラベルのついた学習用データをすべてのカテゴリに対して用意

することが難しい．しかしながら，カテゴリ間の意味的な関係などを利用することで，一部のカテゴリに関する学習用データが無い状態でも，他のカテゴリに関する学習用データの情報を使った学習が可能になることが示されている．われわれは，zero-shot 学習の考え方を，トピックに依存した多数の言語モデルを同時に学習する問題に適用し，学習法を提案している [2]．

3.1 動作の意味的な構成

今回の実験で対象としているのは簡単な人の動作である．観測される動作は，手（両手を含む）と足に関する動作であり，手足を上げる，下げる，そして，その動作方向を示す，前，横，左，右の計 9 つの構成要素の内の 4 つから成り立っている．例えば，「右手を前から上げる」の場合，「右」「手」「前から」「上げる」という要素から成ると考えることができる．図 1 に対象とする 20 種類の動作の構成を示す．図 1 において，縦一列が一つの動作を表し，一動作は 4 つの構成要素を含んでいる．

図 1: 動作の意味的な構成

3.2 Zero-shot 学習の方法

図 1 に示す k 番目の動作に l 番目の要素が含まれていることを $a_{kl} = 1$ で表し，それを成分とする行列を A とする．

各動作に対する言語モデルとして，2 単語ペアの出現確率 $p(w_i, w_j)$ を求めることを考える．動作 k に対する説明文集合から計算される $p(w_i, w_j)$ の値を並べたベクトルを ψ_k として，それを各行とする行列を Ψ とする．ここで，行列 Ψ が，$\Psi = A\Phi + \varepsilon$ のように近似的に分解できることを仮定する．ここで，Φ は動作の構成要素に対する言語モデルを行とする行列である．すなわち，各動作に対する言語モデルが，動作の構成要素に対する言語モデルの線形の重みつき和で近似できると仮定していることになる．この仮定に基づき，以下の手続きに示される zero-shot 学習の方法を提案した．以下では，学習用データ（説明文）が存在しない動作を「データ欠損動作」と呼ぶ．

step1. Ψ のうちの，学習用データが存在する動作に対応する行だけから成る行列を Ψ' とする．また，A のうちの、同じようにデータが存在する行動に対応する行から成る行列を A' とする．

step2. Ψ' と A' から、動作の構成要素に対する言語モデル $\hat{\Phi}$ を最小二乗推定する（式 1）．

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

ここで A'^{+} は A' の一般化逆行列である．

step3. 推定された $\hat{\Phi}$ を用いて $\hat{\Psi} = A\hat{\Phi}$ のように，Ψ 全体を復元することで，データ欠損動作に対する言語モデルを推定する．

4 実験

学習用データが存在しないことの影響を評価するために，一部の動作に対する学習用データを取り除き，最小二乗推定による zero-shot 学習を行うことにより，他の動作に対する学習用データを用いて，データ欠損行動に対する言語モデルの推定を行う．その後に，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した．

4.1 説明文の収集

実験で用いたデータは [3] で用いられたものと同じである．それぞれの動作を Kinec で撮影したビデオを 12 人の被験者に見せ，動画に映っている，人の動作を説明するテキストを収集した各動作ごとに 20 文程度のデータを収集している．

4.2 実験設定

Zero-shot 学習により，データ欠損行動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，収集したデータの一部だけを用いた実験を行った．説明文を欠損させる動作を，動作の意味的な構成ができるだけ均等になるように選んだ以下の 4 つの場合について検討した．

表 1: 「右手を上にあげる」という動作に対する zero-shot 学習された言語モデルからの生成文

動作	生成文
full	• 右手, を, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12, null13, null14 • 右手, を, あげる, . , null5, null6, null7, null8, null9, null10, null11, null12, null13, null14, EOS • 右手, を, 上, に, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12
three quarters	• 右手, を, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12, null13, null14 • 右手, を, あげる, . , null5, null6, null7, null8, null9, null10, null11, null12, null13, null14, EOS • 右手, を, 前, に, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12
half	• 右手, を, 下げる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12, null13, null14 • 右手, を, 下げる, . , null5, null6, null7, null8, null9, null10, null11, null12, null13, null14, EOS • 右手, を, 上, に, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12
min	• 左手, を, あげる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12, null13, null14 • 左手, を, あげる, . , null5, null6, null7, null8, null9, null10, null11, null12, null13, null14, EOS • 左手, を, 上, に, 上げる, . , null4, null5, null6, null7, null8, null9, null10, null11, null12

1. full (言語資源を全て使用)
2. three quarters (4 分の 3 を使用)
3. half (半分を使用)
4. min (文生成が可能な最低限の数を使用)

生成された文の定量的な評価手法として，以下の 2 つを考える．

- **BLEU スコアによる評価**
  full のデータから学習した言語モデルによって生成されたテキストと zero-shot 学習によって推定された言語モデルによって作成されたテキストとの BLEU スコアにより評価する．

- **生成文の尤度評価**
  zero-shot 学習によって推定された言語モデルから生成された尤度の上位 K 件（ここでは，$K = 3$ とした）の説明文の尤度を full のデータから学習した言語モデルを用いて算出した際の平均値．このとき，full の言語モデルの中に推定された言語モデルから生成された文に現れる単語ペアがない場合には，その単語ペアの確率を語彙数の逆数などを取るとして，適切なスムージングを行って補った．

4.3 実験結果

full，three quarters，half，min のそれぞれのケースに対して,「右手を上にあげる」に関するテキスト生成結果を表 1 に示す．表には，尤度が高い 3 つの文をしめしている．各言語モデルにおいて，生成された文を見ると，学習用データが少なくなるほど，full の言語モデルで生成された文の意味から異なる文が生成されている様子がわかる.「右手を上にあげる」という動作に対して，three quaters では「上に」が「前に」に変わっており，half においては「上げる」という動作が「下げる」になってしまっている．また，min に関しては「右手」が「左手」となっており，元々の文意から大きく異なった文が生成されていることがわかる．

4.4 評価結果

4.4.1 BLEU スコアによる評価

Zero-shot 学習により推定された言語モデルを用いて生成された文を，full の言語モデルにより生成した文を正解文とした場合の BLEU スコアを用いて評価した結果について述べる．図 2 に，zero-shot 学習によって推定された言語モデルおよび取り除く対象にならなかった言語モデルの両方を用いて，全動作に対するテキスト生成を行った結果を示す．また，図 3 に，three quarter, half, min すべての場合に共通するデータ欠損動作である動作 1, 5, 20 に対して zero-shot 学習によって推定された言語モデルから生成された文と full のデータから推定された言語モデルから生成された文との一致を評価した結果を示す．どちらに関しても，取り除かれた言語モデルの推定に多くの学習データを使っているものほど，精度の高い文が生成されていることがわかる．

4.4.2 尤度による評価

three quarters, half, min に共通するデータ欠損動作に対して，zero-shot 学習で推定された言語モデルから生成された文の尤度を，full の言語モデルで計算した結果を図 4 に示す．three quarters, half, min の

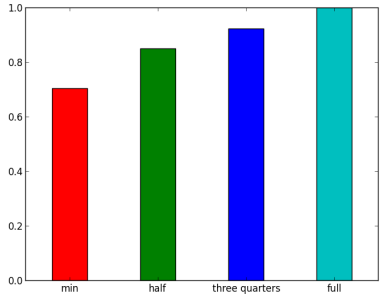

図 2: 全動作に対する BLEU スコア

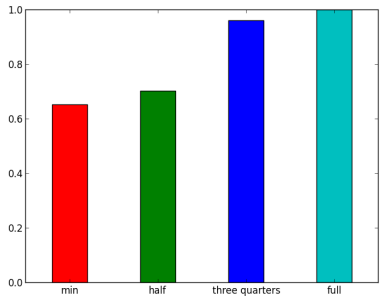

図 3: データ欠損動作に対する BLEU スコア

順に尤度が高くなっており，ここでも，より多くの学習用データを用いて生成したものほど，生成文の精度が高くなっていることがわかる．

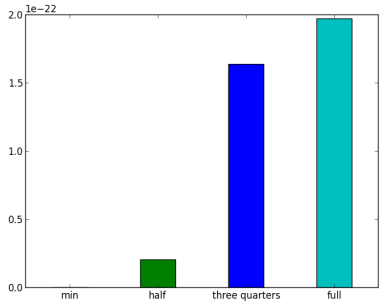

図 4: three quarters, half, min に共通するデータ欠損動作に対する説明文の尤度の平均

full, three quarters, half の 3 つのケースをより詳しく比較するため，three quarters, half に共通するデータ欠損動作 1, 3, 5, 18, 20 についての評価も実施した．図 5 に結果を示す．three quarters, half の順に尤度が高くなっており，より多くのデータを用いて生成したものほど生成文の精度が高くなっている．全体的に，BLUE スコアに比べて尤度による評価のほうが，性能の落ち方が顕著に現れている．

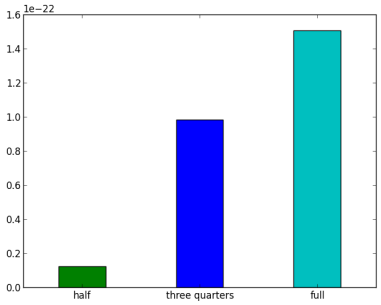

図 5: three quarters, half に共通するデータ欠損動作に対する説明文の尤度の平均

5 おわりに

本発表では，動作の意味的な構成を利用する zero-shot 学習によって推定された言語モデルから生成された文の評価を行った結果について述べた．言語モデルの学習に使用する学習用データの量を変えた際の生成文の評価を BLEU スコアおよび生成文の尤度によって行った．それにより，学習用データの量を削減したときに，どのように性能が劣化するかを明らかにすることができた．

今後の課題としては，より多様なデータ削減方法についての評価を行い，できるだけ性能を劣化させないようなデータ削減方法を明らかにしたいと考えている．また，現在の方法では，動作の意味構成を表す行列 A が対象とする動作すべてについて既知であることを仮定している．また，意味構成に対称性がある簡単な動作を対象としている．こうした点に対して，行列 A の内容も推定しながら言語モデルを推定できる手法などの拡張を検討してゆきたい．

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] 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.

[3] 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.