2F3-2

現象の意味的構造に基づく言語モデルの zero-shot 学習の試み

An Approach to Zero-shot Learning of a Language Model based on the Semantic Structure of Phenomena

樺山 絵里 *1 Eri Kabayama　麻生 英樹 *2 Hideki Asoh　アッタミミ ムハンマド *4 Attamimi Muhammad　小林 一郎 *1 Ichiro Kobayashi　持橋 大地 *3 Daichi Mochihashi
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

Based on an assumption that the meaning of sentences describing humans' simple motions can be represented with the combination of some semantic elements, in this paper, we propose a zero-shot learning method to estimate a language model which describes unknown human motions. We apply our proposed method to estimating the missing linguistic resources of a language model which describes human everyday activities, and report the ability of the method by evaluating the results obtained from the experiments with various conditions of language model used for estimation.

1. はじめに

直接観測される学習用データが全く存在しない状況における機械学習手法として，zero-shot 学習が注目されている．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 [1]．カテゴリ間の属性における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．

一方，近年，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．Ushiku ら [2] は静止画に対する説明文を n-gram モデルを，また小林 [3] らは動画中の人の動作に対する説明文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，我々は，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 [4]，先行研究では，人の動作を記述する簡単な文章を対象にして，手足の動作の左右の対称性をもった意味的な構成から資源を転移させ [5]，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結果について報告を行った．本研究では，４つのカテゴリ（物，動作，場所，人）の意味的な構成に基づき表現される現象を対象にしたより大きなデータを対象にした言語資源の転移を考える．

連絡先: 樺山 絵里，お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース 小林研究室，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，kabayama.eri@is.ocha.ac.jp

2. 言語モデルの学習とテキスト生成

2.1 概要

我々は，現象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，人の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た [3]．また，文章生成に使用される言語資源に対し，zero-shot 学習を用いた資源の拡充 [4][5] を行った．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，言語モデルを学習している．本研究では，[5] における人の動作の意味的な構成を表す小規模なデータから，より一般的な現象を表現する 4 つのカテゴリの構成からなる文章の言語資源に対して，提案手法を適用し，さらに提案する手法の性能評価を行う．

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

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．我々は，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 [4].

3.1 現象の意味的な構成

先行研究 [3, 5] では，人の簡単な動作を対象としており，総計 9 記述要素，総計 20 行為，1 行為あたり 12 文および 4 記述要素が含まれるデータを用いていた．今回の実験で対象としている現象は，4 つのカテゴリ（物，動作，場所，人）を含むものである．各カテゴリにおける全データの構成要素の種類数は，物:33，動作:19，場所:6，人:4 となっている．例えば，「女の子がリビングでガラガラを上下に振る」の場合，「ガラガラ」「上下に振る」「リビング」「女の子」という要素から成ると考えることができる．総計 132 行為，1 行為あたり 5 文および 4 記述要素が含まれるデータである．

図 1 に対象とする 132 種類の現象の構成の一部を示す．図 1 において，縦一列が一つの事象を表し，一事象は 4 つの構成要素を含む．

図 1: 現象の意味的な構成

3.2 zero-shot 学習の方法

図 1 に示す意味的な構成において k 番目の事象に 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．

各事象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．事象 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．また，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 は現象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，現象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）が存在しない事象を「データ欠損事象」と呼ぶ．

step1. Ψ のうちの，学習用データが存在する事象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事象に対応する行から成る行列を A' とする．

step2. Ψ' と A' から、現象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）.

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

step3. 推定された $\hat{\Phi}$ を用いて $\hat{\Psi} = A\hat{\Phi}$ のように，Ψ の削除した行を復元することで，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，一部の事象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，他の事象に対する学習用データを用いて，データ欠損事象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 実験設定

zero-shot 学習により，データ欠損事象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，4 つのカテゴリの意味的な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て取り除くようにして以下の 4 つの場合について検討した．

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．

- **生成文の対数尤度評価**
  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う．

4.2 実験結果

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事象「女の子がリビングでクッキーを取り出す」に関するテキスト生成結果を表 1 に示す．各言語モデルにおいて，生成された文を見ると，取り除かれた言語資源を推定するのに使われた言語資源が少なくなるほど，full の言語モデルで生成された文とは異なる文が生成されている様子がわかる．「女の子がリビングでクッキーを取り出す」という事象に対して，three-quarters および half では「女の子」が「男の子」に変わっており，min では，上位 3 文のうち 2 文において「女の子」という主語がない．このことから，言語資源が減少するにつれて，文章として説明ができて

表 1: 「女の子がリビングでクッキーを取り出す」という事象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	• 女の子 は リビング で クッキー を 開ける null8 null9 null10 null11 null12 null13 null14 null15 null16 null17 EOS • 女の子 は リビング で クッキー 箱 を 開ける null8 null9 null10 null11 null12 null13 null14 null15 null16 null17 • 女の子 は リビング で クッキー を リビング で 開ける null8 null9 null10 null11 null12 null13 null14 null15 null16
three-quarters	• 男の子 は リビング で クッキー 箱 を 開け た null9 null10 null11 null12 null13 null14 null15 null16 null17 • 男の子 は リビング で クッキー 箱 から 中身 を 開け た null9 null10 null11 null12 null13 null14 null15 • 男の子 は リビング で 男の子 は リビング で クッキー 箱 を 開け た null9 null10 null11 null12 null13
half	• 男の子 は リビング で クッキー 箱 を 開け た null9 null10 null11 null12 null13 null14 null15 null16 null17 • 男の子 は リビング で クッキー 箱 から 中身 を 開け た null9 null10 null11 null12 null13 null14 null15 • 男の子 は リビング で クッキー 箱 を 聞く null9 null10 null11 null12 null13 null14 null15 null16 null17 EOS
min	• リビング で クッキー 箱 から 中身 を 開け た null9 null10 null11 null12 null13 null14 null15 null16 null17 • 女の子 が リビング で クッキー 箱 から 中身 を 開け た null9 null10 null11 null12 null13 null14 null15 • リビング で クッキー 箱 から 中身 を 食べる null9 null10 null11 null12 null13 null14 null15 null16 null17 EOS

表 2: BLEU スコアおよび生成文の性能における評価結果

		full	three-quarters	half	min
BLEU	データ全事象	1.0	0.8959	0.7489	0.6152
	欠損事象	(1.0)	0.5789	0.5054	0.4839
対数尤度	min, half, three-quarters 共通欠損動作	-59.82	-94.11	-96.50	-98.36
	half, three-quarters 共通欠損動作	-59.74	-94.25	-96.60	—

いないものが増えることが考えられる．なお，three-quarters と half の生成結果が同じになっているのは，言語資源を削減する際に half で削減した言語資源は three-quarters で削減した言語資源を必ずしも含んでおらず，half であっても，たまたまその意味構成を表現するために必要な言語資源があったためと考えられる．

4.3 評価結果

4.3.1 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．表 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，全事象に対するテキスト生成を行った結果を示す．また，three-quarters, half, min それぞれのデータ欠損事象に対して 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により生成された文との一致を評価した結果を示す．表 2 のデータ欠損事象の full の値を (1.0) と表記したのは，full ではデータが欠損していないため，対象となる文が存在しないが，1.0 とみなすためである．表 2 より，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．先行研究 [5] と比べて，値は低いことから，データに含まれる構成要素の種類数が多くかつ，1 つの事象を説明する文章量が少ない場合，言語資源を削減したときの生成文の精度は低くなっている．

4.3.2 生成文の性能評価

three-quarters, half, min について共通するデータ欠損事象に対して，zero-shot 学習で推定された言語モデルから生成された文の対数尤度を，full の言語モデルで計算した．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 と half に共通するデータ欠損事象についての評価も実施した．表 2 にその結果を示す．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，BLEU スコアに比べて生成文の尤度による評価のほうが，性能の落ち方が顕著に現れるのを観察できることがわかる．先行研究 [5] と比べると，全体的に値が小さい理由は BLEU スコアと同様である．

5. おわりに

本発表では，事象を表現する 4 つのカテゴリの構成を利用する 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．先行研究 [5] との比較によりデータの性質に基づく評価結果の傾向を捉えることができた．

現在の方法では，現象の意味構成を表す行列 A が対象とする事象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 4 つのカテゴリからなる事象を対象としている．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．

謝辞

本研究の一部は,JSPS 科研費 26280096 および人工知能研究振興財団の助成を受けて実施した．

参考文献

[1] Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-data learning of new tasks. AAAI Conference on Artificial Intelligence, 2008

[2] 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.

[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.

[4] 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.

[5] 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，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，言語処理学会第 21 回年次大会，pp.996-999，京都大学，2015