

Title	少ないデータに対する深層学習
Sub Title	Deep learning with less data
Author	斎藤, 博昭(Saito, Hiroaki)
Publisher	慶應義塾大学
Publication year	2023
Jtitle	学事振興資金研究成果実績報告書 (2022.)
JaLC DOI	
Abstract	<p>近年対話 AI が話題に上ることが多くなった。そういった対話応答システムを構築する際、モデル作成に大量の学習データが必要になる。英語や中国語と比較すると、データセットとして公開されている日本語の質の高いコーパスは少ない。そのため小さいデータや質の低いデータで十分な性能の対話応答生成を行うために、今回はTransformerベースのモデルに中国語と英語からの転移学習を利用した。転移学習自体はかなり前から研究されてきた手法であり、無の状態から学習するよりは、たとえ分野が違うモデルであったとしてもその有意性が認められている。今回はそれを1ターンの雑談対話生成というタスクで用い、実験を通して有効性を定量的に求めた。データセットとして3つの日本語コーパスとTwitterから収集したコーパスを利用し、入力文に対しての雑談生成を行った。機械的に計算できる自動評価指標としての、生成対話文の多様性を表す distinct-1 の平均の値は、転移学習なしでは0.368、転移学習モデルの平均は0.412という値となり、転移学習の効果が確認できた。また、人間が主観で評価する、文の繋がり（入力文に対しての出力文の自然さ・入力にどれだけ関連しているか）、情報の多さ（出力文から得られる情報の多さ・出力文の面白さ・独自性）、人間らしさ（入力文関係なく、出力文単体での文としての自然さ）という3つの項目に関して、対話数が9343文の小さな学習データに対し、転移学習なしのモデルと比較してすべての項目で転移学習ありのモデルが大幅に良いスコアを示した。一方、大きな学習データに対しては、多様性は少し上がったものの他の指標に関しては転移学習をしても生成結果にはそれほど大きな効果は表れなかった。</p> <p>Compared to English and Chinese, there are not many high quality publicly available corpora for Japanese in the task of chat dialog response generation. Therefore, in order to achieve a good enough performance in chat dialog generation with small and low quality data, this study utilized transfer learning from Chinese and English in the Transformer based model. Three Japanese corpora and a corpus collected from Twitter were used as the dataset to generate responses. The average value of the distinct-1 as an automatic evaluation index of the generated results was 0.368 without transfer learning, and 0.412 for the transfer learning model. In terms of human evaluation, the model with transfer learning scored significantly better on all three indices: sentence connection, informativeness, and humanness, compared to the model without transfer learning, for a small training dataset with 9343 sentences of dialogs.</p>
Notes	
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=2022000010-20220019

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the Keio Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

研究代表者	所属	理工学部	職名	准教授	補助額	200 (B) 千円
	氏名	齋藤 博昭	氏名 (英語)	Hiroaki SAITO		
研究課題 (日本語)						
少ないデータに対する深層学習						
研究課題 (英訳)						
Deep Learning with Less Data						
1. 研究成果実績の概要						
<p>近年対話AIが話題に上ることが多くなった。そういった対話応答システムを構築する際、モデル作成に大量の学習データが必要になる。英語や中国語と比較すると、データセットとして公開されている日本語の質の高いコーパスは少ない。そのため小さいデータや質の低いデータで十分な性能の対話応答生成を行うために、今回は Transformer ベースのモデルに中国語と英語からの転移学習を利用した。転移学習自体はかなり前から研究されてきた手法であり、無の状態から学習するよりは、たとえ分野が違うモデルであったとしてもその有意性が認められている。今回はそれを1ターンの雑談対話生成というタスクで用い、実験を通して有効性を定量的に求めた。</p> <p>データセットとして3つの日本語コーパスとTwitterから収集したコーパスを利用し、入力文に対しての雑談生成を行った。機械的に計算できる自動評価指標としての、生成対話文の多様性を表す distinct-1 の平均の値は、転移学習なしでは 0.368、転移学習モデルの平均は 0.412 という値となり、転移学習の効果が確認できた。また、人間が主観で評価する、文の繋がり(入力文に対しての出力文の自然さ、入力にどれだけ関連しているか)、情報の多さ(出力文から得られる情報の多さ、出力文の面白さ、独自性)、人間らしさ(入力文関係なく、出力文単体での文としての自然さ)という3つの項目に関して、対話数が 9343 文の小さな学習データに対し、転移学習なしのモデルと比較してすべての項目で転移学習ありのモデルが大幅に良いスコアを示した。一方、大きな学習データに対しては、多様性は少し上がったものの他の指標に関しては転移学習をしても生成結果にはそれほど大きな効果は表れなかった。</p>						
2. 研究成果実績の概要 (英訳)						
<p>Compared to English and Chinese, there are not many high quality publicly available corpora for Japanese in the task of chat dialog response generation. Therefore, in order to achieve a good enough performance in chat dialog generation with small and low quality data, this study utilized transfer learning from Chinese and English in the Transformer based model. Three Japanese corpora and a corpus collected from Twitter were used as the dataset to generate responses. The average value of the distinct-1 as an automatic evaluation index of the generated results was 0.368 without transfer learning, and 0.412 for the transfer learning model. In terms of human evaluation, the model with transfer learning scored significantly better on all three indices: sentence connection, informativeness, and humanness, compared to the model without transfer learning, for a small training dataset with 9343 sentences of dialogs.</p>						
3. 本研究課題に関する発表						
発表者氏名 (著者・講演者)	発表課題名 (著書名・演題)	発表学術誌名 (著書発行所・講演学会)	学術誌発行年月 (著書発行年月・講演年月)			
柳瀬優作, 張逸群, 齋藤博昭	日本語対話応答生成における他言語からの転移学習の適用	人工知能学会第36回全国大会	2022年6月15日			