様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １４ 日現在

機関番号：１２６０８
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１１９

研究課題名（和文）大規模データにおけるエンコーダ・デコーダモデルの効率的な学習

研究課題名（英文）Efficient training for neural encoder-decoders on a large amount of training data

研究代表者
高瀬　翔（Takase, Sho）

東京工業大学・情報理工学院・助教

研究者番号：４０８１７４８３

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究の目的は逐次的にモデルを更新可能なエンコーダ・デコーダモデルの実現である。これを実現するためには、質の良いエンコーダ・デコーダが必要であり、本研究では、主にこれの構築に注力した。研究成果として、質の良いエンコーダ・デコーダ（あるいはデコーダ部分に該当するニューラル言語モデル）を構築する手法をいくつか提案した。これらの研究成果について記した論文は、EMNLP、NAACL、AAAI、NeurIPSといった、自然言語処理、人工知能、機械学習分野でのトップ会議に採択されている。

研究成果の学術的意義や社会的意義
ニューラルネットワークの導入により、機械翻訳や要約において、計算機が流暢な出力を行えるようになってきている。しかしながら、計算機の出力した翻訳や要約と、人手で作成したものとの一致率はまだ低く、改善の余地があることが伺える。本研究課題での成果は、従来の機械翻訳器や要約器の性能を引き上げるものであり、この手法を導入することにより、より良い出力が得られると期待できる。特に、本研究では、要約タスクにおいて、人手で設定した要約率に応じた出力を可能にする手法を提案しており、これにより、計算機の出力の柔軟性が向上すると考えられる。

研究成果の概要（英文）：The purpose of this research is to explore a method to update parameters of neural encoder-decoders every time we obtain an additional training data. For this purpose, we have to construct a sophisticated neural encoder-decoder. In this research, I focused on constructing such sophisticated neural encoder-decoders, and proposed several methods for the construction. Research papers on these methods are accepted at EMNLP, NAACL, AAAI, NeurIPS, that are top-tier conferences on Natural Language Processing. Artificial Intelligence, and Machine Learning.

研究分野：自然言語処理

キーワード：自然言語処理　ニューラルネットワーク　機械翻訳

１．研究開始当初の背景

機械翻訳、自動要約、対話などの自然言語生成タスクは、近年、エンコーダ・デコーダモデルと呼ばれる、ニューラルネットワークを用いた条件付き言語モデルにより、目覚ましい発展を遂げている。例えば翻訳については、2015 年にエンコーダ・デコーダモデルが統計的機械翻訳手法の性能を上回っており、以降も、エンコーダ・デコーダモデルによる翻訳手法の性能は向上し続けている。

一方で、エンコーダ・デコーダモデルは過学習しがちであり、汎化性能向上のためには大量の学習データが必要となる。このため、学習データが増えた場合や適用先の文書の傾向が変わった場合（例えば文書のドメインが「新聞」から「法律」に変わった場合）は、既に性能の高いモデルがあったとしても、そのモデルを再利用することはできず、全ての学習データを用いて再学習しなければならないという問題がある。現実的には、１度学習したモデルを使い続けられることはまれであり、新聞やニュース、Web 上のテキストなど、日増しに増えていくデータに出現する、新たな単語（人名や製品名など）や時事への対応が必要となる。このため、新たなデータを追加する度、過去のデータも含めた、全てのデータを用いてゼロから再学習を行わなければならない。加えて、大量のデータを入力した際のエンコーダ・デコーダモデルの学習自体も極めて遅く、例えば、翻訳で十分な性能を達成するためには、１週間以上を要する。さらに、高い性能を達成するためには、最適なハイパーパラメータの探索も必要となる。これらの要因により、学習データが増えた際の、エンコーダ・デコーダモデルの更新は、多大な労力を割かなければならない、という状況が、本研究開始当初の状況であった。

なお、ドメインへの適応が難しい、という状況は依然として変化していないが、近年では、大規模な訓練データで学習したモデルを、別のデータセットに適用し、高い性能を達成可能であるという報告が多数されている。

２．研究の目的

上記の状況をふまえ、本研究では、学習データが追加された際の、学習コストの大幅な低下を実現する。具体的には、大量のデータを用いて学習を行ったモデルの出力と、追加された学習データでのみ学習を行ったモデルの出力を適切に組み合わせ、大量の学習データから得た高い汎化性能を維持したまま、追加されたデータにも適した出力を実現する。本研究で目指すモデルの概要を図１に示す。

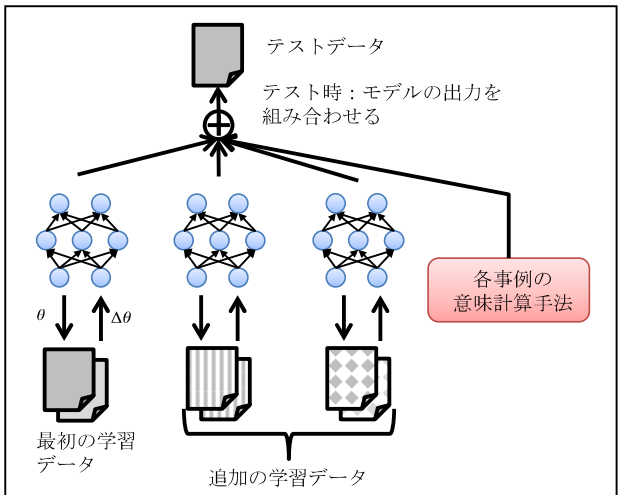

図１　本研究の目標：学習データ毎にモデルを学習、テスト時、各事例毎にモデルの出力を組み合わせる。複数のモデルを独立に学習可能なため、データが追加された際はそのデータでのみ学習したモデルを追加すれば良い。テスト時の各事例と学習データとの意味的な近さを計算し、モデルを適切に組み合わせることで、大量の事例で学習した１個のモデルと同等の性能の達成を目指す。

本研究の手法を実際に運用する際は、学習データが追加される毎に新たなモデルを学習し、テスト時には、複数のモデルから、入力毎に適したモデルを選択する。仮に、入力の事例と全く同じ事例で学習したモデルがあった場合、そのモデルが最適な出力を行えるだろう。入力事例と全く同じ事例が存在しなかった場合には、入力事例と意味的に似た事例を多く学習に用いたモデルの出力が適している。すなわち、モデルの出力の組み合わせは、翻訳や要約対象の文と似た文を多く含む学習データを探せれば良く、対象の文と学習データに含まれる文の意味計算を行い、互いの意味的な類似度を計算できれば良い。しかしながら、句や文は単語の組み合わせにより構成されるため、似た意味の表現を何種類も作成可能である。このため、句や文の意味計算自体、自然言語処理分野で長年取り組まれてきた課題であり、人間の評価結果との間の相関もいまだに低い。従って、本研究のように、与えられた事例について、最適なモデルを選択するという手法は極めて挑戦的であると言える。

３．研究の方法

本研究では、主に、機械翻訳や要約のような系列タスクに着目し、これに用いるためのニューラルエンコーダ・デコーダモデルを対象に、改良・実験を行い、研究を進めていく。研

究の目的としては、新たに学習データが追加された際、全てのデータを用いての再学習を行わず、追加データでの学習のみで、高い汎化性能を維持したまま、追加データにも適応したモデルを実現することであり、具体的には、各データで学習した複数のモデルの出力を適切に組み合わせることによって、学習データ全体を使って 1 個のモデルを得た際と同程度の性能を達成する。本研究の目指す手法では、組み合わせるモデルの数は問わないため、学習データが追加される度に、追加されたデータでのみエンコーダ・デコーダモデルを学習すれば良い。これを実現するためには、1. 質の良いエンコーダ・デコーダモデルの構築手法の探求、2. 新たに入力事例が与えられた際に最適なモデルを選択する手法の探求の 2 つが必要となると考えられる。

質の良いエンコーダ・デコーダモデルの構築については、文字通り、従来のエンコーダ・デコーダモデルの品質を向上させるようなモデルや学習手法を提案することが目的である。ニューラルエンコーダ・デコーダのデコーダ部分に該当する、ニューラル言語モデルの品質向上も含まれる。

新たに入力事例が与えられた際に最適なモデルを選択する手法の探求については、入力事例と訓練事例中のどの事例が近しいかを選択する、意味敵類似性を計算する手法の構築が必要となる。

4. 研究成果

上記のように、本研究の目的は逐次的にモデルを更新可能なエンコーダ・デコーダモデルの実現である。具体的には、大量の学習データで学習済みの、高性能なエンコーダ・デコーダについて、学習データが追加された際に、既存の学習データも含めた全データでの学習ではなく、新規に追加された学習データでのみ学習を行ったモデルを用意し、大量のデータで学習したモデルと適切に組み合わせる手法の実現を目指す。本研究において、高品質なエンコーダ・デコーダモデルは必須である。

高品質なエンコーダ・デコーダモデル実現のために、出力する確率分布の質を高める研究に取り組んだ。具体的には、エンコーダ・デコーダモデル（あるいはデコーダ部分のみのニューラル言語モデル）が出力する確率分布について、通常は 1 つの確率分布のみを使用するが、複数の確率分布を計算し、重み付き和を計算して最終的な確率分布とする手法を導入した。これは、Mixture of Softmaxes という手法として知られており、通常は、ニューラルネットワークの最終層のみから確率分布を計算するが、これを拡張し、複数層のニューラルネットワークの各層から確率分布を計算し、組み合わせる手法を提案した。

また、エンコーダ・デコーダの入力は、単語ではなく、部分文字列が使われるようになって久しいが、文字情報も欠かせない情報として存在していると考えられる。これを利用するために、部分文字列を構成する、文字や部分文字列から単語の分散表現を計算し、エンコーダ・デコーダモデル（や、上記と同様に、ニューラル言語モデル）で利用する手法を提案した。上記の手法は、言語モデルタスクにおける、標準的なベンチマークデータセットで、発表当時、世界一の性能を達成していた。例として、本研究での、ニューラル言語モデルに適用した際の、当時の最先端との研究の比較を表 1 に示す。表 1 では、Perplexity という、負の対数尤度を元にした値で各手法を比較しており、小さいほど良い手法であることを示す。表 1 より、提案手法が最も良い性能を達成していることが分かる。

Model	#Param	Valid	Test
Variational LSTM + IOG (Takase et al., 2017)	70M	95.9	91.0
Variational LSTM + WT + AL (Inan et al., 2017)	28M	91.5	87.0
LSTM with skip connections (Melis et al., 2018)	24M	69.1	65.9
AWD-LSTM (Merity et al., 2018)	33M	68.6	65.8
AWD-LSTM + Fraternal Dropout (Zolna et al., 2018)	34M	66.8	64.1
AWD-LSTM-MoS (Yang et al., 2018)	35M	63.88	61.45
Proposed method: AWD-LSTM-DOC	37M	**60.97**	**58.55**
Proposed method: AWD-LSTM-DOC (fin)	37M	**60.29**	**58.03**
Proposed method (ensemble): AWD-LSTM-DOC × 5	185M	**56.14**	**54.23**
Proposed method (ensemble): AWD-LSTM-DOC (fin) × 5	185M	**54.91**	**53.09**

表1 言語モデルの標準的ベンチマークデータでの性能
性能はPerplexityで評価しており、低いほど良い

加えて、実用的なエンコーダ・デコーダの開発として、生成型要約タスクの一種である、見出し文生成の性能向上に取り組んだ。具体的には、生成中の単語の文内での位置を正弦波と余弦波を用いて表す際に、これらの周期の値を変化させることで、所望の長さの生成を可能にした。既存の研究では軽視されがちであるが、所望の長さの出力を行うことは要約生成の実用化を目指す上では必須の技術である。また、機械翻訳と要約、言語横断要約など複数のタスクについて、タグを用いてタスクを表現することにより、ひとつのエンコーダ・デコーダでの学習・生成を可能とし、複数タスクでの性能向上を実現した。

さらに、実験を繰り返す中で得た知見から、エンコーダ・デコーダモデルに用いる単語の埋め込みについて、質を維持したまま要するパラメータ数を低減させる手法を提案した。機械翻訳や要約タスクにおいて、通常のエンコーダ・デコーダの性能と遜色ないまま、パラメータ数の削減に成功している。

上記の研究成果は、それぞれ、自然言語処理や人工知能分野、機械学習分野のトップ会議である、EMNLP、NAACL、AAAI、NeurIPS で発表している。また、採択されていない論文も arXiv 上に公開済みであり、主要なものについては、実験用のコードも公開し、広く利用可能な状態となっている。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Sho Takase, Sosuke Kobayashi
２．発表標題 All Word Embeddings from One Embedding
３．学会等名 34th Conference on Neural Information Processing Systems
４．発表年 2020年

１．発表者名 Sho Takase, Naoaki Okazaki
２．発表標題 Positional Encoding to Control Output Sequence Length
３．学会等名 North American Chapter of the Association for Computational Linguistics（国際学会）
４．発表年 2019年

１．発表者名 Sho Takase, Jun Suzuki, Masaaki Nagata
２．発表標題 Direct Output Connection for a High-Rank Language Model
３．学会等名 Empirical Methods in Natural Language Processing（国際学会）
４．発表年 2018年

１．発表者名 Sho Takase, Jun Suzuki, Masaaki Nagata
２．発表標題 Character n-gram Embeddings to Improve RNN Language Models
３．学会等名 Thirty-Third AAAI Conference on Artificial Intelligence（国際学会）
４．発表年 2019年

１．発表者名 高瀬翔，岡崎直観
２．発表標題 位置エンコーディングを用いた出力長制御
３．学会等名 言語処理学会年次大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関