3章

ディープラーニングを用いたテキスト生成技術と気象ニュース原稿生成への応用

駒井雅之†, 千葉詩音†, 金　秀明†, 武田光平†

キーワード：自然言語処理，ニューラルネットワーク，ディープラーニング，自然言語生成，エンコーダデコーダ

1．まえがき

ニューラルネットワークやディープラーニングの爆発的流行に伴い，人工知能技術が注目される中，ビジネスへの応用に対する期待が高まっている．代表的な適用事例として，問合せ対応業務や，審査業務がある．

業務という視点で考えると，文書作成業務も効率化が期待される分野であろう．東レ経営研究所の調査によると，文書作成をはじめとした資料作成系の業務が，全業務の26.9%を占めるといった報告もある[1]．文書を一定の品質で自動生成できれば，業務の効率化や働き方改革につながると期待される．このような背景があり，人工知能技術によるテキスト生成技術が大きな注目を集めている．

2016年度，われわれはメディア業界の企業と共同で「人工知能技術を用いた原稿作成支援」プロジェクトを実施した．プロジェクトでは，アナウンサが読み上げる気象ドメインのニュース原稿を，人工知能技術を用いて自動的に生成することを試みた．

本稿では，近年の代表的なテキスト生成手法について説明したのち，気象ニュース原稿の自動生成に向けた実証実験について述べる．

2．テキスト生成の基礎技術

システムによるテキスト生成技術は，二つの主流のアプローチがある．一つはテンプレートベースの手法であり，もう一つは統計的な手法である．

テンプレートベースの手法は，事前にテキストのテンプレートを設計し，必要な値を当てはめることでテキストを生成する方式である．例えば，気象ニュースの分野ならば”，［地方］では［雨量］の雨が降りました．”というテンプレートを設計し，対応する“［地方］”や“［雨量］”の値を当てはめることでテキストを生成する．

テンプレートベースの手法は，単純なテキストならば充分な品質で生成可能であり，またテンプレートの見直しが容易である．しかし，複雑なテキスト生成の仕組みを実現することは難しい．

統計的な手法は，データ駆動型の方式である．古典的な手法としては，N-gram＊1等の言語モデルによるアプローチがある．また，近年ではディープラーニングを用いたテキスト生成手法が主流である．その中でも，プロジェクトで採用したエンコーダデコーダ方式は汎用性が高く，テキスト生成の入力として，画像[2]，時系列データ[3]など広く利用可能である．

エンコーダデコーダ方式は，テキストをはじめとした入力データを符号化(エンコード)することで中間表現を得て，得られた中間表現を復号(デコード)して出力データを得る．エンコードとデコード時の計算にて，ディープラーニングを用いる．

テキストを入力とし，テキストを出力する仕組みでは，sequence-to-sequence（seq2 seq）[4]と呼ぶアルゴリズムが支配的な地位にある．われわれの検証においてもseq2 seqをベースとしたアルゴリズムを用いている．

seq2 seqは教師あり学習のアルゴリズムであり，二つの再帰型ニューラルネットワークを用いて，入力の単語系列$\mathbf{x} = (x_1, ..., x_{T_x})$を，別の単語系列$\mathbf{y} = (y_1, ..., y_{T_y})$へと変換する．ただし，$T_x$と$T_y$は各々のテキストの単語列長を意味し，両者は異なりうる．またx_t, y_tは対応する語の成分のみが1となり，他の成分が0となっているようなone-hot vectorであり，$x_t \in \{0, 1\}^{|V^x|}$, $y_t \in \{0, 1\}^{|V^y|}$である．ここで，V^xは入力データ側の語彙であり，V^yは出力データ側の語彙である．V^xとV^yは同一の場合もある．seq2 seqの具体的な計算式の説明については，原著論文に譲る．

seq2 seqを日本語のテキストに適用する場合，一般には形態素解析＊2によって分かち書きした結果を与える．図1の①を例に挙げると，入力の単語系列$\mathbf{x}$として（“停滞”，

＊1　連続するn個の文字列や単語列を切り出す手法
＊2　単語間の区切り，語の活用系，品詞を決定する処理

†株式会社エヌ・ティ・ティ・データ
"Natural Language Generation using Deep Learning: An application to weather news generation" by Masayuki Komai, Shion Chiba, Hideaki Kin and Kohei Takeda (NTT DATA Corporation, Tokyo)

映像情報メディア学会誌 Vol. 74, No. 1, pp. 12～15（2020）

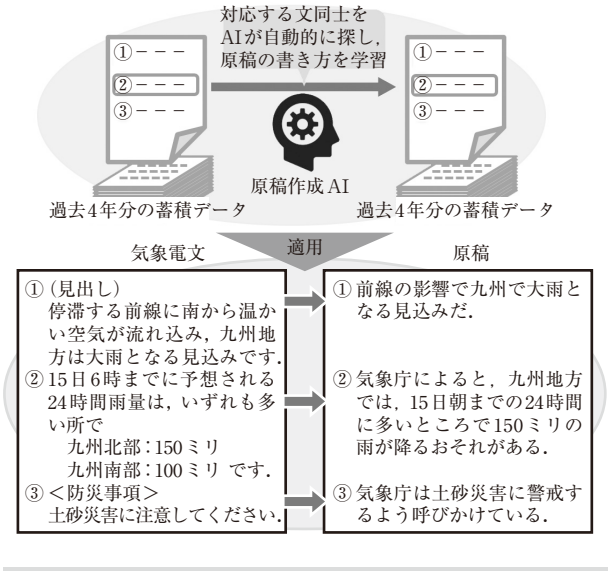


図1　ニュース原稿の自動生成のイメージ

“する”，“前線”，…），出力の単語系列$\mathbf{y}$として（“前線”，“の”，“影響”，…）が対応する．

seq2 seqは教師あり学習の手法であるため，テキストを生成するためには，事前に学習処理が必要となる．学習処理のためには学習データ$\mathbf{D} = \{(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^{|D|}, \mathbf{y}^{|D|})\}$を必要とする．この学習データについて，「すべての事例を学習する」操作を複数回実施することが一般的であり，その時の回数のことepoch数と呼ぶ．

学習処理を終えると，テキスト生成が可能となる．すなわち，何らかの新規の単語系列$\hat{\mathbf{x}}$が与えられたとき，seq2 seqの出力としての単語系列$\hat{\mathbf{y}}$を得ることができる．注意事項として，seq2 seqが真に出力する情報は，各時刻での単語の生成確率である．そこで，最適な単語系列を探索するためのアルゴリズムが必要となるが，探索アルゴリズムとしてビーム探索やグリーディ探索をしばしば用いる．

以上，統計的なテキスト生成手法として，代表例であるseq2 seqを説明した．実際には，オリジナルのseq2 seqではテキスト生成において何らかの問題が発生することが多く，別の発展的なアルゴリズムを用いることが多い．発展的なアルゴリズムとしては，アテンション機構[5]を用いたseq2 seqが著名であり，われわれの検証においても，これを用いている．その他，出力するテキストについて多様性を促進する試み[6]や，入力データの一部を複製するコピー機構[7]を提案している研究もある．さらに，再帰型ニューラルネットワークの代わりに，画像処理の分野で広く用いられている畳み込みニューラルネットワークを用いたアルゴリズム[8]もある．近年，自己アテンション機構[9]を用いたTransformer[10]という別種のアーキテクチャも提案されている．詳細については，原著論文を参考していただきたい．

3．気象分野のニュース生成

ここからは，本記事の主題である「気象ニュース原稿の生成」について述べる．われわれの実験では，2節にて説明したseq2 seq+アテンション機構を用いて「気象電文」から「ニュース原稿」を自動生成することを試みた．本検証を行う上で，以下の三つの課題を解決する必要があった．

課題（1）：入出力データの調達

われわれが直面した最大の問題は，「学習用の入出力データがない」ということであった．具体的には，われわれが参画したプロジェクトでは，入力データと出力データとが紐づけされておらず，入力に対して正解となる出力が判断できない状態であった．入出力データを利用できない場合，学習処理を実行できないため，テキスト生成技術を検証することが難しい．

そこでわれわれのケースでは，入力データとの類似度を用いて，正解となる出力データを紐づけた．類似度の計算時，ベクトル化した入出力データ間のコサイン類似度を計算する．各ベクトルは，語彙の大きさの次元を持ち，各成分は対応する単語の出現頻度などの値を持つ．入出力データ間の類似度が閾値を超えたデータ対を，入出力データとして抽出した．

課題（2）：テキストヴァリエーションの圧縮

課題（1）で述べたように，われわれの実験においては整備された入出力データがなかったために，機械的に収集した入出力データを用いた．

このデータを用いて予備実験をしたところ，テキスト生成モデルが充分に機能しなかった．実際の出力結果を見たところ，数値や地名といった，低頻度語の周辺で多くの誤りが見られた．一般的に，単語のカテゴリーが同一の場合も，別の用語として扱う．例えば，雨量に関する単語“100ミリ”と，“99ミリ”を別物として扱う必要がある．このような具体語の異なりが，テキストのヴァリエーションを肥大化させてしまい，学習処理が不充分となっているのではないか，と仮説を立てた．

そこでわれわれは，固有表現認識*3技術や独自に構築したキーワード抽出技術を用いて入力データのマスキング処理を行った．われわれが行ったマスキング処理は，テキスト中で発見されたキーワードを，キーワードのカテゴリー名に置き換える操作である．マスキング処理の対象は，雨量や地名といった表現，○○日といった時間的表現も含む．マスキング処理の例を図2の左に示す．マスキング処理によって，雨量や地名などの具体語を生成できなくなるものの，マスキングされた文ならば充分な品質で生成できるようになった．また，語彙サイズの圧縮に寄与することからテキスト生成モデルの高速化にも繋がった．

*3　地名や人名といった固有表現を認識する技術

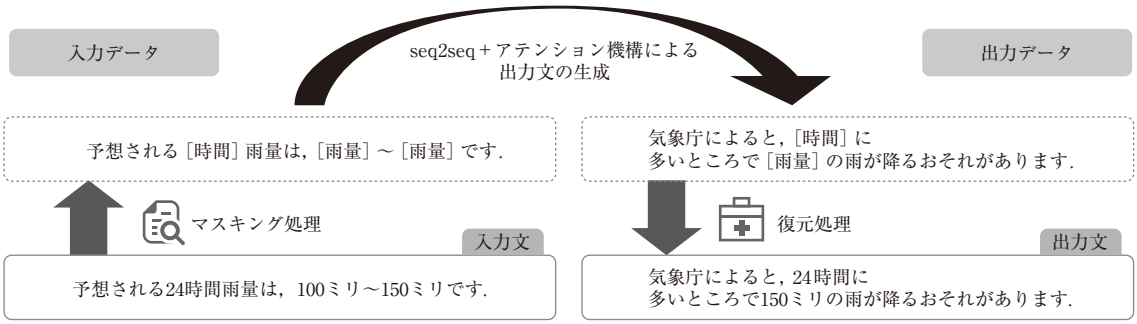


図2　システム全体像
seq2seq+アテンション機構の処理の前後に，マスキング処理（図左）と復元処理（図右）がある．

課題（3）：マスキングの事後処理

マスキングした入出力データでテキスト生成モデルによる学習や生成処理を行った場合，出力するニュース原稿もマスキングされているため，何らかの方法によって正規の文へと変換しなければならない．

本検証では，ルールと文脈情報を併用して復元処理を行った．復元処理の例を図2の右に示す．まず，入力文と出力文のマスキング部分を比較し，それらに1対1の対応があれば，対応する具体語を出力文にコピーする．

1対1以外の場合は，マスキング間の類似度を用いて具体語をコピーする．具体的には，入力文側のすべてのマスキング部と，出力文側の一つのマスキング部をベクトル化し，入出力間の類似度が最大となる具体語をコピーする．ベクトル化時は，マスキング部周辺の単語の情報を用いた．類似度計算の方法としては，課題（1）と同様にコサイン類似度を用いた．

次に実験および実験結果について説明する．われわれの実験において，「テキスト生成技術はどの程度実用に即するのか」，「実用のためにはどの程度のデータが必要か」といった観点で評価する必要があった．そこで，データ量を5,000件，10,000件，20,000件の3パターンで，システムの精度を算出した．評価データとして859件を用いた．定量評価指標として，広く使われているRouge*4という指標を用いた．

実験結果について，システムの精度を図3に示す．全体の傾向として学習繰り返し数が10の時の精度が最も高く，その後に穏やかに悪化するという過学習特有の結果となった．システムが出力したニュース原稿を定性的に評価したところ，何らかの文法的な誤りが発生した事例は，全体の10%程度であった．一方，入力データと矛盾するような内容的な誤りについては，1文あたり平均的に1～2語あった．

誤った事例を分析した結果，いくつかの問題点が明らかになった．よく見られたのは，時制の誤りである．例えば，未来のことを書くべきなのに，「～となっている」といったように，事実のように述べることがあった．また，日程について，「あす」，「きょう」と述べたいのだが，日程に関する意味理解が充分でなく，単に「○○日」といったようにニュースに適さない表現も見られた．ニュースというドメインであるがゆえに，誤りが許容されづらいので，実利用に向けてはこのような誤りを抑制する必要がある．

最も致命的な誤りは，災害に関わる情報について言い誤ることである．ニュースのようなドメインでは，災害など住民にとって危険性があることは何よりも正確に伝えねばならない．しかしながら，現行手法は，例えば，入力データが「土砂災害」と「河川の氾濫」に言及していても，「土砂災害」のみが出力されるなど，伝えるべき情報が欠落する場合があることがわかっており，今後の検討課題である．

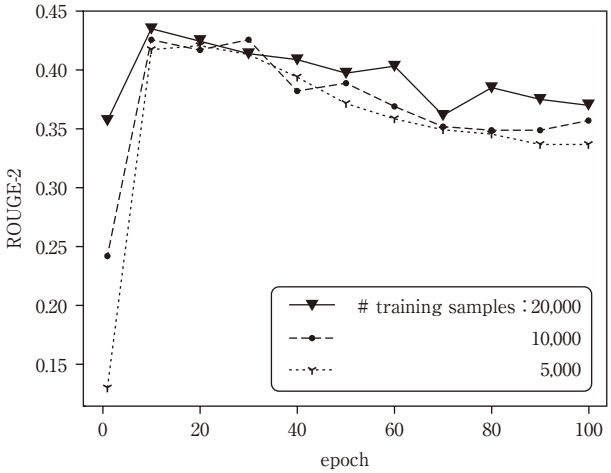


図3　テキスト生成技術の精度推移

4．むすび

本記事は，基本的なテキスト生成の仕組みと気象ニュース原稿の自動生成に関する実証実験について述べた．前者については，エンコーダデコーダ方式のseq2seqというテキスト生成モデルや，アテンション機構といった要素技術について解説した．後者については，プロジェクト中に発生した課題や解決策に加え，実験結果に対する考察を述べた．

*4　テキスト生成技術の評価指標の一種．システムの出力と正解の出力に対し，両者のN-Gramの一致率を測る．

われわれが伝えねばならない最も重要なことは，理論だけで解決できる現実の課題は極めて限られる，ということである．例えば，研究用途で用いられるデータは，充分に整備されており，せいぜい弱いノイズがあるのみ，ということが多い．しかしながら，現実のデータは欠損や誤り，分布の偏りがあるのが一般的である．さらには，データがないケースもあり，今回の実験のように機械的に収集したデータで代用せざるを得ない場合もある．

また，システムの最終的な目的は，多くの場合業務効率化のような，ビジネス上の課題の解決である．その目的を達成するため，アルゴリズム設計も重要となるだろうが，ユーザビリティも意識する必要があると考える．例えば，人工知能技術は一般的に完全な回答をすることは難しく，何らかの誤りを出力することがある．実業務に組み込む場合，人手によるチェックはほとんどのケースで必要となるだろう．そのために，システムはただ結果を出すだけでなく，チェックにおいて有用な情報も出力する必要があるだろう．今回のケースだと，生成文において尤度の低い部分を強調する，災害情報に関する辞書を構築し，入出力間で矛盾が発生したらアラートを出す，といった仕組みが考えられる．

最後になるが，最近のわれわれの取組みについても簡単に紹介したい．ニュース原稿の自動生成の次のステップとして，金融分野の融資稟議書の自動生成に取り組んでいる．これは融資の可否判断に必要な情報を文書化するものである．問題設定としては，本稿で取り組んだテキストからテキストを生成するという問題設定ではなく，構造化されたデータからテキストを生成する，という設定を想定している．ただし，金融ドメインのデータのため，入力となる構造化データは数値情報を多く含み，数値情報に関する高度な意味理解が必要となることが予想される．

研究活動は何らかの形で社会に還元できるものでなければならないと考える．われわれNTTデータは目先の真新しさを求めるだけでなく，真に社会へと貢献できるような技術開発に努めていきたい．

（2019年10月11日受付）

〔文　献〕

1）松井滋樹：“ワーク・ライフ・バランスの実現のために（実践編）－タイムマネジメントの実践－”，東レ経営研究所（2012）

2）O. Vinyals, A. Toshev, S. Bengio and D. Erhan: "Show and Tell: A Neural Image Caption Generator", In CVPR（2015）

3）S. Murakami, A. Watanabe, A. Miyazawa, K. Goshima, T. Yanase and H. Takamura: "Learning to Generate Market Comments from Stock Prices", In ACL（2017）

4）I. Sutskever, O. Vinyals and Q.V. Le: "Sequence to Sequence Learning with Neural Networks", In NIPS（2014）

5）D. Bahdanau, K. Cho and Y. Bengio: "Neural Machine Translation by Jointly Learning to Align and Translate", In ICLR（2016）

6）J. Li, M. Galley, C. Brockett, J. Gao and B. Dolan: "A Diversity-Promoting Objective Function for Neural Conversation Models", In NAACL（2016）

7）J. Gu, Z. Lu, H. Li and V.O.K. Li: "Incorporating Copying Mechanism in Sequence-to-Sequence Learning", In ACL（2016）

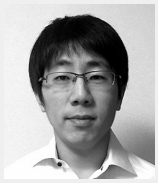
8）J. Gehring, M. Auli, D. Grangier, D. Yarats and Y.N. Jonas: "Convolutional Sequence to Sequence Learning", In ICML（2017）

9）Z. Lin, M. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou and Y. Zhouhan: "A Structured Self-attentive Sentence Embedding", In ICLR（2017）

10）A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin: "Attention is All You Need", In NIPS（2017）

駒井　雅之　2016年，奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年，（株）NTTデータに入社．専門は自然言語処理と情報検索．テキスト生成技術，ランキング学習，文書分類に関する研究開発に従事．

千葉　詩音　2017年，東北大学大学院情報科学研究科修士課程修了．同年，（株）NTTデータ入社．テキスト生成技術や分析自動化アルゴリズムの研究開発に従事．



金　秀明　2013年，京都大学大学院理学研究科物理学・宇宙物理学専攻博士後期課程修了．同年，日本電信電話（株）入社．2017年，（株）NTTデータ転籍の後，2019年より日本電信電話（株）．機械学習，確率過程を用いた統計解析モデルの開発に従事．博士（理学）．

武田　光平　2002年，東京大学工学系研究科航空宇宙工学専攻修士課程修了．同年，NTTデータ入社．マルチメディア処理，情報検索，ディジタルアーカイブ，対話ロボットなどに関するR&Dおよびソリューション開発を経て，現在，次世代のイノベーションの種となる技術開発に携わるチームを率いる．