研　究　報　告　書

「様々な形式のデータを言語で柔軟に記述する汎用的技術の開発」

研究タイプ：通常型
研究期間：　2016 年 12 月～2020 年 3 月
研 究 者：　高村　大也

1.　研究のねらい

本研究の狙いは、様々な形式のデータを言語で表現するための汎用的な方法を開発することである。センサの普及やストレージの拡充により、金融データ、生理指標データ、人間の行動データなどを含む様々なデータが、数値データ、カテゴリカルデータ、テキストデータなど様々な形式で大量に蓄積されている。しかし、こういったデータが何を意味しているかを知るためには、人間が解釈する必要がある。さらにデータの解釈を他の人間に伝えるためには言語表現(テキスト)という形で説明する必要がある。これをコンピュータにより自動的に行うことができるようにすることで、蓄積されたデータが有効利用できるようになる。また、用途などにより、長さを含めた様々な制約が、生成されるテキストに課されることも応用上は多い。そこで本研究では、様々な形式のデータをテキストで柔軟に表現するための汎用的な方法を開発することを目的とする。より具体的には、《課題1》時系列数値データに対する説明テキスト生成手法の開発、《課題 2》複数のデータに対する説明テキスト生成手法の開発、《課題 3》テキスト生成において出力を柔軟に制御する技術の開発、の3つの課題を実施する。

これにより、様々なデータの意味が人間に理解しやすくなる。特にデータのドメインの専門家でない者が労力をかけずにデータを解釈したいときなどに有用である。また、テキスト化されたデータはマイニング手法を適用しやすく、大量のデータにおいて何が起こっているのかを把握しやすくなる。

データをテキストで表現するための手法としては、言語生成能力の高い encoder-decoder ニューラルネットワークを拡張したモデルなどを考えている。

2.　研究成果

（1）概要

いくつかのデータセットを題材に、各課題に関し研究開発を実施した。時系列数値データとしては、気象データと経済指標時系列データを準備し、それぞれに対する説明テキストを生成するニューラルネットワークを構築した。気象データは 2 次元データの時系列であるので、畳み込みニューラルネットワークなどで 2 次元データをエンコードしてから時系列モデルに入れる仕組みを用いた。このような手法を用い、時系列数値データから、高い精度で説明テキストを生成できることを実証した。また、これらのデータは複数の指標からなっており、複数のデータに対する説明テキスト生成手法の開発も同時に行った。これらのデータからのテキスト生成に加え、サッカーのプレーデータからのテキスト速報の生成や、バスケットボールのスタッツからの試合サマリの生成などを通し、技術開発を行った。サッカーのプレーデータは、プレーイベントが起こった時間、選手の位置、またプレーの種類など多様なデータから成る。これらのデータをエンコードし、デコード時には適切なプレーイベントにアテンションを当てる仕組みを

導入することで、より正確なテキスト速報を生成した。また、スタッツからの試合サマリの生成では、スタッツのどの部分に言及するかを明示的に決定する仕組みと、これまでどの部分に言及したかを記憶する仕組みを導入することで、より正確な試合サマリを生成することに成功した。出力テキストの制御技術の開発に関しては、テキストの出力長の制御技術、応答生成におけるテキストのスタイルの制御技術、また試合サマリ生成におけるサマリ著者の執筆スタイルの制御技術について、開発および検証を行った。また、上記 3 つの各課題に加え、近年提案された多くのテキスト生成技術を俯瞰的に整理し、どのような選択肢があり、どのような場合にどの手法を用いるべきかについて議論した。さらに、それらの手法を包含する一般的な枠組みを提案した。

（2）詳細

《課題1》時系列数値データに対する説明テキスト生成手法の開発

時系列数値データとして、気象データを取り上げた(図 1)。これは、日本の各地における気象関連の 11 指標(気温、気圧、湿度など)の値を、現時点から1時間ごとに84時間先まで予測したものである。すなわち、各指標が二次元データの時系列で表現されることになる。これに対し、各時点での情報を畳み込みニューラルネットワークあるいは多層パーセプトロンでエンコードし、これをさらにリカレントニューラルネットワークに入れて時系列としてモデル化した。さらに、天気予報コメント生成時のメタ情報(月、日、曜日、時間、場所)を埋め込みベクトルで表現して追加入力とすることで、メタ情報に依存する言語表現を生成できるようにした[主要な学会発表 1]。

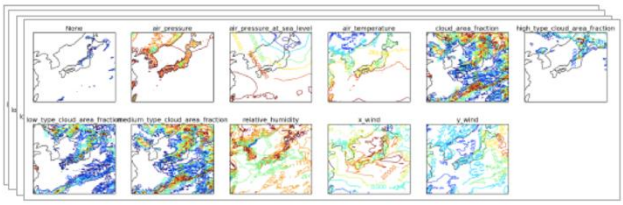

図1: 入力気象データの視覚化

《課題 2》複数のデータに対する説明テキスト生成手法の開発

経済指標時系列データからの市況コメントの生成技術の開発を行った[論文発表 4]。特に、複数の指標を入力として、日経平均の動きを、その変化要因(例:「米株高を好感し」)とともに記述する技術を開発した。具体的には、主な記述対象である日経平均株価に加えて，東証株価指数，ダウ平均株価，S&P500，上海総合指数，FTSE100 種総合株価指数，USD/JPY，USD/EUR の数値時系列を入力として、状況に応じて各指標にアクセスしつつ、日経平均の動きを変化要因とともに説明するニューラルネットワーク(図 2)を構築した。例えば、

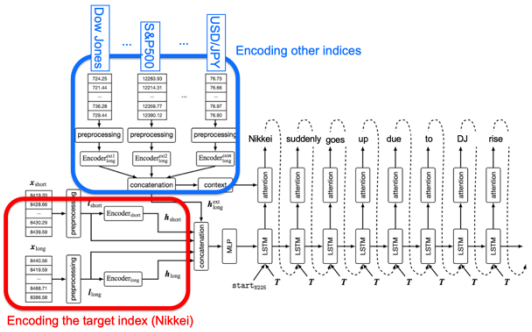

図 2: 複数の経済指標を入力とし説明テキストを生成するニューラルネットワーク

「*日経平均、続落で始まる　米株安で、円高が重荷*」

という市況コメントを出力することができる。これは NEDO 委託事業「人間と相互理解できる次世代人工知能技術の研究開発」との共同研究である。

また、サッカーのプレーデータからテキスト速報を生成する技術を開発した[論文発表 3]。プレーデータとは、試合中の各プレーイベント(シュート、パスなど)に対し、その時間、位置、関わった選手、など様々な情報を加えたデータである。関連しうるプレーイベントが複数存在するので、適合的にプレーイベントにアクセスすることで、より良いテキスト速報を生成することを可能にした。例えば、"*Odion Ighalo has a shot blocked.*"や、"*Chris Smalling is shown yellow card for a foul on Dwight Gayle.*"のようなテキスト速報が、開発手法により実際に生成された。

さらに、バスケットボールの試合のスタッツ(各選手の得点数、リバウンド数などのデータ)から試合サマリを生成する技術を開発した[論文発表 1]。この研究課題の特徴としては、各選手の得点数、リバウンド数など複数の種類のデータが入力となるため、入力データのサイズが大きいこと、また出力テキストも他のデータと比較すると長い文章であることが挙げられる(図 3)。単純なエンコーダ-デコーダでは、適切な文章が生成できないことがわかっている。本研究では、入力データのどの部分に言及するかを明示的に決定する、いわゆる内容選択をモデルに組み込んだ。さらに、サマリ全体として一貫性のある文章が生成できるように、テキストを生成しながら、そこまでに言及した情報をトラッキングし、それをテキスト生成に利用する仕組みを導入した。これは NEDO 委託事業「人間と相互理解できる次世代人工知能技術の研究開発」および産総研・東工大実社会ビッグデータ活用オープンイノベーションラボラトリ「ビッグデータを活用するデータ処理技術の開発」との共同研究である。

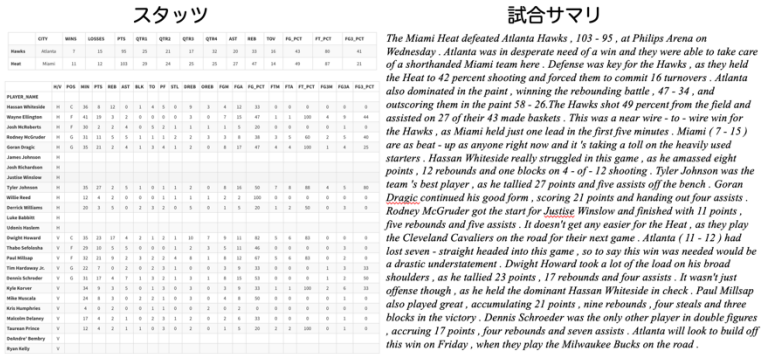

図 3: スタッツと試合サマリの例

《課題 3》テキスト生成において出力を柔軟に制御する技術の開発

対話における応答生成課題において、応答者の発話スタイルを考慮した応答生成モデル(図 3)を開発した[論文発表 2]。特に、応答者の過去の発話傾向と発話スタイルを関連付けてニューラル応答生成に利用することに成功した。ニューラルネットワークの構造を図 4 に示す。提案手法におけるスタイル付で必要なのは過去の発話集合であり、ここから発話傾向を導出することができるので、応答者がニューラルネットワークの訓練データに出現していなくても、利用可能である。これが類似の既存手法との大きな差異である。

また、日本語文圧縮という研究課題において、データセットを作成し、出力長を制御する技術の効果について検証した[論文発表 5]。さらに、上記で言及したバスケットボールの試合のサマリ生成においては、生成時に、該当するサマリ著者の情報を入れることで、テキストのスタイルが制御できることを示した[論文発表 3]。

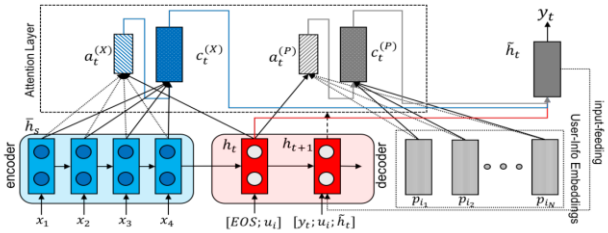

図 4: 応答生成のニューラルネットワーク構成

また、上記 3 つの各課題に加え、近年提案された多くのテキスト生成技術を俯瞰的に整理し、どのような選択肢があり、どのような場合にどの手法を用いるべきかについて議論した[主要な学会発表 2]。特に、入力データの性質に応じたエンコード方法の選択、内容プランの導入方法、エンティティのトラッキング方法などについて、議論した。

3. 今後の展開

本さきがけ研究では、時系列数値データを含む複数の入力に対し、適切な説明テキストを生成する技術、および出力テキストのスタイルなどを制御する技術を開発した。また、新たなデータセットに対して、どのようなテキスト生成モデルで臨むべきかについての戦略を整理することができた。実応用に向けて必要となることの一つとして、生成されたテキストの内容の根拠となるものを示せる技術が挙げられる。それによりユーザが納得して使用してくれるものと思われる。また、より幅広い利用に向けては、現実の問題では、大規模な訓練データが用意できないこともあり、小さな訓練データでモデルが学習できるようにすることが挙げられる。

また、よりリアルタイム性が重要となる実況中継のような応用のためには、いつ生成するのか、どのくらいの長さの出力をするのか、などが適切に決定できる必要があり、基盤技術の開発が必要となる。

4. 自己評価

研究用のデータセットが十分に揃っていない、また自動評価方法も確立しておらず人手での評価が必要である、などの点で、チャレンジングな課題であった。しかし、さきがけのサポートにより、時系列数値データを含む複数の入力に対し、適切な説明テキストを生成する技術が開発できた。また、個々の技術を俯瞰することでテキスト生成課題に臨む際の戦略が整理でき、それより汎用性が確保できた。研究の進め方に関しては、2 つの研究組織にまたがってプロジェクトをすすめることになったため、若干の難しさがあった。

技術としては、実用レベルに達していると考えられ、自動記事生成や速報テキスト生成などに利用できるので、適切な適用の場を見つけることで普及していくだろう。また、開発技術

の、スマートスピーカーへの応用や、実況生成などへの拡張などにより、さらに多くの場面で活用できるようになる可能性がある。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, Hiroya Takamura. Learning to Select, Track, and Generate for Data-to-Text. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. pp. 2102-2113.
2. Abdurrisyad Fikri, Hiroya Takamura, Manabu Okumura. Stylistically User-Specific Response Generation. International Natural Language Generation Conference (INLG). 2018. pp. 89-98.
3. Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, Manabu Okumura. Generating Live Soccer-Match Commentary from Play Data. AAAI Conference on Artificial Intelligence. 2019. pp. 7096-7103.
4. Tatsuya Aoki, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, Yusuke Miyao. Generating Market Comments Referring to External Resources. International Natural Language Generation Conference (INLG). 2018. pp. 135-139. 5. Shun Hasegawa, Yuta Kikuchi, Hiroya Takamura, Manabu Okumura. Japanese sentence compression with a large training dataset. The 55th annual meeting of the Association for Computational Linguistics (ACL). 2017. pp. 281-286.

(2)特許出願

研究期間累積件数:0 件(公開前の出願件名については件数のみ記載)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

主要な学会発表

1. 村上聡一朗, 笹野遼平, 高村大也, 奥村学. “数値予報マップからの天気予報コメントの自動生成”. 言語処理学会年次大会. 2017.
2. 高村大也. “データを言語で記述するための一般的枠組み”. 人工知能学会全国大会. 2020.

招待講演

1. 高村大也. “言語生成技術 -データを言葉で記述する-”. JST-NSF 国際連携シンポジウム. 2019.
2. 高村大也. “言語生成技術”. 2019 年度第 5 回画像応用技術専門委員会. 2020.

3. 高村大也. “言語理解と言語生成”. 医学情報学連合大会. 2019.
4. 高村大也. “自然言語分野におけるグランドチャレンジ”. MIRU 第 21 回 画像の認識・理解シンポジウム. 2018.