データ拡張を用いた固有表現抽出の精度向上

大林 弘明 トランスコスモス株式会社

Oobayashi.Hiroaki@trans-cosmos.co.jp

1 はじめに

コンタクトセンター業務及びヘルプデスク業務では、応対品質向上や VoC 分析ために問い合わせの会話音声を音声認識器を用いてテキストデータ (以下、会話文と呼ぶ。) に変換し記録している。さらに、業務効率化のために、テキストデータ化された会話文から、要約、キーワード抽出、 オペレータの回答支援、及び個人情報の匿名化等を自動化することを検討している。これらの自動化を実現するためには、様々な自然言語処理技術が必要であり、最初の段階として文章の中から氏名や住所といった固有表現の同定が必要である。これは一般に固有表現抽出 [1] と呼ばれる技術であるが、氏名や住所のような固有表現は、電話番号や電子メイルのような固有表現と異なり単純な正規表現では同定できない。そのため、より複雑な固有表現抽出の技術が必要となる。ここではニューラルネットを用いた固有表現技術を適用してみることとした。

近年、自然言語処理技術にニューラルネットワークを用いた事例 [2][3] が報告されている。固有表現抽出も同様に、ニューラルネットワークを用いた事例 [4][5][6][7] が報告されている。通常、ニューラルネットワークを学習させるためには大規模なデータが必要とされる。しかし、上記のような業務の会話文において、氏名及び住所などの固有表現を含む文章の割合は小さい。このような訓練データが少ない問題に対して、コンピュータビジョン分野の物体認識タスクでは訓練画像をランダムにクロップした画像を継ぎ合わせて新たな訓練画像を生成する手法 [8] がよく知られている。また、自然言語処理分野の機械翻訳タスクでは、頻度の少ない単語を別の文章に類似単語と置き換えて、訓練データを拡張する手法 [9] や翻訳したい言語対とは異なる言語対の対訳データを活用する手法 [10] などのデータ拡張の手法が報告されている。

本研究の対象であるコンタクトセンター業務及びヘルプデスク業務の会話文は、氏名、住所などの固有表現を含む文章の割合が小さいだけではなく、氏名、住所などの固有表現も多種多様な表現に渡っている。そこで、本研究では、上記の手法を参考に、文章内の氏名を表す固有表現部分を異なる氏名表現に置き換えることで、固有表現を含む文章のデータ拡張を行い、ニューラルネットワークを用いた固有表現抽出器の精度向上の検討を行った。さらに、データ拡張の増加率と固有表現抽出の精度との関係も検討した。

2 会話文

コンタクトセンター業務及びヘルプデスク業務の会話文の例を表 1 に示す。また、氏名表現の箇所を下線で示す。

表 1: 会話文の例

ユーザ:	すいません。
担当者:	はい。お客様相談センターのヤマダです
ユーザ:	あの、先日購入した製品について聞きたいのですが
担当者:	はい。確認しますので、お客様のお名前をいただけますでしょうか
ユーザ:	はい。タカハシと申します。
担当者:	・・・・・
ユーザ:	・・・・・

本研究では、実際の問い合わせ業務 (約 1700 件) の会話文を利用した。その約 1700 件の会話文の中から、氏名表現を含む文章を 1400 文章を抽出し、1000 文章を訓練用 (以下、訓練データ) に、400 文章を評価用 (以下、評価データ) に振り分けた。訓練データの 1000 文章についてのみ次章で説明するデータ拡張を施すこととした。

3 データ拡張

少ない訓練データを拡張する手法 [9][10] などが提案されている。本研究では、これらを参考にし、訓練データすべてに対し、氏名を表す部分にアノテーションを施し異なる氏名表現に置き換え、さらに、置き換えを繰り返し行うことで氏名表現を含む文章を増加さ

せることとした。本研究では、 繰り返し数を 2 回、5 回、10 回、50 回、100 回とし、訓練データの増加率と固有表現抽出の精度向上の関係を確認することとした。尚、置き換える氏名については、次の 2 点を満たすようにした。

- 氏名表現を繰り返し置き換える際、同じ氏名表現を繰り返さないように置き換えための氏名表現を多重に選択しない。
- 評価用データに含まれる氏名表現と同じ氏名表現で置き換えない。

次に、本研究で用いた音声認識器が氏名を含む発話をテキストに書き起こした場合、氏名を示す部分は、漢字もしくはカタカナでテキストに書き起こされていた。本研究では、音声認識器の特性に合せて、置き換える氏名表現のパターンを次の通りとした。

- 漢字表記の姓のみ (**山田**)
- 漢字表記の姓＋名 (**山田太郎**)
- カタカナ表記の姓のみ (**ヤマダ**)
- カタカナ表記の姓＋名 (**ヤマダタロウ**)

本研究では、置き換えるための異なる氏名表現として、インターネット上の疑似個人データサービス [11] を利用して、上記のパターンに従った擬似的な氏名表現を生成した。

4 固有表現抽出

ニューラルネットワークを用いた固有表現抽出器の有効性が報告 [6][7] されている。本研究では、これらを参考にし固有表現抽出をニューラルネットワークを用いた系列ラベリング問題として扱うこととした。また、本研究に用いた会話文は、比較的に短い文章で構成される傾向にあることから、単語毎にラベリングを行うのではなく、1 文字毎にラベリングを行うこととした。ラベリングのフォーマットには、IOE2[12] を用いた。ラベリングの例を表 2 に示す。

表 2: 文字毎のラベリングの例

入力	タ	ナ	カ	で	す	。
出力	B-PSN	I-PSN	E-PSN	O	O	O

また、ニューラルネットワークのモデルは、上記の報告 [4][5][6][7] を参考にし、LSTM を用いたニューラルネットワークを実装した。尚、最終的な出力は CRF と Softmax の 2 通りで比較検証を行った。本研究に用いたニューラルネットワークのモデル構成を、図 1 と図 2 に示す。

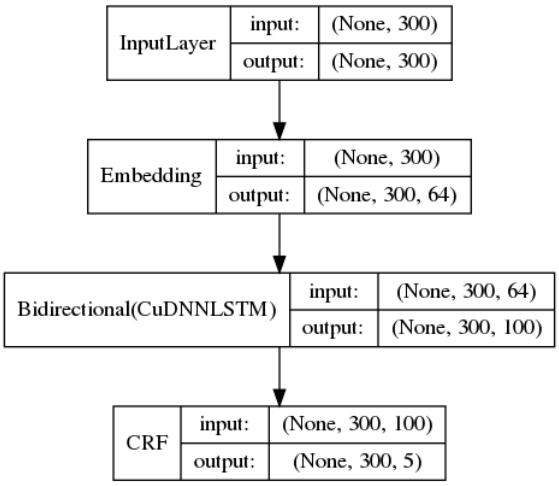

図 1: Embedding+LSTM+CRF

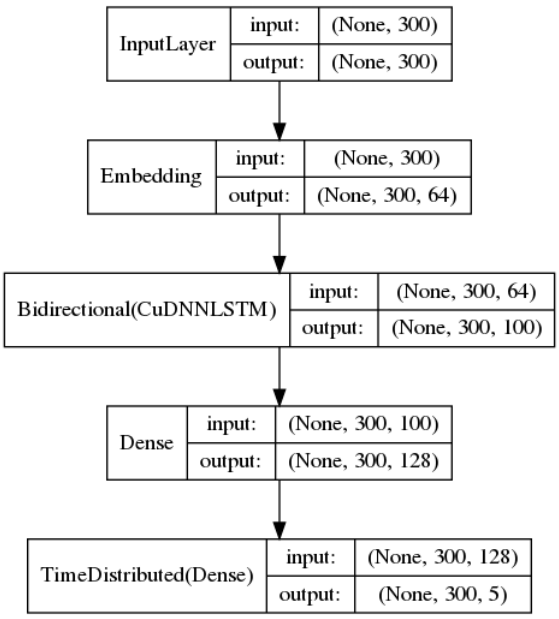

図 2: Embedding+LSTM+Softmax

5 結果

図 1 と図 2 のモデルを上記の手法を用いてデータ拡張した訓練データで学習させ、評価用データに含まれる氏名表現を抽出できるかを検証した。出力が CRF のモデルの検証結果を図 3 に、出力が Softmax のモ

デルの検証結果を図 4 に示す。x 軸はデータ拡張の増加率とし、y 軸は氏名表現の抽出精度とした。 出力が

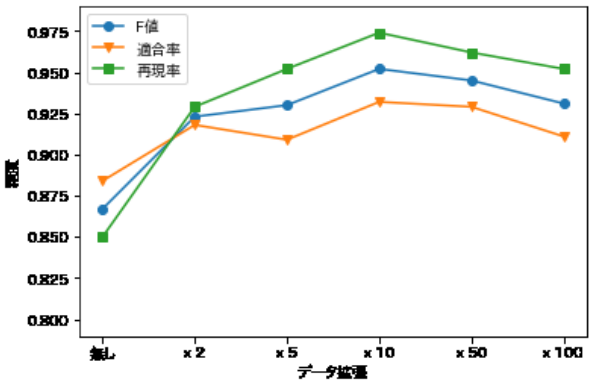

図 3: Embedding+LSTM+CRF

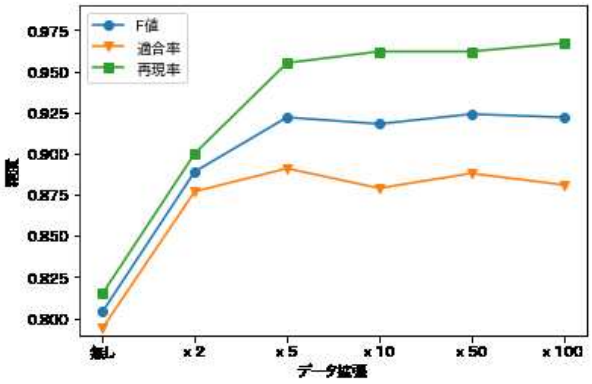

図 4: Embedding+LSTM+Softmax

CRF のモデルの場合、適合率が最大で 0.05、再現率が 0.12、F 値が最大で 0.1 向上することができた。また、出力が Softmax のモデルの場合、適合率が最大で 0.1、再現率が 0.14、F 値が最大で 0.12 向上することができた。いずれの出力の場合においても、データ拡張無しで学習させたモデルより、データ拡張有りで学習させたモデルの方が固有表現抽出の精度が高いことを確認できた。

次に、データ拡張無しで学習させたモデルとデータ拡張有りで学習させたモデルのラベリング結果の具体例を表 3 示す。具体例の文章には、”奥寺”という氏名表現が含まれているが、これは訓練データ及び置き換えるための異なる氏名表現に含まれなかった氏名表現である。下記の具体例 3 に示すように、データ拡張無しで学習させたモデルの場合では、”奥寺”を抽出できなかったが、データ拡張有りで学習させたモデルの場合は、”奥寺”を抽出することを確認できた。

表 3: ラベリング結果の具体例

入力	私	奥	寺	と	申	し	ま	す
正解	O	B-PSN	E-PSN	O	O	O	O	O
データ拡張無し	O	O	O	O	O	O	O	O
データ拡張有り	O	B-PSN	E-PSN	O	O	O	O	O

上記の同様な事例として、次の文章内の氏名は訓練データ及び異なる氏名表現に含まれなかった氏名表現であるが、データ拡張有りで学習させたモデルの場合はそれらの氏名を抽出することが確認できた。

- では渡会の方に
- どこの窪塚さん。
- 私森内と申します

さらに、データ拡張無しで学習させたモデルにおいて、誤って氏名として抽出した具体例を表 4 に示す。データ拡張有りで学習させたモデルでは、期待通り”エリア”を氏名として抽出しなかった。

表 4: 誤抽出した具体例

入力	エ	リ	ア	を	特	定	し	て
正解	O	O	O	O	O	O	O	O
データ拡張無し	B-PSN	I-PSN	E-PNS	O	O	O	O	O
データ拡張有り	O	O	O	O	O	O	O	O

このことから、本研究に用いた会話文において、本データ拡張の手法はニューラルネットワークを用いた固有表現抽出器の精度を向上させるのに有効と考えられる。

6 おわりに

近年、ニューラルネットワークを用いた固有表現抽出器の事例が報告されている。通常、ニューラルネットワークを訓練するためには大規模なデータが必要とされる。一方、コンタクトセンター業務及びヘルプデスク業務の会話文において、氏名、住所などの固有表現を含む文章の割合が小さく、また、固有表現も多種多様な表現に渡っている。このような訓練データが少ない問題に対して訓練データの拡張の方法が提案されている。そこで、本研究では、文章内の氏名を表す固有表現部分を異なる氏名表現に置き換える手法でデータ拡張を行い、ニューラルネットワークを用いた固有

表現抽出器の精度向上の検討を行った。さらに、データ拡張の増加率と固有表現抽出器の精度との関係も検討した。

本研究の結果、データ拡張により固有表現抽出器の精度を向上することが確認できた。また、データ拡張を行うことにより、訓練データ及びデータ拡張に含まれない氏名表現についても抽出できることも確認できた。このことから、本研究に用いたデータにおいては、本データ拡張の手法はニューラルネットワークを用いた固有表現抽出器の精度を向上させるのに有効と考えられる。今後は、氏名だけでなく住所も含めたデータ拡張の効果についても検討を行いたい。

7 謝辞

本研究にあたり直接の御指導を戴いた理化学研究所 革新知能統合研究センター関根 聡先生に深謝する。

8 参考文献

参考文献

[1] Nadeau, David and Satoshi Sekine.A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):326. 2007.

[2] Diksha Khurana, Aditya Koli, Kiran Khatter and Sukhdev Singh.Natural Language Processing: State of The Art, Current Trends and Challenges.*Journal Frontiers in Neuroscience*, 2016.

[3] Daniel W. Otter, Julian R. Medina and Jugal K. Kalita.A Survey of the Usages of Deep Learning in Natural Language Processing.*UniversityofColoradoColoradoSprings,USA*,2018.

[4] Jing Li, Aixin Sun, Jianglei Han, Chenliang Li.A Survey on Deep Learning for Named Entity Recognition.*cs.CL*,2018.

[5] 澤山 熱気, 鈴木 潤, 進藤 裕之, 松本 裕治. 非即時的なタスク設定における固有表現抽出の改善. 言語処理学会 第 *24* 回年次大会 発表論文集,921-924,2018.

[6] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura and Tomoko Ohkuma Fuji Xerox Co., Ltd.Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition.*Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 97102, 2017.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer.Neural Architectures for Named Entity Recognition.*Proceedings of NAACL-HLT*, 260-270, 2016.

[8] Ryo Takahashi, Takashi Matsubara, Member, IEEE, and Kuniaki Uehara.Data Augmentation using Random Image Cropping and Patching for Deep CNNs. *Proceedings of The 10th Asian Conference on Machine Learning*, PMLR 95:786-798, 2018.

[9] M. Fadaee，A. Bisazza and C. Monz.Data Augmentation for Low-Resource Neural Machine Translation.Proc. *ACL*，pp.567-573（2017）

[10] M. Johnson，M. Schuster，Q. V. Le，M. Krikun，Y. Wu，Z. Chen，N. Thorat，F. Vigas，M. Wattenberg，G. Corrado，M. Hughes and J. Dean.Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.*Transactions of the Association of Computational Linguistics* Vol.5，Issue 1，pp.339-351（2017）

[11] https://hogehoge.tk/personal/

[12] Erik F. Tjong Kim Sang and Jorn Veenstra.Representing text chunks.*In Proceedings of EACL' 99*, pp. 173179, 1999.