事前学習モデルを用いた音声認識結果からの固有表現抽出

今藤 誠一郎 [1]　上田 直生也 [1]　岡 照晃 [1]
杉山 雅和 [2]　邊土名 朝飛 [2]　小町 守 [1]
[1] 東京都立大学　[2] 株式会社 AI Shift
{kondo-seiichiro, ueda-naoya}@ed.tmu.ac.jp, teruaki-oka@tmu.ac.jp
{sugiyama_masakazu, hentona_asahi}@cyberagent.co.jp,
komachi@tmu.ac.jp

概要

本研究では自動音声認識（Automatic Speech Recognition; ASR）結果からの固有表現抽出（Named Entity Recognition; NER）に取り組む. 認識結果は音声認識誤りや, 略称や別称による未知の固有表現を含む. それらを頑健に汲み取る手段の一つとして, 事前学習モデルを使った文脈の考慮が挙げられる. 大量のテキストで学習された事前学習モデルの中には, 様々な文脈情報が獲得されていると期待できる. 本研究では, 事前学習済み BERT と T5 を用い, 福井県の道路交通情報に関する対話システムの履歴に対し NER を行なった. 辞書マッチに基づく単純な手法より, BERT と T5 は高精度な抽出が可能であることを確認した. T5 は未知の固有表現[1)]についても抽出できる傾向が見られた.

1　はじめに

音声対話システムはユーザーの発話を入力とし, それに自然言語で応答を返す. 本研究ではユーザーの質問や要求などに対して情報を提供するタスク指向対話に取り組む. タスク指向対話の実現には, 地名や人名などの固有表現を正確に認識する必要がある. 現在の音声対話システムはユーザーの発話を自動音声認識（ASR）でテキストに変換し, そのテキストを処理するのが一般的である. 固有表現の認識もこのテキスト処理の中で行われる. 現在我々が行っている固有表現の認識も, ASR テキストに対してタスクごとに作成した辞書に基づいてリンキングを行うことで実現している（図 1 を参照）. 現状はシステム主導の対話が主であり, 発話内容の多くが

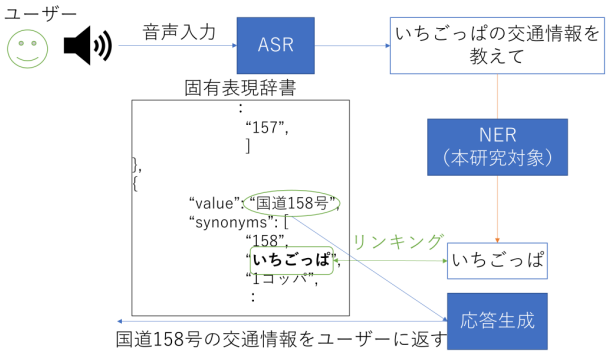

図 1: 想定する音声対話システムのフローチャート

固有表現のみである場合が多いため, ASR テキストをそのままリンキングするのみで事足りる場合が多い. しかし今後の展望としてユーザー主導のシステムを想定する場合, ユーザーは自由な発話をするため, 固有表現の抽出が課題といえる.

テキストにおける固有表現抽出（NER）に関してはこれまでさまざまな研究が行われてきた [1, 2, 3, 4]. 辞書に基づいてマッチングをとる手法やルールベースに基づく手法 [1], 系列ラベリングタスクとして機械学習を用いる手法 [2, 3, 4] などが提案されている. これらの先行研究では, 人手で書かれたテキストを対象としてきた. このようなテキストとは異なり, ASR テキストは音声認識誤りを含むことが想定される. したがって従来の手法, 特に辞書ベースやルールベースのもの以上に困難なタスク設定となっている.

本研究ではテキスト処理部の機能向上を目指し, ASR テキストに対する NER に取り組む. 音声認識結果を処理する場合, End2End で処理した方が良いという先行研究 [5] もあるが, ここでは, モジュールやリソースの柔軟な交換を可能にするため, ASR は既存のものを利用するという制約の下で行う. ASR テキストを扱うにあたって, 抽出の対象となる固有

1)　辞書に対象となる語は含まれているが略語や音声認識誤りが含まれていないという意味の未知の固有表現と, そもそも辞書に対象となる語が含まれていないという意味の未知の固有表現があるが, 本研究ではそれらを区別しない.

表現自体に音声認識誤りが生じることも多いため, 文脈に基づく NER を目指す. 文脈の情報を活用するのに, 大量の文を用いて学習されている事前学習モデルが有効ではないかという考えの下, 大規模事前学習モデルである BERT を用いて NER に取り組む. またノイズの多いテキストには系列ラベリングよりも Seq2Seq として取り組んだ方が良いとの報告 [6] もあるため, エンコーダーデコーダーの事前学習モデルである T5 を用いた実験も行う.

2 関連研究

2.1 事前学習モデルを用いた NER

事前学習モデルを用いた NER はこれまでにも取り組まれている. 現在, 事前学習モデルとして広く利用されている BERT [7] を提案した Devlin らは, CoNLL-2003 の NER タスク [8] において, BERT が fine-tuning を行うことで当時の SoTA 手法と同等の性能を発揮できることを示した. また, Phan ら [9] は T5 [10] に医学文献を用いてドメイン適合を施し, NER を行なった.

2.2 音声認識結果における NER

Wang [11] らは ASR の予測候補上位 N 個の埋め込み表現を合成し, 事前に持っている辞書の各固有表現の埋め込み表現とのドット積を利用することで固有表現を獲得した. 本研究では ASR の予測候補のうち最上位のみを用いる設定で実験を行う. また辞書に基づいて NER を行う際には, 簡単のため埋め込み表現は用いず string match で行なった.

英語を対象にニューラルモデルを用いて音声認識結果から NER に取り組んだ研究がある. Arushi ら [12] は ASR による音声認識誤りを含むテキストから, 音声通話のための人名抽出タスクに取り組んだ. テキスト情報のみではなく音素情報や ASR から得られる候補の上位 N 個, 利用者の所属に関する情報を用いる手法を提案した. その結果, ノイズを多く含むようなテキストからの NER で recall を大きく改善できることを示した. Yadav ら [13] は音声をテキストに変換してから固有表現を抽出するのではなく, RNN モデルで音声スペクトルから直接固有表現を表すラベル付きテキストを End2End で出力する研究を行なった. その結果, より頑健に固有表現を抽出できるようになると主張している. 本研究では, ASR モデルのアーキテクチャーは変更できず, 予測結果の最上位以外は利用できないという設定の下, ASR から得られた日本語文の NER にニューラルモデルを用いて取り組む.

日本語の音声認識結果における NER に関する研究は隠れマルコフモデルや SVM を用いたモデルで見られた. 長谷川ら [14] は音声認識結果から未知の NER に取り組んだ. 隠れマルコフモデルの学習において, 誤り語を特定のシンボルに置き換え, 固有表現クラス付きの bi-gram の学習および誤りを考慮したクラスの導入を提案した. 須藤ら [15] は ASR テキストで SVM を学習する際に, ある単語が正しく認識できているかを表す二値の素性を組み込むことで precision を向上させられることを示した. 一方, 本研究では音声認識誤りを含むテキストから後続タスクに繋げるため, 固有表現を recall 重視で抽出することを目的とし, これに事前学習モデルを利用する.

湯野川ら [16] はコールセンターにおいて ASR で書き起こされた日本語テキストに BERT を用いた系列ラベリングにより, 顧客入電意図の抽出を行なった. 本研究でも固有表現の抽出を目的に BERT で系列ラベリングを行うとともに, T5 を用いた実験も行う.

3 手法

本研究では音声認識誤りを含む ASR テキストからの NER 評価用のデータとして道路交通情報に関するデータを用いる. それに伴い, 道路と住所に関する固有表現の抽出を行う. 前提条件として出力したいラベル（道路と住所）は指定できるものとする.

BERT を用いた抽出　系列ラベリングタスクとして NER に取り組む. サブワード単位に分割されたテキストに, 固有表現の先頭に B, 内部・終端に I, 固有表現以外に O のラベルを付与する BIO モデルでラベリングを行う. 概略図を図 2 に示す. 道路情報に関するラベルは “`{B, I}-route`”, 住所情報に関するラベルは “`{B, I}-address`” として扱う. ラベルを指定するにあたり, 文頭に “route” または “address” トークンを付与し, ラベルとして “B-label” を与える.

T5 を用いた抽出　質問応答タスクとして NER に取り組む. 文頭に特殊タグを付与したテキストを入力文として与え, 特殊タグに対応する固有表現を出力する. このとき, 抽出された各固有表現の末尾にはラベルがハイフンで連結され, 付与されるような設定となっている. 概略図を図 3 に示す. 道路情報に関するラベルは “道路名”, 住所情報に関するラベルは

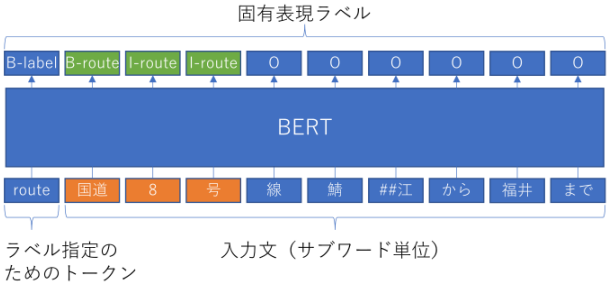

図 2: BERT を用いた NER

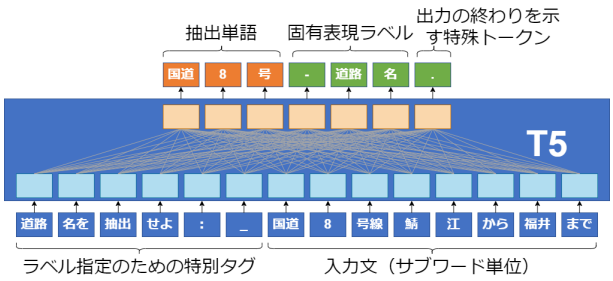

図 3: T5 を用いた NER

“住所名” として扱う. ラベルを指定するにあたり, 文頭に “道路名を抽出せよ:” または “住所名を抽出せよ:” という特殊タグを付与する.

4 実験

4.1 データ

本研究では福井県の道路交通情報に関するシステム主導型の対話ログを用い NER を行なった. 対話ログは発話者が道路名や住所名を発話していると期待されるターンのものを用いた. 抽出したい固有表現の辞書が与えられており, この辞書には固有表現とそれに対する一部の別称や略称, 音声認識誤りが登録されている（図 1 中の固有表現辞書）. 運用上, この辞書を用いて辞書マッチで NER が成功したデータ（match）と, 失敗したデータ（fallback）（表 1）の両方を NER の対象テキストとして扱った. match のデータは辞書マッチでラベル付けを行い[2], fallback のデータに関しては, 人手アノテーションで住所と道路に関する固有表現にラベル付けを行なった[3]. match を, 低コストで入手できて基本的には辞書マッチで抽出できるデータ, fallback を入手にコストがかかり基本的には辞書マッチで抽出できないデータとして扱う（4.4）. fallback にはアノテーションの際に福井県に実在するか否かの基準でアノテーションを行なったため, match とラベルがついている単語の基準に相違がある.

表 1: match と fallback のデータ例
（下線部が固有名詞）

	text
match	鯖江から敦賀市へ向かう高速道路
fallback	えーとサザエさん、サザエ市春江町車が中の運転中ですること

出力するラベルを指定するという設定に合わせて, 各文の先頭に提案手法ごとの前処理を施す. 利用したデータ数について Appendix の表 3 に示す.

4.2 実験設定

辞書に基づいた string match, 事前学習モデルである BERT, T5 による 3 通りの NER について比較を行う. BERT モデルは東北大の公開しているモデル[4]を利用し, token-classification 形式の fine-tuning を行なった. パラメーターはバッチサイズを 8, エポック数を 3, 最大系列長を 258 と設定した. また, T5 モデルは Megagon Labs の公開しているモデル[5]に fine-tuning を行なった. パラメーターは学習率を 0.0005, バッチサイズを 8, エポック数を 20, 最大系列長を 128 と設定した. いずれも fine-tuning には, Huggingface の公開している Transformers[6]のスクリプトを用いた.

4.3 評価方法

完全一致を真陽性とみなした precision, recall, F1 スコアを算出して評価を行う. また match と fallback のラベル付け基準には相違がある（4.1）ため, それぞれのテストセットを別々に評価する. match の評価は辞書マッチで抽出できる固有表現を抽出できるのか, fallback の評価は辞書に含まれない固有表現を抽出できるのかを示す指標とする.

また本研究における NER は下流タスクで entity linking を行うことを想定している. このとき, 抽出した固有表現が本来よりも短い場合, 下流タスクに

2） 明らかに意味的に誤っているラベルに関しては人手でラベルを取り除いた.
3） このアノテーションでは音声認識誤りも考慮して行なった. また, 辞書にエントリーされていない固有表現に対しても福井県に実在する場合には抽出対象として扱った. 表 1 に示すように抽出すべき固有表現がそもそもテキスト中に存在していない場合と, 音声認識誤りや辞書のエントリー不足のために抽出できなかった場合が混在している.
4） https://huggingface.co/cl-tohoku/bert-base-japanese-v2
5） https://huggingface.co/megagonlabs/t5-base-japanese-web
6） https://github.com/huggingface/transformers

表 2: string match, BERT, T5 による実験結果
(“c_”: 固有表現をリファレンスよりも長く予測した場合に真陽性としてスコア計算し直したときの結果)

手法	data	P	R	F1	c_P	c_R	c_F1	P	R	F1	c_P	c_R	c_F1
辞書マッチ	match	96.3	**100**	**98.1**	96.3	**100**	98.1	-	-	-	-	-	-
	fallback	50.0	23.3	31.7	50.0	23.3	31.7	-	-	-	-	-	-
		全データで学習						match のみで学習					
BERT	match	97.3	97.3	97.3	**99.2**	99.2	**99.2**	**97.3**	97.3	97.3	**98.8**	**98.8**	**98.8**
	fallback	67.9	83.7	75.0	67.9	83.7	75.0	**58.8**	46.5	**51.9**	**58.8**	46.5	**51.9**
T5	match	**98.0**	97.7	97.9	**99.2**	98.8	99.0	**97.3**	**97.7**	**97.5**	98.5	**98.8**	98.6
	fallback	**74.0**	**86.0**	**79.6**	**74.0**	**86.0**	**79.6**	41.3	**60.5**	49.1	42.3	**62.8**	50.9

おいて linking が不可能になる可能性が考えられる. 一方で, 本来よりも長く抽出した場合の entity linking における問題は軽微であると考える. 従って, 固有表現をカバーできている場合は真陽性とみなすが, カバーできていない部分一致は偽陽性とみなした評価も行う. 例えばリファレンスの “8 号線” という固有表現にラベルが付与されていた場合, “国道 8 号線” の抽出は認めるが, “8 号” のみの抽出は認めない.

4.4 実験結果と考察

学習データに match と fallback を両方用いた場合と match のみを用いたときの, match と fallback それぞれのテストセットにおける実験結果を表 2 に示す. BERT と T5 それぞれの NER 結果の実例を Appendix に掲載する.

辞書マッチ match のデータに関してはデータ自体が辞書マッチで作成されているため recall が 100 となる. precision が 100 にならないのは match のデータを作成する際に明らかに間違ってラベルが付けられていたものを人手で取り除いたため（4.1）である. 一方で fallback のデータに関してはどの評価指標においても 50.0 以下であり, 固有表現を十分に抽出できていないといえる. 本研究ではこのようなデータに事前学習モデルを用いた NER がどの程度機能するのかを調査した.

BERT match のテストデータに関しては F1 スコアで 0.8 ポイント下回るものの, 辞書マッチで抽出できる固有表現は BERT でも同等に抽出できているといえる. c_F1 では 99.2 ポイントの精度であり, 後続タスクのための NER として望ましい結果である. fallback のデータに関しては, 辞書マッチと比較して match のみで学習した場合+20.2 ポイント, fallback も含めて学習した場合+43.3 ポイントであることが確認できる. 特に recall のスコアの向上が顕著に見られることから, BERT を用いることで辞書マッチでは抽出できなかった固有表現を抽出できていることがわかる. 学習データに match のデータの半分程度の fallback のデータを加えることで, より大きなスコアの向上が見られるが, これは match のデータに含まれない未知語に対しても学習が行われたためだと考えられる.

T5 T5 に関しても BERT 同様の傾向がみられる. match のテストデータに関しては,BERT と同程度の性能を出せており, 辞書マッチと同等に抽出できている. fallback のデータに関しては, match のみで学習を行うと precision が低くなった. しかし学習データに fallback のデータを加えることで precision は改善され, BERT と比較しても F1 スコアで+4.6 ポイントであり, より意図通りに抽出することができた.

5 おわりに

本研究では事前学習モデルである BERT と T5 を用いて, 音声認識結果に対する NER に取り組んだ. 実験の結果, 辞書マッチで作成したデータに関しては事前学習モデルを用いても概ね良好に抽出できた. また学習データに人手アノテーションしたデータを追加することで, 辞書に含まれていない固有表現に関しても抽出できることを確認した. 今後はより文脈を利用することで, 未知の固有表現を含むノイジーなテキストから頑健に固有表現を抽出するための fine-tuning の方法を含めた工夫、例えば固有表現をマスクしたデータを学習データに追加するというような検討を行なっていく.

参考文献

[1] 竹元義美. 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出. 情報処理学会論文誌, Vol. 42, No. 6, pp. 1580–1591, 2001.

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In **Proceedings of the Eighteenth International Conference on Machine Learning**, p. 282–289, 2001.

[3] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均. 最大エントロピーモデルと書き換え規則に基づく固有表現抽出. 自然言語処理, Vol. 7, No. 2, pp. 63–90, 2000.

[4] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[5] Motoi Omachi, Yuya Fujita, Shinji Watanabe, and Matthew Wiesner. End-to-end ASR to jointly predict transcriptions and linguistic annotations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1861–1871, 2021.

[6] Stefan Constantin, Jan Niehues, and Alex Waibel. Incremental processing of noisy user utterances in the spoken language understanding task. In **Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)**, pp. 265–274, 2019.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[8] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In **Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003**, pp. 142–147, 2003.

[9] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. SciFive: a text-to-text transformer model for biomedical literature. **CoRR**, Vol. abs/2106.03598, , 2021.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[11] Haoyu Wang, John Chen, Majid Laali, Kevin Durda, Jeff King, William Campbell, and Yang Liu. Leveraging ASR N-Best in Deep Entity Retrieval. In **Proceedings of the Interspeech 2021**, pp. 261–265, 2021.

[12] Arushi Raghuvanshi, Vijay Ramakrishnan, Varsha Embar, Lucien Carroll, and Karthik Raghunathan. Entity resolution for noisy ASR transcripts. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations**, pp. 61–66, 2019.

[13] Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. End-to-end named entity recognition from English speech. In **Proceedings of the Interspeech 2020**, pp. 4268–4272, 2020.

[14] 長谷川隆明, 林良彦. 隠れマルコフモデルに基づく音声認識結果からの固有表現抽出. 言語処理学会第 9 回年次大会, pp. 533–536, 2003.

[15] Katsuhito Sudoh, Hajime Tsukada, and Hideki Isozaki. Incorporating speech recognition confidence into discriminative named entity recognition of speech data. In **Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics**, 2006.

[16] 湯野川八雲, 久保優騎. BERT によるコールセンタログからの顧客入電意図抽出. 第 92 巻, p. 04, 2021.

A データセット

表 3: 実験に利用したデータ情報
（発話数と各ラベルが付与された単語数）

		train	dev	test
match	発話数	1,757	220	220
	address	1,220	144	147
	route	802	104	110
fallback	発話数	949	118	122
	address	197	30	26
	route	92	8	17

B 出力例

表 4: match のデータにおける NER 失敗例

	text	hyp	ref
BERT（住所）	吉田郡永平寺町	吉田郡, 永平寺町	永平寺町
T5（住所）	田尻町から福井市までの福井市内まで	田尻町, 福井市, 福井市	福井市, 福井市
BERT（道路）	イチゴったー	—	イチゴったー
T5（道路）	イチゴったー	—	イチゴったー

表 5: fallback のデータにおける NER 失敗例

	text	hyp	ref
BERT（道路）	青年の道	—	青年の道
T5（住所）	低い	低い	—
BERT（住所）	あの高みの方のエルパ行きのバスは取った後	高みの	—
T5（住所）	あの高みの方のエルパ行きのバスは取った後	高みの方の	—

T5 で抽出に成功して BERT で失敗した例, BERT で抽出に成功して T5 で失敗した例, 両方で失敗した例を, match と fallback それぞれのデータセットについて表 4 および表 5 に示す.

match のテストデータは辞書マッチに基づいて作成されているため, リファレンスには吉田郡や田尻町が含まれていないが, これらは実際には福井県に存在する地名であり, 実際には抽出した方が良い例となっている. “イチゴったー” は “158 号線” のことを “いちごっぱ” と発話することがあり, それを認識誤りしたものと考えられる. このような例は BERT も T5 も抽出が難しいことが確認できる.

fallback に示す “青年の道” は学習データには含まれていないが, 実際に福井県に存在する道路名である. T5 が抽出できたのは, 末尾についている “道” という単語から道路名であると予測したためと考えられる. T5 のみがこのような文脈から予測をできた理由として, モデル構造の違いと事前学習時のデータ量の違いが考えられるが, この要因の特定は今後の研究課題である. BERT と T5 の両方が “高みの（方の）” を住所を表す固有表現として抽出しているが, 方向を表すような表現から文脈的にこれらのモデルが固有表現を抽出しようとしたことが推測される.