ニュース記事中の企業名の Entity Linking における Question Answering を用いた曖昧性解消

齋藤 慎一朗 髙橋 寛治
Sansan 株式会社
{shinichiro.saito,ka.takahashi}@sansan.com

概要

ニュース記事と、記事に登場する企業名の法人番号を紐付けたい。しかし、同名企業が存在する場合、一意に特定できないことがある。既存手法では、ニュース記事と、企業情報のベクトルの類似度が最も大きい法人番号を出力しているが、企業データベースが持つ他のテキスト情報を有効活用できる手法ではない。そこで、GPT による Question Answering を利用し、他のテキスト情報を適切に活用可能な Entity Linking を提案する。提案手法により、既存手法に対し 65%ポイントの性能改善を確認した。さらに、Web 検索結果のテキストを追加で与えることで、5%ポイントの性能改善を確認した。

1 はじめに

企業に関する正確な情報を知るために、ニュース記事に対して、ニュース記事中に登場する企業名と、その法人番号を紐付けて管理することが重要である。しかし、ニュース記事に登場する企業名が、どの企業の法人番号に紐づくか曖昧性が存在する。この曖昧性の中には、企業名が同じだが、法人番号が異なる企業（同名企業）が存在する場合がある。本研究では、同名企業の曖昧性を解消するタスク(同名企業特定タスク) に着目する。

Entity Linking(EL) は、テキストから Entity を認識し、知識ベースに登録されている Entity に紐付ける技術である。知識ベースとは、特定のデータを体系的に整理したデータベース (DB) のことである。同名企業特定タスクは、テキスト中の Entity として企業名が与えられたときに、知識ベースである企業 DB 中の Entity と紐付けるタスクであると定義する。

同名企業特定タスクにおいて、同名企業でも住所・代表者名・取り組む事業などが異なる点を考慮することで、より正確な特定が可能となる。しかし、ニュース記事には企業情報が必ずしも全て記載されるわけではないため、候補企業となる企業情報のうち、記事に記載された企業情報のみを適切に活用することが重要である。

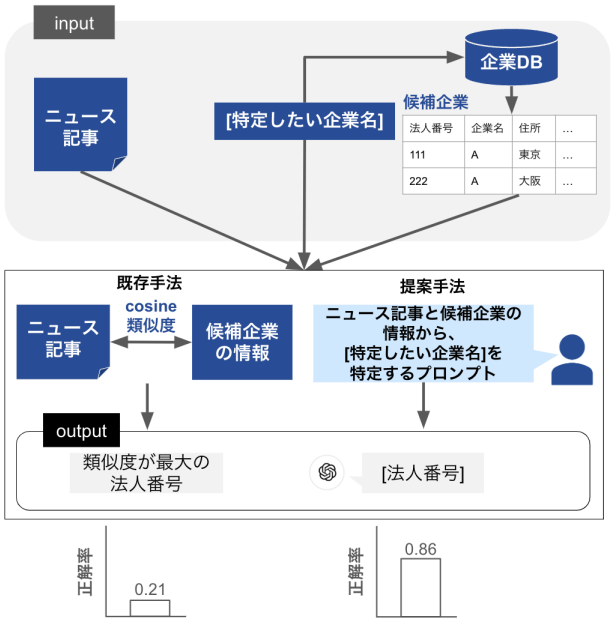

図 1: 提案手法の概要図

髙橋ら [1] は、Fine-Tuning(FT) を行わず、企業の取り組む事業のテキストと、記事テキストの類似度が最も大きい候補を出力する手法を提案している。本手法は、候補となる企業情報とニュース記事を別々にベクトル化するため、候補となる企業の情報のうち、ニュース記事に登場する企業情報のみを取捨選択できる方法ではない。

澤田ら [2] は、記事テキストと特定された企業名を紐付けたデータを用意し、FT することで、類似度が最も高い候補を出力している。この手法は、用意されたデータに含まれない企業を正しく推論することができない。一方、企業の情報は日々変化するため、変化に追従できるよう頻繁にアノテーションされたデータを用意し、モデルを更新するのは時間

的コストがかかる。

また、Ding らは ChatEL[3] において、EL に対して、Question Answering(QA) を利用する手法の性能改善方法を提案している。ChatEL では、候補に関する情報を与え、複数候補からの選択を行うことで EL タスクを解いている。本手法は、EL 対象となるテキストと候補に関する情報をまとめて扱うことで、特定したい候補と関係のある情報のみを考慮することが期待でき、FT を行わないため、データ準備とモデル更新のための時間的コストがかからない。

そこで、本研究では、同名企業特定タスクにおいて、EL に対して GPT を用いた QA を行うことで、FT が必要なく、ニュース記事と企業情報をまとめて扱う手法を提案し、その有効性を示した。本論文における提案手法を図 1 に示す。

2 関連研究

2.1 QA による EL

ChatEL では、以下の 3 段階のフレームワークを利用することで、EL における QA の性能を改善した。

1. Prior[4] と BLINK[5] を利用し、候補 Entity を生成する。
2. LLM を用いて、Entity の情報を要約する。
3. 候補 Entity とその情報を複数選択質問として与え、候補から 1 つを出力する。

本論文では、ニュース記事と企業情報を 1 つのプロンプトとして扱うことで、特定したい企業候補と関係のある企業情報のみを考慮することを期待し、手順 3 と同様の手法を行う。具体的には候補 Entity とその情報を GPT に対して複数選択質問として与え、正解となる Entity を 1 つ出力する。

2.2 ニュース記事における同名企業特定タスク

髙橋ら [1] は、記事内容と、企業の取り組む事業が似ていると仮定し、それらの文書ベクトルの類似度が最も高い企業を出力している。また、澤田ら [2] は、FT を行ったモデルを利用し、記事中における企業名のスパンベクトルと、企業 DB における企業名のベクトルの類似度を計算し、最も類似度が高い企業を出力している。

本論文では、既存の手法に対し、FT を必要とせず、またベクトルによる類似度の計算を行わない。QA を用いてニュース記事と企業情報をまとめて扱う手法を提案する。

3 提案手法

ニュース記事の記事タイトルと記事本文、特定したい企業名、企業情報を含めた企業の候補を GPT に与え、どの法人番号が最も近いかを生成する。図 2 のプロンプトを利用する。本プロンプトは、ChatEL のプロンプトを、同名企業特定タスク向けに改変したものである。

```
以下にニュース記事と、特定したい企業名を与えます。さらに、企業の候補を複数与えます。ニュース記事と、企業の候補に与えた情報を利用し、候補の中から 1 つ企業を特定し、その法人番号を教えてください。

回答の際には、法人番号のみを出力してください。

# ニュース記事
記事タイトル: {ARTICLE_TITLE}
記事本文: {ARTICLE_BODY}

# 特定したい企業名
{COMPANY_NAME}

# 企業の候補
{CANDIDATE_COMPANIES}
```

図 2: 提案手法のプロンプト

4 実験

4.1 実験設定

4.1.1 GPT

GPT として、OpenAI の gpt-4o-2024-08-06[6] を利用した。本バージョンの学習データは 2023 年 10 月までである。

4.1.2 利用データ

企業 DB として、企業情報データベース Musubu[7] が保有するデータのうち、法人番号、住所、代表者名、見出し（企業の活動を一言で表現したもの）を利用した。候補企業として、ニュース記事中の特定したい企業名と同一企業名の企業情報を、企業 DB から取得した。

評価データとして、同名企業が存在する企業名が記事中に登場するニュース記事を 100 件用意し、gBizINFO[8] を用いて法人番号を付与した。1 記事

に対し、同名企業の候補数の中央値は 8 である。評価データの統計量を、付録の表 3 に載せた。また、ニュース記事の作成時期は、2023 年 11 月から 2024 年 11 月までとした。これは、GPT の学習データより後に作成された記事であり、学習データに評価データの記事が含まれていないことを意味する。

4.1.3 評価指標

評価指標として、正解率を用いた。正解率として、評価データ全件の法人番号に対し、何件の法人番号を完全一致で出力できたかの割合を計算した。

4.1.4 比較手法

比較手法として、ランダム、SCDV(Sparse Composite Document Vectors)[9]、openai-text-embedding-3-large[10]、ルールベースを用いた。

ランダムについて、候補となる同名企業からランダムに 1 つ選んだ。

SCDV は髙橋ら [1] において用いられた手法である。文中に含まれる単語の単語ベクトルから潜在的なトピックを考慮した文書のベクトルを取得する。SCDV により得た記事テキストのベクトルと、企業情報のベクトルの類似度が最大の法人番号を推論結果とする。単語ベクトルとして、Wikipedia Entity Vectors[11] の 200 次元のベクトルを利用する。

SCDV は、2017 年の提案手法であるため、最新のテキスト埋め込みモデルとして、日本語テキスト埋め込みベンチマークである JMTEB[12] での平均性能が最も高い openai-text-embedding-3-large(OpenAIEmb と呼ぶ) を利用し、SCDV と同様の検証を行った。OpenAIEmb の学習データは 2021 年 9 月までのため、学習データに評価データの記事が含まれていない。

ルールベースでは、ニュース記事中に企業情報が登場したら加算を行い、最も点数が高い法人番号を推論結果とした。詳細は、付録のセクション A に載せた。

4.1.5 Web 検索

提案手法では、テキストでの情報の追加が容易に可能であると考えられる。その検証のための追加実験として、Web からの検索結果を利用し性能が改善可能か確認した。まず、ニュース記事と企業 DB による情報だけでは判定できない場合に、GPT に「不明」と出力させた。「不明」と出力されたデータについて、Web 検索を行うためのキーワードを gpt-4o-2024-08-06 を用いて抽出した。次に、「企業名 キーワード」のクエリを用いて、Bing Web Search API[13] を利用し、検索結果の上位 10 件それぞれに対し、100 文字程度からなるスニペットを取得した。最後に、スニペットを改行区切りで結合し、プロンプトに加えることで、法人番号の推定を行った。Web 検索を含めた提案手法の概要図を、付録の図 4 に載せる。

4.2 結果

4.2.1 手法ごとの比較

表 1 に、手法ごとの比較結果をまとめる。

表 1: 手法ごとの正解率

手法	利用するデータ	正解率
ランダム	-	0.13
SCDV	見出し + 住所 + 代表者名 + 企業名	0.21
OpenAIEmb	見出し + 住所 + 代表者名 + 企業名	0.58
ルールベース	企業名 + 住所 + 代表者名	0.64
提案手法	見出し + 住所 + 代表者名 + 企業名	0.86

提案手法は、他手法よりも正解率が高い。提案手法では、特定したい企業に対する情報のうち、ニュース記事に登場する情報のみを取捨選択し、性能が改善していると考えられる。

また、OpenAIEmb での性能が、ルールベースよりも低い。このことより、同名企業特定タスクにおいて、ベクトルによる類似度を用いた手法に限界があることが分かる。

4.2.2 データを加えた場合の比較

表 2 に、手法ごとにデータを加えた場合の比較結果をまとめる。

表 2: データを追加した場合の正解率

利用するデータ	正解率		
	提案手法	OpenAIEmb	SCDV
なし	0.28	-	-
見出し	0.75	0.60	0.27
見出し + 住所	0.84	0.77	0.28
見出し + 住所 + 代表者名	0.86	0.75	0.35
見出し + 住所 + 代表者名 + 企業名	0.86	0.58	0.21
見出し + 住所 + 代表者名 + 企業名 + Web	0.91	-	-

SCDV について、見出しに加え、住所、代表者名、企業名を追加しても正解率は大きく改善しない。これは、SCDV で利用した語彙が、住所、代表者名などの固有の語彙に対応していないことが原因と考え

られる。

OpenAIEmb について、見出しに加え、住所の情報を加えた際は正解率が改善するが、さらに代表者名や企業名を追加すると、正解率が悪化することが分かる。これは、ニュース記事に登場しないが、データとして与えた企業情報が余計なベクトルとして加味されてしまい、類似度による推論の性能が低下してしまうことが原因と考えられる。

提案手法では、情報の追加に従い、性能が改善していくことが分かる。よって、提案手法は、ニュース記事と候補の企業情報をまとめて与えることで、テキストから適切な情報を取捨選択し、正しい法人番号を判断できていると考えられる。

さらに、Web 検索の結果を利用した場合も性能の改善が見られた。よって、曖昧性解消に寄与するテキスト情報をプロンプトに渡すだけで、候補から絞りきれなかった企業を限定する情報が増え、正解率を改善できると考えられる。

4.2.3 改善した例とエラー分析

SCDV、OpenAIEmb に対し、提案手法にて改善した例を図 3 の a に示す。本図に登場するテキストは、実際のテキストをもとに、改善・エラーの傾向を示すために用意したダミーのテキストである。

SCDV、OpenAIEmb は住所を正しく考慮できておらず、大阪府の株式会社ダミーを誤って推論したが、提案手法は東京都の株式会社ダミーを正しく推論できている。

次に、提案手法にて、Web 情報を用いて改善した例を図 3 の b に示す。候補のうち、どの株式会社サンプルに関する記事なのかを特定するための情報がない。一方、Web 検索結果より、地域支援サービスを開始したのは、コンサルティングを行う株式会社サンプルであることが分かり、Web 検索結果を用いて正しい推論ができたと考える。

最後に、Web 情報を用いた提案手法において正解できていないデータについて調査する。まず、不正解のデータ数は 100 件中 9 件である。うち、候補の中に正解が存在しないデータが 4 件存在した。これは、企業 DB の拡充が解決策となりうる。また、記事だけでは特定する情報が不足しているが、不明と出力できず Web 検索ができていないデータが 5 件存在した。このことより、GPT が法人番号を特定可能かどうかを判定する方法を改善する必要があると考えられる。

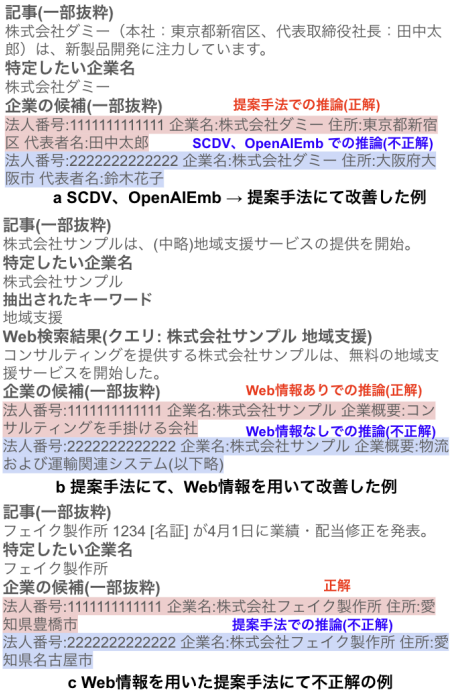

図 3: 改善した例、不正解の例

推論結果が不正解だが、不明と出力できず Web 検索ができていないデータの例を図 3 の c に示す。正解は愛知県豊橋市の株式会社フェイク製作所であるが、異なるフェイク製作所を推論してしまっている。なぜ不明と出力できなかったかについて、名証が名古屋証券取引所を意味し、推論結果の住所は名古屋市であるため、特定できたと誤った判断をしてしまったと考えられる。このデータについて、1234 のような証券コードに紐づいた Web 検索結果を用いることができれば正解が導けると考えられる。

5 おわりに

本研究では、ニュース記事における同名企業特定タスクに対し、FT 用のデータを用意する必要なく、特定したい候補と関係のある情報のみを取捨選択する手法を提案し、その有効性を示した。さらに曖昧性解消に寄与するテキスト情報をプロンプトに渡すだけで、性能改善が可能なことを示した。

今後の課題として、GPT が法人番号を特定可能かどうかの判定を改善する必要があると考えられる。また、本研究ではニュース記事の中から GPT を用いてキーワードを 1 つ抽出し、Web 検索を行う方法に限定して検証を行ったが、どのようなキーワードが良いかに関しての検証は行えていない。そのため、キーワード抽出の改善が、同名企業特定タスクにおける更なる性能改善に繋がると考える。

参考文献

[1] 高橋寛治, 甫立健悟, 奥田裕樹. ニュース記事中の組織名の曖昧性解消. 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 4Rin103–4Rin103, 2020.

[2] 澤田悠治, 安井雄一郎, 大内啓樹, 渡辺太郎, 石井昌之, 石原祥太郎, 山田剛, 進藤裕之. 企業名の類似度に基づく日経企業 id リンキングシステムの構築と分析. 自然言語処理, Vol. 31, No. 3, pp. 1330–1355, 2024.

[3] Yifan Ding, Qingkai Zeng, and Tim Weninger. ChatEL: Entity linking with chatbots. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 3086–3097, Torino, Italia, May 2024. ELRA and ICCL.

[4] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2619–2629, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6397–6407, Online, November 2020. Association for Computational Linguistics.

[6] OpenAI and et al. Gpt-4 technical report, 2024.

[7] Baseconnect 株式会社. Musubu. https://musubu.in/.

[8] 経済産業省. gbizinfo. https://info.gbiz.go.jp/index.html.

[9] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. Scdv : Sparse composite document vectors using soft clustering over distributional representations, 2017.

[10] OpenAI. https://platform.openai.com/docs/guides/embeddings/.

[11] 鈴木正敏. https://github.com/singletongue/WikiEntVec/releases/tag/20190520.

[12] SBIntuitions 株式会社. https://www.sbintuitions.co.jp/blog/entry/2024/05/16/130848.

[13] Microsoft Corporation. https://www.microsoft.com/en-us/bing/apis/bing-web-search-api.

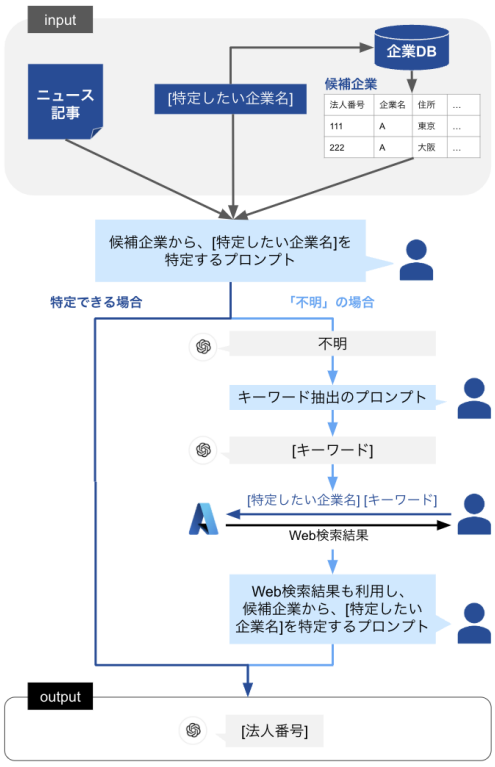

図 4: Web 検索を含めた提案手法の概要図

表 3: 評価データの統計量

大項目	小項目	値
特定したい企業名のユニーク数		40
正解となる法人番号のユニーク数		42
候補企業のデータ件数	最大値	102
	最小値	2
	平均値	13.07
	中央値	8

A　ルールベースの手法

企業 DB から、特定したい企業名に該当する候補となる企業群を取得する。その後、ニュース記事中に企業情報の文字列が含まれていたら、該当する企業に点数を加算し、最も点数が高い法人番号を出力する。点数として、まずは住所の市区町村が存在するか判定し、存在した場合は 1 を与える。市区町村が存在しない場合は、都道府県が含まれるかを判定し、存在する場合は 0.5 を与える。次に、代表者名が存在するか判定し、存在した場合は 1 を与える。最後に、最も点数が高い法人番号を出力する。最も点数が高い企業が複数ある場合は、最も点数が高い企業群からランダムに選択する。

B　提案手法におけるデータなしの正解率について

表 1 におけるランダムの性能が 0.13 であるのに対し、表 2 における利用するデータが「なし」の場合の性能は 0.28 であり、15%ポイント高いことが分かる。このことより、法人番号の数字から企業の情報をある程度特定して正解できている可能性が考えられる。[1)] また、GPT 自体が法人番号を丸暗記している可能性も考慮し、同じ評価データ 100 件に対し、企業名に対する法人番号を生成可能かの実験を追加で行なったが、1 件も正しく生成できなかった。よって、法人番号自体を記憶しているわけではないことが分かる。

1)　例えば、13 桁のうち、先頭から 6、7 桁目が 01 の場合は株式会社、02 の場合は有限会社、03 の場合は合同会社を指す。