検索クエリにおける共起情報を活用した非曖昧ドメイン固有語辞書の構築: ランドマークの事例

西川荘介 山城颯太 浅野広樹 佐野峻平 颯々野学
ヤフー株式会社
{sonishik,soyamash,hiroasan,shsano,msassano}@yahoo-corp.jp

概要

特定ドメインにおける曖昧性のない固有名詞群（非曖昧ドメイン固有語）は固有表現抽出を中心に有用である. 本稿では非曖昧ドメイン固有語辞書構築の一例として, 地図上の一点を表す拠点名称以外の意味ではほぼ出現しない語 (非曖昧ランドマーク語) の辞書の自動構築に取り組む. 提案手法ではこの問題を, エンティティ名が非曖昧ランドマーク語であるか否かの二値分類タスクとして扱い, 検索クエリにおける共起語情報を考慮したモデルを提案する. 実験では提案モデルで 0.907 の F1 値を達成し, ルールベースやその他の機械学習モデルよりも高い性能を示した. さらに, 提案モデルを用いた辞書自動構築により, 固有表現抽出システムの改善を確認した.

1 はじめに

非曖昧ドメイン固有語とは特定のドメインを特徴づけるドメイン固有語 [1, 2, 3] のうち, そのドメイン内のエンティティを指し示す以外の意図で使われることがほぼない語を指す. 非曖昧ドメイン固有語辞書は固有表現抽出器構築において, あるドメインで適合率重視の高速な抽出器を手早く実現したい場合 [4, 5, 6] や, ツイートや検索クエリなどの対象語の文脈が乏しく曖昧性の解消が難しい場合 [7, 8] などに有用である. [1)] しかし, 非曖昧ドメイン固有語は潜在的に大量に存在し, 常に新語が登場するため, 人手により辞書を拡張し続けることは困難である. 一方で非曖昧ドメイン固有語を含むエンティティリストは知識ベースを始めとする言語資源から容易に手に入る. そこで本稿ではこれらの資源を活用し, 拠点名称ドメインを対象とした非曖昧ドメイン固有語（非曖昧ランドマーク語）辞書の自動拡張に取り組む.

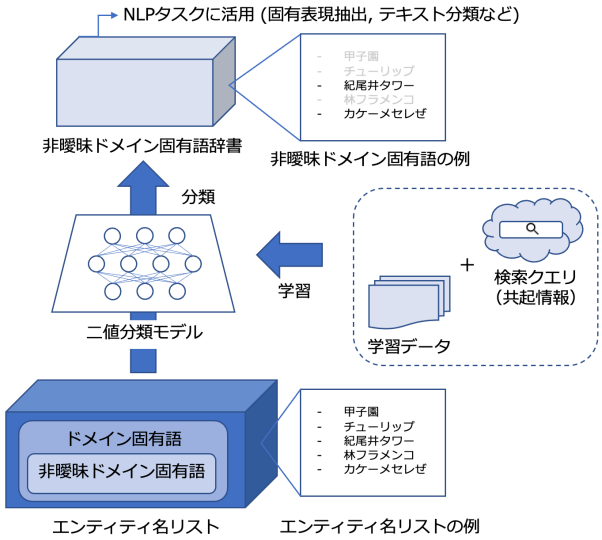

図 1 拠点名称ドメインでの非曖昧ドメイン固有語構築

本研究ではこの自動拡張を行うため, 知識ベースなどから取得されたエンティティ名が非曖昧ランドマーク語か否かの二値分類タスクを考える（図 1）. この分類設定では例えば実在する施設名だが一般語でも利用される語（“だるま”,“チューリップ”）や拠点名称かつ固有名詞だがランドマーク意図以外でも利用される語（“甲子園”）などが混在する中で, 真の非曖昧ランドマーク語（“紀尾井タワー”）を判定する必要がある. そのため, 既存の一般名詞か固有名詞かを区別する設定 [9, 10, 11] やドメイン固有語を分類する設定 [1, 2] とは分類対象が異なり, 非常に曖昧性が高い. また, 非曖昧ランドマーク語に関する知識は一部の著名なランドマークを除いて利用可能でない場合が多く [12], 分類の手がかりとして使用するのは難しい. そこで本研究では非曖昧ランドマーク語が拠点名称ドメインに関連する語と検索クエリ上で共起しやすい点に着目し, 検索クエリの共起語特徴量を考慮した二値分類モデルの提案を行う. こ

1) 例えばヤフーではウェブ検索サービスにおいてランドマーク語を抽出する際,“紀尾井タワー” などの曖昧性なく拠点名称を示す語は抽出するが, 時に店名などの拠点名称を指す “だるま” などの一般名詞は抽出したくない場合に利用できる.

のモデルを用いて, エンティティ名リストから非曖昧ランドマーク語を自動的に抽出し, 辞書拡張を行うことを想定する. なお, 本手法は学習データとエンティティ名リストを目的ドメインのものに差し替えることで, 拠点名称以外のドメインにも適用可能だと考えられる. 関連研究については付録 A で述べる.

実験では人手で構築したデータセットを用いて複数のベースライン手法との比較を行い, 提案手法の有効性を確認する. さらに提案した二値分類器により実際に非曖昧ランドマーク語辞書の拡張を行い, 固有表現抽出システムへの効果を調査する.

2 提案手法

本章ではエンティティ名が非曖昧ランドマーク語か否かを判定する二値分類器の構築法を記す.

2.1 データセット構築

本節では, 非曖昧ランドマーク語か否かの二値分類データセットを構築した手順について説明する. 我々はヤフーの固有表現抽出システムで補助的に用いられてる既存の非曖昧ランドマーク辞書や場所名リストから正例・負例のデータセットを構築した. 具体的には, 非曖昧ランドマーク辞書の語は正例データとし, 場所名リストの語は人目で非曖昧ランドマーク語ではないと判定された語を負例データとした. この負例データには県名・市名などの一般的な地名や, 地図上で一点に定まらないチェーン店名などが含まれる. また,“甲子園”, “だるま” などの必ずしもランドマーク意図があるとは限らない曖昧性の高い語を負例に含めるために, ウェブ検索クエリコーパスに対して形態素解析で前処理後, ある頻度以上出現する語を負例データに追加した.

2.2 検索クエリの共起語情報の導入

本節ではより正確な分類を目指し, 検索クエリを活用した手法について説明する. 分類対象のエンティティ名にランドマーク語の手がかりとなるような部分語（“タワー”,“郵便局”）が含まれない場合, 対象語の表記情報のみから非曖昧ランドマーク語の判定を行うことは困難である. また, 多くの非曖昧ランドマーク語は Wikipedia のような自然文コーパスを始めとする大規模コーパス中には出現しないため [12], これらの文書の周辺文脈から曖昧性判定を行うのは難しい. [2] そこで本研究では検索クエリの活用に着目した. 例えば曖昧性の高いランドマーク語である “甲子園” は検索クエリ上で “優勝”,“結果”,“球場” などのその曖昧性を解消しうる複数ドメインの語と共起する. 一方, 非曖昧ランドマーク語は検索クエリ上で拠点名称ドメインの語とよく共起する傾向があり, 例えば “紀尾井タワー” は “アクセス”,“行き方”,“レストラン” とよく共起する. そこで提案手法では分類モデルとして BERT[14] を用い, (1) 事前学習時, (2) fine-tuning 時の二つの設定で検索クエリ上の共起語情報を導入する.

(1) の設定では BERT の事前学習（Masked LM[14]）時のデータセットにウェブ検索クエリコーパスを用いる. これにより, 対象語についての検索クエリ上の共起語情報が暗黙的にモデルに取り込まれる. また,(2) の設定では対象語に対して, その語と共起するウェブ検索クエリ上の語を全て [SEP] トークンで連結することでクエリ共起語付きデータを生成する（例: “紀尾井タワー [SEP] アクセス [SEP] 行き方 [SEP] レストラン”）. これにより, モデルは対象語とクエリ共起語を区別でき, 一つの対象語に対して複数のクエリ共起語を付与できる. [3] このデータセットを用いた fine-tuning により検索クエリの共起語情報が明示的にモデルに取り込まれる.

3 実験

本章では構築したデータセットを用いて二値分類モデルの学習・評価を行う. さらに, 実際に非曖昧ランドマーク辞書の拡張を行い, 固有表現抽出システムへ導入した場合の影響を報告する.

3.1 実験設定

2.1 節で構築したデータセットからランダムにサンプリングを行い, 正例 13,443 件, 負例 62,902 件を得た. これを層化分割し, 学習データ 56,345 件, 検証データ 10,000 件, テストデータ 10,000 件を得た. 評価指標としては Recall,Precision,F1 値を用いた. また, 以下では構築した各モデルの設定について述べる.

2) 仮に出現する場合も, その周辺文脈のうちどの範囲を分類の手がかりとするかは自明ではない. 仮に出現文書全体を手がかりとする場合は分類モデルの不要な複雑化を招く [13].

3) ユーザーの入力した生の検索クエリ（例: “アクセス 紀尾井タワー”,“紀尾井タワー”）を直接分類対象とする設定も考えられるが, (1) 対象語とクエリ共起語の区別,(2) 検索クエリの出現頻度の偏りの調整など, いくつかの課題を解消する必要があり煩雑となる.

表 1 各モデルの二値分類タスクにおける評価結果

モデル	Precision	Recall	F1
頻度フィルタ	0.512	0.900	0.652
接尾辞一致	0.863	0.562	0.680
fastText	0.875	0.800	0.836
+入力共起語	0.843	0.722	0.778
東北大 BERT	0.876	0.901	0.888
+入力共起語	0.894	**0.920**	0.906
クエリ BERT	**0.904**	0.899	0.901
+入力共起語	0.900	0.915	**0.907**

BERT ベースラインモデルとして日本語 Wikipedia で事前学習された BERT(以下東北大 BERT) を用いる.[4] また, ウェブ検索クエリコーパスで事前学習された BERT（以下クエリ BERT）としてはヤフー社内製のものを用いる.[5] 両モデルともに最終層の [CLS] トークンに対応する表現を線形分類器に入力することで分類する.

次に,2.2 節で述べたクエリ共起語付きデータの構築法について説明する. 直近のヤフー検索[6]のクエリから 300 万件標本抽出し, それらをスペース区切りで分割し, 各語に対する異なる共起語を最大 10 件までランダムに収集する. [7] この共起語辞書を用いて分類対象の語に対応する共起語を取得し,[SEP] トークンで連結させることで, クエリ共起語付きデータセットを構築する. 実験では（1）対象語のみの場合（2）得られた共起語を全て（最大 10 件）付与した場合の 2 つの設定で提案モデルを比較した.[8]

fastText 深層学習を活用して文書分類を解く強力なモデルとして fastText 分散表現の平均ベクトルを入力として線形層で分類する手法 [15] が知られており, 本研究のベースラインとして採用した. このモデルでも BERT での処理 3.1 と同様に学習データに共起語を付与した設定を検証する. この際, 対象語と共起語は単にスペース区切りで連結し, 前処理として MeCab[16] による単語分割を行った.

頻度フィルタ 非曖昧ランドマーク語は一般名詞群やランドマーク以外の意図を持つ語と比較し, ウェブ検索クエリ上での出現頻度が低いことが経験的にわかっている. そこでウェブ検索クエリコーパスでの頻度が一定以下である語を非曖昧ランドマークとするルールを作成した. なお, この閾値としては学習データで最も F1 値が高くなる値を採用した.

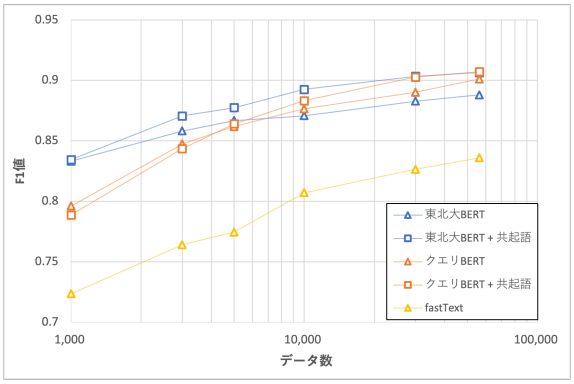

図 2 データサイズの影響. 横軸は対数スケール.

接尾辞一致 非曖昧ランドマーク語となり得る語には “○○病院”,“○○郵便局” など特定の接尾辞を持つ語が多い. そこで事前に既存の非曖昧ランドマーク辞書語に頻出する接尾辞を収集し, その接尾辞リストに対象語の接尾辞がマッチした場合に非曖昧ランドマークとするルールを作成した.

なお, 付録 C にその他の詳細設定を記載した.

3.2 実験結果

表 1 に実験結果を示す. 結果から, ルールベースの手法と比較すると機械学習モデルの特に BERT モデルの性能が高いことが分かる. また, 入力共起語を用いない BERT モデルでは東北大 BERT より, クエリ BERT の性能が高い. また, 両 BERT モデルで入力共起語を付加した場合, さらに性能が向上しており, [9] 特にクエリ BERT + 入力共起語モデルでは 0.907 の F1 値を達成している. これらの結果は, ウェブ検索クエリによる共起語情報を事前学習時,fine-tuning 時のそれぞれで導入する有効性を示している.

3.3 データサイズの影響

非曖昧ランドマーク判定に用いる二値分類モデルを構築する際, 今回のように必ずしも数万件単位の学習データが用意できるとは限らない. 従って, 本節では学習データ数ごとの各モデルの性能を比較することで, 少量の初期データしか手に入らない実践的な設定での提案手法の効果を確認する (図 2). その結果,1,000〜5,000 程度のデータ数[10]でも各 BERT モデルで F1 値 0.8 前後のスコアが確認されたが, この

4） huggingface.co/cl-tohoku/bert-base-japanese-v2
5） 検索クエリは短いため,Masked LM のみで事前学習している. techblog.yahoo.co.jp/entry/2021122030233811/
6） https://search.yahoo.co.jp/
7） 出現頻度が高い共起語を優先する方法も考えられるが, ランダムに収集する場合とほぼ性能が変わらないことが予備実験で確認されている.
8） なお, 入力共起語数の増加に伴い, 分類性能が向上することが実験で確認されている (付録 B).
9） fastText モデルでは入力共起語の付加により性能が低下した. このモデルでは対象語と付加共起語の区別がつかず, 付加共起語がノイズとして働いてしまったためだと考えられる.
10） 正例 (非曖昧ランドマーク語) 数は 180~900 程度である.

範囲のデータ数ではクエリ BERT より東北大 BERT に基づくモデルの性能が高いことが確認された. これはクエリ BERT では一般的な語の知識が欠けており, その獲得のために一定の学習データが必要であることが理由として考えられる. また, クエリの共起情報を保持していない東北大 BERT では, データ数 3,000 ほどから fine-tuning 時にクエリ共起語を付加する効果が現れ始めた. 以上から学習データが少ない場合は東北大 BERT+入力共起語モデルを活用することで効率良く辞書を拡張できることがわかる.

3.4 固有表現抽出でのオフライン評価

本節では構築した二値分類モデルを用いて非曖昧ランドマーク辞書の拡張を行い, 非曖昧ランドマーク辞書を補助的に利用する固有表現抽出システム[11]に実際に導入した際の効果を報告する. 分類対象のエンティティ名リストは拠点名称とそのメタ情報を含む拠点名称データベースから取得する. このデータベースでは図 1 の例のように, 曖昧なランドマーク名が混在している. このデータベースの拠点名称群に対して前処理[12] を行い,3.1 節で構築したクエリ BERT モデルによる非曖昧ランドマーク語の抽出を行うことで, 既存の非曖昧ランドマーク辞書を 10.9%拡張した.

その結果, 直近一週間のウェブ検索クエリに対するランドマーク認識率は 7.5%増加した. この増加分を人手で評価した結果,98.7%が改善,1.3%が改悪と判断された. この結果から, 提案手法の分類モデルによる非曖昧ランドマーク辞書拡張が固有表現抽出システムに有用であることが確認された. また改悪例ではチェーン店を誤認識している例が散見されたため, この改善が今後の課題として考えられる.

4 定性分析

本章では 3.1 節のテストデータを用いて定性分析を行う. まず, ルールベースと機械学習ベースのモデル群の差分について述べる. 頻度フィルタによる分類では「検索されやすい高頻度ランドマーク語」を誤って負例と判定する場合があった. また, 接尾辞一致では “大学病院” などの「拠点名称になり得る接尾辞を持つ一般的な語」も正例判定してしまう場合があった. これらの語について, 提案手法で構築した機械学習モデルは正確に分類できており, 単なる頻度情報や対象語の表記情報のみに依存しない分類が可能であることが確認された.

次に, 検索クエリによる事前学習の効果を確かめるため, 東北大 BERT とクエリ BERT を比較する. その結果,“大社” や “亭” などの「低頻度だがランドマークに用いられやすいサブワードを含む語」をクエリ BERT で正例判定可能となった例が確認された. さらにクエリ BERT では “甲子園” などの「様々な意図で用いられる曖昧語」を負例判定可能となった例がいくつか確認された. これらの結果はクエリ事前学習により, 検索クエリ上における共起語情報が埋め込まれたことが原因と考えられる.

最後に, 入力データに検索クエリを付加する効果を確かめるため, クエリ BERT とクエリ BERT+入力共起語を比較する. クエリ BERT 単体では例えば “カケーメセレゼ”[13]のような「対象語の表記情報にランドマークに関連するサブワードが一切なく, かつウェブ検索クエリ上で低頻度な語」に対して誤分類する例が散見された. 一方で, 入力共起語を用いたモデルではこれらの語において “BBQ”,“食事”,“道の駅” などの拠点名称を想像できる入力共起語により正例判定が可能となった例や “ツイッター”,“本名”,“アルバイト” など様々なドメインの入力共起語により負例判定が可能となった例が確認された. これらの結果はウェブ検索クエリ事前学習時に共起情報がモデルに十分に埋め込まれなかった語に対して, 明示的に入力データに共起語を付加することで,fine-tuning 時にモデルが共起情報を考慮できるようになったためだと考えられる.

5 おわりに

本稿ではエンティティ名リストからの非曖昧ランドマーク語自動抽出を目指し, 検索クエリの共起語情報を考慮した分類モデルを提案した. 実験では提案手法の有効性が確認され, さらに実際に提案モデルにより既存の非曖昧ランドマーク辞書を拡張することで, 固有表現抽出システムの改善を実現した.

本研究では拠点名称ドメインを対象としたが, レシピやショッピング, 医療などの他の異なるドメインにおいても非曖昧ドメイン固有語辞書の需要がある. 従って本研究のアイデアを他のドメインに適用することが今後の発展として考えられる.

11） “ランドマーク”,“チェーン店” などの場所に関するカテゴリを付与する固有表現抽出システムである. 一部辞書とパターンマッチに基づく判定が行われる.

12） 前処理としては既にデータセットにある語の排除, メタデータから判断できるノイズデータの削除などを行った.

13） 説明のために作成した架空のランドマーク名である.

参考文献

[1] Su Nam Kim and Lawrence Cavedon. Classifying domain-specific terms using a dictionary. In **Proceedings of the Australasian Language Technology Association Workshop 2011**, pp. 57–65, December 2011.

[2] Lucas Hilgert, Lucelene Lopes, Artur Freitas, Renata Vieira, Denise Hogetop, and Aline Vanin. Building domain specific bilingual dictionaries. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 2772–2777, May 2014.

[3] Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. Extracting domain-specific words - a statistical approach. In **Proceedings of the Australasian Language Technology Association Workshop 2009**, pp. 94–98, December 2009.

[4] Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. Named entity recognition over electronic health records through a combined dictionary-based approach. **Procedia Computer Science**, Vol. 100, pp. 55–61, 2016.

[5] Reiko Hamada, Ichiro Ide, Shuichi Sakai, and Hidehiko Tanaka. Structural analysis of cooking preparation steps in japanese. In **Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages**, IRAL '00, p. 157–164, 2000.

[6] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. **Journal of Biomedical Informatics**, Vol. 46, No. 6, pp. 1088–1098, 2013. Special Section: Social Media Environments.

[7] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing**, pp. 1524–1534, July 2011.

[8] Woojin Paik, Elizabeth Liddy, Edmund Yu, and Mary McKenna. Categorization and standardizing proper nouns for efficient information retrieval. In **Acquisition of Lexical Knowledge from Text**, 1993.

[9] Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in WordNet. In **Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing**, pp. 168–175, 2003.

[10] 村脇有吾, 黒橋禎夫. テキストから自動獲得した名詞の分類. 言語処理学会年次大会発表論文集, pp. 716–719, Mar 2010.

[11] Judita Preiss and Mark Stevenson. Distinguishing common and proper nouns. In **Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity**, pp. 80–84, June 2013.

[12] 松田耕史, 佐々木彬, 岡崎直観, 乾健太郎. 場所参照表現タグ付きコーパスの構築と評価. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2015, No. 12, pp. 1–10, 01 2015.

[13] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. **arXiv:2004.05150**, 2020.

[14] 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, June 2019.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 427–431, April 2017.

[16] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, July 2004.

[17] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Recent trends in word sense disambiguation: A survey. In Zhi-Hua Zhou, editor, **Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21**, pp. 4330–4338, 8 2021. Survey Track.

[18] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, Vol. 30, No. 1, pp. 3–26, 2007.

[19] Qing Liu, Hongyu Lin, Xinyan Xiao, Xianpei Han, Le Sun, and Hua Wu. Fine-grained entity typing via label reasoning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 4611–4622, November 2021.

[20] Davi Reis, Felipe Goldstein, and Frederico Quintao. Extracting unambiguous keywords from microposts using web and query logs data. In **Making sense of Microposts (at WWW 2012)**, 2012.

[21] Aixin Sun. Short text classification using very few words. In **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '12, p. 1145–1146, 2012.

[22] Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. In **Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '07, p. 231–238, 2007.

[23] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In **Proceedings of the 17th International Conference on World Wide Web**, WWW '08, p. 91–100, 2008.

[24] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. Understanding short texts through semantic enrichment and hashing. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 28, No. 2, pp. 566–579, 2016.

[25] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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. 380–385, June 2019.

[26] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In **WWW '05: Proceedings of the 14th International World Wide Web Conference**, pp. 391–400. ACM Press, 2005.

[27] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In **Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '06, p. 131–138, 2006.

[28] 佐々木彬, 五十嵐祐貴, 渡邉陽太郎, 乾健太郎. 場所参照表現のグラウンディングに向けて. 言語処理学会年次大会発表論文集, pp. 177–180, Mar 2014.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**, 2019.

A　関連研究

本研究で行う非曖昧な語の抽出は, 語義曖昧性解消 [17] や固有表現抽出 [18],Entity Typing[19] のように与えられた文脈に基づき対象語の曖昧性を解消するのではなく, 元来語義曖昧性のない（低い）語を選別する. 最も類似した研究として,Reis らの研究がある [20]. Reis らは英語の SNS テキストを対象とする非曖昧キーワード抽出器の構築に取り組んだ. 彼らはウェブクエリやウェブコーパスから得られる逆文書頻度を始めとした様々な統計的特性を用いたヒューリスティックなフィルタリングルールと, そのルールによる分類結果を用いて学習した SVM 分類器を提案した. 本研究では日本語の拠点名称ドメイン固有語の抽出に対象を絞り, 人手による学習データとウェブ検索クエリを用いた比較的シンプルな設定で汎用言語モデルを学習することで, 実用的な性能を持つ二値分類モデルを構築する. また, 抽出された非曖昧ランドマーク語を固有表現抽出システムに取り込み, システムの性能改善を試みる.

本研究と関連する名詞辞書構築の研究としては, 特定ドメインを特徴づけるドメイン固有語辞書 [1, 2, 3] の構築を目指した研究や一般名詞と固有名詞を分類する研究 [9, 10, 11] などがある. これらの研究と比較すると主に (1)“非曖昧” ドメイン固有語か否かの分類を行う点,(2) 分類の素性にウェブ検索クエリを活用する点で本研究は異なる. (1) について例えば Kim らの研究では石油ドメイン固有語として “gulf” を紹介しているが, これは石油に関係ない文脈で “湾” として登場したり, 一般的に “（意見の）隔たり” という意味でも利用されるため, 本研究では収集対象外の語となる. さらに複数のドメイン意図を持つ固有名詞（“甲子園”）が存在するため, 単に固有名詞を収集するわけでもない. このように名詞辞書構築という観点で特定ドメインの意図以外でほぼ登場しない非曖昧ドメイン固有語を自動収集する方向性は, 我々の知る限り着目されてこなかった. また,(2) について例えば村脇らは対象語のウェブコーパス中の振る舞いを素性として用いることで固有名詞群の分類を行っている. しかし, 一部の有名なランドマークを除いて多くの非曖昧ランドマーク語は自然文コーパスに登場しないため [12], 本研究では分類の素性にウェブ検索クエリを活用する.

また, 本研究は比較的短い文書の分類を行う短文書分類タスクと関連がある. このタスクの有力な解決手段として特徴量拡張による手法が提案されており, 検索エンジンの結果 [21, 22] や大規模知識ベース [23, 24] などの何らかの外部言語資源から得た特徴量を追加で考慮することで分類性能を向上させている. 本研究では非曖昧ドメイン固有語の抽出を短文書分類タスクとして扱い, 汎用言語モデル BERT の Masked LM[14] の事前学習データや fine-tuning 時の付加データ [25] としてウェブ検索クエリのテキスト情報を直接活用することで分類性能の向上を試みる.

また, ユーザーの入力した検索クエリ自体に対してカテゴリ分類もしくは意図推定を行う研究がいくつか報告されている [26, 27]. 本研究の分類対象はあくまでエンティティ名であり, 分類性能を向上させるために検索クエリにおける共起語情報を補助的に用いる.

また, 文書中から場所参照表現を抽出し位置情報と紐付ける研究がいくつか行われている [12, 28]. 本研究は固有表現抽出タスクではなくエンティティ名を分類するタスクである. 本研究の成果により構築された非曖昧ランドマーク語辞書を活用することで, ランドマーク意図を持つ語をどのような文脈でも高い適合率で固有表現抽出することが可能となる.

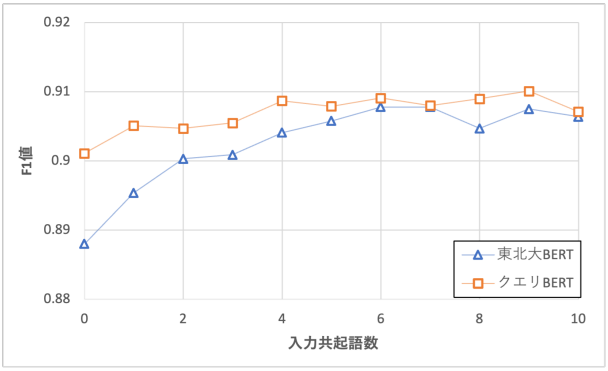

図 3　入力共起語数の影響

B　入力共起語数の影響

本節では入力共起語数の影響を調査するため, 入力共起語数ごとの性能推移を BERT モデルで観測した（図 3）. その結果, 両 BERT モデルで入力共起数の増加に伴い, 性能が向上することが確認された. また, 性能の向上幅は東北大 BERT の方が大きく, 入力共起語 10 個を利用した場合はクエリ BERT の性能とほぼ同等の性能を達成していることが確認された. この結果は, ウェブ検索クエリ事前学習を行わなくても, 共起語を fine-tuning 時に付加することで検索クエリの共起情報を BERT に導入できることを示している.

C　実験設定詳細

本節では本文から省略した実験設定の詳細について説明する. BERT モデルの学習では学習率を 2×10^{-5}, バッチサイズを 64 にし, パラメータ更新の最適化アルゴリズムは AdamW[29] を用いた. また, 検証データにおける F1 値が改善されなくなるまで学習を行った. また,BERT モデルの実装には `transformers`[14],fastText 文書分類モデルの実装には `fastText`[15] を用いた. 機械学習ベースのモデルでは異なるシード値で 3 回実験を行い, その結果の平均を示した.

14）https://huggingface.co/
15）https://fasttext.cc/