日経企業 ID リンキングのための類似度ベース EL システムの構築と分析

澤田悠治 [1] 安井雄一郎 [2] 大内啓樹 [1,3,4] 渡辺太郎 [1]
石井昌之 [2] 石原祥太郎 [2] 山田剛 [2] 進藤裕之 [2]
[1] 奈良先端科学技術大学院大学 [2] 日本経済新聞社
[3] 理化学研究所革新知能研究センター [4] 国立国語研究所
{sawada.yuya.sr7,hiroki.ouchi,taro,hiroyuki.shindo}@naist.ac.jp
{yuichiro.yasui,masayuki.ishii,shotaro.ishihara,tk.yamada}@nex.nikkei.com

概要

本研究は，日本経済新聞社の記事に登場する企業名を日経企業 DB へリンクするタスクを設計し，新聞記事中に企業名に対するリンキング性能を分析する．企業名へのリンキング性能を分析するため，新聞記事に出現する企業名と日経企業 ID を紐づけしたデータセットを作成し，事前学習済み LUKE モデルのスパン類似度学習によるリンキングシステムを構築する．作成したデータセットを用いた評価実験では，実装システムが既存システムを上回るリンキング性能を示すことを確認し，実装システムの企業 ID リンキングにおける課題について整理する．

1 はじめに

日本経済新聞社は，新聞記事・適時開示情報などの過去に作成したテキスト，企業情報などの各業界に関するデータベース (日経企業 DB) を保有している．本研究では，新聞記事と日経企業 DB に含有された知識の統合化に焦点を当て，新聞記事に登場する企業名を日経企業 DB の企業 ID に紐づける企業 ID リンキングシステムを作成する．新聞記事の企業名を企業 ID と紐づけるためには，文中から企業名を抽出する固有表現認識モデルと，抽出した企業名に対して企業 ID を割り振るエンティティリンキング (Entity Linking; EL) モデルが必要になる．日本語の固有表現認識では，GiNZA[1)] などの汎用 NLP ツールが提供されているものの，これらのツールでは企業名は組織名 (Organization) の一部として定義されているため，GiNZA で抽出された組織名から企業名かそれ以外の組織名か分類する必要がある．また，既存の日本語 EL システム [1, 2] は主に Wikification タスクを目的として設計されているため，日経企業 ID を対象にしたリンキングへの適応は難しい．そこで本研究では，企業 ID リンキングシステムの作成に必要なデータセットを整備し，事前学習済み日本語言語モデルを用いた実装システムによるリンキング性能を報告する．また，実装システムの実験結果から企業 ID リンキングタスクの課題を分析する．

2 日経企業 ID リンキング

2.1 タスク概要

日経企業 ID リンキングは，略称を含む直接的に企業名称を指す表現[2)] を企業名と仮定し，新聞記事に記述されている企業名を日経企業 DB に登録された企業 ID とリンキングする．本研究では，2018 年 1 月から 2022 年 6 月までに掲載された朝刊からサンプリングした記事と，日本経済新聞記事オープンコーパス [3] の記事から成る 3,025 件の新聞記事からデータセット (以下，日経データ) を構築する．企業 ID は 2022 年度時点の日経企業 DB のものを使用し，後節のアノテーション定義に従って企業名と企業 ID を付与する．

2.2 アノテーション定義

日経データでは，文中に含まれる特定の企業を指す記述に，それぞれ日経企業 DB の企業 ID をラベル付けする．企業名においては，新聞記事に記述される企業名の最長範囲を抽出対象とし，入れ子や不連続な範囲から成る固有表現は対象から除外する．

1) https://megagonlabs.github.io/ginza/

2) 本研究では “あの企業”, “当企業” などの特定の企業を間接的に示す表現は対象外とする．

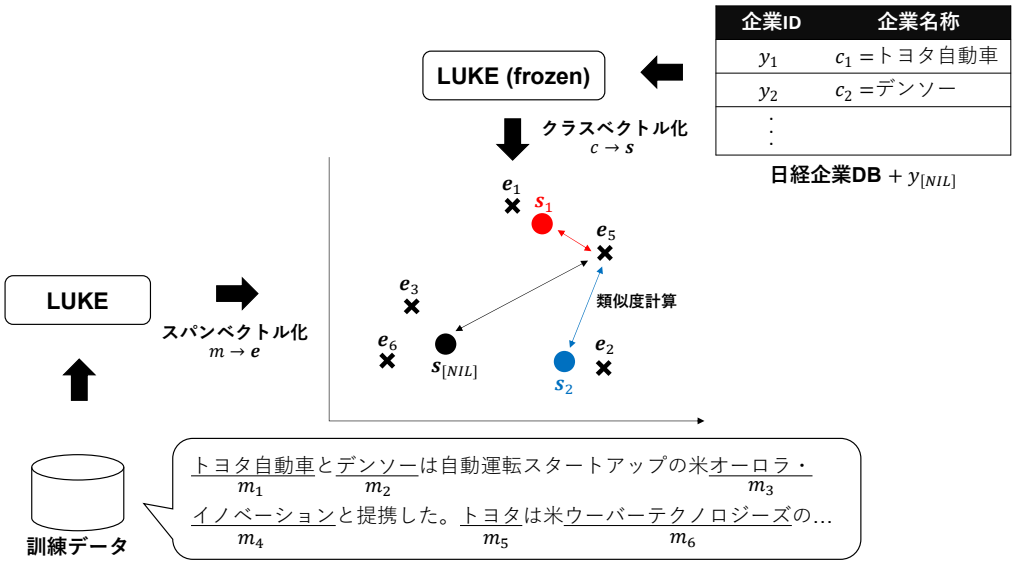

図 1　LUKE を用いた類似度ベース企業 ID リンキング (LUKE-Similarity)

例えば，子会社の企業名は “ネッツトヨタ愛知” のように親会社の企業名が入れ子状に含まれるため，これらの記述は子会社の企業名のみ抽出対象とする．また，株式会社の略語表現 (“ヤンセン社”) は企業名の範囲に含め，“日経平均株価” のような用語内部に含まれる企業名は対象から除外する．

企業 ID を付与する際は，日経企業 DB に登録された各企業の正式名称や会社情報を元に最も当てはまる企業 ID をラベル付けする．企業 ID が一つに絞り込めない場合は，リンク先に成りうる全ての日経企業 ID をラベル付けし，評価時は複数の企業 ID のいずれかの企業 ID が出力されていれば正解とする．

3　企業 ID リンキングシステム

日経企業 ID リンキングシステムは，入力テキストから企業名を抽出する企業名抽出器と，各企業名を該当する企業 ID に紐づける企業 ID リンキング器のパイプラインで構成されている．企業名抽出器は日本語 LUKE モデル [4] を固有表現認識タスク向けに Finetuning したモデル [5] を使用し，企業 ID リンキング器は日本語 LUKE モデルのスパン類似度から企業 ID または [NIL] を分類するモデル (LUKE-Similarity) を作成する．

3.1　LUKE-Similarity

LUKE-Similarity の概要図を図 1 に示す．LUKE-Similarity は，企業 ID リンキングタスクを日経企業 DB C に登録されている企業 ID $y \in C$ ($C = \{y_1, y_2, ..., y_{|C|}\}$) と [NIL] から成るラベル集合 $\mathcal{Y}$ ($\mathcal{Y} = \{y_1, ..., y_{|C|}, y_{[\mathrm{NIL}]}\}$) の分類問題と仮定し，事前学習済み日本語 LUKE モデルの類似度計算によってラベルを付与する．具体的には，文字数 n のテキスト $X = (x_1, x_2, ..., x_n)$ において，任意の始点と終点 i, j ($1 \le i \le j \le n$) のスパンから成る k 個の企業名 $m \in \mathcal{M}$ ($\mathcal{M} = \{m_1, m_2, ..., m_k\}$) に，該当する企業 ID または [NIL] を表すクラスラベル y を推定する．各企業名 m からクラスラベル y に割り当てる確率は以下の式より計算する．

$$\mathrm{P}(y|m) = \frac{\exp\left(\mathrm{score}(m, y)\right)}{\sum_{y' \in \mathcal{Y}} \exp(\mathrm{score}(m, y'))} \tag{1}$$

ここで，企業名 m とクラス y の類似度を表す $\mathrm{score}(m, y)$ は，LUKE から出力される企業名 m のスパンベクトル $\boldsymbol{e}$ と，企業 ID または [NIL] y を表すクラスベクトル $\boldsymbol{s}$ の cosine 類似度である．企業名のスパンベクトル e とクラスベクトル $\boldsymbol{s}$ は以下の式より求められ，[NIL] のクラスベクトルは，ランダム値で作成したベクトルを使用する．

$$\boldsymbol{e} = \mathrm{LUKE}(X, i, j) \tag{2}$$

$$\boldsymbol{s} = \begin{cases} \mathrm{LUKE}(c, 1, |c|) & (y \neq [\mathrm{NIL}]) \\ \mathrm{random}() & (y = [\mathrm{NIL}]) \end{cases} \tag{3}$$

$$\mathrm{score}(m, y) = \frac{\boldsymbol{s} \cdot \boldsymbol{e}}{||\boldsymbol{s}||\,||\boldsymbol{e}||} \tag{4}$$

式 2,3 において，企業名のスパンベクトル $\boldsymbol{e}$ はテキスト X と位置情報 (i, j)，[NIL] 以外のクラスベクトル $\boldsymbol{s}$ は企業 ID y に対応する企業名称 c と企業名称の全体範囲 $(1, |c|)$ を入力してベクトルを取得する．モデル学習時は，クラスベクトル $\boldsymbol{e}$ は更新せず，以

	NER(正解)+EL			パイプライン		
	R@1	R@5	R@10	適合率	再現率	F 値
T-laei	-	-	-	82.7	46.9	59.9
辞書マッチ	69.8	71.4	71.4	51.2	62.4	56.3
LUKE-Similarity (Zero-shot)	47.8	54.3	55.9	35.1	42.9	38.6
LUKE-Similarity (日経データ)	89.0	90.2	91.0	81.3	84.9	83.0

表 1　企業 ID リンキングの実験結果

下の式より正解企業 ID のクラスベクトルと類似度が最大になるように損失を最小化する．

$$\mathcal{L} = -\sum_{n=1}\sum_{n=1} \log(\mathrm{P}(y|m)) \tag{5}$$

推論時は，企業名抽出器から抽出した企業名 m' のスパンベクトルとモデル学習時に作成したクラスベクトルの cos 類似度を計算し，最も高い類似度を示したクラスの企業 ID または [NIL] ラベル $\hat{y}$ をリンク先として推定する．

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, \mathrm{P}(y|m') \tag{6}$$

4　評価実験

4.1　実験設定

	Train	Dev	Test
記事	2,340	586	100
段落	7,992	1,991	331
企業名	6,956	1,813	245
企業 ID	4,269	1,082	159
企業 ID の異なり	1,058	402	81
未知企業 ID	-	254	62

表 2　評価実験時データセットの基本情報

日経データを用いて，各企業名に対する日経企業 ID のリンキング性能を評価する．本研究では，企業 ID リンキングシステムのみのリンキング性能と，企業名抽出器と組み合わせたパイプラインシステムによるリンキング性能を評価する．リンキングの評価では，企業名の全正解スパンを EL モデルに与え，各スパンと最も類似度が高い上位 $k(k \in \{1, 5, 10\})$ 件の企業 ID の中から正解企業 ID が含まれているか評価する．パイプラインシステムの評価では企業名抽出モデルから抽出されたスパンに対して，企業 ID リンキングシステムのスパンとリンク先の一致を適合率・再現率及びそれらの調和平均 (F 値) で評価する．また，新聞記事の段落を実装システムへの入力とし，日経企業 ID データセットは表 2 のように全 3,025 記事のうち 100 記事を評価用として使用する．

4.2　比較手法

日本経済新聞社が採用しているルールベース情報抽出システム (T-laei) をベースラインとして採用する．T-laei は企業名の抽出とリンキングを End-to-End で行うため，リンキングの評価では Levenstein 距離 [6] を用いた辞書マッチを比較手法に追加する．Levenstein 距離は長さが企業名と企業名称の文字数に依存するため，それぞれの Levenstein 距離に対して文字列同士の最大編集距離で正規化し，最短の距離を示した企業エントリをリンク先の日経企業 ID として出力する．また，文字列類似度以外の尺度として事前学習済み LUKE のパラメータから最大スパン類似度を検索する Zero-shot モデルの性能も示す．パイプラインの評価においては，どちらも日経データを使って作成した企業名抽出器の出力を使用する．

4.3　実験結果

実験結果を表 1 に示す．企業名の正解スパンを対象にしたテストデータの評価において，LUKE-Similarity(日経データ) は全てのレベルで辞書マッチより高い再現率を示しており，9 割程度の企業名が LUKE のスパン類似度によってリンキングできている．一方で，事前学習済み LUKE のパラメータをそのまま用いる LUKE-Similarity(Zero-shot) は，辞書マッチと比べて R@1 で 20 ポイント程度，R@10 で 15 ポイント程度の差が見られる．これは [NIL] のベクトルをランダム値で設定したことと，企業ベクトルを日経企業 DB の全企業名称から作成したことで，略語などの企業名称と異なる表現に対してリンク先を分類できていないことが原因として考えられる．実際に，[NIL] をランダム値のベクトルではな

入力記事 (2)	バイオ創薬の**ナノキャリア**は創薬支援を手掛ける アクセリード (神奈川県藤沢市) と 4 月に共同出資会社を立ち上げた.
正解企業名称	ナノキャリア
LUKE-Similarity(記事)	[NIL]

表 3　LUKE-Similarity の出力誤り例

入力記事 (2)	旧**昭和シェル石油**が中心となって 2001 年に創設した.
正解企業名称	RS エナジー
LUKE-Similarity(記事)	昭和シェル船舶

表 4　企業名の時間変化による LUKE-Similarity の出力誤り

く類似度の閾値[3]から求めると R@1 は 23 ポイント程度増加しており，リンク先の存在しない企業名は日経企業 DB 中の企業名称とも類似度が低い傾向にある．しかし，この再現率も辞書マッチと近い値であるため，事前学習済み LUKE のスパンベクトル自体は文字列類似度と同程度の効果と考えられる．

4.4　分析

	未知企業 ID		
	適合率	再現率	F 値
辞書マッチ	64.3	58.1	61.0
LUKE-Similarity(文)	57.7	48.4	52.6
LUKE-Similarity(段落)	60.0	58.1	59.0
LUKE-Similarity(記事)	66.7	58.1	62.1

表 5　訓練事例に出現しない企業 ID のリンキング性能

新聞記事は主に経済や政治に関する内容が多く，一部の大手企業が頻繁に出現する傾向にある．そのため，中小企業やベンチャー企業などの日経企業 ID は登場しにくく，訓練データで出現しない日経企業 ID に対するリンキングは，一般 EL タスクと同様に困難な問題として考えられる．訓練事例に存在しない企業 ID に対するリンキング性能を表 5 に示す．訓練事例に存在しない企業 ID において，LUKE-Similarity は辞書マッチとほぼ同程度となっており，未知な企業 ID に対してリンキング性能が改善されたとは考えにくい．実際に実装システムによる出力 (表 3) では，正解の企業名称と文中の企業名が一致するにも関わらず [NIL] と予測されている．これは日経データが外国企業を [NIL] として扱っているため，文脈と表記によって外国企業と誤って予測された可能性がある．

また，企業の名称は M&A などの組織改編によって変わりやすい特徴があり，記事の掲載時期と日経企業 DB の更新時期の異なりから企業名と企業名称のミスマッチが頻繁に起きる．例えば，2019 年以前の新聞記事で記載される「ソニー」は現在の「ソニーエレクトロニクス」を指しているのに対して，2019 年以降の「ソニー」は「ソニーグループ」を指すなど，同じ企業名でも時間軸の変化によってリンク先の企業 ID が変化することがある．実際に企業名と企業名称の間に時間変化がある事例 (表 4) では，LUKE-Similarity は過去に存在した “昭和シェル石油” の文字列と類似する企業がリンキングされており，一新された企業名称とのリンキングは困難である．これらの結果を踏まえて，LUKE-Similarity は未知企業 ID におけるリンキング性能と企業名の変遷においては依然として課題が残る．

5　おわりに

本研究は，日本経済新聞社の新聞記事と日経企業 DB の知識統合化のための企業 ID リンキングデータセットを作成した．作成したデータセットを元に類似度ベース企業 ID リンキングシステムを実装した結果，既存システムよりも高い抽出性能を達成した．しかし，実装システムの出力から，未知の企業 ID や企業名の時間変化には改善の余地があることが示された．企業の基本情報や業種カテゴリなどを用いた企業ベクトルの作成や，企業 ID の有無以外に外国企業であるかを事前に分類をする機構の導入は今後の課題である．

3)　企業名リストの全組み合わせのなかで最大 cosine 類似度が 0.98 未満の場合，対象企業名のリンク先を [NIL] とする．閾値は訓練データと開発データを用いて探索した．

参考文献

[1] Davaajav Jargalsaikhan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. Building a corpus for japanese wikification with fine-grained entity classes. In **Proceedings of the ACL 2016 Student Research Workshop**, 2016.

[2] 関根聡, 中山功太, 隅田飛鳥, 渋木英潔, 門脇一真, 美佑 宇佐 まや 安藤明波. 森羅タスクと森羅公開データ. 言語処理学会第 29 回年次大会発表論文集, 2023.

[3] 浅原正幸, 高松純子, 若狭絢, 大村舞. 日本経済新聞記事オープンコーパス：新聞記事コーパスと形態・統語情報アノテーション. 言語処理学会第 29 回年次大会 併設ワークショップ JLR2023, 2023.

[4] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2020.

[5] Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. In **ArXiv**, 2023.

[6] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. **Soviet physics. Doklady**, Vol. 10, pp. 707–710, 1966.

[7] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. In **Intelligent Systems**, 2020.

[8] 近江崇宏. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会第 27 回年次大会 発表論文集, 2021.

[9] 橋本航, 笛木正雄, 黒木裕鷹, 髙橋寛治. 日本語固有表現抽出における bert-mrc の検討. 言語処理学会第 28 回年次大会 発表論文集, 2022.

[10] Tjong Kim Sang, Erik F., 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.

A　企業名抽出器の予備実験

A.1　実験方法

文中の各企業名のスパンの一致について，適合率・再現率及びそれらの調和平均 (F 値) で評価する．比較手法に BERT-CRF[7]，データセット作成時に使用した Stockmark 日本語固有表現データセット [8] で LUKE を Finetuning したモデルを追加し，企業名抽出器の抽出精度を比較する．Stockmark 日本語固有表現抽出データセットは全 8 種類[4]の固有表現ラベルが付与されているため，本研究では法人名として予測されたスパンを企業名と仮定し，法人名と正解企業名スパンの一致について評価する．また，橋本ら [9] と同様に学習データ，開発データ，テストデータの比率が 8:1:1 になるようにデータセットを分割し，開発データで抽出精度が最も高いパラメータで企業名を抽出する．

A.2　実験結果

	訓練データ	適合率	再現率	F 値
T-laei	-	92.1	52.2	66.7
BERT-CRF	日経データ	91.3	94.7	93.0
LUKE-NER	Stockmark	76.6	93.5	84.2
LUKE-NER	日経データ	93.0	97.1	95.0

表 6　企業名抽出の実験結果

実験結果を表 6 に示す．テストデータにおいて，BERT-CRF(日経データ) と LUKE(日経データ) はルールベースの T-laei よりも高い再現率・F 値を示している．日経データは全企業名の 4 割のリンク先が [NIL] であるため，企業 ID を持つ企業名のみ抽出する T-laei は約 5 割程度の再現率となっている．BERT-CRF は LUKE と比べて 2 ポイント程度下回り，この結果は Yamada ら [4] の CoNLL2003[10] による実験結果と同様の傾向が見られている．LUKE(Stockmark) は LUKE(日経データ) よりも F 値が 10 ポイント程度下回ったものの，再現率はほぼ同程度の値を示している．Stockmark 固有表現認識データセットの法人名には大学名などの企業名と異なる固有表現が含まれているため，企業名以外の固有表現によって偽陽性の誤りが生じたことが原因として考えられる．

4)　人名，法人名，政治的組織名，その他の組織名，地名，施設名，製品名，イベント名

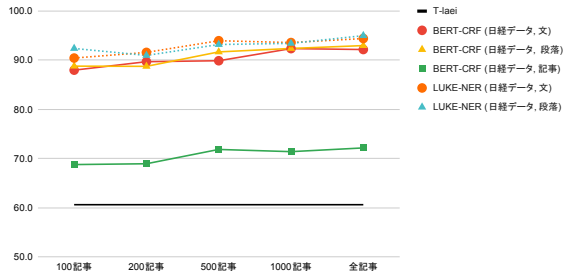

図 2　訓練用データの記事数ごとの F 値の推移．頂点が丸印は文，三角印は段落，四角印は記事レベルに入力テキストを分割して学習したモデルを表す．

企業名抽出器の抽出性能を訓練用の記事数ごとに比較した結果を図 2 に示す．BERT-CRF と LUKE は，100 件程度の記事から 500 記事にかけて F 値がやや増加傾向にあるものの，500 記事からほぼ横ばいの F 値で推移している．日経データの新聞記事は主に大手企業が頻出し，「タタ自動車」の「タタ自」，「現代自動車」の「現代自」のように企業名の略称も規則性が高い特徴が見られる．500 件以上の記事から大手企業を網羅し，略称の規則性を捉えたことで BERT-CRF と LUKE-NER が全記事と同程度の抽出性能が得られたと考えられる．また BERT-CRF を記事レベルで学習・評価した場合，文レベル・記事レベルと比べて性能が著しく低下した．記事を入力とした場合，文・段落レベルと比べてラベルの系列長がそれぞれ 8.9 倍・3.4 倍長くなるため，系列パターンの学習が困難になったことが性能低下の原因として考えられる．LUKE は入力文が長くなるほど計算量が指数的に増加するため，記事レベルに対応した企業名抽出器の検討は学習コストの面で今後の課題である．