Vol.2022-OS-157 No.1
Vol.2022-IOT-59 No.1
2022/9/13

研究資源のメタデータに関する情報を学術論文から抽出する試み

佐治 礼仁[1,a)]　松原 茂樹[1,2]

概要：本論文では，研究データリポジトリの効率的な拡充を目的に，研究資源のメタデータに関する情報を学術論文から抽出する試みについて述べる．具体的には，論文テキストに出現するエンティティ及びそれらの間の関係を抽出し，エンティティを節点，関係を有向辺とする知識グラフを獲得する．論文データを用いて構築した知識グラフを用いて，既存のメタデータリポジトリにおけるメタデータ及びエントリの拡充可能性を実験的に検証した．実験の結果，既存のメタデータに対する新たな情報の追加可能性，ならびに，研究資源を示すエンティティを識別するニューラルモデルによる研究資源エントリの拡充可能性を確認した．

Extracting Information on Research Resource Metadata from Scientific Papers

1. はじめに

オープンサイエンスは，学術論文や研究資源などの共有や利活用を推進する活動である．学術論文や研究資源を，そのメタデータとともにリポジトリに登録し公開することは一つの方策である．論文については，CiteSeerX[*1]，Google Scholar[*2]，Semantic Scholar[*3] などの検索サービスが提供されている．研究資源についても，Google Dataset Search[*4][1]，CiNii Research[*5] などの検索プラットフォームが実現されている．

学術論文の場合，通常，出版する過程において学会や出版社等でメタデータを作成する仕組みが確立されているため，それを用いて機械的にメタデータを整備し，リポジトリに登録できる．一方，研究資源のメタデータは，研究資源の生成者や登録者などが人手で作成することになる．例えば，Google Dataset Search では，Schema.org[*6] のスキーマに従ったメタデータを作成し，それを Web サイトに登録することで，プラットフォーム上でデータセットを参照可能となる．CiNii Research でも，連携プラットフォームである JAIRO Cloud に，人手で作成したメタデータを登録する必要がある．学術論文と比べ，研究資源をリポジトリに登録するための生成者や登録者のコストが大きく，リポジトリの発展あるいはメタデータの充実が進みにくい状況にある [2]．

そこで本研究では，研究資源のメタデータの自動生成を目的に，メタデータに関する情報を学術論文から獲得することを試みる．具体的には，論文テキストから抽出された情報が研究資源メタデータリポジトリの拡充に有用である，という仮説を設定し，実験的にこれを検証する．学術論文では，当該研究で生成あるいは利用された研究資源について言及されることが多く，これらをメタデータの作成に利用できる可能性がある[*7]．

上述の仮説を検証するために，本論文では，学術論文から獲得された情報を利用することの効果に関して，以下の２つの調査について述べる．

メタデータの拡充可能性　リポジトリに登録済みの研究

[1] 名古屋大学大学院情報学研究科
Graduate School of Informatics, Nagoya University
[2] 名古屋大学情報連携推進本部
Information and Communications, Nagoya University
[a)] saji.ayahito.y7@s.mail.nagoya-u.ac.jp

[*1] https://citeseerx.ist.psu.edu
[*2] https://scholar.google.com
[*3] https://www.semanticscholar.org
[*4] https://datasetsearch.research.google.com
[*5] https://cir.nii.ac.jp
[*6] https://schema.org
[*7] 加えて，既存のメタデータには存在しない情報が得られる可能性 [3] もある．

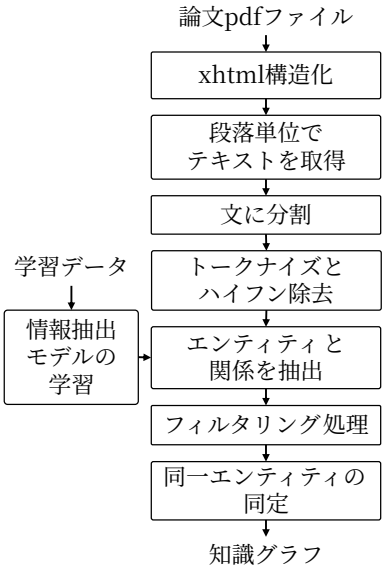

図 1 知識グラフの構築フロー

資源について，そのメタデータの情報が増大すること．

エントリの拡充可能性 リポジトリに登録された研究資源のエントリ数が増加すること．

本論文ではまず，論文テキスト内の特定の参照表現 (以下では，**エンティティ**と呼ぶ) ，及び，エンティティ間の**関係**を抽出する手法について述べる．エンティティを節点，エンティティ間の関係を有向辺として表現する知識グラフを獲得する．仮説の検証では，言語処理分野の国際会議論文 15,721 件から獲得した知識グラフを使用し，言語資源メタデータリポジトリ SHACHI[*8][4] を対象にその拡充実験を実施した．メタデータの拡充では，SHACHI の既存のメタデータに追加可能な情報の件数を調査した．エントリの拡充では，知識グラフを構成するエンティティに対し，それが研究資源を示すものか否かを判別するニューラルモデルを学習し，その性能を評価した．

本論文の構成は以下の通りである．続く 2 章で関連研究に言及し，3 章で知識グラフの構築手法について述べる．4 章でメタデータの，5 章でエントリの拡充可能性を検証する実験について報告する．

2. 関連研究

学術論文では，研究の過程で生成または利用された研究資源について言及される．学術論文から研究資源に関する情報を獲得し，研究資源リポジトリの構築や拡張を目指した研究が行われている．

学術論文における文献リストや注釈等の参照情報を用いた研究として，Ikoma ら [5] は，参考文献リストに記載された書誌情報から研究資源を参照するものを識別する手法を提案している．識別された書誌情報は，研究資源のメタデータ情報として利用できる．また，Tsunokake ら [6] は，論文テキストに記載された URL のうち，研究資源を参照するものを同定する手法を提案している．同定された情報は，研究資源メタデータにおける識別情報として利用できる．

一方，テキストからの情報抽出 [7] を論文テキストに適用することで，研究資源の名称や説明を示す文字列の抽出法が研究されている．Singhal ら [8] は，論文テキスト内の大文字で始まるトークン（標準辞書に出現しないものに限る）に対し，“dataset”との類似度を NGD（Normalized Google Distance）に基づき算出し，類似度が大きいものをデータセットとして検出する手法を提案している．Heddes ら [9] は，人工知能に関する国際会議論文の 6,000 文で学習した抽出モデルに基づき，論文テキストにおけるデータセット示すトークンを認識することを試み，高い認識性能を達成している．Prased ら [10] は，RTCC（Coleridge Initiative’s Rich Text Context Competition）を用い，データセットへの言及部分の抽出と分類の方法を提案している．また，Ikeda ら [11] は，データセット名の，分野に依存しない汎用的な抽出技術を提案している．この手法は，[8] と同様，候補トークンに対する，“dataset”や “database”等の単語との類似度計算に基づいている．他にも，ソフトウェアに対して同様のアプローチを採用している研究 [12], [13], [14] が存在する．さらに，Kozawa ら [15] は，研究資源の「用途」に関する情報を論文から自動で抽出する手法を提案している．この手法では，研究資源メタデータリポジトリに登録された研究資源の名称を手がかりに，学術論文において研究資源に関して言及された文を同定し，構文的パターンと照合することで用途情報を抽出する [4].

これに対して本研究は，論文テキストから研究資源に関する知識グラフを獲得する方法を示し，その知識グラフの利用が研究資源リポジトリの拡充に有用であることの実験的検証を試みる．

3. 論文テキストを用いた知識グラフの構築

本節では，学術論文に記された知識を獲得し，エンティティを節点，エンティティ間の関係を有向辺で表現する知識グラフの構築技法について述べる．図 1 に処理の流れを示す．まず，論文データに対して前処理を施し，文分割された論文テキストに変換する．次に，訓練データを用いてエンティティ関係抽出モデルを学習する．論文テキストにモデルを適用し，エンティティならびにエンティティ間の関係を獲得する．

3.1 前処理

本研究では，pdf 形式の学術論文の利用を前提とする．まず，PDFNLT-1.0[*9][17] によって半構造化テキストに変

[*8] http://shachi.org

[*9] https://www.info-proto.com/2018/05/17/pdfnlt

表 1 エンティティのタイプ [16]

タイプ	説明
Task	アプリケーション，解決するべき課題，構築されたシステム
Method	方法，システムの部品，フレームワークモデル，利用されるシステムやツール
Evaluation Metric	評価指標，システムや方法の質を説明可能なエンティティ
Material	データ，データセット，リソース，コーパス，知識ベース
Other Scientific Terms	科学用語であるが上のいずれでもないもの
Generic	一般的な単語

表 2 関係のタイプ [16]

タイプ	説明
Used-for	B は A のために利用される，A は B によって訓練される，A は B に基づく
Feature-of	B は A に属する，B は A の特徴である
Hyponym-of	B は A の下位語である，B は A の 1 つである
Part-of	B は A の一部である
Compare	2 つのモデルや方法を比較する
Conjunction	似たような 2 つのエンティティを並列化する
Evaluate-for	B は A によって評価される

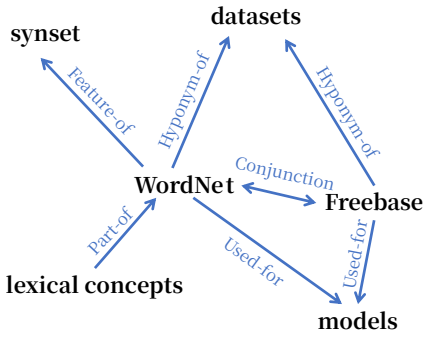

図 2 構築された知識グラフの一部

換する．すなわち，論文の構成要素（タイトル，著者名，本文，キャプション，脚注，参考文献リストなど）を保持した xhtml ファイルを生成する[*10]．xhtml ファイルには段落境界がタグ付けされる．

続いて，各段落のテキストを PySBD[18] を用いて文に分割する．PySBD はルールベースの文境界推定モジュールである．各文中には，行末にハイフンが付与されている場合があるため，dehyphen[*11]を用いて不要なハイフンを除去する．

3.2 情報抽出モデルの学習と適用

文分割された論文テキストからエンティティ関係を抽出するモデルを SciERC[16] を用いて学習する．SciERC は，人工知能分野の論文 500 件のアブストラクトから作成された，学術論文を対象とした情報抽出のためのデータセットである．SciERC のスキーマは SemEval-2017 Task10[19] および SemEval-2018 Task 7[20] 定義を拡張したものであり，文におけるエンティティの位置とタイプ（6 種類），及び，エンティティ間の関係とそのタイプ（7 種類）がアノテーションされている．SciERC[16] におけるエンティティのタイプを表 1 に，関係のタイプを表 2 にそれぞれ示す．

情報抽出モデルとして SpERT[21] を用いる．これは，エンティティとその関係を抽出するためのアテンション [22] に基づくモデルである．BERT[23] や SciBERT[24] 埋め込みを採用しており，CoNLL04 データセット [25] や ADE データセット [26] でも高い性能を達成している．SciERC データセットの学習データを用いてモデルのパラメータを訓練する．作成した抽出モデルを，前処理が施された各文に適用する．

3.3 知識グラフの構築

エンティティを節点，エンティティ間の関係を有向辺とする知識グラフを作成する．知識グラフの作成にあたり，フィルタリング処理及び同一エンティティの同定を行う．フィルタリング処理では，以下の条件を満たすエンティティを取り除く．

- 数式の一部や変数とみなせる一文字のアルファベット・ギリシャ文字・数式用英数字記号[*12]であるトークンを含む[*13]
- 括弧の対応が取れない[*14]

同一エンティティの同定では，大文字・小文字・スペース・数字を無視した文字列比較を行い，一致率の高いエンティティに対応した節点を併合する．図 2 に，本研究で構築された知識グラフの一部を示す．

[*10] 本手法では，これらのうち本文を抽出対象として用いた．
[*11] https://github.com/pd3f/dehyphen

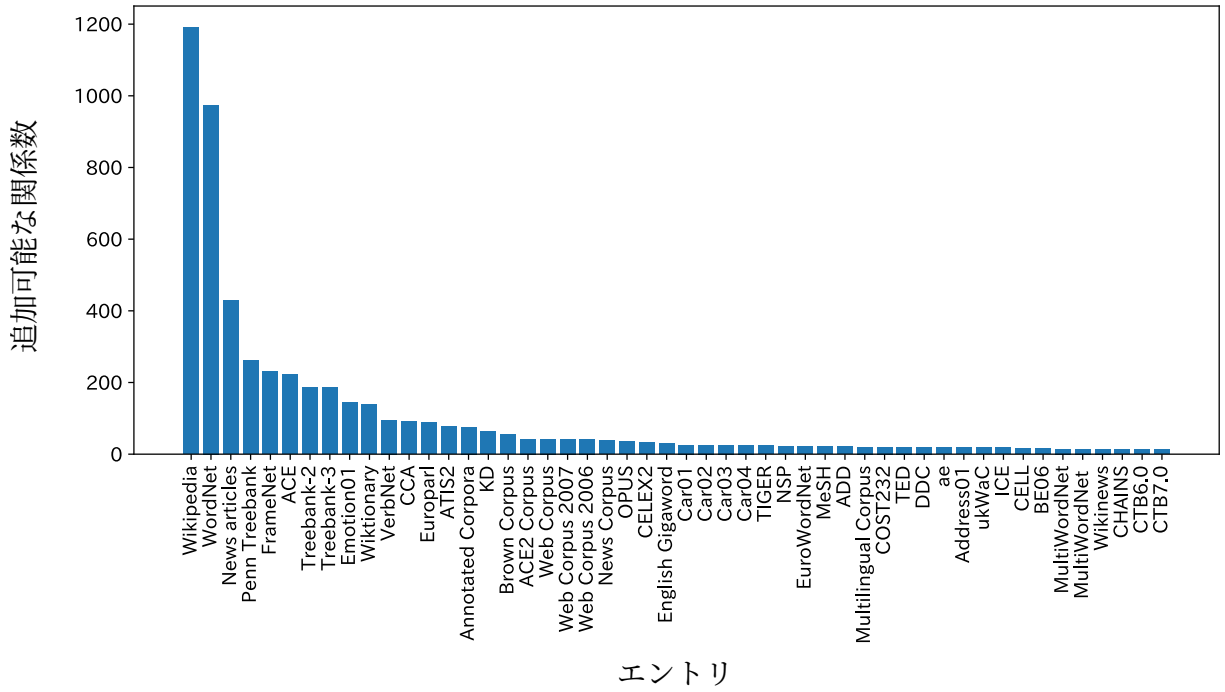

図 3 追加可能な関係数における上位 50 エントリ

4. メタデータの拡充可能性

4.1 準備

メタデータの拡充可能性を検証するために，言語処理分野の論文データを用いて知識グラフを構築した．論文データとして，2000 年〜2022 年 5 月までに開催された国際会議 ACL，NAACL，EMNLP の本会議論文 15,721 件を，ACL Anthology[*15]から収集し使用した．

PDF 形式から xhtml 形式への変換，テキストの文分割等の前処理を実行した結果，文の総数は 1,228,368 文となった．各文に対して，SciERC データセットを用いてパラメータを訓練し，得られた SpERT を用いてエンティティとその関係の抽出を行った．構築された知識グラフの規模は節点数で 763,305，有向辺数で 1,302,589 となった．節点数のうち，他の節点と有向辺で接続する節点数は 506,862 であった．

本研究では，検証に用いる既存のデータリポジトリとして，言語資源メタデータリポジトリ SHACHI[4] を用いた．SHACHI に格納されているメタデータは，Dublin-Core[*16]に準拠した語彙で記述されており [27]，これに従ってメタデータを拡充することで一貫性を保持できる．SHACHI から，エントリ名に関する記述として `title`，`title.abbreviation`，`title.alternative` の情報を収集した．対象とするエントリは 3,319 件であり，エントリ名の異なり数は 4,168 件であった．

表 3 関係と SHACHI データベーススキーマの対応

関係タイプ	プロパティ
Used-For	relation.utilization
Feature-of	description[*17]
Hyponym-of	relation.hypernym
Part-of	relation.part-of
Compare	relation.similar
Conjunction	relation.similar

4.2 検証の方法と結果

4.1 節の知識グラフにより，SHACHI に格納された既存のメタデータ情報をどの程度拡充できるかについて実験的に検証した．知識グラフ中のエンティティと同一であると判定された SHACHI のエントリに対し，エンティティに付与された関係を，それに対応したメタデータ項目の情報として追加することができれば，既存のメタデータを拡充できることが示唆される．エンティティ間の関係のタイプと SHACHI のメタデータ項目との対応を表 3 に示す．

SHACHI に登録されたエントリ名称 4,168 件と，論文から抽出したエンティティ 763,305 件との間で，大文字・小文字・スペース・数字を無視して文字列比較を行ったところ，274 件が同一であると判定された．このうち，1 つ以上の関係が付与されている，すなわち，エントリを拡充できるのは 208 件（エントリ全体の 6.3%）であり，それらに対して合計 5,845 件の追加可能な関係が存在した．追加で

[*12] ユニコード U+1D400 から U+1D7FF を対象とした．
[*13] 冠詞 “a” はエンティティには含まれない．
[*14] 左括弧のみや右括弧のみ含まれるような場合が該当する．
[*15] https://aclanthology.org
[*16] http://dublincore.org
[*17] description は通常文章で記述されているが，この情報はタグとして記述する．

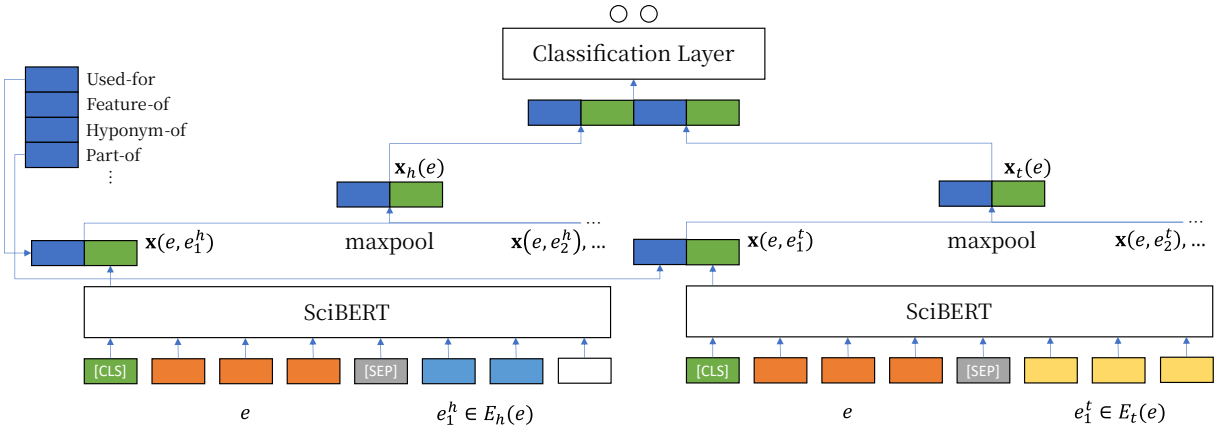

図 4 ニューラルモデルの構造

表 4 研究資源である (1) か否か (0) の判別例

エンティティ	判定	理由
bread	0	一般的な単語
mobile apps	0	一般的な単語
B-CLS	0	研究資源を指し示さない
SMS message	0	研究資源を指し示さない
reduced document	0	研究資源を指し示さない
Arabic data	0	研究資源を指し示さない
LCMC Corpus	1	研究資源
Wikipedia	1	研究資源のためのデータ収集対象

きる関係数において上位のエントリを図 3 に示す．

以上の実験により，言語処理分野の論文から獲得した知識グラフにより，SHACHI のメタデータを拡充できることが示された．

5. エントリの拡充可能性の検証

4.1 節で獲得した知識グラフ，及び，SHACHI を用いて，データリポジトリに登録されたエントリの拡充可能性を実験的に検証する．知識グラフを構成するエンティティには，研究資源の名称を示すものとそれ以外のものがあるため，研究資源を示すものを識別し，それを新たなエントリとしてリポジトリに追加する必要がある．本実験では，まず，モデルの学習に用いるデータセットを作成し，研究資源を示すエンティティを識別するためのニューラルモデルを構築した．検証では，知識グラフを構成するエンティティにモデルを適用し，研究資源名を示すと判定されたエンティティを検証した．

5.1 データセットの作成

データセットを作成するために，4.2 節の知識グラフのエンティティのうち，Material タイプである割合が 0.5 以上のものを用いた．ただし，固有名でなく一般的な言語データを含めることを避けるため，ISO639-3 言語コード[*18]に含まれる言語名と一致した場合にはこれを取り除いた．また，関係が付与されていないエンティティ，及び，1 件の論文のみに出現するエンティティは対象外とした．以下では説明の便宜上，以上の処理により得られたエンティティの集合を E で表す．E のサイズは 16,869 件であった．次に，E から 1,000 件をサンプリングし，研究資源名を示すものを人手で判別しラベル付けした．判別では，以下の全ての条件を満たすことを条件とした．

- 固有名詞である
- 研究資源，または，研究資源の構築に利用された Web サービス等を指す

エンティティが研究資源である (1) か否 (0) かの判定例を表 4 に示す．これらの作業の結果，237 件 (23.7%) が研究資源を示すエンティティであると判定された[*19]．

5.2 識別モデルの作成

研究資源を示すエンティティを識別するモデルを SciBERT[24]，および FFNN を用いて作成した．図 4 にモデルの構造を示す．

関係が付与されたエンティティ $e_1, e_2 \in E$ に対して，e_1 を構成する BPE[28] トークン列を $(t_1^1, \ldots, t_{n_1}^1)$，$e_2$ を構成する BPE トークン列を $(t_1^2, \ldots, t_{n_2}^2)$ とするとき，$(\texttt{[CLS]}, t_1^1, \ldots, t_{n_1}^1, \texttt{[SEP]}, t_1^2, \ldots, t_{n_2}^2)$ を SciBERT へ入力して得られる [CLS] に対する埋め込みベクトルを $\mathbf{c}(e_1, e_2)$ とする．また，e_1，e_2 に対する関係のタイプの埋め込みベクトルを $\mathbf{t}(e_1, e_2)$ とする．このとき，e_1 と e_2 の関係を表す埋め込みベクトルを，

$$\mathbf{x}(e_1, e_2) = [\mathbf{t}(e_1, e_2); \mathbf{c}(e_1, e_2)]$$

とする．ただし，$[\mathbf{a}; \mathbf{b}]$ は $\mathbf{a}$ と $\mathbf{b}$ の連結を表す．

知識グラフにおいて，あるエンティティ $e \in E$ に対して，e が始点である関係の終点エンティティの集合を $E_h(e)$ とし，e が終点である関係の始点エンティティの集合を $E_t(e)$ とする．

[*18] https://iso639-3.sil.org

[*19] すなわち，E 中に約 4,000 件の研究資源名を示すエンティティが含まれている可能性がある

表 5 テストデータに対する混合行列

	Negative	Positive
False	104	5
True	16	25

表 6 テストデータに対するモデルの結果

評価指標	結果
適合率	0.83
再現率	0.61
$F_{0.5}$ 値	0.78

あるエンティティ e および $\{e^h_1,\ldots,e^h_{m_h}\} \subseteq E_h(e)$, $\{e^t_1,\ldots,e^t_{m_t}\} \subseteq E_t(e)$(ただし，$m_h, m_t \leq \psi$) に対して，

$$\mathbf{x}_h(e) = \begin{cases} f\big(\mathbf{x}(e,e^h_1),\ldots,\mathbf{x}(e,e^h_{m_h})\big) & E_h(e) \neq \emptyset \\ \mathbf{0} & E_h(e) = \emptyset \end{cases},$$

$$\mathbf{x}_t(e) = \begin{cases} f\big(\mathbf{x}(e,e^t_1),\ldots,\mathbf{x}(e,e^h_{m_t})\big) & E_t(e) \neq \emptyset \\ \mathbf{0} & E_t(e) = \emptyset \end{cases}$$

とする．ここで，f は maxpool である．最終的な識別層への入力は，$\mathbf{x}(e) = [\mathbf{x}_h(e); \mathbf{x}_t(e)]$ である．

識別層は $\hat{y}(e) = \mathrm{softmax}(\mathrm{W} \cdot \mathbf{x}(\mathrm{e}) + \mathbf{b})$ であり，2 値のスコアを出力する．

5.3 検証の方法と結果

作成した識別モデルを用いて検証実験を行った．まず，識別モデルのパラメータ W, $\mathbf{b}$ の学習を行った．ラベル付けした 1,000 件のデータを，学習データ 700 件，検証データ 150 件，テストデータ 150 件に分割し，訓練には学習データと検証データを用いた．学習における損失関数として重み付きクロスエントロピー損失を利用し，重みとして各クラスのデータセットの件数の割合の逆数 $(1000/237, 1000/763)$ を採用した．また，最適化アルゴリズムとして Adam[29] を採用し，学習率は 1e-5 を基準として，1 エポックごとに 0.9 倍減衰させた．バッチサイズを 16，エポックサイズを 64，関係のタイプの埋め込みベクトルの次元数を 25，ψ を 8 に設定した．さらに，ドロップアウト率 0.1 のドロップアウト層を識別層の前に置いた．

評価指標として，テストデータに対する適合率，再現率，$F_{0.5}$，および，すべてのエンティティ集合 E のうち研究資源を示すと判定されたものからランダムサンプリングした 100 件に対する正解率を用いた．ここで，

$$F_{0.5} = 1.25\frac{precision \cdot recall}{0.25precision + recall}$$

である．メタデータリポジトリでは，登録される研究資源の網羅性よりも妥当性が重要であると考えられるため，適合率を重視した指標である $F_{0.5}$ 値を採用した．検証データにおいて $F_{0.5}$ 値が最大となるパラメータを採用している．

実験結果を表 5 および表 6 に示す．また，研究資源を示すとして識別されたエンティティ 3,514(0.21%) 件からサンプリングした 100 件を調査したところ，80 件が研究資源を示すものであった．これは，SHACHI のエントリ数 3,319 件と同規模のエントリ数が増加する可能性を示唆している．

以上の実験より，言語処理分野の論文から獲得した知識グラフにより，SHACHI のエントリを拡充できることを確認した．

6. おわりに

本論文では，データリポジトリの効率的な拡充を目的とした，研究資源のメタデータに関する情報を学術論文から抽出する試みについて述べた．具体的には，論文テキストに出現するエンティティ及びエンティティ間の関係を抽出し，エンティティを節点，関係を有向辺とする知識グラフを獲得する．言語処理分野の国際会議論文 15,721 件を用いて知識グラフを構築し，言語資源メタデータデータベース SHACHI の拡充可能性を実験的に検証した．メタデータの拡充に関する検証では，SHACHI に格納されたメタデータに対する新たな情報の追加可能性を確認した．エントリの拡充に関する検証では，知識グラフを構成するエンティティ集合から研究資源を示すものを識別するニューラルモデルを作成し，識別実験を行った．実験の結果，SHACHI における研究資源エントリの拡充可能性が示された．

謝辞 本研究は，一部，科学研究費補助金 (基盤研究 (B))(No. 21H03773) により実施したものである．

参考文献

[1] Noy, N., Burgess, M. and Brickley, D.: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem, *28th Web Conference (WebConf 2019)* (2019).

[2] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E. and Groth, P.: Dataset Search: A Survey, *The VLDB Journal*, Vol. 29, No. 1, p. 251–272 (online), DOI: 10.1007/s00778-019-00564-x (2020).

[3] Kozawa, S., Tohyama, H., Uchimoto, K. and Matsubara, S.: Collection of Usage Information for Language Resources from Academic Articles, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, European Language Resources Association (ELRA), (online), available from ⟨http://www.lrec-conf.org/proceedings/lrec2010/pdf/746_Paper.pdf⟩ (2010).

[4] Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S. and Isahara, H.: Construction of an Infrastructure for Providing Users with Suitable Language Resources, *Coling 2008: Companion volume: Posters*, Manchester, UK, Coling 2008 Organizing Committee, pp. 119–122 (online), available from ⟨https://aclanthology.org/C08-2030⟩ (2008).

[5] Ikoma, T. and Matsubara, S.: Identification of Research Data References Based on Citation Contexts, *Interna-*

tional Conference on Asian Digital Libraries, Springer, pp. 149–156 (2020).

[6] Tsunokake, M. and Matsubara, S.: Classification of URLs Citing Research Artifacts in Scholarly Documents based on Distributed Representations, *Proceedings of 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2021) collocated with ACM/IEEE Joint Conference on Digital Libraries (JCDL2021)*, Vol. 3004, pp. 20–25 (2021).

[7] Grishman, R. and Sundheim, B.: Message Understanding Conference- 6: A Brief History, *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, (online), available from ⟨https://aclanthology.org/C96-1079⟩ (1996).

[8] Singhal, A. and Srivastava, J.: Data Extract: Mining Context from the Web for Dataset Extraction, *International Journal of Machine Learning and Computing*, pp. 219–223 (2013).

[9] Heddes, J., Meerdink, P., Pieters, M. and Marx, M.: The Automatic Detection of Dataset Names in Scientific Articles, *Data*, Vol. 6, No. 8 (online), DOI: 10.3390/data6080084 (2021).

[10] Prasad, A., Si, C. and Kan, M.-Y.: Dataset Mention Extraction and Classification, *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 31–36 (online), DOI: 10.18653/v1/W19-2604 (2019).

[11] Ikeda, D., Nagamizo, K. and Taniguchi, Y.: Automatic Identification of Dataset Names in Scholarly Articles of Various Disciplines, *International Journal of Institutional Research and Management*, Vol. 4, No. 1, pp. 17–30 (2020).

[12] Li, K. and Yan, E.: Co-mention network of R packages: Scientific impact and clustering structure, *Journal of Informetrics*, Vol. 12, No. 1, pp. 87–100 (online), DOI: https://doi.org/10.1016/j.joi.2017.12.001 (2018).

[13] Du, C., Cohoon, J., Lopez, P. and Howison, J.: Softcite dataset: A dataset of software mentions in biomedical and economic research publications, *Journal of the Association for Information Science & Technology*, Vol. 72, No. 7, pp. 870–884 (online), DOI: 10.1002/asi.24454 (2021).

[14] Du, C., Howison, J. and Lopez, P.: Softcite: Automatic extraction of software mentions in research literature, (online), available from ⟨https://scinlp.org/history/2020/pdfs/softcite-automatic-extraction-of-software-mentions-in-researchliterature.pdf⟩ (2020).

[15] Kozawa, S., Tohyama, H., Uchimoto, K. and Matsubara, S.: Automatic Acquisition of Usage Information for Language Resources, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, European Language Resources Association (ELRA), (online), available from ⟨http://www.lrec-conf.org/proceedings/lrec2008/pdf/169_paper.pdf⟩ (2008).

[16] Luan, Y., He, L., Ostendorf, M. and Hajishirzi, H.: Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 3219–3232 (online), DOI: 10.18653/v1/D18-1360 (2018).

[17] Abekawa, T. and Aizawa, A.: SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan, The COLING 2016 Organizing Committee, pp. 136–140 (online), available from ⟨https://aclanthology.org/C16-2029⟩ (2016).

[18] Sadvilkar, N. and Neumann, M.: PySBD: Pragmatic Sentence Boundary Disambiguation, *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, Online, Association for Computational Linguistics, pp. 110–114 (online), DOI: 10.18653/v1/2020.nlposs-1.15 (2020).

[19] Augenstein, I., Das, M., Riedel, S., Vikraman, L. and McCallum, A.: SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Association for Computational Linguistics, pp. 546–555 (online), DOI: 10.18653/v1/S17-2091 (2017).

[20] Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H. and Charnois, T.: SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers, *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 679–688 (online), DOI: 10.18653/v1/S18-1111 (2018).

[21] Eberts, M. and Ulges, A.: Span-based joint entity and relation extraction with transformer pre-training (2020).

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is All You Need, NIPS'17, Red Hook, NY, USA, Curran Associates Inc., p. 6000–6010 (2017).

[23] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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)*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[24] Beltagy, I., Lo, K. and Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 3615–3620 (online), DOI: 10.18653/v1/D19-1371 (2019).

[25] Roth, D. and Yih, W.-t.: A Linear Programming Formulation for Global Inference in Natural Language Tasks, *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, Boston, Massachusetts, USA, Association for Computational Linguistics, pp. 1–8 (online), available from ⟨https://aclanthology.org/W04-2401⟩ (2004).

[26] Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M. and Toldo, L.: Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-related Adverse Effects from Medical Case Reports, *Journal of Biomedical Informatics* (2012).

[27] Tohyama, H., Kozawa, S., Uchimoto, K., Shigeki, M. and Hitoshi, I.: SHACHI: A Large Scale Metadata Database of Language Resources, *Proceedings of the First Inter-*

national Conference on Global Interoperabikity for Language resources (ICGL-2008), pp. 205–212 (2008).

[28] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (online), DOI: 10.18653/v1/P16-1162 (2016).

[29] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *International Conference on Learning Representations (ICLR)* (2015).