技術記事投稿サイトを対象とした企業技術情報の抽出・構造化手法の提案

A Proposal for Extracting and Structuring Corporate Technical Information from Technical Article Websites

柿本　侑毅[*1]，太田　光一[*1]，長谷川　忍[*1]
Yuuki KAKIMOTO[*1]，Koichi OTA[*1]，Shinobu HASEGAWA[*1]
[*1] 北陸先端科学技術大学院大学
[*1]Japan Advanced of Science and Technology
Email: y-kakimoto@jaist.ac.jp

あらまし: 企業は技術的な情報の開示手段として技術記事投稿サイトや自社ブログを使用していることが多い．一方で，就職活動で志望する会社が公開している全ての技術記事を個人が調べ，全体像を把握することは容易ではない．本稿では技術記事に含まれる企業内の技術について言及のある「企業技術情報」の抽出，また，その可視化を目的とした RDF の構築を行った．
キーワード：技術記事，SVM，RDF，半教師あり学習

1. はじめに

インターネットを用いた学習が一般的になってきたが，インターネット上の情報量は増え続け，広告誘導を目的とした企業記事や，企業が運営する SEO の高いサイトが存在するため，間違った情報が検索上位になるなどが問題視されている．

同様に就職活動の場面でも企業が求めるスキルが多様化しており，学生が志望する企業に合わせて企業が使用している技術を調査するのは非常に困難である．

一方で最近では企業内部のエンジニアが情報発信の手段として利用しているサイトに社内ブログや技術記事投稿サイトが挙げられる．

IT 人材白書 2020[(1)]によると情報サービス業全体の労働人口は 157 万人であるが，本研究で情報抽出の対象とする Qiita では 2021 年 9 月の段階で会員数が 70 万人を超えており，日本最大級のエンジニアコミュニティと言える．

Qiita には Organization という機能があり，サイト内の説明[(2)]には「簡易技術ブログや採用活動の一環として」利用されているという記述がある．このことから Qiita の記事によっては企業に関する情報が抽出可能であると分かる．

しかしながら Organization の機能は登録されたユーザの記事を Organization ごとにまとめるが，そこに記述された内容が企業に関連する記事である保証をしているわけではない．そのため企業によっては登録されている記事を全て調査する必要が出てくる．

本研究では Qiita の記事に含まれる企業に関連する技術についての情報を「企業技術情報」と定義し，記事から企業技術情報を抽出するために半教師あり学習を実施した．

また，得た情報について概観を理解しやすいよう単純な構造の RDF（Resource Description Framework）として構築し，その可視化を行った．

2. 記事本文の抽出

本研究では企業技術情報の抽出について対象とするデータとして 2015 年から 2020 年までの Qiita Advent Calendar の記事を使用した．

収集した HTML データについて本研究では Web ページからの本文抽出タスクにおける SoTA(State of The Art)である Boiler Net と条件に合致しない文を排除するルールベースの 2 つの手法での抽出を行った．

記事の本文について人間の目で見て抽出したデータを用意し，その文章内でどの程度単語と一致するかを単語一致度として以下のような式で計算を行った．

$$\text{単語一致度} = \frac{\sum_{i=1}^{N} min\left(1, \frac{\text{各手法での単語数}}{\text{真の単語数}}\right)}{\text{真の語彙数} N}$$

単語一致度の比較の結果，ルールベースの手法が Boiler Net に比べて高い単語一致度を示した．

3. 入力データの生成

抽出した本文データには半教師あり学習への入力データとして特徴データ系列への変換を行うためいくつかの方法を検討した．

Bag of words では入力データを生成した場合，記事全体の語彙数が次元数になるため次元数が膨大になり処理の実行が困難であった．

次に，文書全体を数百次元程度のベクトルで表現するために，単語の分散表現を検討した．

この分散表現の代表的な手法として BERT[(3)]があるが一般に公開されているモデルでは文頭と文末の token を合わせた 512token 以上の文章は扱えず，その範囲に収まる文章を選別した場合，文章全体の 4%程しか使用できなくなる．

結果として本研究では手作業で分類した企業記事について，企業カテゴリ記事と企業でないカテゴリの記事という 2 つの文書として，TF-IDF 値を計算し

た．これにより企業カテゴリを代表するような語彙と企業に該当しない要素を示す語彙が得られる．本研究では半教師あり学習において複数の手法と比較するために全ての実装で実行可能であった次元数として TF-IDF 値の上位 500 単語を用いて，文書ごとに求めた単語頻度を入力データとして使用した．

4. 半教師あり学習

本研究では Label Spreading 法，Support Vector Machine(SVM), 混合ガウスモデル(GMM)の 3 つの半教師あり学習の手法について 5 分割交差検証を用いた比較を行った．それぞれの半教師あり学習での結果に加えて，人手でラベルを付け直した結果についても以下に示す．

表 1 Label Spreading 法の結果

	1 分割	2 分割	3 分割	4 分割	5 分割	平均
正答率	0.32	0.42	0.39	0.50	0.50	0.43
適合率	#DIV/0!	0.42	0.39	0.50	0.50	#DIV/0!
再現率	0.00	1.00	1.00	1.00	1.00	0.80
F値(f1)	#DIV/0!	0.59	0.56	0.67	0.67	#DIV/0!

表 2 SVM の結果

	1 分割	2 分割	3 分割	4 分割	5 分割	平均
正答率	0.63	0.74	0.78	0.72	0.67	0.71
適合率	0.88	0.67	0.67	0.83	0.64	0.74
再現率	0.54	0.75	0.86	0.56	0.78	0.70
F値(f1)	0.67	0.71	0.75	0.67	0.70	0.70

表 3 GMM の結果

	1 分割	2 分割	3 分割	4 分割	5 分割	平均
正答率	0.68	0.42	0.39	0.50	0.50	0.50
適合率	0.68	0.42	0.39	0.50	0.50	0.50
再現率	1.00	1.00	1.00	1.00	1.00	1.00
F値(f1)	0.81	0.59	0.56	0.67	0.67	0.66

表 4 人手による再分類の結果

	結果
正答率	0.80
適合率	0.82
再現率	0.78
F値(f1)	0.80

5. RDF の生成

RDF は XML で作られた主語，述語，目的語の 3 つの要素で概念の関係情報を表現する言語である．この 3 要素を合わせてトリプルと呼ぶ．

本研究では Cabocha[4]を用いて係り受け解析を行い，係り受け関係が存在しないものを述語句として，述語句にリンクが有り係助詞を含むものを主語，格助詞を含むものを目的語として図 1 に示す形で RDF を生成した．

```
<rdf:RDF xmlns:schema="http://schema.org/" xm
  <rdf:Description rdf:about="主語">
    <s:述語>目的語</s:述語>
  </rdf:Description>
  <rdf:Description rdf:about="主語2">
    <s:述語2-1>目的語2-1</s:述語2-1>
    <s:述語2-2>目的語2-2</s:述語2-2>
  </rdf:Description>
</rdf:RDF>
```

図 1 生成した RDF の構造

6. RDF についての考察

半教師あり学習で収集したデータに対して 5 節で述べた手法を用いて RDF を生成した．図 2 は生成した RDF に対して特定の言葉を検索したクエリの可視化結果である．この図 2 からエンハンスド LB という単語が，その意味を知らなくても暗号や乱数に関係のある言葉であることが分かる．

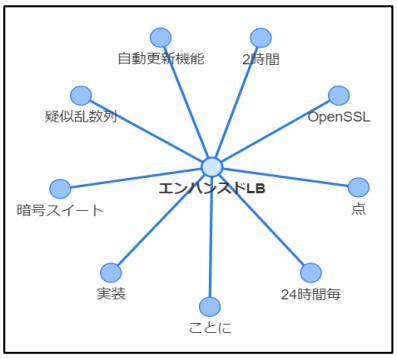

図 2　クエリの可視化結果

7. まとめ

本研究では Qiita Advent Calendar の技術記事を対象とした企業に関連する技術情報の抽出，技術と企業間の関係を RDF として生成し可視化を行った．

その結果，未知の言葉に出会った時にそれがどのような技術に関連するかについて技術概観理解の助けになる可能性が示唆された．

また，半教師あり学習によって生成したラベルデータは人手には及ばないものの良い精度である．しかしながら 500 単語にしている点で情報が抜け落ちている可能性は否めない．加えて，RDF に関しても技術に関係ない一般名詞が存在している．そのため，今後は RDF として採用する一般名詞の取捨選択や，入力データ系列の適切な次元数などについても検討していきたい．

参考文献

(1) 独立行政法人情報処理推進機構 社会基盤センター編．「IT 人材白書 2020」: 59-60.
(2) Qiita Organization とは https://help.qiita.com/ja/articles/qiita-org-1
(3) Devlin, Jacob, et al. "Bert: Pre-training of deepbidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
(4) CaboCha/ 南 瓜 : Yet Another Japanese Dependency Structure Analyzer https://taku910.github.io/cabocha/