1C-6

Web 上の文書を対象とした産学連携研究開発情報抽出の試み

蔵川　圭†　　孫　媛†　　西澤　正己†　　柿沼　澄男†　　相澤　彰子†

国立情報学研究所†

1.はじめに

産学連携の実態を把握するための情報源として，Web 上に公開されている情報は無視できないものとなっている．大学や企業のプレスリリースや研究室のホームページ，新聞記事などに掲載される多種多様な情報の一つとして，産学連携研究開発についての報告が多くみられる．本研究は Web 上にある文書を対象として，産学連携研究開発に関する情報を自動的に抽出する技術の確立を目的とする．

目的を達成するためには，Web 上にある URL が指し示す文書に対して，ある学術機関と企業が産学連携したことを示す根拠を自動で抽出できればよい．抽出できれば，その文書が指し示す URL は産学連携研究開発関連であると判定できる．

産学連携研究開発関連の文書の特徴は，多くの場合集中的に 1 文または前後の 2−3 文にあらわれる．それは，「○○大学は，株式会社○○と○○に関する共同研究を開始しました．」というように表現されることが多い．そのほかの文章は，研究内容の紹介と区別がつかないほどである．

素朴に考えれば，学術機関名と企業名，および産学連携のキーワードが文書に存在すれば，それは産学連携研究開発関連であると考えることができる．キーワードとは，たとえば，「共同」，「協同」，「研究」，「産学」，「連携」，「開発」などである．そこで，ある大学のホームページに限定して文書をクローリングし，大学名 1 件と企業名 43542 件の辞書，および産学連携キーワードリスト 13 件を用いて，文書群に対して文字列マッチングを試みた．その結果，まったく関係のない文字列にヒットして，産学連携研究開発関連ではない文書が多数抽出された．特に，企業名には一般名詞と同じ表記も多くあり，それが文脈を考慮してはじめて企業名であることが判断つくことが多かった．また，キーワードに関しても，たとえば「研究」や「開発」というだけでは，産学連携とは限らないことが多かった．

このことから，産学連携研究開発関連の文書を判定するときには，一文に注目し，形態素を考慮して企業名などの固有名詞を抽出できる必要があることがわかった．

A Trial on Extracting Academia-Industry Collaborative Research and Development Documents on the Web
†National Institute of Informatics

本報では，文書の一文に注目して，大学名や企業名の固有表現と，キーワードを抽出する方法を試行したので，その結果を記す．

2.産学連携文書の固有表現とキーワードの抽出

2.1. 方法

産学連携を意味する特徴的な文に対して，大学機関名と企業名，産学連携キーワードを機械学習によって抽出することを考える．そのために，以下のアプローチをとった．

1. 大学の Web ページにある産学連携関連の文書を検索・発見し，学習に必要な正解文を人手でコピーペーストしてそろえる．
2. 正解文の集合を形態素解析器 MeCab[1]にかけ，形態素と品詞分類を得る．
3. 得られた形態素に対して，大学名，企業名，その他産学連携に関係のある組織名，産学連携キーワードをカテゴリわけし，IOB2 タギングを行う．
4. IOB2 タギングした形態素と品詞分類の文集合に対し，機械学習器 CRF++[2][3]を用いて学習を行う．
5. 学習した結果を用いて，未分類の産学連携関連文書の正解文集合に対して，学習器 CRF++が自動で IOB2 タギングを行う．
6. 自動で付与した IOB2 タギングの精度と再現度，F 値を算出して，性能を測る．

2.2. 実験

まず，産学連携関連の文書を，東京大学の Web ページから人出で検索・抽出し，キーセンテンスを抜き出して正解文の集合を作った．たとえば，「国立大学法人東京大学（総長：小宮山 宏、以下東京大学）とサン・マイクロシステムズ株式会社（本社：東京都世田谷区、代表取締役社長：ダン・ミラー、以下サン）は、両者の IT 分野における基礎研究技術を相互に提供し、新しい技術・価値の創造や産学連携モデル確立を目指す組織的な共同研究に係る協定書を締結しました。」というのが一文である．このような文を 104 文抽出して正解文集合を用意した．

104 正解文集合に対して，形態素解析器 MeCab0.98 にかける．このとき，会社名が形態素の固有名詞と一致させて，単語境界のずれがなくなるようにユーザー辞書として会社名リスト 43542 件を登録した．登録の連接 ID とコスト，品詞は以下のように設定した．

```
いずみ,1292,1292,6849,名詞,固有名詞,組織,*,*,*,いずみ,*,*
丸一産業,1292,1292,6849,名詞,固有名詞,組織,*,*,*,丸一産業,*,*
明和産業,1292,1292,6849,名詞,固有名詞,組織,*,*,*,明和産業,*,*
. . . .
```

図 1 MeCab に会社名リストの追加

形態素解析の結果，形態素（トークン）と POS（品詞分類）のタプルを得る．これによって，6984 形態素を得た．

つづけて，形態素に IOB2 フォーマットでチャンキングを行う．ここでは，カテゴリは 5 個で，それぞれに対して B(Begin)と I(Inside)の 2 種に区分けし，その他は O(Outside)で，11 個のタグを設定した．

表 1 産学連携関連要素チャンキングのためのカテゴリ

カテゴリ	説明（例）
ACADEMIA	アカデミア（東京大学）
INDUSTRY	産業界（住友金属工業 株式会社）
GOVERNMENT	官界（文部 科学 省）
FUNDINGAGENCY	助成団体（財団 法人 メトロ 文化 財団）
COLLABORATION	産学連携（共同 作業 で ある）

チャンキングした結果は，次に示すようである．

```
2009 名詞-数-*-*-* O
年 名詞-接尾-助数詞-*-* O
3 名詞-数-*-*-* O
月 名詞-一般-*-*-* O
まで 助詞-副助詞-*-*-* O
の 助詞-連体化-*-*-* O
3 名詞-数-*-*-* O
年間 名詞-接尾-助数詞-*-* O
に 助詞-格助詞-一般-*-* O
亘り 動詞-自立-*-*-五段・ラ行 O
、 記号-読点-*-*-* O
みずほ情報総研名詞-固有名詞-組織-*-* B-INDUSTRY
㈱ 名詞-サ変接続-*-*-* I-INDUSTRY
と 助詞-並立助詞-*-*-* O
東京 名詞-固有名詞-地域-一般-* B-ACADEMIA
大学 名詞-一般-*-*-* I-ACADEMIA
で 助詞-格助詞-一般-*-* O
共同 名詞-サ変接続-*-*-* B-COLLABORATION
研究 名詞-サ変接続-*-*-* I-COLLABORATION
を 助詞-格助詞-一般-*-* I-COLLABORATION
行う 動詞-自立-*-*-五段・ワ行促音便 I-COLLABORATION
。 記号-句点-*-*-* O
```

図 2 チャンキングの例

チャンキングした結果を学習器にかけ，別の形態素解析した正解文にチャンキングのテストを試みる．ここでは，10 分割交差確認を行うこととし，先ほどのチャンキングしたデータに対し，文を単位に 10 分割し，9 割を訓練データ，1 割をテストデータとして，10 セット作った．学習器は CRF++ 0.54 を用いて，“-f 3 -C 4.0”とオプション指定し，素性のテンプレートは 3 カラム用のものを用いて，10 セットに対して訓練データの学習およびテストをした．

10 セットのテスト結果のそれぞれ，およびその全体に対し，精度，再現度，F 値を算出して，平均は以下のようになった．accuracy: 89.58%; precision: 79.35%; recall: 65.22%; FB1: 71.59．また，カテゴリごとについては表のようになった．

表 2 カテゴリごとのテスト結果

カテゴリ	Precision	Recall	F 値(β=1)
ACADEMIA	84.06%	76.32%	80.00
INDUSTRY	88.89%	67.88%	76.98
FUNDINGAGENCY	0.00%	0.00%	0.00
GOVERNMENT	0.00%	0.00%	0.00
COLLABORATION	65.41%	54.38%	59.39

2.3. 考察

産学連携関連文書を読んだときの特徴は，正解文書のように 2−3 文（句点から句点までを一文）に現れる．今回はこれをひとまとまりとして学習器にかけた．その結果，チャンキングの問題としては 71.59 の F 値を得た．

カテゴリについては IOB2 チャンキングの結果として，FUNDINGAGENCY と GOVERNMENT はゼロである．実際の正解文を眺めると ACADEMIA と INDUSTRY，COLLABORATION で占め，FUNDINGAGENCY や GOVERNMENT は事例がほとんどなかった． 多くの文では，ACADEMIA と INDUSTRY，COLLABORATION のセットとして現れている．

同程度出現している COLLABORATION は，ACADEMIA や INDUSTRY の固有表現とは意味の性質が異なる．学習の結果も思わしくない．「共同研究する」のような「共同研究」というキーワードが「する」というサ変動詞と直結するため，係り受け解析を適用するほうが効果的に抽出できる可能性がある．

3.おわりに

Web 上にある文書を対象として，産学連携研究開発に関する情報を自動的に抽出する技術の確立を目的とし，本報では，文書の一文に注目して，大学名や企業名の固有表現と，産学連携キーワードを，形態素解析器 MeCab と機械学習器 CRF++を用いて抽出する方法を試みた．その結果，固有表現に関して，71.59 の F 値を得た．キーワードの抽出については性能が比較的劣っていたので，今後の展開として，これを改善するためにキーワードの性質を考慮して係り受け解析を行うことが考えられる．

参考文献

[1] MeCab, http://mecab.sourceforge.net/
[2] CRF++, http://crfpp.sourceforge.net/
[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282-289, 2001