6T-07

Wikipediaからの技術やサービス間の関係抽出に用いるフィルタリングの改良

南川　大樹†　　杉本　徹‡

芝浦工業大学大学院　理工学研究科†　　芝浦工業大学　工学部‡

1. 研究の背景と目的

初学者がこれから新しい分野を学習する際、学習意欲を向上させるために、その分野がどのように役立っているかを知る必要がある。また、初学者はカーナビゲーションなどといった身近なサービスを実現するために、どのような技術を学ぶ必要があるか分からない。

これらを解決するために、Wikipedia から「技術とその技術を用いて実現できる応用例という関係」を表す使用関係(例：「形態素解析→かな漢字変換」)を抽出する方法を提案した[1]。この関係の抽出過程で使用したフィルタリング処理では、使用関係候補ペアの各項が技術や応用例を表す用語かどうかを判定する。しかし、人物名や年号が技術名として判断されてしまうことやペアの各項の間の関係性を考慮していない等の問題があった。

本研究では、このフィルタリングをルールに基づいて不適切な候補ペア(ノイズ)を除去する段階と SVM を用いてノイズを除去する段階の 2 段階で行う。また、学習に用いる訓練データへのラベルの付与の負担を減らすため、能動学習を取り入れる。そして、実際にこの手法で使用関係を抽出し、その妥当性を評価する。

2. 提案手法

2.1. 全体の流れ

全体の流れを図 1 に示す。まず、Wikipedia の本文からパターンマッチを用いて使用関係の候補となるペアを抽出する(2.2 節)。次に、抽出した使用関係候補ペアに対して、フィルタリングを行い、ノイズを除去する(2.3 節)。

2.2. 使用関係候補ペアの抽出

2.2.1. 本文からの使用関係候補ペアの抽出

まず、Wikipedia の記事本文の係り受け解析を行う。次に、各係り元文節と係り先文節のペアの形態素列に対してパターンマッチを行い、抽出したペアを使用関係候補ペアとする。また、ペアの片方の項が欠けていた場合は、その項を記事名、見出し名等で補完する。

2.2.2. 箇条書きからの使用関係候補の抽出

次に、直後に箇条書きで項目が列挙されやすい手がかり語(例：「以下」、「示す」、「次」)を含む文を探し、その文中に技術(または応用例)を表す用語があれば、その文の下に箇条書きで列挙される項目をその記事に関連する技術(または応用例)として抽出する。また、箇条書きでは片方の項しか抽出できないため、もう片方の項を補完する必要がある。

Improvement of the filtering method for the extraction of relations between technologies and services from Wikipedia.
† Hiroki Minamikawa, ‡Toru Sugimoto
† Graduate School of Engineering and Science, Shibaura Institute of Technology
‡ College of Engineering, Shibaura Institute of Technology

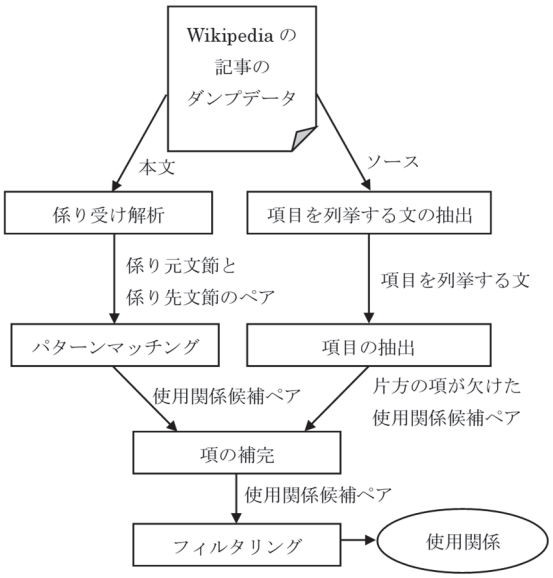

図 1　全体の流れ

2.3. フィルタリングによるノイズの除去

2.3.1. ルールに基づくノイズの除去

まず、2.2 節の手法により抽出した使用関係候補ペアから、以下の条件(1)~(3)のいずれかに該当するペアを除去する。

(1) どちらか一方の項が人物名や組織名を表す。
(2) どちらか一方の項が「<数字> + “年”」を含む、またはストップワード(図 2)である。
(3) 2 つの項のリダイレクト先(無い場合はその項自身)が同じ(同義語関係)。

条件(1)における、「人物名、組織名を表す」とは、その項と同名、またはリダイレクト先の記事 A が存在し、その記事 A の最初の 1 文(例：「A とは、B である。」)の最後の名詞(例における B)、またはその記事 A が属するカテゴリのいずれかが人物・組織キーワード(図 3)と後方一致していることとする。条件(2)のストップワードは出現頻度が高い項の中から、明らかに不適切であるものを選出した。

問題, 機能, インターネット上, 無料, 研究史, 事, その他, 概要, 技術, 手法, 形, 世界, 実装, 特徴, 日本語, 方法, 利用, 外部リンク, 歴史, 登場人物

図 2 ストップワードの一覧

人物, 人, 者, 家, 員, 組織, 所, 会, 大学, 企業, 会社, 団

図 3 人物・組織キーワードの一覧

2.3.2. SVM を用いたノイズの除去

次に、前節の方法で除去できなかったノイズを、SVM を用いて除去する。この段階では、ある使用関係候補ペア「項 X→項 Y」に対して、「項 X が使用関係の技術の項として適しているか」と「項 Y が使用関係の応用例の項として適しているか」を、2 つの独立した分類器で各項に対して分類を行い、両方の項が適切であると判断された場合、そのペアを正しい使用関係として出力する。

本研究では素性として、項の分散表現、関係の分散表現(項 Y と項 X の分散表現の差)、JUMAN により付与された形態素のドメインやカテゴリ、表層格等の 17 種類を用いた。分散表現は Word2Vec[2](Skip-gram/200 次元/前後 5 単語)で学習したものである。また、RFE[3]を用いて特徴選択を行った。

2.3.3. 能動学習

能動学習とは、大量のラベルなしデータ(プール)から、ラベルを付けると分類器の性能が向上する(推測されるラベルが不確実な)データを機械的に選んで、その選んだデータに人手でラベルをつけて再学習を繰り返す手法である。これにより、ランダムサンプリングに比べて少ない訓練データで分類器の性能の向上が期待される。今回は、SVM の分離超平面との距離が近いデータに対して、優先的にラベルを付与する[4]。

3. 評価実験

本研究では、カテゴリ「人工知能」とその子孫カテゴリ(ただし、関係ない子孫カテゴリと関係ない記事を多く含む子孫カテゴリを人手で除いた)に属する記事 1,024 件から関係を抽出した。2.2 節の手法によって抽出された使用関係候補ペアは 38,580 件となり、うち 7,281 件のペアが 2.3.1 節の手法によりノイズとして除去された。

3.1. ラベル付け

まず、2.3.1 節のフィルタリングで残った使用関係候補ペア 31,299 件をプールとし、そのプールの中からランダムに各項 1000 件ずつ「項 X(Y)が使用関係の技術(応用例)の項として適切かどうか」のラベルを付ける。このラベル付きデータをそれぞれ DX1 と DY1 とする。次に、DX1 と DY1 を用いて RFE による特徴選択を行う。そして、選ばれた素性を用いて能動学習を行い、各項 1000 件ずつラベルを付ける。このラベル付きデータをそれぞれ DX2 と DY2 とする。DX1、DY1、DX2、DY2 のそれぞれにおける正例の件数は 55 件、58 件、112 件、105 件となり、能動学習ではランダムサンプリングに比べて正例ラベルが多くなる傾向が見られた。

3.2. 使用関係の抽出結果の評価

使用関係候補ペア 31,299 件に対して、DX2 と DY2 を訓練データとして学習した分類器を用いて、各項が「使用関係の技術(応用例)の項として適切かどうか」を判定した。その結果、両方の項が適切であると判定され、正しい使用関係として出力されたペアの件数は 912 件となった。この 912 件のうち重複を除いた 747 件の中からランダムに選んだ 100 件のペアに対して、6 名の被験者に使用関係として適切かどうかを評価してもらった。正解率を「(正しいと評価したペアの総数) / (評価したペアの総数)」で算出した結果、約 64.2%となった(表 1)。従来手法[1]に比べ、正解率が向上したが、抽出できたペアの数は少なくなった。これは、使用関係として正しいと判断する条件を厳しくしたことに起因すると推測される。

実験で使用したペアとそのペアに対する被験者による正誤判断の例を表 2 に示す。欠けた項に対して使用関係としての関係性が薄いものを項として補完してしまうケース(P6 の「REST 対 RPC」は見出し名から補完)や、文中で技術としての使用が否定されていても抽出されてしまうケース(P7 は本文「浮動小数点数を使うと思われるが、(中略)アンダーフローを生じる危険性がある。」から抽出)など抽出処理に関連する問題が見られた。

表 1 提案手法と従来手法の評価結果の比較

	提案手法	従来手法
正解率(%)	64.2	40.9
出力の件数	912	7,841

表 2 抽出した使用関係の例

P1	自然言語処理 → Siri	正
P2	機械学習 → 情報検索	正
P3	ラグランジュの未定乗数法 → サポートベクターマシン	正
P4	ニューラルネットワーク → 文書分類	正
P5	誤差逆伝播法 → 多層パーセプトロン	正
P6	REST → REST 対 RPC	誤
P7	浮動小数点数 → ビタビアルゴリズム	誤

4. まとめと今後の展望

本研究では、Wikipedia から使用関係の抽出を行った。抽出された使用関係の正解率は 64.2%となり、従来手法よりも改善したものの、抽出できたペアの数が 912 件と少なくなった。また、能動学習では、ランダムサンプリングに比べて正例をより多くラベル付けすることができた。

今後はフィルタリングによって除去されてしまった使用関係として正しいペアを調査し、さらなる改善をしていきたい。

参考文献

[1] 南川大樹, 杉本徹: “Wikipedia からの技術やサービス間の関係抽出”, 情報処理学会第 80 回全国大会講演論文集, pp. 299-300 (2018).

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.: “Distributed representations of words and phrases and their compositionality”, Advances in neural information processing systems, pp.3111-3119 (2013).

[3] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V.: “Gene selection for cancer classification using support vector machines”, Machine learning, Vol.46, No.1-3, pp.389-422 (2002).

[4] Schohn, G., & Cohn, D.: “Less is more: Active learning with support vector machines”, Proc. 17th International Conference on Machine Learning, pp.839–846 (2000).