6P-04

Wikipediaからの技術やサービス間の関係抽出

南川　大樹　　杉本　徹
芝浦工業大学 工学部

１．研究の背景と目的

初学者がこれから学ぶ新しい分野に対する学習意欲を向上するために、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。

2 つ目の問題を解決する先行研究として、ユーザが入力した身近な情報サービスや技術名から、Wikipedia を用いて情報工学分野で学習すべき内容を求めてユーザに提示する研究[1]がある。この先行研究では、サービスと提示された学習内容、それを学習することによって習得できる技術の関係が把握しづらいという問題点がある。

本研究では、Wikipedia から技術とその技術によって実現できる技術・サービスのペアを抽出し、図１のような「技術→技術・サービス」の辺で構成されたグラフを自動生成することを目的とする。

生成されたグラフを用いることで、実現したい技術やサービスから実現するために履修すべき科目を、または履修する科目からその科目で学習した内容を応用して実現できる技術やサービスを、提示することができると期待される。

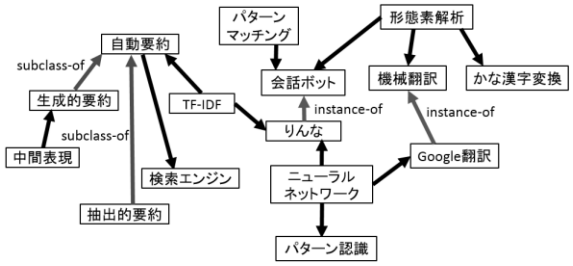

図１　生成されるグラフのイメージ

２．使用関係の抽出

２．１　ノードと使用関係の定義

まず、グラフのノードを情報分野の技術やサービスを表す名詞句と定義する(例：「ダイクストラ法」「LINE」)。次に、「技術とその技術を用いて実現できるものの関係」を使用関係と定義する(例：「形態素解析 → かな漢字変換」)。この使用関係はグラフのノード間の主要な関係(辺)である。

Extraction of relations between technologies and services from Wikipedia.
Hiroki Minamikawa, Toru Sugimoto College of Engineering, Shibaura Institute of Technology

２．２　本文からの使用関係候補の抽出

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う。人手で作成したパターン(表 1)を用いて同一文中から名詞句 X と名詞句 Y が抽出できた場合、使用関係候補のペア「名詞句 X→名詞句 Y」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。

表１　使用関係候補のペアの抽出パターンの例

係り元文節	係り先文節
名詞句 X +「を」	<基本形が「用いる」>
名詞句 X +「による」	名詞句 Y

２．３　箇条書きからの使用関係候補の抽出

前節で述べた方法では、箇条書きで列挙される技術や応用例を抽出することができない。そこで、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技術(または応用例)として抽出する。

２．４　使用関係候補ペアのフィルタリング

2.2 節、2.3 節で抽出した使用関係候補のペアには不適切なものが多く含まれるため、ペアの各項に対して、技術、応用例、サービスを表す名詞句として適切かどうかを SVM で判定し、両方の項が使用関係の項として適切であれば、使用関係として出力する。

本研究では、SVM の学習の素性として、「記事名として存在するか」、「特徴的な単語が含まれているか(各名詞の tf-idf の最大値)」、「品詞が名詞である形態素の割合」等の 20 個を用いる。

３　instance-of、subclass-of 関係の抽出

３．１　instance-of、subclass-of 関係の定義

本研究では、使用関係の他に、グラフのノード間の関係として instance-of 関係と subclass-of 関係も補助的に扱う。instance-of 関係を技術とその実例の関係(例：「りんな instance-of 会話ボット」)、subclass-of 関係を技術とその技術のう

ちの一つという関係(例：「抽出的要約 subclass-of 自動要約」)と定義する。

３．２　上位下位関係抽出ツール

「instance-of 関係と subclass-of 関係は上位下位関係を細分化したものである」という考えのもと、instance-of 関係と subclass-of 関係の抽出を行う。まず、隅田らの手法[2]にもとづく上位下位関係抽出ツールを用いて、Wikipedia から詳細な上位下位関係[3]を抽出した。詳細な上位下位関係とは、上位概念と下位概念の間に、記事名で補完した T-上位概念と、記事名の上位概念で補完した G-上位概念を追加した関係であり、「対象－属性－属性値」として解釈できる。

３．３　instance-of、subclass-of 関係の抽出

抽出した詳細な上位下位関係のうち、T-上位概念の最後の形態素の表層形 X、または表層形 X の上位語が表 2 のいずれかの場合、「下位概念 instance-of(subclass-of) 記事名」として抽出する。上位語は日本語 WordNet を用いて取得する。

表 2　T-上位概念の手がかり語

関係	手がかり語
instance-of	ソフトウェア, システム, 創造物
subclass-of	技法, 手法, 手段, 技術, ツール, 分野

４　グラフの生成

2 節、3 節で抽出した使用関係、instance-of 関係、subclass-of 関係からグラフを生成する。グラフのノードを追加する際、表記揺れに対応するため、半角スペースとアンダースコアを除いたものを追加する。

５　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、関係ない記事を多く含む可能性があるカテゴリとその子孫カテゴリを人手で除いた）に属する記事 1,273 件を対象に関係を抽出する。

５．１　使用関係候補のフィルタリングの評価

抽出した使用関係候補ペアの各項に対する SVM による適切さの判断の精度を評価する。訓練データは 2.2 節、2.3 節で抽出できた使用関係候補のペアの各項に人手で正誤ラベルを付与したもの 939 件（うち正例 355 件)を用いる。RDF カーネルを用い、パラメータは C=256、γ=2^(-8) とした。また、訓練データが不均衡データのため、クラスごとに重みを付けて学習する。テストデータ 147 件（うち正例 41 件）に対し、正例における F 値が 65.2%、負例における F 値が 84.9%、正解率が 78.9%となった。とくに、正例における F 値が低いため、正しい使用関係のみを抽出するのは難しいことが分かった。

５．２　使用関係の抽出結果の評価

抽出された使用関係の一部に対し、それが正しいかどうかを人手で判断し、抽出精度を評価する。今回は 99 件の使用関係に対して、8 名の被験者に判断してもらったところ、正解率の平均は 40.9%となった。実験で使用した使用関係と正誤判断の結果の一部を表 3 に示す。

表 3　抽出した使用関係の例

形態素解析 → 全文検索	正
自然言語処理 → かな漢字変換	正
サポートベクターマシン → 機械学習	正
機械学習 → 検索エンジン	正
バックプロパゲーション → ニューラルネットワーク	正
人工無脳 → ソフトウェア開発者	誤
人工知能 → 1993 年以降	誤
フォルマント合成 → Diphone 合成	誤

5.3　instance-of、subclass-of 関係の抽出結果の評価

対象とした記事から得られた T-上位概念を持つ詳細な上位下位関係は 819 件となり、そこから instance-of 関係は 57 件、subclass-of 関係は 15 件抽出できた。得られた instance-of、subclass-of 関係に対し、人手で正しいかどうか判断をしたところ、正解率は instance-of 関係は 49.1%(28/57)、subclass-of 関係は 46.7%(7/15)となった。実際に抽出した関係を表 4 に示す。誤って抽出された instance-of 関係のうち 27 件は、2 のような「ゲーム名 instance-of ニンテンドーDS」というパターンであり、それ以外の誤抽出は 2 件であった。

表 4　抽出した instance-of、subclass-of 関係の例

1	やねうら王 instance-of コンピュータ将棋	正
2	テトリス DS instance-of ニンテンドーDS	誤
3	距離ベース手法 subclass-of 異常検知	正
4	国際法定計量機関 subclass-of 計量学	誤

６　まとめと今後の展望

本研究では Wikipedia から技術とサービス間の関係を抽出し、それらを用いてグラフを生成した。実際に抽出された関係の正解率はそれぞれ 50%以下であり、改善の余地がある。

今後は、関係の抽出手法を改良するとともに、生成されたグラフを用いた実用的なシステムの作成をしていきたい。

参考文献

[1]太田, 杉本: 学習項目オントロジーを利用した学習内容決定支援システムの構築, 電子情報通信学会総合大会 (2015)
[2]隅田, 吉永, 鳥澤: Wikipedia の記事構造からの上位下位関係抽出, 自然言語処理, 16 巻, 3 号, pp.3-24 (2009)
[3]山田 他: Wikipedia を利用した上位下位関係の詳細化, 自然言語処理, 19 巻, 1 号, pp.3-23 (2012)