特許シソーラスを用いた特許用語定義文の自動生成

Automatic Generation of Patent Term Definitions Using Patent Thesaurus

中央大学理工学部経営システム工学科教授
難波　英嗣

2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。博士（情報科学）。東京工業大学精密工学研究所助手、広島市立大学大学院情報科学研究科准教授等を経て、2019 年より中央大学理工学部教授。自然言語処理、テキストマイニングの研究に従事。

nanba@kc.chuo-u.ac.jp　TEL 03-3817-1833

1 はじめに

日本では、年間に 30 万件以上の特許が出願される。これらの特許には、新しい概念を示す専門用語が多数含まれており、辞書やインターネット検索では意味が調べられないものも数多く存在する。すべての技術分野を包含する専門用語辞書を手作業で構築するのは非常にコストがかかるため、特許データベースから用語の定義文を自動抽出し、特許用語辞書を構築する試みもある [1]。しかし、特許データベース中にそもそも定義文が存在しない用語も少なからず存在する。そこで、本稿では、定義文が存在しない用語について、その定義文を自動的に生成する手法を提案する。

特許データベース中にある用語 X の定義文が存在しなくても、その用語と上位下位関係にある複数の用語に定義文が存在する時、人間であればそれらの定義文を読めば、用語 X のおおよその意味を推測することができる。これを、文書生成技術を用いて自動的に行うことが本研究の最終目的である。

次章では、関連研究について述べる。3 章では、定義文の自動生成手法を提案し、その実現に向けて行った調査結果を 4 章で報告する。最後に 5 章で本稿をまとめる。

2 関連研究

藤井 [1] は、特許データベース（公開特許公報）から、190 万語の見出し語、定義文を抽出している。抽出した定義文は、用語の定義の仕方に応じて例示、目的、要素、機能など、13 種類のカテゴリに自動分類している。本研究において、関連する複数の用語の定義文からある用語の定義文を自動生成する際に、藤井と同様に定義文の自動分類を行う。

筆者は過去の研究 [2] において、特許データベースから定型表現に着目して用語の上位下位関係を抽出することでシソーラスを自動的に構築している。これは、例えば「ダイヤモンドなどの砥粒」という表現がある場合、「などの」という定型表現に着目し、その前後の名詞句（砥粒とダイヤモンド）を上位下位関係として抽出する。本研究では、ある用語の定義文を自動生成する際に、この上位下位関係を利用する。詳細は次節で述べる。

3 特許用語の定義文の自動生成

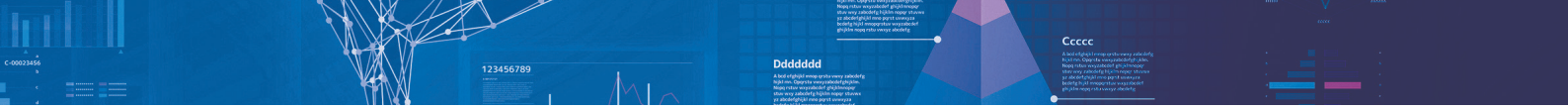
3.1 基本的なアイディア

今、特許データベースから、2 章で述べた定型表現を用いて図 1 に示す上位下位関係が抽出されているものとする。

技術＞自然言語処理＞形態素解析
　　＞構文解析
　　＞機械翻訳

図 1　特許データベースから抽出された用語の上位下位関係の一部（A>B は、A が B の上位語であることを示す）

また、「とは、」という表現に着目し、見出し語＋「とは、」＋定義文というパターンを用いて、特許データベースから 3 つの用語「形態素解析」「機械翻訳」「構文解析」



の定義文が以下（1）～（3）として抽出されているものとする。

- [形態素解析の定義文] 形態素解析とは、文章を文節または単語単位に区切り、それぞれの品詞（名詞、動物、助詞、…）を判別する解析処理である。（1）
- [機械翻訳の定義文] 機械翻訳とは、コンピュータを利用してある言語による文章を自動的に別の言語の文章に変換する技術である。（2）
- [構文解析の定義文] 構文解析とは、与えられた文を解析して、あらかじめ定められた文法からみて、それが正しい文章であるかどうかを整定し、正しい場合に、その文章の構造を抽出することをいう。（3）

上記の定義文を読むと、「形態素解析」「機械翻訳」「構文解析」は、いずれも「文章または文を対象とした（コンピュータを利用した）処理」であり、これが、3 単語の上位語である「自然言語処理」の定義であると推測できる。従って、ある用語の定義文の自動生成は、その用語の下位語の定義文集合の共通項をまとめる、いわゆる複数テキスト要約の一種と考えることができる。

3.2　定義文の分類

2 章でも述べたとおり、用語には様々な定義の仕方が存在する。以下の文は「形態素解析」の定義文であるが、上記のものと異なり、「くるまではこをはこぶ」という例を用いて「形態素解析」はどのようなものであるか説明している。

- [形態素解析の定義文] 形態素解析とは、例えば「くるまではこをはこぶ」と入力された仮名文字列を、辞書に登録された各単語の品詞情報等を参照することで、「くるまで／はこを／はこぶ」と解析する処理をいう。（4）

同じカテゴリに属する定義文であればそれらの共通項をまとめることはできそうだが、このような異なるカテゴリの複数の定義文をまとめるのは困難である。そこで、抽出された定義文をカテゴリごとに分類し、新しい定義文を生成する場合には、同じカテゴリに属する定義文集合から生成する。

3.3　定義文の自動生成

同じカテゴリに属する定義文集合から、ある用語の定義文を自動生成するには、複数の定義文間で共通する個所を自動的に対応付ける必要がある。例えば、3.1 節の例文（1）～（3）において、（1）と（2）の「文章を」および（3）の「文を」を対応付けなければならない。これには、例えば、word2vec などの単語の分散表現を用いて 2 文間の単語レベルの対応付けをとる方法が提案されている [3]。

もし、定義文を生成したい用語の下位語に定義文が存在しない場合には、上位下位関係のみから定義文を生成する。例えば、図 1 において「自然言語処理」の定義文を生成する場合を考える。この時、もし、下位語の「形態素解析」「機械翻訳」「構文解析」のいずれにも定義文が存在しない場合でも、「形態素解析や機械翻訳や構文解析などの技術を指す」という定義文であれば、図 1 の上位下位関係のみから自動生成することができる。

4　特許用語の定義文を自動生成するための調査

本節では、特許用語の定義文を自動生成するために行ったいくつかの調査について報告する。

4.1　特許データベースからの定義文候補および上位下位関係の抽出

1993 年～ 2016 年の公開特許公報を対象に、「とは、」を含む文を抽出した結果、1,333,911 文が得られた。これらを定義文候補とし、4.2 節の実験に用いる。また、「等の」または「などの」を含む文から用語の上位下位関係 14,167,501 件を抽出した。

4.2　定義文の自動分類

実験用データ

4.1 節で抽出された定義文候補を、藤井 [1] の研究に基づいてカテゴリに分類するため、特許データベースから抽出された定義文候補 2858 文を人手で分類した。分類結果を表 1 に示す。なお、ひとつの定義文に複数のカテゴリが付与される場合もある。

表 1 からわかるとおり、カテゴリによって件数がかなり異なる。そこで、1 カテゴリあたり 100 件以上あるカテゴリのみを対象に、定義文候補の自動分類実験を

表 1　定義文候補 2858 文を人手で分類した結果の内訳

カテゴリ番号	カテゴリ名	件数
1	定義	954
2	略語	1
3	例示	600
4	目的	133
5	同義語	31
6	書籍	22
7	製品	8
8	利点	22
9	欠点	15
10	歴史	0
11	要素	454
12	機能	320
13	その他	358
14	説明文でない	254

行った。

評価尺度

再現率と精度を用いて分類器の評価を行った。

実験結果

定義文候補を分類するための学習器として fastText[4] を用いた。word2vec の次元として 100 を用い、5 分割交差検定を行った。実験の結果、再現率 0.485、精度 0.568 が得られたが、定義文を生成するのに利用するには十分な分類精度が得られているとは言えず、今後は、深層学習を用いた文書分類でしばしば用いられる CNN や注意機構の導入なども検討する必要があると考えられる。

5 おわりに

本稿では、特許データベースを用いて特許用語の定義文を自動生成する手法を提案した。本稿で提案した手法を今後実装していくが、現時点ですでに分かっている問題として、多義語の問題が挙げられる。ある特許用語が複数の語義を持つ場合、ある用語の定義文も語義ごとに分割しておく必要がある。この点についても、あわせて今後検討していく。

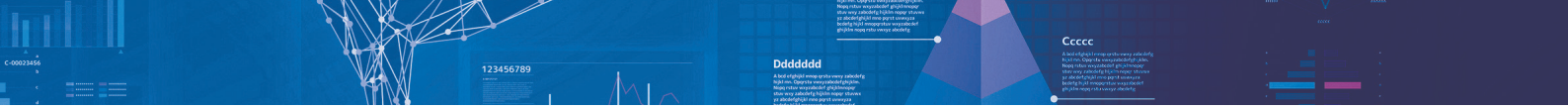
参考文献

[1] 藤井敦“特許情報を用いた辞典検索システム”情報処理学会研究報告データベースシステム , 2008-DBS-145, pp. 9-15 (2008)

[2] 難波英嗣、奥村学、新森昭宏、谷川英和、鈴木泰山“特許データベースからのシソーラスの自動構築”言語処理学会 第 13 回年次大会、pp. 1113-1116 (2007)

[3] Yangqiu Song and Dan Roth “Unsupervised Sparse Vector Densification for Short Text Similarity” Proceedings of NAACL HLT 2015, pp. 1275–1280 (2015)

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov “Bag of Tricks for Efficient Text Classification” arXiv:1607.01759v3 [cs.CL] (2016)


123456789
Dddddd
Ccccc