The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

EP 特許公報を用いた英文同義語辞書の自動作成手法の提案

Automatically building English synonym dictionaries by using European Patent publications

丸崎　恒司[*1]　　津田　和彦[*1]
Koji Marusaki　　Kazuhiko Tsuda

[*1] 筑波大学大学院ビジネス科学研究科

Graduate School of Business Sciences, University of Tsukuba

Recently, conducting patent search becomes increasingly important, especially for foreign patent documents, in order to prepare for many overseas patent disputes and to avoid filing unnecessary patent applications. Therefore, we need to speed up patent research works. For doing so, building English synonym dictionaries is helpful as they can be used to find out suitable keywords during patent surveys and to raise the precision of automatic patent search system. However, developing synonym dictionaries is difficult since it requires a great deal of time and effort. Thus, in this study, we propose the method of automatically generating English synonym dictionaries from European patent documents by using reference signals in the claims and the specifications of them. Consequently, we could extract groups of English synonyms and it showed that the method was useful to create English synonym dictionaries.

1. はじめに

近年，企業の特許戦略が量から質へと変換しつつあること，企業間の特許係争が増加傾向であることを受け，不要な特許出願の抑制や，競合他社の保有する特許の無効化のため，特許調査の需要はますます増え続けている．中でも企業活動の国際化に伴い，外国特許調査の需要は高い．

これらの特許調査においては，漏れのない調査のために，調査対象となる発明に関連した用語のみならず，その同義語や上位概念あるいは下位概念の用語を検索語として調査することが一般的である．

また最近では，機械学習やテキストマイニングの手法により調査対象となる発明と類似する特許文書を自動的に抽出するツールも実用化されている．しかし，このようなツールの多くは，同義語は考慮されていない．その理由の一つに個々の検索対象技術分野に応じた同義語辞書を作成することの困難性がある．

特に特許文書では，「ベルト」に対して「無端伝動部材」など，通常の文書では用いられないような上位概念化した用語，一般化した用語が多く用いられている．

よって，これら特許文書に特有の上位概念化・抽象化した用語を網羅した同義語辞書を作成することができれば，特許調査における最適な検索語の抽出や，類似文書の自動抽出ツールにおける精度向上などに有益と考えられる．

そこで本研究では，欧州特許出願の European Patent publication, European Patent Specification（以下 EP 特許公報と記載する）から自動的に英文同義語辞書を構築する方法を提案する．

2. EP 特許公報を用いた同義語の抽出

2.1 EP 特許公報の特徴

Rule43(7) EPC には，「引用符号を含む図面を有する EP 特許出願では，クレームの理解の助けとなるときは，クレームに記載する技術的特徴には，それらの特徴に関する当該引用符号を括弧に入れて続けることが望ましい」と記載されている．

すなわち，請求項中に記載された技術的特徴が，図面中の各構成のいずれに対応するかを示すために，図面で使用されている番号を引用符号として請求項中の技術的特徴に対応する記載にも付与すべきことを，出願人に要求している[Höhfeld 2013].

通常，請求項に記載された発明は，図面を用いて明細書中でその詳細が説明されている．すなわち図面中の各構成を指し示す引用符号を用いて，特許公報中で当該発明が説明されている．

一方，上述の通り，欧州特許の実務においては明細書中のみならず，請求項に記載の技術的特徴についても当該引用符号が記載されていることが多い．

図1に，EP 特許公報 EP2453476A1 の例を示す．(a)の請求項中の a first contact area (12) は，その引用符号 12 から，(b)の明細書中の記載 a first contact portion 12 に対応することが分かる．

> wherein a raised portion of the leadframe adjacent to said recess defines a first contact area (12);

(a)請求項中の記載

> provided e.g. by stamping or etching with any suitable etch recipe to provide a leadframe 10 in which a first contact portion 12 is defined adjacent to the recess 14.

(b)明細書の記載

図 1　EP2453476A1 中の記載

通常，請求項では権利範囲を広く確保するために，発明の技術的特徴を上位概念で記載し，明細書中では当該技術的特徴の具体的な内容を下位概念で記載することが多い．つまり請求項中で，技術的特徴を表す上位概念の用語が，明細書中では請求項中の記載とは異なる下位概念の用語で記載されることがよく見られる．図 1 は，”area”という上位概念の用語に対して，”portion”という下位の用語が対応していると思われる．

連絡先:丸崎 恒司，筑波大学大学院 ビジネス科学研究科，
東京都文京区大塚 3-29-1，s1740127@u.tsukuba.ac.jp

2.2 引用符号を用いた照合モデル

上記の EP 特許公報の特徴を活かし, EP 特許公報の請求項中の発明の技術的特徴に対応する記載に付された引用符号と, 明細書中の当該技術的特徴を具体的に記載した部分に付された引用符号を対応付けることで, 上位概念化された用語と, より具体的な下位概念の用語とを対応させた同義語辞書を, 自動的に作成することが可能と思われる. さらには, 請求項内や明細書中で同一の引用符号が複数出現するものも多々あるため, これらも全て抽出して照合することで, 表記の揺れや俗称など様々な同義語が抽出できる可能性がある.

これまでも, 請求項中の発明の技術的特徴とそれに対応する明細書中の記載から上位概念, 下位概念の用語対を, 定型表現等を用いて抽出し, 同義語辞書やシソーラスを作成する手法は研究がなされてきた[難波 2007] [間宮 2011]. しかし, 請求項の記載箇所と, それに対応する明細書の記載箇所を自動的に特定することが困難であった.

そこで本研究では, EP 特許公報に特有の請求項中の引用符号と明細書中の引用符号を対応させることで, 請求項中の発明の技術的特徴に対応する明細書中の詳細説明部分を容易に対応させることができる点に着目し, 請求項中のフレーズと, それに対応する明細書中のフレーズを元に, 上位概念の用語と下位概念の用語, それらの表記の揺れや俗称などからなる用語対を作成し, それらを集めて同義語辞書を作成する方法を検討した.

3. 引用符号の抽出手法と評価

3.1 同一引用符号を持つフレーズの抽出

以下に, EP 特許公報, EP 2453476A1 を例に, 請求項と明細書から, 同一引用符号を持つフレーズを抽出する方法について, 説明する.

EP 特許公報は, ヨーロッパ特許庁(EPO)の提供する European Publication Server[data.epo]から pdf や xml の形式でダウンロードすることができる. 今回は同様に EPO の提供する European Patent Register[register.epo]の Advanced search を用いて, Keyword(s) in title で“semiconductor”かつ“packaging”をタイトルに含む 202 件(2017.10.14 確認)を選定し, そこから European Patent Publication を有し, 請求項中に引用符号を含む 10 件を検討対象とした. また当該 10 件については, Publication No.を European Publication Server に入力することで, xml 形式のファイルを取得した. 当該 xml ファイルはテキストデータとして読み込んだのち, それらをすべてトークン化した. 引用符号を持つフレーズは, その記載ルール従い, フレーズの後に引用符号が記載されている. それゆえ, 当該請求項中の丸括弧()で囲まれた数詞(引用符号), 及び明細書中の数詞(引用符号)を見出し, そこから前方に単語をたどればフレーズを抽出することができる. この際, 課題となるのは, 数詞から前方へ何単語目までがフレーズかという判断である. この判定は表1に示す抽出ルールに基づき, 行った. ただし丸括弧()で囲まれた対応する数詞(引用符号)を持たない数詞, すなわち明細書中にのみ記載されている引用符号を持つフレーズ, また英語以外の言語を含むフレーズはノイズとして対象外とした. また数詞であっても, Fig 1 や claim 1 などの数詞は引用符号とならないため, 数詞の直前に”Fig”(“Figs”)や”claim”(“claims”)を持つフレーズもノイズとして抽出対象外とした.

表1 フレーズの抽出ルール

	抽出終点	抽出起点
A	冠詞(a,the,said)	数詞
B	前置詞	数詞
C	and(数詞の直前の and は無視する)	数詞
D	接続詞	数詞
E	, () <>などの記号(ただし数詞直前の記号は除く)	数詞

請求項や明細書中の数詞(引用符号)から前方にどこまでを抽出すべきフレーズとするか, を定めたものが表1のルールである. 表1に従い, 実際の作業では請求項や明細書中の引用符号の位置から, 前方に, 例えば A のルールでは冠詞の直前までをフレーズとして抽出する.

図 1(a) (b)ではいずれも, 数詞(12)や 12 から前方に進むと冠詞が存在するので, 表1の A に従い, それぞれ first contact area, first contact portion が抽出されるべきフレーズとなる.

そのようにして得られたフレーズから, さらに表 2 に示す同義語抽出ルールを用いて, 同義語対を作成する. 表 1 のルールに基づき得られたフレーズは, 序数詞や分詞などの修飾語を含むため(先述の first contact area, first contact portion の例では, first), このままでは, 同義語辞書として活用することはできない. そこで, 抽出したフレーズにはさらに表 2 のルールを適用した.

表 2 同義語抽出ルール

フレーズから削除する項目
・序数詞・数詞 ・形容詞 ・分詞 ・副詞 ・接続詞 ・,や(などの記号

その結果, 先述の引用符号(12), 12 に対応するフレーズから, contact area, contact portion の2つの同義語からなる同義語対が抽出できた.

表 3 に EP2453476A1 から表 1, 表 2 のルールを適用して抽出した同義語対を示す.

表 3 EP2453476A1 から得られた同義語対

得られた同義語対
leadframe portion, leadframe
semiconductor device package, semiconductor device packages, package
contact area, contact portion, contact portions, contact surface, surfaces
recess, recesses
semiconductor device, dies
contact, contacts, carrier contact
surface, contact surface, contact surfaces, contact area, surfaces
layer, resin
layers, layer

表 3 から, 今回提案した手法を用いることで, layer と resin のように, “層”の材質を特定した, すなわち下位概念化した“樹脂”が抽出されていることが確認できた. よって layer と resin については, 上位概念と下位概念の同義語対が抽出できていると考えられる.

また例えば contact area と, contact surface のような, 上位・下位の関係とは言い難く, 言い換え表現となっている同義語対も得られていることが確認できた.

3.2 抽出ルールの正答率

表 4 に本手法を用いて抽出した 2354 フレーズの正答率を示す. ルール E までを用いることで, 最終的に 90%を超える正答率が実現され, 当該手法による同義語辞書作成の可能性が示された.

表 4 各抽出ルール適用時の正答率

	A	+B	+C	+D	+E
正答数	1506	1838	1889	1999	2126
抽出数	2354	2354	2354	2354	2354
正答率	0.640	0.781	0.802	0.849	0.903

また実際に得られた同義語対の結果から, 例えば表 3 に示す semiconductor device, dies の他に die, electronic device が得られることが確認できた. それぞれ semiconductor device, electronic device が dies, die の上位概念になっている. 以上から, dies と die という複数, 単数の相違はあるものの, dies, die を共通の用語とみなし, semiconductor device, dies と die, electronic device を1つの同義語対とみなすことで, dies(die)の上位概念として semiconductor device, electronic device の2つを対応させ得ることが確認できた. このように共通用語を介して2つの同義語対を結合することで, 本手法で得られる同義語辞書をシソーラスに発展させられる可能性があることが示された.

4. おわりに

本研究では, EP 特許公報の請求項及び明細書の記載に対して共通に付されている引用符号を用いることで, 英文同義語辞書を自動作成する手法について, 提案した.

今回, 提示した抽出ルールによって, 一定の正答率での同義語対を抽出可能なことが分かった. これによって本手法により自動的に英文同義語辞書を作成する可能性が示された.

今後は, 当該同義語辞書を用いて, 意味的な階層構造を持つシソーラスの作成が可能か, またこれらを実際の特許調査に応用して, 類似文書抽出精度を上げることが可能か検証していく.

参考文献

[register.epo] https://register.epo.org/advancedSearch?lng=en, accessed:2017-10-14

[data.epo] https://data.epo.org/publication-server/?lg=en, accessed:2017-10-14

[Höhfeld 2013] Jochen Höhfeld, 田中 紫乃: 欧州特許出願においてすべきこと,すべきでないこと, パテント, 2013.

[難波 2007] 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, 2007.

[間宮 2011] 間弓沙織, 難波英嗣, 竹澤寿幸: 日英特許データベースからのシソーラスの自動構築論文, 言語処理学会 第17 回年次大会 発表論文集, 2011.