1H2-4

製品名と特許における名称の言い換え知識の抽出

Extracting Paraphrase Knowledge of Product Names in Patent Documents

坂地泰紀 *1 Hiroki Sakaji　所佳祐 *1 Keisuke Tokoro　酒井浩之 *1 Hiroyuki Sakai

*1成蹊大学
Seikei University

This paper proposes a method that extracts paraphrase knowledge of product names in patent documents. Our method extracts paraphrase knowledge of the product names from patent documents by using IDF and word2vec. We evaluated our method and confirmed that our method outperform similarity of word2vec only.

1. はじめに

日本国内における特許出願は年間約 30 万件にも達している．特許出願される技術は，出願人である企業等にとって非常に重要なものである場合が多く，その出願動向を調査することは，企業における技術開発戦略，及び，知財戦略の策定や国，地方自治体における技術開発推進政策立案に大きく寄与する．しかしながら，特許文書中には日常生活では使用しない専門的な用語で製品を表しているものもあるため，一般的な製品名称を用いた特許検索は困難な場合がある．例えば，製品名称である「液晶テレビ」に対して，特許文書中では「画像表示装置」や「液晶表示装置」と表される．これらの特許文書中で出現する製品名称に対する専門的な用語（以降，言い換え表現とする.）の辞書が構築できれば，製品名称を用いた特許検索が可能となり，ある製品の開発に利用された特許の検索ができるようになると考えられる．

そこで我々は，製品名称に対する特許文書中での言い換え表現の辞書作成を目的とする．しかしながら，人手によって特許文書中から言い換え表現を探すのは難しいこと加え，言い換え表現の候補も数多く存在することから，自動的に辞書を構築するのは困難である．そのため，本研究では，製品名称に対する言い換え表現の候補を自動的に抽出する技術を開発する．言い換え表現の候補に言い換え表現らしさを示すスコアを付与し，スコアの上位に言い換え表現が存在すれば，人手によって容易に言い換え表現の辞書を構築できるようになる．

2. 製品名称の抽出

特許文書から言い換え表現を抽出するにあたり，まず，製品名称を抽出する必要がある．そこで，我々は，製品発表に関する記事が存在する日経プレスリリースより製品名称を抽出する．表 1 の製品名称抽出パタンを用いて，2012 年から 2014 年の日経プレスリリース 61,390 記事のタイトルから製品名称を抽出した．表 1 は，Python の正規表現で記述しており，<product> に製品名称が当てはまる．

例えば，製品名称抽出パタンを適用した場合，以下のような文から製品名称が抽出される．

表 1: 製品名称抽出パタン

.* 楽しめる (?P<product>.+?) (を発売|を開発) .*
.* した (?P<product>.+?) (を発売|を開発) .*
.* の (?P<product>.+?) (を発売|を開発) .*

- 使いやすさを追求した**ノートパソコン**を開発
- 高品質なデザインの**電子ピアノ**を発売

製品名称抽出パタンを用いて製品名称を抽出したところ，620 個の製品名称を抽出することができた．抽出した製品名称の例を表 2 に示す．

表 2: 製品名称の例

カップめん　デジカメ　缶コーヒー　スマートフォン 液晶テレビ　清涼飲料水　デジタル一眼レフカメラ 入浴剤　ゴルフクラブ　腕時計　シューズ

3. 言い換え表現候補の抽出

本節では，言い換え表現候補の抽出手法について述べる．製品名称に対応する言い換え表現候補を得るために，我々は，製品名称が含まれる特許文書には，言い換え表現が含まれるという仮定をし，各製品名称が含まれる特許文書をそれぞれ獲得する．その後，各製品名称が含まれる特許文書から言い換え候補となる名詞 N-gram を抽出する．各製品名称に対して，以下に示す手続きを行い，言い換え表現候補を抽出する．

Step 1: 製品名称が含まれる特許文書を獲得する．

Step 2: Step 1 で獲得した特許文書集合から，言い換え表現候補となる 2 文字以上の名詞 N-gram を抽出する．

Step 3: IDF 値に基づき，抽出した名詞 N-gram の選別する．

Step 4: Step 3 で選別した名詞 N-gram と，製品名称の両方を含む文を特許文書から獲得する．

連絡先: 坂地泰紀，成蹊大学，東京都武蔵野市吉祥寺北町 3-3-1，hiroki_sakaji@st.seikei.ac.jp

表 3: 評価結果

製品名称	frequency	word2vec	本手法
液晶テレビ	液晶表示装置 (6)	液晶モニタ (16)	液晶ディスプレイ (11)
LED 電球	LED 照明 (73)	LED 照明 (58)	LED 照明 (16)
キャディバッグ	ゴルフバック (29)	ゴルフバック (513)	ゴルフバック (18)
消臭芳香剤	臭気吸着剤 (204)	臭気吸着剤 (582)	臭気吸着剤 (279)
デジタル一眼レフカメラ	撮像装置 (41)	撮像装置 (133)	撮像装置 (19)
チューハイ	アルコール飲料 (97)	アルコール飲料 (7)	アルコール飲料 (7)
平均出現順位	180.181	155.636	55.3

Step 5: Step 4 で獲得した文集合から各名詞 N-gram の頻度と，word2vec[*1] による類似度を計算する．頻度と word2vec 類似度を掛け合わせた値をスコアとし，スコアの上位に出現する名詞 N-gram を言い換え表現候補として抽出する．

3.1 IDF 値による選別 (Step 3)

Step 3 で行う IDF 値による名詞 N-gram の選別について述べる．製品名称を含む特許文書から全ての名詞 N-gram を抽出すると，数多くの不適切な名詞も抽出してしまう．そこで，本手法では，IDF 値を用いて名詞 N-gram の選別を行う．

以下の式 1 に基づき，名詞 N-gram n の $IDF(n)$ を計算し，IDF 値が 5 未満の名詞 N-gram を削除する．

$$IDF(n) = \log_2 \frac{N}{df(n)} \quad (1)$$

ここで，N は全特許文書の数，$df(n)$ は名詞 N-gram n が出現する特許文書の数となる．

4. 評価実験

本手法を評価するために，評価実験を行う．評価実験には，1996 年から 2003 年に出願された特許 2,468,599 件を使用する．また，製品名称には，2. 章で抽出した 620 個の製品名称を用いる．形態素解析器には MeCab[*2] を用いる．

比較手法として，本手法 Step 5 で算出した頻度をスコアとした手法 (frequency) と，word2vec の類似度のみをスコアとした手法 (word2vec) を用いる．

4.1 評価結果

本手法と，比較手法を評価した結果の一部と，それぞれの平均結果を表 3 に示す．表 3 では，それぞれの手法で抽出できた言い換え表現と，その言い換え表現が出現した順位を括弧内に示している．この数値が低いほど，本タスクにおいて良い手法であることを示している．また，製品名称が「LED 電球」であった場合の言い換え表現と，その出現順位を表 4 に示す．

表 4: 製品名称「LED 電球」の評価結果

LED 電球	LED 照明用 (15)，LED 照明 (16) LED 照明用電球 (17)，照明用電球 (18)

*1 https://code.google.com/archive/p/word2vec/
*2 http://taku910.github.io/mecab/

5. 考察

表 3 より，比較手法である frequency や word2vec に比べ，本手法の平均順位が高くなっている．このことより，本手法の有効性を示すことができた．また，表 4 より，製品名称が「LED 電球」であった場合の言い換え表現の出現位置が上位にあることがわかる．しかしながら，表 3 の製品名称「消臭芳香剤」の結果が示す通り，いくつかの製品名称では比較手法の方が良い結果となった．これは，頻度が有効であったの対して，word2vec 類似度が有効ではなかったことから，本手法のスコアが word2vec に影響を受けてしまったことが起因する．今後の課題として，上記のように片方の指標が有効でなかった場合の対処を考える必要がある．

6. 関連研究

関連研究として，寺田らは「同義語は同じような文脈で使用される」という仮定から同義語を得る手法を提案している [寺田 07]．難波らは，同義語抽出手法を利用し，論文用語を特許用語に自動的に変換する手法を提案している [難波 10]．特許を対象とした研究として，酒井らは製品発表プレスリリースと，その製品と関連のある特許を判定する手法を提案している [酒井 13]．これらの研究に対して，本研究では，word2vec を利用することで，製品名称に対応する特許文書中の言い換え表現の抽出を試みている．

7. まとめ

本研究では，製品名称に対する言い換え表現の候補を自動的に抽出する手法を提案した．頻度を word2vec 類似度を組み合わせることで，比較手法よりも良い結果となった．今後の課題として，片方の指標が有効でなかった場合の対処が考えられる．

参考文献

[寺田 07] 寺田 昭, 吉田 稔, 中川 裕志 : 同義語の類似度に関する考察, 言語処理学会第 13 回年次大会, pp. 1097–1100 (2007)

[酒井 13] 酒井 浩之, 増山 繁 : 製品特徴に基づく製品発表プレスリリースと特許との関連性の判定, 言語処理学会第 19 回年次大会, pp. 725–728 (2013)

[難波 10] 難波 英嗣, 竹澤 寿幸 : 同義語抽出手法を利用した論文用語の特許用語への自動変換, 言語処理学会第 16 回年次大会, pp. 772–775 (2010)