特許文書のための形態素解析辞書の構築

橋本 泰一 †　　藤井 敦 ‡

† 東京工業大学 総合プロジェクト支援センター　‡ 東京工業大学 大学院情報理工学研究科 計算工学専攻

†hashimoto.t.ab@m.titech.ac.jp

1 はじめに

近年の電子化テキストの普及により，テキストデータが爆発的に増加している．現在は，膨大なテキストデータから欲しい情報を検索したり，テキストデータから抽出した情報を二次利用したりといったことが必然である．情報検索やテキストマイニングといった技術を支える基礎的な自然言語解析技術として，形態素解析がある．形態素解析で重要な要素技術が大きく二つある．一つは解析アルゴリズムであり，もう一つは形態素解析用辞書である．

形態素解析用辞書として有名なのは，情報処理振興事業協会 (IPA) の IPA 品詞体系 (THiMCO97) に基づいて作成された IPA 品詞体系日本語辞書（IPADIC）[1]である．現在，IPADIC の後継として開発されている日本語辞書が NAIST Japanese Dictionary（NAIST-JDIC）[2]である．その他には，京都大学が形態素解析器 JUMAN[3]のための辞書や国立国語研究所が開発している形態素解析用日本語辞書 Unidic[4]などがある．

従来の形態素解析用辞書は日本語を形態素解析する上で基本的な単語を中心に辞書を構築していた．しかし，ウェブや特許などの様々なテキストを解析する上で，従来の辞書が十分な範囲の単語をカバーしているとは言いがたい．さらに，一般的に形態素解析用辞書の構築には，人手による非常に多大なコストが必要である．なぜならば，形態素の区切りをどのように定義すべきか，新語の品詞をどのように定義するかといった問題を，各新語ごとに文章内での使われ方やその言語的機能を分析しながら検討しなければならないためである．

本研究では，特許文書に特化した形態素解析辞書の拡張についての事例研究について報告する．特許文書は，専門的な用語や造語が漢字のみ構成されやすいといった特徴を持っている．その特徴を生かし，15年分の特許公報から自動的に漢字列を抽出し，人手で作成した規則で形態素候補を選定する．そして，周辺文脈を素性としたベクトルを用いて，従来の形態素解析器の辞書と k 近傍法アルゴリズムで新語の品詞推定を行う．

2 関連研究

特定分野のコーパスから専門用語を自動的に抽出する研究が盛んに行われている．Sekine ら [2] はタグ付きコーパスから専門用語の出現位置を学習し自動抽出する手法を提案した．教師付き機械学習とは異なるアプローチとして，「専門用語は対象の分野に多く出現し，他の分野においては一般的な語でない」という特徴をいかし，統計的に専門用語抽出を行う手法も提案されている [1,3]．長町ら [4] はコーパス中の文字 N グラムの統計量を利用して専門用語の抽出する手法を提案している．中川らの FLR 法 [6] は，前後の単語の統計量をベースとして抽出を行い，村上ら [5] は，専門用語と助詞の関連性について指摘している．

これらの研究は主に専門用語の抽出手法であり，形態素解析用の辞書への応用は検討されていないため，品詞の推定をしていない．また，専門用語抽出の精度はそれほど高くない．なぜならば，様々な表現パターンの専門用語を単一の枠組みで抽出しようとしていることが原因であると考えられる．

3 特許文書を用いた形態素解析用辞書の拡張

本研究では，特許文書から新語（専門用語）を抽出し，その語の品詞を推定し，既存の形態素解析用辞書を構築する．この研究では，抽出する新語を名詞に限定する．なぜならば，特許文書において誤解析をしやすいのは，名詞もしくは複合名詞であるためである．さらに，ありとあらゆる表現パターンの用語を高い精度で抽出することは困難であると判断したためである．特許文書では，下記のような既存の形態素解析用辞書に含まれにくい表現が多数出現する．

分野特有の専門的な用語

装置の部品や図中の部分を指し示す参照表現

変化，状態，特性に関する表現

例えば，特許文書（特開 2007-04801，図 1）中の要約では，「シェーディング」「レジストパターン」「ＣＭＰ法」

[1]http://sourceforge.jp/projects/ipadic/
[2]http://sourceforge.jp/projects/naist-jdic/
[3]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4]http://www.tokuteicorpus.jp/dist/

画素が 微細化 され、かつ、画像感度、色 シェーディング、感度 シェーディング が良好な、画像特性の良い固体撮像装置を実現するための固体撮像装置の製造方法を提供する。

層間絶縁膜１０の表面のうち、ゲート電極６と、光電 変換部３を保護する保護膜７とが重畳形成され、さらに、遮光膜９が形成されている部分には、面積が広く高さが高い 凸部 が現れる。層間絶縁膜１０の表面を 平坦化 するために、まず、この凸部上に 開口 ２２を有する レジストパターン ２１を形成する。そして、レジストパターン ２１から露出した部分を ドライエッチング して 凸部 の大きさを予め小さくしておいた後に、ＣＭＰ法 で層間絶縁膜１０の表面を 平坦化 する。そして、平坦化 後の層間絶縁膜１０上に 光路 変更レンズを形成する。

図 1: 特許文書の例：特開 2007-042801 から引用

「光電」といった専門用語（紫字）,「凸部」「開口」といった装置のある部分を表す語（青字）,「微細化」「平坦化」といった変化を表す語（赤字）が出現している．このような単語は，特許文書に現れやすいが，新聞記事などではあまり出現しない．そのため，従来の形態素解析用辞書には登録されていないことが多い．

3.1 新語候補の抽出

特許文書では，比較的，漢字のみで構成された複合語を使いやすい．そこで，特許文書中の漢字列に注目して，新語を抽出する．まず，特許文書から下記の条件に当てはまる漢字列をパターンマッチで新語候補として抽出する．

抽出条件 1

- 漢字列の直前の文字が,「,」「、」「.」「。」「の」「を」「に」「る」「が」「て」「た」のいずれか
- 漢字列の直後の文字が,「,」「、」「.」「。」「(」「の」「を」「に」「る」「が」「と」「で」のいずれか

しかし，以上の条件だけでは,「水等（みずなど）」「紙等（かみなど）」「色毎（いろごと）」「年毎（としごと）」といった「名詞＋接尾」や「十色」「一例」といった「数字＋名詞」のような誤ったパターンの漢字列を新語として抽出してしまいやすい．そのため，以下の条件を追加して絞り込む．

抽出条件 2

- 漢数字を含まない[a]
- 漢字列の先頭の文字が,「又」「程」「際」「用」「他」ではない
- 漢字列の最後の文字が,「毎」「等」「別」ではない

[a]ただし，3 文字かつ先頭が「一」の場合を除く

加えて，低頻度の漢字列は削除し，新語候補を決定する．

3.2 新語の品詞推定

まず，前節で抽出した新語候補を，既存の形態素解析用辞書に登録されているものとそうでないものに分類し，辞書に含まれていない漢字列を新語とする．既存の辞書に登録されている漢字列（登録語）は品詞推定に利用する．

各漢字列は，直前直後の 1 文字の組み合わせを次元とし，その出現頻度を値とするベクトルで表現する．

$$\text{端部} = (\text{“}\_\text{の}\text{”}:5, \text{“.,”}:2, \text{“.　で”}:1, \text{“た,”}:4, \ldots)$$

そして，k 近傍法（k-Nearest Neighbor Algorithm, k-NN）を用いて，新語の品詞推定を行う．新語と類似する登録語を，先に述べたベクトルとコサイン類似度で計算する．類似度が高い上位 10 登録語の品詞に対して，新語と登録語の類似度をスコアとして加える．もっともスコアが高かった品詞を新語の品詞として登録する．

4 評価実験

4.1 新語抽出実験

前節で述べた手法により，過去 15 年間（1993 年から 2007 年まで）に登録された特許公報を対象として，新語の抽出と品詞の推定を行った．この実験では，特許文書中の「名称」「要約」「詳細な説明」「請求項」に含まれるテキストを対象として抽出を行った．

抽出した文字列長が 2 もしくは 3 の漢字列に対して抽出条件 1 を適用した．抽出した結果を用いて，特許文書における新たな漢字列の増加具合について確認した．図 2 は，文字列長が 2 の漢字列の異なり数の推移であり，図 3 は，文字列長が 3 の場合である．横軸は対象とした特許公報の登録期間で，縦軸は漢字列の異なり数である．図中の「漢字列」は抽出条件 1 を適用した漢字列の数，「NAIST」は抽出した漢字列のうち，NAIST Japanese Dictionary（NAIST-JDIC）に登録されている単語の数，

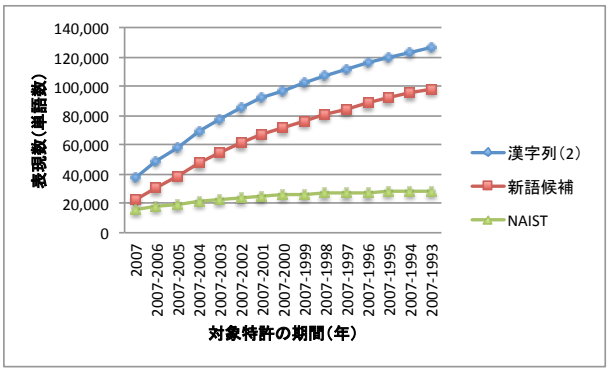

図 2: 長さ 2 の漢字列の頻度分布

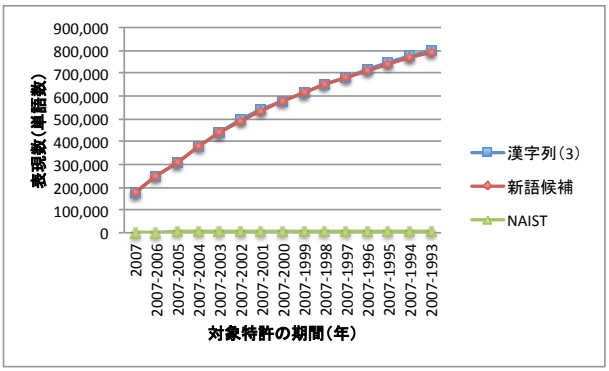

図 3: 長さ 3 の漢字列の頻度分布

表 1: 抽出された漢字列の例

2 文字列		3 文字列	
新語候補	NAIST	新語候補	NAIST
端部	場合	本発明	自動的
同図	特徴	実施例	周波数
下側	記載	具体的	複数個
底部	状態	化合物	主成分
凹部	複数	本実施	自動車
膜厚	可能	先端部	不純物
本例	方法	混合物	被写体
基材	必要	外周面	短時間
図中	構成	軸方向	誘導体
金型	所定	一般的	紫外線

「新語候補」は,「漢字列」から「NAIST」を覗いた漢字列の数を表す.

文字列長が 2 の漢字列は 12 万種類以上，文字列長が 3 の漢字列は 79 万種類以上が抽出されている．1993 年から 2007 年までの 15 年間で，2 文字もしくは 3 文字の漢字列は増加傾向にあり，以前として収束する傾向にない，従来，特許文書では新語生成が盛んであると言われているが，この結果からも同様の傾向があることがわかる．また，抽出された漢字列のうち，NAIST-JDIC に登録されている単語は，文字列長 2 では約 3 万語，文字列長 3 では約 4 千語程度であった．NAIST-JDIC に登録されていない漢字列すべてが新語である可能性はないが，辞書へ収録される語を増強する必要性はおおいにある．

高頻度の抽出例を表 1 に記載する．

4.2 品詞推定実験

抽出条件 2 を適用し，k 近傍法により品詞推定をし，新規に形態素解析用辞書として登録する語を決定する．実験では，出現頻度 10 以上の新語候補を対象とし，コサイン類似度が 0.8 以上の NAIST-JDIC に登録されている単語（最大 20，k=20）から品詞を推定した．類似度が 0.8 以上となる単語がない場合は品詞推定されず新語として登録しない．抽出した新語の例を表 2 に示す．

表 3: 抽出された漢字列（新語）の数

	抽出条件 1	抽出条件 2	品詞推定
文字列長 2	97,812	92,882	20,477
文字列長 3	792,621	726,786	82,871

自動的に抽出した新語を既存の形態素解析用辞書に加え，実際の特許文書を解析した．特許文書からランダムに抽出した新語を含む文（2 万文）を解析対象文とする．形態素区切りが異なる箇所は，66,978 箇所であった．最も顕著な修正箇所としては，従来の辞書では，漢字一文字を動詞の語幹と解析していたところを修正しているところであった（10,782 箇所）.

良修正

拡張前

折	動詞, 自立,*,*, 五段・ラ行, 体言接続特殊 2
曲	動詞, 自立,*,*, 五段・ラ行, 体言接続特殊 2
端	名詞, 形容動詞語幹,*,*,*,*

拡張後

折曲端	名詞, 一般,*,*,*,*,

ただし，単語が増加したことにより，長い複合語の解析が比較的誤りやすくなる傾向にある．

誤修正

拡張前

径	名詞, 一般,*,*,*,*
方向	名詞, 一般,*,*,*,*
端	名詞, 形容動詞語幹,*,*,*,*
部	名詞, 接尾, 助数詞,*,*,*

拡張後

径方	名詞, 一般,*,*,*,*
向	接頭詞, 名詞接続,*,*,*,*
端部	名詞, サ変接続,*,*,*,*

表 2: 抽出された新語の例

2 文字列				3 文字列			
端部	名詞, サ変接続	同図	名詞, 一般	本発明	名詞, 一般	実施例	名詞, 一般
下側	名詞, 一般	底部	名詞, 一般	具体的	名詞, 一般	本実施	名詞, 一般
凹部	名詞, 一般	本例	名詞, 一般	先端部	名詞, 一般	混合物	名詞, 一般
端面	名詞, 一般	図中	名詞, 副詞可能	外周面	名詞, 一般	一般的	名詞, 形容動詞語幹
金型	名詞, 一般	光軸	名詞, 一般	電気的	名詞, 形容動詞語幹	範囲内	名詞, 一般
外径	名詞, 一般	回動	名詞, サ変接続	開口部	名詞, 一般	選択的	名詞, 形容動詞語幹
粘度	名詞, 一般	筒状	名詞, 一般	実質的	名詞, 形容動詞語幹	中央部	名詞, 一般
光束	名詞, 一般	流路	名詞, 一般	断面図	名詞, 一般	置換基	名詞, 一般
凸部	名詞, 一般	係合	名詞, サ変接続	幅方向	名詞, 一般	連続的	名詞, 形容動詞語幹
板状	名詞, 一般	開度	名詞, サ変接続	一体的	名詞, 一般	回転数	名詞, 一般
周面	名詞, 一般	粒径	名詞, 一般	反射光	名詞, サ変接続	両端部	名詞, 一般
軸線	名詞, 一般	各色	名詞, 一般	問題点	名詞, 一般	下端部	名詞, 一般
内周	名詞, 一般	光路	名詞, サ変接続	周方向	名詞, 一般	相対的	名詞, 形容動詞語幹
縁部	名詞, 一般	頂部	名詞, 一般	基板上	名詞, 一般	上端部	名詞, 一般
肉厚	名詞, 一般	軸心	名詞, 一般	効果的	名詞, 形容動詞語幹	付勢力	名詞, 一般
負圧	名詞, 一般	潜像	名詞, 一般	最終的	名詞, 形容動詞語幹	外周部	名詞, 一般
後側	名詞, 一般	角部	名詞, 一般	横方向	名詞, 一般	一時的	名詞, 形容動詞語幹
横軸	名詞, 一般	縦軸	名詞, 一般	耐久性	名詞, 一般	絶縁膜	名詞, 一般
全周	名詞, 一般	要部	名詞, サ変接続	駆動力	名詞, 一般	平面図	名詞, 一般
純水	名詞, 一般	正極	名詞, 一般	抵抗値	名詞, 一般	平均値	名詞, 一般
巻線	名詞, 一般	液滴	名詞, 一般	一端部	名詞, 一般	上流側	名詞, 一般
圧油	名詞, 一般	反力	名詞, 一般	密着性	名詞, 一般	部分的	名詞, 形容動詞語幹
長尺	名詞, 一般	溶湯	名詞, 一般	小型化	名詞, 一般	反対側	名詞, 一般
別体	名詞, 一般	基端	名詞, 一般	添加剤	名詞, 一般	高精度	名詞, 一般
脚部	名詞, 一般	軸部	名詞, 一般	効率的	名詞, 形容動詞語幹	先端側	名詞, 一般
治具	名詞, 一般	後面	名詞, 一般	後端部	名詞, 一般	説明図	名詞, 一般

また，新語としては正しいが非常に類似した語があり，誤解析される場合もある．例えば,「試打（しだ）」という新語に対して「試打ち（ためしうち）」という表現が存在し,「試打ち」を誤解析されやすい．この研究では，漢字のみから構成される新語を対象としていたが，送り仮名付きの新語も同時に抽出しなければ，解析精度を下げる場合もある．

5 まとめ

本研究では，特許文書を対象に既存の形態素解析辞書に含まれない新語（専門用語）を抽出し，品詞推定することで，辞書を拡張する手法について報告した．特許文書は，専門的な用語や造語が漢字のみ構成されやすいといった特徴を生かし，1993 年から 2007 年の間に登録された特許公報から自動的に漢字列を抽出し，人手で作成した規則で形態素候補を選定する．そして，周辺文脈を素性としたベクトルを用いて，従来の形態素解析器の辞書と k 近傍法アルゴリズムで新語の品詞推定を行った．実験により，約 10 万語の新語を抽出できた．

今後の課題として，新語の品詞推定の精度や辞書の拡張による形態素解析精度の評価を行わなければならない．加えて，カタカナや送り仮名付きの新語を対象とした抽出にも取り組む必要がある．この研究成果をベースに新語登録を拡張し，人手により確認・修正後，一般公開する予定である．

参考文献

[1] Y. Fukushige and N. Noguchi. Ntcir experiments at matsushita: Tmrec task. In *Proc. of the 1st Conference of NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 467–474, 1999.

[2] S. Sekine, R. Grichman, and Hiroyuki Shinnou. A decision tree method for finding and classifying names in japanese texts. In *Proc. of the Sixth Workshop on Very Large Corpora*, 1998.

[3] K. Uchimoto, S. Sekine, M. Murata, H. Ozaku, and H. Isahara. Term recognition by using corpora from different field. *Terminology*, Vol. 6, No. 2, pp. 233–256, 2001.

[4] 長町健太, 武田善行, 梅村恭司. 文書拡張によるキーワード抽出. 自然言語処理, Vol. 14, No. 1, pp. 67–86, 2004.

[5] 村上浩司, 乾孝司, 橋本泰一, 内海和夫, 石川正道. 専門用語抽出における助詞情報の利用に関する一考察. 情報処理学会自然言語処理研究会 (2007-NL-182), 2007.

[6] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.