特許関連業務支援のための技術用語自動抽出の試み

柚木山 駿, 太田 貴久, 小林 暁雄, 増山 繁
豊橋技術科学大学 情報知能工学専攻
{yukiyama，kikyu} @ la.cs.tut.ac.jp，a-kobayashi@cs.tut.ac.jp，masuyama@tut.jp

1 はじめに

現在，日本を含め世界各国で日夜新たな技術の開発，発明が数多くなされている．そしてそれらの保護とさらなる発明の推奨のため日々新たな特許の申請がなされており，その中には海外で申請されていても日本語の特許文書が存在しないものも存在する．その情報を日本語で得たい際には翻訳する必要があるが，特許文書には技術用語，新語や文献独自の定型表現が多く出現するため，翻訳精度の向上には，これらの表現を内包した辞書が不可欠である．この辞書によって，特許文書に対する検索精度の向上や，翻訳者に訳語の選択肢を提供し翻訳作業を補助することが可能となる．

辞書作成の第一段階としてまず日本語の特許文書に出現する技術用語を収集する必要がある．この対象は最初から日本語で記述された特許文書である．これにより，別の言語から日本語に翻訳された文書に出現する技術用語について,「完全な新語」,「対応する語があるが翻訳者が新語を作成した」,「対応する語があるが別の技術用語に翻訳した」といった新語，及び，誤りの判別が可能となる．

そこで技術用語の自動抽出，特に抽出する技術用語に漏れが無いようなシステムの開発を行った．

技術用語抽出に関して中川ら [1] は単語の出現頻度と連接頻度に基づいた複合語の抽出手法を考案している．またその実装及び検証を行ない [2]，Web 上で利用可能にしたものも存在する [3].

2 使用データ

本技術用語抽出手法（以下，本手法とする）は一般財団法人日本特許情報機構（以下，Japio と略記する）の特許関連業務の支援のため開発された．そこで，本手法における正解データとして，Japio が特許関連業務に用いている日本語技術用語集を使用した．日本語技術用語集は英日辞書から抽出した見出し語約 250 万件，及び，Japio の日本語キーワードシステムから抽出した見出し語約 150 万件である．この正解データに関して効率良く再現率の高い自動構築を行うことが本手法の目的である．

本手法では特許庁で公開されている公開特許公報を使用した．また，公開特許公報について，IPC（国際特許分類）のうち電気セクションに該当する特許データに対象を絞り対象コーパスとした．以下調査毎にコーパスの文書数，及び，公開された年を表記する．

3 言選による用語の抽出

中川ら [1] は，技術用語の自動抽出手法について次の様に述べている．

「技術用語の多くは複合名詞である事が多く，それらは少数の基本的，かつ，これ以上分割不可能な名詞（以下，単名詞とする）の組み合わせで形成されている．そこで対象コーパスの各文から形態素解析によって単語を切り出し，連続する名詞を用語候補として抽出する．次に用語候補に用語としての重要度を反映するスコア付けを行う．複合名詞は単名詞の組み合わせで表現されるので，ある単名詞に連接する別の単名詞の種類，及び，ある単名詞を含む複合名詞の出現頻度を用いることでその単名詞の重要度を表現できる．これらを組み合わせることでスコアを計算し，スコアの高いものを技術用語とする.」

[1]，[2] を元にシステムを実装した．言選 Web の抽出結果との比較を行い，ほぼ同等の技術用語を抽出した．抽出の対象コーパスは 4,461 件，公開された年は 2006 年となっている．図 1 に実験で抽出した技術用語のスコアの分布を示す。

ヒストグラムを見る限り，技術用語とそれ以外の語を分ける明確なしきい値は見られない．また本タスクではより多くの技術用語の抽出を目標としているため，再現率をより重視し重要度並びにスコア付けについては取り扱わない．

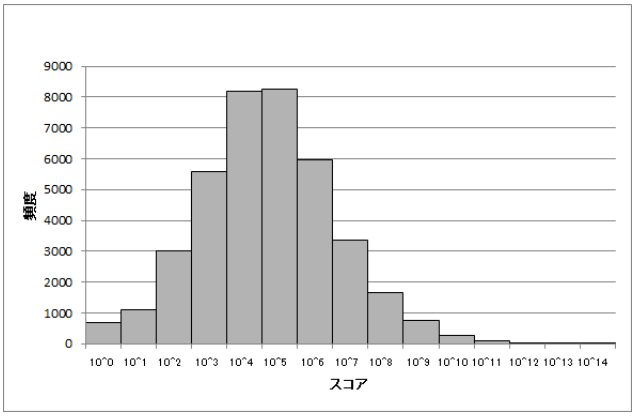

図 1: 技術用語のスコアの分布

4 改善案

4.1 削除すべき接頭辞，接尾辞の調査

予備調査において，特許コーパスから複合語を抽出し，辞書に登録されているか否かを調査した際に，複合語が辞書の語句と完全一致する割合が約 70 %，複合語の一部に辞書の語彙を含むものを合わせると約 97 ～99 %と高かった．完全一致しなかった語彙を調べると，辞書に登録されている語にノイズとなる接頭辞，及び，接尾辞が付随している．よって，抽出した複合語からそれらの不要な語を削除することで抽出精度の向上を見込める．そこで抽出した技術用語内に削除するべき語句が含まれているか否かの調査を行った．

以下に削除するべき語を含む技術用語，及び，その語の一例を示す．表 1 は接頭辞に削除するべき語がある用語の一覧，表 2 は接尾辞に削除するべき語がある用語の一覧となっている．対象コーパスは 4,461 件，公開された年は 2006 年となっている．

接頭辞としては「該当」や「次」,「前記」など特許文書内での指示語が付随した複合語が存在しそれらを削除することが有効だと考えられる．接尾辞にはあまり削除すべき語が出現せず，ABC やイロハ，数字などの文書においてのみ使用される語の削除が考えられる．

4.2 削除可能な名詞間に挿入される表現

3 で技術用語は単名詞の組み合わせによって形成されているとあった．予備調査において，技術用語を記述する際に名詞と名詞との間を接続する語句が用いられていることが確認されている．そこで名詞と名詞の間に，名詞以外の品詞の語句が存在するものを抽出し，それらの種類を分類する．接続する語句を削除するこ

表 1: 専門用語及び削除するべき語

専門用語	削除するべき単語
該当 接 部 同士 該当 エントリ 該当 下向 光 信号 該当 ハイ スループット 端末	該当
次 分割 線 次 側 検出 用 巻 線 次 ドライ エッチング 用 ガス 次 分割 端 面	次
前記 メール 処理 前記 一 酸化 炭素 除去 部 前記 多孔 質 ガラス 層 形成 工程 前記 スパイラル 露光	前記
上記 素子 終端 部 上記 板 状 フラクタル 構造 体 上記 電子 ファイル 管理 システム 上記 フッ素 系 樹脂 絶縁 体	上記
それら 磁気 抵抗 素子 それら 階 調 曲線 それら 酸化 物 透明 電極 層 それら 発光 層	それら

とで辞書に登録された技術用語になるものと，削除が不適切であるものを調査し，削除するか否かの判別を行う．

以下に複合語において名詞同士の接続に用いられていた語を示す．表 3 は名詞同士を接続する語句を含む複合語の一覧となっている．対象コーパスは 4,461 件，公開された年は 2006 年となっている．

助詞の「の」については除去し複合語を生成しても良いと考えられる．一方,「を」や「な」等の助詞や形容詞については削除した方がよい場合と，不適切な場合が混在している．する，される等は削除することで複合語を形成できるが不要な語を生成する可能性があ

表 2: 専門用語及び削除するべき語

専門用語	削除するべき単語
設計 支援 システム A 金属 元素 B 空間 群 C 設計 支援 システム C	A～Z 他記号
八木アンテナ 等 軟質 樹脂 材 等 ポリエチレン 層 等 ゲート 絶縁 膜 等	等
ネットワーク プロセッサ 自体 弾性 表面 波 素子 自体 半導体 チップ 自体 金属 水酸化物 自体	自体

るため削除処理を行わない方が良いと考えられる．

表 3: 複合語と挿入される表現

複合語	削除した場合の表現
つなぎ目 の 部分	つなぎ目部分
データ通信 用の 機器	データ通信用機器
伝送 装置の 回線	伝送装置回線
低音 発生用の 振動	低音発生用振動
データ通信 を 開始	データ通信開始
不純物 を 導入	不純物導入
不安 を 軽減	不安軽減
ブラウン管 を 映像	ブラウン管映像
データ通信 可能な 端末	データ通信可能端末
低減 可能な 構造	低減可能構造
不安定 な 状態	不安定状態
不正 な無線 通信	不正無線通信
アミノ酸 を含有する 物質	アミノ酸含有物質
データ通信 を遂行する 受信	データ通信推敲受信
フッ素 を除去する 工程	フッ素除去工程
中央 を縦貫する 中空	中央縦貫中空
イオン化 された 不純物	イオン化不純物
一体 形成された 操作	一体形成操作
上部 に形成された 領域	上部形成領域
付与 された識別 番号	付与識別番号

4.3 定型句

電気分野について選択した特許コーパスについて技術用語の抽出を行い、特許コーパスのどのタグに出現しやすいかの調査を行った．1,888 件，2010 年度公開の該当コーパスから抽出を行った．表 4 に調査結果を示す．

要約書，及び，請求項，明細書【符号の説明】で多くの語句が出現し，それらのコーパス全体での出現頻度も高かった．

しかしながら実際に抽出した語句を見ると，ほとんどを定型文に用いられる語が占めており，技術用語として有用なものはそれほど多く抽出できなかった．出現件数に関わらず調査してみると，明細書【発明を実施するための形態】，明細書【課題を解決するための手段】に技術用語が多く含まれていた．よってこれらのタグのみを抽出対象とすることで，処理の効率化と技術用語のノイズの除去に効果があると考えられる．

5 考察

これまでの調査を元に，抽出漏れの少ない技術用語の抽出手法を考案する．技術用語について構成する単

表 4: タグ別複合語抽出

複合語 出現頻度 (件数)	タグ
図　1758 提供 1260 部 799 形成 733	要約書
項 1841 請求 1840 請求項 1840 記載 1797	請求項
発明 1555 図 1552 実施 1541 説明 1526	明細書【符号の説明】
発明 27 特許 27 場合 27 使用 27	明細書 【発明を実施するための形態】
特許 28 これら 28 発明 28 場合 28	明細書 【課題を解決するための手段】

語の品詞に着目し，名詞のみで構成される語，名詞を一部含む語，名詞を一切含まない語で分類した．日本語見出し語について特許コーパス内に含まれているものを正解データとし，それぞれ抽出した語をどの程度網羅しているか調査した．なお，特許コーパスは 2010 年度の電気分野，5000 件となっている．表 5 に調査結果を示す．なお，該当する正解データは 45407 語存在した．

[1]，[2] では複合語のほとんどは名詞のみによって構成されているという前提に基づき手法を構成していた．しかしながら，本タスクの対象となる技術用語の中には名詞を一部含む複合語も一定の割合を占めており，無視できない．取得できた複合語数で比較すると，一部に名詞を含む複合語は膨大な数が取得できる一方で，有効な語数は 0.09%程度と極端に少ない．当然処理速度にも関わってくるのでこれらのノイズを再現率を落とさずに減らせればシステムとして有用と考えられる．TF-IDF で重み付けを行ない閾値でフィルタリングを試してみたが抽出精度は減少した．図 1 でのスコア付けと同様に特許文書中の専門用語には出現頻度によるアプローチは効果が低いと考えられる．

6 おわりに

電気分野について選択した特許コーパスについて技術用語の抽出を行い、抽出できなかった語についての

表 5: 複合語取得についての調査結果

	A [1]	B [2]	C [3]	D [4]
名詞のみ	115,882	36,798	81.04	31.7
名詞含む	8,117,852	7,407	16.3	0.09
名詞含まず	1,148	22	0.04	1.9
合計	8,234,882	44,224	97.38	0.05

[1] 取得できた複合語数
[2] 複合語と正解データの一致
[3] 再現率 (%)
[4] 正解データの割合 (%)

調査を行った。表 6 に抽出できなかった複合語の例を示す。対象 2010 年度の電気分野，5000 件の特許コーパスである．分野は明細書【発明を実施するための形態】とした

技術分野では慣例的にカタカナ語の末尾の長音符を省略するため，表記も混在している．両方の記述に対応した辞書を使用することで抽出できる．未知語中に含まれる語については，最長一致する語句が抽出できるべきだと考えられるので取得できなくても問題はないと考えられる．形態素分割によって抽出できない語句もまた，分割に関係なく最長一致する語句が抽出できるべきなので取得できなくても問題はないと考えられる．記号を含む語句は出現数が極端に少ないため優先度は低く，今回は取り扱わない．数値を含む複合語については請求項や図など文書に表記に使われるものがある一方，技術用語にも含まれているため抽出できる必要がある．

表 6: 抽出できなかった複合語

抽出できなかった複合語	備考
黄フィルタ 電磁波遮蔽フィルタ 非接触リーダ 液晶シャッタ	末尾の長音符が 省略されている
高周波パターン 高パフォーマンス 酸化亜鉛バリスタ 選択コンテンツ	カタカナの未知語は 最長で区切られる 例： 高周波パターンメモリ
音制御手段 電界質 電子回路基 重なり合	形態素分割によっては 取得できない
金属 - 金属 炭素 - 炭素二重結合 請求項 9 図 2	記号を含むものは取得しない
IEEE1394 RS-232 2 次元 2 値デジタル信号	数値を含む複合語

謝辞

データを提供頂き，有益な議論をして頂いた調査研究部 大塩只明氏を始めとする一般財団法人日本特許情報機構の方々に深謝する．

本研究の一部は日本学術振興会科研費（C）26330359 の支援を受けた．

参考文献

[1] 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出”, 自然言語処理,Vol.10No.1, pp. 27 - 45, 2003 年 1 月

[2] 小島浩之, 前田朗: “キーワード (専門用語) 自動抽出システムの構想とその展開”, 第 51 回日本図書館情報学会研究発表要綱, pp.17-20, 2003.10

[3] 専門用語（キーワード）自動抽出サービス「言選Ｗｅｂ」 http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html