特許文書の日英クロスリンガル wikification

綱川 隆司　　千 浩　　梶 博行

静岡大学大学院情報学研究科

{tuna,kaji}@inf.shizuoka.ac.jp
gs14051@s.inf.shizuoka.ac.jp

1 はじめに

特許の利用者や審査官は、専門用語を多く含む特許明細書を迅速に理解する必要がある。一般に、特許明細書の読者はその特許の分野に精通しているとは限らず、未知の専門用語が現れれば必要に応じて調べなければならないが、これには多大な労力を要する。

本稿では、テキスト内の語句に、それを説明する Wikipedia 記事へのリンクを自動的に付与する wikification [1] を特許明細書に対して行うことで、特許明細書中の専門用語の理解に要する労力を軽減することをねらいとする。大規模な Web 上の百科事典である Wikipedia は、一般的な概念や人名・地名などの固有名詞だけでなく、特許明細書に現れるような幅広い分野の専門用語もカバーしている。

特許明細書の wikification においては、特許の内容に関連した専門用語に対するリンクを付与すべきである。そこで本稿では、特許明細書に付与された国際特許分類 (IPC) を手掛かりに、IPC タグと関連の強い Wikipedia カテゴリをあらかじめ獲得することにより、特許明細書の分野に関連する専門用語をアンカーとして特定する方法を提案する。IPC タグと Wikipedia カテゴリの関係は、アンカーが多義語の場合に適切なリンク先記事を決定するためにも有効である。

また、Wikipedia は多言語百科事典であり、各言語版は独立して編集されていて規模が異なる。英語版は日本語版の 5 倍以上の記事を持ち、専門用語に関する記事もより充実している。そこで、リンク先とする Wikipedia 記事を日本語版だけでなく英語版も対象とすることにより、カバーできる専門用語の範囲を広げることを試みる。このようにテキストと異なる言語の Wikipedia 記事へのリンク付けはクロスリンガル wikification [2] と呼ばれる。本研究では専門用語を特許の日英パラレルコーパスから得た対訳対を用いて日本語に訳して行う方法を検討する。

2 関連研究

Wikification は、アンカーテキスト（テキスト中でリンク元となる語句）の特定と、各アンカーテキストのリンク先記事の決定の二つのステップから構成される。アンカーテキストの特定には、アンカーテキストとなる語句が入力テキストの中で重要かどうか判断することが必要である。重要性の主な指標として、Wikipedia 全体において語句がリンクのアンカーテキストになっている確率（キーフレーズネス [1]）が挙げられる。リンク先記事の決定には、アンカーテキストの語義曖昧性解消 [3] が必要となる。そのアンカーテキストから各リンク先記事候補にリンクされる確率 [4]、周辺文脈の類似度 [1]、および、周辺リンクのリンク先記事との関連性 [5] が曖昧性解消に有効な特徴である。

3 提案方法

3.1 基本アイデア

図 1 は特許明細書の wikification の例を示しており、専門用語である“アーク”や“アーク抑制”についてその意味を説明する適切な記事へのリンクを付与している。“アーク抑制”については、それを直接説明する日本語版記事が存在しないため、英語版記事へリンクしている。Wikification の結果はリンクのリストであり、一つのリンクはアンカーテキストとリンク先記事の組で表される。

Wikification におけるアンカーテキストの特定およびリンク先記事の決定のいずれにおいても、入力テキストに現れる語句とリンク先記事の関連性 [5] は有効である。テキストの内容に関連した記事はリンク付けする重要性が高く、かつ、アンカーテキストが曖昧であっても内容と関連する意味の記事をリンク先に選ぶことで適切なリンク先が決定できることが期待される。関連性の判定には、テキストが Wikipedia 記事の場合においては、既存リンクのリンク先記事の類似性を用いる方法が用いられている。特許明細書の場合、そのような従来方法は適用できないが、特許明細者が国際特許分類により分類されて

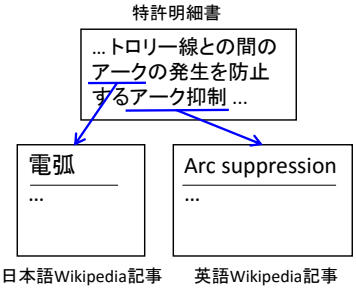


図 1　特許明細書の wikification

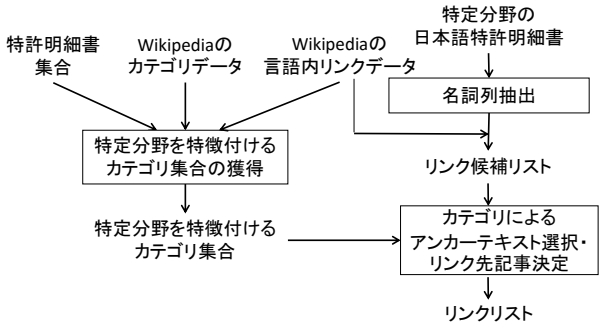


図 2　提案方法の概要

いることに着目した。特許コーパスを利用して、IPC タグと関連の強い Wikipedia カテゴリを獲得しておくことにより、特許明細書→IPC タグ→Wikipedia カテゴリ→Wikipedia 記事のルートで関連性を判定することが提案方法の特徴である。

3.2　提案方法の概要

図 2 に提案方法の概要について示す。予め、特許明細書集合から特定分野を特徴付けるカテゴリ集合を獲得しておく。入力とする同じ分野の特許明細書から名詞列を抽出し、Wikipedia のリンクデータからリンク候補リストを生成する。リンク候補リストのうち、特定分野を特徴付けるカテゴリに関連するもののみを残すことでリンクのリストを生成し出力する。

特許明細書には、1 つ以上の国際特許分類 (IPC) が付与されている。全技術分野は A～H の 8 つのセクションに分類されており、各セクションは 2 桁の数字で示されるクラスに分類され、以下階層的に分類されている。例えば、G06 で始まる IPC を持つ特許は“計算;計数”クラスに分類されている。

一方で、Wikipedia 記事は、その記事に関連する 1 つ以上の Wikipedia カテゴリに対応付けられている。Wikipedia カテゴリは、最上位カテゴリ（日本語版では“主要カテゴリ”）から体系化して整理されている。そこで、国際特許分類のある分野に頻出する Wikipedia カテゴリを、その分野を特徴づけるカテゴリとして獲得する。以下の各節で、特定分野を特徴づける Wikipedia カテゴリの獲得およびリンクリストの出力について詳しく説明する。

3.3　特許の特定分野を特徴付ける Wikipedia カテゴリの獲得

特許の特定分野を特徴づける Wikipedia カテゴリは、文書を特徴づける語句の重み付け方法の一つである tf-idf に基づいて獲得する。語句の tf-idf 値は語句の文書内の出現頻度 (tf) と逆文書頻度 (idf) の積で求められる。これに倣い、特許のある分野に出現する各カテゴリの出現頻度 (cf) と逆文書頻度 (idf) を求め、それらの積をその分野に対するカテゴリの重みとする。このとき、カテゴリは特許明細書に直接出現するわけではないことから、次のように頻度を推定する。

まず、ある分野の特許明細書集合を抽出し、各明細書 d に出現する全ての名詞列 s ($\in d$) のうち、Wikipedia においてアンカーテキストとして用いられたことのあるものをアンカーテキスト候補として出現頻度 $\mathrm{freq}_d(s)$ とともに列挙する。アンカーテキスト s が記事 a をリンクする確率 $\Pr(a \mid s)$ を予め Wikipedia から求めておく。

ある特許明細書 d におけるカテゴリ c の出現頻度 $\mathrm{cf}_d(c)$ を、以下の式で推定する。

$$\mathrm{cf}_d(c) = \sum_{s \in d} \left(\mathrm{freq}_d(s) \times \sum_{a \in A(c)} \Pr(a|s) \right).$$

同様に、カテゴリ c の文書頻度は、カテゴリがある明細書 d に出現する尤度 $\mathrm{df}_d(c)$ の和であり、$\mathrm{df}_d(c)$ は以下の式で求める。

$$\mathrm{df}_d(c) = \max_{s \in d} \sum_{a \in A(c)} \Pr(a|s).$$

これらを用いて、カテゴリ c が特定の分野 t を特徴づける重み cf-idf(c) を以下の式で求める。

$$\text{cf-idf}_t(c) = \sum_{d \in D(t)} \mathrm{cf}_d(c) \times \log \frac{|D|}{\sum_{d \in D} \mathrm{df}_d(c)},$$

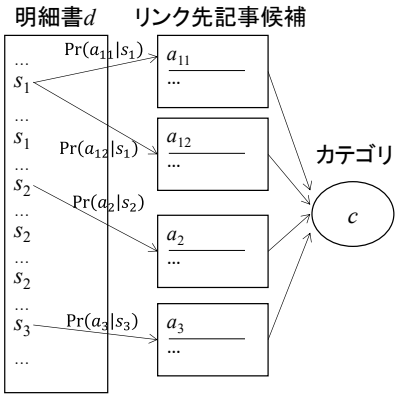
ここに、$D(t)$ は分野 t の特許明細書集合、D はすべての特許明細書の集合とする。

図 3 は、ある特許明細書 d にアンカーテキスト s_1, s_2, s_3 が現れ、各アンカーテキストがカテゴリ c と図のように対応しているときのカテゴリ c の出現頻度 $\mathrm{cf}_d(c)$ および文書頻度 $\mathrm{df}_d(c)$ の計算例を示している。

Wikipedia カテゴリすべてについて cf-idf 値を計算し、cf-idf 値が上位 θ %のカテゴリを分野 t を特徴づけるカテゴリ集合として得る。

3.4　リンクリストの出力

入力した分野 t の特許明細書に対して、wikification 結果となるリンクリストを出力する。まず、入力した特許



$$\mathrm{cf}_d(c) = 2 \times (\Pr(a_{11}|s_1) + \Pr(a_{12}|s_1)) + 3 \times \Pr(a_2|s_2) + 1 \times \Pr(a_3|s_3)$$

$$\mathrm{df}_d(c) = \max(\Pr(a_{11}|s_1) + \Pr(a_{12}|s_1), \Pr(a_2|s_2), \Pr(a_3|s_3))$$

図 3　カテゴリの出現頻度・文書頻度の推定

明細書から、3.3 節と同様にアンカーテキスト候補を抽出する。各アンカーテキスト s がリンクする可能性のある記事を列挙し、各記事が属するカテゴリの中で、分野 t に対する cf-idf 値が最大のカテゴリを選択する。このカテゴリに属する記事 a の中で、$\Pr(a \mid s)$ の値が最大のリンク先記事をそのアンカーテキストのリンク先とするリンクを出力する。一方、各記事が属するカテゴリの中に分野 t を特徴付けるカテゴリが一つも含まれていない場合は、そのアンカーテキスト候補のリンクを出力しない。

さらに、英語記事へのリンクを付与するため、アンカーテキスト候補のうち、対応する日本語記事がないものについて、日英特許パラレルコーパスから得た日英対訳フレーズ対を用いて英語に翻訳する。英訳したアンカーテキスト候補について、対応する英語版記事を列挙する。英語版記事は英語版の Wikipedia カテゴリに属しているため、英語特許と英語版 Wikipedia を用いて分野 t を特徴付ける英語版カテゴリを求め、日本語版と同様の方法でリンクリストを出力する。

4　評価実験

4.1　実験設定

本実験では、特許の特定分野を特徴付ける Wikipedia カテゴリを求め、当該分野の特許明細書の wikification を行う。実験に用いた特許データは NTCIR-7 PATMT テストコレクションから得た。Wikipedia は 2013 年 3 月時点のダンプデータを用い、カテゴリは隠しカテゴリ以外のすべてのカテゴリを対象とした。

特定分野として、IPC タグが G06 で始まる“計算;計数”クラスを採用した。分野 G06 を特徴付けるカテゴリを求めるため、2000 年に出願された日本の特許明細書のうち、

表 1　分野 G06 を特徴付ける日本語カテゴリ

順位	カテゴリ	cf-idf ($\times 10^3$)
1	資料学	1131
2	コンピュータのデータ	1116
3	コンピュータグラフィックス	980
4	画像処理	972
5	コンピュータの仕組み	966
6	情報学	774
7	コンピュータネットワーク	663
8	コンピュータのユーザインタフェース	616
9	記憶装置	576
10	ラジオの情報・ワイドショー番組	516
…	…	…
25	生態域	373

分野 G06 のものを 10000 件、任意の分野のものを 10000 件、それぞれ無作為に抽出した。また、英語版カテゴリを求めるため、2001 年に出願された米国特許明細書を同様に 5000 件ずつ抽出した。特許 Wikification のテストセットとして、2001 年出願の分野 G06 の日本語特許明細書 15 件を抽出し、人手で日本語記事へのリンクを付与した。

特許明細書からアンカーテキスト候補を抽出するため、MeCab[1]による形態素解析を行い、名詞（非自立等は除く）と接頭詞からなる単語列をすべて抽出した。また、アンカーテキスト候補を英訳するため、上記のコレクションに含まれる日英特許パラレルコーパス [6] から抽出された日英対訳フレーズテーブルを用いた。

特定分野を特徴づけるカテゴリの閾値 θ については、$\theta = 10$ (%) を用いた。

4.2　特許の特定分野を特徴づける Wikipedia カテゴリの獲得

表 1 に、分野 G06 について cf-idf 値が上位となった日本語 Wikipedia カテゴリを示した。1 位のカテゴリは“資料学”であり、これは当該分野で頻出する語“データ”のリンク先記事“データ”が属しているカテゴリである。以下、上位にはコンピュータ関連のカテゴリが並んでおり、当該分野にコンピュータを用いたシステムの特許が多いことを反映していると考えられる。

25 位のカテゴリ“生態域”は、当該分野の頻出語“ステップ”から得られたと考えられる。段階・手順を示す一般語である“ステップ”に関する記事はなく、リンク先記事候補として“ステップ (植生) ”があるため、その記事が属するカテゴリの cf-idf 値が上昇した。この結果、特許明細書中に現れる“ステップ”はすべて記事“ステップ (植生) ”にリンクされてしまい、不適切な結果となる。この問題に対処するためには、カテゴリの出現頻度

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

を求める際に単独のアンカーテキストから得られたカテゴリを無視する、文脈に応じたリンク先記事の選択を行う、といった方法が考えられる。

4.3 特許明細書の wikification 結果

表 2 に、提案方法によって得られたリンクに対して、テストセットのリンクと比較した結果を示した。得られたリンクがテストセットに含まれないケースが多いが、これはテストセットに比べ、提案方法は一般語に近い“情報”のような語にもリンクしているためである。

表 3 に、ある特許明細書中のアンカーテキスト“インタフェース”について、リンク先記事候補の一部と記事が属するカテゴリおよびその cf-idf 値を示した。人手で付与したリンク先記事は“インタフェース (情報技術)”であるが、提案方法では最も大きい cf-idf 値をもつカテゴリ“コンピュータグラフィックス”に属する記事“グラフィカルユーザインタフェース”を選択した。本提案方法は分野のみに依存してリンク先記事を決定するため、この例のようにより細かい分類が必要な曖昧性解消を行うには、国際特許分類の細分類（サブクラスなど）から得られた cf-idf 値を組み合わせる、あるいは文脈情報など他の wikification 手法と組み合わせるといった改善法が考えられる。

テストセット上で提案方法によって得られた英語記事へのリンクの例を表 4 に示す。一つ目の例は、“セルラ”という語からは適切な日本語記事が見つからないため、英訳して “cellular” にすることで携帯電話を表す記事と対応付いたもので、概ね適切である。二つ目の例は、データ転送の記事が選択されており、記事中で非同期転送についての記述があることから、関連した記事に対応付けることができたものである。一方で、アンカーテキストを英訳した時点で異なる意味の記事に対応しやすくなるために不適切な記事が選ばれる例も散見された。

5 おわりに

本稿では、特許明細書中の専門用語の理解を容易にするため、明細書の内容とリンク先記事の関連性を国際特許分類と Wikipedia カテゴリの対応付けから得ることによるクロスリンガル wikification 方法を提案した。

今後の課題として、一般語の頻出から得られる不適切なカテゴリの除去が挙げられる。このようなカテゴリは同一明細書中の他のカテゴリとの関連性が低いため、カテゴリ間の関連性を求めることで除去できる可能性がある。また、リンクリスト生成のときにも他のリンクのカテゴリとの関連性が高いものを優先的に選ぶことで入力特許明細書ごとに適した記事を選ぶことも考えたい。

表 2 提案方法で得られたリンク数

分類	リンク数
テストセットと一致	302
テストセットとリンク先記事が異なる	144
テストセットに含まれない	4055
テストセットのみに含まれるリンク	204

表 3 アンカーテキスト“インタフェース”に対するリンク先記事の選択

リンク先記事候補	記事が属するカテゴリ	cf-idf ($\times 10^3$)
インタフェース (情報技術)	ソフトウェア	497
	電子工学	207
	インタフェース規格	121
	インタフェース	104
グラフィカルユーザインタフェース	コンピュータグラフィックス	**970**
	コンピュータのユーザインタフェース	589
	グラフィカルユーザインタフェース	161
	ソフトウェアアーキテクチャ	39

表 4 提案方法で得られた英語記事へのリンクの例

特許明細書（下線はアンカーテキスト）	提案方法で得られた英語リンク先記事
…この発明は、例えばセルラ無線通信システムの加入者が…	Mobile phone
…シードのアシンクロナスパケットを受信して…	Data transmission

謝辞

本研究は JSPS 科研費 15K16096 の助成を受けたものです。本研究を進めるにあたり、NTCIR データセットに含まれる日英特許パラレルコーパスから構築した日英対訳フレーズテーブルをご提供頂いた筑波大学宇津呂武仁教授および山本幹雄教授に深く感謝致します。

参考文献

[1] Mihalcea, R. and Csomai, A. (2007), “Wikify!: linking documents to encyclopedic knowledge,” in *Proc. of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 233–242.
[2] McNamee, P., Mayfield, J., Lawrie, D., Oard, D.W., and Doermann, D. (2011), “Cross-language entity linking,” in *Proc. of the 5th International Joint Conference on Natural Lan-guage Processing (IJCNLP)*, pp. 255-263.
[3] Navigli, R. (2009), “Word sense disambiguation: a survey," *ACM Comput. Surv.*, 41(2):10:1-10:69.
[4] Milne, D. and Witten, I.H. (2008), “Learning to link with Wikipedia,” in *Proc. of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 509–518.
[5] Ratinov, L., Roth, D., Downey, D. and Anderson, M. (2011), “Local and global algorithms for disambiguation to Wikipedia,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 1375–1384.
[6] Utiyama, M. and Isahara, H. (2007), “A Japanese-English patent parallel corpus,” in *Proceedings of Machine Translation Summit XI*, pages 475-482.