付与根拠箇所推定に基づく特許文書へのFターム付与

榊原 隆文† 笹野 遼平† 高村 大也† 目黒 光司‡ 奥村 学†
†東京工業大学 ‡特許庁
tsakaki@lr.pi.titech.ac.jp, meguro-koji@jpo.go.jp,
{sasano,takamura,oku}@pi.titech.ac.jp

1 はじめに

日本の特許庁に出願された特許文書には，検索などの用途のためにFタームという分類が人手で付与される．Fタームはテーマコード (e.g., 2H033) と，英字二桁の観点と数字二桁[1](e.g., AA01) で構成される．そのうち，テーマコードは特許文書の属する技術分野を表しており，観点はその技術分野を目的，機能，材料，用途などの切り口で細分化している．現在，日本の特許庁に出願される特許文書の数は年間30万件以上[2]であり，Fタームを人手で付与するコストは大きい．そこで本研究では，特許文書にFタームを自動付与する技術に焦点を当てる．

Fタームの自動付与に関する研究はいくつか存在する [3, 4]．これらの研究では特許文書全体をひとまとまりにして，あるFタームがその文書に付与されるかどうかを機械学習を用いて自動分類している．しかしながら，Fタームは文書単位で付与される分類指標ではあるが，あるFタームが付与されている特許文書内の全ての記述がそのFタームに関連しているわけではない．例えば図1に示した例では,【請求項1】と【0001】段落と【0002】段落はFターム CA02(形態素解析) と関連しているが，他の箇所はほとんど関連していない．したがって，そのFタームに関連している可能性が高い部分の情報を重視することにより自動付与の性能が高くなると考えられる．そこで，本研究ではFタームの付与根拠箇所を推定し，自動付与の手がかりとして利用する手法を提案する．

付与根拠箇所を考慮して特許文書の自動分類を試みた研究として，笹野 [5] や小林 [6] の研究がある．これらの研究では，人手でアノテートされた付与根拠データを利用して特許分類の性能向上を試みている．一方，本研究ではこのようなアノテートされたデータがない場合にも適用可能な手法を提案する．

[1]本研究では，観点と数字の部分のみを指して**F**タームと呼ぶ．
[2]https://www.jpo.go.jp/shiryou/toushin/nenji/nenpou2016/toukei/0101.pdf

テーマコード　5B091(機械翻訳)
Fターム　AA03(2 言語間) CA02(形態素解析) CA05(構文解析)
【特許請求の範囲】
【請求項 1】コンピュータに，自然言語の文を入力する入力ステップと，該入力ステップで入力された入力文の形態素解析を行う形態素解析ステップと，……を実行することを特徴とする情報処理装置．
【請求項 2】前記解析ステップは，……ことを特徴とする請求項 1 に記載の構文解析プログラム．
……
【発明の詳細な説明】
【0001】本発明は，コンピュータにより自然言語の単語分割を行う形態素解析プログラム，コンピュータにより自然言語の構文を解析する構文解析プログラム，……に関する．
【0002】形態素解析ステップでは，取得したテキストデータを単語ごとに分解する．
この手法としては既知の形態素解析法を用いる．
【0003】主語述語抽出ステップでは，取得したテキストデータの係り受け関係を解析し，テキストデータ中の主語と述語を抽出する．この分析の手法としては，既知の構文解析法を用いる．
【0004】並べ換えステップでは，原言語の形態素を目的言語の語順に近くなるように並べ換える．
……

図 1: 特許文書の例

2 提案手法

2.1 手法の概要

本研究では，特許文書へのFタームの自動付与タスクを，あるテーマコードに属する特許文書が与えられたときに，各Fタームが対象の特許文書に付与されるかどうかを判定する2値分類問題として扱う．以降，本稿では対象のFタームが付与された文書を**正例文書**，それ以外の文書を**負例文書**と呼ぶことにする．

本研究では特許文書の項目のうち,「特許請求の範囲」とその説明である「発明の詳細な説明」の項目に記載された内容を利用する．図1に示したように,「特許請求の範囲」の項目は請求項という単位で分割されており,「発明の詳細な説明」の項目は段落の単位で分割されている．本稿では，各単位をいずれも単に段落と呼ぶことにする．

本稿の研究課題は Multiple-instance 学習の枠組みで捉えることができる．Multiple-instance 学習は Dietterich ら [2] が提唱した概念で，複数の事例から構成される集合のラベルを教師あり学習で推定する問題である．Multiple-instance 学習では，それぞれの

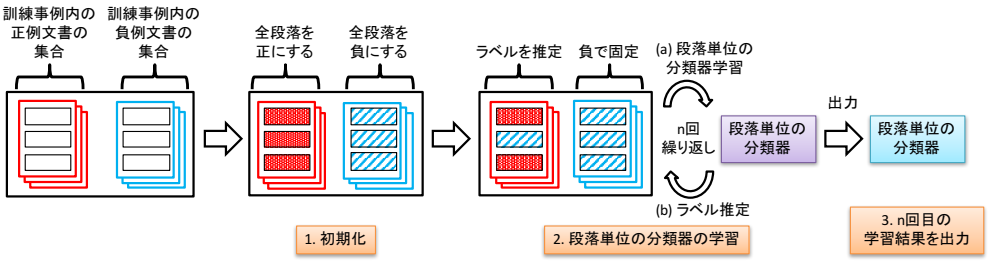

図 2: 段落単位の分類器学習の概要

事例に隠れたラベルが存在し，集合内の事例のラベルによって集合全体のラベルが決定される．具体的には，集合内に 1 つでも正のラベルを持つ要素があるとき集合のラベルは正となり，それ以外では負となる．本研究で扱う特許文書に対する F タームは，ある F タームに関する記述が特許文書中に 1ヶ所でも存在していれば文書全体にその F タームが付与されるため，Multiple-instance 学習での集合が特許文書に対応し，集合内の各事例が段落に対応すると考えることができる．

本研究の手法は，まず各段落が対象としている F タームの付与根拠箇所かどうかを推定するために，段落単位の分類器を学習する．次に，学習した段落単位の分類器を用いて文書内の各段落が F タームの付与根拠箇所であるかどうかを推定する．その上で，段落単位の推定の結果を利用して，文書全体に F タームが付与されるかどうかを次の 2 つの方法のいずれかを用いて推定する．1 つ目の方法は付与根拠箇所と推定された段落が存在するかどうかで一律に決定する方法で，2 つ目の方法は付与根拠箇所と推定された段落から得られる素性を手がかりとして文書単位の分類器を学習して推定する方法である．これらの詳細は 2.3 節で説明する．

なお，本稿では，F タームの付与根拠箇所であると推定された段落を**正のラベルを持つ段落**，そうでない段落を**負のラベルを持つ段落**と呼ぶことにする．また，文書に F タームが付与されるかどうかを推定した結果を**文書ラベル**と呼び，文書に F タームが付与されると推定された場合，文書ラベルを正と呼び，そうでない場合，文書ラベルを負と呼ぶ．

2.2 段落単位の分類器の学習

本研究では Multiple-instance 学習の手法のうち，基本的に Andrews らが提案した mi-SVM [1] を用いる．mi-SVM は，「正のラベルを持つ集合内には正のラベルを持つ事例が少なくとも 1 つ含まれ，負のラベルを持つ集合内には正のラベルを持つ事例が 1 つも含まれない」という制約を満たすように，集合内の事例単位の分類器を繰り返し学習する手法である．分類器の学習のために用いるラベルは観測できないため，まずは正例文書内の段落ラベルは全て正に，負例文書内の段落ラベルは全て負に初期化した後，正例文書内の段落ラベルを再帰的に更新していく．負例文書に含まれる段落のラベルは制約により全て負であるため，負例文書に含まれる段落のラベルは更新しない．図 2 に段落単位の分類器学習の概要を，以下に手順をそれぞれ示す．

1. 訓練事例の正例文書内の段落ラベルを全て正に，負例文書の段落ラベルを全て負に初期化
2. 以下の手順を n 回繰り返す
   (a) 現状の段落ラベルを利用して段落単位の分類器を SVM で学習
   (b) 2-(a) で学習した分類器を用いて，訓練事例内の正例文書中の各段落のラベルを推定
3. n 回目に学習された分類器を段落単位の分類器として出力

2-(b) において，正例文書内の段落でラベルが負と推定された段落は以降の処理から除外する[3]．ただし，全段落のラベルが負と推定された正例文書があった場合は，SVM の出力[4]が最も大きい段落のラベルを正と推定されたものとして扱う．また，Andrews らの手法では集合内の各事例のラベルの変化がなくなるまでイテレーションを繰り返しているが，本研究ではイテレーションの回数 n をパラメータとし，開発データを用いて最適なパラメータを決定する．

2.3 文書ラベルの推定

与えられた文書全体のラベルを決定する方法として，学習した段落単位の分類器で，与えられた文書に含まれる各段落のラベルを推定し，正のラベルと推定された段落が文書内に 1 つでも存在する場合に文書ラベルを正に，文書内に正のラベルと推定された段落が 1 つも存在しない場合に文書ラベルを負にするという方法が考えられる．これは，Andrews らの研究において文

[3] Andrews らの研究 [1] では，正例文書内の段落のうち，ラベルが負と推定された事例は負の訓練事例として利用している．

[4] SVM が学習した重み $\boldsymbol{w}$ と段落の素性ベクトル $\boldsymbol{x}$ の内積にバイアス項 b を足した値．

書ラベルを決定する方法と同じである．しかし，この方法では文書内の情報を局所的にしか用いることができていない．そこで，本研究では付与根拠箇所であると推定された段落から得られる素性を分類の手がかりとして用いて文書単位の分類器を学習し，未知の文書ラベルを推定する手法を提案する．

具体的な手順は次の通りである．まず，訓練事例内の全文書内の段落のラベルを 2.2 節で学習した段落単位の分類器で推定する[5]．その上で，各文書から学習で使用する素性ベクトルを生成する際に，文書全体から生成される素性ベクトルに加えて，ラベルが正であると推定された段落だけからも同様の素性ベクトルを生成し，前者と区別して扱うことで，文書単位の分類器を学習する．

未知の文書ラベルを推定する際には，まず未知の文書内の段落ラベルを段落単位の分類器で推定する．次に，文書全体から生成される素性ベクトルに加えて，ラベルが正であると推定された段落だけからも同様の素性ベクトルを生成し，前者と区別して扱うことで，文書単位の素性ベクトルを生成する．その上で，生成した文書単位の素性ベクトルに対して先に学習した文書単位の分類器を適用することで未知の文書の文書ラベルを推定する．

多くの F タームは，それが付与されている文書より付与されていない文書のほうが多いため，通常の方法で学習すると負のラベルばかりが出力されるように学習されてしまう．このようなデータの不均衡に起因する問題を解決するため，本研究では文書単位の分類器の学習の際に，訓練事例内の負例文書数のほうが正例文書数より多い場合には，負例文書数を正例文書数で割った値で正例のクラスに重みを付けるようにした[6]．

2.4 素性

段落単位の分類器と文書単位の分類器を学習するための素性として，形態素ユニグラムと形態素バイグラムの bag-of-words を 2 値素性として用いた．計算時間の削減のため，文書頻度が 5 以上のユニグラム，バイグラムのみ使用した．形態素解析には MeCab[7] バージョン 0.996 を用い，辞書は mecab-ipadic-2.7.0-20070801 を使用した．また，形態素ユニグラムは名詞，動詞，形容詞，連体詞の 4 種類の品詞の形態素のみ使用した．

2.5 調整が必要なパラメータ

提案手法では 4 種類のパラメータの調整が必要となる．具体的には，段落単位の分類器を学習する際の SVM の正則化パラメータ (para_c)，段落のラベルを決定するための閾値 (θ)，文書単位の分類器を学習する際の SVM の正則化パラメータ (doc_c)，mi-SVM のイテレーション回数 (n) を開発データを用いて調整する．これらのパラメータのうち，para_c は 0.000001 から 0.001 まで 10 倍刻み，θ は-0.6 から-0.3 まで 0.1 刻み，doc_c は 0.000001 から 0.001 までの 10 倍刻みで探索した．また，イテレーション数 n は，他のパラメータを固定した状態で順々に増やしていき，開発データでの F 値が下がったところで探索を打ち切っている．

3 実験

3.1 実験設定

本研究では，テーマコード 2H033，2C056，3H130，3K107，3K243，4C083，4C167，4H011，4K029，5C164，5F151，5K030，5K033，5K201 の 14 のテーマコードに属する特許文書のうち，2005 年から 2014 年の 10 年間に出願された特許文書を使用した．このうち，2005 年から 2011 年の 7 年間の特許文書を訓練データ，2012 年の特許文書を開発データ，2013 年と 2014 年の文書を評価データとして使用した．使用する F タームは訓練事例中に F タームが付与された文書が 700 文書 (1 年あたり 100 文書) 以上ある F ターム[8]の中から 20 個を無作為に抽出した[9]．なお，AA00 のように F タームの数字が 00 である F タームは除外した．評価尺度には F 値のマクロ平均とマイクロ平均を用いた．

3.2 比較手法

実験では次の 3 つの手法を比較する．1 つ目は，段落単位のラベル推定をせず，文書内に含まれる情報を素性化して SVM で学習する手法 (SVM_doc) である．SVM の学習の際，提案手法と同様の方法で正例のクラスに重みを付ける．2 つ目は，mi-SVM で学習した段落単位の分類器のみを用いて文書のラベルを決定する手法 (mi-SVM) である．すなわち，学習した段落単位の分類器で評価データの文書に含まれる段落のラベルを推定し，文書内に 1 つでも正のラベルと推定した段落が存在した場合，文書全体のラベルを正とする手法である．これは，Andrews らの研究でのラベル決定の方法と同じである．3 つ目は，段落単位の推定結果に基づく文書単位の分類器を用いる手法 (mi-SVM$^+$) である．

[5] SVM が学習した重み $\boldsymbol{w}$ と段落の素性ベクトル $\boldsymbol{x}$ の内積にバイアス項 b を足した値が，設定した閾値よりも高いかどうかでラベルを決定する．

[6] 実装としては，LIBLINEAR の-w オプションで指定した．

[7] http://taku910.github.io/mecab/

[8] F タームには階層構造が存在しているが，本研究では下位の F タームが与えられている場合，機械的にその上位の F タームも付与されているものとして扱った．

[9] ただし，テーマコード 3K243 に関しては条件を満たす F ターム数が 12 個だったため，12 個の F タームのみを使用した．

テーマ	マクロ平均			マイクロ平均		
コード	SVM_doc	mi-SVM	mi-SVM$^+$	SVM_doc	mi-SVM	mi-SVM$^+$
2H033	0.514	0.500	**0.531**	0.568	0.557	**0.579**
2C056	0.472	0.450	**0.505**	0.559	0.562	**0.598**
3H130	0.544	0.530	**0.565**	0.682	0.633	**0.684**
3K107	0.590	0.538	**0.619**	0.652	0.607	**0.671**
3K243	**0.546**	0.504	0.539	0.655	0.633	**0.661**
4C083	0.570	0.525	**0.580**	0.631	0.584	**0.636**
4C167	0.501	0.490	**0.512**	**0.595**	0.569	0.594
4H011	0.591	0.550	**0.605**	0.659	0.612	**0.669**
4K029	0.665	0.660	**0.671**	0.710	0.706	**0.715**
5C164	0.429	0.397	**0.448**	0.436	0.420	**0.452**
5F151	0.606	0.591	**0.609**	0.659	0.650	**0.665**
5K030	0.428	0.421	**0.450**	0.472	0.479	**0.496**
5K033	0.409	0.396	**0.421**	0.444	0.445	**0.474**
5K201	0.474	0.461	**0.494**	0.569	0.553	**0.585**
平均	0.524	0.501	**0.539**	0.592	0.572	**0.606**

表 1: F ターム自動付与実験の結果

3.3 パラメータの探索

SVM_doc で用いるパラメータは文書単位の分類器を学習する際の SVM の正則化パラメータである．このパラメータは 0.000001 から 0.001 まで 10 倍刻みで探索した．mi-SVM で用いるパラメータは 2.5 節で説明した mi-SVM$^+$ のパラメータのうち，para_c と θ と n の 3 種類である．文書ラベルの決定方法が mi-SVM$^+$ とは異なっていることから，開発データにおける F 値も異なるため，別途パラメータ探索を行った．para_c は 0.00001 から 0.01 まで 10 倍刻み，θ は-0.3 から 0.1 まで 0.1 刻みで探索した．イテレーション数 n は mi-SVM$^+$ と同様に探索する．mi-SVM$^+$ のパラメータは 2.5 節で説明した通りである．ただし，計算時間の削減のため，para_c は全テーマコードで共通する値を使用することとし，その値はテーマコード 2H033 の開発データに対する F 値のマクロ平均に基づき決定した．また，SVM の実装は LIBLINEAR[10]を利用した．

3.4 実験結果と考察

実験結果を表 1 に示す．全体として，mi-SVM$^+$ が最も良い性能となっており，付与根拠箇所を考慮することで分類性能が向上することが確認できた．一方，SVM_doc と mi-SVM を比較すると，SVM_doc のほうが良い性能を示した．この結果は，段落ごとのラベルを推定するメリットよりも，局所的な情報しか考慮できないデメリットのほうが大きかったことを示唆していると考えられる．

推定した付与根拠箇所を評価するために，専門知識を持つアノテータ 2 名[11]によってテーマコード 2H033 と 5K033 に属する特許文書 10 件ずつに対して付与根拠箇所のアノテーションを行い，段落ラベルの評価を行った．評価尺度は F タームごとに求めた再現率，適合率，F 値のマイクロ平均を用いた．結果を表 2 に示す．参考のため，アノテートされた正解データの段落ラベルを無作為に入れ替えたもの (shuffle) に対しても値を算出した．mi-SVM$^+$ はいずれの評価尺度でも mi-SVM よりも良い結果となった．しかし，この結果は，文書中の大半の段落が付与根拠であるとアノテートされた文書の影響を大きく受けたものである可能性が考えられる．そこで，マクロ平均による評価も行った．再現率と適合率[12]のマクロ平均の結果は表 2 の右側に示している．mi-SVM と mi-SVM$^+$ を比較すると，適合率は mi-SVM が mi-SVM$^+$ よりも高く，再現率はその逆である．これは，mi-SVM は 1 つでも正のラベルを持つ段落があれば文書のラベルも正と判定することから厳しめの閾値が採用されるのに対し，mi-SVM$^+$ は段落ラベルの情報を素性として利用するため，できるだけ多くのラベルを出力する傾向があることと対応していると考えられる．

	マイクロ平均			マクロ平均	
手法	R	P	F	R	P*
mi-SVM	0.203	0.234	0.218	0.225	**0.382**
mi-SVM$^+$	**0.640**	**0.271**	**0.380**	**0.577**	0.294
shuffle(参考)	0.374	0.374	0.374	0.178	0.202

表 2: 段落ラベルの評価

4 まとめ

本研究では F タームの付与根拠箇所を考慮して特許文書に F タームを自動付与する手法を提案した．実験の結果，F タームの付与根拠箇所を推定し，その結果を文書単位の分類器の素性に反映することで，自動分類の性能が向上することが確認できた．また，本手法を用いることで，特許文書中の各段落がどの F タームと関連性が高いかを提示することが可能であり，人手による特許分類業務の補助への応用が考えられる．

参考文献

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support Vector Machines for Multiple-Instance Learning. In *Proc. of NIPS'02*, pages 561–568.

[2] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997.

[3] Kazuya Konishi and Toru Takaki. F-term Classification System Using K-Nearest Neighbor Method. In *Proc. of NTCIR-6*, 2007.

[4] Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. SVM Based Learning System for F-term Patent Classification. In *Proc. of NTCIR-6*, 2007.

[5] 笹野 秀生. 特許分類の自動推定に向けた取り組み—機械学習による自動分類技術の特許文献への適用—, 2012. Japio YEAR BOOK 2012.

[6] 小林 英司. 特許分類の自動推定に向けた取り組み—機械学習による自動分類技術の実用化に向けて—, 2013. Japio YEAR BOOK 2013.

[10] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[11] それぞれのアノテータがテーマコード 1 つずつを担当した．

[12] システムが 1 つも正のラベルを出力していない場合には適合率は算出できないため，適合率のマクロ平均を計算する際にはどの手法を用いても 1 つ以上の正のラベルが出力された F タームと文書の組のみを利用して評価を行った．表 2 では P*と表記した．