特許請求項と詳細説明の自動対応付け

広島市立大学大学院情報科学研究科准教授 **難波　英嗣**

PROFILE 1996 年東京理科大学理工学部電気工学科卒業。2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。同年、日本学術振興会特別研究員。2002 年東京工業大学精密工学研究所助手。同年、広島市立大学情報科学部講師。2010 年広島市立大学大学院情報科学研究科准教授。現在に至る。博士（情報科学）。言語処理学会、情報処理学会、人工知能学会、ACL、ACM 各会員。

広島市立大学大学院情報科学研究科修士課程 **畝田　将登**

PROFILE 2011 年広島市立大学情報科学部知能情報システム工学科卒業。同年、広島市立大学情報科学研究科知能工学専攻修士課程入学、現在に至る。

広島市立大学大学院情報科学研究科教授 **竹澤　寿幸**

PROFILE 1984 年早稲田大学理工学部電気工学科卒業。1989 年早稲田大学大学院博士後期課程修了。同年（株）国際電気通信基礎技術研究所入社。2007 年広島市立大学大学院情報科学研究科教授、現在に至る。工学博士。音声対話翻訳の研究開発に従事。平成 18 年度電子情報通信学会 ISS 論文賞受賞。電子情報通信学会、人工知能学会、日本音響学会、言語処理学会各会員。

1 はじめに

近年、特許出願件数の増加にともない、知的財産の非専門家でも特許を検索したり、出願したりするケースが増えてきている。新しく特許を出願する際、過去に同じ技術を用いた特許が出願されていないか調査する必要がある。この調査を行うには、発明を特定するための事項が記載されている「特許請求の範囲」（以下、特許請求項）を読む必要がある。特許請求項は一般的な技術文書と異なり、独特の記述スタイルで記述される。そのため、特許請求項は可読性が低く、知的財産の非専門家では発明の内容が理解しにくいという問題がある。

この問題を解決するための第一歩として、発明の技術に関して詳細な説明が記載されている「発明の詳細な説明」（以下、詳細説明）と特許請求項を自動で対応付けさせることを目指す。本研究では、複数の類似性尺度に基づく手法と、素性として、類似性尺度と手掛かり語を用いた、機械学習に基づく手法の２つを、特許請求項と詳細説明の自動対応付け手法として提案する。

本論文の構成は以下のとおりである。２節では自動対応付け手法と類似性尺度についての関連研究を述べる。３節では、提案手法について述べる。４節では、提案手法の有効性を調べるために行った実験について述べ、５節で考察を行う。６節で結論を述べる。

2 関連研究

2.1 特許請求項と詳細説明の自動対応付け

新森ら[1]は、特許請求項の記述が非専門家の人にとっては読みにくいという問題に対し、特許請求項と詳細説明の自動対応付け手法を提案している。特許請求項と詳細説明の対応付けが行われることにより、特許請求項に対する作用と効果を明確にする、特許請求項の重要箇所を明確にする、特許請求項で使われている表現に関する言い換えを取得するという点が期待される。対応付け手法には、特許請求項を構造解析し、「用言文節を起点としたローカルアラインメント」を行うことにより、特許請求項と詳細説明の対応付けを行っている。

新森らは、手掛かり語を用いて、詳細説明内に記述される、発明の効果が説明されている「発明の効果」節を、

対応付け対象文として抽出し、特許請求項と対応付けを行っている。しかし、本研究では、詳細説明内の全ての節に対して対応付けを行っているため、新森らの手法と異なる。

2.2 文書間の類似性尺度

2.2.1 圧縮距離

相澤[2]は、Ziv_Merhav crossparsing と単純ベイズ法による確率計算を組み合わせた圧縮プログラム、ZM-Bayes の計算法を提案している。圧縮距離全般に共通する問題に、ファイルサイズのばらつきへの対応が難しいという点がある。ZM-Bayes では、競合的 N グラム選択と呼ぶ仕組みを導入することにより、この問題を解決している。

文書間の類似性比較手法として、正規化圧縮距離（Normarized Compression Distance、以下 NCD）が提案されている。NCD では、圧縮プログラムを用いて、文書間の近似的な距離を求めている。本研究では、圧縮プログラムとして、ZM-Bayes を用いている。NCD は Li ら[3]を参考に、以下のとおり定義した。

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

xy：入力 x と y をつなげたファイル
C（x）：x に対する圧縮後のファイルサイズ

2.2.2 ROUGE

ROUGE[4]は要約評価用の尺度として知られている。そのうちの一つである ROUGE-N は、要約システムの自動評価法として広く用いられている手法である。ROUGE-N は、参照要約とシステム要約の間で一致する N グラムの割合で類似性を計算している。

本研究では、文書ペアの類似性を比較することにより、対応付けを行っている。ROUGE は要約評価用の尺度であるが、2 つの文書ペアに対する類似性の比較においても有効であると考え、本研究において ROUGE を用いる。

2.2.3 DP マッチング

DP マッチングは、動的計画法を用いて、2 つ情報のパターンを比較することにより、文書間の対応付けを行う手法である。DP マッチングに基づく手法は、現在、音声認識や画像認識など幅広い分野で使用されている。本研究では、文書比較に、以下の式で定義される DP マッチングを用いる。

● DP マッチングの類似度

$$\sqrt{\frac{j}{\text{文書 A の形態素数}} \times \frac{j}{\text{文書 B の形態素数}}}$$

なお、上述の手法に加え、本研究では、以下の式で定義される DP_mod でも実験を行う。

● DP_mod の類似度

$$\frac{j}{\text{文書 A, B の内形態素数の少ない文書}}$$

DP_mod の特徴は、形態素の一致率を一方の文書に依存させることである。すると、ある文書に対して、抜き出した表現（部分一致表現）である文書を取得できると考えた。

3 提案手法

本節では、特許請求項と詳細説明の自動対応付け手法として、複数の類似性の組み合わせに基づく手法と、機械学習に基づく手法の 2 つを提案する。

●類似性尺度の組み合わせ

類似性尺度には、DP マッチング、DP_mod、圧縮距離、ROUGE の 4 種類を用いる。この類似性尺度の各類似度の値に対して、重み掛けることにより優劣をつけ、その優劣をつけた類似度を合計する手法を提案する。類似性尺度の組み合わせ手法の具体的な手順を以下に示す。

1）以下の式を用い、類似度を算出する。なお、各類似度と重み w は正規化されている。

$$score = \sum_{i=1}^{4} s(i)w(i)$$

s（i）：i の類似度

w（i）：i に対する重み

2）学習データにおいて、F 値の最も高くなる閾値 μ、重み w を選択。

3）評価データにおいて、2）で得た閾値 μ と重み w を用いることにより評価。

●機械学習

機械学習に Tiny SVM を使用した。機械学習の素性には、以下の３種類の素性を用いる。

①　類似性尺度

2.2. 節で示した類似性尺度の、圧縮距離、ROUGE、DP マッチング、DP_mod において、類似度を計算した値を用いる。各類似度は正規化されている。

②　文書ペアの形態素数の比

詳細説明は、発明の内容を詳しく説明しているため、特許請求項と比べ、形態素数が非常に多い場合がある。2.2. 節で示した類似性尺度は、文書間の形態素数の差が大きいと、類似度の値が低下する。それにより、文書ペアの形態素数の差が小さい時と比べると、対応付けが行われにくいという問題が考えられる。そこで、本研究では、機械学習の素性として、文書ペアの形態素数の比を用いた。文書ペアの形態素数の比は、特許請求項と詳細説明の形態素数を比べ、形態素数の少ない文書の形態素数を、形態素数の多い文書の形態素数で割ることにより、計算している。

③　手掛かり語

本研究では、「具備する」や「有する」のような、特許請求項に頻出する表現を多く含む詳細説明の段落は、特許請求項と対応付けられる可能性が高いと考え、特許請求項と詳細説明に頻出する用語の頻度数を素性として用いた。図１に、本研究で用いた手掛かり語リストの一部を示す。

特徴	上記	目的	形成
記載	課題	可能	解決
発明	構成	有する	係る
本発明	防止	配置	実施
請求項	提供	効果	検出

図１　手掛かり語リスト

4 実験

4.1　データセット

実験には、国立情報学研究所主催の評価ワークショップ NTCIR-3 において提供されている、特許データコレクションを用いる。このデータに含まれる 1999 年の公開特許公報から、公開日が 1999 年１月から３月までの 59,956 件の中から、無作為に 100 件抽出したデータを用いる。この際、化学物質に関する発明の場合、詳細説明において、化学式だけが特許請求項に記載されることが多い。よって特許請求項と対応付けができないことが考えられる。そこで、本研究では、「特許公報及び公表特許公報発行区分表」において、「Ⅲ化学・冶金・繊維」部門に属する特許については、評価の対象外とした。

正解データは、特許請求項と詳細説明を、段落単位で対応付けされている。また、ある請求項に対して対応付けられる詳細説明の数は、段落が複数存在するものもあれば、存在しないものもある。なお、正解データ数の内訳は、正例数約 7,908 対、負例数約 222,925 対である。評価尺度には、精度、再現率、F 値を用い評価する。また、４分割交差検定を実施する。

4.2　比較手法

提案手法の有効性を確かめるため、３種類の提案手法、５種類の比較手法で実験を行った。なお、実験に用いるデータセットにおいて、正解データに含まれている正例数に対して、負例数が明らかに多く、機械学習においてうまく分類されないことが考えられる。そこで、予備実験の結果から、負例数を正例数の３倍の数だけ無作為に抽出し、それらを機械学習に用いた。

提案手法

- ４種類の類似度：４種類の類似性尺度を組み合わせた手法
- SVM（４種類の類似度）：SVM の素性に、４種類の類似性尺度、形態素の比を使用
- SVM（４種類の類似度＋手掛かり語）：SVM の素性に、４種類の類似性尺度、形態素の比、手掛かり語を使用

ベースライン手法

- GETA：汎用連想検索エンジンの一つで、類似度の計算に SMART を使用
- DP マッチング：2.2.3 節に示す類似度を使用
- DP_mod：2.2.3 節に示す類似度を使用
- 圧縮距離：2.2.1 節に示す類似度を使用
- ROUGE：2.2.2 節に示す類似度を使用

4.3　実験結果

以下に、4.2. 節で示した、２種類の実験結果を示す。

表１　特許請求項と詳細説明の対応付け実験の結果

手法	精度(%)	再現率(%)	F値(%)
４種類の類似度	27.10	16.32	20.37
SVM（４種類の類似度）	37.19	14.38	20.74
SVM（４種類の類似度＋手掛かり語）	31.32	29.01	30.28
GETA	22.37	26.22	24.14
DP マッチング	17.21	32.01	22.38
DP（改）	19.60	11.61	14.58
圧縮距離	25.29	18.31	21.24
ROUGE	20.14	21.38	20.74

表１から、提案手法はいずれも、精度の値がベースライン手法を上回った。また、SVM（４種類の類似度＋手掛かり語）は、F 値が最も高い結果となった。

5　考察

特許請求項と詳細説明の対応付け結果について、機械が誤って検出した事例と、システムが検出できなかった事例について分析する。

・システムが誤って検出した事例

人手では対応付けを行わなかったが、システムが誤って対応付けを行った事例について分析を行った。その結果、特許請求項と詳細説明を比較すると、表層的には類似しているが、発明の内容が異なっているペアを誤って検出されていることが多かった。以下に、誤って検出した例を示す。

【特許請求項】

この環境条件検出手段の検出結果と前記キャップ手段によって前記記録ヘッドをキャッピングした時からの経過時間に応じて前記所定時間を設定する手段とを備えたことを特徴とするインクジェット記録装置

【詳細説明】

この環境条件検出手段の検出結果と前記キャップ手段によって前記記録ヘッドをキャッピングした時からの経過時間に応じて前記記録ヘッドから吐出させるインク量を設定する手段とを備えた。

図２　システムが誤って検出した例

図２では、同じインクジェット記録装置についての説明がされている例である。この例では、同じ手段についても書かれているが、下線部のように、異なる手段や効果について説明もしている場合がある。本手法では、文書ペアの類似度から対応付けする手法を提案している。そのため、上記のような例では誤って対応付けされてしまう。改善策としては、本手法を用いることで、表層的な対応付けを行い、その後、意味的な対応付けを行うことが考えられる。

・システムが検出できなかった例

人手では対応付けを行ったが、システムが対応付けを行わなかった事例について分析を行った。その結果、特許請求項と詳細説明を比較すると、表層的には類似していないが、発明について同じ内容が記述されているペアを検出できないことが多かった。以下に、システムが検出できなかった例を示す。

【特許請求項】

前記記録媒体として布帛が用いられる請求項８ないし 16 のいずれか１項に記載の記録装置

【詳細説明】

前処理において上記物質等を布帛に含有させる方法は、特に制限されないが、通常行われる浸漬法、パッド法、コーティング法、スプレー法などを挙げることができる。

図３　システムが検出できなかった例

図３では、「布帛に含有させる方法」について説明されている例である。この例では、特許請求項、詳細説明ともに、同じ内容の説明がされているが、詳細説明において、より専門的な用語を用いて説明がされている。よって、同じ内容を示しているが、表層的には類似していないため、本手法では対応付けを行うことができなかったと考えられる。改善策としては、上位下位概念を用いて対応付けを行うことにより、表層的な一致では対応付けを行うことができない対を獲得できると考えられる。

6 おわりに

本研究では、特許請求項と詳細説明の自動対応付け手法を提案した。対応付け手法には、複数の類似性尺度に基づく手法と、類似性尺度と詳細説明に頻出する用語を素性とした、機械学習に基づく手法を提案した。実験の結果、提案手法である、SVM の素性に、４種類の類似性尺度、形態素の比、手掛かり語を使用する手法は、F 値が最大で 15.70% の向上という結果が得られ、本研究の有効性が示された。

参考文献

[1] 新森明宏、奥村学：特許請求項読解支援のための「発明の詳細な説明」との自動対応付け、自然言語処理 Vol.12, No.3, pp.111-128 (2005).

[2] 相澤彰子：情報検索における圧縮距離の適用に関する考察、情報処理学会研究報告 , Vol.2010-NL-199, No.8 (2010).

[3] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. : The Similarity Metric, IEEE Trans. on Information Theory, Vol.50, No.12, pp.3250-3264 (2004).

[4] Lin, C.-Y: ROUGE: A Package for Automatic Evaluation of Summaries. Proc. the ACL-04 Workshop “Text Summarization Branches Out”, pp.74-81 (2004).