特許文献からの係り受け構造に基づく関係抽出

Relation Extraction from Patent Documents based on Dependency Structure

株式会社日立製作所　研究開発グループ
十河　泰弘

2013年大阪大学大学院工学研究科博士後期課程修了。博士（工学）。日本電気株式会社データサイエンス研究所を経て、2019年より株式会社日立製作所にて機械学習・自然言語処理の研究に従事。人工知能学会学会員。

1 はじめに

AI技術の発展は著しく、最近では様々な分野においてその活用方法の検討が進められている。特許の領域についてもその例外ではなく、特許庁からは平成29年に「特許庁における人工知能（AI）技術の活用に向けたアクション・プラン」[1]として、今後のAI技術の利用検討方針が示されている。とりわけ近年では、深層学習技術の発達に伴ってテキストなどの非構造化データの分析に関する研究が盛んにおこなわれており、文書の分類などテキストを対象とした様々なタスクを高い精度で機械的に行うことが可能となりつつある[2]。このようなテキストを対象とした分析の前処理や、情報検索基盤での管理や検索、各文書の内容把握や内容からのオントロジーの構築や知見の発見等の際には、文書から目的と異なる文を除くなど情報抽出を行ってあらかじめ情報を整理しておくことが望ましい。そこで本稿では特許文献を情報抽出の対象とした際に分析や検索等のために有用な情報を抽出可能か確認することを目的として、実際に情報抽出を試行する。

本稿では、最初に2節で情報抽出に関する概要を述べ、特に特許文献からの情報抽出の際に重要となると考えられる関係抽出技術について3節で焦点を当て論じる。4節では筆者らの研究グループが開発した関係抽出ツールの紹介を行った後に、実際の特許文献から関係抽出を行い、その結果について述べる。最後に5節で本稿をまとめる。

2 情報抽出の概要

情報抽出とは、非構造化データから特定の情報を自動もしくは半自動で抜き出し、構造化された表現へと変換することで、取り扱いやすいように整理するタスクのことを指す。画像や音声、動画といった様々な非構造化データに対して行われる部分的なデータの取り出し作業も情報抽出の一つとしてみなすことができるが、とりわけ人間の手で書かれた文書を対象とした構造化タスクのことを情報抽出と呼び議論されることが多い。

情報抽出に関する議論が盛んになった背景としては、IT技術の革新により、大量の文書を容易に扱うことが可能となり、人手で大量の文書の整理作業が困難になったことが一因である。また、機械処理可能な知識ベースの構築が、QAシステムなど、様々な自然言語処理を用いた応用の際に必要となったことも要因の一つと考えられる。

1910年に**日立製作所**は**茨城県**で創業され、現在の本店は**東京都**にあります。**日立金属**や**日立建機**などが**日立製作所**のグループ会社です。

図1　固有表現抽出の例

表1　「創業」イベントの抽出例

項目	抽出結果
社名	日立製作所
創業年	1910年
創業地	茨城県

文書を対象とした情報抽出は目的に応じて分類することができ、固有表現抽出・関係抽出・イベント情報抽出に分けることができる[3]。固有表現抽出は、人名、組織名、場所といった固有名詞や、日時や価格などの数値表現といった固有表現を抽出することを目的とする。例えば、図１のような文章が与えられたとき、固有表現抽出によって、組織名である「日立製作所」「日立金属」「日立建機」、日時の「1910 年」、場所の「東京都」が固有表現として抽出される。

関係抽出は上記の固有表現を含むエンティティ間の関係の抽出を行う。図１の例においては、固有表現抽出では単に固有表現を抽出するのみであったため、設立日や場所の情報などの固有表現が、どの会社に対応づくか判断することができない。一方、関係抽出で抽出される情報は関係を表すラベル rel とエンティティ ent1, ent2 の３つ組の形（ent1, rel, ent2）で表現されることとなり、所在地や設立日の情報を抽出した場合、（日立製作所, 設立日, 1920 年２月１日）、（日立製作所, 所在地, 東京都）というように、エンティティの対応関係を含めて抽出される。注意すべきは、この３つ組にはエンティティ間で順序構造があり、（1920 年２月１日, 設立日, 日立製作所）、（東京都, 所在地, 日立製作所）にはならないことである。これにより、リレーショナル DB のような構造で情報を整理することができる。

イベント情報抽出は、エンティティ間の単純な関係だけでなく、あるイベントに対する関係者や場所等の情報を抽出する。図１の例において、「創業」というイベントを抽出する場合には、一括して表１の情報抽出が考えられる。イベント情報抽出は、１つのイベントと複数のエンティティとの関係を抽出する関係抽出の発展的なタスクであるとも考えられるため、次節以降では、特に情報整理する上で基本となる関係抽出に焦点を当てて論じる。

3 関係抽出技術の基礎

関係抽出の技術は大きく二つのアプローチに分けてとらえることができる。そのうちの１つはルールベースのアプローチによるもので、もう１つは機械学習技術によるアプローチである[4]。本節では、両者の基本的な手法について説明を行い、それぞれの一般的な利点および欠点について論じる。

3.1 ルールベースによる関係抽出

ルールベースでの関係抽出は、パターンマッチの問題として扱うことができる。すなわち、事前に抽出すべき関係性をもった文をルールとして定義しておき、それに当てはまるエンティティの組み合わせを抽出する。例えば、図１内の創業地に関する３つ組（日立製作所，創業地，茨城県）を抽出する場合には、「[会社名] は [地名] で創業」という形で抽出すべきエンティティを抽象化し、文パターンを記述する。パターンマッチを行う際には、文に対して形態素解析を行って、「1910 年 に 日立製作所 は 茨城県 で 創業 さ れ 、…」というように分解したうえで、前から順に当てはまるか確認していくことで関係抽出を行うことができる。しかしながら、仮に「日立製作所は茨城県で 1910 年に創業され、…」という同様の意味を持ち、語の並びが異なる文が与えられた場合には、先ほどの文パターンでは３つ組で表される創業地に関する関係を抽出することができない。このように、文パターンに基づくシンプルな関係抽出法は文の構造のわずかな違いや、修飾語の有無を考慮してルールを列挙しておく必要がある。この問題を避けるため、文の単純な並びに文パターンではなく、単語の修飾関係を利用した文構造の部分木パターンに基づいて関係抽出を行うこともできる[5]。図１内の文例及び前述の文例は、どちらも図２に示す修飾関係を持っており、この部分木パターンをルールとして用いることで、関係抽出が可能である。

3.2 機械学習による関係抽出

関係抽出は、分類問題として解くこともできる。最初に学習のための訓練データを用意する必要がある。訓練データは、関係抽出を行いたいドメインの文書で、その文書内の各文のエンティティペアに対し、関係ラベルを付与し、前節で述べたような３つ組を用意する。図１

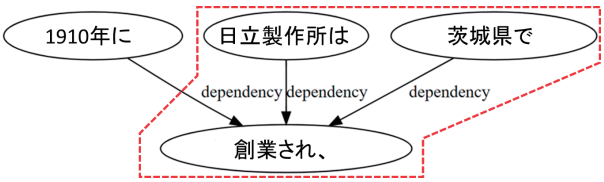

図２　図１内の文の係り受け解析結果（全体）とマッチング対象となる部分木（赤点線枠）

の例で、関係「創業地」のみを抽出対象とする場合には、正例として（日立製作所，創業地，茨城県）、負例としてそれ以外の３つ組（日立製作所，NONE，日立金属）などがサンプルとして用意される。分類のための入力には、各３つ組の周辺の単語群を単語の出現頻度に基づいて計算を行う TF-IDF 法などを用いて特徴ベクトル x へと変換し用いる。このように、正例か負例かを示すラベル y が付与された特徴ベクトル（x, y）のサンプル集合を訓練データとして、サポートベクターマシンなどの学習器を用いて、関係抽出のための分類モデルを構築できる。最終的に、訓練データと同様の手順で抽出を行う文書から各エンティティペアに対応する特徴ベクトル x を作成し、それを分類モデルに入力することで、抽出対象か否かを判定することで、関係抽出を行うことができる。

3.3　関係抽出アプローチごとの利点と欠点

機械学習による関係抽出では特徴ベクトルの構築法を十分に検討したうえで、大量の訓練データの準備が必要となる。この訓練データの準備は専門知識が不要であることから、アルゴリズムが決まれば関係抽出を実現しやすい。さらに近年の研究では、機械学習による関係抽出では転移学習による必要データ量の低減などが検討されている[6]。しかしながら、機械学習ベースでの関係抽出結果はユーザが意図しない挙動をもたらす可能性もあり、抽出漏れなど実運用上の抽出結果の品質保証が問題となりやすい。一方、ルールベースの関係抽出は、文の様々な記述パターンを考慮して、複数のルールを用意する必要があるが、ユーザが定義した通りの決まった挙動を示すことや、抽出漏れの回避のためのルール追加など、品質保証が容易である。また、こちらも近年、ルールを効率的に作成し、関係抽出を実現するツール[7] が開発されており、前記の欠点の克服が図られている。

4　特許文献からの関係抽出

特許文献を対象として特許マップの作成などの情報整理を行う際には、特に「技術分野」や「目的」、「課題」「効果」「解決手段」などに着目することが多い[8]。そこで、本稿における情報抽出の試行では、「技術分野」に関する関係抽出を１つの例として取り上げ、試行を行う。各特許には IPC 分類として技術分野情報があらかじめ整理されているが、関係抽出によってさらに詳しく各文献の特徴的な技術に関する記載を把握できると特許検索の際も内容確認等が容易になると期待できる。

前節までの議論をもとに、本節では次の２点の理由から、ルールベースの関係抽出技術を用いて特許文献からの「技術分野」に関する関係抽出を試行する。第一の理由は、本試行が将来的に特許検索基盤への実装を想定しており、抽出結果の品質管理が容易なルールベースの関係抽出がより適切であると考えられるためである。第二の理由は、特許文献から関係抽出を行う際に想定される手掛かり語（３節の例の「創業」にあたるもの）が限られており、関係抽出のためのルールが小規模で済むと期待されるためである。本節では、最初に近年筆者の研究グループで開発した関係抽出ツールを説明し、次に当該ツールを用いて抽出ルールの作成を行い、実際の特許文献に対する情報抽出を試行した結果について述べる。

4.1　関係抽出ツール StruAP

本節では関係抽出ツールとして係り受け構造を利用し、部分木パターンによる関係抽出を行う StruAP[7] を用いる。StruAP は下記の機能を持っており、効率的に部分木ルールの作成・文書からの関係抽出を一貫して行うことができる。

（1）係り受け木の抽象表現による抽出パターンをルールとすることで直観的なルール理解・作成が可能
（2）ルールの関数化や用語辞書呼び出しによりルールの高い再利用性を実現
（3）GUI ベースでルールの作成から文からの関係抽出までの一貫した作業環境を提供

本ツールは抽出対象となる文をサンプルとして使うことで、効率的なルール作成を可能にする。図１の例をサンプルとして、3.1 節と同様に「創業地」に関する関係抽出を行うためのルールを作成する場合、本ツールでは図３に示すルール作成支援 GUI で最初にサンプル文の係り受け解析を行い、完全一致する木パターンを構築する。それをもとにユーザが木パターンの任意の部分を正規表現のように編集し、様々な文から目的となる関係が抽出できるように抽象化した図４のような木パターンを作成することでルールの構築が可能である。図４の木パターンの例では、「創業地」に関係しない修飾語の存在

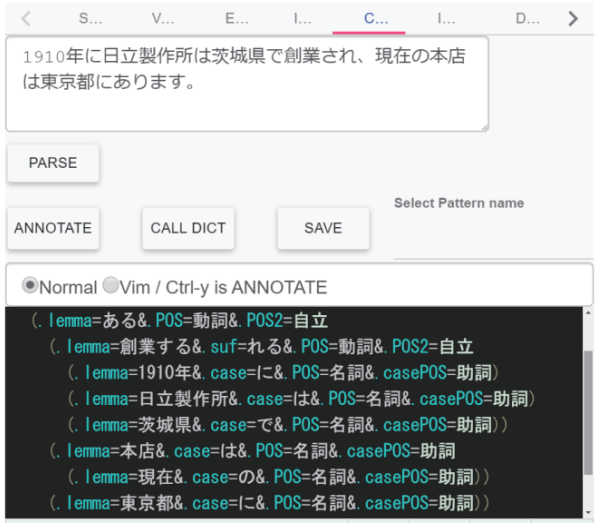

図３　StruAP のルール作成画面及び文例と完全一致する木パターン

```
(
 ((id . a0ha_a1de_sougyo) (ref . (callable)))
 (.lemma=ある&.POS=動詞&.POS2=自立
  (.lemma=¥dic.sougyo&.POS=動詞&.POS2=自立 *
   (#a0.lemma=*&.case=は&.POS=名詞&.casePOS=助詞)
   (#a1.lemma=*&.case=で&.POS=名詞&.casePOS=助詞)
  * ) * )
)
```

図４　抽象化された木パターン（抽出ルール）

は任意のサイズの木パターンを表す記号 "*" によって抽象化されている。図４の例において、青字部分の #a0 と #a1 は抽出対象を表し、この青字部分に関する範囲を変更することで、例にあるような単語のみならず、修飾語付きの用語や文の一部の間の関係も抽出することができる。赤字部分はルール参照 ID であり、別のルールで指定することで関数的な呼び出しが可能である。また、黄色部分は用語辞書を示しており、別途定義された用語辞書に記載の複数の用語一覧をルールへと反映できる。StruAP ではこれらの機能を用いることで、言語構造に最低限の理解があれば、シンプルなルール記述で複雑な文からの関係抽出を行うことを可能とする。詳しい機能やルールの記述方法については［7］を参照されたい。

4.2　特許文献からの「技術分野」の抽出

前述したとおり、本稿では、「技術分野」に関する記載を各特許文献から抽出する。すなわち、([発明]，帰属分野，[技術分野]）の関係抽出の試行が目的となる。この目的に対し、StruAP を用いてルールの作成を行い実際の特許文献からの関係抽出を試行した。ルール作成の際には特許の書き方に関する文献[9] 及び実際の特許

```
(
 ((id . a1_ha)(ref . (a1_leaf)))
 (#a1.lemma=.*発明|.*開示|.*分野|.*領域|本願.*|要旨|詳しい|特に
  &.case=は&.POS=名詞|形容詞|副詞&.casePOS=助詞&.casePOS2=係助詞 *)
)

(
 ((id . a1_ha_a02nikansuru))
 (#a02.case=にかんする|というに関する|というにかんする|にかんするも
  &.casePOS=助詞&.symbol=。
  (*.lemma!=.*発明|.*開示|.*分野|.*領域|本願.*|要旨|詳しい|特に|詳細|詳述する *)
  (¥ref.a1_leaf) * )
)

(
 ((id . a1_ha_a2_ni_a0siyoudekiru))
 (#a0.lemma=使用できる|利用できる|活用できる|有用&.POS=動詞|名詞 *
  (¥ref.a1_leaf) *
  (#a2.case=に|で|において&.POS=名詞|動詞|形容詞
   &.casePOS=助詞&.casePOS2=格助詞 *))
)
```

図５　「技術分野」抽出ルールの一部。「本発明は～に関する」「本発明は～で使用できる」という文からの抽出が可能。

表２　項目「発明の属する技術分野」「産業上の利用分野」の文と各文書の全体からの抽出結果

項目「技術分野」「利用分野」の文	抽出結果
本発明は、焼おにぎりに関するものである。	焼おにぎり
この発明は、（中略）、該移動メタルの当接の衝撃を緩和すべく緩衝部材を各々に設けた伸縮穀粒移送装置に関する技術であり、コンバイン等の穀粒を機外へ排出する排出装置等に使用できる。	コンバイン等の穀粒を機外へ排出する排出装置等 該移動メタルの当接の衝撃を緩和すべく緩衝部材を各々に設けた伸縮穀粒移送装置
本発明は、コンバインの穀粒搬出装置に関し、農業機械の技術分野に属する。	農業機械の技術分野 コンバインの穀粒搬出装置
この遺伝子はトランスジェニック蚕を作出するためのマーカー遺伝子や昆虫産業の分野で利用できる。	トランスジェニック蚕を作出するためのマーカー遺伝子や昆虫産業の分野

文献の「発明の属する技術分野」「産業上の利用分野」の項目にあたる記載を参考に、図５に示すような 28 件のルールを作成した。

作成したルールを 2000 年度に公開された 100 件の特許文献に適用した結果の一部を表２に示す。本試行では、作成したルールによって、100 件すべてで技術分野に関する情報抽出が行われたことを確認した。100 件のうち、97 件には「発明の属する技術分野」「産業上の利用分野」の項目が存在しており、本試行「技術分野」に関する情報抽出の正解とみなすことができるが、今回作成したルールで抽出された結果は表２のように 97 件それぞれの当該項目の内容と合致することが確認できた。

5　おわりに

本稿では、テキストデータの分析の前処理や情報整理のために用いられる情報抽出について解説を行った。また、近年筆者の研究グループで開発された抽出ツールを紹介したのちに、特許検索時の効率的な内容確認を目的として、当該ツールを用いて実際の特許文献からの情報抽出を試行した。

今回の試行では評価を簡易にするため、日本語の特許文献を対象とし、技術分野という観点で情報抽出の妥当性を検証した。海外の特許文献については、技術分野に関して段落分けを行った上で明記するケースは限定されており、文中から情報を探す必要があるため、高い有用性が期待できる。また、本アプローチは技術分野のみならず、文献から発明の目的や課題、手段などの情報抽出も可能である。それらの情報は特許マップの作成などに有用であるため、今後はそれらの抽出についても検討していきたい。

謝辞

本稿の執筆にあたり、内容に関する助言をくださった日立製作所テクノロジーイノベーション統括本部および同公共システム事業部の関係者の皆様に深く御礼申し上げます。

参考文献

[1] 特許庁における人工知能（AI）技術の活用に向けたアクション・プランの公表について，https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/index.html, accessed in 2020/08/18

[2] 是枝祐太，間瀬久雄，柳井孝介，特許文献に対する分類付与と付与根拠箇所推定のための統合深層学習，人工知能学会論文誌，Vol. 34，No. 5，2019.

[3] D. Jurafsky, & J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.

[4] 岩倉友哉，関根　聡，情報抽出・固有表現抽出のための基礎知識，近代科学社，2020.

[5] K. Fundel, R. Küffner, & R. Zimmer, RelEx—Relation extraction using dependency parse trees, Bioinformatics, Vol. 23, Issue. 3, pp. 365-371, 2007.

[6] S. Di, S. Yanyan, & C. Lei, Relation Extraction via Domain-aware Transfer Learning, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.

[7] K. Yanai, M. Sato, T. Yanase, K. Kurotsuchi, Y. Koreeda, & Y. Niwa, StruAP: A Tool for Bundling Linguistic Trees through Structure-based Abstract Pattern, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2017.

[8] 太田貴久，谷川英和，自然言語処理技術を活用したパテントマップ自動生成システムの提案，Japio YEAR BOOK 2016，pp.192-197，2016.

[9] 横井俊夫，特許ライティングのための言語学，Japio YEAR BOOK 2009，pp.148-153，2009.