自然言語処理技術を用いた特許オントロジーの自動構築および知識発見

中央大学　理工学部　経営システム工学科
教授　難波 英嗣

令和2年9月15日

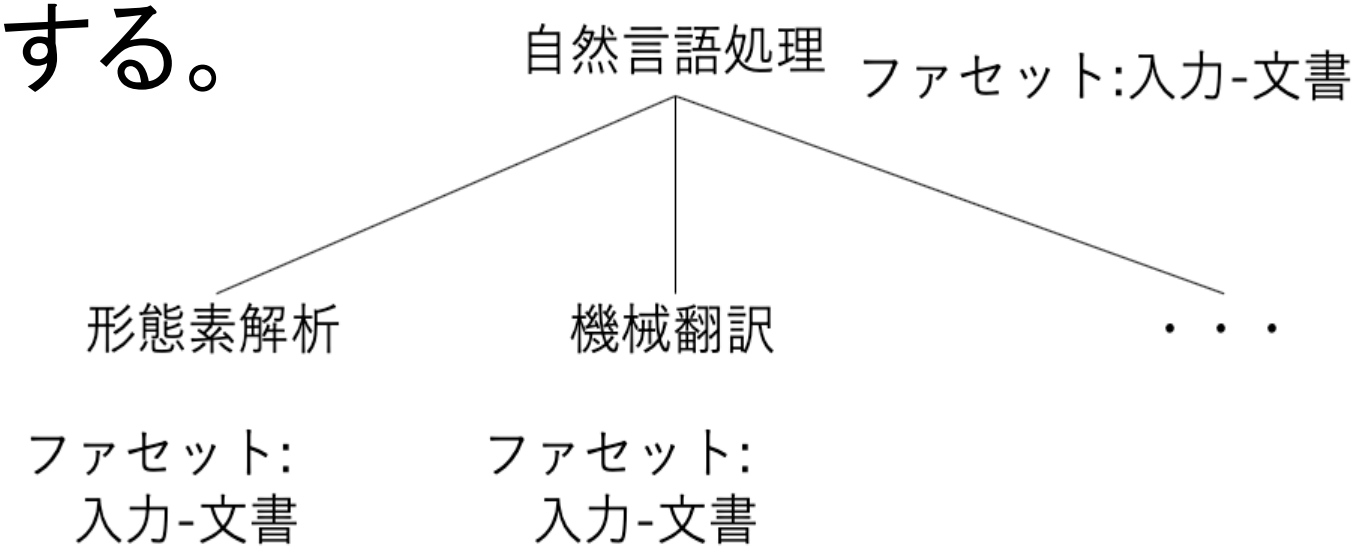
背景

- 特許オントロジーとは、特許や論文等の技術文書を検索したり、執筆する上で有用な情報源であるが、人手で構築し、更新することは非常にコストがかかる。
- 自然言語処理技術を用いて自動構築する手法も提案されてきたが、従来技術では扱える用語間の関係が限られており、また品質面の問題から、広く利用されるまでには至っていなかった。

研究の目的

- 本研究では、自然言語処理技術を用いて特許データベースから専門用語の用語間の関係を解析し、特許オントロジーを構築する。
- このオントロジーを用いて新たな知識を発見する手法を提案する。

本研究で作成する特許オントロジーの例



想定される用途

- 特許オントロジーを用いた特許検索時の検索質問拡張
- 特許を対象とした自然言語処理のための言語資源
- 技術文書の執筆支援
- 特定分野の技術動向分析

従来技術とその問題点 (1)

「などの」「等の」といった定型表現(英語の場合、such as)に注目し、テキストデータから用語の上位下位概念を獲得する手法が提案されているが、定型表現で必ず正しい関係が抽出できるとは限らないという問題があった。[Hearst 1992]

成功例）形態素解析や構文解析などの自然言語処理

→　自然言語処理の下位概念：形態素解析、構文解析

失敗例) パソコンなどのキーボード

→　キーボードの下位概念：パソコン？

Marti A. Hearst, “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proceedings of the 14th International Conference on Computational Linguistics, pp. 539-545, 1992.

従来技術とその問題点 (2)

- 少数の正解事例(例えば、正しく上位下位関係にある用語の対)を入力し、ブートストラッピング法を用いて、定型表現と上位下位関係にある用語対を相互に再帰的に抽出 [Pantel 2006]
- ブートストラッピング法とBERTを組み合わせた手法[Nouraoui 2020]

従来手法では、低頻度語が抽出できない。特許の場合、低頻度語である新概念(新語)の抽出が重要。

Patrick Pantel and Marco Pennacchoitti, “Espresso: Leveraging Generic Pattern Extraction by Multi-Level Bootstrapping,” Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp.113-120, 2006.
Zied Bouraoui, Jose Camacho-Collados, Steven Schockaert, “Inducing Relational Knowledge from BERT,” Proceedings of AAAI 2020, pp. 7456-7463, 2020.

新技術の特徴･従来技術との比較

- 「復号処理」―「誤り訂正復号」のような上位下位関係の用語対を従来よりも正確に抽出することが可能になった。
- 上位下位以外の様々な関係(部分全体、対象) を扱うことが可能になった。
- 低頻度の新語でも抽出可能。

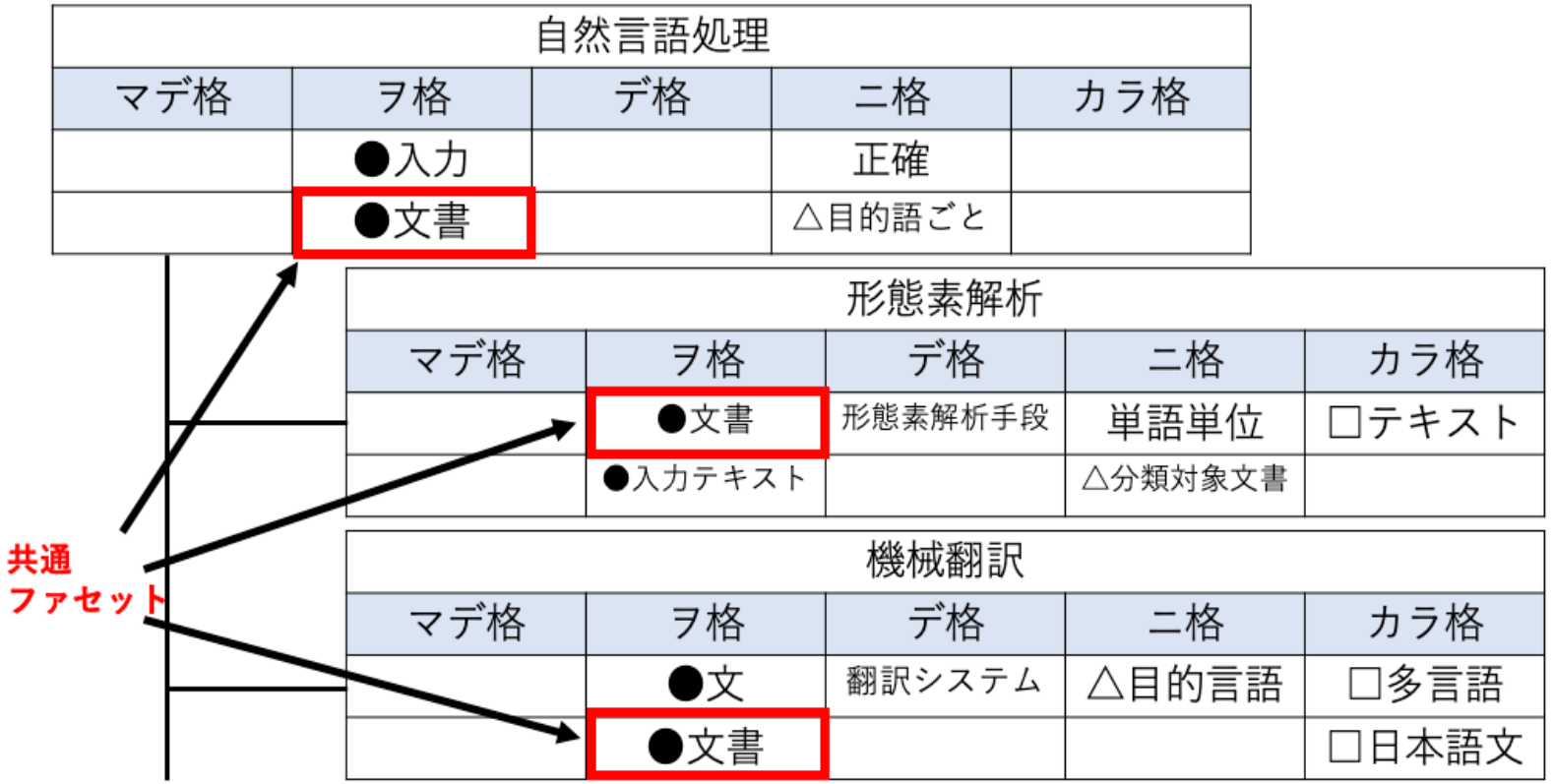
特許オントロジー構築の手順

- (手順1) 定型表現と言語モデルを用いた上位下位関係の候補の抽出
- (手順2) 格文法を考慮したファセット抽出および用語の上位下位関係の抽出

(手順1) 定型表現と言語モデルを用いた上位下位関係の候補の抽出

二つの定型表現「(Aや)BなどのC」と「(Aや)B等のC」を用いて、上位下位関係の候補を抽出する。今回は、この候補の中で上位語、下位語共に名詞サ変接続のものを対象とする。

(手順2) 格文法を考慮したファセット抽出および用語の上位下位関係の抽出



- 言語モデルBERTを用い、定型表現の前後の用語が上位下位関係にあるかどうか判断する。
- 手順1で得られた上位下位関係の候補の中から、共通ファセットを持つ用語対を上位下位関係があると判断する。この共通ファセットの有無を調べるため、格文法を利用する。

実験結果

- 上位下位関係の抽出は、従来手法(再現率1、精度0.84)に比べ、提案手法はより高い精度(再現率0.64、精度0.89)を達成。
- 再現率は従来手法よりも低いが、格文法を利用することで、上位下位以外の関係を抽出できる点が従来と異なる。
- 提案手法を1993年～2017年の特許に適用し、7981件の上位下位関係、延べ22,427,356件の格要素(8種類の格助詞)を抽出。

実用化に向けた課題

- 提案手法は、従来手法よりも再現率が低い問題を解決する必要がある。今後は、格文法を用いて抽出された用語間の関係を利用し、再現率の改善を図る。
- 英語や中国語などの他の言語への適用を検討している。

企業への期待

社内での特許情報の活用(例えば技術動向分析)、特許以外の技術文書の活用を考えている企業との共同研究を希望。

本技術に関する知的財産権

- 発明の名称：関連用語取得装置、関連用語取得方法、及びプログラム出願番号
- 発明者：難波英嗣
- 出願人：広島県広島市/公立大学法人広島市立大学
  (現在の権利者：学校法人中央大学)
- 出願番号：特願2008-505065
- 公開番号：WO2007/105530
- 登録番号：特許5078164

産学連携の経歴

- 2020年-　株式会社アイ･アール･ディー技術顧問
- 2018年-　株式会社ジー･サーチと共同研究実施
- 2017年-　株式会社アイ･アール･ディーと共同研究実施
- 2017年-2019年　株式会社ブリヂストンと共同研究実施
- 2016年-2017年　パナソニックIPマネジメント株式会社と共同研究実施
- 2016年-2017年　科学技術振興機構と共同研究実施

お問い合わせ先

中央大学　研究推進支援本部

ＴＥＬ　０３－３８１７－１６０３
ＦＡＸ　０３－３８１７－１６７７
e-mail　ksanren-grp@g.chuo-u.ac.jp