

S17 特許データを用いた手順オントロジーの自動構築

菊池泰成, 福田悟志, 難波英嗣 (中央大学)

研究背景・目的

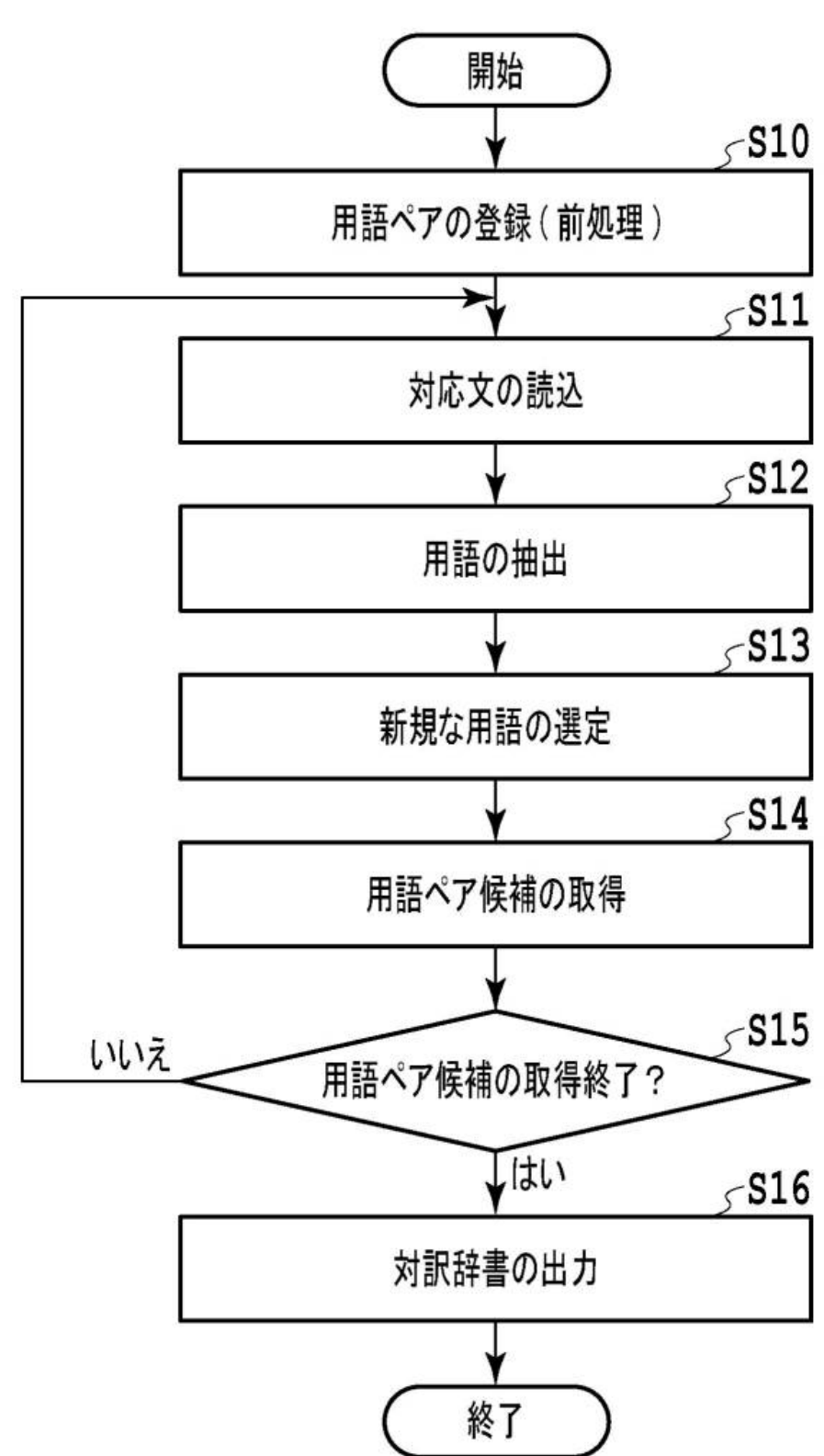
- 幅広い分野の一連の手続きに関する知識を特許から自動抽出し、それらを体系化した手順オントロジーの自動構築を目指している。
- 我々は、これまでに特許中の画像からフローチャートか否かを自動的に識別し、さらにその説明文を自動生成する研究を行っている。[1]
- 本研究では、(1)特定分野の特許中のフローチャート画像と要約から、グラフ形式の手順情報を抽出し、(2)それらを統合・整理することで、手順オントロジーを構築する。

[1] Nanba, H., Kubo, S., and Fukuda, S. (2023) "Automatic Generation of Explanatory Text from Flowchart Images in Patents". In Proceedings of the 4th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2023) in conjunction with SIGIR 2023.

(手順1) 特許中のフローチャート画像と要約からの手順情報の抽出

本研究では、GPT-4 Turbo with visionを用いて、フローチャート画像と要約から手順情報を抽出する。

Prompt: この画像からノードとエッジを抽出し、列挙してください。以下は、この画像の説明文です。ノードとエッジの表現は、できる限り以下のテキストを利用してください。



対訳コーパスから複数の対応文を読み込みS11、複数の対応文から用語を抽出しS12、抽出された用語が用語ペアテーブルに登録されている用語ペアを構成する用語以外である場合には、当該用語を新規な用語として選定するS13。複数の対応文のマッチングに基づいて、新規な用語のペアを用語ペア候補として取得しS14、用語ペア候補の出現頻度に応じて、当該用語ペア候補を構成する新規な用語ペアを対訳辞書として出力するステップS16。取得するステップでは、複数の対応文の順序をランダムに変更して前記マッチングを繰り返し行う。

Nodes (Steps):

- S10: 開始
- S11: 対訳コーパスから複数の対応文を読み込み
- S12: 複数の対応文から用語を抽出
- S13: 抽出された用語が用語ペアテーブルに登録されている用語ペアを構成する用語以外である場合には、当該用語を新規な用語として
- S14: 複数の対応文のマッチングに基づいて、新規な用語のペアを用語ペア候補として取得
- S15: 用語ペア候補の検証? S16: 用語ペア候補の出現頻度に応じて、当該用語ペア候補を構成する新規な用語ペアを対訳辞書として出力
- 終了 (End)

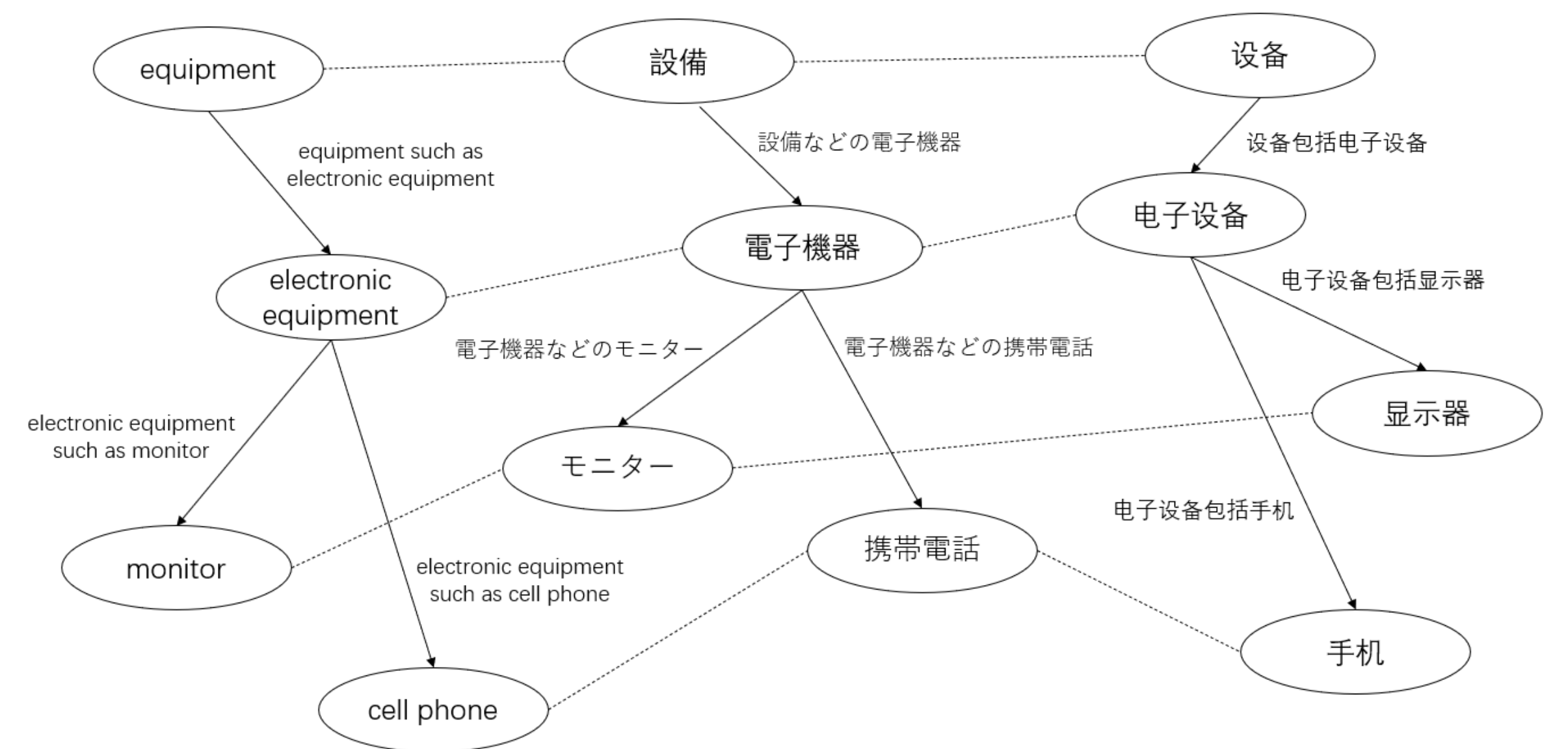
Edges

(Transitions):

- From S10 to S11
- From S11 to S12
- From S12 to S13
- From S13 to S14
- From S14 to S15
- From S15 to S16 (Yes)
- From S15 back to S14 (No)
- From S16 to End

(手順2) 複数の手順情報の統合

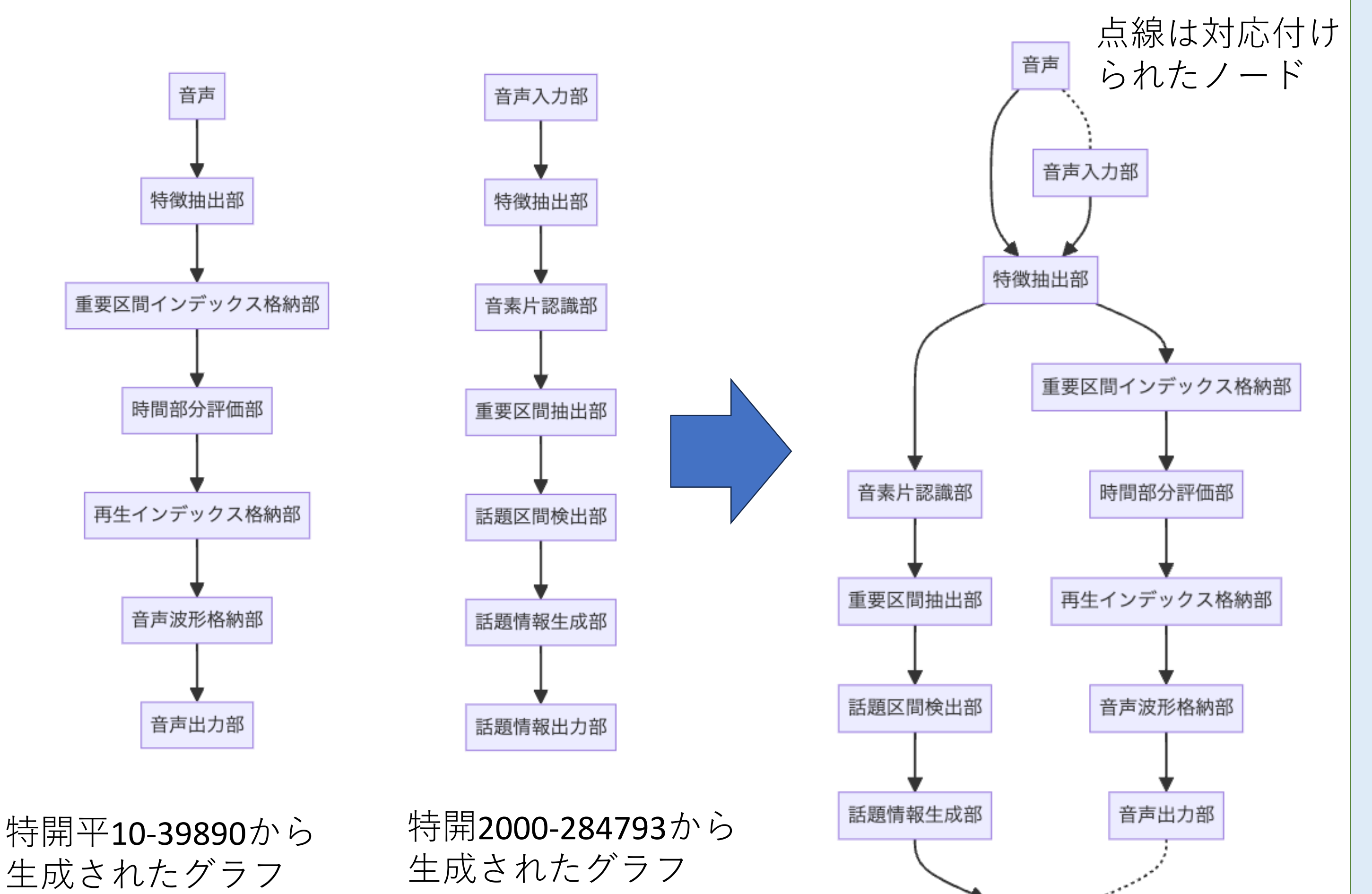
我々は、これまでに、日本語、英語、中国語で記載された特許から上位下位関係にある用語対を抽出した後、ConvEとGraphSAGEを組み合わせたリンク予測により、言語間の用語を対応付けることで、多言語シソーラスを自動構築している。[2]



ConvEとGraphSAGEを組み合わせたリンク予測により構築された多言語シソーラス

[2] Iwakuma, K., Gong, Y., Nanba, H., and Fukuda, S. (2023) "Automatic Multilingual Hypernym-Hyponym Relation Extraction Using a Link Prediction Model". In Proceedings of the 14th International Conference on Advanced Applied Informatics (IIAI AAI 2023). Competitive Paper Award

この手法を特定分野の複数の特許から抽出したグラフ形式の手順情報に適用することで、特定分野の手順情報を統合する。



特開平10-39890から生成されたグラフ

特開2000-284793から生成されたグラフ

音声要約に関する特許からのグラフ生成

特開平10-39890と特開2000-284793を対応付けて生成されたグラフ

今後の課題

様々な特許の対応付け

- 同一組織から出願された特許の対応付け
- 拒絶された特許とその根拠となった特許との対応付け
- 同一IPC、Fタームの特許の対応付け

評価方法の検討

- 対応付けられたノードの一致率による評価
- 情報検索などのタスクベースの評価