一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

複数手順テキストからの手順オントロジーの自動構築

難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1
E-mail: †{nanba, takezawa}@hiroshima-cu.ac.jp

あらまし 本研究では，特許における類似の手順テキスト集合から，特定の目的達成のための典型的な手順を抽出することで手順に関するオントロジーを自動構築する手法を提案する．手順テキストを大量に収集し，目的別に分類し，統計的機械翻訳技術によって同じ目的で類似する複数の手順テキストを比較すれば，典型的な手順を見つけることができる．複数テキスト要約の手法により，このような典型的な手順を大量に集め，体系化することで，手順オントロジーの自動構築を目指す．

キーワード オントロジー，手順，統計的機械翻訳，特許，レシピ

Automatic Construction of Procedure Ontology from Multiple Procedure Texts

Hidetsugu NANBA† Toshiyuki TAKEZAWA†

†Graduate School of Information Sciences, Hiroshima City University 3-4-1, Ozukahigashi, Asaminamiku, Hiroshima 731-3194 Japan
E-mail: †{nanba, takezawa}@hiroshima-cu.ac.jp

Abstract In this paper, we propose a method that constructs a procedure ontology from multiple procedure texts in patents. If we collect procedure texts, classify them for each purpose, and compare similar procedure texts using a statistical machine translation technique, we can identify a typical procedure. We construct procedure ontology by identifying typical procedures based on a multiple document summarization method and organizing them.

Keywords Ontology, Procedure, Statistical Machine Translation, Patent, Recipe

1. はじめに

料理レシピは料理を完成させるための一連の手続きを記したものである．特許においても新しい技術や発明を説明するために，それを実現する手順を記載することがしばしばある．

図 1 は，食器洗浄乾燥機に関する特許の請求項である[1]．この図から，この装置は，(1)「水を吸水し」(2)「外気を吸引し」(3)「洗浄ポンプ駆動させ」(4)「ヒータを発熱させる」という，4 つの手順から構成される食器洗浄乾燥機であることが分かる．

このように，ある特定の目的を達成するための一連の手続きを記したものを，手順テキストと呼ぶ．本研究では，類似の手順テキスト集合から，目的を達成するにいたる典型的な手順を抽出することで，手順に関するオントロジーを自動構築する手法を提案する．

手順テキストを大量に収集し，目的別に分類し，同じ目的で類似する複数の手順テキストを比較すれば，典型的な手順を見つけることができる．さらにこのような典型的な手順を大量に集め，体系化できれば，手順オントロジーを構築することが可能になる．

給水された洗浄槽内の水を吸水し(1)，噴射ノズルを介して洗浄槽内の食器類に噴射する洗浄ポンプと，洗浄槽内の水を加熱する第１のヒータと，外気を吸引し(2)，送風口を介して洗浄槽内に送り込む送風モータと，該送風モータと送風口との間に設けられた第２のヒータと，予め設定された複数のすすぎ工程のうち最後のすすぎ工程の前までは、前記洗浄ポンプを駆動させ(3)，最後のすすぎ工程においては前記洗浄ポンプを駆動させると共に，前記第１のヒータを発熱させる第１の制御手段と，乾燥工程時，前記送風モータを駆動させると共に，前記第２のヒータを発熱させる(4)第２の制御手段とを備えたことを特徴とする食器洗浄乾燥機．

図 1 特許における手続きの記載例(特開 1999-178777)
Fig 1. Example of a procedural description in a patent (Japanese published unexamined application 1999-178777)

[1] なお，下線部および数字は筆者が付与した．

手順オントロジーを構築する処理を，本研究では複数テキスト要約と捉える．入力された複数のテキストからひとつの要約を作成する，いわゆる「複数テキスト要約」では，入力テキスト間の類似点と相違点を検出することが必須の処理のひとつであると言われている[1]．今，ある目的に関する複数の手順テキストを複数テキスト要約システムの入力と考えるならば，その典型的な手順と個々の手順テキストの違いを認識することは，複数テキスト要約における類似点と相違点の検出に該当する．そこで，本研究では，複数テキスト要約という観点から，ある目的に関する典型的な手順を出力するシステムの開発を目指す．

本論文の構成は以下のとおりである．2 節では，関連研究について述べる．3 節では，手順オントロジーを自動的に構築する手法について述べる．4 節では，手順オントロジー構築のための基礎的な実験について報告し，5 節で本稿をまとめる．

2. 関連研究

近年，複数の類似した手順テキストから，共通手順を抽出する研究が行われるようになってきている．山肩ら[2]は，「肉じゃが」や「カルボナーラ」などのクエリを用いて検索した料理レシピ集合に対し，各レシピをその調理手順を表したフローチャートに変換・統合することで，典型的な調理手順(レシピツリー)を導出する手法を提案している．さらに，典型的なレシピツリーと個々のレシピを比較することで，個々のレシピの特徴を抽出している．これらは，1 節で述べた複数テキスト要約における類似点と相違点の検出の一種と捉えることができる．

料理レシピを対象にしたこの他の研究に，瀧本ら[3]のものがある．瀧本らは，複数の類似レシピから，その共通手順を抽出するタスクを，施設配置問題と捉えている．

高木ら[4]は，「バジルの育て方」などが記載された複数の手順テキストから，その類似点と相違点を検出し，それをひとつのフローチャートとして自動的にまとめ，出力する手法を提案している．

フローチャートを対象とした関連研究もある．近年では，myExperiment[2]や SHIWA[3]など，フローチャートを共有するサービスがはじまっており，これに伴い，あるフローチャートと類似するものを検索する技術の需要が出てきている．Starlinger ら[5]は，あるフローチャートと別のフローチャートがどの程度似ているのかを算出するため，2 つのフローチャート間の対応関係を取る様々な手法について検討している．

[2] http://www.myexperiment.org/
[3] http://www.shiwa-workflow.eu/

3. 手順オントロジーの自動構築

3.1. 特許からの手順テキストの抽出

本研究では特許から手順テキストを抽出する．特許から，手順について記載された請求項を検出し，本研究で扱える形にするために，新森らの請求項構造解析ツール[6]を利用した．

請求項は，一般に，「～し，～し，～した，～」のように，処理を順序的に記述する順序列挙形式や，「～と，～と，～とからなる，～」のように，構成要素を列挙する形で記述する構成要素列挙形式など，特許固有のいくつかの記述スタイルが存在する．新森らは，請求項の構造解析を修辞構造解析の一種と捉え，手がかり語に基づいた請求項構造解析手法を提案している．例えば，図 1 の請求項を，新森らのツールを用いて解析すると，図 2 のような解析木が得られる．図 2 は，図 1 の請求が 5 つの部分文書に分割され，そのうちの最初の 4 つが，この解析木のヘッダ用語「食器洗浄乾燥機」と係り受け関係にあることを示している．また，その関係として“Procedure”(手順)というラベルが付与されており，以上から，この請求項は 4 つの手順から構成される食器洗浄乾燥機に関するものであることが分かる．

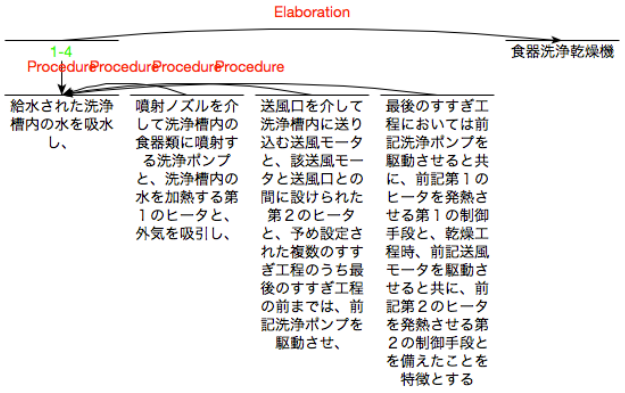

図 2 新森らのツールを用いて図 1 の請求項を解析した結果[4](特開 1999-178777)

Fig 2. Analysis result of a patent claim in Fig. 1 using Shinmori's tool (Japanese published unexamined application 1999-178777)

以下に，特許からの手順テキストの抽出手順について述べる．まず，新森らのツールを用い，1993～2013 年の公開公報に含まれるすべての第一請求項を解析した．次に，ひとつの請求項に Procedure 関係を 3 つ以上含む請求項を抽出し[5]，それらをヘッダ用語ごとに分

[4] 解析木の可視化には RSTTool (http://www.wagsoft.com/RSTTool/)を用いた．
[5]類似請求項と対応付けを行う際，手順の数があまりに少ないと，適切な対応付けができないと考えたため．

類した.

以上述べた処理の結果, 1,226,498 個の手順型の請求項が抽出された. これらの請求項に含まれるヘッダ用語の異なり数は 77,486 であった. 図 3 に, 請求項数の多いヘッダ用語の上位 10 件を示す. なお, 各文字列の後ろの数値は請求項数を示す.

画像形成装置 (23091)
半導体装置 (15057)
半導体装置の製造方法 (12903)
画像処理装置 (6626)
液晶表示装置 (7952)
記録媒体 (7752)
遊技機 (6444)
半導体記憶装置(5353)
情報処理装置 (4893)
画像処理方法 (4327)

図 3　請求項数の多いヘッダ用語上位 10 件
Fig. 3　Top 10 header terms having many patent claims

3.2. 複数の手順テキストの要約

3.1 節で述べた手法でヘッダ用語ごとにまとめられた請求項をいくつか調べたところ, 同一のヘッダ用語であっても, 請求項には多様性があることが分かった. 例えば, 図 3 の「画像形成装置」の場合, 画像形成装置にはコピー機, ファックス, プリンタなどが含まれている. また, 同じプリンタでも, レーザプリンタとインクジェットプリンタでは, 仕組み自体が異なるため, 手順の対応付けにそもそも馴染まないという問題がある. そこで, ヘッダ用語ごとにまとめられた請求項の集合を, bayon[6]を用いてクラスタリングし, 内容の近いものごとにまとめた. これらの請求項を対象に, 要約を行った.

一般的な複数テキスト要約と同様, テキスト間の類似点を検出する. ここで, 手順テキストの場合は, 以下の問題を考慮する必要がある.

- ある手順テキストと別の手順テキストの各手順が 1 対 1 で対応するとは限らず, 場合によっては 1 対多や多対多で対応する可能性がある.
- ある手順テキストでは A→B の順で出現した手順が, 別の手順テキストでは B→A の順で出現する可能性がある.

以上の問題を考慮した類似点検出を実現するため, 本研究では, 統計的機械翻訳技術を利用する. 統計的機械翻訳とは, 大量の対訳文から統計情報に基づいてモデルを学習し, そのモデルを用いて翻訳を実現する技術のことである. 統計的機械翻訳の中でも, 特に句に基づく機械翻訳では, 句の順序の入れ替えを考慮しつつ, 文単位の対訳を句単位の対訳に分解して, 翻訳モデルを構築する. 今, 統計的機械翻訳の入力となる対訳文の代わりに, 類似する手順の対を入力とすれば, 上記の問題を考慮した 2 つの手順テキスト間の類似点の検出が実現できると考えられる.

ここで, 対訳文の代わりに手順テキストを統計的機械翻訳の入力とするには, そもそも手順テキストをどのような形式で表現するのかを検討する必要がある. 今回は, 手順テキスト中の各手順を, その手順の最後に出現する動詞(自立語)またはサ変名詞とし, 手順テキスト全体を動詞列として表現した. 例えば, 図 2 の例は,「吸水 吸引 駆動 制御[7]」の動詞列として表現される. なお, 手順の最後に出現する動詞が「行う」「実行」「動作」といった手順の内容を示さない一般的な動詞の場合は, それよりひとつ前の動詞を用いる. また, 対訳文の代わりに入力とする手順テキストの対は, 前述のクラスタリングの結果でまとめられた請求項の任意の 2 対を用いるが, その際, 手順数に 2 以上差がある対は除外した. また, 2 つの手順テキストから生成された動詞列間で, 動詞が 2 つ以上一致する場合のみ統計的機械翻訳の入力として用いた. なお, 統計的機械翻訳システムとして, cicada[8]を利用した.

手順テキスト間の類似点を検出した後, 各クラスタの代表手順テキスト(クラスタの中心ベクトルから最も近いテキスト)の各手順が, 上述の cicada により, クラスタ内の他の手順テキストと対応付けられた場合に, その手順を複数テキスト要約の結果として出力する.

4. 手順オントロジー検索システムの構築

4.1. システムの動作例

3 節で述べた手法に基づいて, 手順オントロジー検索システムを構築した. 図 4 は「乾燥機」で検索した結果を示している.「乾燥機」を含むすべての用語が検索結果として表示される.

[6] https://code.google.com/p/bayon/wiki/Tutorial_ja bayon 実行時のオプション “-idf -l 1.5”

[7] 「を備えたことを特徴とする」などの定型表現は事前に削除する.

[8] http://www2.nict.go.jp/univ-com/multi_trans/cicada/

図 4　システム動作例 1

Fig. 4　System snapshot 1

図 4 において，ユーザが「真空乾燥機」という用語の(手順)をクリックすると，真空乾燥機の手順の要約が図 5 のように表示される．

```
<rst>
 <header>
  <relations file="d:/USR/SHINMORI/research/pat_browser/RSTTool27Mod
/Relation-Sets/Patent.rel"></relations>
 </header>
  <segment id=201 parent=2001 relname="procedure">真空ポンプ（２０）により吸引することによって、外気に対して気密の真空乾燥室（１０）を減圧して、真空乾燥室（１０）内に保持した衣料その他の物体の水分の蒸発を促進し、</segment>
  <segment id=202 parent=2001 relname="procedure">真空乾燥室（１０）で発生した水蒸気を含む気体を、真空乾燥室（１０）と真空ポンプ（２０）との間に介在させる冷凍機（３０）に導き、</segment>
  <segment id=203 parent=2001 relname="procedure">その冷却コイル（３１）に接触させて、水蒸気を凝固させて除去すると共に、前記冷却コイル（３１）を通った冷媒を冷凍用コンプレッサー（３２）で圧縮し、</segment>
  <segment id=204 parent=2001 relname="procedure">その圧縮した冷媒を液化コイル（３３）に通し、</segment>
  <segment id=205 parent=2001 relname="procedure">外部に放熱させるか、水冷するかして液化させると共に、その液化した冷媒を液化コイル（３３）の出口から冷却コイル（３１）の入り口に導き断熱膨張させて、降温させるようにしてなる真空乾燥機において、前記液化コイル（３３）を真空乾燥室（１０）内の衣類その他の物体と接触させる如く配置したことを特徴とする</segment>
  <segment id=206>真空乾燥機</segment>
<group id=2001 type="multinuc" parent=2002 relname="span">
<group id=2002 type="span" parent=206 relname="elaboration">
</rst>
```

図 5　システム動作例 2

Fig. 5　System snapshot 2

なお，図 4 において，(構成要素)というリンクをクリックすると，各用語の典型的な構成要素が表示される．これは，新森らのシステムを用いて請求項を解析し，手順を抽出するのと全く同じやり方で，構成要素(Component というラベルが付与された文字列)を抽出し，類似請求項間で類似構成要素を検出し，それらが要約として出力される．

4.2. 考察

評価用データが出来ていないため，実際のシステムの出力例を見て気づいた点についていくつか述べる．まず，統計的機械翻訳を用いた手順テキスト要約作成手法について，入力となる手順テキスト対は，かなり類似度の高いものを準備しなければ，類似手順の検出結果はかなり悪い．現状では，手順テキスト対は類似度が非常に高いものだけを用いているが，その結果，同一組織から出願された別の特許が手順テキスト対として選択される傾向にあり，手順オントロジーとしての一般性に欠けるという問題点がある．

次に，手順オントロジーを構築する用語について述べる．3.1 節で述べたとおり，今回は 77,486 語に関する手順オントロジーを構築した．この 77.486 語を詳しく見ると，同義語が複数存在していることが分かった．例えば，図 4 の例では，食器洗い乾燥機と食器洗浄乾燥機は同義語であるが，現在は別の用語として扱われている．このような同義語は事前に何らかの方法で統合した上で手順オントロジーを構築する必要があると考えられる．

最後に，同義語問題に関連して，用語間の上位，下位関係にも配慮する必要があると考えている．図 4 に表示されている用語の中で，乾燥機，衣類乾燥機，ドラム式衣類乾燥機の間には上位−下位関係がある．2 つの用語間に上位−下位関係があれば，それぞれの用語から生成される手順オントロジーにも何らかの関係があるはずだが，現在はその点については全く考慮していない．今後は，ある用語とその手順テキスト集合だけでなく，その用語と上位，下位関係にある用語についても何らかの配慮をして手順オントロジーを構築する必要があると思われる．

5. おわりに

本研究では，統計的機械翻訳技術を用いて，類似する複数の特許に関する手順テキストから典型的な手順を抽出することでオントロジーを構築する手法を提案した．今後は特許だけでなく，料理レシピにも提案手法を適用する．

文　献

[1] 奥村学，難波英嗣，“テキスト自動要約，” コロナ社，2005.

[2] 山肩洋子，今堀慎治，杉山祐一，田中克己，“レシピフローグラフを介したレシピ集合の要約と特徴抽出，” 電子情報通信学会技術研究報告，DE 研第 1 種研究会 データ工学と食メディア，Vol. 113, No. 214, DE2013-36, pp.43-48，2013.

[3] 瀧本洋喜，笹野遼平，高村大也，奥村学. (2015) “施設配置問題に基づく同一料理のレシピ集合からの基本手順の抽出” 言語処理学会第 21 回年次大会発表論文集，pp. 1092-1095.

[4] 高木優，藤井敦．(2015) “手順テキストを対象とした比較対象要約” 言語処理学会第 21 回年次大会発表論文集，pp. 573-576.

[5] Johannes Starlinger, Bryan Brancotte, Sarah Cohen-Boulakia, and Ulf Leser. (2014) “Similarity Search for Scientific Workflows” Proceedings of the VLDB Endowment, Vol. 7, No. 12, pp.1143-1154.

[6] 新森昭宏，奥村学，丸山雄三，岩山真．(2004) “手がかり句を用いた特許請求項の構造解析” 情報処理学会論文誌，Vol.45，No.3，pp.891-905.