手順テキストの概要把握を目的としたアウトライン化手法

An Outlining Method for Grasping the Overview of Procedural Text

東京工業大学大学院情報理工学研究科准教授 **藤井　敦**

PROFILE: 1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より現職。自然言語処理、情報検索、Web マイニング、特許情報処理の研究に従事。

1 はじめに

我々が何らかの手順を理解したり、また他者に説明する際は、枝葉末節にこだわるよりも全体の大まかな流れを把握することが重要な場合がある。例えば、東京の自宅から大阪にある道頓堀までの行き方は、以下に示す 3 つの段階で説明することができる。

（1）自宅の最寄り駅から東京駅へ向かう。
（2）東京駅から新幹線で新大阪駅へ向かう。
（3）新大阪駅から難波駅へ向かう。

各段階で乗車する路線や乗り換えに関する情報は必要に応じて説明すればよい。相手によるものの「乗車券を買う」や「靴を履く」のような当たり前の動作を逐一説明することは稀であろう。

手順に関する別の例として、以下に示すカレーライスのレシピを考える。

（1）にんじん、じゃがいもを一口大に切る。玉ねぎはくし形に切る。
（2）牛肉を炒め、焼き色がついたらにんじん、じゃがいも、玉ねぎを加え、炒め合わせる。
（3）水を加えて、野菜が柔らかくなるまで煮込む。
（4）カレー粉を加えて、弱火でさらに煮込む。

上記レシピの大まかな流れは、例えば以下のように各段階を代表する動作の列で表現することができる。

（1）野菜を切る。
（2）材料を炒める。
（3）材料を煮込む。
（4）カレー粉を加える。

以上の例は、手順を構成する動作を特定し、必要に応じて、「野菜を切る」のように複数の動作を抽象化することで、対象となる手順の骨格を抽出する処理である。この処理を本稿では「アウトライン化」と呼び、コンピュータによる手順テキストのアウトライン化手法を提案する[1]。現在は料理レシピを対象として基礎研究を行っている。長期的には、特許明細書に書かれた方法論、学術文献に書かれたアルゴリズム、生物学や化学の実験プロトコル、製品の操作説明書など様々な手順テキストへの応用を目指している。

2 提案するアウトライン化手法

本研究で提案するアウトライン化手法は、料理レシピのテキストを構成する各段階（セグメント）から、「野菜を切る」や「材料を炒める」のような動作表現を「見出し」として抽出し、それらを連結することでアウトラインを生成する。料理レシピがセグメントに分割されていない場合は、既存のテキスト分割手法を利用する。見出しの形式は「野菜を切る」のように「名詞＋助詞＋動詞」で構成される述語項関係を想定している。また、名詞と助詞が省略された動詞だけの場合もある。

次に、見出しの形式となる述語項関係を各セグメントから抽出する。各述語項関係の特徴を表す素性ベクトルを作成し、サポートベクターマシン（SVM）を用いて述語項関係が見出しとなるべきか否かの二値分類を行う。料理の各段階に見出しが付与されたレシピの書籍から必要な項目を電子化し、学習データとして利用する。

見出しには大きく二種類ある。まず、セグメント内の表記がそのまま見出しに用いられる「抜粋型」がある。また、セグメント内の表記が必要に応じて編集されている「抽象型」がある。例えば、1 章に示したカレーライスのレシピでは「にんじん」、「じゃがいも」、「玉ねぎ」が「野菜」に抽象化されている。本研究は抜粋型を主に扱い、複数の食材を「野菜」や「肉」のような上位概念にまとめるような抽象型を併用する。

以下、本研究で使用する素性について説明する。前半の 4 つは既存の要約手法で用いられている素性であり、それ以外は本研究で提案する素性である。

単語の重要度

単語の重要度を示す指標として、情報検索で用いられる TF・IDF がある。これは、多くの文書に出現する単語は重要度が低く、特定の文書だけに出現する単語は重要度が高いという考えに基づいている。本研究では一つのセグメントを一つの文書とする。

位置情報

段落の開始文が重要であるなど、文の位置によって重要度が異なる。本研究では「述語項関係の位置情報」として用いる。述語項関係の位置をセグメント単位及び料理レシピ単位でそれぞれ求める。

タイトル

文書のタイトル、章、節などの見出しに現れる内容語を含む文が重要であることが多い。本研究では料理名をタイトルとして扱い、述語項関係内にある名詞が料理名に含まれれば、素性値を 1 にし、含まれなければ 0 にする。

文間・単語間のつながり

テキスト中で他の文と強い関連がある文を重要視する。本研究では、「述語項関係間のつながり」という素性を設け、料理レシピ内にある述語項関係が他の述語項関係と同じ単語（名詞か動詞）を共有すれば、共有する述語項関係の数を求め、名詞の数と動詞の数の総和を素性値とする。

加熱調理の動詞

「炒める」などの加熱調理を表す動詞を含む述語項関係は見出しになりやすい。加熱に関する動詞が述語項関係内に出現した場合は素性値を 1 にして、出現しない場合は 0 とする。

見出しにおける動詞の頻度

セグメント内にある動詞に対し、「見出しになったセグメント数」を「出現したセグメント数」で割って素性値に変換する。

セグメント内における動詞の出現頻度

セグメント内に繰り返し出現する動詞は見出しになりやすい。そこで、述語項関係内に含まれる動詞に対して、「動詞がセグメント内に出現する回数」を「セグメント内にある全ての動詞が出現する回数の最大値」で割った値を素性値とする。

動詞の出現履歴

述語項関係に含まれる動詞、特に加熱調理の動詞は、そのセグメントで初めて出現するか否かで見出しになる可能性が変わる。動詞が既出の場合は、前のセグメントで見出しになった可能性があるので、そのセグメントでは見出しになりにくい。述語項関係内の動詞がそのセグメントで初出の場合は素性値を 1、既出の場合は 0 とする。

レシピ全体における動詞の DF

ある動詞が一つの料理レシピ内で複数のセグメントに現れると、その動詞が料理内で重要な動作である可能性がある。DF は、述語項関係内の動詞について一つの料理レシピ内の「動詞を含むセグメント数」を「全セグメント数」で割った値とする。

名詞の属性

述語項関係内の名詞が「豚肉」などの食材であれば見出しになりやすく、「フライパン」などの調理道具、「こしょう」などの調味料は食材と比べると見出しになりにくい。「強火」、「中火」、「弱火」の 3 つは火加減を表す名詞であり、ほとんど見出しにならない。そこで、名詞

辞書を作成し、述語項関係に含まれる名詞の属性によって異なる素性値を与える。

上位概念語

「野菜」や「材料」などの上位概念語は複数の対象を総称するため、見出しに用いられる傾向にある。上位概念語の辞書を作成し、上位概念語を含む述語項関係の素性値を1とし，それ以外は0とする。

セグメント番号

同じ動詞であっても、セグメント番号によって見出しになる可能性が変わる。料理レシピ内の「セグメントの通し番号」を「全セグメント数」で割った値を素性値とする。

3 評価実験

セグメントごとに見出しが付与されているレシピの書籍から合計 306 件のレシピを収集して評価実験に使用した。当該レシピ集合には、1465 件のセグメントが含まれている。このレシピ集合を用いて 10 分割の交差検定を行った。この際、書籍によって見出しの付け方に差異があるため、特定の書籍から収集したレシピが特定の分割に集中しないようにした。述語項関係の抽出に失敗した場合は、本研究で提案した素性の有効性だけを評価することが困難になる。そこで、正解となる述語項関係が得られなかったセグメントはテストデータから削除した。ここで、正解の定義として「完全一致」だけでなく、動詞だけが一致すればよい「部分一致」も用いた。その結果、完全一致と部分一致の場合でそれぞれ 128 件と 264 件のセグメントがテストデータとして使用された。比較した手法を以下に示す。

- ランダム：セグメントから無作為に抽出した述語項関係を見出しとする。
- 既存の素性： 既存の素性だけを用いて SVM による二値分類を行う。
- 本手法： 既存の素性と提案した素性を用いて SVM による二値分類を行う。

以下にランダム、既存の素性、本手法の正解率を示す。本手法と既存の素性は共にランダム抽出での正解率を上回った。また、本手法は既存の素性より部分一致と完全一致の両方で正解率を上回ったことから、本手法の素性が有効であることが分かった。

正解の基準	ランダム	既存の素性	提案手法
部分一致	21.32	35.79	64.50
完全一致	20.07	28.95	50.12

4 おわりに

本研究は料理レシピを対象としたアウトライン化手法を提案した。料理レシピ特有の素性を提案し、評価実験によって有効性を示した。今回、抽象型の見出しでは上位概念語への置換だけを行った。今後の課題として、調味料をふる動作を表す「調味する」や料理の準備段階を示す「下準備をする」など、様々な抽象型による見出しの生成を実現する必要がある。また、本手法を料理レシピ以外の手順テキストに応用する必要がある。

参考文献

[1] 西原 弘真、苅米 志帆乃、藤井 敦 . 料理レシピを対象としたアウトライン型自動要約 . 情報処理学会第 89 回デジタル・ドキュメント研究会 第 110 回情報基礎とアクセス研究会合同研究会、2013.