D-001

ハウツー情報の典型性と独自性の把握のための集約的提示
Aggregation of How-to Information for Understanding Typicality and Uniqueness

湯本高行†
Takayuki Yumoto

1. はじめに

料理のレシピやソフトウェアのインストール方法などのハウツー情報を検索する需要は高い．ハウツー情報に特化した検索・閲覧手法は実現されていない[1]．そこで，目的は同じだが，手順の異なるハウツー情報を集約して提示する手法を提案する．

本研究では，ハウツー情報を構成する手順を操作と対象のペアとして表現し，そのリストとしてハウツー情報を表現する．これに対して，料理のレシピについては，Wang ら[2]や山肩ら[3]はグラフによって表現する方法を提案しているが，これは複数の材料に対する操作が最終的に 1 つの料理としてまとまることを前提としているためである．一方，ハウツー情報の中にはこのようにならないものも存在するため，本研究では手順のリストを用いる．手順はサポートベクトルマシン[4](以下，SVM)を用いて抽出する．次に抽出した手順に対して，Wikipedia などのデータを基に作成した辞書を用いて同定を行う．その後，手順の前後関係から手順の最適な順序を決定し，集約を行う．これらの手法を応用し，ハウツー情報の集約提示システムを開発した．このシステムを用いることで，典型性に基づくフィルタリングや特定の手順を含むページの検索が可能である．

2. ハウツー文からの手順の抽出

2.1 機械学習による操作と対象の個別抽出

文に対して係り受け解析を行い，各文節が対象または操作を含むかどうかを判定する分類器を SVM で構築する．操作の方が,判定が容易であると考え，先に操作を含むかどうかの判定を行い，その結果を用いて対象を含むかどうかの判定を行う．対象を含むと判定された文節から操作を含むと判定された文節に対して係り受け関係がある場合，それぞれの文節から対象および操作を抽出し，そのペアを手順として抽出する．

2.1.1 操作分類器の構築

操作分類器では，任意の動詞句および一部の名詞句を判定対象とする．対象の名詞句は末尾または末尾の文節と並立関係にあるものに限定する．操作分類器の素性としては以下を使用する．

- 文節の末尾の形態素の品詞(末尾が句読点の場合はその前の形態素の品詞)
- 文節に含まれる助詞，助動詞，非自立動詞．これらは，学習データを素性ベクトルに変換する前に，助詞，助動詞，非自立動詞を列挙しておくことで，含むか否かを素性ベクトルの独立した成分に対応させる．
- 係り先の種類(体言，用言，末尾)
- 主辞の品詞
- 助詞の「を」を含む文節から係っているか
- 助詞の「は」を含む文節から係っているか

2.1.2 対象分類器の構築

対象分類器においては，対象と操作が同時に出現する文節はないと考え，操作を含まない名詞句を対象にする．
対象分類器の素性としては以下を使用する．

- 数詞を含むかどうか
- 名詞の品詞細分類
- 文節に含まれる助詞：助詞ごとに別の素性として表現し，それぞれ 0/1 の 2 値で表現する．なお，並立関係で係っている場合は係り先の文節の助詞を使用する．
- 操作を含む文節に係るかどうか：係り先の文節が操作を含むかどうかを 0/1 の 2 値で表現する．なお，並立関係を経由して間接的に操作に係る場合も直接係っている場合と同様に扱う．
- 表 1 に示すストップワードを含んでいるか

表 1 ストップワード

こと	事	場合	時	とき
もの	モノ	物	者	

2.1.3 対象分類器の構築

対象を含むと判定された文節から操作を含むと判定された文節に対して係り受け関係がある場合，それぞれの文節から主辞の語を抽出し，そのペアを手順として抽出する．なお，係り受け関係は直接的なものではなく，間接的なものも含む．本研究で扱う間接的な係り受け関係は，体言を含む文節 A が文末でない述部 B より前に出現し，文末の述部 C に係る場合において，A と B の関係である．たとえば，「A は B して C する」という文では，A と C は直接的な係り受け関係にあり，A と B は間接的な係り受け関係にある．

2.2 機械学習による操作-対象ペアの同時抽出

2.1 では，操作と対象を順に抽出していた．この方法では，操作の抽出で失敗があった場合，その影響を対象が受けやすいと考えられる．そこで，操作と対象をそれぞれ含む可能性がある文節のペアに対して，操作と対象の両方を含む文節ペアかどうかの判定を行う．判定には，SVM で構築した 2 値分類器を用いる．

この分類器の判定対象は，操作分類器の対象としていた文節と対象分類器の対象としていた文節のうち，それに係る文節のペアとする．また，ペアの素性ベクトルは操作分類器の素性ベクトルと対象分類器の素性ベクトルから「操作を含む文節に係るかどうか」の素性を除いたものを連結したベクトルを使用する．

対象と操作の両方を含むと判定された文節ペアのそれぞれから主辞の語を抽出し，そのペアを手順とする．

† 兵庫県立大学大学院工学研究科
Graduate School of Engineering, University of Hyogo

2.3 実験

肉じゃがの作り方および Ubuntu のインストールの方法について書かれたページを 5 ページずつ用意した．これらのページでハウツー情報が書かれた部分に対して係り受け解析を行い，正しく解析できた 110 文を対象とした．これらの文に対して，各タスクについて人手で正解を作成したデータを使用した．また，形態素解析器には Mecab[5], 係り受け解析器には Cabocha[6]を用いた．

2.3.1 操作と対象それぞれの抽出手法の評価

2.1.1 と 2.1.2 のそれぞれの手法に対して，10 分割交差検定を行い，操作と対象のそれぞれを抽出した．評価指標としては，以下で定義される正解率(acc)，適合率(P)，再現率(R)，F 値(F)を用いた．

$$\mathrm{acc} = \frac{A+D}{A+B+C+D}, P = \frac{A}{A+B}, R = \frac{A}{A+C}, F = \frac{2PR}{P+R}$$

なお，A, B, C, D は表 2 の混同行列の各要素に分類されたアイテムの数である．また，SVM のカーネル関数には RBF カーネルを用い，パラメータは F 値を最大化するようにグリッドサーチを行って求めた．結果を表 3 に示す．なお，対象分類器に用いる素性「操作を含む文節に係るかどうか」は操作が正しく抽出できている状態の値を用いた．操作の抽出結果を用いた場合，文献[7]で示したように各指標が 5 ポイント程度低下すると考えられる．

表 2 対象(操作)の判定結果の混同行列

	対象(操作)	対象(操作)でない
対象(操作)と判定	A	B
対象(操作)でないと判定	C	D

表 3 操作, 対象の抽出結果(%)

	正解率	適合率	再現率	F 値
操作	85.5	88.2	86.4	87.3
対象	90.5	80.5	89.6	84.8

2.3.2 手順の抽出手法の評価

2.3.1 で用いたデータセットのうち，間接的なものを含む係り受け関係にある 2 語を候補とし，そのうち，対象と操作のペアであるものを正解とした．適合率は抽出した手順のうち正解であるものの割合，再現率はデータセットに含まれる手順のうち抽出できたものの割合，F 値は適合率と再現率の調和平均と定義し，これらを評価指標に用いた．

2.1(個別抽出)の手法の結果としては，実験 2.3.1 で抽出した操作および対象の候補に対して，2.1.3 の手法でペアを作成して用いた．2.2(同時抽出)の手法の評価では 10 分割交差検定を用いた．SVM のカーネル関数には RBF カーネルを用い，パラメータは F 値が最大になるようにグリッドサーチを行って求めた．結果を表 4 に示す．同時抽出は個別抽出に比べて再現率が高く，その結果として F 値も高くなっている．

表 4 手順の抽出手法の比較(%)

	適合率	再現率	F 値
個別抽出	75.7	64.7	69.8
同時抽出	72.6	82.4	77.2

3. ハウツー文書の要約の生成

3.1 手順の同定

2 つの手順において対象と操作のそれぞれが同義関係にあるとき，2 つの手順は同一であると見なす．対象および操作の語の同義性は，Wikipedia コーパス，日本語 WordNet[8]，動詞含意関係データベース[9]のそれぞれから構築した同義関係辞書を用い，いずれかに同義関係が定義されていれば同義と見なす．各同義辞書の構築方法は以下のとおりである．

Wikipedia コーパスを用いた同義辞書の構築では，リダイレクトされる記事名とリダイレクト先の URL，Wikipedia 記事にリンクするアンカーテキストとリンク先の URL のそれぞれでペアを抽出し，同一の URL とペアになっている文字列を同義関係にあると見なす．適合性を高めるためには，頻度などによって候補を絞り込むことが考えられるが，今回は網羅性が重要だと考え，頻度による制限を行わない．この辞書は主に対象に対して用いることを想定している．

また，日本語 WordNet では同一の概念に属する語を同義語と見なす．この辞書は対象および操作に対して用いることを想定している．

動詞含意関係データベースを用いた同義辞書の構築では，当該データベースに登録されている含意関係のうち，直接の親概念が共通している語は同義関係にあると見なす．ただし，「行う」は異なる動作を表す動詞と含意関係が定義されていたため，対象から除く．この辞書は主に操作に対して用いることを想定している．

3.2 手順の順序の決定

手順の順序の決定では，任意の 2 つの手順のどちらが先に出現しやすいかを表現する順序行列を用いる．順序行列では，それぞれの行と列が手順に対応し，行と列で対応する手順は同じ順に並んでいるとする．また，要素は，行に対応する手順が列に対応する手順より前に出現する回数を行に対応する手順が列に対応する手順より後に出現する回数から引いた値とする．これにより，値が大きいほど，行に対応する手順が列に対応する手順より先に出現しやすく，値が小さいほど，行に対応する手順が列に対応する手順より後に出現しやすいことを表す．値が 0 に近い場合は順序関係を特定できる情報が少ないことを表す．なお，対角成分は以降の計算には使用しないが，便宜的に 0 とする．また，順序行列は交代行列である．

手順の順序の決定は以下のように行う．

(1) 順序行列の作成
(2) 手順の順序の仮決定
(3) 手順の順序の最適化

まず，順序行列の作成では，行と列の順番は出現するページ数(以下，DF)の降順に並んでいるとする．値の計算は，各ページに含まれる任意の手順の組合せにおいて，行に対応する手順が列に対応する手順より先に出現する場合

は値に 1 を加え，行に対応する手順が列に対応する手順より後に出現する場合は値から 1 を引くことにより行う．

次に，手順の順序の仮決定では，DF の降順に手順の挿入位置を決定する．Lを手順の仮の順序を表すリストとし，初期値として，DF が最大の手順 1 つのみからなるとする．Lの先頭から順序行列を用いて，注目している手順との順序関係を調べていき，注目している手順の方より後方にある手順が見つかったらその位置に挿入する．見つからなかった場合はLの末尾に追加する．このアルゴリズムは図 1 のように表現できる．なお，L_jをLのj 番目に位置する手順の番号とする．Sを DF の降順に手順が並んだリストとし，各手順を$S_1, S_2, \ldots$，手順の数を$|S|$と表すとする．また，Mは順序行列である．この処理の後，順序行列の行と列はLの順序と同じになるように入れ換えを行う．

```
for i = 1,…,|S|
  for j = 1,…,i − 1
    if M_{L_j,i} > 0
      Lのj番目にiを挿入
      次のiに進む
    end if
  end for
  L のi 番目にiを追加
end for
```

図 1 順序の仮決定のアルゴリズム

最後に，手順の順序の最適化を行う．手順の行および列が理想的な順序で並んでいる場合，順序行列の下三角部分はすべて負になっている．しかし，実際には手順が逆転している場合があるため，正になっている部分が存在する．また，仮決定のアルゴリズムでは手順のすべての組合せについて順序を比較しているわけではないため，得られた順序が最適であるとは限らない．そこで下三角部分に正の要素がなるべく少なくなるように手順を前方に移動させる．具体的には順序行列の対角成分から左方向に向かって値を足していき，累積和を算出する．この累積和が，最大となった位置が最も矛盾が少ないと考え，その位置に移動させる．

ここで例を示す．図 2 に示す順序行列Mにおいて，手順 1～5 の順に並んでいる手順 5 の位置を変更することを考える．$M_{5,5}$から左側に向かって累積和を求めると順に-1, 0, 2, 1 となり，$M_{5,2}$までの累積和が最大となるので，手順 5 を手順 2 の位置に移動させる．このとき手順は 1, 5, 2, 3, 4 の順になる．この入れ換えを順序行列Mに反映させたものが図 3 に示す行列である．下三角部分に正の値は残っているが，1 箇所だけであり，値の総和も小さくなっていることがわかる．

0	4	2	3	1	…
−4	0	3	2	−2	…
−2	−3	0	2	−1	…
−3	−2	−2	0	1	…
−1	2	1	−1	0	…
…	…	…	…	…	0

図 2 最適化前の順序行列の例

0	1	4	2	3	…
−1	0	2	1	−1	…
−4	−2	0	3	2	…
−2	−1	−3	0	2	…
−3	1	−2	−2	0	…
…	…	…	…	…	0

図 3 最適化後の順序行列の例

表 5 手順の決定の実行列

対象	操作
じゃがいも	むく
じゃがいも	切る
牛肉	切る
じゃがいも	水にさらす
玉葱	切る
にんじん	むく
にんじん	切る
糸コンニャク	切る
水	入れる
サラダ油	熱す
牛肉	入れる
牛肉	炒める
玉葱	入れる
糸コンニャク	入れる
玉葱	炒める
じゃがいも	入れる
だし汁	入れる
にんじん	入れる
砂糖	入れる
醤油	入れる
酒	入れる
*	煮る
*	入れる

3.3 実行例

2.3 の実験で使用した肉じゃがのレシピのうち，2 件以上のページに出現した手順の順序を決定した結果を表 5 に示す．なお，使用した手順の抽出は人手で行ったため，現在のアルゴリズムでは抽出できない「水にさらす」という操作や調理中の材料全体を表す「*」という対象が含まれている．(水, 入れる)の後に(サラダ油, 熱する)があるなど，一

部で順序の逆転が見られる．これには 2 つの原因が考えられる．まず，他の手順との頻度の低い場合に，順序の決定に使用できる情報が不足したというものである．この場合の「水」は代わりに「だし汁」が使われる場合があるため，出現頻度が低くなっている．もう 1 つの原因としては，同じ手順が別の文脈で使われることにより，対応する順序行列の値が 0 に近づき，共起頻度が少ないのと同じ状況が発生するというものである．この例では，(水, 入れる)という手順が糸コンニャクの下準備のために行われる場合と肉じゃが全体の調理として行われる場合がある．

4. ハウツー情報の集約的提示システム

以上の手法を応用し，ハウツー情報の集約的提示システムを開発した．図 4 にそのインターフェースを示す．①の部分がハウツー情報の入力部分である．②では入力されたハウツー情報について，2 節で述べた手法で手順を抽出し，3 節で述べた手法で順序を決定して一覧を提示する．このとき，提示される手順は①のスライダーバーにより頻度によってフィルタリングを行うことができる．②の各手順にはチェックボックスがついており，これを選択することにより，その手順を含むページのみが③に提示される．また，③に表示されたページにはラジオボタンがついており，これを選択することによって，そのページに含まれる手順を②に表示することが可能である．

利用のシナリオとしては以下が考えられる．

- 典型的な手順を把握するために，①を用いて，②に表示される手順を絞り込んで表示する．
- 上記とは逆に①を用いて②から他では用いていない，独自性の高い手順を発見する．その手順を含むページを③で検索する．
- ③に表示されるページのタイトルから閲覧するページの目星をつけ，そのページに含まれる手順を②に表示することで，概要を把握する．つまり，要約として②を使用する．

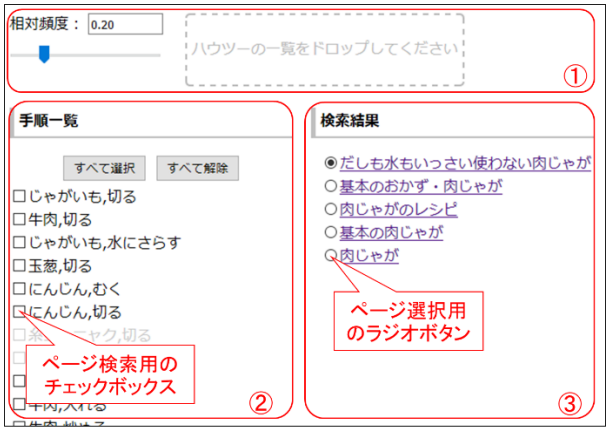

図 4 ハウツー情報の提示システム

5. おわりに

本研究では，料理のレシピやソフトウェアのインストール方法などのハウツー情報を対象に，目的は同じだが，手順の異なる情報を集約して提示する手法を提案した．まず，ハウツー情報を構成する手順を操作と対象のペアとして SVM を用いて抽出する．ここでは，操作と対象を個別に抽出する手法と同時に抽出する手法の 2 つを提案し，実験により後者の方が再現率および F 値が高いことを示した．次に抽出した手順に対して，Wikipedia などのデータを基に作成した辞書を用いて同定を行う．その後，手順の前後関係から手順の最適な順序を決定し，集約を行う．これらの手法を応用し，ハウツー情報を集約して提示するシステムを開発した．このシステムでは，典型性に基づくフィルタリングや独自性の高い手順を含むページの検索が可能であることを示した．今後は，作成したシステムを用いた被験者実験を行う．

謝辞

本研究の一部は，科研費若手研究(B)「情報の詳細関係に基づく Web ページの組織化」（課題番号：24700097）によるものである．

参考文献

[1] 野中諒志, 湯本高行, 新居学, 高橋豐, “概要・詳細の見やすさに基づく手法情報のランキングと閲覧支援”, WebDB Forum 2010, 2A-1 (2010).
[2] Liping Wang, Qing Li, Na Li, Guozhu Dong, and Yu Yang, Substructure similarity measurement in Chinese recipes. In Proceedings of the 17th international conference on World Wide Web (WWW '08), pp.979-988 (2008).
[3] 山肩洋子, 今堀慎治, 森信介, 田中克己, “ワークフロー表現を用いたレシピの典型性評価と典型的なレシピの生成”, 電子情報通信学会論文誌 D, Vol.J99-D, No.4, pp.378-391 (2016).
[4] Vladimir N. Vapnik, “Statistical Learning Theory”, Wiley-Interscience (1998).
[5] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004).
[6] 工藤拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析”，情報処理学会論文誌，Vol.43, No.6, pp.1834-1842 (2002).
[7] 湯本高行，“機械学習によるハウツー情報の手順の抽出とその応用”, 情報処理学会第 77 回全国大会, pp.507-508 (2015).
[8] “日本語 WordNet”, http://compling.hss.ntu.edu.sg/wnja/
[9] “ALAGIN 言語資源・音声資源サイト”, https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-2