3C-01

機械学習によるハウツー情報の手順の抽出とその応用

湯本　高行†

兵庫県立大学大学院工学研究科†

1. まえがき

料理のレシピやソフトウエ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が，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]．ハウツー情報は手順のリストとして表現することが可能で，手順は「対象」と「操作」により，構成される．

我々はルールベースでハウツー情報の手順を抽出する手法を開発したが，精度に問題があった[2]．本研究では，機械学習を用いて Web ページからハウツー情報の手順を抽出する方法を提案する．手順の抽出では，Support Vector Machine(以下，SVM)[3]を用いて操作と対象のそれぞれを抽出，それらを組み合わせることで手順を構成する．

また，抽出した手順をハウツー情報の検索や閲覧に応用する手法について述べる．

2. 提案手法

本研究では，ハウツー情報を手順のリストとして表現し，手順を対象と操作のペアとして表現する．たとえば，「肉を切ります」という文では，対象は「肉」，操作は「切る」である．対象は名詞，操作は名詞または動詞で表現される．

手順の抽出の際は，まず，操作の抽出を行い，その結果を用いて対象の抽出を行う．また，操作および対象の抽出では，文節に操作および対象を含むかどうかを判定する分類器をそれぞれ構築して用いる．その後，操作と対象を含む文節が以下のいずれかの条件を満たす場合に操作と対象のペアを手順として抽出する．

- 対象を含む文節から操作を含む文節に対して直接の係り受け関係があり，並立関係にない．
- 対象を含む文節から係る文節が並立関係にあり，その文節から並立関係以外の係り受け関係が操作を含む文節に対して存在する．

ここで，並立関係は，係り受け関係にある 2 つの文節が共に体言を含む，もしくは共に用言を含む場合に，係り元の文節に含まれる助詞を考慮して判定している．

後者の例として「肉と野菜を炒める」という文の係り受け関係を図 1 に示す．この例では，「肉と」が「野菜を」に並立関係で係る．また，「野菜を」は「炒める」に係るが，並立関係ではない．ここで，「肉と」が対象，「炒める」が操作を含む文節である場合，「肉と」は「野菜を」を介して「炒める」に係っており，後者の条件を満たす．

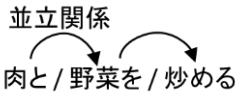


図 1 係り受け関係の例

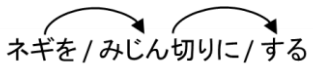


図 2 操作の補完の例

2.1 操作の抽出

文節が操作を含むかどうかを判定する分類器を SVM で構築し，この分類器で操作を含むと判断された文節から操作を抽出する．具体的には，操作を含むと判断された文節で最初に現れる自立語の名詞または動詞を操作として抽出する．また，抽出された操作が「する」の場合は，その文節に係る直近の文節から自立語の名詞を抽出し，これを操作とみなす．たとえば，「ネギをみじん切りにする」は図 2 に示すように「ネギを/みじん切りに/する」のような文節に分かれ，「ネギを」と「みじん切りに」が「する」に係る．このうち，「みじん切りに」の方が「する」に近いため，「みじん切り」が操作として抽出される．

文節が操作を含むかどうかを判定する分類器の素性として以下を使用する．

(1) 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)
(2) 係り先の種類(体言，用言，末尾)
(3) 文節に動詞を含むか
(4) 助詞の「を」を含む文節から係っているか
(5) 助詞の「は」を含む文節から係っているか

このうち，(1)，(2)は多値をとるため，排他的に 1 となる複数の 2 値の素性に分割する．たとえば，(2)は係り先が体言を含むか，用言を含むか，係り先がない(末尾かどうか)のそれぞれを 2 値で表す 3 つの素性で表現する．なお，形態素解析の結果が，自立語の名詞または未知語を体言，動詞，形容詞，形容動詞を用言とした．また，(3)～(5)はそれぞれ 2 値をとる素性である．

2.2 対象の抽出

対象は操作の抽出を行った後に抽出する．具体的には，SVM により文節が対象を含むかどうかを判断する分類器を構築し，対象を含むと判定された文節から自立語である名詞を抽出する．

文節が対象を含むかどうかを判定する分類器で使

Extraction of Processes in How-to Information by Machine Learning and Its Application

† Takayuki Yumoto, Graduate School of Engineering, University of Hyogo

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

用する素性は以下のとおりである．

- 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．
- 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．

前者の素性は，複数の素性で表現されるが，1 となる要素は多くとも 1 つであり，すべてが 0 になることもありうる．後者の素性では，係り先の文節との関係が並立関係の場合は，その先の係り先の文節が操作を含むかどうかで判断する．

3. 実験

提案手法の有用性を示すため，操作，対象，手順のそれぞれの抽出精度を評価した．形態素解析器には Mecab，係り受け解析器には Cabocha を使用した．

3.1 操作の抽出方法の評価

まず，学習および評価に使用するデータセットの作成方法について述べる．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．分割できた文の文節のうち，文末または動詞を含む文節に対して，操作を含むか否かで 2 値のラベルづけを行った．このデータセットを用いて素性ベクトルを作成したところ，素性の数は 11 であった．

評価として，10 分割交差検定を行い，操作抽出の正解率，適合率，再現率，F 値を算出し，比較する．正解率は文節を操作を含むか否かの 2 値に分類する問題として算出し，適合率，再現率，F 値は対象を発見する問題として算出する．なお，カーネルには 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た結果を用いた．結果は，正解率 79.7%，適合率 82.3%，再現率 82.3%，F 値 82.3%となった．

3.2 対象の抽出方法の評価

データセットは操作の抽出に使用した文を用い，対象を含む文節に対して 1，体言を含む文節のうち，ラベルがつけられていない文節は対象を含まないものとし，機械的に-1 とラベルづけした．このデータセットを用いて素性ベクトルを作成したところ，素性の数は 14 であった．

評価では，操作の抽出を提案手法で行った場合と手作業で行った場合の 2 通りについて，10 分割交差検定を行い，対象抽出の正解率，適合率，再現率，F 値を算出し，比較する．正解率は文節を対象を含むか否かの 2 値に分類する問題として算出し，適合率，再現率，F 値は対象を発見する問題として算出する．なお，カーネルには 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た結果を用いた．結果を表 1 に示す．

いずれも提案手法によって操作を発見した場合(表 1 の手法)に比べ，操作が正しく抽出されている場合(表 1 の手作業)の方が，良好な結果が得られている．この結果から，操作の抽出精度を上げることによって対象の抽出精度も向上することが期待できる．一方，適合率については操作が正しく抽出できている場合も 70%程度であり，別の素性を追加するなどの改善が必要である．

表 1 対象の抽出実験の結果

	正解率	適合率	再現率	F 値
手法	81.8	64.7	85.3	73.6
手作業	85.3	70.4	87.0	77.8

3.3 手順の抽出結果の評価

3.1 と 3.2 で抽出した操作と対象のうち，前述の条件を満たす係り受け関係のあるものを手順として抽出したところ，適合率は 64.7%であった．また，誤って抽出した手順はすべて対象でない語を誤って対象として抽出していた．このため，対象の抽出精度を向上させることにより，改善が見込まれる．

4. 手順抽出の応用

抽出した手順はハウツー情報を検索，閲覧する際にユーザの助けとなる．たとえば，手順をハウツー情報の要約として利用することが考えられる．さらに，ページに含まれる手順の数はハウツー情報としての詳しさを判断するための指標になりうる．たとえば，同一の目的に対するハウツー情報の中から，用途に応じて，詳しいページを選んだり，端的にまとまったページを選んだりすることに利用することが考えられる．

5. まとめ

SVM を用いて文節にハウツー情報の操作および対象を含むか否かを判定し，これらのペアとして手順を抽出する手法を提案し，その応用について述べた．今後は，対象の抽出精度の向上を目指し，使用する素性について再検討を行う．

謝辞

本研究の一部は，平成 26 年度科研費若手研究(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] 野中他，“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] 湯本，“Web 上のハウツー情報の構成要素の抽出”，情報処理学会研究報告，2011-DBS-153(3) (2011).

[3] V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience (1998).

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.