Web コーパスからのノウハウの獲得

小澤　俊介†　　内元　清貴‡　　松原　茂樹†

†名古屋大学　　‡情報通信研究機構

1　はじめに

Web 上には，病気の対処法や料理のレシピなど，様々なノウハウが蓄積されている．そのため，ノウハウを知るために Web を参照することも多くなっている．QA サイトにおいても，ノウハウが回答となる How-to 型の質問は約 2 割程度を占めており [1]，これらの質問に答えることは重要なタスクであると言える．従来の Web 検索でも「方法」などのキーワードを用いることで多少回答を絞ることができるものの，ノウハウのみを選定するには不十分である．この問題に対し，ノウハウを予め整理できれば，How-to 型の質問に対する回答が容易になると考えられる．また，ノウハウを蓄積・分析することにより，不測の事態に対する対策や対処法を必要に応じて事前に知らせることも期待できる．

本研究で対象とするノウハウとは物事の手順やアドバイスである．ノウハウの例を図 1 と図 2 に示す．従来研究では,「方法」や「手順」を含むテキストを対象に，手がかり表現などを用いてノウハウの獲得が行われてきた．しかし，図 2 の例のように,「方法」や「手順」が含まれていないノウハウも多数存在しており，手がかり表現を用いるだけではこれらを効率良く獲得することは困難である．では，どうすれば効率良く獲得できるだろうか．図 1 の例では「扇風機」で「風を送る」ことにより放熱を促すこと，図 2 の例では「ドライヤー」で排水口を「温める」ことが，それぞれノウハウかどうかを判断する手がかりになると考えられる．このように，ノウハウにはモノとその使われ方に関する記述が現れることが多い．

そこで本研究では，モノとその使われ方に着目することにより効率良くノウハウを獲得する手法を提案する．まず，モノを含むパッセージを獲得し，ノウハウの候補を抽出する．そして，モノと用途表現を併用することによりノウハウを獲得する．

文 ID	文
33	熱中症の対処法
34	熱中症の手当ての方法ですが、1．安静、2．冷却、3，水分補給が基本となります。
35	第一に涼しい場所に体を横にして安静にさせることです。
36	衣服を緩め体熱が放散しやすくします。
37	この時に呼びかけに応答するか、簡単な質問に答えられるかなどで意識状態を確認したり、皮膚の色調、汗のかき方などを確認したりします。
38	脈が弱い、皮膚蒼白があれば足を高くしたほうがよいでしょう。
39	第二に冷却ですが、団扇や扇風機、衣服などにより風を送り放熱を促します。
40	水をかけて冷やすことも一法です。
41	氷嚢、氷枕、アイスパックなどで太い血管が体表面近くを走っている首の両側、両脇の下、両側そけい部を冷やします。
42	冷やしたタオルで体幹、四肢をマッサージすればより効果的です。
43	本人が寒いと言うまで徹底的に冷やすことが大事です。
44	第三に水分補給ですが、意識状態が悪い場合は無理ですが、意識があり誤嚥の危険性がなければ水分を飲ませます。
45	冷たいものの方がよく、過量の糖分を含まない電解質が含まれているものを摂らせるとよいでしょう。
46	いずれにしても早めに対処することが肝要で、重症化する前に医療機関に搬送することが大事です。

図 1: ノウハウの例 1

文 ID	文
4	排水口の掃除には、ドライヤーが大活躍！
5	排水口にドライヤーで熱風をかけて温めたあと、水道水を勢いよく流すだけで、こびりついた汚れがはがれます。
6	熱湯をかけて汚れを取るよりこちらのほうが温かさが持続するため、効果が高いそうです。
7	ドロドロに汚れた排水口。
8	ドライヤーで熱風をかけて温めます。
9	勢いよく水道水を流します。
10	きれいになりました！
11	ドライヤーの温風を排水口に直接あてて、よく温めた後、一気に水を流すと、簡単に汚れが落ちました。
12	ドロドロしたヌメリがきれいにはがれて、悪臭の原因も解消できそうです。

図 2: ノウハウの例 2

2　関連研究

質問応答の How-to 型の質問に答えるために，手順が書かれたテキストの分析が行われている [2, 3]．これらの研究では，修辞関係などに着目した分析や，手順が書かれたテキストに対してタイトルや注意，助言などのタグの付与が試みられている．しかし、手順が書かれたテキストはすでに獲得されたことを前提としている。

これに対し，Web から手順が書かれたテキストを獲得する先行研究も行われている．武智らは単語 N-gram を用いることにより，HTML のリストタグが付与されたテキストを手順タイプと非手順タイプに分類する手法を提案している [4]．しかし，リストタグが付与されているテキストのみを対象としており，あらゆるテキストから手順を獲得することはできない．また，Ling らは「方法」または「手順」という単語を含むテキストを対象に構文情報，形態素情報，手がかり表現を用いることにより，テキストの手順らしさを計測する手法を提案している [5]．しかし，十分な獲得精度は得られていない．

本研究では，物事の手順またはアドバイスが書かれたテキストをノウハウとして獲得する．我々は，手がかり表現だけでなく，モノとその用途表現に着目することにより精度良くノウハウを獲得する手法を提案する．

3　ノウハウの獲得手法

Aouladomar はノウハウを 3 種類に分類している [6]．1 つ目はレシピやマニュアルなどの手順，2 つ目は規則

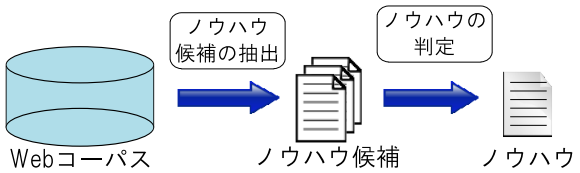

図 3: ノウハウ獲得の流れ

などの命令，3 つ目は健康法などのアドバイスである．本研究では，このうち，手順とアドバイスをノウハウとする．

従来研究の手がかり表現を利用して，ノウハウ獲得の予備実験を行ったところ，精度が約 40%，再現率が約 30% と低い結果であった．これに対し，本研究ではモノとその使われ方に着目する．モノの使われ方については，例えば,「ドライヤー」に対し「温める」といったモノの用途についての表現 (用途表現) を用いる．

ノウハウ獲得の流れを図 3 に示す．ノウハウは通常複数の文から構成されているが，テキスト中の全ての文から構成されているとは限らないため，本手法ではパッセージをノウハウの単位とする．多くのパッセージにはノウハウが含まれていないことが予想されるため，まずノウハウの候補を抽出する．Web コーパスとして河原ら [7] が収集した「Web 上の 5 億文の日本語テキスト」を利用する．そして，モノと用途表現，手がかり表現パターンを用いて，ノウハウか否かを判定する．

3.1 ノウハウ候補の抽出

まず，各モノについてモノを含むパッセージを次の方法により抽出する．以下の HTML タグを利用して，対象のモノを含む最小のタグ領域を抽出する．

body, div, table, span, p, blockquote
h1, h2, h3, h4, h5, h6

抽出したタグ領域に含まれる文数が α 以下の場合，パッセージとして抽出し，そうでない場合は，TextTiling 法 [8] によりパッセージ分割を行う．これは HTML タグが正しく利用されていない場合を考慮したためである．窓の幅を α として TextTiling 法を適用し，パッセージを抽出する．

次に，ノウハウを含みやすいと考えられるパッセージを抽出し，ノウハウの候補とする．ただし，1 種類の手法のみを利用した場合，獲得できるノウハウに偏りが生じる可能性があるため，以下の 5 種類の手法によりパッセージを絞り込み，それらの和集合を取ることによりノウハウの候補を抽出する．

(A) 文字列「方法」を含むパッセージを獲得する．

(B) リスト表現として ⟨ul⟩ または ⟨ol⟩ タグを含むパッセージを獲得する．

(C) 従来研究 [3, 5] を参考に，助言や注意，条件などを表す 47 表現を定義し，それらを含むパッセージを獲得する．

(D) 日本語評価極性辞書 (用言編)[9] 中の経験タグが付与された 638 表現を利用し，それらを含むパッセージを獲得する．

(E) モノの用途表現中の動詞の集合を利用し，それらを含むパッセージを獲得する．モノ n の用途表現とは,「n を使う」「n を楽しむ」といった表現の言い換えとして特徴づけられ，助詞と動詞の対で表現される．例えば，本の用途表現には ⟨ を，読む ⟩ などが挙げられる．用途表現の獲得方法については後述する．

3.2 ノウハウの判定

獲得したノウハウ候補がノウハウであるか否かを判定する．自動判定には，機械学習モデルを利用する．機械学習モデルで考慮する素性としては，モノと用途表現，手がかり表現パターンを利用する[1]．本節では，用途表現と手がかり表現パターンの獲得方法について述べる．

3.2.1 用途表現の獲得

鳥澤は，ある名詞 n の用途表現の性質として，1) 共起しやすい，2) 一人称代名詞が主語となりやすい，3) 助詞「で」の場合，用途表現になりやすい，という 3 つの仮説を立てた [10]．これらを反映した以下の式を利用して，用途表現を獲得する．

$$U(n) = argmax_{<v',p'>\in V\times A}\{Uscore(n,p',v')\}$$

ここで，V は用途表現になり得る動詞またはサ変名詞の集合であり，Web コーパス中で「たい」と共起する 6,485 語の動詞またはサ変名詞を利用した．ただし，用途表現であることが自明な「使う」「利用」や用途表現になりそうにない「ある」「なる」などの 20 語をあらかじめ除外した．A は助詞の集合である．

$$Uscore(n,p',v') = P(n,p',v')P(S|AP,v')Bias(p')/P(n)$$

$P(n,p',v')$ は n が $\langle p',v'\rangle$ とともに現れる確率である．$P(S|AP,v')$ は動詞 v' の主語が一人称代名詞となる条件付き確率である．S は一人称代名詞であり,「私」「私たち」などのような 17 個の語からなるものと仮定した．また，AP は主語を表現できる助詞の集合であり，$AP = \{$が，は$\}$ と仮定した．さらに，$Bias(p')$ は助詞「で」に関するバイアスを与える項であり,「で」の場合を 25，それ以外を 1 とした．

[1] 実験により，単語 n-gram は有効に働かなかったため利用していない．

表 1: 分析データ.

モノ	ノウハウを含むパッセージ数						パッセージ数
	A	B	C	D	E	計	
ドライヤー	43	44	41	16	40	147	439

3.2.2 手がかり表現パターンの獲得

手がかり表現パターンを獲得するために，3.1 節の手法を用いて Web コーパスから分析データを構築した．窓幅 α を 20，モノをドライヤーとし，A〜E の手法によりそれぞれ 100 パッセージずつ抽出した．A と B についてはランダムで，C〜E については頻度上位のパッセージをそれぞれ抽出し，ノウハウを含むか否かを人手により判定した．用途表現には，3.2.1 節の手法を用いて獲得した上位 100 の助詞と動詞の組の中から人手により選定した 24 種類の用途表現を利用した．分析データの規模を表 1 に示す．表 1 の各列の数字はそれぞれ 100 パッセージ中のノウハウを含むパッセージの数を表す．

分析データを分析することにより，手がかり表現パターンを獲得した．ただし，パターンは文あるいは文内の文節列に対して適用することを想定し，文をまたぐパターンは作成しない．作成したパターンの一部を表 2 に示す．パターン中の | は OR，+ は単語境界，.* は任意の単語列を表す．また，3 列目の対象はパターンの適用対象を表し，S は文，B は文頭文節，E1 は文末文節，E2 は文末の 2 文節，E3 は文末の 3 文節を表す．

分析の結果，ノウハウを含むパッセージには，1〜3 のような順序を表す表現，あるいは，4 や 5 のような助言を表す表現，6〜8 のような注意を表す表現，9 や 10 のような条件を表す表現などが含まれていた．また，15 はノウハウのタイトルを表す表現であり，[方法] には対策や仕方などの 23 表現，[説明] には紹介や教えるなどの 21 表現が含まれる．これらのパターンを含む 72 種類の手がかり表現パターンを作成した．

表 2: 手がかり表現パターン.

No.	パターン	対象
1	まず | 先ず | 初めに | 初め+は | 最初+は	B
2	それ+から | 次に | この後 | その後 | 次は | そして	B
3	出来上がり | 仕上がり | 完成 | 終了 | 完了	E1
4	動詞+方+が.*良い | 動詞+の+が.*安心だ	S
5	動詞+やすい	E2
6	禁物 | だめだ | 厳禁 | 危険	S
7	動詞+ない+ようだ.*動詞 | 動詞+ぬ+に.*動詞	S
8	気+を+付ける | 手+を+抜く+ない | 注意	E3
9	を+対象 | に+限定 | に+限る | のみ+と+限定	E3
10	必要だ | 用意 | 欠かせる+ない | 必須だ	E3
11	役立つ | 活躍 | 便利だ | 効果 | 効率	E2
12	丁寧だ | 慎重だ | 均一 | しっかり | 念入りだ	S
13	試す ください | 試す みる	E2
14	？| か 。	E1
15	[方法]+を+[説明]	S

表 3: 評価データ.

モノ	ノウハウを含むパッセージ数						パッセージ数
	A	B	C	D	E	計	
アイロン	28	38	23	8	54	132	457
エアコン	5	3	6	1	12	26	485
オーブン	27	79	24	6	47	175	462
携帯電話	6	3	5	0	1	15	489
洗濯機	5	15	3	0	23	43	478
扇風機	14	18	21	5	29	71	457
掃除機	13	28	20	5	37	90	470
デジタルカメラ	3	5	11	1	18	35	481
電子レンジ	19	59	16	9	63	152	473
冷蔵庫	20	44	10	0	38	113	489
合計	140	295	139	35	322	852	4741

4 実験

4.1 設定

本評価実験で着目するモノとしては，Wikipedia の電気製品の一覧に掲載されており，かつ，Web コーパス中で頻度が 10,000 回以上出現するモノの中から 10 種類の電気製品を選択し利用した．これらのモノに対し，3.1 節の手法を用いてノウハウ候補を抽出し，ノウハウか否かを人手で判定することにより，分析データと同様に評価データを構築した．用途表現には人手選定による平均 25 種類の用途表現を利用した．本研究では，ノウハウを含むパッセージを正解とし，ノウハウが一部しか含まれていない場合，不正解とした．評価データの規模を表 3 に示す．

実験では，以下の 3 種類の素性を用いる．手がかり表現パターンのみを利用した手法と手がかり表現パターンにモノと用途表現を加えた手法を比較することにより，モノと用途表現を併用する方法の有効性を検証する．機械学習モデルとしては，Support Vector Machines (SVM) を用い，SVM の学習には，TinySVM[2]を利用した．

パターン 各パターンがマッチした頻度と全パターンがマッチした総頻度，マッチしたパターンの種類数を利用する．

用途表現 (自動) モノと用途表現の 3 つ組の頻度を利用する．用途表現には，3.2.1 節の手法により計測した $Uscore$ が上位 25 の助詞と動詞の組を利用した．

用途表現 (人手) モノと用途表現の 3 つ組の頻度を利用する．用途表現には，データセットの構築時に利用した人手選定による用途表現を利用した．

4.2 評価実験

各モノのデータをそれぞれ 5 分割し，10 種類のモノのデータを用いて 5 分割交差検定による評価実験を行った．実験結果を表 4 に示す．パターンのみを用いた手法に対して，モノと用途表現を併用することにより，精度，再現率ともに向上していることが分かる．このこと

[2]http://vhsdrn.org/ taku/software/TinySVM/

表 4: 5 分割交差検定による実験結果.

	精度	再現率	F 値
パターン	73.4% (201/274)	23.6% (201/852)	35.7
パターン+ 用途表現 (自動)	73.4% (307/418)	36.0% (307/852)	48.4
パターン+ 用途表現 (人手)	73.7% (329/446)	38.6% (329/852)	50.7

表 5: 10 分割交差検定による実験結果.

	精度	再現率	F 値
パターン	67.6% (152/225)	17.8% (152/852)	28.2
パターン+ 用途表現 (自動)	72.3% (274/379)	32.2% (274/852)	44.5
パターン+ 用途表現 (人手)	71.2% (282/396)	33.1% (282/852)	45.2

から，モノと用途表現に着目することにより，効率よくノウハウの獲得ができると言える．また，人手選定した用途表現に比べ，自動獲得した用途表現ではやや性能が落ちるものの，ノウハウ獲得において十分な効果が得られることが分かった．

上記の実験では，学習データとテストデータの両方に 10 種類のモノに関するデータが含まれている．しかし，学習データに含まれていないモノに対しても同様の効果が得られるとは限らないため，9 種類のモノに関するデータを用いて学習し，残りの 1 種類のデータでテストを行う 10 分割交差検定を行った．実験結果を表 5 に示す．上記の実験と同様に，用途表現を利用することにより，精度・再現率ともに向上していることが分かる．また，上記の実験よりも大きな効果が出ている．このことから，学習に利用していないモノに対してもモノと用途表現が有効に働くことが示された．

4.3 考察

評価データ中には，ノウハウを一部しか含まないパッセージが 108 パッセージ存在した．これはパッセージの獲得誤りが原因であるが，パッセージの獲得が正しくできれば，これらは正解となるパッセージである．そこで，これらを正解と見なした場合の評価実験を 4.2 節と同様に行った．その結果，いずれの手法においても再現率で 5% 以上の向上が見られた．このことから，パッセージの獲得手法について今後検討が必要である．

評価データ中で正解数の多い電子レンジやオーブンでは，いずれの手法においても全体の性能より F 値が 5% 以上良い結果がであったのに対し，正解数の少ない携帯電話やエアコンでは全体の性能より F 値が 5% 以上低い結果であった．また，デジタルカメラに至っては，パターンのみの手法ではノウハウを獲得できなかった．しかし，用途表現を用いた手法では，少ないながらもノウハウの獲得ができていた．このことはモノと用途表現を利用することで様々なモノに対してノウハウの獲得ができることを示唆している．

5 まとめ

本研究では，モノとその使われ方に着目することにより，効率よくノウハウを獲得する手法を提案した．まず，ノウハウの候補を獲得する．そして，手がかり表現パターンとモノ，用途表現を利用することにより，ノウハウを含むパッセージを獲得する．

本実験では，評価データ上での有効性を示すことができたが，評価データとは異なる性質のデータや異なるドメインに対して有効であるかは検証できていない．そのため，今後の課題としてオープンドメインでの評価実験が挙げられる．また，パッセージの獲得手法の洗練化も必要となるだろう．

参考文献

[1] 田村晃裕, 高村大也, 奥村学: 複数文質問のタイプ同定, 言語処理学会第 11 回年次大会 (2005).

[2] Delpech, E. and Saint-Dizier, P.: Investigating the Structure of Procedural Texts for Answering How-to Questions, In Proceedings of the 6th International Conference on Language Resources and Evaluation, pp.544-550 (2008).

[3] Fontan, L. and Saint-Dizier, P.: Analyzing the Explanation Structure of Procedural Texts: Dealing with Advices and Warnings, In Proceedings of the 2008 Conference on Semantics in Text Processing, pp. 84-93 (2008).

[4] 武智峰樹, 徳永建伸, 松本裕治, 田中穂積: WWW ページからの手順に関する箇条書きの抽出, 情報処理学会論文誌, Vol.44, No.12, pp. 51-63 (2003).

[5] Yin, L. and Power, R.: Adapting the Naive Bayes Classifier to Rank Procedural Texts, In Proceedings of the 28th European Conference on Information Retrieval Research, pp. 179-190 (2006).

[6] Aouladomar, F.: Towards Answering Procedural Questions, In Proceedings of the IJCAI Workshop on Knowledge and Reasoning for Answering Questions, pp. 21-31 (2005).

[7] Kawahara, D. and Kurohashi, S.: A Fully-lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, In Proceedings of HLT-NAACL 2006, pp.176-183 (2006).

[8] Hearst, M. A.: TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, Vol. 23, pp. 33-64 (1997).

[9] 小林のぞみ, 乾健太郎, 松本祐治, 立石健二, 福島俊一: 意見抽出のための評価表現の収集, 自然言語処理, Vol. 12, pp.2003-222 (2005).

[10] 鳥澤健太郎: 対象の用途と準備を表す表現の自動獲得, 自然言語処理, Vol. 13, No. 2, pp. 125-144 (2006).