情報処理学会研究報告
IPSJ SIG Technical Report

モノの用途表現を手がかりとしたWebからのノウハウの獲得

小澤俊介[†1]　内元清貴[†2]　松原茂樹[†1]

Web 上には，病気へ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．しかし，従来の Web 検索ではノウハウのみを検索することは困難である．この問題に対し，ノウハウを整理し，提供することができれば，災害に対する予防策など，様々な事象への対処・対策が容易になる．そこで本論文では，Web から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．モノと用途表現に着目することにより，モノを含むパッセージから精度よくノウハウを含むパッセージを獲得する．

Acquisition of Know-How Information from Web

Shunsuke Kozawa ,[†1] Kiyotaka Uchimoto [†2]
and Shigeki Matsubara[†1]

A variety of know-how such as recipes and solutions for troubles have been stored on the Web. However, it is not so easy to appropriately find certain know-how information since it is difficult to discriminate the know-how information from non-know-how information. If know-how could be appropriately detected, it would be much easier for us to know how to tackle unforeseen situations such as disasters. This paper proposes a method for acquiring know-how information from the web. First, we extract candidates for know-how. Then, passages containing the know-how are acquired by focusing on each object and its typical usage.

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 情報通信研究機構
National Institute of Information and Communications Technology

1. はじめに

Web 上には，病気へ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，QA サイトにおいても，ノウハウが回答となる How-to 型の質問が約 2 割を占めている[10]．しかし，How-to 型の質問に対する回答を Web から見つけることは容易ではない．従来の Web 検索において，質問に関連するキーワードと「方法」などのキーワードを併用することにより回答を見つけることも可能だが，多くの場合，ノウハウが記されていないテキストも検索されることになり，回答を見つけるのに多くの時間を要する．この問題に対し，ノウハウを予め収集し，整理できれば，How-to 型の質問に対する回答の発見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 に示す．従来研究では,「方法」や「手順」という文字列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[12]．しかし，図 1 の 2 つ目の例のように,「方法」や「手順」が含まれていないノウハウも多数存在している．また，ノウハウの獲得に必要な様々な手がかり表現を用意することは容易ではないため，手がかり表現に着目するだけではこれらをもれなく獲得することは困難である．

本研究では，この問題に対し，モノとその使われ方に着目する．というのも，ノウハウには少なくとも 1 つはモノが含まれていることが多く，それらの利用がノウハウにおいて重要な役割を果たしているからである．例えば，図 1 の 1 つ目の例では「扇風機」で「風を送る」ことにより放熱を促すこと，2 つ目の例では「ドライヤー」で排水口を「温める」ことが重要な役割を果たしている．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ため，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．

そこで本論文では，モノとその使われ方に着目することにより精度よ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．

本論文の構成は以下の通りである．まず 2 章で，手順が書かれたテキストに着目した先行研究について述べる．3 章では，モノとその使われ方に着目してノウハウを獲得する手法について述べ，4 章でノウハウの獲得実験の結果とその考察を示す．最後に，5 章で本論文のまとめと今後の課題について述べる．

ID	文
1	熱中症の対処法
2	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。
3	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。
4	衣服を暖め体熱が放散しやすくします。
5	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。
6	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。
7	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。
8	水をかけて冷やすことも一法です。
9	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。
10	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。
11	本人が寒いと言うまで徹底的に冷やすことが大事です。
12	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。
13	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。
14	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。

ID	文
1	排水口の掃除には、ドライヤーが大活躍！
2	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。
3	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。
4	ドロドロに汚れた排水口。
5	ドライヤーで熱風をかけて温めます。
6	勢いよく水道水を流します。
7	きれいになりました！
8	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。
9	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 1 ノウハウの例

2. 関 連 研 究

How-to 型の質問に答えるシステムを開発するために，手順が書かれたテキストの分析が行われている[1)–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．また，Quarteroni らは How-to 型の質問に答えるためのモデルを提案している[7)]．しかし、これらの手法は手順が書かれたテキストがすでに獲得されていることを前提としている．

これに対し，手順が書かれたテキストを自動獲得する先行研究も行われている．Schwitter らは手順に関する質問に対する回答を技術文書から抽出する手法を提案した[8)]．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[9)]．しかし，これらの手法は適用できるテキストが限定されており，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[12)]．しかし，少数の手がかり表現のみでは手順が書かれているかどうかを判定することは困難であるため，十分な獲得精度は得られていない．

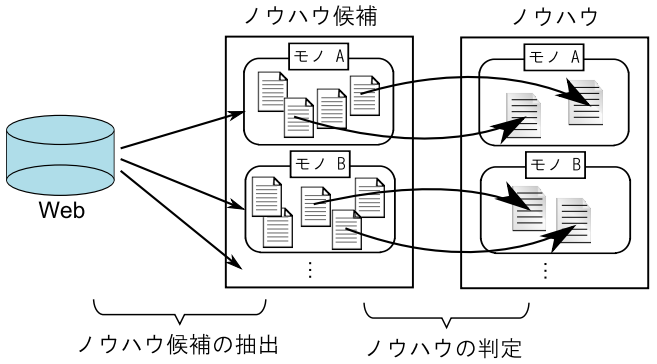

図 2 ノウハウ獲得の流れ

一方，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，モノとその使われ方に着目することにより精度よくノウハウを獲得する手法を提案する．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．

3. ノウハウの獲得手法

Aouladomar は手順に関するテキストを 3 種類に分類している[1)]．1 つ目はレシピやマニュアルなどの「手順」であり，2 つ目は規則などの「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，QA サイトでよく質問される「手順」と「アドバイス」をノウハウとする．

ノウハウ獲得の流れを図 2 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でもって一つのノウハウを構成す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．まず，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，対象のモノに対して Web コーパスからノウハウの候補を抽出する．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．Web コーパスとして「Web 上の 5 億文の日本語テキスト」[5)] を利用する．Web コーパスには形態素情報と構文情報が自動付与されている．

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する．そうでない場合は，窓の幅を α として TextTiling 法[4] を適用することによりパッセージに分割する．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．

次に，ノウハウを含む可能性が高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それらをノウハウの候補とする．

(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),12)] を参考に，助言や注意，条件などを表す 47 表現 (「ないように」,「方が良い」,「場合」など) を定義し，それらのいずれか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)[6] 中の経験タグが付与された 638 表現 (「よくなる」,「治る」など) を利用し，それらを含むパッセージを獲得する．これは，ノウハウには経験を表す表現が含まれやすいと想定したためである．
(E) モノの用途表現中の動詞の集合を利用し，それらのいずれか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，モノと用途表現及び，手がかり表現パターンを利用する[⋆1]．本節では，用途表現の獲得方法と手がかり表現パターンの作成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[11)]．本研究では，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．

3.2.2 手がかり表現パターンの作成

ノウハウを含むパッセージを分析することにより，手がかり表現パターンを作成する．分析のために，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．パッセージとして抽出するときの文数の閾値 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 1〜5 列目の数字はそれぞれ 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．

表 1 分析データ．

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

分析データをもとに，手がかり表現パターンを作成した．分析の結果，ノウハウを含むパッセージには，順序や助言，注意，条件などを表す表現が含まれていた．そこで，これらの表現をパターン化し，72 種類の手がかり表現パターンを作成した．また，ノウハウを含まないパッセージに頻出した 7 種類の手がかり表現パターンも作成した．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

⋆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 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

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．パターン 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を作成した．また，パターン 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．

4. 評　　価

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[⋆1]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

⋆1 http://ja.wikipedia.org/wiki/電気製品の一覧

表 3 評価データ．

	ノウハウを含むパッセージ数						
モノ	A	B	C	D	E	計	パッセージ数
アイロン	34	43	31	9	57	154	458
エアコン	14	16	33	5	40	99	486
オーブン	28	79	32	7	51	188	463
携帯電話	11	5	11	0	6	31	489
洗濯機	13	26	14	3	30	78	479
扇風機	17	29	37	5	39	105	458
掃除機	14	34	31	5	36	106	471
デジタルカメラ	8	14	19	5	21	62	482
電子レンジ	24	61	27	13	69	178	474
冷蔵庫	28	56	22	1	62	165	490
合計	201	363	258	53	407	1165	4750

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した．

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．

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の 3 つ組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．

4.2 実 験 結 果

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

⋆2 http://www.chasen.org/taku/software/TinySVM/

表 4 5 分割交差検定による実験結果.

素性	精度	再現率	F 値
パターン (ベースライン)	72.50% (543/749)	46.61% (543/1165)	56.74
パターン + 用途表現 (自動)	74.52% (579/777)	49.70% (579/1165)	59.63
パターン + 用途表現 (人手)	75.25% (593/788)	50.90% (593/1165)	60.73

表 5 10 分割交差検定による実験結果.

素性	精度	再現率	F 値
パターン (ベースライン)	71.99% (514/714)	44.12% (514/1165)	54.71
パターン + 用途表現 (自動)	74.24% (562/757)	48.24% (562/1165)	58.48
パターン + 用途表現 (人手)	74.29% (572/770)	49.10% (572/1165)	59.12

ことから，モノと用途表現に着目することにより，精度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，モノと用途表現の 3 つ組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．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

4.3.1 モノと用途表現の効果

モノと用途表現の 3 つ組による効果を調べるため，パッセージに 3 つ組が含まれているか否かに着目して表 5 の結果を調査した．評価データに含まれる 4,750 パッセージのうち，1,289 パッセージに 3 つ組が含まれていた．3 つ組を含むパッセージに着目した場合の結果を表 6 に，3 つ組を含まないパッセージに着目した場合の結果を表 7 に，それぞれ示す．3 つ組を含まないパッセージに対する F 値がやや下がっているものの，3 つ組を含むパッセージに対する F 値が大きく向上していることがわかる．このことはノウハウの獲得においてモノと用途表現が有用であることを示している．

表 6 10 分割交差検定による実験結果 (3 つ組を含むパッセージのみ)

素性	精度	再現率	F 値
パターン (ベースライン)	82.61% (266/322)	47.50% (266/560)	60.32
パターン + 用途表現 (人手)	79.12% (360/455)	64.29% (360/560)	70.94

表 7 10 分割交差検定による実験結果 (3 つ組を含むパッセージを除く)

素性	精度	再現率	F 値
パターン (ベースライン)	63.27% (248/392)	40.99% (248/605)	49.75
パターン + 用途表現 (人手)	67.30% (212/315)	35.04% (212/605)	46.09

表 8 モノごとの実験結果 (10 分割交差検定).

	パターン			**パターン+用途表現 (自動)**			**パターン+用途表現 (人手)**			3 つ組を
モノ	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値	含む割合
アイロン	78.75	40.91	53.85	81.05	50.00	61.85	78.50	54.55	64.37	54.55
エアコン	68.97	40.40	50.96	72.92	35.35	47.62	72.34	34.34	46.58	13.13
オーブン	78.05	51.06	61.74	79.85	56.91	66.46	78.03	54.79	64.38	63.30
携帯電話	35.48	35.48	35.48	34.62	29.03	31.58	37.93	35.48	36.67	12.90
洗濯機	65.00	50.00	56.52	69.49	52.56	59.85	72.73	51.28	60.15	30.77
扇風機	66.67	36.19	46.91	68.00	32.38	43.87	67.31	33.33	44.59	20.95
掃除機	61.04	44.34	51.37	68.42	61.32	64.68	69.47	62.26	65.67	61.21
デジタルカメラ	45.45	16.13	23.81	48.00	19.35	27.59	52.17	19.35	28.24	22.58
電子レンジ	80.37	48.59	60.56	81.58	52.54	63.92	80.70	51.98	63.23	58.76
冷蔵庫	84.85	50.91	63.64	80.18	53.94	64.49	81.90	57.58	67.62	66.04
合計	71.99	44.12	54.71	74.24	48.24	58.48	74.29	49.10	59.12	47.64

表 8 に 10 分割交差検定におけるモノごとの実験結果を示す．ほとんどのモノについては，3 つ組を利用することによって F 値が向上していることが分かる．ただし，エアコンと扇風機に関しては F 値が低下した．そこで，ノウハウを含むパッセージが 3 つ組を含むか否かを調べた．表 8 の最右の列はノウハウを含むパッセージのうち，3 つ組を含むパッセージの割合を示している．各モデルの性能と 3 つ組を含むパッセージの割合を比較したところ，3 つ組を用いたモデルの性能と 3 つ組を含む割合に関連があることが分かった．つまり，3 つ組を含む割合の小さいモノについては性能がやや低下もしくは向上するに留まったものの，3 つ組を含む割合の大きいモノについては大きく性能が向上していた．この結果はノウハウの獲得においてモノと用途表現の 3 つ組が大きな影響を果たしていることを示している．

4.3.2 誤り分析

ノウハウの獲得誤りの原因を調べるため，パターンおよび人手で選定した用途表現を利用したモデルでの 10 分割交差検定の結果を分析した．

誤って獲得した 198 パッセージについて分析したところ，以下に示す，主に 4 つの原因があることが分かった．

- ノウハウを一部しか含まないパッセージ
  35.4% にあたる 70 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．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 190 パッセージ存在した．そこで，これらを正解とみ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 F 値で約 8% 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．
- 手順に関するテキストの一種である規則やガイドラインを含むパッセージ
  13.6% にあたる 27 パッセージ存在した．本研究では対象外としたノウハウであるが，これらを正解とみなすことも考えられるため，獲得したとしても大きな問題が生じるわけではないと考える．
- 日記が書かれたパッセージ
  10.6% にあたる 21 パッセージ存在した．これらにはモノの利用に関する記述は存在するものの，ノウハウは記述されていなかった．これらに関しては，パターンや 3 つ組の共起や順序などを考慮することで対処できると考えている．
- 商用サイトなどのモノの機能に関する説明を含むパッセージ
  7.1% にあたる 14 パッセージ存在した．これらについては，商用サイトを特定できるパターンの追加が必要となるだろう．

獲得できなかった 593 パッセージについて分析した．593 パッセージのうち，3 つ組を含むパッセージは 33.7% の 200 パッセージに過ぎなかったものの，用途表現を含むパッセージは 74.9% の 444 パッセージ存在した．そこで，評価データ中のパッセージが用途表現を含むか否かについて分析した．その結果，全パッセージのうち，用途表現を含むパッセージの割合は 61.0% であった．これに対し，ノウハウを含むパッセージのうち，用途表現を含むパッセージの割合は 82.3% だった．このことから，用途表現の頻度や他の素性との組み合わせを考慮することにより，再現率の向上が期待できる．また，593 パッセージのうち，88.4% の 524 パッセージでは，対象のモノとは異なるモノが利用されていた．このことは，本実験で獲得できなかったノウハウが他のモノを対象としたときに獲得できる可能性があることを示している．さらに，複数のモノを考慮することにより，再現率が向上することが期待できる．

表 9 オープンドメインでの実験結果.

モノ	パッセージ数	パターン	パターン+用途表現 (自動)	パターン+用途表現 (人手)
カーテン	486	55.56% (30/54)	54.55% (30/55)	54.55% (30/55)
香水	439	68.29% (28/41)	76.92% (30/39)	76.32% (29/38)
鏡	490	36.84% (14/38)	42.42% (14/33)	40.00% (14/35)
カレンダー	481	39.29% (11/28)	48.00% (12/25)	50.00% (11/22)
ライター	473	58.14% (25/43)	65.63% (21/32)	65.71% (23/35)

4.4 オープンドメインでの実験

4.2 節の実験により，モノと用途表現を利用する有効性を評価データ上で示すことができた．しかし，評価データとは異なるドメインに対しても有効であるとは限らない．そこで，オープンドメインにおける実験を行った．対象とするモノには電気製品とは異なる 5 つのモノを選択した．評価データの構築と同様に，3.1 節の手法を用いてノウハウ候補のパッセージを抽出し，評価データを用いて学習したモデルにより，それらがノウハウであるか否かを判定した．

ノウハウの獲得精度を表 9 に示す．4 つのモノに対して精度が向上していることが分かる．しかし，カーテンに対しては，精度が下がっている．この理由として，カーテンがノウハウにおいて重要な役割を果しているケースが少ないことが挙げられる．例えば,「カーテンを閉める」というカーテンの利用に関する記述は獲得されたノウハウにおいて重要な役割を果たしていることが少なかった．しかし,「カーテンを閉める」と「直射日光を避ける」が共起している場合には,「カーテンを閉める」が重要な役割を果たしていた．このことから，モノの使われ方だけでなく，利用目的にも着目すれば，効果的にノウハウの獲得ができると考えられる．実験結果より，一部のモノに関しては利用目的などを考慮する必要があるものの，異なるドメインのモノに関しても，モノとその使われ方に着目することにより，ノウハウを効果的に獲得できる可能性を示した．

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，精度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，対象としたモノが限定されているものの，対象のモノを広げることによって様々なノウハウを獲得できる可能性を示した．

今後の課題としては，用途表現中の動詞の目的語を考慮することが考えられる．例えば，図 1 の 1 つ目の例では,「扇風機で送る」よりも「扇風機で風を送る」の方がより重要な役割を果たしており，ノウハウの判定において，より有用な素性になることが考えられる．また，本研究では，1 つのモノとその使われ方に着目したが，ノウハウには複数のモノが利用されていることも多いと考えられる．そのため，複数のモノとそれらの使われ方に着目するにより，より精度よくノウハウを獲得できるだろう．

謝　　辞

本研究は一部、財団法人堀情報科学振興財団研究助成により実施した．

参 考 文 献

1) Aouladomar, F.: Towards answering procedural questions, *Proceedings of the IJCAI Workshop on Knowledge and Reasoning for Answering Questions*, pp.21–31 (2005).
2) Delpech, E. and Saint-Dizier, P.: Investigating the structure of procedural texts for answering how-to questions, *Proceedings of the 6th International Conference on Language Resources and Evaluation* (2008).
3) Fontan, L. and Saint-Dizier, P.: Analyzing the explanation structure of procedural texts: dealing with advice and warnings, *Proceedings of the 2008 Conference on Semantics in Text Processing*, pp.115–127 (2008).
4) Hearst, M.A.: TextTiling: Segmenting text into multi-paragraph subtopic passages, *Computational linguistics*, Vol.23, No.1, pp.33–64 (1997).
5) Kawahara, D. and Kurohashi, S.: A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis, *Proceedings of the 7th Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp.176–183 (2006).
6) Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. and Fukushima, T.: Collecting evaluative expressions for opinion extraction, *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pp.584–589 (2004).
7) Quarteroni, S. and Saint-Dizier, P.: Addressing How-to Questions using a Spoken Dialogue System: a Viable Approach?, *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, pp.19–23 (2009).
8) Schwitter, R., Rinaldi, F. and Clematide, S.: The Importance of How-Questions in Technical Domains, *Proceedings of the Question-Answering Workshop of TALN2004*, pp.451–460 (2004).
9) Takechi, M., Tokunaga, T., Matsumoto, Y. and Tanaka, H.: Feature selection in categorizing procedural expressions, *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pp.49–56 (2003).
10) Tamura, A., Takamura, H. and Okumura, M.: Classification of multiple-sentence questions, *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pp.426–437 (2005).
11) Torisawa, K.: Automatic acquisition of expressions representing preparation and utilization of an object, *Proceedings of the 5th Recent Advances in Natural Language Processing*, pp.556–560 (2005).
12) Yin, L. and Power, R.: Adapting the Naive Bayes classifier to rank procedural texts, *Lecture Notes in Computer Science*, Vol.3936, pp.179–190 (2006).