

物事の用途とその準備を表す表現の自動獲得

鳥澤健太郎

北陸先端科学技術大学院大学情報科学研究科

torisawa@jaist.ac.jp

1 はじめに

本稿では、プラン認識 [1] などの知的エージェントによる推論を最終的な目標として、人工物の用途とその用途を達成するための準備を表す自然言語表現を獲得する手法について述べる。プラン認識の研究は 90 年代初めに盛んに行われたが、その後、下火になったといわざるを得ない。一つの原因は、大量のプランないしはその構成要素の定義を書き下ろす事が困難であったことであろう。本研究では、Generative Lexicon Theory [8] に見られる telic role などの対象固有のゴールや、agentive role に見られるような対象の起源の概念を、再定義、拡張し、コーパスから自動的に抽出する事を狙う。本研究で自動獲得を行うのは、自然言語表現のペアであり、これを準備/用途対と呼ぶ。この準備/用途対は名詞あるいは名詞句に対して定義されるものであり、例えば、「本」という名詞に対しては、「本を書く」-「本を読む」というペアを準備/用途対と見なす事ができる。本稿では、ペアの最初の要素（「本を書く」）を「準備表現」と呼び、2番目の要素（「本を読む」）を「用途表現」と呼ぶ。これらの表現は明らかに Generative Lexicon Theory における agentive role と telic role に非常に類似した概念であるが、以下の2点において相違する。すなわち、I) 用途表現は、telic role のように「目的」とか「機能」といった漠然とした概念を用いず、「言い換え」として特徴づけられる、また II) 準備表現は agentive role よりも広い範囲の表現を含む。まず、対象 X の用途表現は、「X を使う」「X を利用する」「X を活用する」「X を楽しむ」といった表現の「主要」かつ「直接的な」言い換えとして特徴づけられる。ここで「主要」というのは、例えば、「X を利用する」の通常の方法を反映した言い換えでなければならないということである。例えば、「本で殴る」「本をプレゼントする」というのは、「本を利用する」の言い換えであるということとは可能であるかもしれないが、「本を利用する」の「通常」の方法であるとは言えない。したがって、「本で殴る」などの表現は本の用途表現と見なすことはできない。また、「直接的な」言い換えというの、「X を使う」などの出来事と同一の出来事を指す表現を指す。例えば、「本を買った」場合、通常その後「本を利用する」出来事が続くため、「本を買う」は「本を利用する」の言い替えであるとみなす場合があるかも知れない。しかしながら、本稿では「本を買う」と「本を利用する」は全く別個の出来事であるとみなし、よって、「本を買う」は「本を利用する」の直接的な言い替えであるとはみなさない。従って、「本を買う」は本の用途表現ではない。

また、一方で agentive role と準備表現との差であるが、agentive role が「対象が存在する起源、あるいは存在するようになるという出来事」を参照する表現として特徴づけられるのに対して、準備表現は対応する用途表現を実現するための準備の一部を表す表現と特徴づけられる。

例えば、「本を買う」とか「本を持ってくる」といった表現を agentive role と見なす事は難しいが、「本を読む」という用途表現のための準備と見なせるので、準備表現と見なす事ができる。一方で、通常の agentive role つまり、「本を書く」「本を出版する」といった表現も準備表現とみなすことができ、つまるところ、agentive role は準備表現の部分集合であると考えられる。（ここで重要なことは「準備表現」と「用途表現」は同一の行為者によって実行される行為である必要はないということである。つまり、本稿では「本を書く」というのは、行為者が一致しないものの「（誰かが）本を読む」ことの準備であるとみなす。）

類似した意味的關係、あるいは表現の自動学習に関する研究はいくつか存在する。乾らは、ある行為 Act1 とその行為を行うための準備 Act2 の対を means relation とよび、接続詞「ため」に注目して、それらを自動獲得する手法を提案した [5]。我々の準備/用途対は実際にはこの means relation の特殊なケースとみなすことができる。差は「目的」である Act1 を対象の用途表現に限定した事である。この制限は推論の範囲を狭める一方で、推論結果の妥当性については望ましい帰結を持つ。例えば、「本を買う」という表現と「本をプレゼントする」という表現の間の関係は、means relation の一種であると見なす事ができるが、本を含むような状況であって、実際に「本をプレゼントする」という出来事が発生するのはごく一部の状況であろう。一方で、本の用途表現である「本を読む」という出来事は、実際に本に関係する状況の大部分で生じているとみなすことができる。つまり、「本を読む」を含むような means relation すなわち、用途/準備対は本に関係する状況の大部分において、より妥当な推論結果をもたらすものと期待できる。また、乾と我々の研究とのもう一つの差は、乾の獲得手法においては一般化が行われないということである。これによって、獲得できる means relation の範囲は非常に限定されたものとなる。一方で我々の獲得手法は、自然な形で名詞に関する一般化を行っており、より多様な関係を獲得できる。これ以外の類似した関係、表現の自動獲得に関する研究としては、出来事間の時間的前後関係に関するもの、telic あるいは agentive role の学習に関するものなどが存在する [4, 2, 9, 7, 11]。

2 獲得アルゴリズム

本稿で提案する準備/用途対の獲得アルゴリズムは以下の二つのステップからなる。

Step 1 語の共起関係を用いた用途表現の獲得

Step 2 Step 1 で獲得した用途表現を前提とした準備/用途対の獲得

ここで用語の再定義が必要になるが、ある対象 X の用途表現は、動詞の項位置 (argument position) で、 X がその項位置に来た場合に X の用途を表せる表現になるものと定義する。例えば、対象「本」の用途表現は、動詞「読む」の項位置の内、助詞「を」でマークされるものであって、その用途表現は、助詞と動詞の組、〈を、読む〉であると見なすことになる。準備表現にたいしても同様の定式化を行い、助詞と動詞の対によって表せるものとする。

以下では提案手法の各ステップについて詳しく述べる。

2.1 Step 1: 用途表現の獲得

対象 X の用途表現である動詞の項位置は、以下の仮説に基づいて獲得される。

- X が用途表現である項位置にくる頻度は大きい。
- 項位置を指定する助詞が、道具、場所などを指定する「で」である場合、その項位置は用途表現になりやすい。
- 「私」などの一人称代名詞が主語となりやすい動詞は、用途表現を構成しやすい。

最初の仮説は、 X への言及が行われる場合には、 X の用途が言及されることが多いという観察に基づいている。また、助詞の「で」に注目する理由は、「で」が出来事で使われる道具や、場所を指定することが多く、また、 X が出来事で使われる道具、場所を指定する場合、その表現は用途を表しやすいという理由による。また、一人称代名詞に注目する理由は、それによって「普通」の出来事が用途として抽出されやすくなるからである。例えば、新聞などにおいては、交通事故、犯罪などの「普通でない」出来事がよく現れる。その結果として、「車」と「逃走する」などの「普通でない」動詞が高い頻度で共起することになり、これが「普通の」用途表現を獲得する上で障害となった。一方で、「私」などの一人称代名詞が参照するのは、通常、「普通の人々」であって、それらと高い頻度で共起する動詞は「普通の出来事」を参照している可能性が高いため、そのような共起頻度を考慮する事によって、普通の出来事を表す動詞を用途として獲得する事ができる。実際の獲得は、以上の仮説を反映した以下のスコアによって行われる。

$$U_{score}(X, p, v) = P(X, p, v)P(S|v)Bias(p)/P(X)$$

我々の獲得アルゴリズムは、名詞 X に対してこのスコアでトップとなった、項位置 $\langle p, v \rangle$ を用途表現として出力する。 $(p$ は助詞、 v は動詞である。) ここで $P(X, p, v)$ は X が項位置 $\langle p, v \rangle$ に現れる確率、また $P(S|v)$ はあらかじめ指定した一人称代名詞の集合 S の要素のいずれかが動詞 v の主語になる確率、 $P(X)$ は X の出現確率である。さらに $Bias(p)$ というのは、助詞「で」に関するバイアスを与えるための項であり、以下のように定義される。

$$Bias(p) = \begin{cases} 25 & p = \text{de} \\ 1 & \text{otherwise} \end{cases}$$

2.2 Step 2: 準備 / 用途対の獲得

Step 2 では、Step 1 で獲得された用途表現を含むような、用途 / 準備対の獲得を行う。これは次の仮説に従って行われる。

- ある名詞 X に対する用途 / 準備対に現れる項位置は X で占められる頻度が大きい。
- ある名詞 X' に対する用途表現、ないしは準備表現は、他の名詞 X に対する用途表現、準備表現にもなりやすい。
- ある名詞 X に対する用途 / 準備対は特定の言語的パターンに現れやすい。

今回提案する獲得手法は、まず第一の仮説に従って、準備表現を表す項位置を X と高頻度で共起しているものに限定する。さらに、第 3 の仮説に従って、Step 1 で獲得された用途表現候補と次のようなパターンで共起しているものを準備表現の候補とする。(実際にカウントされている共起は、項位置間の共起ではなく、動詞間の共起である。また、これらのパターンはすでに先行研究で動詞間の意味的關係を得るために使われていることに注意。)

並列動詞句 (本を買い) 準備表現候補、(読んだ) 用途表現候補

関係代名詞句 (買った) 準備表現候補 本を (読んだ) 用途表現候補

接続詞「ため」(本を読む) 準備表現候補 ため (買った) 用途表現候補

具体的な獲得アルゴリズムは以下ようになる。アルゴリズムへの入力の名詞の集合であり、その要素を X で表すことにする。

1. まず、入力の名詞集合中の各々の要素 X に対して、項位置の候補を以下のように求める。まず、 X と共起している項位置をその共起頻度によってソートする。ついで、Step 1 で求めた用途表現と上記 3 パターン中で共起している項位置のうち、各パターンごとに共起頻度で上位 5 個の項位置を求める。さらにパターン中での共起に関して何ら制限をもうける事なく上位 5 個を取り出す。最後にこれらの項位置のすべての和集合をとり、これを準備表現の候補とする。名詞 X に対して以上の方法で計算された項位置の集合を $C(X)$ で表すこととする。
2. 後述する方法にしたがい、 $C(X)$ 中の準備 / 用途表現の候補にたいして、SVM[10] のための素性ベクトルを生成し、前もってトレーニングした SVM の決定関数の値を求め、それを候補に対するスコアとする。
3. 決定関数の値を求める時点で、注目している準備 / 用途表現が学習データ中で、 X 以外のある名詞に対する適切な準備 / 用途表現であるとラベルづけされている場合には、前のステップで求めたスコアにバイアス (実験では 10^6) を加える。
4. 求めたスコアに従って、準備 / 用途候補の表現をソートし、そのトップ N を出力として生成する。

2 番目のステップでの素性ベクトルは以下のように生成する。まず、学習セット、テストセットに現れている準備表現に対してユニークな ID をふる。同様に、用途表現に対してもユニークな ID をふるが、準備表現に対する ID とぶつからないような ID をふる。仮に同一の項位置

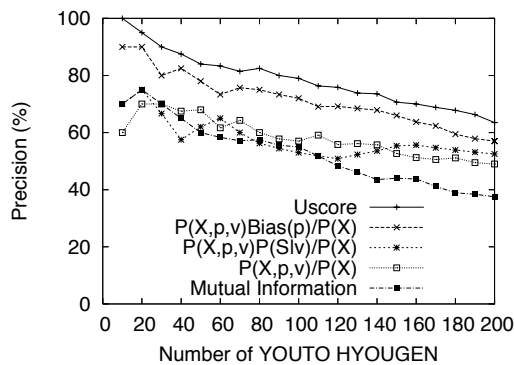


図 1: 用途表現の精度

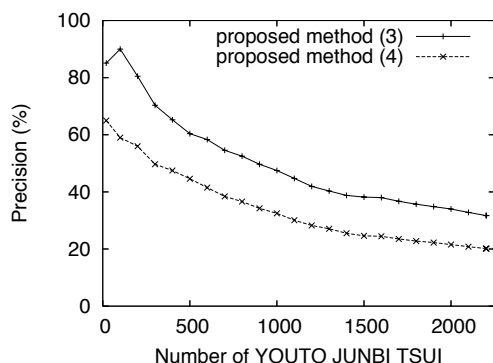


図 2: 準備 / 用途対の精度

が用途表現と準備表現の両方に出現したとしても、それらには異なる ID をふる。ついで、もとの ID に対応する素性を 1、それ以外の素性を 0 とするような素性ベクトルを生成し、SVM の学習、および、決定関数の計算に用いる。一見して明らかなように、このような素性ベクトルを用いるという事は単純に学習データにおいて適切であるとラベル付けされた表現を SVM に記憶させることで判別を行っている事になる。前述の仮説に従えば、パターンにおける共起頻度や、名詞と項位置との共起頻度を考慮にいれれば、性能が向上するように思われるが、少なくとも現在の設定では目立った性能向上は見られなかった。また、提案手法は単純に正解の用途表現、準備表現を記憶させておく場合にくらべて、高い精度を得ることができた。これらのポイントについては後ほどより詳しく述べる。SVM の実装としては TinySVM¹ を用いた、また、Kernel 関数としては 2 次の ANOVA kernel を用いた。これらのパラメータは予備実験において、学習が収束し、また、高い精度が得るものを経験的に選んでいる。

3 実験

実験ではまず 35 年分の新聞記事 (計 3.24GB) と 92.6GB のウェブ文書をパースし、提案手法に必要な頻度を計算した。また、今回の実験では、用途表現ならびに準備表現が定義されやすいと考えられる人工物をさす名詞を対象を限定した。このため先ほどの 35 年分の新聞記事のサブセットである 33 年分の新聞記事に 500 回以上現れ、なおかつ日本語語彙体系 [3] で著者が指定した 10 個の概

念の下位に含まれる 2,766 語をリストアップし、これを人工物をさす名詞と仮定した。以後これらの語を人工物名詞と呼ぶ事にする。次いで、先ほどの 35 年分の新聞記事とは別個の新聞 1 年分から 1000 件の記事をランダムに選び、その最初に現れた人工物名詞の出現 300 回を抽出し、デベロップメントセットとした。このデベロップメントセットは 226 個の異なる人工物名詞を含んでいた。ついで、Step 1 を評価するため、先ほどの人工物名詞でデベロップメントセットに含まれていないものをランダムに選んで 200 語抜き出し、それに対して Step 1 を適用して 200 個の用途表現を出力させ、その結果を 3 名の被験者に判定してもらった。判定に際して被験者は以下のような指示を受けている。まず、提示された項位置は「X を利用する」「X を使う」「X を用いる」「X を活用する」「X を楽しむ」の言い換えになっているかどうかを判定するよう指示された。ついで、各々の人工物名詞はそれが出現する新聞記事と一緒に提示され、もし語が意味的に曖昧な場合には新聞記事の文脈で判断を行うように指示された。判定の結果、200 個の用途表現の内、3 名ともが適切と判断したものが 127 個 (63.5%) あった。出力された用途表現をそのスコア順にソートし、precision をプロットしたのが、図 1 である。比較対象として、人工物名詞と項位置の Mutual Information をスコアとして用いた場合 [9] と、我々のスコアから一部の項を取り除いたものの precision カーブも同じグラフにプロットした。我々のスコアは Mutual Information と項を取り除いたスコアのいずれよりも高い精度を出している。特に後者との差は、スコア中のすべての項が精度向上に貢献していることを示している。ちなみに、判定結果の被験者間での一致度を示す kappa 統計量は 0.57 で [6] によれば、中程度 (moderate) の一致を見ている。

ついで、準備 / 用途対の獲得の性能評価を行うための学習セットを以下の手順で用意した。まず、これまでに用いてきた 426 個の人工物名詞に対して、前節で説明した準備 / 用途対候補の集合 $C(X)$ を計算し、その結果を 3 名の被験者によってその要素が適切な準備 / 用途対であるかどうかを判定してもらった。ついで、3 名の被験者がすべて適切と判断した準備 / 用途対候補のみを正例、それ以外を負例として学習データを作った。この学習データの総数は、4,570 個、また、そのうち正例は 1,106 個 (=24.2%) であり、被験者間の一致を示す kappa 統計量は 0.605 で良い (good) 一致であった。

ついで、これまでに用いた合計 426 個の人工物名詞以外の人工物名詞 200 個をランダムに選び、テストセットとした。これらの名詞に対して準備 / 用途対候補の集合 $C(X)$ (合計 2,204 個) を計算し、今度は 4 名の被験者によって判定を行ってもらった。kappa 統計量の値は 0.61 で今回も良い (good) 一致である。ついで、提案手法によってすべての人工物名詞に対する $C(X)$ の要素のスコアを計算し、準備 / 用途対をソートした。このソート結果を元にトップ N を出力させたときの precision をプロットさせたのが図 2 にあるグラフである。グラフ中の “proposed method (3)” とあるのは、3 名以上の被験者が適切と判断した準備 / 用途対候補を “正しい” ものと見なした場合の精度のグラフであり、“proposed method (4)” とあるのは、4 名の被験者がすべて適切と判断したもののみを “正しい” とした場合のグラフである。たとえば 200 個の準備 / 用途対を出力した場合には、81% が 3 名以上の被験者によって適切と判断され、56 がすべての被験者によって適切と判断されている。ついで、400 個の準備 / 用途対を出力した場合には、65% が 3 名以上によって、48% が 4

¹<http://chasen.org/~taku/>より入手可能

名の被験者によって適切と判断されている。入力された人工物名詞の数が 200 個であることを考えると、これらは妥当な精度であると考えている。また、素性ベクトルに、他の有効であると考えられるもの、具体的には、前節でのべたパターン中での準備 / 用途対候補の共起頻度ならびに、準備表現と人工物名詞との共起頻度、ならびに、用途表現を求める際に計算した U_{score} の値を素性として加えた場合や、それらの共起頻度だけを素性ベクトルとした場合にも精度向上が見られなかったことを付記しておく。なお、それらの共起頻度に従って、準備 / 用途対をソートし、precision をプロットした場合にはいずれの場合も右肩下がりのグラフが得られたため、それらの共起頻度は一定の手がかりにはなるようであるが、提案手法で用いた素性の方がより有効であるということであろう。また、kernel の次数を増やしてもこの傾向は変わらなかった。

提案手法によって出力された準備 / 用途対の例を図 3 に示す。ここで順位とあるのは、提案手法によって計算されたスコアによる順位であり、被験者数とあるのは、提示した準備 / 用途対が適切であると判断した被験者の数である。

提案手法では学習データで入力とは異なる名詞に対して適切な準備 / 用途対と見なされた項位置の対にバイアスを与えていた。このような項位置の対はこのバイアスによって、最終的なソートされた準備 / 用途対では独占的に上位の位置を占めている。実験でバイアスを与えられた準備 / 用途対候補は計 255 個あったが、そのうち、テストセット中で 3 名の被験者が適切であると判断したものは 75% であった。この割合はバイアスを与えられた準備 / 用途対候補をランダムにソートしてトップの要素を出力したときの精度の期待値であるが、これはたとえば、提案手法がトップ 100、トップ 200 を出力した時の精度、それぞれ 90%、81% を下回っている。これは可能な準備 / 用途対候補をリストアップしただけでは高い精度を達成することはできず、機械学習によって得られるような統計的選好が有効であることを示している。また、たとえば、トップ 400 を提案手法で出力した場合には、学習データ中に見つからないか、あるいは適切ではないと見なされた 69 個の準備 / 用途対候補が 3 名の被験者によって適切であると見なされている。この点も機械学習の有効性を示すものと考えられる。

最後に乾の means relation を獲得する手法 [5] で、提案手法で獲得された準備 / 用途対がどの程度獲得できるかの実験を行った。乾らの手法で準備 / 用途対が獲得されるとすれば、少なくとも接続詞「ため」、準備 / 用途対にあらわれる二つの動詞、ならびに入力となる人工物名詞の計 4 語が現れる文が必要である。そのような文が我々の実験で語の頻度を求めたコーパスにどれだけ出現するかを調べた。その結果、獲得される可能性のある人工物名詞と準備 / 用途対の組み合わせは 12 個であり、提案手法は乾らの手法では獲得できない準備 / 用途対を大量に獲得できていることになる。これは、提案手法では名詞-項位置の共起、動詞間の共起を独立に求めており、一般化と同様の効果が得られているためと考えられる。

4 まとめ

本稿ではプラン認識などの高レベルの推論を最終的な目的として、人工物をさす名詞に対して準備 / 用途対という表現の組を定義し、その自動獲得を行う手法につい

順位 / 被験者数	準備表現	用途表現
9/4	書店に行く	書店で買う
12/4	プリペイドカードを発行する	プリペイドカードで買う
80/3	時代劇を作る	時代劇で見る
81/3	初版を出版する	初版で読む
149/4	豆を買う	豆で作る
150/4	点字を覚える	点字で書く
152/3	時代劇に登場する	時代劇で見る
352/4	モスクに行く	モスクで祈る
355/3	ピザを発行する	ピザで滞在する
357/4	フルートを習う	フルートを演奏する
359/4	記録映画を撮影する	記録映画で知る

図 3: 獲得された準備 / 用途対の例

てのべた。最終的な精度は 200 個の人工物を示す名詞に対して、200 個および 400 個の準備 / 用途対を出力した場合には 4 名中 3 名がそれぞれ 81%、65% を適切であると判断した。今後は、さらなる精度向上と具体的に獲得された準備 / 用途対を用いたアプリケーションの開発を行う。

参考文献

- [1] Sandra Carberry. *Plan Recognition in Natural Language Dialogue*. MIT Press, 1990.
- [2] Timothy Chklovski and Patrick Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP-04*, 2004.
- [3] Satoru Ikehara et al. *Goi-Taikei-CDROM*. Iwanami Shoten, 1999. in Japanese.
- [4] Toshiaki Fujiki, Hidetsugu Namba, and Manabu Okumura. Automatic acquisition of script knowledge from text collection. In *Proceedings of The Research Note Sessions of EACL'03*, 2003.
- [5] Takashi Inui. *Acquiring Causal Knowledge from Text Using Connective markers*. PhD thesis, NAIST, Japan, 2004. NAIST-IS-DT0161005.
- [6] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.
- [7] Mirella Lapata and Alex Lasacariades. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):263–317, 2003.
- [8] James Pustejovsky. *The Generative Lexicon*. MIT Press, 1995.
- [9] James Pustejovsky, Peter Anick, and Sabine Bergler. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):221–358, 1993.
- [10] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [11] Ichiro Yamada and Timothy Baldwin. Automatic discovery of telic and agentive roles from corpus data. In *Proceedings of PACLIC 18*, 2004.