特定コミュニティーの文書群からの言語知識の獲得

西野 文人[†], 出羽 達也[†], 福重 貴雄[†], 辻井 潤一[‡]
[†] (株) 日本電子化辞書研究所 [‡] 東京大学
[†]{nishino,izuha,fuku}@edr.co.jp [‡] tsujii@is.s.u-tokyo.ac.jp

1 はじめに

電子化辞書の普及が進み，基本語レベルでは様々な辞書が利用できるようになってきた．しかし，専門用語辞書に目を転じてみると，情報処理用語，医学用語といった大きな分類での専門用語辞書の作成はなされているものの，小さな分類単位でみると，まったく語彙が不足しているというのが現状である．そして，家庭への計算機の浸透，情報検索要求の増大などに伴って科学技術系の用語のみでなく家庭一般，スポーツ，趣味など様々な分野の用語辞書の必要性が高まってきているが，コスト効果の点からこれらの分野の用語辞書の整備はあまり行なわれていない．そこで我々は比較的小さなコミュニティーで使用されている文書群（言語仕様が限定されている少量のコーパス）から，部分言語固有の言語知識（その分野で独特な専門用語や固有名詞などの語彙知識や特殊な記法など）を人間の介在を含めて漸近的に獲得することを目標に研究を進めてきた．本稿では，このような特定コミュニティーの少量のコーパスからの言語知識獲得の考え方を示す[1]．

2 特定コミュニティーの言語知識とは

我々は言語知識の獲得といっても広く一般的な言語知識を獲得するのではなく，分野を絞って特定コミュニティー内で使われている部分言語固有の言語知識の獲得に焦点をあてた．これにはまず一般的な用語に対するオンライン辞書は整いつつあるのに対して，特定のコミュニティーで使用される言語知識ベースの開発はその数の多さや一般性の少なさから非常に遅れていること，しかもこのような言語知識を個別のコミュニティーごとに手で獲得するのは困難であり，特定コミュニティーの言語知識を自動的に獲得することが切に望まれているという社会的背景がある．そしてもうひとつには分野を絞ることによって語の用法が限定されるので言語知識獲得がやりやすくなるという技術的背景もある．

では部分言語固有の言語知識にはどのようなものがあるだろうか．そこでまず将棋の分野に注目し（実際の新聞記事の例を図 1に示す），特定のコミュニティー固有の言語知識にはどのようなものがあるのかを見てみることにしよう．

序盤で角を交換－－第４０期将棋・王将戦七番勝負第一局
米長邦雄王将と挑戦者・南芳一棋王の第四十期王将戦七番勝負（毎日新聞社、スポーツニッポン新聞社主催）の第一局は十七日午前九時、神奈川県箱根町のホテル花月園で始まった。午前九時前、南、米長の順に入室。駒を並べ終えた後、記録の深浦康市三段が歩を五枚取って振り駒を行い、歩が三枚で米長が先手になった。局面は序盤早々角が交換され、前期もよく指された角換わりの戦型が組み上がった。南は棒銀に出た。南の攻勢を目指す棒銀に、米長は一時間の長考で６六歩と応じた。南はすぐに７五歩。まず７筋で駒がぶつかった。南が三十手目を考慮中の午後零時半、昼食休憩に入った。持ち時間各九時間のうち、消費時間は米長一時間五十六分、南一時間十六分。午後一時半再開。

図 1: 新聞記事例（CD- 毎日新聞 1991 年 1 月 17 日の記事より）

1. 語構成は一般的だが一般には使われない語

   形態素解析で通常未登録語として処理されるもの（語範囲認定を誤ることはあるが）であり，固有名や通常の専門用語である．

   例）米長，棋王，振り駒，棒銀

[1] この研究は IPA の創造的ソフトウェアプロジェクトである「知識ベース増殖のためのソフトウェアの開発プロジェクト」として進めてきたものである

2. 一般用語だが特定のコミュニティーの中で特別な意義を持っているもの

   形態素解析では辞書にも登録されており通常の語として解析されるが，その語義は通常と違う（特殊な用途のみに用いられる）ので，共起する語などが通常と異なるものである．

   例）角（将棋の駒の一つである），
   [駒が] ぶつかる．

3. 特殊な意味をもった文字列

   形態素解析では通常一つの語としては認識されない特殊な文字列構成をしているものである．

   例）６六歩

   将棋に限らず，実際の文章の中を見ると様々な意味をもった特殊な文字列が存在する．形態素解析の多くはこのような特殊文字列の存在に対して弱いことが多い．しかし，実用的なシステムでは形態素解析でやるかその前処理や後処理でやるかはともかく，何らかの形でこのような特殊文字列を取り扱う必要がある．

3 言語知識獲得のアプローチ

大量のコーパスに対しては，統計処理を施すことによって様々な言語処理を獲得することができるが，それほど多くの量のないコーパスからでは生のままでは言語知識獲得に充分な統計量が集まらないことも多く，雑音の影響が非常に大きくなる．その結果，たとえばクラスタリングのような処理を行なっても期待するようなクラスタリング結果を得ることはできない．このような少量のコーパスからの処理では以下のようなことを考えることが必要であろう．

1. 複数の知識処理を融合した漸近的処理

   単一の知識処理では有効でなくても，複数の知識処理を組み合わせることで効果がでるかもしれない．そのような処理を行なうためには各種の言語処理がある定まった形式の入出力を取り扱えるようにし，得られた知識を利用してまた新たな処理を行なうような漸近的な言語知識獲得ができるようにすることが必要である．

2. 人間の介在

   少量のコーパスからの言語知識獲得には高度な知識が必要となる．そのような知識はそのコミュニティーの知識を有する人間が与えるのがリーズナブルなものであろう．むしろ，人間の方が中心で，人間の不完全な部分をコーパスを使って補ってやるという考え方が必要かもしれない．

このようなアプローチを考えた時，各種の言語処理をどう統合的に扱うか，すなわち，人間を含めて，各処理が与えたデータや知識をいかに統合的に扱うかが重要になってくる．この問題に対して我々は中央知識データベースと呼ぶ機構を作成し，各種の言語処理がある定まった形式の入出力を取り扱えるようにしている [1].

さて，人間とシステムが共同して言語知識を獲得するにはそれぞれがどのような役割を担ったらいいであろうか．我々はそのコミュニティー特有の知識を有する人間が言語知識獲得の種となる情報を与え，それをもとにシステムが，そのコミュニティーの中のコーパスあるいは一般文のコーパスを参照して，関連する情報を提示するという使い方を想定することにした．例えば，以下のような使い方があるだろう．

1. 限定された意味で使用される語を提示する．

   例えば，駒の名前として，「歩」，「香」，… を与えることで，「成る」，「打つ」，「ツギ」のような将棋の世界で特殊な共起関係を提示する．

2. 同じクラスの要素のものを提示する．

   あるクラスの要素を示し，そのクラスに属する別の要素を提示する．例えば，駒の名前として最初に与えた知識では，「と金」，「角行」あるいは，「ビショップ」，「卒」といったようなものを人間は忘れているかもしれない．そこで，そのようなヒントさえ与えられれば，そのコミュニティーの中の人間ならば，あとはコーパスがなくてもさらに知識を補充していくことが可能であろう．

この他にも，与えられた種が出現するパターンから算用数字＋漢数字＋駒名 [＋ action(「成」，「不成」，「右」など）] のような統語的な規則を見つけるというようなものも考えられるが，今回は以上の２点に絞った．

4 実験

実際に特定コミュニティーの少量のコーパスから言語知識を獲得することがどのくらい可能であり，また何が困難であり，どのような支援が可能なのかを調べるために簡単な調査実験を行なった．

実験には CD- 毎日新聞 91,92,93 版からキーワード「将棋」で検索した 214 記事を利用した．この中には「将棋倒し」など本題が将棋と関係ない記事も含まれている[2]．そして，形態素解析をして，名詞連続や連体修飾，係り受けなどの関係にある二つの単語をその間の表層関係とともに取り出した．

実験としては，ユーザが「歩」，「香」，「桂」，「銀」，「金」，「角」，「飛」，「玉」を駒名であることを指示してやることにより，将棋という分野における特徴的な述語の取り出し，およびユーザが与えたクラス（駒）の中の不足している要素（駒名）の補完を行なった．

まず，限定された意味で使用される語を特定するには，一般的な用法で使用されているもの（例えば金曜日の意味の「金」など）を除外する必要がある．そこで，一般文との使用例の相違を見て，実際の駒名として使われている部分と将棋に特徴的な語を取り出した（図 2）．ここで数値は tscore の値であり，これは単語 w_1 と単語 W の同時出現可能性と単語 w_2 と単語 W の同時出現性との相違性を見るためのスコアとして，以下の計算式によるものである [2].

$$t = \frac{P(W|w_1) - P(W|w_2)}{\sqrt{\sigma^2(P(W|w_1)) + \sigma^2(P(W|w_2))}}$$

将棋のことを知っている人ならば，この結果を見て，「同桂」，「玉頭」，「銀成」，「2四歩と突く」といったような使い方がわかる．

次に，同類のものの補完であるが，これはあらかじめ与えられた知識が本来の知識のかなり多くの部分（頻度的に）を占めるならば，それと同等の使いかたをする語を探すことで補完が可能なのではないかと考えた．そこで，駒名と共起する単語に共起する単語（関係まで含めて同じもの）を取り出すことにした．しかし，将棋の駒に共起する語に共起する語が駒名とは限らないので，共起しているものが駒名である可能性の高いものを残すことにした．図3にその結果を示す（なお 100% 駒名のケースは新しい知識獲得につながらないので除去してある）．ここで，「#」のあとの数値は，駒名での共起回数と全体での共起回数である．この結果から，「香車」のような語を補完すればよいことがわかる．

[2]新聞紙上に毎日掲載されている観戦記は CD に含まれていない

```
4.82   修飾 / 同:koma
2.69   名詞連続 / 頭:koma
2.52   名詞連続 / 成:koma
1.96   と / 突く:koma
1.75   と / 出る:koma
1.75   名詞動詞連続 / 換わる:koma
1.75   名詞動詞連続 / 交換:koma
1.53   を / 回る:koma
1.53   を / 突く:koma
1.29   で / 取る:koma
1.29   と / 取る:koma
1.29   と / 突っかける:koma
1.29   に / 長考:koma
1.29   を / 取る:koma
1.29   名詞連続 / 得:koma
1.02   と / する:koma
1.02   と / 引く:koma
1.02   に / あてる:koma
1.02   を / 引く:koma
1.02   を / 交換:koma
1.02   名詞連続 / 損:koma
```

図 2: 将棋の駒に共起する特徴語

```
# 11/12
11 名詞連続 / 成,koma
1 名詞連続 / 成, 右
--
# 6/7
6 を / 回る,koma
1 を / 回る, 苑内
--
# 5/6
5 で / 取る,koma
1 で / 取る, 熟考
--
# 4/6
4 を / 引く,koma
2 を / 引く, 香車
--
# 6/10
6 を / 突く,koma
4 を / 突く, 筋
--
# 5/9
5 を / 取る,koma
1 を / 取る,title
2 を / 取る, 位
1 を / 取る, 攻勢
```

図 3: 将棋の駒の割合が多い共起語

しかし実際にはこのような駒名の割合が多いところだけでなく，割合が低いところにも駒名がでており（図 4），単独の証拠（共起語）だけでなく，複数の証拠を組み合わせることが必要であることがわかる．そこで，各語の共起頻度を要素とするベクトルと 各共起語の駒である割合であるベクトルとの内積を語の長さで割ったものでソートした結果を図5に示す．

```
#  1/2
1 も / 切る,koma
1 も / 切る, 竜
--
#  2/5
2 名詞連続 / 攻め,koma
3 名詞連続 / 攻め, と金
--
#  1/4
1 を / 寄せる,koma
2 を / 寄せる, と金
1 を / 寄せる, 陣
--
#  1/4
1 名詞連続 / 左,koma
1 名詞連続 / 左, 士
1 名詞連続 / 左, 写真
1 名詞連続 / 左, 進
--
#  1/4
1 名詞連続 / 先,koma
1 名詞連続 / 先, いずれも
2 名詞連続 / 先, 飛車
```

図 4: 将棋の駒の割合が少ない共起語

```
4.39 sei
2.70 筋
1.98 成
1.98 ホテル
1.77 将棋
1.70 と金
1.61 手
1.45 棋士
1.33 香車
1.32 馬
```

図 5: 将棋の駒名候補

この結果では，「香車」以外にも「と金」と「馬」が示され，これらの語を補う必要があることがわかる（「竜」，「飛車」なども補うべきであることは，将棋を知っている人間なら容易に思いつくであろう）そこで与えられた手がかりをもとに成り駒の名前や二文字の名前を補って同じ処理を繰り返すと（前回提示したもので補充されなかったものは除去してある），図6のようになった．

```
1.55 日夜
1.25 攻勢
1.22 タイ
1.15 封じ手
1.11 位
1.07 右
1.05 考
1.01 title
1.00 同所
0.92 局面
0.89 シリーズ
0.87 兵
```

図 6: 将棋の駒名候補 2

ここからは，「兵」という中国象棋の駒名が新たに提示され（最初の処理では 0.07 でずっと下位の方にあった），これを手がかりにまた知識が漸近的に増殖されることになる．

5 おわりに

本稿では限られたコミュニティーの中で使用されている少量のコーパスから，そのコミュニティーに特有な言語知識の獲得に関して述べ，人間を介在した漸近的な言語知識獲得について簡単な実験結果を紹介した．少量のコーパスからの処理では，頻度のような統計量はあまりあてにならない．今回の適合率ではまだ不十分かもしれないが，数式を工夫してもあまり効果はないようである．色々な処理を積み重ねること，あるいは，負の情報をうまく導入してやって二度め以降の提示で以前と同類の余分な提示はしないようなユーザインタフェースの工夫が大事であろう．

参考文献

[1] 西野文人, 杉山健司, 辻井潤一：知識ベース増殖のための中央データベース，言語処理学会第 3 回年次大会, D1-3, pp. 123–126 (1997).

[2] Church, K., Gale, W., Hanks, P. and Hindle, D.: Using Statistics in Lexical Analysis, in Roche, E. and Schabes, Y. eds., *Finite-State Language Processing*, pp. 115–164, The MIT Press (1997).