基般

3.7 知識獲得

柴田 知秀（京都大学）

知識獲得とは

計算機によるテキスト解析やアプリケーションを高度化するためには，人間が持っている常識的な知識を計算機に与えなければならない．文法や基本的な語に関する知識はある程度人手で記述できるが，固有名詞に関する知識や語と語の関係などは人手で記述しきれないので，自動獲得する必要がある．知識のタイプとしてはさまざまなものが考えられるが，以下のようなものが挙げられる．

- 同義・上位下位：同じ意味を表すものや，ある語がある語の上位／下位概念を表すもの
  例）MacBook Air ＝ MBA →ノートパソコン
- 固有表現：人名，地名，組織名など
  例）ネイマール：人名，バンプレスト：組織名
- 格フレーム：「誰が何をどうした」という名詞と動詞の間の関係
  例）{人, 男, 彼}が{犯罪, 犯行} を犯す
- 事態間知識：2 つの事態（「誰が何をどうする」）の間の関係．“「誰が何をどうする」としばしば「誰が何をどうする」”という時間経過の関係や，“「誰が何をどうした」結果，「誰が何をどうする」”という因果関係などが含まれる．
  例）x:{男，容疑者}が犯罪を犯す⇒x:{男，容疑者} が逮捕される☆1

知識獲得の方法

知識獲得源としては大規模な Web テキストや Wikipedia などが挙げられる．構造化されたテキストからはパターン，もしくはルールを用いて知識を獲得することができる．構造化されていない大規模テキストからは共起関係を手がかりに知識を獲得することができる．たとえば格フレームの場合，「男が犯罪を犯した」「彼が犯罪を犯した」のような表現が多数現れることをもとに，それらをクラスタリングすることにより，知識を獲得することができる．

Winograd Schema Challenge

知識獲得の研究において難しいのは評価の問題である．システムが獲得した知識の中からランダムにサンプリングして精度を算出し，その精度が高かったとしても，必ずしも獲得された知識がほかのタスクで有用であるとはいえない．そこで，獲得した知識をあるタスクに適用し，その精度が向上するかによって知識の有用性を確認することが考えられる．

近年，常識的な知識の獲得を評価するものとして，Winograd Schema Challenge（以降，WSC と呼ぶ）という評価セットが構築されている[1), 2)]．日本語の WSC の例を**表 -1** に示す．この評価セットは照応解析と呼ばれるタスクであり，たとえば，表 -1 の（1-a）の問題では，“彼女”（照応詞と呼ばれる）に対して，“デビー”と“ティナ”（先行詞候補と呼ばれる）が与えられ，“彼女”が“ティナ”を指していることをシステムが正しく認識できるかどうかが問われる．この評価セットでは，常識的な知識が必要な問題が集められており，上記の問題では，「X さんが Y さんに水をかけると，Y さんがびしょびしょになる」という事態間知識が必要となる．エラー分析ワークショップでは日本語 WSC を題材とし，必要な知識の分析と現状の知識獲得の誤り分析を行った．

☆1 左辺の x と右辺の x は対応付いていることを示す．

(1-a)	デビーが**ティナ**に水をかけた.	彼女はびしょびしょになった.
(1-b)	**デビー**がティナに水をかけた.	彼女はめんどうをおこしたのだ.
(2-a)	猫は**犬**より賢い.	彼らは理由なく吠えるからだ.
(2-b)	**猫**は犬より賢い.	彼らはいつも足から着地するからだ.
(3-a)	**男**は隣人の自転車を盗んだ.	彼は 1 台必要だったからだ.
(3-b)	男は**隣人**の自転車を盗んだ.	彼が 1 台余分に持っていたからだ.

表 -1　日本語 WSC の例（下線を引いた語は先行詞候補，太字の語は正解，波線を引いた語は照応詞を示す）

必要な知識の分析と現状

日本語 WSC を分析したところ，問題を解くために必要な知識は以下のように分類できることが分かった（下記で括弧内の数字は 100 問あたりの問題数を示す）．それぞれの知識とその知識獲得の現状を述べる．

1. 選択選好（26）

選択選好とは，ある動詞のある格（「が」「を」「に」など）がどのような名詞をとりやすいかという知識のことをいう．たとえば，動詞「吠える」の「が」は，「猫」よりも「犬」の方をとりやすいというもので，この知識を用いることにより，表 -1 の（2-a）の問題を解くことができる．選択選好は格フレームという形で精度高く自動獲得されている．

2. 事態間知識（22）

事態間知識は最初に挙げたとおり，「誰が何をどうする」としばしば「誰が何をどうする」という関係で，以下のような知識を用いることで，表 -1 の（1-a）の問題を解くことができる．

X:{私, 彼, ...}がY:{彼女, ...}に水をかける
⇒Y:{彼女, ...} がびしょびしょになる

事態間知識を Web から自動獲得する研究があるが，精度はそれほど高くなく，また，カバレッジが十分ではないのが現状である．

3. 難問（29）

たとえば，(3-a) や (3-b) のような問題を解くには複数の知識を組み合わせる推論を必要とし，現状の技術では大変難しい問題である．

現在の知識獲得技術ではまだ 60% 強ほどしか解けていない．詳細については参考文献 2) を参照せよ．

今後の展望

常識的な知識獲得はまだまだ始まったばかりで，今後，知識獲得の精度を上げることもさることながら，どれくらいの知識が必要なのか，どのような粒度で知識を獲得すればいいのかなど，課題がたくさんある．常識的な知識獲得が進み，ここに紹介した問題が結構解けるようになると，言語処理が一段進んだと言えるであろう．

参考文献

1) Levesque, H. J. : The Winograd Schema Challenge. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning (2011).
2) 柴田知秀，小浜翔太郎，黒橋禎夫：日本語 Winograd Shema Challenge の構築と分析，言語処理学会第 21 回年次大会論文集，pp.493-496 (2015).

(2015 年 10 月 6 日受付)

柴田 知秀（正会員）shibata@i.kyoto-u.ac.jp
2007 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．2014 年より京都大学大学院情報学研究科特定講師，現在に至る．自然言語処理，特に知識獲得や情報検索の研究に従事．言語処理学会，ACL 各会員．