自然言語処理用辞書における語義文

5F−6

野村直之　小泉敦子　野口喜洋　横田英司

(株)日本電子化辞書研究所

1. はじめに

自然言語処理用辞書では、登録語に対して複数の語義を設定することがある。これにより解析時に語の意味を同定する機能の基盤が提供できる。設定した語義を1文で表現したものを語義文と呼ぶ。本稿ではこの語義文が備えるべき機能について考察する。

語義文の初期セットの開発にあたっては、伝統的な紙の辞書の場合と同様に人手で記述する方法がEDR88に報告されている。語義文の初期セットを評価し改良するにあたって機械の助けを借りて品質の向上をはかるならば、改良中の語義のセット自身を用いて解析した結果を利用することになる。即ち必然的にboot strap方式となる(例えばMuraki91)。

boot strap方式によって語義文の品質を改良するには、機械で診断できるような評価の尺度ならびに人手による評価の尺度を与える必要がある。これらの評価の尺度を用いて適切に品質の管理を行なわなければ、改良の効率の低下はおろか、語義文の品質低下の可能性すら生じる。

以下では、まず語義文の品質評価の尺度を考察するにあたって語義文の機能の整理を試みる。次にそれらの機能の実現に伴う課題とその解決案、そしてその実現の手順を取り上げて考察する。さらに、いくつかの解決案に共通する具体的な目標として語彙制限と文体統一を取り上げてその効果について考察する。

2. 語義文に求められる機能

語義文の利用目的は、第1に見出し語に対してその語義文で表現された語義を設定することの妥当性が評価できるようにすることである。第2に、他の辞書情報(品詞,訳語等)と語義との整合性の評価である。第3に、その辞書を用いてテキストを解析した結果の評価が考えられる。これらの利用目的に照らして、以下に示す可読性、完結性、非循環性、省検索性、想起性、一意性、弁別性の7つの機能が語義文に望まれる。

機能1)「可読性」
定義:「文法的、意味的、語用的、談話的に適格な文であること。」
説明:文法的誤りや、「赤い手触り」のような意味的破綻があってはならない。「雨がおりる×(cf. 霜がおりる○)」のような不自然なコロケーションがなく、且つ自然な談話の流れに沿っている。

機能2)「完結性」
定義:「同一の辞書内で定義された語句とその語義のみを使用して、語義文が記述されていること。」
説明:語義文に使用されている語句の検索が同一辞書内で可能という意味で辞書システムとして閉じていること。

機能3)「非循環性」
定義:「語義文中の使用語句の語義文を再帰的に検索した際に、一度検索した語句が再び現われて循環しないこと。」
説明:完結性が実現されても、複数の語の語義文が互いの語を参照するという循環があると語義文の利用価値が減ずる。
例:【主観的】客観的でないこと;
【客観的】主観的でないこと;

機能4)「省検索性」
定義:「参照者が語義文を理解するのに必要な、語義文中の使用語句の再帰的な検索の回数が少ない。」
説明:「非循環性」が達成されていても、参照者が元の語義文を理解するのにあまりに多数の参照回数を必要とするのでは語義文の利用価値が減ずる。
例:【内蔵】(ウチクラ) 官物を納めた倉庫　;　語義文1
→大和朝廷の所有物を納めた倉庫　;　語義文2
上例では、語義文2の方が、改めて官物を検索することなく語義を理解できる可能性が高いと考えられる。そこで相対的に語義文2が省検索性が高いことになる。

機能5)「想起性」
定義:「なんらかの概念を想起させられる表現となっていること。」
説明:機能1)～4)が達成されていても、下例のように、語義文単独では単一の概念を想起し難いものが許されてしまう。
例:【お早う】　おはようございます
上例のように概念の説明を意図した文体からはずれるものや、「月と石のこと」のように単一の概念が想起できないものが、想起性の低い語義文に該当する。

機能6)「一意性」
定義:「語義文が一意に解釈できること。」
説明:機能1)～5)が達成されていても、想起できる概念が複数あれば、見出し語や訳語等の辞書情報との対応の適否を判定するという目的に支障が生じる。
例:【価値】　もののねうち
上例は、「ねうち」が金銭面の価値のことなのか、ある物事の本質的な意義のことを指すのか不明な点で一意性に反する。

機能7)「弁別性」
定義:「複数の語義文の間で同一性、差異、重なりが判別できること」
説明:　単独では各々一意に読める語義文であっても、記述された対象が互いに同一であるかどうかが判別できなくては、語義の改廃に支障が生ずる。また、記述された対象が異なるならば、重なる部分や差異の部分が判別できなくては、複数の語義を統廃合して改良するのに支障が生ずる。
例:【腰】　腰のあたり;　【腰】　腰のまわり;
上例の2つの語義文は、共に身体の部位または周辺を指していることはわかるが、同一の対象を指しているかどうかの判定が困難である。従って、互いに弁別性の悪い語義文に該当する。

これらの7つの機能は、各々単独で実現できる性質のものではない。例えば、可読性1)が損なわれたならば、他の全ての機能の実現に支障をきたす。同様に想起性5)の実現は、一意性6)、弁別性7)を判定するための前提である。一方、計算機で再帰的な検索を行なうという支援を経て完結性2)、非循環性3)が実現されていれば、可読性1)、想起性5)、弁別性7)の評価と改良は容易になる。これらの依存関係をもとに、個々の改良の手順や共通に参照する資源を吟味することによって機能全体の改良を効率化する戦略が設定できる見込みがある。

3. 各機能を実現するための課題と解決法

本節では、個々の機能における特有の課題と解決法を論じて依存関係を具体的に考察する。可読性1)については、「どのような人間(あるいは機械)にとって可読か」という評価の規準を求め

How to develop word meaning descriptions in a dictionary for natural language processing; Naoyuki NOMURA, Atsuko Koizumi, Yoshihiro Noguchi and Eishi Yokota
Japan Electronic Dictionary Research Institute Ltd.

るのが課題となる。同じ母国語話者であっても、文法的適格性や語用の自然さの判断には個人差がある。これらの判断規準を網羅するのは困難であるが、語義文の記述に必要な範囲に限ってでも、規準の収集と明確化の努力がなされるべきである。

完結性2)は、使用語句だけの完結性であれば、語義文を形態素解析し、未知語の有無を調べることによって評価することができる。しかし、語義文に使用された語句の語義の存在も保証するには、解析時の語義選択が必要となるため、boot strap方式では、人手による判断が欠かせない。辞書とそれを用いた解析プログラムの精度が不十分ならば、機械には語義選択を行なわせず、辞書検索の効率化にのみ利用する方が効率が高い可能性がある。

非循環性3)は、2)と同様の方法で辞書を再帰的に検索する過程で、前述の循環を検出し、それを断ち切ることによって実現される。循環を断ち切るには、他の語句による言い換えや、前節の例のような否定表現などを避けた語義文の記述を行なうことが1つの課題となる。抽象的な語義の場合は困難が予想されるが、語義文の具体例を収集・整理するなどによる解決を目指す必要がある。

省検索性4)は、判断する人間の知識の個人差によって評価結果が食い違うという問題がある。前述の例では、「官物」を検索しなくとも「大和朝廷の所有物」と書き換えなくとも理解できる評価者もいるであろう。また、1度理解したつもりでも使用語句を検索して誤解に気づいたり、一意性6)が実現されていないことに初めて気づくような事態も生ずる。

想起性5)が実現されない典型例は、見出し語の用例だけによる記述や、類義の語句に言い換えただけの記述、あるいは、その語句が発話される典型的な状況の記述（決まり言葉や慣用句に多い）である。これらは、類義の語義文との弁別性を評価する作業を困難にする等の問題を生じさせる。様々な理由から語義文が書きにくかったこと自体が想起性の低い原因であるため、解決法も各種の事例の収集と整理が中心となる。例えば、相の副詞「まだ」の語義文として「ある状態がこれから先にも存在していることを示す状態属性」を示して時相を表す語義文の記述の指針とする、等の対策である。

一意性6)についても、省検索性4)や想起性5)と同様に客観的な判断が難しい。その原因は、1つには一意性の判定が、記述された説明文内の文脈に依存する点に求められる。

例：　光子(ﾐﾂｺ)の買物　v.s.　光子(ｺｳｼ)の質量

そこで、一意性の確保には、ある程度以上、説明文を詳細に記述する必要がある。しかし過度な詳細さは、語義を不必要に狭く限定したり曖昧にする可能性を招く。これらの両条件のバランスがとれた適切な記述法の追及が1つの課題となる。

一意性の判断が揺れる原因は、十分高い尤度をもつ曖昧性の存在自体に気付かないことにも求められる。そこで他の解釈を顕在化するような支援環境の構築がもう1つの課題となる。

弁別性7)の実現のための課題は、同一見出し語の語義文の場合と異なる見出し語の場合とで異なる。前者であれば、一意性実現の課題と同様、語義文の記述の詳細さの規準が重要となる。

例：【化粧紙】相撲取りが身体を拭く紙　；　語義文1
　　【化粧紙】化粧室で手を拭く紙　；　語義文2

上例では、2つの語義文の構文的複雑度は同等であり、且つ組み合わせ的には、「相撲取りが化粧室で手を拭く紙」が可能なことから重なりが考えられる。実際には、語義文1は「土俵上で」という限定が不足していたという意味で一意性が不十分なことが、この弁別性評価の際に検出される。このように、一意性実現のための詳細さの規準が弁別性の判定に役立ち、逆に弁別性の判定作業から一意性判定へのフィードバックがあるため、両機能の実現は互いに連携させる必要がある。

異なる見出し語の語義文間の弁別性ならば、まず互いに類義の語義文の候補を集めることが課題である。これについては、第1に既存のシソーラスや2言語間の両方向の対訳データから候補を得る方策が考えられる。専門用語などで全く新規の語彙群を対象にする場合は、第1の方策が困難である。この場合、次節で述べる語彙制限と文体統一によって語義文そのものの類似性を出していくという第2の方策が考えられる。

4．語彙制限と文体統一の効果

前節に記した個々の課題のいくつかは、語義文を記述する際に使用する語彙を限定する**語彙制限**を目標とする戦略によって効率良く解決できる見込がある。従来、Longman78に、約2000語で中型の英語辞典の語義を記述した実例がある。ここでは、語義の制限までは達成していないが、それが達成できなくとも完結性や可読性については、「未知語／未知語義」が無くなるという観点から効果がある。語義の制限を行なった場合、それは即ち非循環性を実現するための規準となる語義のプリミティブのセットを定めたことになる。また、語義文中の使用語句の語義が定まることから、各機能、とくに一意性と弁別性の判断に恣意性が入り込むのを効果的に防止できる。

もう1つ、各機能の改良の際に共通に目標とすべき戦略として、**文体統一**がある。これは、語義文における機能語の使い方や修飾句のパターン等を制限して統一を目指す戦略である。文体統一は、第一に語義文を読む際の解釈の仕方を一定にすることによって、全ての機能の評価の効率化、高精度化に貢献する。特に、一意性や弁別性の評価の効率化が期待される。また、語義文として妥当な文体統一がなされれば、定義から想起性は自動的に達成される。可読性を実現するための支援ソフトウェアの開発も容易となる。

5．おわりに

自然言語処理用辞書における語義文に求められる7つの機能、**可読性、完結性、非循環性、省検索性、想起性、一意性、弁別性**を指摘し、それらを達成する際の手法や課題を取り上げて考察した。特に、これらのいくつかに共通に貢献する語彙制限と文体統一の効果を指摘した。

我々は、現在、EDR88に記された計画の延長で数10万個の語義文のセットの改良に取り組んでいる。品質の漸進的な向上を目指して、文法性の判断規準や使用語彙の緩やかな統制に関するガイドラインを開発し、個々の機能の向上をはかっている。また、ボトムアップに収集された語義文のパターンを分析、整理して文体の分類を改良しそれを適用しながら緩やかに文体統一へ向かうアプローチを追及している。

実際の辞書開発においては、7つの機能を同時に追及するのが得策とは限らない。また、特定の機能、たとえば「完結性」を先に実現したとして、それが他の機能の実現や改良にどの程度貢献するかにも未知の部分が多いため、今後の検証が必要である。特定の機能実現を保留してとしても、必ずしも本稿で論じた機能間の依存関係故に他の機能の低下を招くというわけではない。例えば語義文単独による一意性の実現を保留し、代わりにその語義文の理解に用いた例文を併記するという対策をとれば、弁別性の実現の停滞を防止できる見込みがある。今後とも、これらの方策を通じて、本稿に記した課題を解決する具体的な方法論の開発と検証を行なっていきたい。

参考文献

[Longman78] Longman Group Limited, "Longman Dictionary of Contemporary English", ISBN 0 582 52571 3, 1978

[EDR88] EDR Technical Report, 「概念辞書（第2版）」より 4.1 「概念見出し」, EDR TR-012, 1988

[Muraki91] Muraki, K. "Machine Translation Systems and Large-Scale Electronic Dictionaries", Proc. of the Int'l Workshop on ED, EDR TR-031.