日本語の意味タグ体系を定義する試み
FrameNet の視点から

黒田　航　　井佐原　均
独立行政法人 通信総合研究所
{kuroda,isahara}@crl.go.jp

2004 年 2 月 12 日

概要

本論文は Berkeley FrameNet (FN) [5, 6] に (緩やかに) 準拠して日本語のための意味タグ体系を定義する可能性について考察する．そのような目的のために FN が適している理由の一つとして，一貫した意味タグ体系の定義する際に不可欠な「意味要素の自然な分類特徴」が得られる点を指摘する．

1　背景

過去十数年間，自然言語処理は目覚ましい進歩を遂げた．その要因の一つは言語資源の充実によって可能となった機械学習アプローチの普及と定着である．品詞・統語情報などの付加情報（アノテーション）つきのコーパスを訓練データに用いた機械学習に基づいて，以前は不可能だった様々な成果が達成された．

このような手法によって達成された解析技術の高度化には目を見張るものがあり，同様の効果が意味処理にも期待しうるが，一つ大きな障壁がある．現時点では意味処理に適切な言語資源が存在しない．

このような空隙を埋めるのは急務であると考え，通信総研の自然言語グループでは，次のような企画を開始した: (i) Berkeley FrameNet [5, 6] に準拠した日本語のための意味タグ体系 S の開発; (ii) S (の部分タグ) のついた日本語コーパスの構築と公開 (ただし，現時点で公開するコーパスの規模は決まっていない).

以下では，特に (i) に関して FN が意味タグ体系の定義に有効だと考えられる理由を論じる．

2　FrameNet が意味タグ体系の定義に有効である理由

議論の始めに，[13] の意見を参考にしつつ，意味タグ体系が満足すべき一般的性質を幾つか上げておく．有効な意味タグ体系とは，

(1) 十分な体系性と一貫性が備わっていて，機械学習可能である
(2) 品詞情報や統語情報の体系から分離されていると同時に，それらと統合されている (か統合可能である)
(3) 特定の知識領域や目標課題 (e.g., 機械翻訳) に限定されない一般性と網羅性を有し，多くの分野の研究者が有用な情報を引きだせる
(4) 必要に応じて拡張可能である

これらは条件としては網羅的ではないが，要点は尽くしていると思われる．

FN は特に (3, 4) の問題に関して有効なアプローチである．以下では，この点に関して，詳しく説明する．この論文では扱わないが (1) の問題は [8] で検討されている．

2.1　課題としての意味タグづけ

品詞タグづけ POS tagging は，形態素解析の出力である形態素列に最適な品詞タグ列を割り当てる手順である．**意味タグづけ** *semantic tagging/marking*

も，本質的には同様な手順として表現することが可能であるが，それには品詞タグづけにはない問題がある．例えば，

(5) 仮に意味タグ体系が閉じているとして，その空間 (おそらく数百から数千のオーダー) は品詞タグの空間 (数十から数百のオーダー) より広大である
(6) 品詞タグづけの単位が比較的明瞭 (e.g., 形態素, 語) であるのに対し，意味タグづけの単位は (長年の言語学の研究にも係わらず) 今もって十分に明瞭だとは言いがたい
(7) (5, 6) の当然の結果として，意味タグづけの場合，最適なタグ列を決定する過程で生じる相互依存性がケタ違いに大きい
(8) 品詞タグ体系を定義するのに使用される分類特徴 (±countable, ±inflectional, ...) ほど自明な分類特徴が，意味タグ体系の定義の場合には得られない．つまり「意味要素の自然な分類特徴」の発見は自明ではない

(7) は計算的な側面を含み，FN が直接解答を与える問題ではないが，(6, 8) の問題に関して，FN は非常に有効な答えを提供しうる．以下では特に (8) の分類基準発見の問題に関して，その理由を述べる．

2.2 FS/FN の基本概念

FN は Fillmore のフレーム意味論 Frame Semantics (FS) の応用であり [4]，その意味で，日本語 NLP でなじみの深い格文法 Case Grammar [3] の発展形でもある．

FS/FN は「理解には基本単位が存在する」という仮定を立て，その単位を (意味) フレーム (semantic) frame と呼ぶ．**この意味でのフレームは非言語的な単位で，ヒトが理解できる状況を定義する構造体である**．

フレームが特定し表現しているのは「何が何のために何をどうした」という理解の単位である．**フレームが特定された時，(あるレベルの) 理解が達成される**．この際，フレームの特定性の程度の差によって「浅い理解」と「深い理解」の差が生じる．

「何が」「何のために」「何を」のような項の性質はフレームが決定し，それらの意味タイプによっては定まらない．つまり，フレームは項の状況における役割を定める．このようにして定まる状況相対的な意味役割を FN では**フレーム要素** *Frame Element (FE)* と呼ぶ．

語のタイプに係わらず，語は様々なフレームを喚起 evoke するが，その喚起の強度は語のタイプによって異なる．**動詞は特にフレームの特定に大きく貢献するが，それでも完全に一つのフレームを特定はしない**．動詞と名詞 (群) との同一文内の組み合わせによってしかフレームは定まらない．例えば動詞「襲い手が犠牲者を襲う」には，[10] が記述するように，次の (i)-(iv) のような基本フレームとその上位/下位フレームが幾つか存在する: (i) ⟨(主に捕食を目的とした) 動物による襲撃 ⟩ のフレーム，(ii) ⟨(主に資源の強奪を目的とした) 人間による攻撃 ⟩ のフレーム，(iii) ⟨ 自然災害発生 ⟩ のフレーム，(iii) ⟨ 活動への打撃発生 ⟩ のフレーム．これらのフレームが存在することで，(9)-(12) のそれぞれの表現で曖昧な指示 “それ” の意味タイプが (決定可能でなくても) 推定可能となる:

(9) 人食い鮫がそれを襲った [“それ” の型= { 人, 魚，ケガをしたイルカ, ...}]
(10) 強盗がそれが襲った [“それ” の型= { 銀行, 現金輸送車，コンビニ, ...}]
(11) それが東京を襲った [“それ” の型= { 地震, 台風，インフルエンザ, ...}]
(12) それが市場を襲った [“それ” の型= { 株価の暴落，恐慌, ...}]

また，道具の使用が含意されるのは，(ii) のフレームのみである．

このことを一般化して言うと，**どんな語も単独ではフレームを特定する力はない**．これが語の多義性の原因となる．別の言い方をすれば，**語の多義性，曖昧性が解消されるとは，(ほかの語との共起によって) 意味フレームが特定され，フレーム内でのその語の意味役割が定まること**である．

FN が FS と異なっている点は，FN では多数のフレームが継承関係などによって (OOA 風に) 組織化

された構造である点に注目している点にある．

FS/FN の考える意味役割は，一方でフレーム相対的，状況相対的である．それが状況相対的である理由は，それがモノの物理的，客観的特性には還元しえないからである．他方で，FS/FN の考える意味役割は，多分に文化相対的である．FS/FN はもはや，格文法の頃のような「普遍的な意味役割の目録」に基づく意味記述は目指していない．

これらの点から明らかなように，*FS/FN* **は理解の記述を指向しており，真理条件の記述を指向する意味論とは一線を画する**ものである．

FS/FN に基づく意味タグ体系には限界もある．例えば，修飾部に現われる形容 (動) 詞の扱いは自明ではない．これらの意味の基盤を何に求めるかは，現時点では見通しが立っていない．

2.3 FS/FN が意味記述に関して示唆すること

以上の議論から解るように，意味タグとして FE を採用するのは有効である．これが正しいならば，FN/FS は (8) の問題に対して (間接的には (6) に対しても) 有望な答えを出している．一方でこれが示唆するのは，有用な意味タグ体系は客観的特徴の集大成としてのシソーラスよりも，理解の単位として意味フレームの詳細な記述に基づくべきでだということである．これにより，古典的な「フレーム問題」，つまり記述量の爆発を避けることが可能となると考えられる．

3 類似の枠組みとの比較

動機や目標は異なるが，意味タグ体系を定義する試みは幾つも存在する．その代表的なものは 橋田浩一が提唱している Global Document Annotation (GDA), WordNet [2], Resource Description Framework (RDF) [11] とその利用形態としての Semantic Web [1], Ontologies [12] などである．

また，機械翻訳などの特定の言語処理に役立つ辞書構築の企画の一つと FN を見なすならば，それは格フレーム辞書構築 [9] と明白な関連をもつ．

このような試みと FN とのあいだには興味深い類似点，相違点が存在するが，FN は次の点で際立っている．

(13) 知識構造の直接表現でなく，それを資源として達成される理解のモデル化を目指す

(14) (13) の結果として，型ベースではなく役割ベースのコーディング体系を採用する

(15) 言語学者の意味に関する優れた直観と工学者の優れた技術力を統合する

第一の点に関しては，すでに述べた．ほかに二点に関しては，分量の制限もあり，この論文で詳しく論じるのを避けるが，ポスター発表では，これらの点に関しても十分な説明を行いたい．

4 結論に代えて: 言語科学者が FN に期待しているもの

以上，FN の有効性に関して，主に工学的な観点から論じてきた．だが，第一筆者の背景は言語学/認知科学であり，関心の中心は必ずしも工学的なものではない．以下，彼が FN に強い期待をもっている理由を説明する．

意味タグはコーパス利用者の関心を反映したものでなければならないことは，最初に述べた．だが，これが日本語のコーパスの従来の構築法の延長線状に起こるとは考えにくい．京大コーパスであれ何であれ，言語学者が積極的に構築に関わってこなかったという理由もあって，それらの言語学的，認知科学的有用性は限られている．

言語学者はコーパス利用に関して，今までは「工学者の作った便利なものを使わせてもらう」という受身な発想をすることが多かった．だが，**これからは「本当に自分の必要にあったコーパスを自分でデザインし，それを工学者に注文する」という能動的な姿勢を取るべき**であり，そのような形で工学者と積極的に係わってゆくべきだと第一著者は考える．

このような共同作業を通じて言語学が受ける恩恵は絶大なものである．**現時点での言語学は，実験生物学成立以前の生物学のような状態にある**．言語学者の一部には一部の先導者の意見に躍らされて，言語学を物理学になぞらえる人々がいるが [7]，これは明らかに言語学自体にとって好ましい結果を生んでいない．いわゆる「チョムスキー革命」以来，言

語学者は体系的にデータを収集し，それを理論的バイアスを回避しながら記述するという自然科学的に基本的な研究態度を取るのを止めてしまった．その結果，**言語学者はすっかり怠惰になり，言語データを真剣に見なくなり，自分の理論に都合のいい例を作例し，気に入った現象を恣意的に「説明」している**．現在，データ収集の方法は行き当たりばったりで，ご都合主義的であり，完全に非科学的である．そのような劣悪な記述に基づいて（例えば UG に関する）「深遠」な説明を提案するのに言語学者は忙しい．これが現在の「科学的」言語学の実態である．

だからと言って，第一著者は「伝統的」言語学にありがちな，見通しのない，瑣末主義的な現象の記述に回帰すれば良いと主張しているわけではない．言語の記述が言語資源と呼べるためには，まず，それが効果的に (i) 再利用可能であり，(ii) 共同利用可能であることが必要である．効果的に再利用可能であるためには，(iii) 記述のフォーマットが定まっていて，利用者に解釈のために最低限の前提知識しか要求しないことが必要である．更に言えば，(iv) 記述が電子化され，(v) データベース化されていて，(vi) オンラインで利用可能であることが望ましい．

言語記述という問題において，言語学がこれまで分野を越える共有資源の構築になした貢献は実質的に無に等しい．特に意味記述の分野でこの傾向は顕著であり，それが認知科学的には意味の実証的理論の立ち後れ，工学的には意味処理の立ち後れに結果していると思われる．FN は，このような事情に歯止めをかける枠組みとして有望である．それは，言語学が過去数十年間の怠惰から失った関連研究分野との実りある連携関係を取り戻すきっかけを与えるかも知れない．

参考文献

[1] Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*. May 2001.

[2] Fellbaum, Christiane, *Ed*. 1987. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[3] Fillmore, Charles J. 1968. The case for case. In *Universals in Linguistic Theory*, pp. 1-88. Ed. W. Bach and R. T. Harms. New York, Holt, Rinehart & Winston.

[4] Fillmore, Charles J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pp. 111–137. Ed. Linguistic Society of Korea. Seoul, Hanshin Publishing.

[5] Fillmore, Charles J., C. Wooters, and C. F. Baker. 2001. Buildinig a large lexical databank which provides deep semantics. In *Proceedings of the 15th Pacific Asia Conference on Language Information and Computation*.

[6] Fontenelle, Thierry, *Ed*. 2003. *International Journal of Lexicography, 2003 Sep Special Issue: FrameNet and Frame Semantics*.

[7] 福井直樹. 2001. 自然科学としての言語学: 生成文法とは何か. 東京: 大修館.

[8] Gildea, Daniel, and Jurafsky, Daniel. 2002. Automatic labelling of semantic roles. *Computational Linguistics* 28 (3): 245–288.

[9] 河原大輔・黒橋禎夫. 2002. 用言と直前の格要素の組を単位とする格フレームの自動獲得. 自然言語処理. 9 (1).

[10] 黒田　航・野沢　元. 2004. 比喩理解におけるフレーム的知識の重要性: FrameNet との接点. [http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/papers/metaphor-and-frames.pdf].

[11] Lassila, Ora, *et al*.. 1999. *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Recommendation ⟨http://www.w3.org/TR/REC-rdf-syntax⟩

[12] 溝口理一朗. 1999. オントロジー研究の基礎と応用. 人工知能学会誌 14 (6). 45–56 [977-988]

[13] Wilson, Andew and Thomas, Jenny. 1997. Semantic annotation. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Ed. R. Garside, G. Leeach, and A. McEnery. London: Longman.