意味分類体系 'IPA *STaX*' の実装とその応用

緒方典裕
筑波大学大学院文芸・言語研究科
ogata@stc.ipa.go.jp

橋本三奈子
情報処理振興事業協会 (IPA) 技術センター
hasimoto@stc.ipa.go.jp

1 序

IPA技術センターでは、「計算機用日本語辞書IPAL」のサブプロジェクトとして自然言語の意味理論および情報科学の理論にもとづいた意味分類体系 IPA *STaX* (Semantic TaXonomy)、およびその WWW 上での実装 IPA *STaX* for WWW、そしてその応用である *DiSTaX*、*ASTaX* を実装した。

IPA *STaX* は、構造を持った意味カテゴリーの集合で、その構造はリンク・エンジンによって動的に構成され, 辞書の変更に対して頑健に対応できる。意味カテゴリーとは、プログラミング言語のデータ型 (**data type**) に相当する概念で、全ての自然言語表現の意味カテゴリーは、名詞の持つ基本的な意味カテゴリーと constructer から構成される。また、この基本的な意味カテゴリーの判定は、生物学的な分類や恣意的な分類ではなく、名詞と述語との意味的な共起関係を基礎的なデータとして構成される。この意味的な共起関係は、**カテゴリー誤謬** (Ryle (1965))、**カテゴリー一致** (緒方・橋本 (1997)) という概念でとらえられる。

本稿では、カテゴリー体系の記述法および構成法を述べ、それを WWW 上で実装した IPA *STaX* の仕様およびその応用である *DiSTaX*、*ASTaX* について述べる。

2 カテゴリー一致・誤謬に基づいた意味分類構成法 IPA *STaX*

IPA *STaX* の基本的単位である**カテゴリー** (*category*) とは述語の項・名辞・その他表現一般に割り当てられている**意味の種類**もしくは**型**で、プログラミング言語の**データ型** (*data type*) に対応する概念である。[1]全ての自然言語表現の意味カテゴリーは、次のように名詞の持つ基本的な意味カテゴリー *BaseCat* から構成される。ただし、*BaseCat* は対象言語 L の部分集合である。また、カテゴリーのメタ概念を kind という。(cf. Barendrecht (1992))

$$c \in BaseCat$$

$$\alpha \in L - BaseCat$$

$$\delta \in \{ \text{その, この, あの, ある} \}$$

$$\pi \in \{ \text{が, を, に, から, へ, より} \}$$

$$\xi \in \{ \text{種, タイプ, ロール, 段階, 状態, 関係} \}$$

$$\mathcal{C} ::= []|@|t|c|c\pi|c\xi|c\xi\pi|\mathcal{C}_1 \rightarrow \mathcal{C}_2$$

δは変数を構成し、πは格助詞付きの名詞句を構成する。[] は kind をあらわすコンスタント、@ は *BaseCat* を表すコンスタント t は文をあらわすコンスタント、$\rightarrow$は関数カテゴリーを構成する。IPA STaX の言明を表現したものを式と呼び、次のように定義する。

$$\varphi \in \Phi$$

[1] category という用語の歴史的概観については緒方 (1997) を参照のこと。

$$\varphi ::= \alpha : \mathcal{C} | \lambda\delta\mathcal{C} : \mathcal{C}.\varphi | \alpha_1\alpha_2 : \mathcal{C} | \mathcal{C}_1 \sqsubseteq \mathcal{C}_1$$

例 1 「ポチ」「歩く」は次のように表現される。

$$\text{ポチ} : \text{犬}$$

$$\text{歩く} : \text{場所へ} \rightarrow \text{場所から} \rightarrow \text{動物個体が} \rightarrow t$$

□

注意 1 犬、会社など *BaseCat* に属するカテゴリーはほとんどが個体レベルのカテゴリーである。種レベルのカテゴリーは上の規則にあるように**犬種**、**会社種**のように、*BaseCat* に**-種**が付加したものである。同様に、属性を表すものは**-タイプ**、役割・機能を表すものは**-ロール**、ある時間に限定された属性は**-段階**、ある状態は**-状態**、関係は**-関係**を付加する。これらを「**派生カテゴリー (derived categories)**」と呼ぶことにする。□

例 2 「犬」の派生カテゴリー

属する名詞	派生カテゴリー
柴犬、土佐犬	犬種
野良犬、白犬	犬タイプ
子犬	犬段階
軍用犬、盲導犬	犬ロール
狂犬、痩せ犬	犬状態
親犬	犬関係

□

派生カテゴリーは繋辞が付加して名詞述語・形容動詞述語を構成する。

式のある集合を文脈という。ある文脈$\Gamma \subseteq \Phi$で対象言語 L の表現αがカテゴリー$\mathcal{C}$に属するという判断 (judgement) を '$\Gamma \vdash \alpha : \mathcal{C}$' として表し、次のように定義する。

$$\emptyset \vdash @ : [] \quad \emptyset \vdash t : [] \quad \emptyset \vdash c : @$$

$$\frac{\varphi \in \Gamma}{\Gamma \vdash \varphi} \qquad \frac{\Gamma \vdash \alpha : @}{\Gamma, \delta\alpha : \alpha \vdash \delta\alpha : \alpha}$$

$$\frac{\Gamma \vdash \varphi \; \Gamma \vdash \alpha : @}{\Gamma, \delta\alpha : \alpha \vdash \varphi} \qquad \frac{\Gamma \vdash \alpha_1 : s \; \Gamma \vdash \alpha_2 : s}{\Gamma \vdash \alpha_1 \rightarrow \alpha_2 : s}$$

where $s \in \{[], @\}$,

$$\frac{\Gamma \vdash \alpha_1 : \mathcal{C}_1 \rightarrow \mathcal{C}_2 \; \Gamma \vdash \alpha_2 : \mathcal{C}_1}{\Gamma \vdash \alpha_2\alpha_1 : \mathcal{C}_2}$$

$$\frac{\Gamma, \delta\alpha_1 : \mathcal{C}_1 \vdash \alpha_2 : \mathcal{C}_2 \; \Gamma \vdash \mathcal{C}_1 \rightarrow \mathcal{C}_2 : s}{\Gamma \vdash \lambda\delta\alpha_1 : \mathcal{C}_1.\alpha_2 : \mathcal{C}_1 \rightarrow \mathcal{C}_2}$$

$$\frac{\Gamma \vdash \alpha : \mathcal{C}_1 \; \Gamma \vdash \mathcal{C}_1 \sqsubseteq \mathcal{C}_2}{\Gamma \vdash \alpha\pi : \mathcal{C}\pi}$$

$$\frac{\Gamma \vdash \alpha : \mathcal{C} \; \Gamma \vdash \mathcal{C} : @}{\Gamma \vdash \alpha\pi : \mathcal{C}\pi}$$

また、この基本的な意味カテゴリーの判定は、生物学的な分類や恣意的な分類ではなく、名詞と述語との意味的な共起関係を基礎的なデータとして構成される。この意味的な共起関係は、**カテゴリー誤謬** (Ryle (1965))、**カテゴリー一致** (緒方・橋本 (1997)) という概念でとらえられる。

(1) ＊この素数は赤い。

上の文は「文である」ことは理解できるが，何を意味しているかは理解できない。これを**カテゴリー誤謬 (category mistake)** という。

また、インド・ヨーロッパ系の言語等では、主語と述語、修飾語と被修飾語、先行詞と照応詞の間で性・数・人称などの文法的な一致が強制される。一方、日本語ではこのような文法的一致はないが、意味的な一致が強制される。

(2) a. The number of planets is necessarily *nine*.

b. 惑星の数（かず）は必然的に＊九/九つ/九個である。

このような意味的な一致を**カテゴリー一致 (categorial agreement)** と呼ぶことにする。[2]

3 IPA *STaX* for WWW の仕様

IPA *STaX* for WWW は、次のような、カテゴリー一致を反映する文法的項目の集合からなるカテゴリーの集合である。

- 属する名詞
- カテゴリー固有述語 (proper predicate): カテゴリーだけを項に取るような格助詞付き述語[3]

[2] 詳しくは、緒方・橋本 (1997), 緒方 (1997) 参照。

[3] 固有述語の概念は, Carlson (1977) の *kind-level predicate*、*stage-level predicate*、などのカテゴリー特有の述

- カテゴリー一致述語 (agreeable predicate): そのカテゴリーを項に取る格助詞付き述語
- カテゴリー形態素 (category morpheme): そのカテゴリーを表す形態素
- 助数詞 (classifier)
- リンカー (linker): カテゴリー間のリンク構成に使われるカテゴリー一致述語、upは上位カテゴリー、downは下位カテゴリー、upkindは上位種カテゴリー、downkindは下位種カテゴリー、uppartは上位部分カテゴリー、downpartは下位部分カテゴリーのリンクに使われる。

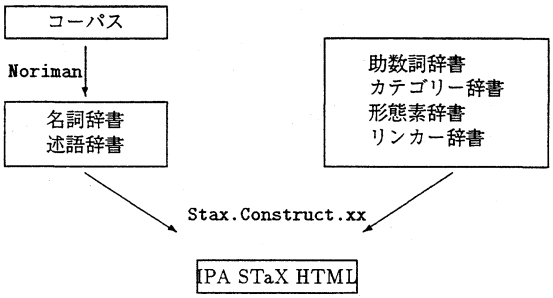
例 3

カテゴリー	会社
属する名詞	東宝映画株式会社 紀伊国屋書店 ...
カテゴリー固有述語	_に入社する _を退社する _に入社する _に来社する _に在社している _へ出社する _は会社タイプである
カテゴリー一致述語	_がつぶれる _が発足する _が倒産する _を創設する _を設立する _を辞める _に勤める _に勤務している _に在職している _へ出勤する
カテゴリー形態素	-社
助数詞	-社
リンカー	_に勤める:up _に所属している:up _に入社する:down

語という概念と共通している。しかし，固有述語は，「猫」それだけに使われる述語「黒猫だ，白猫だ，どら猫だ，野良猫だ」などの述語のようなものも含まれ，さらに一般的な，あるカテゴリーに固有に用いられる述語という広い概念である。

4 IPA STaX for WWW = IPA *STaX* + リンク・エンジン

IPA STaX for WWWは、*STaX*をhtmlおよびcgiプログラミングによってWWW上に実装したものである。それは次の図のように、IPA日本語辞書グループにおいて利用可能なコーパス、およびいくつかの国語辞典から、専用のプログラム (Noriman、Stax.Consturct.xx) を用いて所属する名詞・カテゴリー一致述語・カテゴリー形態素・助数詞を抽出し、各カテゴリーを構成した。



このときに問題になるのが、分類体系の更新を繰り返すとカテゴリー間のリンクの管理が煩雑になるという点である。これを解決するために、リンカーからカテゴリー間のリンクの構成をcgiプログラミングによって実現した。この結果、カテゴリー間のリンクは、共有されたカテゴリー一致述語を検索することにより、ダイナミックに構成される。また、リンカーを複数指定することにより、木構造的な分類体系ではなく、ネットワーク型の分類体系を構成することが可能となった。

例えば、上位カテゴリーの決定は次のような定義を手続き化したものである。

(3) カテゴリー Cのカテゴリー固有述語 p が、カテゴリー Dのカテゴリー一致述語ならば、$D < C$(Cは D の**上位カテゴリー** (*super-category*)，もしくは、Dは Cの**下位カテゴリー** (*sub-category*) という)。

5 *DiSTaX*: Disambiguator exploiting *STaX*

(4) a. マイクロソフトを購入した。

b. マイクロソフトを買収した。

上の文の「マイクロソフト」は, それぞれ製品、会社というカテゴリーをもち、曖昧であるが、次のようにカテゴリー形態素を付加すると曖昧性がなくなる。

(5) a. マイクロソフト製品を購入した。

b. マイクロソフト社を買収した。

*DiSTaX*は(4)のようなカテゴリー的に曖昧な文を入力すると、IPA STaX for WWW の情報からカテゴリーを決定し、そしてカテゴリー形態素を付加した表現(5)を返す。

6 *ASTaX*: Associator exploiting *STaX*

*ASTaX*はIPA STaX for WWW の述語辞書のデータから、固有述語において共起するカテゴリーという観点で連想されるカテゴリーを検索するプログラムである。例えば、ユーザーが「学校」を入力すると、システムはその固有述語である「で教える」を検索し、そこで「が」格に共起する「教師」を検索してユーザーに返す。そのときの述語辞書は次のような形態をとる。

(6) 教える:教師#が&学校[p]#で

ただし、X[p]#$Case$ は Xに関する固有述語がその述語に $Case$ を付加したものであるということを表す。

7 まとめ

以上のように、カテゴリー一致・誤謬という観点からの意味分類体系のIPA *STaX*の構成の基本的アイデアとカテゴリー間の関係をダイナミックに構成する、WWW 上に実装されたIPA STaX for WWW、そしてその応用システムである*DiSTaX*、*ASTaX*についての概観を述べた。現在、IPA STaX for WWW はバージョン2.4で、収録名詞数約4000、収録述語数約1000、収録カテゴリー数約350であり、さらに更新中である。

また、IPA STaXのドメイン限定版の構成法とそれをWWW上でリンクさせる'*Open STaX*'を構想中である。

参考文献

BARENDRECHT, H. P. 1992. ' Lambda Calculi with Types', in S. Abramsky et al eds. *Handbook of Loic in Computer Science*, iClarendon Press: Oxford, pp. 118–309.

CARLSON, G. N. 1977. *Reference to Kinds in English.* PhD thesis, University of Massachusetts. published by Garland Publishing, New York, 1980.

CRUSE, D. A. 1986. *Lexical Semantics.* Cambridge: Cambridge University Press.

DRANGE, T. 1966. *Type Crossings.* The Hague: Mouton and Co.

LAPPIN, S. 1981. *Sorts, Ontology, and Metaphor: The Semantics of Sort Structure.* Berlin: Walter de Gruyter.

緒方典裕 1997(forthcomming). 「IPA STaX の仕様とその応用システム」『ソフトウェア文書のための日本語処理の研究 - 13』東京: 情報処理振興事業協会.

緒方典裕・橋本三奈子 1996. 「意味分類の言語学的構成法とそのWWW 上のシソーラス構築」『情報処理学会研究報告』97-NL-117-7, pp.45–50. 東京: 情報処理学会.

PUSTEJOVSKY, J. 1995. *Generative Lexicon.* Cambridge: The MIT Press.

RYLE, G. 1965. Categories. In A. Flew, ed., *Logic and Language*, New York: Doubleday, pp. 281–298.

VENDLER, Z. 1967. *Linguistics in Philosophy.* New York: Cornell University Press.