自然言語処理 118－4
（1997. 3. 21）

市販電子化辞書からの自然言語の意味抽出

小田 誠雄　西村 靖司　小田 まり子†　横田 将生‡

福岡工業短期大学　† 久留米工業大学　‡ 福岡工業大学

心像意味論 (MIDST: Mental-Image Directed Semantic Theory) に基づいて、人間の言語の学習過程をシミュレーションする自然言語概念学習システムの作成を試みている。今回、我々は簡単な自然言語を用いて語概念の説明を行い、学習を行うシステムについて考察した。

国語辞典に記載された語義文は我々のシステムへの典型的な入力例であり、電子化された辞書の語義文を使えば一時に大量の知識の入力が可能となる。

だが、一般的な国語辞典には数多くの語とその語義文が掲載されており、全てを一度に対象とするのは難しい。心像意味論では意味記述の中心は事象概念である。そこで、我々は、まず国語辞典の動詞の語義文を取り出し、文法構造、意味の記述方法の調査を行った。

Extraction of Natural Language Concepts from a Published Electrical Dictionary

Seio Oda　Yasushi Nishimura　Mariko Oda†　Masao Yokota‡

Fukuoka Juniour College of Technology
† Kurume Institute of Technology
‡ Fukuoka Institute of Technology

Basing on MIDST(Mental-Image Directed Semantic Theory), the authors have been constructing a learning system of natural language concepts, which simulates human knowledge acquisition. In this paper, we study a learning method to utilize the meaning descriptions of word.

The descriptions in Japanese dictionaries are assumed to be the typical inputs for our system. In order to a large scale of knowledge acquisition, electrical dictionaries are quite helpful.

The authors have employed an electrical dictionary for a large scale of knowledge acquisition. In accordance with MIDST and for the simplicity, only the description of verbs are taken into consideration, though a dictionary in general contains various kinds of words and their meaning descriptions.

The descriptions of verbs contained in the electrical dictionary have been analyzed syntactically and semantically.

1 まえがき

我々は、人間の知識の獲得過程、特に幼児期における自然言語概念の獲得は、言語的知識だけでなく視覚や聴覚からの非言語的知識が重要である、という考えに基づき、幼児期の言語の学習過程をシミュレーションする自然言語概念学習システムの作成を試みてきた [1]-[9]。

幼児期の言語学習は、まず物の名前からはじまり、事物の名前の学習が初期に行われる。が、一旦述語を中心とした言語構造を理解すると知識獲得の速度が急に早まる事が知られている。そこで我々は、その期間、つまり自然言語を使った自然言語概念の獲得過程について考えを進める事とした。

自然言語を使って単語の説明を行う実例には、国語辞典に記載されている説明文 (語義文) がある。語義文は我々の考えるシステムへの典型的な入力例であり、電子化された辞書を使えば一時に大量の知識の入力が可能となる。近年マルチメディアの普及によって電子化された国語辞典の入手が容易になったことから、国語辞典からの知識獲得を目標とした。

概念学習のためには、まず語義文が、どのようにして意味を記述しているか知る必要がある。そのため、まず語義文の構造の調査を行い、その結果を基に、我々のシステムの中で、知識を獲得する方法を考えねばならない。対象とする国語辞典には沢山の語とその語義文が記載されているので、それらを一度に調査するのは難しい。我々のシステムは、心像意味論 (MIDST: Mental-Image Directed Semantic Theory) に基づいて作成されており、心像意味論では意味記述の中心は事象概念である。そのため今回は国語辞典の中から動詞の語義文を取り出し、その意味記述を調査する事にした。対象の国語辞典は岩波国語辞典第五版である。

電子化辞書からの知識獲得の研究としては、鶴丸らによる語義文から定義語を抽出しシソーラスを作成する試みがある [10] が、我々の研究は語の持つ概念そのものを記述しようとする点で特徴がある。

2 国語辞典の構造

岩波国語辞典第五版には約6万2千語の語彙が収録されている。採録されている語の例を以下に示す。

あい-する【愛する】〖サ変他〗
それに対し愛をそそぐ。(ア) かわいがり、いつくしむ。「子を—」。心から大切に思う。「国を—」(イ) 異性を恋い慕う。(ウ) 物事を強く好む。「酒を—」

語は見出しと説明の2つの部分からなる。見出しの内容は、以下の3項目である。

1. 読み。平仮名 (外来語はカタカナ) で書かれ、活用語の場合は語幹と語尾の間に "-" が入る。
2. 表記。"【】" の中に示され、漢字の読みの種類や送り仮名の曖昧性などにより、種々の記号が付されている。
3. 品詞や文法上の性質。"〔〕" の中に記される。

一語に複数の意味がある場合、説明に次のような区分記号が付される。

1. (1),(2),(3),···: 最も普通の分類
2. (ア),(イ),(ウ),···: 上の内容を細分する時
3. [一],[二],[三], ···: (1),(2),(3),··· よりも大きな分類が必要な時

また説明の中には、語義の説明文 (語義文) 以外に "《》" 内には文法的な説明、"()" 内にはルビや意味の分かりにくいものの解釈、"「」" 内に用例など、様々な補捉説明が入る他、文字の大きさの指定、書体の指定など表示を行う際に必要な情報も一緒に入っている。

3 意味の記述方法の調査

国語辞典には、語義文以外にも色々な付加情報がそえられているので、まず語義文だけを抜き出す前処理が必要である。

次に形態素解析を行い、品詞情報を付加した単語列にする。

その後、通常の自然言語解析処理では構文解析、意味解析と続くが、今回はこれらの解析は人手で行う事とした。ただし、分析対象の単語列は大量になるので単語列を含まれる品詞によって機械的に分類し分析の助けになるようにした。

3.1 前処理

辞書から、動詞の見出し語と語義文を取り出す。見出し語には多義があるので、分類記号に基づいて3桁の番号を振る。語義文には、ルビ、用例などの付加的情報がついているが、今回はそれらを全て削除した。取り出した語の一部を以下に示す。

愛する (000) それに対し愛をそそぐ
愛する (001) かわいがり、いつくしむ
愛する (002) 異性を恋い慕う
愛する (003) 物事を強く好む
相対する (000) 互いに向かい合う
相次ぐ (000) 次から次へと続く
相手取る (000) 争いの相手とする
合う (100) 物・事が一つになり、離れていない、また矛盾がない
合う (110) 寄り集まって一つになる
合う (120) 互いに同じ動作をする

岩波国語辞典第五版からは、動詞の語義文として 8,228 文を取り出す事ができた。

3.2 形態素解析処理

3.2.1 解析エンジンと単語辞書

前処理によって取り出された語義文は、形態素解析処理により単語列に変換する。今回、我々は形態素解析処理に、入力済単語列から状態状態を計算し、その状態から次に入力される単語により、新たな状態に遷移する、一種のオートマトンと考えられる解析エンジンを作成して利用した。入力の単語列には曖昧性があるので、単語入力前の状態と入力単語の組合せ毎にコストを設け、最良優先探索技法を用いて、最低のコストで最終状態に遷移する単語列を見付けるようになっている。

表 1: 形態素解析の成功割合

項目	個数	割合
語義文	267	100%
解析失敗	20	7.5%
得られた単語列	349	131%
正しい単語列	217	81%

また、単語辞書も語義文を取り出したのと同じ岩波国語辞典から、見出し語と品詞情報を取り出して作成した。単語は表記に漢字を使うか否か、送り仮名の曖昧性などがあるので、作成した単語辞書に登録した項目 (相異なる表記) は約 12 万項目となった。

3.2.2 解析結果

動詞の語義文 8,228 文すべての形態素解析を行ったが、入力文の中には 50 文字以上の長文も含まれるため、1 文あたり 5 分間の制限時間を設けた。またコストで曖昧性の解決できなかった単語列は、曖昧のまま出力するようにした。

今回使用した解析エンジンは開発途中で完全なものではない。また近視的なコストだけで優先順位を決定しても、正しい解析結果が得られるとは限らない。そのため、8,228 文の入力の内、 7,451 文から 10,607 個の単語列が得られた。

解析に失敗した 777 文中, 728 文が制限時間により打ち切られた。残りの 49 文のほとんどは単語辞書の不備により失敗した。

次に得られた解析結果の妥当性を調べるため、語義文の先頭から 267 文の解析結果を調査した。その結果を表 1に示す。

解析が失敗したものの原因について以下に示す。

- 単語辞書の不備、国語辞書から単語辞書を生成した時に予想しなかった表記が現われているもの、例えば「見付ける」に対し「見つける」のように仮名表記と漢字表記が混ざって現われるものと、国語辞典に記載されていない単語の場合がある。後者の例に

は、接尾語の「かかる」、動詞連用形が名詞として使われる「つぐない」、固有名詞「京都」などがある。

- 単語の曖昧性が解消できないもの。サ変動詞の「する」と五段活用の「する (摺る)」の曖昧性は意味解析をしないと解消できない。名詞の「よる (夜)」と動詞の「(なわを) よる」も同様である。

- 国語辞典の他の項目を参照しているもの。「遊ばす (010) 遊ぶ (1) の状態にさせる」は、「遊ぶ」という語への参照であり、動詞として解析できない。

3.3 形態素解析結果の粗分類

日本語文では述語は文末にあるので、語義文の意味は文末の語のパターンに依存すると思われる。そこで文末の品詞パターンを調べてみた。パターン毎の個数分布を表2に示す。

この表から、まず文末に動詞が来るものが圧倒的に多い事が分かる。そして、その動詞は見出し語の上位語、もしくは同意語になっている事が容易に予想される。

次に多いのは2個の事象概念が連続して現われるものである。これらは、それぞれの事象概念の連言になっている場合が多いと思われる。

次に全体のパターンの内、個数の多いもの10パターンを表3に示す。この中で目立つのは、「名詞＋格助詞」の並びが述語の前に来ているものである。そこで「名詞＋格助詞＋動詞」型と「名詞＋格助詞＋詞＋格助詞＋動詞」型について使われている格助詞の種類を調査した。その結果を表4に示す。この中で「の」および「と」は「名詞＋格助詞＋動詞」型には現われなかった。

4 意味学習の手順

前章の調査結果より、

1. ほとんどの場合、上位語または同意語が文末に来る。

表 4: 格助詞の種類別数

格助詞	個数	格助詞	個数
を	388	に	317
が	158	の	77
で	49	から	25
へ	24		

2. 上位語の持つ事象構成要素が、「名詞＋格助詞」という形の文節によって詳細化もしくは特定される。

3. 事象概念が文末に複数現われる場合、それらの連言になっている場合が多い。

と言う事が分かる。従って、

1. どのように事象構成要素を詳細化、もしくは特定するか。

2. 事象概念間の接続方法。

の2つが分かれば、意味学習は可能である。

4.1 事象構成要素の特定

表4の7個の格助詞について、事象構成要素をどのように限定しているかを以下に示す。

が 対象を指示する。ただし、次に「で格」が続く場合は続く事物を修飾する。

から 動作の起点を指示する (上がる (115) ふろから出る)。ただし、強意で用いられる場合 (祈る (020) 心から望む) もある。

に 程度、動作の終点、対象など様々で意味処理を必要とする。ただし、「明ける (223) 留守にする」の場合のようにサ変動詞の「する」や「なる」が続く場合は、事象の終点の状態を指示している。

を 動作・作用の対象を表す (仰ぐ (011) 上方を見る)。サ変動詞「する」が続く場合は、事象概念そのものを表している (息衝く (030) 息をする)。

表 2: 語義文の文末のパターン

パターン	個数	例
動詞+動詞	1,176(833)	愛する (001)　かわいがり、いつくしむ
動詞+助動詞+動詞	559(—)	震わせる (000)　震える-ように-する
動詞+接続 (助) 詞+動詞	1,153(—)	立てる (121)　まっすぐに-すえる-または-差す
形容詞+動詞	604(—)	打つ (123)　強く-刺激する
形容動詞+動詞	399(—)	叩き起す (020)　眠っている者を-むりに-起こす
動詞+助動詞	539(372)	飽かす (010)　あき-させる
(その他)+動詞	5,364(—)	赤らむ (000)　赤みを-帯びる
その他動詞類	33(24)	弾む (120)　調子-づく
形容詞	22(20)	切れる (201)　切れ味が-鋭い
形容動詞	10(5)	及ぶ (030)　…ことが-必要だ
名詞+助動詞	11(6)	窮まる (002)　…が-最上-だ
名詞	427(27)	拝する (020)　受けることをへりくだって言う-語
		呉れる (021)　好意をもって…する-意
		過ぎる (040)　断定を強める言い-方
		上がる (320)　敬語としての-用法
副詞	10(5)	思いきる (022)　思う存分
その他	300(0)	—
合計	10,607(—)	—

() 内は、正しい解析結果の個数、— は未調査

表 3: 語義文の品詞パターン (一部)

パターン	個数	例
名詞+格助詞+動詞	832(704)	赤らむ (000)　赤みを-帯びる
動詞	810(810)	合う (132)　一致する
動詞+動詞	235(207)	愛する (001)　かわいがり、いつくしむ
名+格助+名+格助+動	230(176)	泡立つ (000)　表面に泡が立つ
動詞+助動詞+動詞	230(137)	余す (000)　余るようにする
形容詞+動詞	173(140)	青む (000)　青くなる
動+接続助+動	169(150)	上げる (420)　取り出して-言う
名+格助+動+接続助+動	146(—)	上げる (600)　潮が満ちて来る
名+格助+動+動	128(109)	愛する (002)　異性を恋い慕う
形容動詞+動詞	128(—)	明かす (010)　あきらかにする

() 内は、正しい解析結果の個数、— は未調査

で 物概念の場合は道具、事概念の場合は手段、方法、場所概念の場合は事象が生じる場所を表す。例外的に適応される分野が示される場合もある (割切る (020) 数学で除算する)。

へ 動作の終点 (込める (000) 中へ入れる) を表す。

の 目的の事象には掛からず、次に続く事物を修飾している (賜る (010) 目上の人からもらう)。

ただし、語義文の中には成句を含んでいるものもあり (骨を折る、しびれが切れる等)、その場合は成句単位で意味を考える必要がある。

4.2 事象概念の接続

文末に複数の事象概念が書かれる場合の接続の種類を示す。以下で L_1, L_2 は事象 E_1, E_2 の軌跡式、$\sqcap$ は同時的連言、$\bullet$ は経時的連言を表す。

・事象 E_1 と E_2 が同時に起きるもの (かわいがり、いつくしむ)

$$L = L_1 \sqcap L_2$$

・事象 E_1 と E_2 が連続して起きるもの (なぐり倒す)

$$L = L_1 \bullet L_2$$

・事象 E_1 の開始 (荒れ始める)
・事象 E_1 の終了 (習い終る)
・事象 E_1 を強めて言う (笑いころげる)
・事象 E_1 が E_2 の程度を表す (固く閉じる)
・接続詞「また」,「または」が入る場合は、連言ではなく複数の事象どちらも上位語、同意語であることを示す。

5 まとめ

岩波国語辞典に記載された動詞の語義文における意味の記述方法を調査し、格助詞による事象構成要素の限定と、複数事象概念間の関係を記述する事によって、語義文から自然言語概念の学習が可能である事を示した。

その過程で国語辞書と言えども辞書中に記載されていない語を使って説明を加えたり、単なる言い替えで済ませていたりする事が分かった。

とはいえ、現在使用している形態素解析プログラムは問題も多く、また、意味処理も行っていない。これらの処理を改良・追加して、機械による意味学習シミュレーションを行うことが今後の課題である。

参考文献

[1] 横田 将生: 心像に基づく自然言語意味論の提案, 福工大言語情報工学研究所彙報,Vol.1,PP.1–12(1990)

[2] 横田 将生: 心像に基づく自然言語談話理解システムの試作, 言語情報工学研究所彙報, Vol.2, PP.1–12(1991)

[3] 横田 将生, 白石 正人：自然言語概念獲得に関する考察, 情報処理学会九州支部研究会報告,Vol.3-7,PP.53–62(1991)

[4] Masao Yokota, Masato Shiraishi, Koichi Ryu and Seio Oda: Mental-image Directed Semantic Theory and its Application to Natural Language Understanding Systems, Natural Languege Understanding Pacific Rim Symposium (NLPRS'91), PP.280–287 (1991)

[5] Seio Oda, Yasushi Nishimura, Masato Shiraishi, Masao Yokota: An Experimental System for Learning "Color" Concepts based on Mental-image Directed Semantic Theory, Natural Languege Understanding Pacific Rim Symposium (NLPRS'93), PP.264–271 (1993)

[6] 小田 誠雄, 横田 将生: 概念学習システムのための色に関する語彙の概念分析, 福工大言語情報工学研究所彙報,Vol.3, PP. 31–36(1992)

[7] 小田 誠雄, 横田 将生: 色に関する自然言語の概念学習システムの試作, 言語情報工学研究所彙報,Vol.4, PP. 25–32(1993)

[8] 小田 誠雄, 白石 正人, 横田 将生:心像意味論に基づく色に関する語彙の概念分析, 電気関係学会九州支部連合大会論文集, 45, PP.787(1992)

[9] 小田 誠雄, 白石 正人, 横田 将生：自然言語概念獲得システムのための色に関する動詞の分析, 電気関係学会九州支部連合大会論文集, 44, PP.567(1991)

[10] 鶴丸弘昭：国語辞典の解析に関する研究, 科研費特定研究「言語情報処理の高度化のための基礎的研究」研究発表資料集 (5),PP.36–46(昭和63年)