北海道大学工学部研究報告
第167号（平成6年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 167 (1994)

自然言語処理のための意味表現形式とデータの蓄積について

佐藤　嘉高　　宮永　喜一　　栃内　香次

（平成5年8月31日受理）

On semantic representation and data storage for natural language processing

Yoshitaka SATO Yoshikazu MIYANAGA Koji TOCHINAI

(Received August 31, 1993)

Abstract

According to the rapid increase of computer power and cost performance, a large number of studies on natural language processing are carried out. However, many problems are not yet settled. For example, integrated language processing systems, such as, machine translation systems or dialogue understanding systems need a large amount of knowledge or rules. Relationships for these data must be described with no contradictions, and it is generally required a great deal of labor to construct a whole set of those knowledge and rules.

In this paper, we propose a method to construct a network structure for a semantic group of sentences combined with syntax trees of them. In this network, local classes of concepts in the sentence group can be constructed from nouns, and relations of concepts are expressed by verbs.

We also propose a procedure for data structure construction and an efficient method for getting data. These methods use mutual dependence or interactions between grammatical categories and syntax rules, and a hierarchical structure of data for three levels—that is, paragraphs, sentences, and words are combined.

1．はじめに

近年，計算機の性能の飛躍的な向上や低価格化による普及にともない，自然言語の処理に関する研究もますます活発に行なわれるようになった。しかしいまだ未解決の領域も数多くある。例えば機械翻訳や対話理解，文章の要約など，自然言語処理を統合的に行なうことを目的とするシステムにおいて，構文あいまいさの解消や省略の補完，照応の処理などを的確に行うためには，一般に膨大な量の知識やルール記述が不可欠であり，また概念間の関係を意味的に無矛盾に，か

電子工学科　電子機器工学講座

つ機能的にも整合のとれる形式で記述する必要がある。一方，そのような解析に必要な単語辞書や文法辞書のデータを人手により蓄積するのには，一般にかなりの労力を必要とする。最近ではCD-ROMなどの形式で電子化された辞書も市販されているが，それらのデータからは作成するシステム上で統合的処理を行なうために必要十分な情報がすべて得られるわけではない。従って程度の差はあるが人手による入力が必要となる。これをできるだけ軽減するために種々の手法が考えられているが，逐次処理によるものが多く，効率が良いとは言い難い。

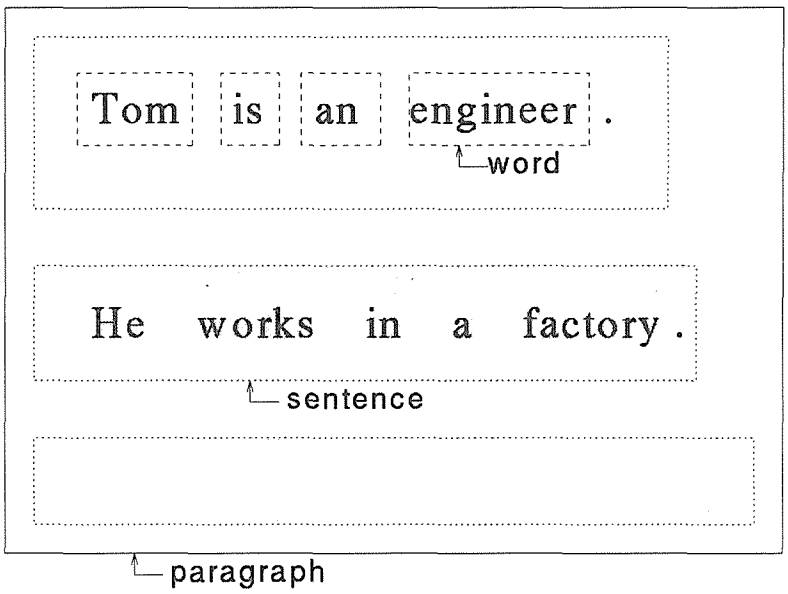
これらの点に関し我々は，パラグラフなど複数の文のまとまりに対する構造を1文の表層構造を表現するデータをネットワーク状に組み合わせることによって表現し，それを文脈解析の手がかりとする方針で研究を行なっている。本稿ではデータの構造と意味表現形式の構築時におけるデータ参照の手法について述べる。さらに，単語の品詞と文法規則との相互依存性を利用し登録単語の品詞推定を構文解析時に同時に行ない，構文ルールの半自動獲得を行なう軽減手法を提案する。

以下，2章で文脈処理のためのデータ構造について，3章では実験システムの構成について，4章ではデータの蓄積と参照の手法について，5章では辞書登録について，6章では実験方法について，それぞれ述べる。

2．文章の表現構造

2.1 文脈処理

機械翻訳システムなどにおいて，文を逐次的に処理する場合，一般に構文はある時点で処理対象となっている一文だけに着目すればよい。ところが，文章が全体として何について述べたものであるか，あるいは段落ごとの主題は何かなどを，入力された文章から把握しようとする場合，



object hierarchy

図1：オブジェクトの階層

代名詞の照応などの解析が必要となり，前後の文を参照する必要がある。

そこで，処理対象の個々の文間のつながりをどのような観点でとらえどのような形式で計算機上に表現するか，が重要となる。本研究では表層構造における単語間の結合関係，すなわち単語の結束性に着目する。結束構造を扱うシステムの構築において，単語，文，文章のそれぞれに対しての異なった処理のスコープが存在するので，それらの分類の明確化が望ましい。本研究では，そのそれぞれをオブジェクト指向言語で記述し，処理のプロトコルのうち共通なものと異なったものとを明確に分離している。この階層構造を，図1に示す。なお，あらかじめ名詞概念の階層構造などの知識を用意し，それを解析時に参照するという方法も考えられるが，現段階では入力された文章が表現している局所的な世界のみを動的に記述することを考えており，用意する知識は最小限にとどめる。

2.2 パラグラフネットワーク

文と文との関係を記述するための方法として，我々は表層で形式的に区分することのでき，かつ意味的にもまとまりのある単位（一般に段落）の大きさに対して一つのデータ構造を割り当てた。この構造はそれぞれの文中に含まれる単語を中心として形成される意味ネットワークの一種であり，オブジェクト指向言語上のオブジェクトで表現する。これを本報告では『パラグラフネットワーク』と呼ぶ。以下で，ネットワークの構造とその生成について述べる。

2.3 ネットワークの構造

ネットワークはオブジェクトで表現されたノードとリンクから構成される。このネットワークは構文木のデータ構造とその一部を共有している。この様子を摸式的に図2に示す。以下にそれぞれの機能と文を構成する要素との関係を示す。

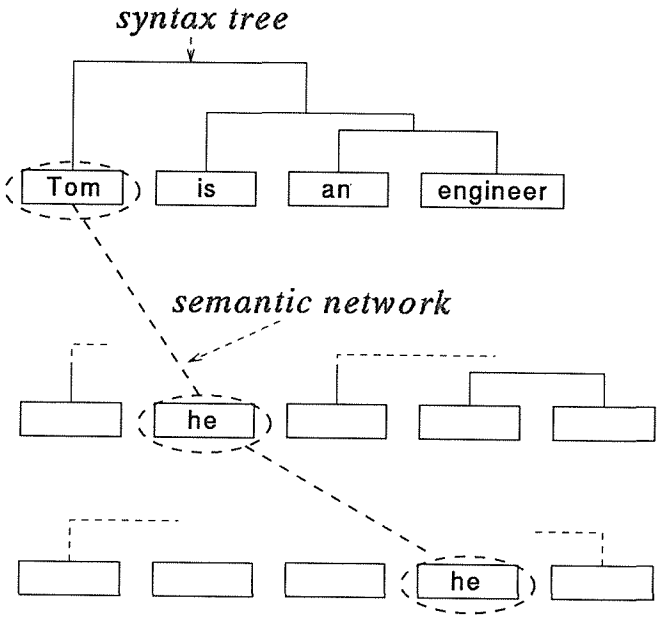


図2：構文木と意味ネットワーク

● ノード
文中の名詞の指示する概念が割り当てられ，その内部状態として単語の文字列と構文木オブジェクトへのポインタを持つ。文を構文的，意味的に構成する要素のうち，静的，受動的なものを表現する。

● リンク
動詞により記述されるノード間の関係を内部状態として保持する。文を構成する要素のうち，動的，能動的なものを表現する。

2.4 ネットワークの生成と参照

ネットワーク構造の形成は解析対象の文の先頭から順に行なう。このプロセスにおいて，文を構成する要素のうち名詞と動詞に着目する。名詞については，それが文中で意味的にどのような概念を表現し，機能を果たしているかを解析する。動詞については，名詞，名詞句の指示する概念間の関係がそれによって記述されていると考える。ネットワーク形成と参照における名詞句の制御手続きとしては，以下のものを考慮する。

● 1文中の名詞句相互の関係
本研究で取り扱う文は句構造規則 S → NP　VP で解析可能なもののみとする。すなわち，文中に必ず名詞句と動詞句が含まれる。
- 主語に含まれる名詞
その文の主題であるとみなす。この名詞は文が他の文からアクセスされる際のキーとなる。
- その他の名詞
主語に対して何らかの関係を持つ。その関係は動詞句で判別する。構文木中でつながりを持つと同時に，意味ネットワークにおけるリンクで主語に接続される。

● 他の文の名詞句との関係
多数考えられるが，本報告では以下のヒューリスティックスに基づくものを用いている。
- ある代名詞の指示対象は直前の文に記述されている可能性が極めて高い。
- ある文において初めて出現した名詞とすでに述べられている名詞とを表層で区別する手がかりとして，冠詞（不定冠詞／定冠詞）に着目する。

入力された文章が表現している局所的な世界を記述し，それを動的に最適化することを考えている。

3．処理システムの構成

以上のような観点に基づき，我々は英文の文脈処理のために実験のためのシステム[1]を作成した。このシステムの構成を，図3に示す。

本システムはワークステーション上でオブジェクト指向言語 Objectworks\Smalltalk により構築されている[3]。単語および文法辞書のセットは，オブジェクトとして実現されている。

この実験システムは，図に示すように機能別にいくつかのモジュールに分けることができる。以下，それぞれについて詳しく述べる。

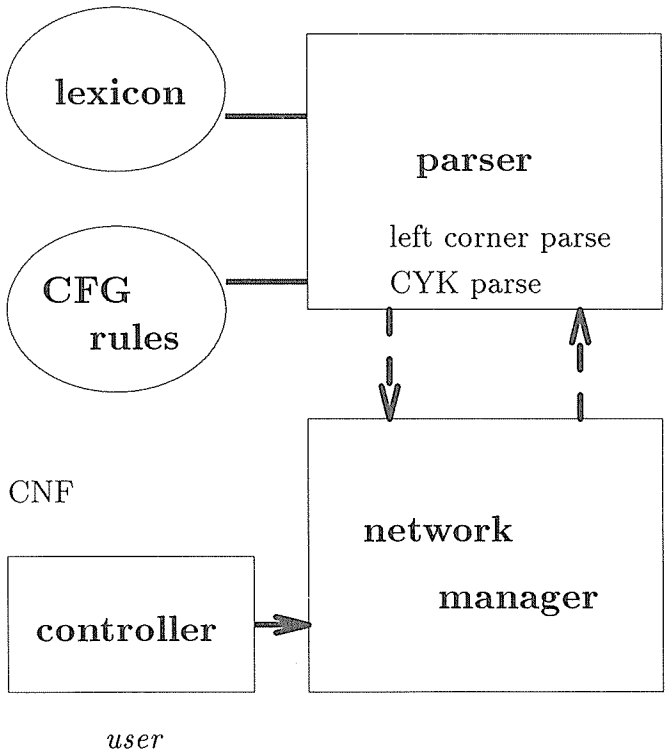


図3　システムの構成

3.1　構文解析部

一般の自然言語処理システムにおいては，文脈自由文法（ＣＦＧ）に基づいたパーサが広く採用されている。その解析アルゴリズムについてもすでに優れたものが多数発表されている。本システムの構文解析部においては，左隅統語解析法[5]に基づくパーサがすでにインプリメントされている。このアルゴリズムはトップダウン予測を行ないながらボトムアップの解析をするものであり，文法の到達可能性をあらかじめ計算しておくことによって単純なボトムアップ法に比べると効率のよい解析が可能である。

3.2　ネットワーク管理部

入力文より生成される概念のネットワーク構造の動的管理を行なう。このモジュールは，構文解析の結果として作成される構文木をもとにして複数の文を連結し，ネットワーク形式の構造を形成するものである。また，ネットワークの状態や動的な変化などをユーザに対して示すとともにユーザからの入力も受け付ける。

3.3　統合環境部

構文解析部とネットワーク管理部とは機能的に独立したモジュールである。これら２つの部分を統合し，またユーザとそれらとのインタフェースとなる部分である。通常はそれらの処理が後述のように逐次的に実行されるが，ユーザがそれぞれを単独で動作させることも可能である。

現時点では，構文解析，ネットワーク生成の各過程はそれぞれ独立したプロセスとして処理されている。したがって構文解析を行わずにネットワークを形成することも可能である。

3.4 単語辞書

本システムにおける単語辞書は単語の文字列とそのカテゴリーとの1対多の対応表である。その詳細に関しては後述する。

3.5 文法辞書

本システムにおける文法辞書は，文脈自由文法の句構造規則のセットである。構文解析の処理をできるだけ簡略なものにするため，データとして扱う規則はチョムスキー標準形，すなわち

$$A \rightarrow B\ \ C,\ A \rightarrow \alpha\ (A, B, C：非終端記号,\ \alpha：終端記号)$$

の形式のもの[5]のみとしている。このような制限は文の構文的複雑度の増加に対する対応を困難なものにする。しかし，本研究の目的は構文解析を完璧に行なうことではなく，また現段階では文の構文構造をどの程度まで解析すればよいのかが明確でない。したがって，データ形式を極力簡単なものにし，あとは必要に応じて構文解析の処理手続きでレベルの高い処理を行なうようにする。

4．データの蓄積と参照

4.1 単語辞書における品詞の分類

単語のカテゴリーには，表層的なものから意味的カテゴリーにいたるまで種々のレベルのものが存在する。ところが，種々の解析に必要なカテゴリーデータが具体的にどのようなものかは，データ獲得の段階では不明である。また，データを登録する過程やその基準などはできるだけ客観的なものであることが望ましい。ここでは，単語を登録する際に必要な情報として，市販の英和辞書と同じ品詞の分類を用いる。ただし，そのうちの名詞，代名詞，動詞，冠詞については，さらに細かく分類する[2]。この分類を表1に示す。

基本カテゴリー名	細分類	実　例
名詞	一般名詞 固有名詞	engineer William
代名詞	人称代名詞 指示代名詞 再帰代名詞	he this itself
動詞	一般動詞 be 動詞	make was
冠詞	定冠詞 不定冠詞	the an
助動詞		will
形容詞		pretty
副詞		always
前置詞		in
接続詞		or
間投詞		oh
句読点		, :

表1：品詞の分類

このように，単語辞書における品詞の分類は階層構造になる。

4.2 処理手法の解釈に基づく分類

前述の分類に基づいて蓄積した単語データをネットワーク構造を生成する際に参照するが，それぞれを元にしてどのような処理をするか，で品詞を以下のように分類する。

1．ノードを生成するもの：名詞，代名詞
2．リンクを生成するもの：動詞
3．名詞を修飾するもの：形容詞，冠詞など
　構文木において名詞，代名詞を修飾するため，間接的にノードの一部となる
4．動詞を修飾するもの：副詞，助動詞など
　構文木において動詞を修飾するため，リンクの一部となる
5．構文的に作用するもの：前置詞，接続詞など
　文全体の句構造の解析の際に参照する

4.3 分類のレベルと相互関係

単語辞書における品詞の分類とは別に処理手法の解釈に基づく分類を行なう理由を以下に述べる。ネットワーク生成や句構造解析のプログラムが単語データを参照する際の，それぞれに適切なデータ分類のレベルがどのようなものかということは，処理内容に依存しているので，データの蓄積の際の階層的な分類とは別の基準に基づき品詞のグルーピングが行なわれた方が，統合的処理が簡潔になると考えられるためである。すなわち，辞書データの品詞の階層とは独立に全てのカテゴリーの中でグループを形成している。

5．辞書データの登録

5.1 辞書構築の手順

ある単語についてそのすべての品詞を一度に登録することは，人間にかなりの負担をしいる作業となる。文法規則についても同様で，構文解析に必要な句構造規則をあらかじめすべてシステムに与えるというのは事実上不可能であり，言語の性質を考慮すると本質的ではない。本手法では，単語の品詞の登録と文法の句構造規則の登録とを独立に扱うのではなく登録の課程において相互に補間させる。このことによってより柔軟なデータの管理ができると考えられる。また，データの蓄積のフェーズとそのデータを用いることにより実際に行なわれる解析のフェーズとを分離せず両者を融合させた形で行なう。

5.2 登録実験

今回実験の対象とした文は大学入試用の問題集[4]である。
単語辞書登録の手順を以下に示す。

1．対象とする文章（文字列）を入力する。
2．文章の始めから単語を1語ずつスキャンし，それが未登録語であればユーザに対しその品詞の入力を求める。
　この際に，ユーザはその単語のすべての品詞を一度に登録する必要はない。また，その単

語が特に多品詞語である場合などにもとの文における品詞を意識する必要がなく，その単語だけをみて直観的に想起した品詞を入力すればよい。

その後，以下のようにして単語，文法辞書の構築を行なう予定である。

- 未登録品詞の推定
  左隅構文解析法におけるトップダウン予測により品詞の推定を行なう。
- 構文規則の半自動獲得
  すでに登録されている規則で解析可能な部分木をユーザに示し，人手により句を構成することによりルールを取得させる。

6．実　　験

前述のような観点に基づきネットワーク構造生成の実験をおこなう。実験の対象とする文は参考文献[4]である。この文献の数量に関するデータを表2に示す。以下に示すように，入力文に対して単語辞書を参照することにより品詞情報を付加し，それがネットワーク構造生成のプログラムを起動する。

文献名	パラグラフ数	総文数	総単語数
英文和訳演習（入門編）	15	194	3305
英文和訳演習（基礎編）	32	211	3597
英文和訳演習（中級編）	24	196	4065
英文和訳演習（上級編）	20	184	4917
総　計	91	785	15584

表2　実験対象文の数量データ

1．入力文を辞書を参照し名詞句と動詞句に大別する
2．名詞句中の中心となる語に着目する
　トピックス・スタックに入れる
3．代名詞の内容を前方に照応する

以上をパラグラフ全体について行う。この様子を以下に示す。

入力文：	Tom	is	an	engineer.
品　詞：	固有名詞	be動詞	不定冠詞	一般名詞
役　割：	ノード生成	リンク生成	名詞を修飾	ノード生成

7．おわりに

7.1　課題

統合的処理を行なう場合，各処理モジュール間の関係があらかじめ明確にわかっているものではなく，実験を繰り返すうちにある程度アドホックに決まることは避けられないと考えられる。したがって，問題点を解決するうえで重要なのは，それぞれのモジュール内部での処理能力をで

きるだけ高めることである。本システムの各モジュールで問題となっている点は，以下の点である。

- 構文解析部

  多品詞語に起因するあいまい性を効率よく処理することが難しい。バックトラックなどのアルゴリズムの機構を手続き的に記述しているため，制御が繁雑である。そのため，現在ボトムアップ並列構文解析法であるCYK法など他のアルゴリズムに基づくパーサについても作成中であり，その検討を行なっている。

- ネットワーク管理部

  意味ネットワークをデータの表現効率がよく，かつ利用しやすい形式で保持するのが難しい。ノードとリンクはデータ型の設計の段階ではそれぞれが機能的に完結したオブジェクトであるのが望ましいが，両者をネットワーク状に組み合わせた時，全体としての情報の表現形式として最適な構造にはならない。この最適性はネットワーク構造を作成する本来の目的が何であるかに依存し，かつ実験をある程度の分量を行なわないと明確にはならない。

7.2　今後の方向

今後はさらに多量のデータに対し実験を行ない，本手法の有効性を確かめる。また，本報告で提案した手法を自然言語処理システムにおける各種の処理を統合化するための中心とすることを考えている。本手法の応用として機械翻訳，CAIなどの対話型システムを考えている。

参考文献

1）佐藤，宮永，栃内：“複数の文に基づくネットワーク形式の構造について”，情報処理学会自然言語処理研究会，情処研報 NL88-5(1992)．
2）安藤貞雄，樋口昌幸（共編著）：“英文法小事典”，北星堂書店(1991)．
3）佐藤理史，田渕　篤，長尾　真：“CFGパーサのオブジェクト指向言語による実現法とその並列化”，WOOC '87(1987)．
4）伊藤和夫：“英文和訳演習”（入門編，基礎編，中級編，上級編），駿台文庫(1989)．
5）田中穂積：“自然言語解析の基礎”，産業図書(1989)．