

論文表題を言い換える

佐藤理史^{†,††}

「ある表現を意味内容を保ったまま別の表現に変換する」という言い換えは、自然言語を巧みに操るためには不可欠な能力と考えられるが、その機械的実現については、ほとんど研究されてこなかった。本論文では、言い換え研究の1つの例題として、科学技術論文表題の言い換えを取り上げ、比較的簡単な構文的言い換えがどのように実現できるかを検討する。論文表題は、長さの制約により凝縮した表現をとることが多い。これは複合名詞の形をとる。この複合名詞が動詞的働きを持つ名詞を含む場合、複合名詞の内部構造を解析し、適切な形に展開することによって、動詞句の形に言い換えることができる。本論文では、その具体的方法を提案するとともに、その方法に基づいて作成したシステムが、実際の論文表題をどのように言い換えるかを示す。

Automatic Paraphrase of Technical Papers' Titles

SATOSHI SATO^{†,††}

This paper proposes a method of paraphrase, which is defined as transformation of an expression to a different expression of the same meaning. I choose paraphrase of technical papers' titles in Japanese as a target problem, because they often contain complex compound nouns. In case a compound noun contains the noun that works as a verb semantically, the compound noun can be expanded by analyzing its internal structure, and by using the syntactic transformational rule that is appropriate to the structure. This paper describes the detail of the method, an implementation of the paraphrase system based on the method, and the system outputs.

1. はじめに

「あることを表現するのに、いろいろな言い方がある」ということは、自然言語が持つ重要な特徴の1つである。いろいろな言い方、つまり、同じ意味内容を表す複数の表現（文章）を結び付ける変換は、言い換え（パラフレーズ）と呼ばれ、我々人間は、1つの文章をいくつもの違った形に言い換えることができるのが普通である。

この言い換え能力を、我々人間はたまたま持っているのであろうか。それとも自然言語を操る能力の重要な一部分なのだろうか。筆者は後者の方がより自然だと考える。難解な文章をやさしく言い換えることによって理解したり、相手に言いたいことが伝わらなかった場合に言い換えて伝え直すといったことを、我々人間は日常的に行っており、もし、この能力がないとする

ならば、自然言語によるコミュニケーションは立ち行かなくなるように思えるからである。つまり、「1つのことを表すのに複数の表現が存在する」とことと、「ある表現を、意味内容を保持したまま別の表現に変換する能力」は表裏の関係にあり、切り離すことができないのではないか。もし、これが正しいとするならば、言い換えを研究することは、自然言語、あるいは、人間の自然言語処理能力を解明する1つの切り口となるはずである。

一方、工学的にも、言い換え能力を持ったシステムを実現することは大きな意味を持つ。先に述べたように、自然言語では1つのことを表すのに複数の表現が存在するため、表層上異なる表現（テキスト）がほぼ同じ意味を持つかどうか判定すること（意味的等価性の判定）が、自然言語の意味内容を扱うあらゆる処理において必要となる。シソーラスによって定義される同義語をほぼ同じ意味の語として扱うこと（シソーラスを用いた単語間の意味的等価性の判定）は、近年広く行われるようになってきたが、句や文といったそれより大きな単位の意味的等価性を判定する方法は確立されていない。たとえば、現在のテキスト検索では、

† 北陸先端科学技術大学院大学情報科学研究科

School of Information Science, Japan Advanced Institute of Science and Technology

†† 科学技術振興事業団さきかけ研究 21「情報と知」領域グループ
"Information and Human Activity", PRESTO, JST

「日英機械翻訳システム」という検索入力から、「日本語を英語に翻訳するコンピュータ・プログラム」という表現を含むテキストを見つけ出すことはできないが、これは、これら2つの表現が意味的にはほぼ等価であることを判定する機械的な方法が存在しないことによる。この問題に対する1つの解決法は、言い換えの機械化を実現し、2つの表現の意味的等価性を表現間の言い換え可能性によって定義する方法である。言い換えは同義な表現への置換であるから、この方法は、シソーラスを用いた単語間の意味的等価性の判定の素直な拡張となる。

このように、言い換えを研究することは、科学的にも工学的にも大きな意義があると考えられるのにもかかわらず、これまで言い換えに関する研究は少なく^{1)~4)}、その全貌はほとんど解明されていない。このため、言い換えの機械化を実現するためには、言い換えるのあらゆる側面に関して精力的に研究を進める必要がある。

本論文では、言い換え研究の1つの例題として、科学技術論文表題の言い換えを取り上げ、比較的簡単な構文的言い換えがどのように実現できるかについて検討し、その実現のために何が必要かを明らかにする。また、その検討に基づいて作成したシステムの概要と実際の論文表題に対して適用した結果についても述べる。

2. 何を例題とするか

2.1 問題設定の困難さ

言い換えの研究を行う際の最初の障害は、対象問題を設定することが難しいという事実である。このことは、次の3つの質問に集約されよう。

- (1) なぜ言い換えが必要か?
- (2) どんな言い換えを行うか?
- (3) 正解は?(どうやって評価するか?)

言い換えは、いわば単言語内翻訳である。通常の翻訳(たとえば日本語から英語への翻訳)がなぜ必要なのか説明する必要がないのに対し、言い換えは、なぜ言い換えが必要なのか(言い換えるの目的)を必要とする。あるいは、目的がないところには言い換えは存在しないと言った方がよいかもしい。つまり、言い換えは、ある意味内容を表すテキストを生成するだけでは不十分であり、ある目的に合致した属性(より分かりやすい、より短い、より堅い感じ、など)を持ったテキストを生成する必要がある。この意味において、言い換えるの実現には、意味内容以外のテキストの属性を何らかの方法で扱う必要が生じてくる。

ここでの最大の問題は、言い換えるの目的(とそれに合致するテキスト属性)がそれほどはっきりとは定義できないという事実である。たとえば、「より分かりやすくする」という目的は、テキストの分かりやすさとは何かということが分かっていない現状においては、非常にあいまいである。目的がはっきりしなければ、どんな言い換えるを行えばよいかははっきりしなくなり、その結果、何をもちて正解とすればよいかもはっきりしなくなる。

このように、言い換えるは明らかに不良設定問題(ill-defined problem)であり、正解が客観的に明確であるような問題とは異なった扱いが必要である。このため、研究が十分進むまでの当面の間は、各種の判断を主観的に行わざるをえない。

2.2 言い換えるの便宜的な分類

ある表現をほぼ等価な表現に変換するという「言い換える」がどこまでの範囲のものを指すのかは、それほど明確ではない。ここでは、まず、言い換えるの種類を大まかに分類することを試みる。ここで参考になるのが、以下のような翻訳の分類である⁵⁾。

- (1) シンタックスにたよる翻訳: 言語学に立脚するもので、与えられた文の単語と構造を他言語の対応する単語と構造に移し変える(マッピング)という考え方に立つ。
 - (2) セマンティックスに基づく翻訳: コミュニケーション理論の立場に立つもので、発話文の持つ意味が同一になるように他言語の文を作り出すという考え方に立つ。
 - (3) プラグマティックスを中心とする翻訳: 社会言語学的立場に立つもので、ある発話文がその社会で理解され、種々の効果をあたえるのとまったく同じ効果を他の言語の社会においても生じさせるような翻訳文をつくるという考え方に立つ。
- 言い換えるもこの分類に対応する3つのクラスを考えることができる。

- (1) 構文的言い換える: 言葉に関する知識によって実現可能な言い換える。単語を同義語や類義語に置き換える言い換えるや、構造のマッピングに基づく言い換えるをこのクラスに分類する。典型例を以下に示す。

- [1] システムを生成する → システムを作る
- [2] システムの生成 → システムを生成すること

- (2) 意味的言い換える: 参照表現などを、それが指す内容で置き換える言い換えるがこのクラスに含まれる。また、省略されているものを意味的に補

うような言い換えもこのクラスに含める。以下に例を示す。

- [3] 去年の出来事 → 1998年の出来事
- [4] 筆者の考え → 佐藤の考え
- [5] 去年の『知恵蔵』には載っていない → 去年出版された『知恵蔵』には載っていない

(3) プラグマティックな言い換え：(1), (2) 以外のより複雑な言い換え。ある状況において同じ効果を持つような文に言い換えるものが、これに含まれる。以下に例を示す。

- [6] どなたか gcc のソースのありかをご存知ないでしょうか → gcc のソースが置いてある ftp サイトを教えてください

明らかに、構文的言い換えが最も実現が容易であると考えられる。以下、本論文では、研究対象を構文的言い換えに限定する。

2.3 論文表題の言い換え

ここでは、構文的言い換えを研究するための1つの例題として、論文表題の言い換えを取り上げる。日本語の科学技術論文の表題は、短い文字数の中に多くのことを押し込むことが多いため、凝縮した表現が多く見られ、言い換えを考える格好の材料を提供する。また、後で述べるように、工学的応用にもつながる可能性がある。

論文表題の例を図1に示す。この例からも分かるように、日本語の科学技術論文の表題には、以下のよ

1	クラス概念によるモデリングを用いた知識型3次元ビジョンシステム
2	仮説選定機構の一実現法
3	構文的予測の分析とその構文解析への応用
4	フレーム型データ構造の一論理的記述について
5	文脈自由言語パーザへの Prolog プログラム変換の応用
6	優先度つきトークンバッシングプロトコルを用いた分散型探索機構
7	演繹的網型データベースシステム
8	オブジェクト指向概念を導入した知識表現言語: 姿
9	対話領域の独立性を指向した日本語対話理解システム
10	動詞の構文—意味属性による日本語動詞句内の多義語の同定
11	Prolog に関するいくつかの性質について
12	関係データベースに基づく演繹データベースの推論実行方式
13	継承階層 Prolog の高速化機構
14	簡単なパルス回路における不連続変化の定性的解析法
15	知識ベース更新の論理的モデル

図1 論文表題の例 (人工知能学会誌より)

Fig. 1 Examples of technical papers' titles.

うな特徴が見られる。

- (1) 論文表題は、ほとんどの場合、名詞句である。
- (2) 動詞的概念が名詞として埋め込まれる。たとえば、3, 5に見られる「応用」や、10の「同定」は、見かけ上は名詞であるが、動詞的概念を表している(以下では、このような名詞を動詞性名詞と呼ぶ)。
- (3) 複合名詞が多く使われる。
- (4) 動詞性名詞が複合名詞に埋め込まれることが珍しくない。たとえば、2の「仮説選定機構」は「仮説を選定する機構」であり、12の「推論実行方式」は「推論を実行する方式」である。
- (5) いくつかのよく使われる表現がある。たとえば、1, 6に見られる「を用いた」や、10の「による」、14の「における」などは、非常に多用され、連語として名詞句を関係づける働きをする。これらは、動詞としての扱いをする必要がない。

これらの特徴を考慮して、以下のような言い換えを実現することを考える。

- (1) 動詞性名詞が含まれている複合名詞を展開する。たとえば、「構文的予測」を「構文的に予測する」に、「日本語対話理解システム」を「日本語対話を理解するシステム」に展開する。この言い換えは、「何をどうした」という構造を明確化する働きを持つ。

(2) 末尾の複合名詞のみに適用する

上記の展開を表題に含まれるすべての複合名詞に適用すると、埋め込み文の段数が増え、かえって分かりにくくなる。そのため、上記の展開を、表題の末尾の複合名詞に制限する。たとえば、「仮説選定機構の一実現法」は、「一実現法」のみを展開し、「仮説選定機構を実現する一つの方法」と言い換える。「仮説を選定する機構を実現する一つの方法」とはしない。

ここで考える言い換えは、おおよそ、名詞句に凝縮された表現を動詞句に言い直すことによって、より分かりやすい(やさしい)表現にするもの、ということができる。

3. 複合名詞の言い換え

上記の言い換えの主要な部分は、動詞性名詞を含む複合名詞を動詞を使った言い方に変換することである。これを、複合名詞の解析、複合名詞の展開、複合名詞に係っている要素の変更、の3つの処理によって実現する。

表 1 複合名詞の解析例とそれに基づく言い換え

Table 1 Analysis of compound nouns and thier paraphrases.

複合名詞					言い換え				
S?	A*	N*	V	T?	A*	N*	V	S?	T?
—		仮説	選定	機構		仮説を	選定する		機構
—	論理的		実現	法			実現する	一つの	方法
	分散型		記述		論理的に		記述する		
		日本語対話	探索	機構	分散型で		探索する		機構
			理解	システム		日本語対話を	理解する		システム
	縦統的	問題	分解		縦統的に	問題を	分解する		

3.1 複合名詞の解析と展開

複合名詞の内部構造の解析に関する研究は数多くなされておられ (たとえば文献 6), 7)), その中には言い換えを念頭において内部構造を設定したものもある⁸⁾. しかし, 筆者の知る限り, 実際に複合名詞を文字列として入力し, それを言い換えたもの (文字列) を出力するシステムは報告されていない. ここでは, 過去の研究成果をふまえ, 実際の論文表題 339 編 (人工知能学会誌 1986-1995 年掲載論文) を用いて言い換えに必要な複合名詞の解析法と展開法を検討し, 以下の方法を得た.

複合名詞の解析

動詞性名詞を含む複合名詞を, 以下のパターンに従って要素に分割する.

S? A* N* V T?

ここで, ‘*’, ‘+’, ‘?’ は, それぞれ正規表現における標準的な意味 (*: 任意個, +: 1 個以上, ?: あってもなくてもよい) を表し, アルファベットは, 以下の意味を表す.

- S ... 数詞.
- A ... 形容詞/副詞的に働く要素. ナ形容詞語幹 (「等価」), 様態を表す名詞 (「自動」), 特定の接尾辞 (「~型」, 「~的」) を持つものなど.
- N ... 名詞.
- V ... 動詞性名詞. サ変名詞, 動詞からの派生語 (「試み」), 「~化」など.
- T ... 動詞性名詞の後ろに現れる名詞.

複合名詞の展開

このようなパターンを用いて解析された複合名詞を 図 2 のように展開する. すなわち,

- (1) 数詞 (S) を, 末尾の名詞 (T) にかける. ただし, T がなく, S だけが存在する場合は, S の係り先がないので, これを削除する.
- (2) 動詞性名詞 (V) を動詞化する. T が存在する場合は, これにかける (連体修飾).
- (3) 名詞 (N*) は, ヲ格化し, 動詞 (V) にかける.
- (4) 形容詞/副詞的要素 (A*) は, 連用修飾に直し,

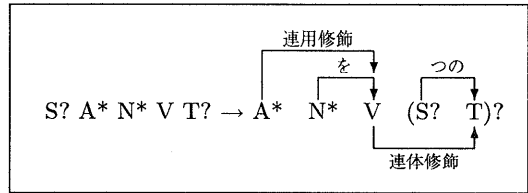


図 2 複合名詞の展開法

Fig. 2 Expansion of compound noun.

動詞 (V) にかける.

複合名詞の解析例とそれに基づく言い換え例を表 1 に示す.

3.2 複合名詞に係っている要素の変更

上記の展開により, 複合名詞は複数の文節に展開される. そのため, 元の複合名詞に係っていた要素が存在する場合は, その要素を展開後のどの文節にどのようにかけるかを決定する必要が生じる. 係り先の候補は, 展開によって現れたヲ格要素 (N) と動詞 (V) であり, 係り方の候補は, 連体修飾, 連用修飾, 格要素などである. これについては, 次の方法で決定する.

- (1) 複合名詞内にヲ格となる名詞 (N*) がない場合
 - (a) 直前の「~の」はヲ格要素とし, 動詞にかける
 - (b) それ以外は, 連用修飾に変更し, 動詞にかける
- (2) 複合名詞内にヲ格となる名詞 (N*) がある場合
 - (a) 直前の「~の」は, 連体修飾のままとし, ヲ格要素にかける
 - (b) それ以外は, 連用修飾に変更し, 動詞にかける

以下にそれぞれの例を示す.

- (1) 簡単なパルス回路における不連続変化の定性的解析法 → 簡単なパルス回路において不連続変化を定性的に解析する方法
- (2) 関係データベースに基づく演繹データベースの推論実行方式 → 関係データベースに基づいて演繹データベースの推論を実行する方式

3.3 類似の言い換え

次のような言い換えも、上記の方法とはほぼ同様の方法で実現可能である。

- (1) 名詞 + の + 動詞性名詞 ⇒ 名詞 + を + 動詞性名詞 (プロダクション・システムによる線画の解釈 → プロダクション・システムを用いて線画を解釈する^{*})
- (2) 動詞性名詞 + の + 一方法 ⇒ 動詞 + 一つの方法 (非単調推論による学習者の理解構造のモデル化の一方法 → 非単調推論を用いて学習者の理解構造をモデル化する一つの方法)
- (3) 解法 ⇒ 解く方法 (スケジューリング問題の解法 → スケジューリング問題を解く方法)

4. 実装

4.1 システムの概要

上記の方法を実際の論文表題に適用するためには、あらかじめ、論文表題(文字列)を文節係り受け構造に変換しておく必要がある。論文表題の係り受け解析には、専用のプログラムを作成する方法⁹⁾もあるが、今回は、汎用の自然言語処理ツールである、JUMANおよびKNPを用いた^{**}。KNPを用いる場合の1つの問題点は、KNPは、原則として、連体修飾要素を直後の文節にかけることである。たとえば、「知識を用いた建築図面の理解」において、「用いた」は「建設図面」に係ると判定される。このため、連体修飾要素の係り受けの変更が必要になる。

作成したシステムの概要を図3に示す。本システムは、すべて自動化されており、入力として受け付けた論文表題(文字列)に対して、以下の処理を行い、言い換え結果(文字列)を出力する。

- (1) 文字列処理による整形
ここでは、不要な部分の削除と最上位の並列句の分割を行う。たとえば、「MBT1:実例に基づく訳語選択」、「フレーム型データ構造の一論理的記述について」の下線部が削除される。また、「発見的学習とその支援環境」は、「発見的学習」と「その支援環境」に分割される。
- (2) 構文解析
JUMAN(juman3.5)とKNP(knp2.0b5)を用いて、文字列を文節係り受け構造に変換する。形態素解析辞書および構文解析文法は標準のものを用いる。

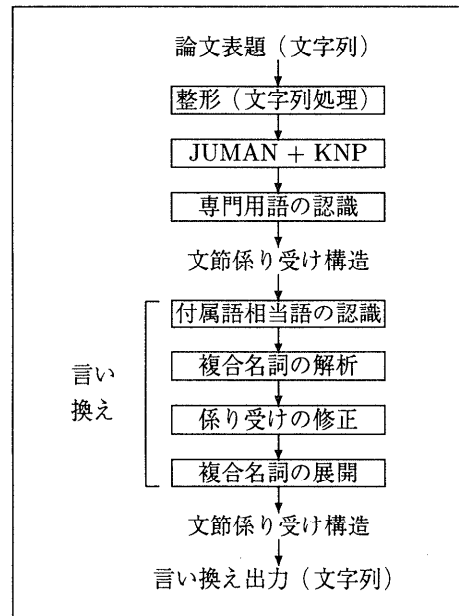


図3 システム構成

Fig. 3 System configuration.

- (3) 専門用語の認識
それぞれの文節において、専門用語となる部分にマークをつける。専門用語辞書として、岩波情報科学辞典¹⁰⁾を用いる。マークされた専門用語は、言い換えを抑制する(4.2節)ために用いられる。
- (4) 付属語相当語の認識
「に基づく」、「を用いた」などの連語を付属語相当語の扱いとする。たとえば、KNPは「実例に/基づく/訳語選択」を3文節(/は文節の切れ目を表す)と見なす^{***}が、これを「実例に基づく/訳語選択」のように2文節に変換する(「に基づく」を付属語として扱う)。
- (5) 複合名詞の解析
3.1節に述べたパターンに基づいて、複合名詞を解析する。
- (6) 係り受けの修正
連体修飾の係り先をヒューリスティックスを用いて変更する。たとえば、「知識を用いた/建築図面の/理解」における「知識を用いた」の係り先を「建設図面」から「理解」に変更する。
- (7) 複合名詞の展開
表題末尾(文節係り受け構造の最上位)の複合

^{*}「により」よりは「を用いて」の方が自然だと考えられるので、このように変換する。

^{**} <http://pine.kuee.kyoto-u.ac.jp/nl-resource/>

^{***} KNPの文節は、1つの自立語とそれに続く任意個の付属語から構成される。名詞、動詞、形容詞などは自立語と認識され、助詞や判定詞などは付属語と認識される。

名詞が動詞性名詞を持つ場合、3.1 節に述べた方法で複合名詞を展開し、あわせて複合名詞に係っていた要素の係り先を変更する(3.2 節)。ただし、ある条件が成り立つ場合には、複合名詞の展開(言い換え)を行わない(次節参照)。また、末尾の複合名詞が動詞性名詞を持たない場合も展開を行わない。

4.2 言い換えの抑制

論文表題の末尾の複合名詞が動詞性名詞を持つ場合でも、3 章で述べた言い換えを適用すると不自然な言い換えとなってしまう場合がある。そのため、本システムでは、以下の場合には、言い換えを行わないことにした(この判定もシステムが自動的に行う)。

(1) 動詞性名詞が複数の語からなる専門用語の一部である場合

たとえば、「仮説推論」は「仮説を推論する」に展開せず、そのまま残す。専門用語はそれ自身独自の意味を持つようになるため、一般の構文的規則で展開すると意味が異なってくる場合が多い。たとえば、「仮説推論」は「仮説を用いた推論」の意味である。

(2) 複合名詞に動詞の連体修飾が係っている場合
たとえば、「対話領域の独立性を指向した日本語対話理解システム」を「対話領域の独立性を指向して日本語対話を理解するシステム」に言い換えるのは明らかに誤りであり、言い換えるべきではない。これは、「指向した」が「システム」に係っているからである。可能な言い換えは、「対話領域の独立性を指向した、日本語対話を理解するシステム」であるが、これは日本語として好ましくない。なお、「に基づく」や「を用いた」などは、付属語相当語扱いとするので、ここでの動詞の連体修飾には含まれない。

(3) 動詞にヲ格要素がなく、かつ、複合名詞から連用修飾要素も得られない場合

たとえば、「算術問題解答システムにおける学習」を「算術問題解答システムにおいて学習する」に言い換えると、目的語(ヲ格要素)がないため、奇異に感じられる。このような言い換えは抑制する。ただし、ヲ格要素がなくても複合名詞から得られる連用修飾要素がある場合は展開する(「知的 CAI における直接的指導方略」→「知的 CAI において直接的に指導する方略」)。

4.3 実行例

図 4 に本システムの出力例を示す。この出力例は、図 1 の 15 表題に対する出力である。15 表題のうち、9 表題(番号の前に「*」がついているもの)が言い換えられている。残りの 6 表題に対しては、言い換える

1	クラス概念によるモデリングを用いた知識型 3 次元ビジョンシステム
*2	仮説選定機構を実現する一つの方法
*3	構文的予測を分析する それを構文解析へ応用する
*4	フレーム型データ構造を論理的に記述する
*5	文脈自由言語パーザへ Prolog プログラム変換を応用する
*6	優先度付きトークンパッシングプロトコルを用いて分散型で探索する機構
7	演繹的網型データベースシステム
8	オブジェクト指向概念を導入した知識表現言語
9	対話領域の独立性を指向した日本語対話理解システム
*10	動詞の構文—意味属性を用いて日本語動詞句内の多義語を同定する
11	Prolog に関するいくつかの性質
*12	関係データベースに基づいて演繹データベースの推論を実行する方式
*13	継承階層 Prolog を高速化する機構
*14	簡単なパルス回路において不連続変化を定性的に解析する方法
15	知識ベース更新の論理的モデル

図 4 プログラムによる言い換え結果
Fig. 4 Paraphrase results.

べきではないと判定され、入力そのまま出力されている☆。

5. 実験と評価

5.1 実験

作成したシステムを用いて、実際の論文表題を言い換える実験を行った。実験対象には以下の論文表題を用いた。

- システム作成用: 人工知能学会誌 1986-1995 年に掲載の 339 編(先に調査に用いたものと同一)
- システム評価用: 人工知能学会誌 1996-1998 年に掲載の 131 編(新たに用意した未知のデータ)

2.1 節で述べたように、ある言い換えが正しいものであるかどうかを客観的に決定することはかなり難しい。ここでは、「真に客観的な評価を目指すのではなく、言い換えシステムの性能の目安が得られればよい」という立場に立って、筆者の主観的な判断に基づき、それぞれの論文表題に対して正解(複数の場合もある)を与え、これとシステムの出力を比較することを行った。

理想的な場合を考えるならば、ある入力に対して、正しい言い換えが存在するか、あるいは、存在しない(言い換えない方がよい)かのどちらかである。一方、

☆ ただし、8 と 11 は、文字列処理による整形により、表題の末尾の一部が削除されている。

表2 言い換えの評価
Table 2 Evaluation of paraphrase.

	言い換えた		言い換えなかった
	正しい	誤り	
正解=言い換える	A	B	C
正解=言い換えない	-	D	E

システムの出力は、言い換えたか、言い換えなかったか(無変換)のどちらかであり、前者はさらに、正しい言い換えか否かに分けられる。つまり、結果は、表2のように整理されることになる。ここで、言い換えるべきではないものを言い換えた場合に、それが正しい言い換えとなることはないので、この表において値が入るのはAからEの5カ所となる。

ここで、次の2つの精度を考える。

- 実用精度：このシステムはどの程度使いものになるか

$$\text{実用精度} = \frac{A + E}{A + B + C + D + E}$$

- 書き換え精度：言い換えられたもののうち、正しいものはどれくらいか

$$\text{書き換え精度} = \frac{A}{A + B + D}$$

実際に言い換えシステムを作成した場合、そのシステムに求められるのは、言い換えるべきものを正しく言い換え、言い換えるべきではないものをそのまま残すことである。実用精度はこれを表している。しかしながら、実際には、言い換えられたものが正しいかどうかを調べるのに比べて、言い換えられなかったものが言い換えるべきではなかったかどうかを判定することはより主観的になる傾向が高い。このため、実際に言い換えたもののうち正しいものはどれくらいであるかをもって、精度を見積もることが実際的な方策となる。書き換え精度はこれを表している。

表3、表4にそれぞれ、システム作成用、システム評価用のデータに対する実験結果を示す。

5.2 評価

表5に誤りの原因を示す^{*}。最も多い誤りは、係り受けの修正の誤り(c)で、これは、連体修飾の係り先を適切に変更する処理(ヒューリスティクス)が不十分であることを示している。このヒューリスティクスはシステム作成用のデータに基づいて作成したので、システム評価用(未知)のデータに対して誤りがより多く発生している。しかし、このタイプの誤りは、形態素解析の誤り(a)や並列構造の認定誤り(b)と同

表3 実験結果：人工知能学会誌1986-1995(339編)
Table 3 Evaluation: JSAI 1986-1995 (339 titles).

	言い換えた		言い換えなかった
	正しい	誤り	
正解=言い換える	155	16	4
正解=言い換えない	-	4	160

(実用精度 = 93% 書き換え精度 = 89%)

表4 実験結果：人工知能学会誌1996-1998(131編)
Table 4 Evaluation: JSAI 1996-1998 (131 titles).

	言い換えた		言い換えなかった
	正しい	誤り	
正解=言い換える	61	12	5
正解=言い換えない	-	6	47

(実用精度 = 82% 書き換え精度 = 77%)

様、基本的には係り受け解析の誤りであり、言い換える誤りではない。表5からも分かるように、係り受け解析の誤り(a~c)が全体の約半数を占めている。

言い換える誤りとして解決策を模索すべきものは、複合名詞の解析誤り(d)と展開時の誤り(e)である。前者には、以下のようなものがある。

- [7] ニューラルネットを用いた緩和ラベリング法 → ×ニューラルネットを用いて緩和をラベリングする方法
- [8] 設計支援のための統合モデリング環境 → ×設計支援のために統合をモデリングする環境
- [9] 組合せ利用 → ×組合せを利用する(○組み合わせさせて利用する)
- [10] 新正規化法 → ×新正規化する方法(○正規化する新しい方法)
- [11] モデル事例ベースを用いた定性的多目的最適設計 → ×モデル事例ベースを用いて定性的多目的を最適に設計する

最初の2つは、それぞれ「緩和」「統合」を形容詞/副詞的に働く要素として認定できないことが誤りの直接の原因であるが、実際には、どちらも言い換えない方がよい。これは、「ラベリング法」や「統合～環境」がある意味で準専門用語化しているからだと考えられる。次の2つは、複合名詞解析に用いているパターンの不備である。前者は、「動詞性名詞 + 動詞性名詞」が並列の関係となる場合であり、後者は、「新」が「法」に係る場合である。最後の誤りは、「多目的」を形容詞/副詞的に働く要素と認定できていないことに直接の原因があるが、たとえ認定できたとしても、どう言い換えるべきかよく分からない。

一方、複合名詞の展開時の誤りには以下のようなものがある。

^{*} この表において、B, C, Dは、表2のカテゴリに対応する。

表5 誤りの原因の分析
Table 5 Error analysis.

誤りの原因	システム作成用				システム評価用			
	B	C	D	計	B	C	D	計
a. 形態素解析の誤り (JUMAN)	1	2		3				0
b. 並列構造の認定誤り (KNP)	3		2	5	1		2	3
c. 係り受け修正の誤り	5			5	6	2		8
d. 複合名詞解析の誤り	2	1	2	5		1	3	4
e. 複合名詞展開時の誤り	3			3	1			1
f. 専門用語関連の誤り		1		1		1	1	2
g. その他	2			2	4	1		5
合計	16	4	4	24	12	5	6	23

- [12] プロダクションシステムの高コストルール対処法 → ×プロダクションシステムの高コストルールを対処する方法
- [13] 機器・センサ多重故障に関する定量的高信頼診断手法 → ×機器・センサ多重故障を定量的に高信頼で診断する手法
- [14] ソフトウェア自動運用のためのポリシー記述言語 → ×ソフトウェア自動運用のためにポリシーを記述する言語

最初の誤りは、「対処する」が二格をとる自動詞であることによる。現在の方法は、直前の「～の」を格要素に変換する場合は、無条件にヲ格としているが、これは、動詞の種類に基づいて適切に決定する必要がある。2番目の誤りは、「高信頼で」が奇異であることに原因がある。「信頼度高く」のように言い換えることが望ましい。最後の誤りは、「～ための」を無条件に動詞への連用修飾要素に変換していることに原因がある。この例のように、名詞（「ポリシー」）に係る場合がある。

その他の誤りには、以下のものがある。

- [15] 屋外電話回線故障位置推定エキスパートシステム → △屋外電話回線故障位置を推定するエキスパートシステム（○屋外電話回線の故障位置を～）
- [16] ロバスト解探索型遺伝的アルゴリズムの基礎提案 → ×ロバスト解探索型遺伝的アルゴリズムの基礎を提案する（○ロバスト解探索型遺伝的アルゴリズムを提案する）

前者の例は、適切な位置に「の」を挿入することが望ましい。一方、後者の「基礎提案」は、「基礎的に提案する」とするよりは、「基礎」を思い切って削除した方がよい。

6. 検 討

本研究において、論文表題の構文的言い換えが機械的に実現可能であることが実証されたと考える。先に

述べた客観的評価の問題は存在しているが、図4の言い換え結果が示すようなはば妥当な言い換えが、比較的簡単な規則で実現できたといつてよいだろう。

言い換えの規則が比較的簡単で済んだ理由は、次の2つの理由によると考えられる。

- (1) 構文的言い換えに限定した。より具体的には、複合名詞の構文的言い換えに限定した。
- (2) 対象とした複合名詞の内部構造の種類が限定されていた。

特に後者は、予想外に種類が少なかった。論文表題は、大きく、名詞型主題（「演繹的網型データベースシステム」）と動詞型主題（「知識を用いた建築図面の理解」＝「知識を用いて建築図面を理解する」）に分けられるが、人工知能の分野では、後者の動詞型主題で使われる動詞のほとんどが他動詞であることが、複合名詞の内部構造の種類少なさにつながっている。情報科学の分野の論文においては、これと同様のことが成り立つと考えられるが、これは実際に、より多くの論文表題に対してシステムを動かし、確かめてみる必要があるだろう。

本論文で提案した言い換えは、以下の3つの技術に支えられている。

- (1) 係り受け解析
- (2) 複合名詞解析
- (3) 構文的要素の変形操作

(1)の係り受け解析は、JUMANおよびKNPの恩恵を受けている。これらのツールは完全ではないが、実験システムのモジュールとして十分に機能した。(2)の複合名詞解析は、パターン自身は単純であるが、それに必要な辞書、すなわち、名詞の下位分類に関しては、問題含みである。既存の辞書などを利用して、動詞性名詞の網羅的リストを作ることは可能であると考えられるが、形容詞/副詞的に働く名詞のリストを作ることは、そのような情報が明記されていないため、はるかに困難である。今後作られる機械可読辞書には、

このような情報も掲載すべきである。(3)の変形操作とは、名詞を動詞化したり、連体修飾を連用修飾に変換したりする操作を指す。これらに関しても、それを行うアルゴリズムが明文化されたものが乏しく、それに必要なプログラムや辞書といったものほとんど整備されていない。このため、今回の実験では、必要なものを試行錯誤的に作成したが、今後は、これらのプログラムや辞書を体系的に整備していくことが必要となろう。

本研究の興味は、基本的には、言い換えをいかにして実現するかということにあったが、論文表題の言い換えは、少なくとも2つの実際の応用がある。1つは、翻訳への応用である。日本語論文表題をこねられた英語に翻訳するためには、複合名詞に含まれる動詞性名詞に対して、その格要素を認定することが不可欠である。本研究の言い換え手法は、それを実現する1つの方法となる。また、それをさらに発展させると、言い換えによる翻訳¹¹⁾につながっていくことになる。もう1つの応用は、表題解析による論文の自動分類¹²⁾への応用である。言い換えを用いて表題を標準化することにより、自動分類の精度を高めることが期待できる。

謝辞 本研究の一部は、文部省科学研究費補助金(課題番号 10878060, 萌芽的研究, 「論文表題の一般化による科学技術論文の階層的詳細分類」)の助成による。

参考文献

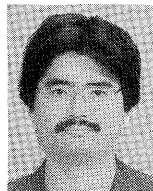
- 1) 白井 諭, 池原 悟, 河岡 司, 中村行宏: 日英機械翻訳における原文自動書き替え型翻訳方式とその効果, 情報処理学会論文誌, Vol.36, No.1, pp.12-21 (1995).
- 2) Robin, J. and McKeown, K.: Empirically Designing and Evaluating a New Revision-Based Model for Summary Generation, *Artificial Intelligence*, Vol.85, No.1-2, pp.135-179 (1996).
- 3) Dras, M.: Reluctant Paraphrase: Textual Restructuring under an Optimization Model, *Proc. Pacling97*, pp.48-55 (1997).
- 4) Sato, S. and Sato, M.: Rewriting in Auto-

mated Editing of QA-Pack, *JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and its Application*, pp.59-64 (1998).

- 5) 長尾 真: 機械翻訳はどこまで可能か, 岩波書店 (1986).
- 6) 影山太郎: 文法と語形成, ひつじ書房 (1993).
- 7) 小林義行: コーパスを用いた日本語複合名詞の解析に関する研究, 技術報告 TR96-0002, 東京工業大学理工学研究科情報工学専攻 (1995).
- 8) 宮崎正弘: 係り受け解析を用いた複合語の自動分割法, 情報処理学会論文誌, Vol.25, No.6, pp.940-979 (1984).
- 9) 松村 敦, 高須淳宏, 安達 淳, 池田和幸: 構造化インデクスを用いた情報検索システム, アドバンスト・データベース・シンポジウム '97 論文集, pp.151-158, 情報処理学会 (1997).
- 10) 長尾 真ほか(編): 岩波情報科学辞典, 岩波書店 (1990).
- 11) 佐藤理史: アナロジーによる機械翻訳, 共立出版 (1997).
- 12) 今井 俊, 佐藤理史: 表題解析による科学技術論文の詳細分類, 第57回情報処理学会全国大会講演論文集 (3), pp.211-212 (1998).

(平成10年7月27日受付)

(平成11年4月1日採録)



佐藤 理史 (正会員)

1983年京都大学工学部電気工学科第二学科卒業。1988年同大学院博士課程研究指導認定退学。同大学工学部助手を経て、1992年より北陸先端科学技術大学院大学情報科学研究科助教授。1997年より科学技術振興事業団研究員を兼任。1998年より京都大学大学院情報学研究科社会情報学専攻助教授を併任。京都大学博士(工学)。自然言語処理, 機械学習, 情報の自動編集等の研究に従事。言語処理学会, 日本認知科学会, AAI, ACL各会員。著書: 「自然言語処理」(共著, 岩波書店, 1996), 「アナロジーによる機械翻訳」(共立出版, 1997), 「言語情報処理」(共著, 岩波書店, 1998)等。