情報学基礎 64−19
自然言語処理 145−19
(2001. 9. 11)

英文科学技術文における単純名詞句決定法の比較

武田 薫† 石野 明‡ 竹田正幸‡ 松尾文碩‡
† 九州大学大学院システム情報科学府
‡ 九州大学大学院システム情報科学研究院

概要

英文科学技術文の理解を目指し，我々は，統計的手法により得られる被修飾度と前方修飾度によって中核名詞と前方修飾語を決定し，これに基づいて単純名詞句を決定する手法の開発を行ってきた [3, 4]. 一方，Voutilainen ら [6] によって開発されたルールに基づく品詞タグ付け器 EngCG により付与されたタグを用い，辻井ら [9] の手法をもとにした単純名詞句決定手続きを作成した．これら二つの手法を英文科学技術文献データ INSPEC の抄録文に適用したところ，統計的な手法による結果のほうがルールに基づく手法による結果よりも優れていた．

Comparing the Methods for Determining Simple Noun Phrases in the English Scientific Sentences

Kaoru Takeda† Akira Ishino‡ Masayuki Takeda‡ Fumihiro Matsuo‡
†Graduate School of Information Science and Electrical Engineering, Kyushu University
‡Faculty of Information Science and Electrical Engineering, Kyushu University

Abstract

We have developped a statistical method for determining headwords by their modificant degrees, premodifiers by their premodifying degrees and then simple noun phrases[3, 4]. On the other hand, based on the method given by Tsujii et al.[9], we made out another procedure for determining simple noun phrases with the tags given by the rule-based part-of-speech tagger EngCG developped by Voutilainen et al.[6] We applied these two methods to the abstracts in the English scientific documents data INSPEC, and the result by the statistical method was sperior to the one by the rule-based method.

1 はじめに

英文科学技術文の理解を目指して，入力文を，動詞を述語記号，名詞句をそのまま項とした論理式に変換することを考える．この変換を行うためには，文の動詞句の統語構造と名詞句の範囲を決定する必要がある．そのなかでも本稿では，特に名詞句の決定についての考察を行う．

以下の例文の下線部で示しているように，名詞句は，太字で示したようなより基本的な単位である名詞句が前置詞などによって結合したものと考えることができる．

> **The values** of **the registration parameters** are automatically calculated by maximizing **an integer similarity measure** selected for **robustness**.

ここで，この基本単位となる名詞句を，単純名詞句と呼ぶことにする．すなわち，単純名詞句は，ひとつの中核名詞とその前方修飾語からなる名詞句である．

この単純名詞句を決定する手法として，本稿では 2 通りの手法を提示する．ひとつは統計的な手法に基づくものであり，もうひとつはルールに基づく方法である．具体的には，前者は被修飾度・前方修飾度を用いる手法，後者は品詞タグ付け器 EngCG により得られる情報を利用する手法である．ルールに基づく手法では，対象文書に対応して制約規則を作成することにより高い判定精度を期待できるものと考えられるが，その制約規則の生成は人手に頼らなければならず多大な手間がかかる．一方，被修飾度や前方修飾度は機械的に算出することが可能であり，統計的な手法ではあまり人手がかからない．

本稿では，第 2 節において被修飾度・前方修飾度を用いて単純名詞句の範囲を決定する手法を，第

3 節において EngCG により品詞を決定し単純名詞句の範囲を決定する手法を提示する．そして，第 4 節において実際に英文科学技術抄録文に対してそれらの手法を適用し，単純名詞句決定のそれぞれの結果を比較し，考察を行う．最後に第 5 節においてまとめを行う．

2 被修飾度・前方修飾度による手法

この節では単語の被修飾度・前方修飾度を用いた単純名詞句決定の手法について説明する．

手順としては，まず文の動詞句を決定する [1]．そして残った動詞句以外の部分の単語列において被修飾名詞を決定する [2]．最後に単純名詞句の範囲を決定する [3, 4]．

2.1 動詞句の決定

この項では，単文における主動詞の決定法を述べる．この決定法は，動詞候補選出手続きと動詞決定手続きの 2 つの手続きからなる．

2.1.1 動詞候補選出手続き

動詞候補選出手続きは，基本的には，研究社英和中辞典 [5]（以下，中辞典と略称する）において動詞の品詞を持つ語を候補として選出する．すなわち，動詞の原形，3 人称単数現在形，過去形を候補として選ぶ．現在分詞形と過去分詞形は，受動形，進行形，完了形をなしている場合にのみ，候補として選出する．いずれの場合も，直前に冠詞や “to” を伴うものは候補としない．すなわち，ここでは主動詞を決定するのが目的であるため，不定詞を形成する動詞は候補としない．

2.1.2 動詞決定手続き

単文において，2 個以上の動詞候補が選出された場合には，そのうちのひとつを動詞として決定しなければならない．

まず，動詞候補を次の 3 つに分類する．

A 助動詞を伴う候補，受動形，進行形，完了形

B 動詞の過去分詞形と同形でない候補

C 動詞の過去分詞形と同形の候補

A の候補は，高い確率で動詞となる．そこで，A の候補には B や C の候補よりも高い優先度を与える．

B や C の候補に対しては，抄録－索引生起比を優先度として用いる．ある語 w の抄録－索引生起比 $r(w)$ とは，w の抄録文での生起頻度を $f_{\mathrm{a}}(w)$，自由索引語中での生起頻度を $f_{\mathrm{i}}(w)$ とすると，

$$r(w) = \begin{cases} \frac{f_{\mathrm{i}}(w)}{f_{\mathrm{a}}(w)} & f_{\mathrm{a}}(w) \neq 0 \text{ のとき} \\ 1 & \text{それ以外のとき} \end{cases}$$

で定義される値のことである．抄録文に主動詞として現れる動詞は機能語的なものが多く，その抄録－索引生起比は比較的低いことが分かっている．すなわち，抄録－索引生起比の小さい候補に高い優先度を与える．そこで動詞決定手続きは次のようになる．

1. 非極大候補（別の候補の部分単語列となる候補）を除去する．
2. A の候補があればそれを動詞とする．
3. A の候補がないとき，抄録－索引生起比の最小の候補を動詞とする．

2.2 被修飾名詞の決定

英語には，名詞を他の名詞の修飾語として使う用法がある．本稿では他の名詞を修飾する名詞を修飾名詞，修飾される名詞を被修飾名詞と呼ぶことにする．例えば，database system において database が修飾名詞で system は被修飾名詞である．また，the document database では database が被修飾名詞である．本稿では，名詞が修飾名詞として生起していないとき，その名詞は被修飾名詞として生起しているという．

次の 3 つの例文で被修飾名詞の生起を太字で示した．また，括弧内の単語列は動詞句である．

例文 2.1

The starburst **observations** (are) a major **motivation** for the **consideration** of this **model** since the extreme **conditions** (are observed) .

例文 2.2

The **properties** of the black **hole** and the nonthermal **radiation** from its **environment** (are calculated) under the **assumption** that the mass **influx** (is) constant.

例文 2.3

An approximate **solution** to these **equations** (is determined) using a Galerkin **technique** involving polynomial and trigonometric **functions**.

例文 2.1 では，被修飾名詞の後置語は動詞句と for, of, since である．例文 2.2 では，後置語は動詞句, of, and, from, that であり，例文 2.3 では，動詞句, to, involving, 文末のピリオドである．このように，被修飾名詞の後置語の多くは，動詞句，現在分詞，過去分詞，of, for, from のような前置詞, and, since のような接続詞，文末ピリオドである．文末ピリオドも語とみなすことにすれば，被修飾名詞は多くの場合，後置語によって決定できる．したがって，後置語による被修飾名詞決定手続きは，基本的には名詞に非名詞が後続する 2 単語列を決定する手続きということになる．

しかし，動詞の ing 形は現在分詞および名詞として出現する．例えば，動詞 engineer の現在分詞形 engineering は,「工学」を意味する場合には通常名詞に分類される．句

the software **engineering** of two important classes of computer systems

では，engineering は被修飾名詞として生起している．この場合，後置語 of で被修飾名詞であることが判定できる．一方，句

information system **engineering** the design, implementation and evaluation of the human-machine interface

では，engineering は現在分詞であり，この場合は後置後 the で現在分詞であることが分かる．ところが，句

two injection moulded semicrystalline **engineering** thermoplastic materials

や

X11 computer assisted software **engineering** integrated tool sets

の場合，engineering の後置語は形容詞あるいは過去分詞であるが，このことにより engineering が被修飾名詞と判断することはできない．したがって，ing 形の後置語が冠詞，形容詞，過去分詞の場合，ing 形を被修飾名詞とするわけにはいかない．

名詞の品詞を持つ語は，後置語が非名詞であれば被修飾語であるものと後置語が冠詞，形容詞，過去分詞形の場合は被修飾名詞としてはならないものに分けられる．そこで，被修飾名詞となりうる語を次の 2 つの範疇に分ける．

A 単品詞名詞

B 多品詞名詞と ing 形

上記以外の非名詞は，B の語の被修飾名詞決定のために次の 2 つの範疇に分ける．

C 冠詞，形容詞，過去分詞形

D 前置詞，副詞，接続詞，動詞など，A，B，C 以外の非名詞および句読点

文中の単語を A，B，C，D に分類したあと，被修飾名詞を決定する手続きは前に述べた．すなわち，A の単語が被修飾名詞であるのは，C，D の単語が後置された場合であり，B の単語が被修飾名詞であるのは D の単語が後置された場合である．

単語が被修飾名詞として生起する相対頻度を被修飾度と呼ぶことにする．この被修飾度が大きい単語は，名詞として生起する確率が高いと考えられる．そこで本節の名詞決定法では，被修飾度がある閾値以上となるような単語を名詞とみなす．

2.3 単純名詞句の範囲の決定

後方修飾と連言と選言を扱わないとき，単純名詞句の文法は次のような簡単な形で定義できる．

$$SNP \rightarrow N_m \mid M\ SNP$$

SNP : 単純名詞句

M : 前方修飾語

N_m : 被修飾名詞

単純名詞句の範囲決定はこの文法を満たす適当な範囲を探すことといえる．このとき，被修飾名詞と前方修飾語が決定されている必要がある．被修飾名詞については 第 2.2 項の手法によって決定できるので，前方修飾語の決定が問題になる．

この前方修飾語の決定を考える際に単純名詞句の先頭語に着目する．さらに単純名詞句の先頭語の中でも確実に識別できる冠詞 the の次の語を対象とする．2 語以上からなる単純名詞句において冠詞 the の次の語はその後ろの語を修飾しているので前方修飾語となる．また前置詞のように単純名詞句の中に前方修飾語としてあらわれないような語が冠詞 the の次の語として現れることはない．

したがって冠詞 the の次の語が前方修飾語と同じような傾向を示すと考えると，ある単語 w が表れる頻度を $f(w)$，"the" の後に現れる頻度を $f_{\text{the}}(w)$ としたとき，

$$m(w) = \frac{f_{\text{the}}(w)}{f(w)}$$

は単語の前方修飾語らしさのひとつの指標と考えられる．この指標 $m(w)$ をこれ以降，前方修飾度と呼ぶ．

前方修飾語を用いた単純名詞句の範囲決定の方法として，上記の文法から M^*N_m に一致する部分を単純名詞句の範囲とする方法が考えられる．また，英文科学技術文では "natural language database" のように名詞が名詞を修飾している場合があるので $(M+N_m)^*N_m$ に一致する部分という考え方もできる．こうすることで単純名詞句内の非前方修飾語かつ被修飾語である語を単純名詞句の一部として正しく決定できるようになるが，冠詞のない単純名詞句の直前に非前方修飾語かつ被修飾語である語があるとき単純名詞句の一部に誤って取り込んでしまう．

3 EngCG による手法

この節では EngCG を用いた単純名詞句決定の手法ついて説明する．

3.1 EngCG とは

EngCG (English Constraint Grammer)[6, 7][1] タガーとは，英文を構成する各トークンに対して，ルールに基づいて決定された品詞タグと文中での役割を示す機能タグを付加して返すものである．

この処理は，次に述べる 3 段階の手順を踏むことにより実現されている．

はじめに，英文をトークンごとに切り分ける．基本的には，ひとつの単語あるいはひとつの句読点をトークンとみなす．ただし，複数の単語からなっている成句でその成句全体でひとつのトークンと考える場合もある[2]．

次に，各トークンに付けるべきタグの候補を決定する．これには約 70 万の項目を持つ辞書が使われる．辞書に載っていないトークンが出現した場合はそのトークンをルールに基づいて解析することによって付けるべきタグを推測する．なお，この段階ではまだトークンが置かれている前後の文脈は考慮されない．そのため，各トークンには考えられる限りにおいてさまざまなタグの候補が付くことになる．

最後に，前段階で提示されたタグ候補の中から最もふさわしいタグを選択する．これは，各トークンの置かれている文脈をもとに，約 4,000 の制約規則を用いて行われる．なお，この制約規則はすべて人手により作成されたものである．

EngCG タガーはルールに基づくタグ付け器であり，従来からある統計的な手法を用いたタグ付け器よりも判定精度の面で優れた性能をもつとされている [8]．

3.2 単純名詞句の範囲の決定

辻井らは，英文の各トークンの情報について EngCG タガーを用いて詳細に判定し，その情報を用いて名詞句を決定するという手法を示している [9]．

英文を EngCG タガーに与えると，EngCG タガー はその文の各トークンに対してタグをつけ，結果として返す．ただし，このタグはかなり細かな分類になっている．本節の EngCG による単純名詞

[1] もとは Altro Voutilainen 氏がヘルシンキ大学にて開発したもので，現在はバージョン 2 となり同氏の所属する Conexor 社が開発・販売する各種製品の基盤技術として使われている．

[2] 例えば "in order to" や "as well as" などの成句は，その成句自体をひとつのトークンとみなす．

句の判定においては，これほど細かな分類は必要ない．そこで EngCG のタグよりも簡略化したタグを各トークンに付け直す．EngCG のタグから簡略化したタグへの対応は表 1 に示した．こうして得られたタグ列で，名詞句を表す文法に合致する部分タグ列があれば，それに対応する部分単語列が名詞句ということになる．

表 1: EngCG のタグから簡略化タグへの対応表

EngCG のタグ	簡略化タグ	説明
@DN>	D>N	冠詞
@NN>, @A>, @QN>, @AD-A>	M>N	前方修飾語
(@NH, @VOC, @SUBJ, @F-SUBJ, @OBJ, @I-OBJ, @PCOMPL-S, @PCOMPL-O, @<P) && (NOM, NUM, PRON)	HEAD	中核名詞
(@NH, @VOC, @SUBJ, @F-SUBJ, @OBJ, @I-OBJ, @PCOMPL-S, @PCOMPL-O, @<P) && (not(NOM, NUM, PRON))	M>N	前方修飾語

辻井らの手続きで決定される名詞句は後方修飾なども考慮したより一般的なものである．第 2 節で示した手法で決定されるのは単純名詞句であるので，結果を単純に比較することができない．そこで，辻井らの手続きにおいて名詞句を判定する文法を簡略化し，単純文字列を判定する文法にした手続きを作成した．

単純名詞句の文法は，後方修飾・連言・選言を扱わないとき，各トークンについて決定した簡略化タグを用いて次のように定義できる．

$$
\begin{aligned}
SNP &\rightarrow M' \ N_m \mid Det \ M' \ N_m \\
M' &\rightarrow \varepsilon \mid M \ M' \\
Det &\rightarrow \texttt{D>N} \\
M &\rightarrow \texttt{M>N} \\
N_m &\rightarrow \texttt{HEAD} \\
SNP &: \text{単純名詞句}
\end{aligned}
$$

4 実験

本研究では、調査対象として、INSPEC (Information Service in Physics, Electro technology and Control) に掲載されている抄録文を利用した。INSPEC とは，英国 IEE (Institution of Electrical Engineering) が提供している代表的英文二次文献データである．一次刊行物から得られたデータから機械可読の二次文献情報を作成し，速報誌と抄録誌の編集・製版の機械化と文献情報の検索を行うことを主目的として 1965 年に開発がはじまり，1969 年にほぼ完成された．

被修飾度および前方修飾度は INSPEC の 1984 年から 1993 年までの 10 年間分の抄録文データ 2,408,118 文献 10,482,511 文から算出したものを，単純名詞句の決定に使用した．

単純名詞句決定の評価は、INSPEC 抄録文から抽出した 1,487 文を用いて行った．これに，被修飾度・前方修飾度による手法と，EngCG による手法とを適用して単純名詞句の決定を行った．その決定結果が，あらかじめ正解として人手により決定しておいた単純名詞句（全部で 8,618 句）と一致しているかどうか調査した．その結果，表 2 のようになった．

表 2: 単純名詞句決定の評価結果

		被修飾度・前方修飾度による手法		EngCG による手法
		M^*N_m	$(M+N)^*N_m$	
判定した句の総数		9,071	8,995	8,780
正しく決定	実数	7,291	7,395	7,047
	精度	80.38%	82.21%	80.26%
	再現率	84.60%	85.81%	81.77%
その他	実数	1,790	1,600	1,733

被修飾度・前方修飾度による手法のほうが，EngCG による手法よりもよい結果が得られた．

以下は，被修飾度・前方修飾度による手法による単純名詞句の決定に失敗した例である．${ と } に囲まれた単語列が，この手法により決定された単純名詞句である．

例文 4.1

(前略) ${fracture mechanics} for **${soil}, ${rock} or ${concrete materials}**.

例文 4.1 では，本来太字部分の単語列全体がひとつの単純名詞句であるが，その判定に失敗している．これはもともと単純名詞句を決定する文法が連言や選言に対応していないためである．従って，ルールに基づく手法でも同じく起こる問題である．

例文 4.2

(前略) in **such ${long transmission lines** having **distributed parameters}**.

例文 4.2 でも，本来は太字部分の各単語列が個別の単純名詞句であるが，やはり判定に失敗している．この場合は，分詞の having の決定失敗が元で前方修飾語の決定に失敗している．

5 まとめ

本稿では，単純名詞句の範囲の決定を行う手法として，単語の被修飾度・前方修飾度を算出することによる方法と EngCG tagger を用いる方法との 2 つを示した．

それらの手法を INSPEC の抄録文を用いて検証すると，被修飾度や前方修飾語による単純名詞句決定法のほうが EngCG による手法よりも精度と再現率ともに良い結果が得られた．

EngCG において用いられている制約規則は，科学技術文に対応して作成されたものではない．したがって，科学技術文を処理するために新たに制約規則を作成すれば，精度や再現率を向上させることは可能であると思われる．しかし，人手でそのような制約規則を作成する作業はかなり困難である．一方，被修飾度や前方修飾度というのは，ともにタグのない大量のコーパスから機械的に算出することができるため，人手による作業を必要としないという利点をもつ．

今後の課題として，連言や選言への対応，現在分詞形や過去分詞形を含む単純名詞句の決定精度の向上などが挙げられる．

参考文献

[1] 竹田正幸，松尾文碩：『英文科学技術抄録文における動詞の決定』，情報処理学会論文誌 34(9)，pp. 1931-1936，1995

[2] 竹田正幸，松尾文碩：『英文科学技術抄録文における名詞の決定』，情報処理学会論文誌 36(8)，pp. 1827-1837，1995

[3] 丸木，柴田，日昔，竹田，松尾：『英文科学技術文における単純名詞句の範囲決定』，情報処理学会第 53 回全国大会講演論文集 (2)，pp. 23-24，1996

[4] 河崎裕司，丸木健次，竹田正幸，松尾文碩：『英文科学技術文における単純名詞句の決定について』，電気関係学会九州支部第 51 回連合大会，1998

[5] 小稲義男ほか編：『新英和中辞典』，第 5 版，研究社，1985

[6] http://www.conexor.fi/analysers.html

[7] Altro Voutilainen: "EngCG tagger, Version 2", Sprog og Multimedier, Aalborg Universitetsforlag, Aalborg, 1997

[8] Christer Samuelsson and Atro Voutilainen: "Comparing a Linguistic and a Stochastic Tagger", ACL-EACL97, ACL, 1997

[9] Takeshi Sekimizu, Hyun Seok Park and Jun'ichi Tsujii: "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts", Genome Informatics, pp. 62-71, 1998