

意味範疇の散らばりに基づいた名詞の統語範疇の分類

田 中 省 作[†] 富 浦 洋 一[†] 日 高 達[†]

名詞句「 NP の NP 」は、2つの名詞句が助詞「の」で結合した名詞句である。日本語文中では、このような名詞句が頻繁に現れ、多様な意味構造を持つことが知られている。このような名詞句に対して、その統語構造から形式的に意味構造を構成する文法体系が提案されている。この文法体系では、名詞句を4つの統語範疇（普通名詞句、項句、関係名詞句、事象名詞句）に細分化しており、この文法体系に基づいて、意味構造を推定するには、名詞がどの統語範疇に属すのかという情報が必要となる。これらの統語範疇のうち項句と事象名詞句に属する名詞はすべて、既存の辞書の文法情報により分類できる。しかし、普通名詞句に属する名詞（普通名詞）と関係名詞句に属する名詞（関係名詞）は、このような情報からは分類することが難しい。そこで、本論文では、コーパスから獲得した名詞句の用例から求めた名詞句「 NP_1 の N_2 」における N_2 を固定したときの NP_1 の意味範疇の散らばりの度合いに基づいて、普通名詞と関係名詞を統計的に決定する手法を提案する。分類実験の結果、79.2%の精度で分類された。

Classification of Syntactic Categories of Nouns by the Scattering of Semantic Categories

SHOSAKU TANAKA,[†] YOICHI TOMIURA[†] and TORU HITAKA[†]

The noun phrase “ NP ‘no’ NP ”, that consists of two noun phrases NPs connected by an adnominal particle ‘no’, is frequently used in Japanese sentences. The surface structure of this pattern is simple, but it has various semantic structures. For such noun phrases, there has been a grammar proposed, which gets semantic structures of noun phrases from their syntactic structures systematically. This grammar fractionates noun phrases into four syntactic categories (CN, T, RN, EN). To estimate semantic structures of noun phrases on this grammar, it is essential to get the information about syntactic category of each nouns in the noun phrases. The nouns that belong to T or EN category are automatically classified by referring grammatical information on existing dictionaries. On the contrary, it is difficult to classify automatically the nouns (*common noun, relation noun*) that belong to CN or RN category by such information. This paper proposes the method for statistically deciding syntactic category of the latter nouns by noun phrase NP_1 ’s scattering of semantic categories when noun N_2 is fixed in examples of noun phrases “ NP_1 ‘no’ N_2 ” in corpora. As a result of this experiment on classification, the accuracy is 79.2%.

1. はじめに

日本語文では、2つの名詞句を助詞「の」で結合した名詞句「 NP の NP 」が頻繁に現れ、しかも、多様な意味構造を持つことが知られている。そのため、名詞句「 NP_1 の NP_2 」の意味構造および表層表現には明示されない NP_1 , NP_2 の間の意味関係を求めるることは、自然言語処理の基本的な問題の1つといえる。たとえば、日本語から英語などの他言語への機械翻訳を考えた場合では、名詞句「 NP_1 の NP_2 」における NP_1 と NP_2 の間の意味関係を推定したうえで翻訳す

る必要がある。「私の車」というような場合、「私」と「車」の間に『所有関係』が成り立っていると考えられるため、“my car”と翻訳され、「駐車場の車」という場合、『位置関係』が成り立っていると考えられるため、“a car in the parking lot”と翻訳される。また、データベースの自然言語インタフェースなどでは、名詞句「 NP_1 の NP_2 」における NP_1 と NP_2 の意味関係を推定し、「 NP_1 の NP_2 」が何を指示しているのか求める必要がある。

このような多様な意味関係を持つ名詞句「 NP_1 の NP_2 」に対して、Montagueの形式化³⁾に基づいて形式的に意味構造を与えるような文法体系が提案されている⁸⁾。この文法体系では、名詞句を意味的観点からさらに4つの統語範疇（普通名詞句、項句、関係名詞

[†] 九州大学大学院システム情報科学研究科知能システム学専攻
Graduate School of Information Science and Electric Engineering, Kyushu University

句、事象名詞句)に細分化するが、名詞句の意味構造推定のシステムを計算機上に実装するには、単語レベルの名詞句すなわち名詞の統語範疇に関する情報が必要となる。細分化された4つの統語範疇のうち、項句や事象名詞句に属す名詞は、表記や既存の辞書の文法情報から分類することができる。しかしながら、残る普通名詞句と関係名詞句に属す名詞(それぞれを普通名詞、関係名詞と呼ぶ)については、このような情報から機械的に分類することは困難だった。そこで、本論文ではコーパス中の用例を基に、名詞句「 NP_1 の N_2 」において名詞 N_2 を固定したときの NP_1 の意味範疇の散らばりに着目し、普通名詞と関係名詞を統計的に分類する手法を提案する。

2章では、まず文献8)で提案されている文法体系を概観し、細分化された名詞句の統語範疇、および統語構造と意味構造との対応について述べる。3章では、名詞句「 NP_1 の N_2 」において、 N_2 が普通名詞の場合と関係名詞の場合とで、 NP_1 の意味範疇の散らばりに異なる傾向があることを説明する。また、普通名詞と関係名詞を分類するために、意味範疇の散らばりをコーパス中の共起情報を用いて定量化する。4章では、意味範疇の散らばりを普通名詞および関係名詞に対する1つの特徴量とし、k-NN推定法に基づいた分類手法を述べ、EDRコーパスおよびRWCテキストデータベースを用いた名詞の分類実験および、その結果を示す。また、普通名詞と関係名詞の半自動分類システムへの応用について考える。

2. 名詞句「 NP の NP 」の意味構造

2.1 名詞句の統語範疇

文献8)で提案されている文法体系では、名詞句「 NP の NP 」に対して、Montagueの形式化に基づいて、統語規則と意味構造への翻訳規則とを対応づけるために、従来、单一の統語範疇として扱われていた名詞句を意味的観点から4つの統語範疇に細分化する。ここでは、この文法体系について概観する。この文法体系では名詞句の意味構造を型論理で記述しているが、本論文では、簡単のために名詞句の意味構造を一階述語論理で記述することにする。統語規則、翻訳規則および正確な意味構造の記述など、詳細は文献8)を参照していただきたい。

2.1.1 普通名詞句(CN)

「色白の美人」における「色白」や「美人」は、それぞれ“色白である”、“美人である”という性質を表している。このような名詞句は、普通名詞句(common noun phrase; CN)という統語範疇に属す。統語的に

は、「ある」「その」「すべての」などの英語の冠詞に相当する連体詞(本論文では、これらを限定詞と呼ぶ)が結合し、項句になるような句が CN である。 CN に属す名詞を普通名詞(common noun)と呼ぶ。

この統語範疇に属す名詞句 n の意味構造は、 n に対応する一変数述語 ' n^* ' によって、

$$n^*(x) \quad (1)$$

と表される。たとえば、「色白」であれば、その意味構造は、

$$\text{色白}(x)$$

となり、「個体 x が色白である」ことを表す。

2.1.2 項句(T)

ある特定の個体や事象を指示しているような名詞句は、項句(term phrase; T)という統語範疇に属す。統語的には、「太郎」や「彼」や CN に限定詞が結合した「その人」や「すべての男性」というようなものである。この統語範疇の名詞句は、動詞の格要素や後述する関係名詞句の補項になりうる句である。

この統語範疇に属す名詞(固有名詞や代名詞など) n の意味構造は、一階述語論理では、個体記号

$$n^* \quad (2)$$

として表される。たとえば、「太郎」は、

$$\text{太郎}^*$$

という個体記号で表す。

2.1.3 関係名詞句(RN)

個体や事象間の関係を表す名詞句、すなわち、「 T の NP 」が指示する個体と T が指示する個体の関係を表す名詞句 NP は、関係名詞句(relation noun phrase; RN)という統語範疇に属す。

たとえば、「兄」という名詞句は、 RN である。このとき、 T である「二郎」を「の」で結合して「二郎の兄」というような名詞句を構成することによって、「二郎と兄弟関係にある個体で、かつ年上の個体」を指示することになる。ここで、「 T の RN 」で T が指示する個体を、その RN の補項と呼び、「 T の RN 」で指示される個体を RN の主項と呼ぶこととする。先の例では、「個体二郎」が「兄」の補項であり、「二郎の兄」で指示される個体が「兄」の主項である。 RN に属す名詞を関係名詞(relation noun)と呼ぶ。

この統語範疇に属す名詞句 n は、ある個体や事象間の関係を表していることから、その意味構造は二変数述語 n^* によって、

$$n^*(x, y) \quad (3)$$

と表される。ただし、 n^* の第1項の x が n の主項を、第2項の y が n の補項である。または、 RN である n 自身が表す個体や事象間の関係を R_n という

二変数述語として, n の意味構造を,

$$n^*(x) \wedge R_n(x, y) \quad (4)$$

とも表せる。たとえば、「兄」は,

$$\text{兄}^*(x, y)$$

と表され, 兄 $^*(x, y)$ は“個体 x は個体 y の兄である”ことを表す。または, 式(4)に従うと,

$$\text{兄}^*(x) \wedge R_{\text{兄}}(x, y)$$

となる。

2.1.4 事象名詞句 (EN)

事象を表している名詞句は, 事象名詞句 (event noun phrase; EN) という統語範疇に属す。「勉強」や「考え」といったサ変動詞の語幹や, 動詞が名詞に転化した名詞などは EN である。これらの EN は, 「太郎の英語の勉強」「太郎の考え」のように, 「の」と結合して元の動詞のように格要素をとり, 事象を指示する。

2.2 名詞句「NP の NP」の統語構造と意味構造

名詞句「 NP_1 の NP_2 」の意味構造は, 細分化した名詞句の統語範疇と密接に対応しており, その意味構造は NP_1 が T か CN で大別される。

2.2.1 「CN の NP」の意味構造

「CN の」は形容詞的に機能し, NP で指示する対象を制限することになる。「CN の NP」の統語範疇は NP の統語範疇と同一である。統語構造が「 CN の CN 」である名詞句「 n_1 の n_2 」の意味構造は,

$$n_1^*(x) \wedge n_2^*(x) \quad (5)$$

となる。たとえば, 「大型の犬」という名詞句では, 「大型の」は「犬」で指示する“犬である”という性質を持つ個体を, “大型である”という性質を持つ個体に制限し, その意味構造は,

$$\text{大型}^*(x) \wedge \text{犬}^*(x)$$

という論理式で表される。このとき, 「大型の犬」の統語範疇は CN である。また, 「色白の妻」というように NP の統語範疇が RN である場合は, 「色白の」は「妻」の主項を制限することになり, 意味構造は,

$$\text{色白}^*(x) \wedge \text{妻}^*(x, y)$$

となる。このとき, 「色白の妻」の統語範疇は RN である。

2.2.2 「T の NP」の意味構造

「 T の」は, 「 T が」や「 T を」のような格要素的に機能し, NP の翻訳された論理式の引数に代入された形になり, 「 T の NP 」の統語範疇は T である。

NP が CN および RN の場合について, それぞれ説明する。

(1) 「 T の CN 」

この統語構造の名詞句の意味構造は, T で指示する個体と CN という性質を持つ個体の集合の間の

意味関係を推定し, その意味関係を表す二変数述語 R (本論文では, この R を意味関係述語と呼ぶ) を補ったうえで構成される。統語構造が「 T の CN 」である名詞句「 n_1 の n_2 」の意味構造は,

$$n_2^*(x) \wedge R(x, n_1^*) \quad (6)$$

となる。このような意味関係述語 R は, 表層表現には明示されていない。たとえば, 「私の車」であれば, 「私」「車」の名詞句間に「所有関係」が一般に成立するので, 「私の車」の意味構造は, 二変数述語‘所有’を補い,

$$\text{車}^*(x) \wedge \text{所有}(x, \text{私}^*)$$

と表される。ただし, 所有 (x, y) は, “個体 y が個体 x を所有している”ことを意味する。このような意味関係は, 人間が容易に推論できることから, 名詞句中の語彙に関する知識から導き出されているものと考えられる。

(2) 「 T の RN 」

統語構造が「 T の RN 」の名詞句では, 「 T の」が (i) RN の補項に割り当てられる用法と, (ii) RN の主項を制限する用法とがある☆。まず, 「 T の」が RN の補項に割り当てられる用法の名詞句「 n_1 の n_2 」の意味構造は, n_2 自身が意味関係となり,

$$n_2^*(x, n_1^*) \quad (7)$$

となる。または, n_2 の意味構造を式(4)で表せば, 「 n_1 の n_2 」の意味構造は,

$$n_2^*(x) \wedge R_{n_2}(x, n_1^*) \quad (8)$$

となる。たとえば, 「太郎の母」の意味構造は,

$$\text{母}^*(x, \text{太郎}^*)$$

となる。ただし, 母 (x, y) は“個体 x は個体 y の母親である”ことを意味する。式(8)に従うと,

$$\text{母}^*(x) \wedge R_{\text{母}}(x, \text{太郎}^*)$$

となる。

一方, 「 T の」が RN の主項を制限するような用法の名詞句は, たとえば「台所の母」などである。この場合は, ちょうど「 T の CN 」と同じように, 意味関係述語 R を導出してその意味構造を構成する。「台所の母」の場合の意味構造は,

$$\exists z \text{ 母}^*(x, y) \wedge \text{位置}(z, x) \wedge \text{台所}^*(z)$$

となる☆。ただし, 位置 (x, y) は“個体 y が地点 x に位置する”ことを表す。

☆ 文献 7) では, 関係名詞を不完全名詞と呼び, 「 T の RN 」という場合に, 「 T の」が RN の主項を制限する用法を連体修飾用法と呼び, 「 T の」が RN の補項に割り当てられる用法を連体補充用法と呼んでいる。

☆☆ 「台所の母」という名詞句は, 「[(あの) 台所] の母」というように「台所」に「あの」という限定詞が結合し T となり, その限定詞「あの」が省略されていると考える。

しかし、*RN*は個体や事象間の関係を表している名詞句であり、*RN*の補項が割り当てられていないような名詞句単体では実質的な意味を欠く。そこで、「*T*の*RN*」という統語構造では、「*T*の」は*RN*の主項を制限する用法に比べ*RN*の補項に割り当てられる用法の方が多いと予想される[☆]。そこで、「*T*の*RN*」という統語構造では、ほとんどが「*T*の」は*RN*の補項への割り当てられる用法となっていると仮定する。

2.3 既存の文法情報を用いた統語範疇の分類

名詞句「*NP* の *NP*」の意味構造を正しく推定するには、まず単語レベルの名詞句すなわち名詞が、細分化された4つの範疇のいずれに属すかが正しく決定されなければならない。従来、学校文法などの多くの文法では、名詞句については基本的に文献8)のような細分化は行われていないが、単語レベルの名詞については固有名詞、代名詞、時詞といった形で細分化されている。たとえば、九州大学の公用データベース日本語単語辞書¹⁰⁾では、名詞を普通名詞、固有名詞、数詞、人称代名詞、指示代名詞の5つに分けている。また、EDRの日本語単語辞書⁵⁾では、名詞を普通名詞、固有名詞、数詞、時詞、形式名詞の5つに、RWCのテキストデータベース⁶⁾では、さらに詳しく32種類に分けており、たとえば固有名詞については人名なのか組織名なのか地域なのかということまで区別している。そこで、これらの既存の辞書に示されている名詞に関する情報を用いることで、文献8)で考える統語範疇への対応が一部とれることを示す。

(1) 項句(*T*)との対応

固有名詞は、名詞単体である特定の個体を指示しているものであるので、*T*に属すと考えられる。また、代名詞についても、その代名詞が出現するまでの文脈でその指示対象が固定されている。よって、*T*に属す。*T*に属す名詞は固有名詞や代名詞のみに限られるので、*T*についてはすべて網羅される。

(2) 事象名詞(*EN*)との対応

事象名詞は、動詞が名詞化したものであるから、辞

[☆] 「母」は、比較的主項を制限するような用法が現れていた関係名詞である。EDRコーパス中の「*T*の母(母親)」という名詞句のうち、「*T*の」が「母」の主項を制限していた数と「母」の補項に割り当てられる用法の数の割合は、それぞれ18.0%, 82.0%であった。抽象的な関係を表す「理由」では、主項を制限する用法が5.7%, 補項に割り当てられる用法が94.3%であった。同様に、抽象的な関係を表す「原因」や「結果」、また位置的関係を表す「隣」や「横」「東」「西」といった関係名詞では、「*T*の」が主項を制限していた用法は、ほとんど見られず、多くが補項への割り当てられる用法となっていた。

書中の動詞に関する情報からすべて網羅される。動詞が名詞化する場合、その表記は動詞の品詞によって次のようになることが分かっている。

• サ変動詞の場合

サ変動詞の語幹が、名詞化した場合の表記となる。たとえば、サ変動詞「講義する」であれば、名詞化した場合の表記は、その語幹「講義」となる。

• サ変動詞以外の動詞

動詞の連用形が、名詞化した場合の表記となる。たとえば、五段動詞「走る」であれば、名詞化した場合の表記は、その連用形「走り」となる。

よって、ある名詞が*EN*に属するもののかどうかは、その名詞の表記と同じ語幹を持つサ変動詞が存在するかどうか、または、同じ表記の連用形となる動詞が存在するかどうかで決定することができる。

(3) 一部の関係名詞(*RN*)との対応

「美しさ」や「大きさ」というような名詞は、*RN*に属す。これらの名詞は、形容詞の語幹に「さ」が連接して名詞化したものである。また、「重み」や「苦しみ」などは、形容詞の語幹に「み」が連接して名詞化したものである。このよう、「(形容詞の語幹)+さ」や「(形容詞の語幹)+み」といった形態の名詞は、すべて*RN*に属す。

さらに、「勇敢さ」や「大胆さ」という名詞も、*RN*に属す。これらの名詞は、形容動詞の語幹に「さ」が連接して名詞化したものである。このように、「(形容動詞の語幹)+さ」といった形態の名詞もすべて*RN*に属す。

実際に、九州大学の公用データベース日本語単語辞書¹⁰⁾に登録されている名詞61270語について調べてみると、22,290語が*T*, *EN*, *RN*のいずれかに分類される。*T*, *EN*は網羅されていることから、残りの63.6%の名詞は、普通名詞または関係名詞である。しかし、それらの名詞がいずれの統語範疇に属するもののかは、表記や品詞などの文法情報からでは判定できない。

3. 普通名詞と関係名詞の統計的性質

細分化された統語範疇のうち、*T*, *EN*に属すすべての名詞、および一部の関係名詞については表記や既存の辞書に示されている文法情報から抽出される。そこで、それらの情報からは決定できない普通名詞と関係名詞の分類のため、まず、その統計的性質として意味範疇の散らばりを導入する。

3.1 意味範疇の散らばり

統語構造が「*T*の*N*」また「*CN*の*N*」である名

詞句「 NP_1 の N_2 」において^{*}, N_2 が普通名詞である場合と, 関係名詞である場合とで, NP_1 の意味範疇の散らばり具合がどのように異なるかについて考える。

3.1.1 「 T の N 」の場合の意味範疇の散らばり

統語構造が「 T の RN 」である名詞句「 np_1 の n_2 」では, T が指示する個体が RN の補項に割り当てられ, その意味構造は, 式(8)に従うと,

$$n_2^*(x) \wedge R_{n_2}(x, np_1^*)$$

となる。つまり, 統語構造が「 T の RN 」という場合には, RN 自身の表す関係 R_{n_2} が np_1 と n_2 の 2 つの名詞句の間の意味関係となる。よって, 統語構造が「 T の RN 」である名詞句「 NP_1 の N_2 」の場合, N_2 を関係名詞 n_2 に固定したときの NP_1 の意味範疇は, R_{n_2} の補項にとりうる意味範疇に集中する^{☆☆}。

たとえば, 統語構造が「 T の RN 」である名詞句「 NP_1 の母」を考えてみる。このとき, 意味関係は ' $R_{\text{母}}$ ' である。すると, NP_1 の意味範疇は, ' $R_{\text{母}}$ ' の補項にとりうる意味範疇である《人間》や《動物》に限られる。ただし, 《 α } は α という意味範疇を表す。

一方, 統語構造が「 T の CN 」である名詞句「 np_1 の n_2 」では, T の np_1 と CN の n_2 の 2 つの名詞句の間で, 表層表現には明示されない意味関係 R が成立し, n_1 は意味関係述語 R の引数に割り当てられ, 次のような意味構造となる。

$$n_2^*(x) \wedge R(x, np_1^*)$$

したがって, 統語構造が「 T の CN 」である名詞句「 NP_1 の N_2 」の場合, N_2 を普通名詞 n_2 に固定したとき, NP_1 の意味範疇は, R の引数にとりうる意味範疇に集中する。だが, 意味関係は, n_2 から一意には決定できず, NP_1 にも依存して決まる。つまり, 意味関係の候補は一般に複数考えられ, その分, NP_1 の意味範疇は散らばりは大きくなる。

たとえば, 統語構造が「 T の CN 」である名詞句「 np_1 の n_2 」として, n_2 が「車」である場合を考え

^{*} ここで、「 NP_1 の NP_2 」という形の名詞句ではなく、「 NP_1 の N_2 」という名詞句のみに限ったのは, NP_2 の主辞の名詞が普通名詞や関係名詞であっても, NP_2 の統語範疇がその主辞の名詞の統語範疇のまとは限らないという理由からである。たとえば、「 NP_1 の NP_2 」の形の名詞句として「高齢の [私の母]」では, NP_2 の主辞の名詞の「母」は関係名詞であるが, NP_2 の「私の母」の統語範疇は, T である。このような場合を避けるため、「 NP_1 の N_2 」という形の名詞句に限定した。

^{☆☆} 「 T の RN 」では、「台所の母」というように「 T の」が RN の主項をある意味関係を介して制限するという用法もあった。この場合の T の意味範疇の散らばりは、「 T の CN 」における T と同程度になるが, 2.2.2 項で述べたように「 T の」が RN の主項を制限する用法の割合は、「 T の」が補項に割り当てる用法に比べ非常に小さいため, NP_1 の意味範疇は, R_{n_2} の補項にとりうる意味範疇に集中する傾向が見られた。

てみる。 np_1 が「太郎」の場合、「太郎」と「車」の間には『所有関係』が成立し, 意味構造は, 意味関係述語‘所有’を用いて,

$$\text{車}^*(x) \wedge \text{所有}(x, \text{太郎}^*)$$

となる。しかし, n_2 が同じ「車」であっても np_1 が「あの駐車場」という名詞句である場合, 意味関係は『位置関係』が推定され, 意味構造は意味関係述語‘位置’を用いて,

$$\exists y \text{ 車}^*(x) \wedge \text{位置}(y, x) \wedge \text{駐車場}^*(y)$$

となる。よって, 統語構造が「 T の CN 」の名詞句「 NP_1 の車」では, 表層表現には明示されない意味関係を『所有関係』と仮定すると, NP_1 の名詞句の意味範疇は, ‘所有’の引数にとりうる意味範疇である《人間》, 《組織》といったものに集中し, 意味関係を『位置関係』と仮定すると, NP_1 の名詞句の意味範疇は, 《地名》, 《具体物》といったものに集中する。

このように, 統語構造が「 T の N 」である名詞句では, N が関係名詞である場合は, 意味関係は唯一であるのに対して, N が普通名詞である場合には意味関係が複数考えられる。よって, NP_1 が T である名詞句「 NP_1 の N_2 」では, N_2 を n_2 に固定したとき, n_2 が関係名詞である場合よりも n_2 が普通名詞である場合の方が NP_1 の意味範疇の散らばりが大きくなることが予想される。

しかし, 現時点では利用可能なコーパスでは, 名詞句を文献 8) のように T と CN のような区別を行っているものはない。その結果, コーパスより収集される名詞句「 NP の N 」は, 統語構造が「 T の N 」という名詞句だけでなく「 CN の N 」というのも混在することになる。そこで, 次に, 統語構造が「 CN の N 」である場合の意味範疇の散らばりについて考える。

3.1.2 「 CN の N 」の場合の意味範疇の散らばり

まず, 統語構造が「 CN の N 」である名詞句「 NP_1 の N_2 」が言語表現として意味的に妥当である必要条件について考える。統いて, それらの条件から N_2 を普通名詞, 関係名詞のいずれの名詞に固定した場合でも NP_1 の意味範疇の散らばりが同程度であることを導く。

統語構造が「 CN の N 」という名詞句「 NP_1 の N_2 」が言語表現として意味的に妥当であるための必要条件は,

$$\text{「}NP_1 \text{ の}\text{」は「}N_2\text{」と矛盾しない} \quad (9)$$

すなわち, 「 NP_1 の」によって「 N_2 」の指示対象が制限されるわけだが, そのような「 NP_1 の N_2 」の指示対象が少なくとも 1 つは存在するということである⁴⁾.

統語構造が「 CN の CN 」である名詞句「 np_1 の n_2 」

では、 np_1 が n_2 の指示対象を制限することになる。たとえば、 n_2 が「犬」である場合を考えてみる。この場合、 np_1 が CN である名詞句として「白の犬」「雑種の犬」「大型の犬」などがあげられる。このとき、名詞句「 np_1 の犬」が意味的に妥当であるには、「 np_1 の」が、 n_2 の「犬」が指示する対象、すなわち“犬である”という性質を持ち、かつ“ np_1 である”という性質を持つ個体が少なくとも 1 つは存在しなければならない。したがって、名詞句「 np_1 の n_2 」が意味的に妥当であるための必要条件は、

$$\exists x [np_1^*(x) \wedge n_2^*(x)] \quad (10)$$

となる。

一方、統語構造が「 CN の RN 」である名詞句「 np_1 の n_2 」の例として、 n_2 が「母」の場合について考えてみる。このとき、「 np_1 の母」といった表現が意味的に妥当であるためには、「 np_1 の」が、「母」の主項（つまり、母親自身）を制限することになる。たとえば、「色白の母」や「病身の母」などがこの場合である。このとき、名詞句「 np_1 の母」が意味的に妥当であるためには、「 np_1 の」が、 n_2 の「母」の主項が指示する対象、すなわち“(誰かの) 母親である”という性質を持ち、かつ“ np_1 である”という性質を持つ個体が少なくとも 1 つは存在しなければならない。ここで、意味関係 $R(x, y)$ の主項 x のとりうる領域を一変数述語 $Im_R(x)$ で表す。すると、意味的に妥当であるための必要条件は、

$$\exists x [np_1^*(x) \wedge Im_{R_{n_2}}(x)] \quad (11)$$

となる。

「 CN の CN 」と「 CN の RN 」とを比較すると、「 CN の RN 」での CN は RN の補項に何ら作用しないため、言語表現として意味的に妥当である必要条件としては同等なもの式 (10), (11) が導かれる。よって、統語構造「 CN の N 」の名詞句「 NP_1 の N_2 」では、 N_2 が普通名詞、関係名詞のいずれの場合でも NP_1 の意味範疇の散らばりは同程度になることが予想される。

3.1.3 「 NP の N 」の場合の意味範疇の散らばり

名詞句「 NP_1 の N_2 」において、 NP_1 は T である場合と CN である場合が考えられる。統語構造が「 T の N 」では、 N が関係名詞では意味関係は唯一であるのに対して、 N が普通名詞では T の NP_1 によって複数の意味関係が仮定されるという違いから、 N が普通名詞の場合の方が意味範疇の散らばりが大きくなる。また、統語構造が「 CN の N 」では、意味範疇の散らばりに大きな相違はない。したがって、名詞句「 NP_1 の N_2 」で N_2 を固定した場合の NP_1 の意味範疇の散らばりは、「 T の N 」における N が普通名

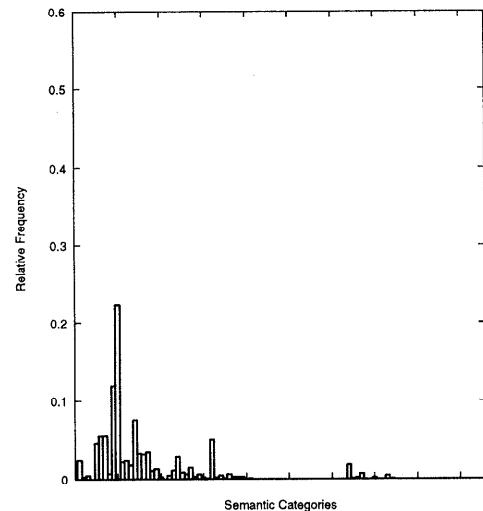


図 1 「自動車」の意味範疇の散らばり

Fig. 1 Scattering of semantic category of "car".

詞、関係名詞の場合の違いが反映される。さらにコーパス中の名詞句「 NP の N 」を見てみると、「 CN の N 」よりも「 T の N 」という統語構造の方が頻出している☆。結果として、 N_2 が普通名詞である場合には、 NP_1 の意味範疇の散らばりは大きくなり、 N_2 が関係名詞である場合には、 NP_1 の意味範疇の散らばりは小さくなる傾向があることが期待できる。

実際に、EDR コーパス⁵⁾から抽出した名詞句「 NP_1 の N_2 」において、 N_2 を普通名詞の「自動車」に固定した場合と、 N_2 を関係名詞の「母」に固定した場合の NP_1 に出現する名詞句の主辞の名詞の意味範疇ごとの出現（相対）頻度を図 1、図 2 に示す☆☆。

3.2 エントロピーを用いた意味範疇の散らばりの定量化

名詞句「 NP_1 の N_2 」において、 N_2 が普通名詞か関係名詞かによって、 NP_1 の意味範疇の散らばりが異なる傾向があることを説明した。名詞を普通名詞と関係名詞に分類するために、コーパス中の「 NP_1 の N_2 」の共起情報を基に、情報理論で定義されるエントロピーを用いて、名詞句「 NP_1 の N_2 」で N_2 を名詞 n に固定したときの NP_1 の意味範疇の散らばりを定量化する。ただし、 NP_1 の意味範疇は、 NP_1 の主辞の名詞 N_1 の意味範疇と同じものとする。

☆ EDR コーパス中で、名詞句「 NP の N 」15,138 組に対して、統語構造が「 CN の N 」という名詞句が 1,013 組 (6.7%), 「 T の N 」という名詞句が 14,125 組 (93.3%) であった。

☆☆ ただし、横軸の意味範疇の集合は、後述する分類語彙表から構成した意味範疇の集合である。

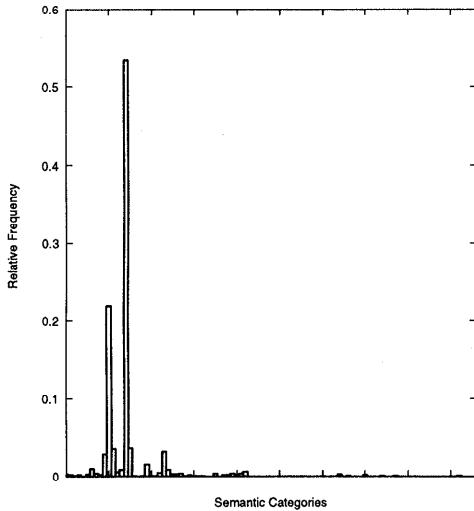


図2 「母」の意味範疇の散らばり

Fig. 2 Scattering of semantic category of "mother".

定義1（意味範疇の散らばり）

意味範疇の集合を $C = \{c_1, c_2, \dots, c_m\}$ とする。名詞句「 NP_1 の N_2 」において N_2 を名詞 n に固定したときの NP_1 の意味範疇の散らばりは次式で計算される^{*}。

$$\mathcal{H}_C(n) = - \sum_{c \in C} Pr(c|n) \log_m Pr(c|n) \quad (12)$$

ただし、

$$Pr(c|n) = \frac{\sum_{\substack{n';c \text{ の下位語} \\ c \in C}} f(\langle n', n \rangle)}{\sum_{\substack{c \in C}} \sum_{\substack{n';c \text{ の下位語}}} f(\langle n', n \rangle)}$$

また、 $f(\langle n', n \rangle)$ は、主辞が n' である名詞句 np と n がコーパス中で「 np の n 」という形で出現した頻度を表す。

エントロピーは、情報理論において不確実性を表すものである。 $\mathcal{H}_C(n)$ は、名詞句「 NP_1 の N_2 」という形の名詞句が出現し、 $N_2 = n$ のみを観測した場合に、 NP_1 の主辞 N_1 の意味範疇がどの程度予測できるかという、その不確実性を表す。 N_2 が関係名詞であれば、出現する N_1 の意味範疇はある程度小数に集中する傾向があると考えられるので、 N_1 の意味範疇の予測の不確実性は減り、式(12)の値は小さくなる。また、 N_2 が普通名詞であれば、 N_1 の意味範疇の散らばりは大きくなる傾向があると考えられるので、 N_1

の意味範疇の予測の不確実性は増し、式(12)の値は大きくなる。このように、式(12)は N_2 を固定したときの N_1 の意味範疇の散らばりの度合いに対応する。

先の図1、図2の例では、 NP_1 の分布が、「母」では特定の意味範疇に集中しているのに比べ、「車」の方は散らばっている傾向が見られる。次節の実験データを用いて、 $\mathcal{H}_C(\text{車})$ 、 $\mathcal{H}_C(\text{母})$ を計算すると、 $\mathcal{H}_C(\text{車}) = 0.646$ 、 $\mathcal{H}_C(\text{母}) = 0.382$ となり、 $\mathcal{H}_C(n)$ が意味範疇の散らばりの度合いに対応していることが分かる。

4. 実験

3章で、名詞句「 NP_1 の N_2 」を考えた場合、 N_2 が普通名詞か関係名詞かによって、 NP_1 の意味範疇の散らばりが異なる傾向があることを説明した。統いて、普通名詞と関係名詞を分類するために、名詞句「 NP_1 の N_2 」という形の共起情報を用いて、 NP_1 の意味範疇の散らばりを量化した。そこで、式(12)の量化に基づいて意味範疇の散らばりを計算し、その値を普通名詞および関係名詞の特徴量ととらえ、普通名詞と関係名詞の分類実験を行った。

4.1 実験データ

本実験のデータとして、まず、EDR コーパス⁵⁾および RWC テキストデータベース⁶⁾より「 NP_1 の N_2 」という形の名詞句を抽出し、それらの名詞句から「 NP_1 の主辞、 N_2 」という共起情報を得た。その結果、133,102組の共起情報を得た。その共起情報の中で「 $*, n$ 」という形で100回以上出現した普通名詞556個、関係名詞457個を人手で取り出した。これらのデータを本実験での分類の対象とした。

また、意味範疇は、シソーラス中の概念を基に設定した。本実験では、シソーラスとして、分類語彙表²⁾を用いた。分類語彙表は、6層の階層構造から構成され、最下層には全部で約800強の概念が設定されている。本実験では、意味範疇の集合 C を分類語彙表のトップの概念（「体の類」、「用の類」、「相の類」、「その他の類」）から深さ2の概念の集合 ($|C| = 96$) とした。

4.2 k -NN 推定法を用いた分類手法

量化した意味範疇の散らばり $\mathcal{H}_C(n)$ を名詞 n の特徴量として、文献9)と同様の手法 (k -NN 推定法を用いた Bayes 決定法) により普通名詞と関係名詞の分類を試みた。

名詞句「 NP_1 の N_2 」において N_2 を n に固定したときの NP_1 の意味範疇の散らばりが $\mathcal{H}_C(n)$ である場合に、 n の統語範疇が $Cat \in \{CN, RN\}$ である確率 $Pr(Cat|\mathcal{H}_C(n))$ を計算し、これを最大とするよ

* 通常、エントロピーは、 \log の底を2とするが、 \log 関数は、底が1より大きければ単調増加関数なので、本論文では意味範疇の数(m)とした。よって、式(12)の最大値は1である。 $Pr(c_i|n) = 1/m$ ($1 \leq i \leq m$) の場合である。

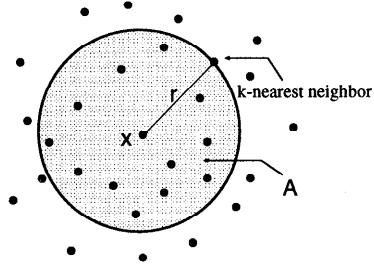


図3 k -NN 推定法
Fig. 3 k -Nearest Neighbor estimation.

うな統語範疇 Cat を名詞 n の統語範疇と決定する (Bayes 決定法). Bayes の定理より,

$$Pr(Cat|\mathcal{H}_C(n)) = \frac{Pr(Cat)Pr(\mathcal{H}_C(n)|Cat)}{Pr(\mathcal{H}_C(n))}$$

であり, 分母は Cat に依存しないので, $Pr(Cat|\mathcal{H}_C(n))$ を最大とする Cat を求めるには,

$$Pr(Cat)Pr(\mathcal{H}_C(n)|Cat) \quad (13)$$

を最大とする Cat を求めればよい. ここで, 式(13)の第1項 $Pr(Cat)$ は, 次のように最尤推定する.

$$Pr(Cat) \simeq \frac{Cat \text{ 中の名詞の異なり語数}}{\text{全名詞の異なり語数}} \quad (14)$$

本実験では, 普通名詞, 関係名詞のみの名詞の集合よりランダムに約1,000個の名詞をサンプリングし, 人手で普通名詞か関係名詞かを調べ, $Pr(Cat)$ を求めた. その結果, $Pr(CN) = 0.72$, $Pr(RN) = 0.28$ としている.

第2項の $Pr(\mathcal{H}_C(n)|Cat)$ は, ノンパラメトリックな確率密度推定法である k -NN 推定法¹⁾により推定する. まず, k -NN 推定法の定義を述べる.

定義2 (k -NN 推定法)

大きさ N のサンプル \mathcal{S}^N における x の確率密度の推定値は,

$$\frac{k-1}{N} \frac{1}{A(k, \mathcal{S}^N, x)}$$

である. ただし, $A(k, \mathcal{S}^N, x)$ は x と \mathcal{S}^N における x の k -nearest neighbor との距離 $r(k, \mathcal{S}^N, x)$ を半径とする超球の体積である.

定義から分かるように k -NN 推定法では, x における確率密度を, x を中心とする半径 $r(k, \mathcal{S}^N, x)$ の超球におけるサンプルの平均の密度として推定する(図3). このような k -NN 推定法は, 推定法に要請される重要な性質の1つである一致が成立しており, サンプルが大きくなれば確率密度の推定誤差を限りなく小さくすることができるという性質を持っている.

Cat のサンプルを \mathcal{S}_{Cat} とすると $Pr(\mathcal{H}_C(n)|Cat)$

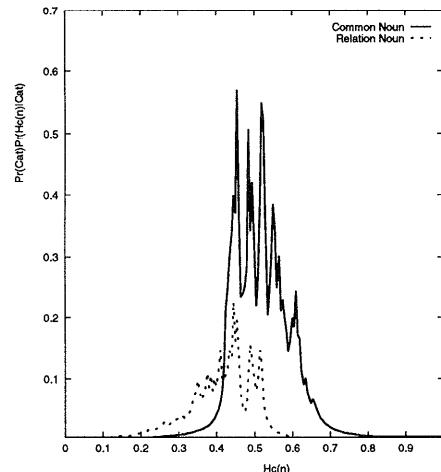


図4 意味範疇の散らばり $\mathcal{H}_C(n)$ と $Pr(Cat)Pr(\mathcal{H}_C(n)|Cat)$
Fig. 4 Scattering of semantic categories $\mathcal{H}_C(n)$ and $Pr(Cat)Pr(\mathcal{H}_C(n)|Cat)$.

表1 正しく分類された関係名詞の例
Table 1 Example of nouns classified to relation noun correctly.

頻出した NP_1 の意味範疇	名詞
「空間・場所」	「風速」「気圧」
「成員・職」	「母」「弟」
「われ・かれ」	「略歴」「氏名」
「心」	「骨子」「核心」
「位置・地点」	「遺物」「風土」

は k -NN 推定法を用いて,

$$Pr(\mathcal{H}_C(n)|Cat) \simeq \frac{k-1}{|\mathcal{S}_{Cat}|} \frac{1}{A(k, \mathcal{S}_{Cat}, \mathcal{H}_C(n))} \quad (15)$$

と推定される(実験では $k = 20$ とした).

よって, 普通名詞と関係名詞の決定は, 共起情報から決定したい名詞 n の意味範疇の散らばり $\mathcal{H}_C(n)$ を計算し, 式(14), (15)より, $Pr(Cat)Pr(\mathcal{H}_C(n)|Cat)$ を計算, この値の高い Cat を名詞 n の属す統語範疇と決定する.

4.3 実験結果

横軸を $\mathcal{H}_C(n)$, 縦軸を $Pr(Cat)Pr(\mathcal{H}_C(n)|Cat)$ としたときのグラフを図4に示す. このときの正解率は, 79.2% であった. 正しく分類された関係名詞の例を表1に示す.

誤分類された関係名詞を見てみると, 「構造」「姿勢」「要因」「概念」「特性」「実力」「謎」などがあった. 正しく分類されている関係名詞について, その関係名詞 n が表す関係 (R_n) の補項がとりうる領域(本論文では, この領域を R_n の定義域と呼ぶことにする)は, 人間が考えた場合でも明解なものとなっている傾向がある.

たとえば、「風速」や「気圧」といったものは、 $R_{\text{風速}}$ の定義域がちょうど意味範疇《空間・場所》に対応している。逆に、誤分類された「構造」という関係名詞では、「構造」が NP_2 のとき NP_1 には、《生産物》や《組織》といった意味範疇に属す名詞句が頻出しそうであるが、実際は、具体物に限らず「腐敗政治の構造」や「近代社会の構造」というように、 NP_1 に抽象的なものを指示する名詞句も多く出現していた。「構造」という関係名詞では、意味範疇の散らばりが大きくなるが、これは分類語彙表の概念の分類体系では、「 $R_{\text{構造}}$ 」の定義域が小数の意味範疇に対応できていないことによる。他の誤分類された関係名詞についても同様のことがいえた。

4.4 考 察

普通名詞と関係名詞は、従来、その違いを定量的に扱うことが困難であったために、計算機を用いた自動分類の試みはなされていない。本実験では、普通名詞と関係名詞に対する統計的性質から導いた意味範疇の散らばりが、普通名詞と関係名詞を分類するための特徴量として有効かどうかを確かめるものである。

本実験で用いた Bayes 決定法は、全体の誤認識率を最も低くなるように分類する。つまり、本実験では次のような判定手法で普通名詞と関係名詞の決定を行つたことになる。

判定手法 1 (Bayes 決定法に基づいた場合)

$$\begin{aligned}\mathcal{H}_C(n) > \zeta &\Rightarrow n \text{ は } CN \\ \mathcal{H}_C(n) < \zeta &\Rightarrow n \text{ は } RN \\ \text{otherwise} &\Rightarrow \text{エラー}\end{aligned}$$

ただし、 ζ は図 4 のグラフにおいて、 $Pr(CN)Pr(\zeta|CN) = Pr(RN)Pr(\zeta|RN)$ となる ζ である。

この普通名詞と関係名詞の分類問題は、 $Pr(CN) > Pr(RN)$ より、正解率は最低でも $Pr(CN)$ の 72% を超えなくては意味がない。判定手法 1 による分類の結果では、正解率は $Pr(CN)$ より 7.2% 上昇しており、意味範疇の散らばりが普通名詞と関係名詞に対する特徴量の 1 つとして有効であるといえる。しかし、普通名詞と関係名詞を完全に自動分類するシステムを想定した場合には、分類の精度からも分かるように、優れた特徴量であるとはいえない。また、今回の普通名詞と関係名詞の分類は、名詞句の意味処理を行ううえで基本的な言語知識として必要となるものであるから、より高い精度での分類が望まれる。

そこで、意味範疇の散らばりを用いた分類に対して、次のような判定手法を考える。これは、判定手法 1 の閾値 (ζ) に対して普通名詞と関係名詞の決定を保留す

る領域を設定するものである。

判定手法 2 (未決定領域を設けた場合)

$$\begin{aligned}\mathcal{H}_C(n) > \zeta + \varepsilon_c &\Rightarrow n \text{ は } CN \\ \mathcal{H}_C(n) < \zeta - \varepsilon_r &\Rightarrow n \text{ は } RN \\ \text{otherwise} &\Rightarrow \text{未決定}\end{aligned}$$

ただし、 ζ は判定手法 1 の ζ と同じで、 $0 \leq \varepsilon_c < 1 - \zeta$ かつ $0 \leq \varepsilon_r < \zeta$ である。

この判定手法 2 は、普通名詞と関係名詞の半自動分類システムに応用できる。判定手法 2 において、未決定となった名詞は、人手で判定する。未決定となる名詞は、人手で判定するのですべて正しく決定できると仮定する。未決定となった名詞の数を R 、自動分類で正しく決定できた名詞の数を C 、分類する名詞の数を U で表すと、判定手法 2 の正解率は、次のようになる。

$$\text{正解率} = \frac{R + C}{U} \quad (16)$$

名詞の分類の精度を高くしたい場合には、 ε_c 、 ε_r を大きくとればよいことになる。しかしながら、このように ε_c 、 ε_r を操作すると、当然未決定となる名詞の割合、すなわちリジェクト率が高くなる。リジェクト率は、次のように計算される。

$$\text{リジェクト率} = \frac{R}{U} \quad (17)$$

式 (16) の正解率と式 (17) のリジェクト率は、負の相関関係にある。よって、判定手法 2 では、分類結果に求められる精度および分類に掛けられるコストに応じてリジェクト率および ε_c 、 ε_r を決める必要がある。一例として、先の実験データに対して $\varepsilon_c = 0.05$ 、 $\varepsilon_r = 0$ とした場合の判定手法 2 の正解率とリジェクト率を示す。

未決定となった名詞の数：253 個

自動分類で正しく決定できた名詞の数：661 個
より、

正解率：90.2%

リジェクト率：25.0%

となった。

意味範疇の散らばりを普通名詞と関係名詞の特徴量とし、判定手法 1 によって分類実験では、その分類精度は $Pr(CN)$ よりもわずかに上昇した。以上から、意味範疇の散らばりは、普通名詞と関係名詞の統計的性質を少なからず反映していると考えられる。しかし、十分な精度で自動分類できるほどには、普通名詞と関係名詞を特徴づけることはできおらず、判定手法 1 の拡張として、未決定領域を設けた判定手法 2 を示

した。判定手法 2 は、意味範疇の散らばりが偏った値をとる名詞のみを自動分類し、未決定となる名詞を人手で判定するような、普通名詞と関係名詞の半自動分類システムに援用することができる。

5. おわりに

本研究では、名詞句「 NP_1 の NP_2 」の意味構造推定に必須となる名詞の統語範疇を、意味範疇の散らばりに着目し統計的に決定する手法を示した。

現在、名詞句の表層表現中には現れないが NP_1 , NP_2 間に仮定される意味関係の獲得および推定法について検討している。

参考文献

- 1) Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, Academic Press Inc., Orlando (1972).
- 2) 国立国語研究所：分類語彙表，秀英出版 (1964).
- 3) Montague, R.: *The Proper Treatment of Quantification in Ordinary English, Approaches to Natural Language*, Reidel, Dordrecht, Hintikka, J., Moravcsik, J. and Suppes, P. (Eds.), pp.221-242 (1974).
- 4) Nakamura, T., Tomiura, Y. and Hitaka, T.: Semantic Validity of Japanese Noun Phrases with Adnominal Particles, *Proc. PRICAI'92* (1992).
- 5) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1995).
- 6) 新情報処理開発機構：RWC テキストデータベース報告書 (1996).
- 7) 寺村秀夫：日本語のシンタクスと意味 III, くろしお出版 (1991).
- 8) 富浦洋一, 中村貞吾, 日高 達：名詞句「 NP の NP 」の意味構造, 情報処理学会論文誌, Vol.36, No.6, pp.1441-1448 (1995).
- 9) 富浦洋一, 日高 達： k -NN 推定法に基づく統語的あいまいさの解消法, 電子情報通信学会論文誌, D-II, Vol.J80, No.9, pp. 2475-2481 (1997).
- 10) 吉田 将, 日高 達, 稲永紘之, 田中武美, 吉村賢治：公用データベース日本語単語辞書の使用について, 九州大学大型計算機センター広報, Vol.16, No.4 (1983).



田中 省作（学生会員）

昭和 47 年生。平成 7 年岡山大学工学部情報工学科卒業。平成 9 年九州大学大学院システム情報科学研究科知能システム学専攻修士課程修了。現在同大学院システム情報科学研究科知能システム学専攻博士後期課程在学中。平成 9 年度電気学会論文発表賞受賞。自然言語処理、計算言語学に関する研究に従事。



富浦 洋一（正会員）

昭和 36 年生。昭和 59 年九州大学工学部電子工学科卒業。昭和 61 年同大学院工学研究科電子工学専攻修士課程修了。平成元年同大学院工学研究科電子工学専攻博士後期課程単位取得退学。同年九州大学工学部助手、平成 7 年同助教授、現在同大学院システム情報科学研究科助教授。工学博士。平成 3 年度情報処理学会研究賞受賞。自然言語処理、計算言語学、人工知能に関する研究に従事。人工知能学会、電子情報通信学会、言語処理学会各会員。



日高 達（正会員）

昭和 14 年生。昭和 40 年九州大学工学部電子工学科卒業。昭和 42 年同大学院工学研究科電子工学専攻修士課程修了。昭和 44 年同大学院工学研究科電子工学専攻博士後期課程中退。同年九州大学工学部助手、昭和 48 年同講師、昭和 55 年同助教授、昭和 63 年同教授、現在同大学院システム情報科学研究科教授。工学博士。形式言語の方程式論、自然言語処理、手書き文字認識の研究に従事。電子情報通信学会、人工知能学会、言語処理学会各会員。

(平成 10 年 9 月 16 日受付)

(平成 11 年 6 月 3 日採録)