自 然 言 語 処 理 107−11
（1995. 5. 26）

実例に基づいた入力文と格フレームの類似度

内山将夫† 板橋秀一‡

† 筑波大学 大学院 工学研究科
‡ 筑波大学 電子・情報工学系

概要

単語と単文の類似度に情報量を用いることを提案した．単語同士が類似するとは，それらが同一の単語集合に属することであると定義した．その単語集合を切り出すためにシソーラスを用いることを述べた．このとき，“単語間類似度 = 単語集合全体のエントロピー − 切り出された単語集合のエントロピー”である．また，単文の類似度は，近似的には単語の類似度の和である．

次に，名詞を，それと共起する格フレームの集合として定義した．これにより，単語同士の類似度と同様な手法で，名詞と格フレームとの類似度が処理できることを述べた．

The example-based similarity between a sentence and a case frame.

Masao UTIYAMA† Shuichi ITAHASHI‡

† Doctoral Program in Engineering, University of Tsukuba.
‡ Institute of Information Sciences and Electronics, University of Tsukuba

Abstract

This paper describes the use of the quantity of information to compute the similarity between words/sentences. Two words are similar when they are both in a word subset, which is obtained utilizing a thesaurus. The similarity R between the words is:

$$R = (\text{the entropy of the word set}) - (\text{the entropy of the subset}).$$

The similarity between sentences is the sum of the word similarities.

This paper also describes the similarity between a noun and a case frame. By defining a noun as the set of the case frames which co-occur with the noun, we can compute the similarity between them by the same method.

1 はじめに

コーパスを利用した自然言語処理が活発である [1]. その処理のための基本技術として，言語表現の類似性を判定するというものがある [2]. 本稿では，その中でも，単文と単文，単文と格フレームとの類似性について述べる.

黒橋ら [3] は，格フレームの選択に実例とシソーラスを利用する手法の有効性を示した．そして，同時に，入力文と実例との類似度を評価する関数も提案した．本稿では，シソーラスを情報理論的に解釈し，評価関数を導出する．第2章で単語の類似度について述べ，第3章では，単文の類似度について述べる．そして，これらの類似度と黒橋らの評価関数とを比較する．最後に，第4章で，名詞と格フレームとの共起関係を評価関数に取り入れる方法を示す.

2 単語の類似度

まず，単語の集合 U における距離と類似度を定義する．次に，単語の集合を分割するときの指針としての多分肢系統樹 [4] を導入する．すると，黒橋らの単語間類似度は，多分肢系統樹 T により分割される U の部分集合 $V(T)$ 上の類似度に対応する.

2.1 距離と類似度

単語集合 U の部分集合 V における距離と類似度を定義する．まず，単語の集合 U を考える．U の要素 u_i には確率 $p(u_i)$ が与えられている.

$$\begin{aligned} p(u_i) &\geq 0, \\ \sum_{u_i \in U} p(u_i) &= 1. \end{aligned}$$

また，U の部分集合 V の確率は次のものである.

$$p(V) = \sum_{v_j \in V} p(v_j).$$

すると，V のエントロピー，あるいは，不確定性 $S(V)$ は次式である.

$$S(V) = -\sum_{v_j \in V} \frac{p(v_j)}{p(V)} \lg \frac{p(v_j)}{p(V)}. \qquad (1)$$

ただし，lg は 2 を底とする対数.

このとき，単語 x と単語 y とが，U において類似するとは，x と y が U の部分集合 V に含まれることであると定義する．そして，V における x と y との距離 $K(x,y,V,U)$ と類似度 $R(x,y,V,U)$ を以下のように定義する.

$$\begin{aligned} K(x,y,V,U) &= S(V), \\ R(x,y,V,U) &= S(U) - S(V). \end{aligned} \qquad (2)$$

$K(x,y,V,U)$ は，x と y が V に含まれていることが確定した時点において，未だ，存在する不確定性である．$R(x,y,V,U)$ は，x と y が V に含まれていることが確定したことにより獲得された情報量 (減少した不確定性) である.

次に，相対的距離 $SK(x,y,V,U)$ と相対的類似度 $SR(x,y,V,U)$ を次式で定義する.

$$\begin{aligned} SK(x,y,V,U) &= K(x,y,V,U)/S(U), \\ SR(x,y,V,U) &= R(x,y,V,U)/S(U). \end{aligned}$$

なお，$0 \leq SK(x,y), SR(x,y) \leq 1$.

2.2 多分肢系統樹

集合から部分集合を選ぶための指針として多分肢系統樹 [4] を導入する．多分肢系統樹は，集合を再帰的に分割する仕方を規定する．1 回の分割において，元の集合 $E = B^{(0)}$ と分割後の集合 $B_1^{(1)}, B_2^{(1)}, ..., B_m^{(1)}$ とに (3) 式の関係が成り立つ場合に，その分割を“集合 $B^{(0)}$ の m 度の多分肢形”と呼ぶ.

$$\begin{aligned} &B^{(0)} = B_1^{(1)} \cup B_2^{(1)} \cup ... \cup B_m^{(1)}, \\ &B_\mu^{(1)} \cap B_\nu^{(1)} = \phi \qquad (\mu \neq \nu), \\ &B_\mu^{(1)} \neq \phi. \end{aligned} \qquad (3)$$

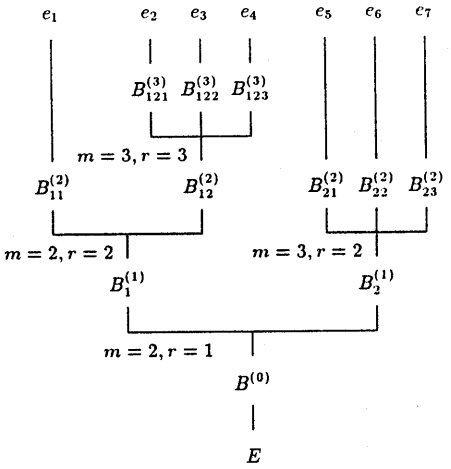
部分集合 $B_\mu^{(1)}$ は，それぞれ，さらに多分肢形をとることができる．この分割を，多分肢形の部分集合に含まれる要素数が 1 となるまで続けたときの分割の過程を表わす木を多分肢系統樹と呼ぶ (図 1).

多分肢系統樹には，B の集合に相当する“枝”があって，それぞれ“分岐点”でつながっている．それぞれの枝 B には，ランク数 $r(B)$ が付随する．ランク数 $r(B)$ とは，幹 $B^{(0)}$ から B に到達するときに辿る分岐点の数である．また，B を分割する分岐点には，その分割の度数 $m(B)$ が割当てられる.

2.3 多分肢系統樹の情報量

多分肢系統樹の情報量に関して (4) 式が成立する [4].

$$I_\mu^{(1)} + I_{\mu\nu}^{(2)} + ... + I_{\mu\nu\kappa\lambda...\omega}^{(r)}$$



$E=\{e_1,e_2,...,e_7\}$ の多分肢系統樹．$B_{\mu\nu\kappa\lambda...\omega}^{(r)}$ において，r は B のランク数であり，$\mu\nu\kappa\lambda...\omega$ は，それぞれの分岐において選択された枝の番号である．
(文献 [4] より修正して引用)

図 1: 多分肢の度数 (m) とランク数 (r)

$$\begin{aligned} &= (S^{(0)}-S_{\mu}^{(1)})+(S_{\mu}^{(1)}-S_{\mu\nu}^{(2)})+... \\ &\quad +(S_{\mu\nu\kappa\lambda...\rho}^{(r-1)}-S_{\mu\nu\kappa\lambda...\omega}^{(r)}) \\ &= S^{(0)} \end{aligned} \tag{4}$$

ただし，$I_{\mu\nu...o}^{(i)}$ は i 番目の分岐において獲得される情報量である．また，r は多分肢系統樹の末端の集合のランク数[1]であり，$S_{\mu\nu...o}^{(i)}$ は，(1) 式により計算される，集合 $B_{\mu\nu...o}^{(i)}$ のエントロピーである．つまり，(4) 式は，多分肢系統樹に従って分岐を行い，一つの要素を得たなら，そのときの獲得情報量は分岐の各段階の獲得情報量の和であることを示している．

また，各段階 i の分岐において，その度数が $m^{(i)}$ であるとき，その分岐における獲得情報量 $I^{(i)}$ は，

$$I_{\mu\nu...o}^{(i)} \leq \lg m^{(i)}$$

である．したがって，ランク数が r，各段階の分岐の最大度数が m の多分肢系統樹では，

$$S^{(0)} \leq r \lg m$$

が成立する．

[1]多分肢系統樹の末端の集合のランク数の最大値を多分肢系統樹のランク数とする．

2.4 多分肢系統樹における距離と類似度

"多分肢系統樹 T における単語 x と y との距離・類似度" は，T に従って単語集合 U の分割を進め，x と y の両方を含む最小の部分集合 $V(T)$ を求めることにより得られる．このときの $V(T)$ のエントロピー $S(V(T))$ が距離であり，$S(U)-S(V(T))$ が類似度である (図 2)．

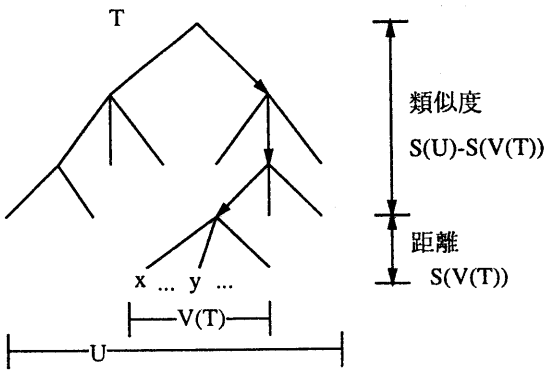


図 2: 多分肢上の類似度

従来の研究では，シソーラス上の距離として，共有する親ノードのレベルが 1 レベルあがると，距離が 1 単位[2]増えるというものが使われてきた [2]．これは次のように定式化できる．

もし，各分岐において度数が等しく m であり，各単語に割り当てられた確率が等しいなら，多分肢系統樹のランク数を r とすると，$S(U) = r \lg m$ である．さらに，集合 $V(T)$ を得るまでの分岐の回数を r_1 とする，すると，$R(x,y,V(T),U) = r_1 \lg m$，$K(x,y,V(T),U)=(r-r_1)\lg m$ であるから，

$$\begin{aligned} SR(x,y,V(T),U) &= \frac{R(x,y,V(T),U)}{S(U)} = \frac{r_1}{r}, \\ SK(x,y,V(T),U) &= \frac{K(x,y,V(T),U)}{S(U)} = \frac{r-r_1}{r}. \end{aligned}$$

したがって，共有する親ノードのレベルが 1 上がれば，相対的距離は $1/r$ だけ増える．つまり，従来の研究で使われてきた距離は，分岐の度数が全て等しく，シソーラスの それぞれの葉に至る分岐の回数が一定

[2]単位の例として，ランク数の逆数がある．

で，かつ，全ての単語に等確率が割当てられているときに，$SK(x,y,V(T),U)$ と同等となる．

2.5 黒橋らの単語間類似度との比較

黒橋ら [3] は，単語間類似度を，分類語彙表 [5] での，共有する親ノードのレベルにより定義している (表1：黒橋)．分類語彙表は，分類語彙表に採録されている語彙集合の分割を規定する，多分肢系統樹と考えられる．したがって，各レベルまでノードを共有する場合の類似度を (2) 式により求めることができる．表1の R_1 は，分類語彙表の語彙全てに等確率を割り振ったときの類似度であり，R_2 は，国立国語研究所の語彙調査 [6] での頻度に比例する確率を各語彙に割り当てたときの類似度である．黒橋らの類似度との相関係数は，それぞれ，0.95 と 0.97 である．

level	0	1	2	3	4	5	6	一致
黒橋	0	0	5	7	8	9	10	11
R_1	0	1.2	3.1	5.8	8.4	9.4	12.0	15.2
R_2	0	1.2	2.8	5.4	7.3	7.9	9.2	10.3

表 1: 単語間類似度

3 単文の類似度

単文の類似度は単語の類似度を情報量を元に統合することにより得られる．まず，単語集合 U から，一つの単語 u_i を取り出すという事象を U で表す．すると，複数の単語集合 $U,V,...,Z$ から，一つずつ単語を取り出すという事象が考えられる．これを結合事象と呼び，$U \otimes V \otimes ... \otimes Z$ で表す．結合事象のエントロピー $S(U \otimes V \otimes ... \otimes Z)$ は次式である．

$$\begin{aligned} & S(U \otimes V \otimes ... \otimes Z) \\ &= S(U) + S(V) + ... + S(Z) - J(U \otimes V \otimes ... \otimes Z) \end{aligned}$$

ここで，$J(U \otimes V \otimes .. \otimes Z)$ は有機性 (冗長度) と呼ばれる量であり，集合 $U,V,...,Z$ の間の確率的な関係の強さ (≥ 0) を示す．$U,V,...,Z$ が，互いに，確率的に独立であるとき，$J(U \otimes V \otimes .. \otimes Z) = 0$ である．

単文 (や格フレーム) は，複数の格と一つの動詞からなる．したがって，これを，それぞれの格に対応する単語集合から一つずつ単語を選び出すという，結合事象と考えることができる (動詞は固定)．すると，同じ動詞からなる二つの単文の類似度を，結合事象の情報量を用いて定義できる．

二つの単文 X と Y が，それぞれ n 個の格 $x_1,x_2,...,x_n$ と $y_1,y_2,...,y_n$ からなり，x_i と y_i が格 C_i の要素であるとする．次に，格 C_i には，単語集合 C_i と多分肢系統樹 T_i が付随するとする．すると，格 C_i について，x_i と y_i を含む最小の部分集合 $V(T_i)$ が求まるので，それを利用して，x_i と y_i の類似度と距離を求めることができる．

このとき，結合事象のエントロピーは次式である．

$$\begin{aligned} & S(C_1 \otimes C_2 \otimes ... \otimes C_n) \\ &= \sum_{i=1}^{n} S(C_i) - J(C_1 \otimes C_2 \otimes ... \otimes C_n) \\ &\leq \sum_{i=1}^{n} S(C_i) \\ &= \sum_{i=1}^{n} (R(C_i) + K(C_i)) \\ &= \sum_{i=1}^{n} R(C_i) + \sum_{i=1}^{n} K(C_i). \end{aligned}$$

ただし，$R(C_i) = R(x_i,y_i,V(T_i),C_i)$，$K(C_i) = K(x_i,y_i,V(T_i),C_i)$．

したがって，単文 X と Y の類似度と距離とを，単語の類似度・距離と同様に，獲得情報量と残された不確定性とで表現するなら，単文の類似度は単語の類似度の和，距離は単語の距離の和で表せる．ただし，有機性を考慮していないため，これらの和は近似となる．同様に，相対的類似度 $SR(X,Y,C_1 \otimes ... \otimes C_n)$，相対的距離 $SK(X,Y,C_1 \otimes ... \otimes C_n)$ は次式で近似できる．

$$\begin{aligned} SR(X,Y,C_1 \otimes ... \otimes C_n) &= \sum_{i=1}^{n} R(C_i) / \sum_{i=1}^{n} S(C_i), \\ SK(X,Y,C_1 \otimes ... \otimes C_n) &= \sum_{i=1}^{n} K(C_i) / \sum_{i=1}^{n} S(C_i). \end{aligned} \quad (5)$$

3.1 黒橋らの評価関数との比較

黒橋らは，二つの単文の類似度 H を次式により求めている．

$$H = (\text{対応格要素の類似度の和}) / n^{1/2} \times (\frac{n}{l})^{1/2} \times (\frac{n}{m})^{1/2}$$

ただし，l と m は，それぞれの文の格要素数であり，n は二文間で対応付けが得られた格要素数である[3]．ここで，H の両辺に関して lg を取り，整理すると次式になる．ただし，sum は，対応格要素の類似度の和である．

$$\lg H = \lg sum - \frac{1}{2}(\lg l + \lg m - \lg n) \qquad (6)$$

次に，(5) 式の SR を同様に変形する．このとき，二つの文を合せた全格要素数は $l + m - n$ である．したがって，SR は次式になる．

$$SR = \sum_{i=1}^{n} R(C_i)/(l + m - n)S$$

$R(C_i)$ は，対応付けられた n 個の要素についてのみ 0 以上の値をとり，また，全格要素に同一の多分肢系統樹を適用する場合には，$S(C_i)$ は一定なので，S と置けるからである．次に，両辺の lg を取ると，

$$\lg SR = \lg SUM - \lg(l + m - n)S \qquad (7)$$

ただし，$SUM = \sum_{i=1}^{n} R(C_i)$.

(6) 式と (7) 式とは，対応付けられた格要素数が多いほど大きい値を取るという点で似ている．

3.2 単文の有機性

もし，情報量による評価関数が有効ならば，単文の有機性を類似度の評価に利用できる．

類似度とは，本稿では，獲得情報量である．獲得情報量は，初期の不確定性 (エントロピー) と残りの不確定性との差である．このうち，初期の不確定性は決まっている．それは，それぞれの格の単語集合の不確定性の和である．したがって，残りの不確定性が最も小さくなる単文同士が最も類似したものとなる．よって，例えば，単文 $A \cdot B \cdot C$ について，A と B，A と C の類似度を比較するときに，格ごとでは，a_i と b_i，a_i と c_i の類似度に差がない場合でも，単文としての類似度には差がでる場合がある．例えば，共起的な制約により，エントロピーが更に減少することが考えられる．これは，個々の格要素が確率的に独立ではないときに利用できる有機性の例である．

4 共起関係を利用した評価関数

情報量を利用した評価関数が有効であるという仮定のもとで，名詞 n_{ij} と格フレーム f_{ijk} との共起関係を利用した評価関数 I について考察する．この評価関数 I は，ある名詞が入力されたことにより減少した，格フレーム選択に際しての不確定性の度合を示し，その値が大きいほど名詞と格フレームとの整合性が高いことを示す．

まず，名詞の集合 C_i は それぞれの格 C_i ごとに設定する．そして，名詞を，当該の格において，その名詞が共起した格フレームの集合として定義する[4]．つまり，

$$n_{ij} = \{f_{ij1}, f_{ij2}, ..., f_{ijk}, ...\}.$$

すると，特定の格の名詞集合 C_i を構成する要素は，名詞と格フレームとの二つ組である．

$$C_i = \{< n_{i1}, f_{i11} >, < n_{i2}, f_{i12} >, ..., < n_{ij}, f_{ijk} >, ...\}$$

それぞれの二つ組には，生起頻度に比例した確率を割り当てる．次に，名詞 n_{ij} が入力されたとする．すると，n_{ij} を含む二つ組からなる集合 $V_i = \{< n_{ij}, f_{ij1} >, ..., < n_{ij}, f_{ijk} >, ...\}$ が得られる．このとき，格 C_i における，格フレーム f_{ijk} に対する，名詞 n_{ij} の格要素としての適切さを示す評価関数 $I(n_{ij}, f_{ijk}, C_i)$ は次式となる．

$$I(n_{ij}, f_{ijk}, C_i) = \begin{cases} \frac{S(C_i) - S(V_i)}{S(C_i)} & f_{ijk} \in V_i \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

入力文と格フレームとの一致の程度を示す評価関数は (5) 式と同様にして求める．

現実的には，入力された名詞が，入力文中の動詞の特定の格フレームと，前もって，共起関係にあるとは限らない．その場合には，前 2 章でのべたように，適当なシソーラスを用いて，名詞の集合から適当な部分集合を切り出す．このときでも，その部分集合は名詞と格フレームとの二つ組からなるので，同様な方法により評価値を計算できる．入力された名詞と入力文中の動詞の格フレームとが共起関係にあるかないかの差は，切り出される部分集合の大きさに現われる．このように，情報量を評価関数に用いることにより，意味的な制約と構文的な制約とを融合できる．

4.1 適用例

IPAL[7] 中の名詞について，それぞれの格ごとに，名詞と動詞の意味分類 (図 3：(1)-(17)) との二つ組を

[3] 簡単のため，必須格のみを考える．

[4] 前 2 章においては最小の構成要素であった名詞に構造を導入したことになる．

作った．その二つ組の集合について，次の三つの事項を述べる．

- それぞれの格の名詞集合のエントロピー，
- 同一の格における異なる名詞のエントロピー，
- 異なる格における同一の名詞のエントロピー．

ただし，名詞集合とは二つ組の集合であり，名詞のエントロピーとは，その名詞を含む二つ組からなる集合のエントロピーである．

それぞれの格のエントロピー

表2に，それぞれの格の名詞集合のエントロピーを示す．この値は，格の用法の多様さを示す値である．例えば，最もエントロピーの小さいヨリ格では，係りうる動詞の分類項目は，(2),(4),(5),(6),(11) の 4 項目に過ぎない．

格ごとの名詞集合のエントロピーは，その格の一般的な重要性を示すと考えられる．なぜなら，残された不確定性が等しいときには，格自体のエントロピーが大きいものほど，(8) 式により，評価値が高くなるからである．実際，表2では，大抵の動詞の必須格である“ヲ，ニ，ガ”のエントロピーが大きい．

同一格における異なる名詞のエントロピー

同一の格であっても，名詞が異なれば，共起する分類項目も異なるので，そのエントロピーは異なる．ガ格の例として，“彼女”と“スミスさんの日本語”とを比べてみる．“彼女”は，図3の分類項目のほぼ全てと共起し，エントロピーは 2.85 である．よって，(8) 式より“彼女”の評価値は $(11.35 - 2.85)/11.35 = 0.75$ となる．一方，“スミスさんの日本語”は，“(状態) 変化”と共起する[5]だけであるので，エントロピーは 0，評価値は 1 となる．つまり，同一の格を取る名詞であっても，それと共起する分類項目が広い名詞ほど，評価値は低くなる．

異なる格における同一名詞のエントロピー

ガ格では ほぼ全ての分類項目と共起する“彼女”も，ヲ格になると共起する分類項目の数が減る．図3の (4),(6),(10),(11),(12),(15),(16) の 7 項目のみである．エントロピーも 2.59 になり，評価値は 0.80 になる．つまり，同一の名詞であっても，格が異なれば，共起する分類項目も異なるため，評価値は異なる．共起する分類項目が広いほど評価値は低くなるといえる．

状態
- (1) 存在・所有
- (2) 関係認定
- (3) 単純状態

動作（動き）
- (4) (抽象的) 関係
- (5) 時間
- (6) (状態) 変化
  - 移動 (位置変化)
    - 出発・帰着
    - 授受
  - 出現・発生
  - 消滅
  - 生産
  - もようがえ
- (7) 設置 (とりつけ)
- (8) 離脱 (とりはずし)
  - (衣服) 着脱
- (9) 接触
- (10) 加力
- (11) 生理・心理
- (12) 知覚・思考
- (13) 発見
- (14) 経済活動
- (15) 社会活動
- (16) 言語活動
- (17) 自然現象

図 3: 動詞の意味的分類

以上，本節では，

- 格ごとの重み，
- 同一格における，異なる名詞の重要性の差，
- 同一名詞の，異なる格における重みの差，

が評価関数 I を利用して表現できることをみた．本節では，名詞と分類項目との二つ組について述べたのみ

[5]スミスさんの日本語は日本人に近付いてきた。

格	ヲ	ニ	ガ	デ	カラ	ト	ヘ	ϕ	ヨリ
エントロピー	12.77	12.08	11.35	10.40	9.81	9.51	8.92	7.48	5.46

表 2: 格ごとのエントロピー

であるが，格フレームとの二つ組に対しても同様なことが言えるであろう．

5 おわりに

本稿では，格フレームの選択に，情報量を用いた評価関数を利用することを提案した．そして，情報量を用いることにより，

- 単語と単文の類似度が統一的に扱えること，
- 意味的な制約 (シソーラス) と構文的な制約 (共起関係) とが融合できること，

を述べた．

今後の課題は，実験により，その有効性を示すことである．

参考文献

[1] 野美山, 浦本, 渡辺ほか. 「コーパスを利用した自然言語処理」サーベイ. 自然言語処理研究会 104-12, 情報処理学会, 1994.

[2] 飯田仁. 人口知能におけるスーパーコンピューティング. 情報処理, Vol. 36, No. 2, pp. 164–167, 1995.

[3] 黒橋禎夫, 長尾真. 格フレーム選択における意味マーカと例文の有効性について. 自然言語処理研究会 91-11, 情報処理学会, 1992.

[4] 著者：渡辺慧, 訳者：村上, 丹治. 知識と推測 1 情報の構造. 東京図書株式会社, 1975.

[5] 国立国語研究所. 分類語彙表 [フロッピー版], 国立国語研究所言語処理データ集, 第 5 巻. 秀英出版, 1994.

[6] 国立国語研究所. 中学校・高校教科書の語彙調査 [フロッピー版], 国立国語研究所言語処理データ集, 第 6 巻. 秀英出版, 1994.

[7] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL(Basic Verbs), 1987.