

# 7E-7 共起データを用いた単語の意味ネットワークの作成

松平正樹 山本由紀雄 坂本仁  
沖電気工業株式会社 関西総合研究所

### 1. はじめに

我々は、日英/英日機械翻訳システムをすでに商品化し、引き続きより高度な機械翻訳システムの研究を行っている。その一環として自然言語の共起や意味に関する研究を行っているが、今回、共起データを用いて単語の意味ネットワークの作成を試みたので報告する。

### 2. 基本的な考え方

共起データを用いて、例えば、

「雨が降る」、「雪が降る」、「みぞれが降る」

のように同じ単語(「降る」と共起関係にある単語(「雨」、「雪」、「みぞれ」)は意味的に近いと考えることができる。この考え方を利用してすべての単語間の意味的な近さを求め、共起データと統合して単語の意味ネットワークを作成することを考える。

### 3. 共起関係の強さ

意味的な近さを求める前に、まず、共起関係の強さというものを考える。

例えば、

「風邪をひく」: 4, 「手をひく」: 4

のように同じ出現頻度でも実際は「手」と「ひく」より「風邪」と「ひく」のほうがより共起関係が強いという印象がある。それは、それぞれの単語(例では、「風邪」と「手」)の出現頻度の違いによって共起データの出現頻度が変化するからである。そこで、共起データの出現頻度を正規化することを考え、共起関係の強さというものを以下のように定義する。

共起関係の強さとは、それぞれの単語が全く独立(一次独立)に出現すると仮定してそれぞれの単語の出現頻度から共起データの出現頻度の期待値を計算し、その計算値から実際の共起データの出現頻度がどの程度ずれているかを数値で示したものである。

共起関係の強さを求めるために、まず、単語 $T_1$ と単語 $T_2$ の共起データの出現頻度の期待値 $E_{12}$ を求める。

$T_1$ の出現頻度を $C_1$ 、 $T_2$ の出現頻度を $C_2$ とすると、 $T_1$ の共起データを1つとってきたときにそれが $T_2$ である確率 $p_{12}$ は、

$$p_{12} = C_{12} / \sum C_k \dots\dots\dots (1)$$

となる。従って、 $E_{12}$ は、

$$E_{12} = C_1 * p_{12} \\ = C_1 * C_{12} / \sum C_k \dots\dots\dots (2)$$

となる。

この場合、共起データの出現頻度が $k$ である確率 $P_{12}(x=k)$ は、poisson分布で近似でき、そのときの分散 $V_{12}$ は、

$$V_{12} = E_{12} \\ = C_1 * C_{12} / \sum C_k \dots\dots\dots (3)$$

となる。

ここで、実際の共起データの出現頻度を $C_{12}$ として、共起関係の強さ $W_{12}$ を以下の式によって求める。

$$W_{12} = (C_{12} - E_{12}) / (V_{12})^{1/2} \\ = (C_{12} - C_1 * C_{12} / \sum C_k) / (C_1 * C_{12} / \sum C_k)^{1/2} \dots\dots\dots (4)$$

$W_{12}$ が大きい(+である)ほど $T_1$ と $T_2$ の共起関係があり、小さい(-である)ほど共起関係がない。

コンピュータ関係の例文から抽出した「が」、「を」、「で」、「に」の前後に現われる名詞と動詞の共起データの出現頻度を表1に、それをもとに(4)式により計算した共起関係の強さを表2に示す。

### 4. 意味的な近さ

2. で述べたようにある単語との共起関係によってグルーピングする方法で意味的な近さを求めることを考える。しかし、2. の例のように1つの単語についての共起関係だけでは、「いん石が降る」という共起データがあると「いん石」が「雨」や「雪」と意味的に近いことになってしまう。そこで、他の単語との共起関係を調べてみると、

「雨がやむ」、「雪がやむ」、「みぞれがやむ」

のように「雨」や「雪」などは「やむ」と共起関係があるが、「いん石」は「やむ」と共起関係がない。また逆に、

「いん石を発見する」

のように「いん石」には共起関係があるが、「雨」や「雪」などには共起関係がない単語がある。これらの点から、単語間の意味的な近さは共起関係にある単語(群)

の分布を比較することによって得ることができると考えられる。

単語  $T_i$  と単語  $T_j$  の意味的な近さ  $D_{ij}$  を求めるため、まず、 $T_i$  と共起関係のある単語（共起関係の強さ  $W_{ik}$   $>$   $0$  である  $T_k$ ）について、 $T_i$  と共起関係がある（ $W_{jk}$   $>$   $0$ ）場合は近さのポイントを大きく（+に）、共起関係がない（ $W_{jk} \leq 0$ ）の場合は近さのポイントを小さく（-に）なるようにして近さのポイントを求める。  
1つの特異なデータによってポイントが高くなるならない

		月	ネット ワーク	コンピ ユータ	パソコン	年	LAN
を	導入する	0	4	21	12	0	5
を	構築する	0	45	0	0	0	13
を	接続する	0	3	13	8	0	0
に	接続する	0	8	2	7	0	2
に	設立する	32	0	0	0	35	0
に	発表する	27	0	0	0	14	0

表1

		月	ネット ワーク	コンピ ユータ	パソコン	年	LAN
を	導入する	-2.3	0.4	10.2	5.6	-1.5	3.1
を	構築する	-1.7	30.9	-1.4	-1.3	-1.1	13.5
を	接続する	-1.2	2.1	12.8	8.1	-0.8	-0.6
に	接続する	-1.2	7.4	1.2	7.1	-0.8	2.6
に	設立する	31.3	-0.8	-0.8	-0.7	52.7	-0.5
に	発表する	28.5	-0.7	-0.7	-0.7	22.5	-0.5

表2

	月	ネット ワーク	コンピ ユータ	パソコン	年	LAN
月	****	---	---	---	---	---
ネットワーク	-29.5	****	---	---	---	---
コンピュータ	-29.7	-2.3	****	---	---	---
パソコン	-28.6	-1.5	10.0	****	---	---
年	28.1	-21.2	-20.0	-20.9	****	---
LAN	-19.6	10.1	-0.2	4.2	-14.9	****

表3

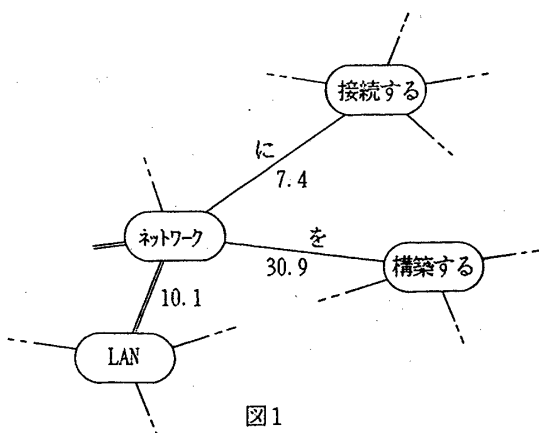


図1

ように、

$$W' = 2/(1+e^{-W})-1 \dots\dots\dots (5)$$

によって共起の強さを  $[-1, 1]$  の範囲にし、ポイント  $W'_{ik} * W'_{jk}$  として、 $W'_{ik} > 0$  であるすべての  $k$  に対して総和を  $D'_{ii}$  とすると、

$$D'_{ii} = \sum (W'_{ik} * W'_{jk}) \quad (k: W'_{ik} > 0) \dots\dots (6)$$

同様に  $D'_{jj}$  は、

$$D'_{jj} = \sum (W'_{ik} * W'_{jk}) \quad (k: W'_{jk} > 0) \dots\dots (7)$$

$D_{ij}$  はその和とし、

$$D_{ij} = D'_{ii} + D'_{jj} \\ = \sum (W'_{ik} * W'_{jk}) \quad (k: W'_{ik} > 0) \\ + \sum (W'_{ik} * W'_{jk}) \quad (k: W'_{jk} > 0) \dots\dots (8)$$

$D_{ij}$  が大きい（+である）ほど  $T_i$  と  $T_j$  は意味的に近く、小さい（-である）ほど意味的に遠い。

表2の共起関係の強さから(8)式によりすべての  $i, j$  ( $i \neq j$ ) について計算した意味的な近さのマトリックスを表3に示す。

5. 意味ネットワークの作成

表2から共起関係の強い ( $W_{ij} > 0$ ) ペア、表3から意味的に近い ( $D_{ij} > 0$ ) ペアを抽出し、それを組み合わせて意味ネットワークを作成した。作成した意味ネットワークの一部を図1に示す。

6. まとめ

例えば、「コンピュータ」と「パソコン」、「月」と「年」などが意味的に近くなっており、かなり良好な結果が得られた。

今回は助詞の前後に現われる名詞と動詞の共起関係だけから意味的な近さを求めたが、今後、名詞と名詞（連続名詞）、助詞の前後に現われる名詞と名詞、形容詞と名詞などの共起関係も抽出して意味ネットワークを再度作成する予定である。また、出現頻度の低い単語について意味的に近い単語から共起データを補完することにより実用的な意味ネットワークを作成していきたい。

参考文献

- [1] 仲尾、初山 係り関係による単語のクラスタリングの試行 情報処理学会自然言語処理 65-1 1988
- [2] 田中、吉田 自然言語の知識獲得 情報処理学会自然言語処理 65-2 1988
- [3] 藤原 共起パターン分布に基づく単語間類似度を用いた動詞・名詞のクラスタリング法 第2回人工知能学会全国大会 1988
- [4] 松川他 共起関係に注目したDM分解と確率的推定による単語のクラスタリング 情報処理学会自然言語処理 72-6 1989
- [5] 田中 語と語の解析用資料 1990
- [6] 杉村他 意味コード体系の自動生成 情報処理学会自然言語処理 78-4 1990