連想概念辞書を用いた意味ネットワークの構築と多義性解消への応用

岡本 潤　石崎 俊

慶應義塾大学 政策・メディア研究科
{juno,ishizaki}@sfc.keio.ac.jp

1　はじめに

コンピュータでテキスト処理を行うとき, 文脈による語の多義性や音声合成分野での同表記異義語の「読み」および「語義」の違いを解消する必要がある.例えば「金」は文脈により貴金属の「きん」や貨幣の「かね」などが考えられる.

本研究では, 連想概念辞書中の概念の連想関係と距離情報を用いて, 文書中の語から意味ネットワークを作成し, 活性拡散によって「読み」や「語義」を決定する手法を提案する.また, 毎日新聞の記事を用いて本システムの評価を行い, 本手法とナイーブベイズ法, ベースライン法との比較によって有効性を確認する.

2　連想概念辞書

本論文で使用した連想概念辞書 [5] は, 人間を被験者として大規模な連想実験を行い, 実験より得られたデータをもとに構築している.

2.1　連想実験

連想実験は自由連想ではなく, 被験者に名詞を刺激語として呈示し,「上位概念」「下位概念」「部分・材料概念」「属性概念」「類義概念」「動作概念」「環境概念」の 7 つ連想関係に関して連想させ, 任意の個数の連想語を 1 単語ずつキーボード入力させる.

刺激語は, 小学校の国語の教科書の学習基本語彙中 [4] の名詞と, 学習基本語彙以外で実験に用いる名詞の計約 840 語である.

また,1 刺激語に対し被験者 50 人で実験を行った.

2.2　距離の定量化と連想概念辞書の記述書式

連想概念辞書では刺激語と連想語間での連想のしやすさを, 概念間の距離として定量化している. 刺激語と連想語との概念間の距離 D は連想実験から得られる連想頻度 F, 連想順位 S のパラメータによる線形結合で表現し, 線形計画法を用いて (1) 式のように最適解が求められている [5].最適解はパラメータをもとに境界条件を距離 D の値が最大で 10.0 程度, 最小で 1.0 程度になるように定め, シンプレックス法で計算している.

$$
\begin{aligned}
D &= 0.81 \times F + 0.27 \times S, \qquad (1)\\
F &= \frac{N}{n+\delta}\\
\delta &= \frac{N}{10} - 1 \quad (N \geqq 10)\\
S &= \frac{1}{n}\sum_{i=1}^{n} s_i
\end{aligned}
$$

ここで刺激語を A, 連想語を B とした時,F は連想語 B を連想した被験者の割合を補正した値,S は連想語 B が連想された順位の平均した値, n は連想人数 ($n \geqq 1$),N は刺激語 1 語に対する被験者数,s_i は被験者 i が連想した語の順位である. 多くの被験者が同一の語を連想している場合は, その連想語は刺激語にとって連想しやすい語であると考えられ, 概念間の距離も短くなる.

刺激語,	連想関係,	連想語,	頻度,	連想順位,	距離
⋮	⋮	⋮	⋮	⋮	⋮

図 1: 連想概念辞書の記述形式

図 1 は連想概念辞書の記述形式である.刺激語の総数は 840 語, 連想関係は「上位概念」「下位概念」「部

分・材料概念」「属性概念」「類義概念」「動作概念」「環境概念」と「関連語」になる．「関連語」は 7 つの連想関係に分類できない連想語がある場合にもうけた連想関係である.たとえば刺激語「犬」に対しての連想語「猫」などは「関連語」とする.また, 頻度（連想者数を被験者数で割った値）, 連想順位, 概念間の距離の情報も含まれている．

3 意味ネットワークの構築と活性拡散による多義性の解消

多義文の理解において, ネットワーク表現を用いた超並列統語解析モデル [8] などがある.また, ネットワーク内の活性値の計算は活性化拡散モデル [6] を用いたものがあげられる.日本語の多義性解消に関する研究はサポートベクターマシンなど様々な機械学習手法 [3] を用いる方法やニューラルネットワークを用いた研究などがある [7].

本研究では, 多義語 (同表記異音異義語) を含むパラグラフに対して以下のステップで連想概念辞書の連想関係と距離情報の情報を用いて意味ネットワークを作成し, 活性拡散を用いて多義性の解消を行う．

3.1 意味ネットワークの構築

同一パラグラフにある多義語は 1 通りの「読み (語義)」を持つものとし, そのパラグラフに対して以下の手順で意味ネットワークを作成する．

1. Chasen[2] を用いて形態素解析を行い, 必要な修正を加え単語ごとに基本形, 品詞などの情報を得る．
2. 単語が名詞で連想概念辞書の刺激語の場合, 以下の条件で意味ネットワークを作成する
   (a) 単語が多義語の場合は, そのすべての候補を辿る（例えば, 単語が「額」の場合刺激語「額（がく）」と「額（ひたい）」の両方を辿る）
   (b) 累積距離が 8 になるまで各連想関係を辿る．
   ただし,
   i. 環境概念で連想された語の部分・材料概念は辿らない
   ii. 上位概念として連想された語の下位語のさらに上位語は辿らない
   iii. 下位概念として連想された語の上位語のさらに下位語は辿らない
3. 文書中の単語が連想概念辞書の連想語の場合
   (a) 上位概念, 下位概念, 部分・材料概念, 属性概念, 類義概念, 動作概念, 環境概念で, その刺激語が上記の意味ネットワークにあれば追加
4. 多義語となる刺激語（例えば「額（がく）」と「額（ひたい）」）同士は抑制リンクを追加する

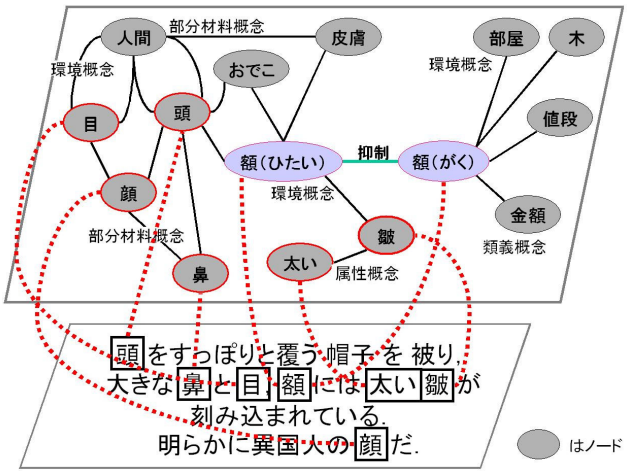

図 2: 意味ネットワークの例

3.2 活性拡散による多義性の解消

構築した意味ネットワークをもとに活性拡散を用いて各ノードの活性値を計算し, 同表記異音異義語の「読み（語義）」を決定する.活性拡散ではパラグラフ中の単語の影響を大きくするために, パラグラフ P_k に出現する単語の個数をもとに意味ネットワーク中のノード N_i の初期値 $V_i(0)$ を求める ((2) 式).次に (3) 式により各ノード N_i の活性値を計算する．

$$V_i(0) = 1000 \times C(P_k, N_i) \tag{2}$$

$$V_i(t+1) = \frac{V_i(t)}{2} + \sum_{j=1} \frac{V_j(t)}{\alpha D_{ij}} \tag{3}$$

ここで,$\sum (Vi(t) - V_i(t+1))^2$ が十分に小さい値となるまで計算を繰り返す.$V_i(t)$ は時間 t でのノード N_i の活性値,$C(P_k, N_i)$ はパラグラフ P_k に出現するノード N_i の個数,D_{ij} は接続する概念 (ノード) 間の距離で, 抑制リンクの場合は $D_{ij} = -1.0$ とする.$V_j(t)$ はノード N_i に接続するノードの活性値,α はリンクの総数を距離の最大値で割った値とした.また, ノード N_i と N_j が接続していない場合は D_{ij} は無限大となる.

解析対象とするパラグラフから作成された意味ネットワークで活性値を計算したときに多義語のノード (たとえば「額」であるならば,「額（がく）」と「額（ひたい）」の 2 つのノード）の活性値の比較し高いほうを「読み」として選択する.

4 本手法と既存の多義性解消手法との比較

前節で説明した多義性解消手法を実装し, ナイーブベイズ法とベースライン法と比較することで本手法の評価を行う.

4.1 既存の多義性解消手法

ナイーブベイズ法は文書分類の一手法としてよく用いらている.これは, ベイズの定理に基づいて各分類になる確率を推定し, その確率値が最も大きい分類を求める分類とする手法である [3]. 近年では迷惑メールのフィルター機能や文書分類 [1] などをはじめ様々な分野に応用されている.

ナイーブベイズ法で多義語の解消をするにあたって, 分類のための語義を c, パラグラフ p_k に現れる単語を $\{w_1, \cdots, w_n\}$ とすると各語義のもとで単語は独立であると仮定して,

$$P(w_1, \cdots, w_n|c) = \prod_{i=1}^{n} P(w_i|c) \tag{4}$$

とし, 次式により可能な分類の集合 C から, その分類（語義）c を得る.

$$c = \underset{c_j \in C}{\operatorname{argmax}} P(c_j|w_1, \cdots, w_n) \tag{5}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(w_1, \cdots, w_n|c_j)P(c_j) \tag{6}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{n} P(w_i|c_j) \tag{7}$$

ナイーブベイズでは $P(c_j) \prod_{i=1}^{n} P(w_i|c_j)$ を最大化する分類を出力する.また, 単語の出現頻度の補正（ディスカウンティング）には, ジェフリー・パークス法を用いた.

次にベースライン法では, 学習データで最も頻度が多かった語義を正解として選ぶものとする.

4.2 評価実験と結果

ここでは連想概念辞書中の刺激語である「額（がく）」と「額（ひたい）」を用いて, パラグラフ中の同表記異音異義語「額」の「読み」の正解率を比較しその結果を示す.

まず, 毎日新聞 CD-ROM 版の 93 年から 95 年の記事に対して Chasen を用い形態素解析を行い, 品詞が「名詞-一般」となっている「額」を含むパラグラフを抽出した (1330 パラグラフ).「累積額」などのように「累積（名詞-サ変接続）」＋「額（名詞-接尾-一般）」の「額」は多義語として用いない.次に, 抽出した 1330 パラグラフを「額」を「がく」と読むグループと「額」を「ひたい」と読むグループの 2 つに人手で分類し, 正解データとした.ただし,「がく」については「金額」「額縁」などの語義が考えられるが今回は同表記同音異義語ついては考慮しないものとする.評価実験に用いたパラグラフは全体の 5%（66 パラグラフ）で, ナイーブベイズとベースラインでは残りの 95% を学習データとして用いる.

1330 パラグラフからランダムに 66 パラグラフを選択した時の正解率の平均は, 本手法では 84.9%, ナイーブベイズ法は 90.0%, ベースライン法は 80.3% になった.

次に, 多義性解消に対して連想概念辞書の効果を調べるために, 刺激語が多く含まれている上位 66 パラグラフ, 上位 57 パラグラフ, 同様に 39,30,24,16 パラグラ

フの 6 つの場合に関して,3 手法を比較した結果を表 1 に示す.

表 1: 刺激語を多く含むパラグラフを対象とした評価

パラグラフ数	本手法	NB	B
66(全体の約 5%)	77.3%	75.8%	59.1%
57(全体の約 4.5%)	78.9%	77.2%	57.9%
39(全体の約 3%)	79.5%	79.5%	53.9%
30(全体の約 2.5%)	80.0%	76.7%	53.5%
24(全体の約 2%)	83.3%	75.0%	54.2%
16(全体の約 1%)	87.5%	62.5%	56.3%

NB:ナイーブベイズ法,　　B:ベースライン法

4.3　考察

パラグラフ中の刺激語の個数に関係なく,66 パラグラフをランダムに選択した場合はナイーブベイズ法での精度は高く, ナイーブベイズ法は多義性の解消にも有効な手段の一つであると考えられる.

表 1 において, 本手法では, 刺激語を多く含むパラグラフの場合は正解率が高いが, パラグラフに含まれる刺激語が少なくなるにつれて正解率が低くなる.一方ナイーブベイズ法については, 刺激語を多く含むパラグラフでは正解率が低くい. 連想概念辞書は小学校の学習基本語彙を刺激語とした連想実験により得られた結果をもとに構築しているので, 基本語を多く含むと考えられる.今回用いた記事では「わいろ」「献金」のような「お金」に関する語の出現率が高いと考えられる.しかしこれらの語は学習基本語彙とはなっていない. 基本語を多く含むパラグラフの場合はナイーブベイズ法の精度が落ちる傾向があると考えられる.

またベースライン法の正解率が低くなるのは,「額（ひたい）」を正解とするパラグラフが増えるためであると考えられる.「額（ひたい）」から連想される語は, 連想概念辞書中で刺激語となっている場合が多いためパラグラフ全体として刺激語の数が増えたと考えられる.

今後, 連想概念辞書が整備され刺激語の数が増えていけば, 刺激語を含むパラグラフの数か増えていくので将来的にはナイーブベイズを用いた手法よりもより多くの場合で正解率が上回る可能性があると考える.

5　今後の展開

本論文では連想概念辞書に中の「額（がく）」「額（ひたい）」について多義性の解消を行った.今後は「金（きん）」「金（かね）」や「文（ぶん）」「文（ふみ）」や「角（つの）」「角（かど）」や「札（ふだ）」「札（さつ）」などについて評価を行う予定である.連想実験では刺激語の読みは被験者に提示しているが, 語義を提示しておらず, 連想概念辞書では「額（がく）」という刺激語には「額縁」と「金額」に関する連想語が混在している.今後は,「額（がく）」関して語義の違いが分かるように連想実験を行って, 同表記同音異義語に関しても解析できるように拡張していきたい.

参考文献

[1] 阿部倫子, 田中久美子, 中川裕志,”コメントを用いた映画の分類” 情報処理学会 NL 研究会 NL-150, pp.105-110, 2002.

[2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸, ”日本語形態素解析システム『茶筌』version2.0 使用説明書第二版”, NAIST-IS-TR99012, 1999.

[3] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均, ”SENSEVAL2J 辞書タスクの CRL の取り組み－日本語単語の多義性解消における種々の機械学習手法と素性の比較－”, 自然言語処理, Vol.10, No.3, pp.115-133, 2003 .

[4] 甲斐睦朗 松川利広, 語彙指導の方法, 光村図書,1996.

[5] 岡本潤, 石崎俊,”概念間距離の定式化と既存電子化辞書との比較”, 自然言語処理,Vol.8,No.4,pp37-54,2001.

[6] McCelland J.L. and Rumelhart D.E.,”An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings”,Psychological Rev.,Vol.88,No.5,pp375-407,1981.

[7] 高橋直人, ”階層型ニューラルネットによる語彙的曖昧性の解消”, 情報処理学会誌, Vol36, No.9, pp.2102-2112, 1995.

[8] Waltz, D.L. and Pollack, J.B.,”Massively parallel parsing: A strongly interactive model of natural language interpretation”,Cognitive Science, Vol.9,pp51-74,1985.