連想概念辞書とコーパスを組み合わせる曖昧性解消手法の検討

堤田恭太[1], 岡本潤[2], 内山清子[3], 石崎俊[1]

[1]慶應義塾大学　環境情報学部

[2]慶應義塾大学　ＳＦＣ研究所

[3]慶應義塾大学　政策・メディア研究科

〒252-8520 神奈川県藤沢市遠藤 5322

e-mail：{t04549kt, juno, kiyoko, ishizaki}@sfc.keio.ac.jp

１．はじめに

近年, 電子化されたテキストデータが増えるにつれ, そこに含まれる情報の活用にも一層の関心が向けられるようになった. これまで人間がコストをかけて処理してきた情報を効率的に扱えるようになることにより, 機械翻訳や情報検索など多方面で言語情報の利用が考えられてきている.

その一方で, コンピュータで自然言語処理を行なうとき, 人間にとって使いやすいシステムを構築するには, 人間が持つ一般的な知識や扱う分野の背景的知識などの情報を電子化して使用可能にする必要がある. たとえば, 単語の多義性解消に代表される曖昧性解消の問題は様々なアプローチから研究されている[2]. 多義語を含む文の理解において, ネットワーク表現を用いた超並列統語解析モデルなどがある[5]. 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を用いたものがあげられる[7]. また, 語の共起関係に基づく類似度の利用, ナイーブ・ベイズやサポートベクターマシンなど様々な機械学習手法を用いる方法, ニューラルネットワークを用いた研究などがある[6].

中でも, コーパスでの共起語を用いたナイーブ・ベイズ法などの統計的に曖昧性を解消する手法がよく使われている. しかし, 十分に有効な共起語を得られない場合に精度が低くなる. そのような場合でも, 分類先のカテゴリについて関連する語を網羅的に増加すれば精度の改善が期待できると考えられる.

そこで本稿では, 人がもつ単語間の連想関係を実験によって定量的にデータ化した連想概念辞書とコーパスを用いて学習を行い, ナイーブ・ベイズ法による分類を行って語の多義性解消の精度を向上させるための手法を提案する.

2. 連想概念辞書

本論文で使用した連想概念辞書は, 小学校の学習基本語彙を刺激語とした連想実験を行ない, 大量の連想語を収集して構造化すると同時に, その連想語との距離が定量化されている. また刺激語約 1100 語, 連想語数約 28 万語, 異なり語数約 6 万語の大規模な辞書データとなっている.

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 つ課題に関して連想させ, 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[3]の基本名詞と, 学習基本語彙以外で連想実験で得られた基本名詞の計約 1100 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. 概念間距離の定量化と辞書の記述形式

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の度合いを, 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[4]. ここで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.

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，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，si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．

$$D = 0.81F + 0.27SC \quad \cdots(1)$$

は連想概念辞書の記述形式である．

「関連語」は 7 つの課題に分類できない連想語がある場合にもうけた課題である．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．次に，頻度（連想者数を被験者数で割った値），連想順位，正規化された連想時間，概念間の距離である．概念間の距離は，連想順位 Sの値にもよるが，おおよそ 1～10 の間にある．

図 1 は，刺激語「いす」を中心とした概念辞書の構造である．「いす」から連想関係ごとに多くの語が連想されており，概念間の距離も定量化されている．また，連想語の一部は刺激語として連想実験が行われており，比較的密なネットワーク構造をなす概念辞書となっている．

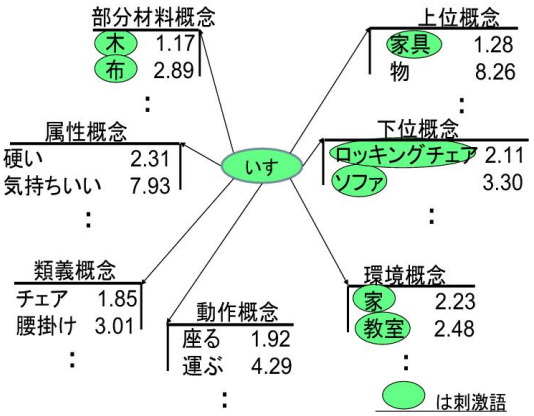

図 1 ：連想概念辞書の概念構造

3. 連想概念辞書とコーパスを用いた語彙の拡張

3.1 多義語の定義と語義の決定

本論文では，連想概念辞書の連想関係に注目して多義語の語義を決定する．ここでは，ある語を「部分材料」という連想関係で連想している刺激語を語の語義として定義している．7 つの連想関係でも「部分材料」を用いる理由は多義性を部分全体関係を用いて解消するためである．

たとえば，針 という語について，連想概念辞書内での刺激語と連想語の関係は図のようになっており，共通の連想語(針)をもつ「刺激語」を多義語(針)の語義とする．

刺激語	:連想語(部分材料)	:針の英訳例
裁縫	:針,糸,裁縫箱	:sewing needle
時計	:針,長針,短針	:clock hand
釣り	:針,竿,糸,えさ	:fish hook
注射器	:針,ガラス	:injection needle
ピアス	:針,金属,宝石	:earring needle
東洋医学	:針,漢方,灸	:acupuncture

図 2 ：針を連想している刺激語の例

ここでは語義の決定に以下の条件を用いることで，針の語義を裁縫・時計・釣り・注射器・ピアス・東洋医学の 6 つに限定した．

- 連想概念辞書において連想語である針と語義となりうる刺激語の概念間の距離が 5.0 未満である
- コーパス中での出現数が確保できる

また，これらが多義的であることは，針それぞれの英訳例をあてると明らかであると考えられる．

3.2. 連想語を用いた語彙の拡張

ここでは語義とする語が連想概念辞書内では「刺激語」であることを利用し，それらの連想語との共起語をコーパスから収集した．

連想概念辞書を用いることで語義となる「刺激語」と関連のある語を収集することができる．これにより語義に関連する単語を効率的に拡張して収集することができる．

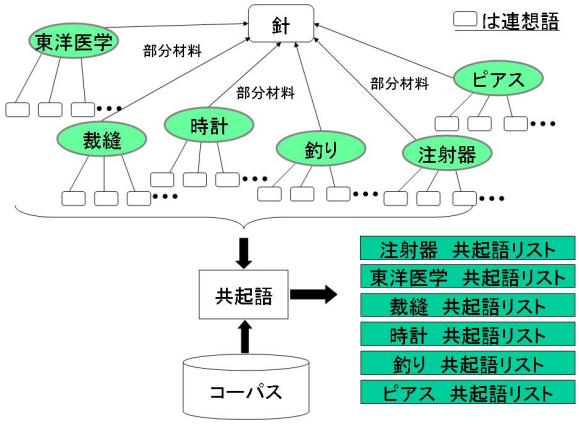

図 3: 語彙拡張の概念図

4. 語彙の拡張とナイーブ・ベイズ法による曖昧性の解消

ナイーブ・ベイズ法とは，あらかじめ学習したデータに基づいて文書の分類などを行うアルゴリズムであり，近年はスパムメールのフィルタリングなどに広く用いられている．文書中に含まれる単語をその文書を特徴づける素性とし，ある文書がどのカテゴリに属するかを，事前に確率・統計的に求めた値を用いて決定し，文書分類の手法に用いられるものの一つである[1][8].

語義を $\{t_i;t_1,\cdots,t_6\}$，語義 (t_i) となる「刺激語」から連想される語を含むパラグラフを $\{s_{ij};s_{i1},\cdots,s_{ij}\}$ とおき，s_{ij} にあらわれる単語を $\{w_i;w_1,\cdots,w_n\}$ とおくと，s_{ij} に対しての語義は，事後確率 $P(t_i \mid s_{ij})$ を最大化するような語義 (t_i) を選ぶことにより決定でき，$\hat{t}$ は以下の式のように求めることができる．

$$\begin{aligned}\hat{t} &= \arg\max_{t_i} P(t_i \mid w_1,\cdots,w_n) \\ &= \arg\max_{t_i} P(w_1,\cdots,w_n \mid t_i)P(t_i) \quad \cdots(2)\end{aligned}$$

さらに，各語義のもとでの単語が独立に生起すると仮定すると，それぞれに存在する総単語数 N とある単語 w の頻度 r をもとに，最尤推定によって事前確率にあたる出現確率は $P(w)=r/N$ と求められる．また，

$$P(w_i,\cdots,w_n \mid t_i)=\prod_{k=1}^{n} P(w_k \mid t_i)$$

とし，分類は次式によって行うことができる．

$$\hat{t}=\arg\max_{t_i} P(t_i)\prod_{k=1}^{n} P(w_k \mid t_i) \quad \cdots(3)$$

ここでは出現確率を，

$P(t_i)=t_i$ に含まれる単語数/全単語数

とし，t_i に出現する単語総数を N_i，t_i において w_i が出現する回数を F_{ik} とおき，ゼロ頻度問題では単語列中の異なり総数を V_{all} とおくと $P(w_k \mid t_i)$ が以下の算式で表される予期尤度推定法(ジェフリース・パークス法)を採用した．

$$P(w_k \mid t_i)=(F_{jk}+0.5)/(N_i+0.5*V_{all})$$

w_i の出現数 F_{ik} が 0 であることにより，成り立つ次式を用いた．

$$P(w_k \mid t_i)=0.5/(N_i+0.5*V_{all}) \quad \cdots(4)$$

5 実験と評価

ここでは，語義の正解データを人手で付与して学習データを作成した方法，連想概念辞書を用いずに語義とコーパスだけで共起語の収集を行った方法，本稿で提案する連想概念辞書とコーパスを組み合わせた方法とでの，正解率の比較を行った結果を示す．

まず，毎日新聞 CD-ROM 版の 93 年から 95 年の記事に対して，形態素解析器 MeCab を用いて解析し，品詞が「名詞,一般」となる針を含むパラグラフをコーパス中から抜き出した．針の語義として裁縫・時計・釣り・注射器・ピアス・東洋医学の 6 つのいずれにも当てはまらないと考えられるものを除いた，計 236 パラグラフについて正解データを与え，全体の 1 割に当たる 26 パラグラフをテストデータとしてランダムに選び，残りの 210 パラグラフを用いて学習データとした．

語義の正解データを人手で付与して学習データを作成した手法では，同様に針の語義の与えられている学習データから品詞が名詞・動詞・形容詞であるものを取り出し，それぞれの語義における単語の出現頻度から事前確率を求め，ナイーブ・ベイズ法を用いてテストデータにおける正解率を出した．

連想概念辞書を用いずに語義とコーパスだけで共起語の収集を行った手法では，コーパスから針の語義である裁縫・時計・釣り・注射器・ピアス・東洋医学の 6 つについて，それぞれを含むパラグラフを抜き出した．そのパラグラフ中に含まれる名詞・動詞・形容詞の単語の頻度から事前確率を求め，ナイーブ・ベイズ法を用いてテストデータにおける正解率を出した．

本稿で提案する連想概念辞書とコーパスを組み合わせた方法では，3 章で述べてきたように，針の語義となる語の連想語を用いて語彙を拡張し，語義における連想語それぞれについてパラグラフを抜き出した．語義だけを用いる手法と同様に，その語義における単語の出現頻度から事前確率を求め，ナイーブ・ベイズ法を用いてテストデータにおける正解率を出した．

表 1　3 手法の正解率

手法	正解率
正解データのみ	0.81
語義として利用する語の拡張	0.54
語義として利用する語と連想語の拡張	0.50

6　考察

本手法で連想語のみでコーパスから語を拡張した場合, 一つ一つの連想語に広い用法があるために, 語義として利用したもとの刺激語と直接関わるとは考えにくいパラグラフが大量に引き出されてしまった. それらを含むパラグラフの集合からの単語の頻度を計算して事前確率を出した場合, テストデータにおける単語そのものの網羅数は大きくなるものの, 語義として利用する語の間での事前確率に差が出にくくなるため, 正解率が下がってしまったと考えられる.

今後, 連想語を用いて語彙を拡張する際に, 語義として利用する刺激語とその連想語を含むようなパラグラフおよびセンテンスから共起語を抽出する手法で, 追試を行いたい. 例えば,「エサを変えると何も釣れない.」という文では, 刺激語「釣り」の連想語として,「エサ」や「釣る」などがあり, 文中には出ていないものの, 釣り針との関わりがより強い文を抽出できるようになることが期待される.

7 おわりに

本研究は, 連想概念辞書の「部分材料」関係に注目して語の多義性を扱った. また, ある語が曖昧性をもつことについて概念辞書から判別でき, コーパスと組み合わせることで曖昧性を解消するための学習が行える可能性があることを示した.

本稿で扱った針のような語義の曖昧性解消は, 語の下位語の同定と似た性質がある. 下位語のような概念そのものについて記述することは, 抽象度が低い分だけ定義がしやすいと考えられるが, その量は膨大であることが推測され, ある程度自動的な拡張が期待される.

今後, より汎用な曖昧性解消を可能にする規則を考え, 学習されたデータを蓄積していくことへの応用が考えられる.

また, 連想概念辞書において定義されている他の関係を用いた応用について, 例えば,「上位／下位」関係に注目することで語の「観点」の違いなどについての研究が行える可能性がある.「りんご」という語の「上位／下位」関係には,「商品」「植物」「食べ物（果物）」などがあり, どの文脈においてどの観点を扱っているのかを考えることは, 人間が行っている柔軟な知識処理に近づくことになると考えられ, より高度な言語情報処理への応用が考えられる.

謝辞

連想概念辞書構築にあたり協力して頂いた連想実験の被験者の皆様に感謝いたします. また, 学習データを作成するに当たって様々な助言を行ってくれた同研究室の粟飯原俊介氏に感謝致します.

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, NL 研究会, NL-150, pp.105-110, 2002.

[2] 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 の取り組みー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ー ", 自然言語処理 ,Vol.10,No.3, pp.115-133, 2003 .

[3] 甲斐睦朗,松川利広編, 語彙指導の方法, 光村図書, 1996.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 北研二,言語と計 4 確率的言語モデル, 東京大学出版会,1999.