E-010

連想概念辞書とコーパスを組み合わせる曖昧性解消法

A Study of Word Sense Disambiguation using Associative Concept Dictionary and Corpora

堤田恭太† 岡本潤‡ 内山清子† 石崎俊†
Kyota Tsutsumida Jun Okamoto Kiyoko Uchiyama Shun Ishizaki

1．はじめに

近年，電子化されたデータが増加するにつれ，そこからの情報抽出などの応用が考えられる自然言語処理への期待は，一層大きなものとなってきた．その一方で，コンピュータで自然言語を処理するにあたっては，いくつかの大きな問題があり，例えば，多義語の曖昧性解消は様々なアプローチからの研究がなされている[1][2][3]．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分類精度が低くなること，正解データを人手で付与するコストがかかること，作成した学習データの汎化能力に問題があることなどが知られている．

そこで本研究では，データ作成のコスト，分類の精度，学習データの汎化性能の 3 点において，多義語の同定に有効な手法を提案する．まず，人間を被験者とした連想実験により得られた結果を構造化した連想概念辞書[4]とコーパスを用いて，曖昧性解消に有効な関連語を自動的に収集した．次に，その頻度情報をパラメータにしたナイーブ・ベイズ法による分類を複数のテストデータに対して行い，分類の精度と学習データの汎化能力の検証を行って，本手法の有効性を示した．

2．連想概念辞書とコーパス

本研究で用いた連想概念辞書は，小学校の学習基本語彙[5]を刺激語とした連想実験を行い，大量の連想語を収集・構造化すると同時に，刺激語と連想語との距離を連想頻度と連想順位を用いて定量化した辞書である．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．

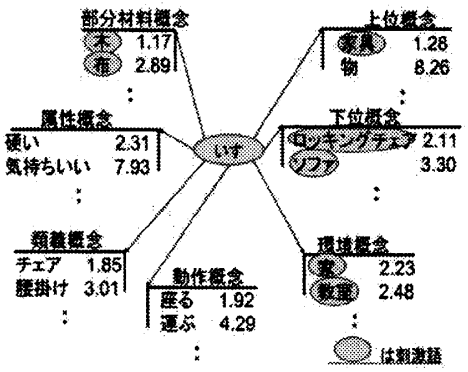

図 1: 連想概念辞書の概念構造

連想実験は，提示した刺激語において，上位概念・下位概念など図 1 の 7 種の概念関係について連想する語を問う形式で行われている．また図 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，概念間の距離は，近いものほど値が小さくなるように定量化されている．連想語のなかの基本的な語はさらに刺激語として連想実験が行われ，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．

コーパスは，パラメータの学習のために毎日新聞 CD-ROM 版の 93 年から 95 年の記事を用いた．また，テストデータには毎日新聞コーパス中の学習データ(210 文)に含めない文(22 文)と，他コーパスにおける汎化性能を比較する目的で，青空文庫のテストデータ(100 文)を用意した．

3. ナイーブ・ベイズ法

ナイーブ・ベイズ法[6]とは，事前に統計的に求めた値に基づいて文書の分類などを行うアルゴリズムであり，語の曖昧性解消に用いられる手法の一つである．近年ではスパムメールのフィルタリングにも利用されている．本研究においては，文中に含まれる単語の出現頻度を学習のパラメータとし，多義語がもつ複数の語義を指す語を分類先として，ナイーブ・ベイズ法による曖昧性解消を行う．

ここで，多義語の分類先を $\{t_i;t_1,\cdots,t_m\}$ とし，出現する単語を $\{w_j;w_1,\cdots,w_n\}$ としたとき，ある分類先 t_i に単語 w_j が出現する事前確率 $P(w_j \mid t_i)$ を用いて，次式(1)で表される $\hat{t}$ を分類先として選択する．

$$\hat{t} = \arg\max_{t_i} P(t_i \mid w_1,\cdots,w_n) = \arg\max_{t_i} P(w_1,\cdots,w_n \mid t_i)P(t_i) \quad \cdots(1)$$

さらに，各分類先のもとでの単語が独立に生起すると仮定し，出現確率 $P(t_i)$ を，

$$P(t_i) \approx (t_i \text{に含まれる単語数}) / \text{全単語数}$$

として得られる次式(2)を用いる．

$$\hat{t} = \arg\max_{t_i} P(t_i)\prod_{k=1}^{n} P(w_k \mid t_i) \quad \cdots(2)$$

また，「ゼロ頻度問題」への対応には，予期尤度推定法(ジェフリース・パークス法)[6]を採用した．

4. 実験手順と評価方法

本論文では，曖昧性を持つ語として「針」という語を扱った．語義の分類先には，単語間の部分・全体関係を利用するため，連想概念辞書において「針」を「部分・材料概念」に連想している刺激語である，裁縫，時計，釣り，注射器，ピアス，東洋医学の 6 つとした．

本研究は，データ作成のコスト，分類の精度，学習データの汎化性能の 3 点において有効性を示すことを目的とした．その為，まず，2 種類のコーパスのテストデータ(毎日

†慶應義塾大学大学院 政策・メディア研究科
‡慶應義塾大学 ＳＦＣ研究所

新聞・青空文庫) を用意した．学習データは，表 1 に示す 3 種類の手法で用意した．それぞれの手法で集めた文を用いて，該当する分類先ごとに語の出現頻度を集計し，各分類先における単語の事前確率値を計算した．その値をナイーブ・ベイズ法で用いる学習データとした．

評価は，2 種類のテストデータにおける各学習データの分類精度を比較した．

表 1: 学習データ作成の 3 つの手法

手法 1: 多義語を含む文に対して，正しい分類先を人手で付与した文の集合を用いる方法
手法 2: 分類先にあたる語を含む文を，その分類先の文とする方法
手法 3: **[提案手法]** 連想概念辞書とコーパスを用いる方法

手法 1 では，針という多義語を含む文に対して正しい分類先を示す刺激語を人手で付与し，品詞が名詞，動詞，形容詞であるものを対象に単語の出現確率を求め，ナイーブ・ベイズ法を用いてテストデータにおける分類精度を出した．

手法 2 では，まず，コーパスから針の分類先とする刺激語である裁縫，時計など 6 つの刺激語を含む文を集めた．含む刺激語ごとに，事前確率値を計算し，手法 1 と同様にテストデータにおける分類精度を出した．

手法 3 では，針の分類先となる語が連想概念辞書において刺激語となっている事を利用し，それぞれの連想語を用いて，その刺激語と連想語の両方を含む文を，その分類先の刺激語と関わりの強い文であるとして集めた．含む刺激語ごとに，手法 2 と同様に学習データを作成し，テストデータにおける分類精度を出した．

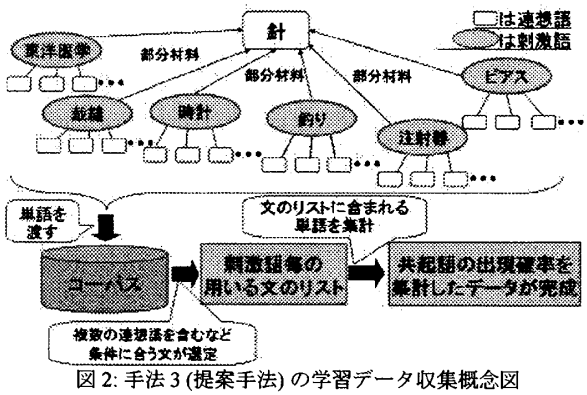

図 2: 手法 3 (提案手法) の学習データ収集概念図

5. 結果と考察

手法ごとに作成した学習データと，2 種類のテストデータ(毎日新聞・青空文庫)における分類精度を表 2 に示す．

表 2: 手法のテストデータに対する正解率の比較

手法	毎日新聞	青空文庫
手法 1	0.59	0.37
手法 2	0.54	0.27
手法 3	**0.71**	**0.67**

結果，毎日新聞と青空文庫のテストデータともに，手法 3 の学習データが最も優れた分類精度を示した．また，テストデータが異なるコーパスにおいては，手法 1,手法 2 は提案手法に比べて精度が著しく低下し，汎化性能において提案手法が優れることを示した．

また，手法 2 と手法 3 を比較することによって，刺激語に加えてその連想語を含む文を選定することの意義が明らかになる．刺激語と連想語を含む文を選ぶことで，分類精度を上げるために有効な共起語の密度が高くなり，曖昧性解消の精度が向上したと考えられる．さらに，手法 3 において含めるべき連想語量の比較を行った結果，コーパスの規模によって適切に設定する必要があることが分かった．本来，連想語を含む量が増える事により，より正確に分類先カテゴリを説明する文を収集することが期待される．しかし，学習に用いられる文が減りすぎ，単語の網羅数が下がりすぎた場合には，分類精度が悪くなることがある．

誤答の傾向は，手法 1 と手法 3 において違いが見られた．手法 1 では，付与したデータ中にごく類似した記事があった場合，特定の記事の文についても分類できたが，手法 3 では正しく分類できなかった．その原因としては，「裁縫の針」や「注射器の針」の用い方としては一般的なものでなかったためであると考えることができる．

6. まとめと展望

本研究では，連想概念辞書を用いてコーパスから取り出すべき文を自動的に選定することで，正しい分類先の付与などの人手で行われる作業コストを大きく減らすことができた．また，多義語についてのコーパスによる学習量が十分に確保できない場合にも，連想語を用いた関連の強い文の選定・収集により，学習に用いる有効なデータを確保し，分類精度を改善できることを示した．さらに，連想概念辞書が人間の持つ常識に近い一般的な知識を提供することで，学習データに汎用性を持たせた．このことから，今後，複数の専門分野にまたがる共通の知識基盤構築に用いて，データに汎用性を持たせることなどへの応用が考えられる．

7. 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．

8. 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－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 .
[3] 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.
[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] 甲斐睦朗, 松川利広編, 語彙指導の方法, 光村図書, 1996.
[6] 北研二, 言語と計算 4 確率的言語モデル, 東京大学出版会, 1999.