多義語の語義ベクトル分解

3R－5

服部 直人　佐藤 健吾　中西 正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

自然言語処理における重要な問題の一つに，言語に関する様々な曖昧性の問題がある．一般に，意味的な曖昧性を解消するためには，意味に関する様々な情報を規則化し記述しておく必要がある．しかし，意味は文脈に依存して決まるため，あらゆる文脈に対応できる全ての意味を予め規則として網羅的に記述しておくことは難しい．また，その曖昧性は，理想的な確率モデルを統計的に求める事の妨げとなっている．

2. 本研究の目的

本研究では，表層上は一つの要素である多義語名詞を複数要素と捉え，辞書情報を利用することにより，これを一つ一つの意味に対応させた要素（仮想名詞ベクトルと呼ぶ）に分解する．また，分解された名詞の有用性をクラスタリングのエネルギー低下から示し，仮想名詞ベクトルにより多義解消に必要な情報を注出できることを示す．

3. 特徴ベクトルの作成

本研究では，名詞をベクトルと捉え，その名詞を直接目的語として共起する d 個の動詞を軸とする d 次元動詞空間上でこれを示した．軸 i における名詞ベクトルの要素は，i 軸で示される動詞とその名詞の条件付き確率，式 (1) の値を用いた．

$$x[i] = p(x|v_i) = \frac{count(x, v_i)}{count(x)count(v_i)}, \qquad (i = 1, \cdots, d) \tag{1}$$

4. 特徴ベクトルの分解

多義語の問題を解消するために多義語 w の特徴ベクトルを語義別に分解する手法を提案する．

Vector Analysis of Ambiguity using Dictionary Information
Naoto HATTORI Kengo SATO Masakazu NAKANISHI
Department of Computer Science, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

4.1 辞書の定義文の利用

辞書において見出し語を定義するために使用されている単語の多くは，見出し語あるいはその語義との間に何らかの連想関係をもつと考えられる．例えば，*Longman Dictionary* から引用した "spring" に関する定義文を見てみると，"春" の語義では "*the season between winter and summer when leaves and flowers appear*" と書かれており，"spring" と共起する名詞として，"season","winter","summer","leaf","flower" などを抽出することができる．

4.2 共起名詞多重集合の作成

辞書には語義別に定義文が書かれているので，語義別に共起名詞多重集合 $\{S\}$ を作成することができる．しかし，一つの定義文に出現する単語の数は少なく，データとしては信用性があまりにも低い．そこで，データを補うためにあらかじめ定義文内に出現する名詞をコーパスを用いてクラスタリングし (クラスタリングアルゴリズムは後に記述)，共起名詞多重集合 $\{S\}$ が属するクラスのメンバーを $\{S\}$ に追加し，$\{S'\}$ を作成する．

4.3 Deterministic Annealing (DA) 法を用いたベクトル分解

作成した $\{S'\}$ の分布を基に，分布の中心を語義の数だけ求めることにより，語義数分の仮想名詞ベクトルを作成する．

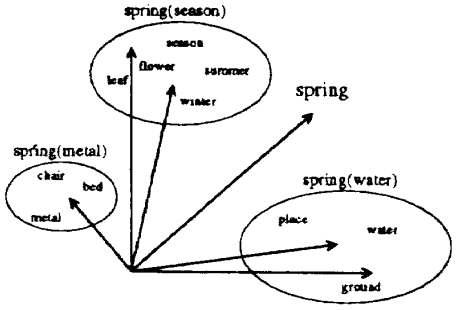

図 1: ベクトル spring の分解

4.4 ベクトル分解への DA 法適応

d 次元ベクトル $\boldsymbol{x}$ の集合 $\mathcal{X}$ を K 個の d 次元ベクトル $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K$ で代表される K 個の中心値を作成する．中心値の性能を，$\boldsymbol{x} \in C_i$ を中心値 $\boldsymbol{y}_i$ で近似する際の誤差の $\mathcal{X}$ にわたる平均値 L で評価し，誤差基準として確率分布の距離を表す式 (3)，$Kullback - Lieblar$ (KL) 距離を採用する．

$$L = \sum_{\boldsymbol{x}} \sum_{i=1}^{K} p(C_i|\boldsymbol{x}) D(\boldsymbol{x}||\boldsymbol{y}_i) \tag{2}$$

$$D(\boldsymbol{x}||\boldsymbol{y}) = \sum_{j=0}^{j=d} \boldsymbol{x}(j) \log \frac{\boldsymbol{x}(j)}{\boldsymbol{y}(j)} \tag{3}$$

$p(C_i|\boldsymbol{x})$ に関して何ら事前知識がないので，常套手段として最大エントロピー原理を適用して $p(C_i|\boldsymbol{x})$ を求めると，求める解は次式の Gibbs 分布となる．

$$p(C_i|\boldsymbol{x}) = \frac{\exp(-\beta D(\boldsymbol{x}||\boldsymbol{y}_i)}{\sum_{j=1}^{K} \exp(-\beta D(\boldsymbol{x}||\boldsymbol{y}_i))} \tag{4}$$

$$F = -\frac{1}{\beta} \sum_{\boldsymbol{x}} \log\Big(\sum_{j=1}^{K} \exp(-\beta D(\boldsymbol{x}||\boldsymbol{y}_i)\Big) \tag{5}$$

平衡状態では、系は自由エネルギーの式 (5) が最小となる状態に落ち着くので，$\boldsymbol{y}_i$ は $\partial F/\partial \boldsymbol{y}_i = 0$ を満たさなければならない．これより，以下の反復式 (6) が導かれる．

$$\boldsymbol{y}_i^{(t+1)} = \frac{\sum_{\boldsymbol{x}} \boldsymbol{x} p(C_i^{(t)}|\boldsymbol{x})}{\sum_{\boldsymbol{x}} p(C_i^{(t)}|\boldsymbol{x})} \tag{6}$$

5. クラスタリング

クラスタリング問題の中で，多義語は以下のような振舞いをする．[1][4].

- ハードクラスタリングを行なった場合
  多義語は，あまり好ましくないと思われるクラスに所属することが多い．
- ソフトクラスタリングを行なった場合
  それぞれの語義の特徴に引かれ，確率的に曖昧な位置をとる．

ソフトクラスタリングでは，多義語を複数のクラスに所属させることが可能であり，確率的な曖昧性を持たすことにより，多義語の問題を解決していると考えることもできる．

本研究では，多義語が決定的な過ちを侵すハードクラスタリングを採用し，クラスタリングアルゴリズムは，確率分布の距離を表す KL 距離式 (3) を使用して，以下のアルゴリズムによってクラスタリングを行なう．

1. すべての単語に対して，一つのクラスを割り当てる．
2. KL 距離が最も近い 二つのクラスを選択し，これらを一つのクラスに併合する．
3. ステップ 2 をクラスが一つになるまで繰り返す．

6. 評価方法

Pereira [4] は自由エネルギー F 式 (5) を定義し，自由エネルギーが低下するようにクラスタリングを行ない，その結果，得られたクラスから単語の統語的，意味的な共通性をとらえることに成功している．

多義語の特徴ベクトル分割前におけるクラスタリングのコストと，分割後のコストを比較する．コストの低下を示すことによって，多義語問題による統語的，意味的共通性の獲得困難を解消していることを示す．

結果に関しては現在実装中であり，発表時に報告したい．

参考文献

[1] Pietra V.J.D. deSouza P.V. Lai J.C. Brown, P.F. and R.L. Mercer. Class-based n-gram models of natural language. In *Computational Linguistics*, Vol. 18, pp. 467–479, 1992.

[2] 福本文代, 辻井潤一. コーパスに基づく動詞の多義解消. In *Jurnal of Natural Language Processing*, 第 4 巻, pp. 21–39, 1997.

[3] Y. Niwa, Y.and Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th COLING*, pp. 304–309, 1994.

[4] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993.

[5] Ido Dagan.Lillian Lee.Fernando Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the ACL*, 1997.