# 文脈ベクトルを用いた語義の曖昧性解消*

1R−7

山下 浩一† 吉田 敬一‡

静岡大学大学院理工学研究科§

## 1 はじめに

自然言語処理において、単語の多義性は非常に大きな問題の一つである。これに対し、近年、大規模コーパスからの共起情報や統計情報を利用して多義解消を図る研究が盛んに行われている [1, 2, 3, 4]。本稿ではより高い精度で語義の曖昧性解消を行うため、これらの情報に加え、コーパスから得られる語と語の依存関係の情報を利用する手法について報告する。また、本手法を用いて多義解消に関する実験を行ったので、その結果についても報告する。

## 2 語義の曖昧性解消法

### 2.1 文脈ベクトルを用いた手法

本研究では、単語の意味を EDR 概念辞書における約 10000 の主要な概念で表されると仮定した。これにより語義の曖昧性解消は、各単語が 10000 のクラスのどれに属するかを同定する問題に帰着できる。この問題の解決策として近年、大規模コーパスが利用可能となったことを背景に、文脈ベクトルと呼ばれる次のベクトルをコーパスから抽出して語義の曖昧性を解消する手法が提案されている [1]。

$$C(w) =< |w^1|, |w^2|, \cdots, |w^\omega| > \qquad (1)$$

ここで、$|w^i|$ は単語 $w$ の文脈サイズ内に単語 $w^i$ が現れた回数である。文脈サイズは通常、多義解消の対象となっている単語の前後 $n$ 語ずつ、計 $2n$ 語である。

式 (1) で定義される文脈ベクトルはコーパス全体にわたって各単語ごとに抽出される。さらに各単語の各意味ごとに文脈ベクトルを合計し、次のベクトルを生成する。

$$C(w_{sense}) =< |w^1|, |w^2|, \cdots, |w^\omega| > \qquad (2)$$

今、多義解消の対象となる単語を $W$ とすると、$W$ の語義は以下の式より導かれる。

$$arg\, max_{sense} sim(C(W), C(W_{sense})) \qquad (3)$$

ここで、$arg\, max_x f(x)$ は $f(x)$ を最大にする $x$ であり、$sim(C_1, C_2)$ は、ベクトル $C_1$,$C_2$ の類似度である。文献 [3] では、類似度の計算法として次の式を用いている。

$$sim(C_1, C_2) = \frac{C_1 \cdot C_2}{|C_1| \cdot |C_2|} \qquad (4)$$

ここで、$|C_i|$ はベクトル $C_i$ の長さである。

### 2.2 提案される手法

式 1 で定義される文脈ベクトルの問題点は、文脈サイズ内のどの単語も一律に同じ重みで扱っていることにある。一般に、単語は同一文中のほかの単語と何らかの依存関係を持っている[1]。そのような依存関係を持つ単語は、語義の曖昧性解消において重要な手がかりとなるものであり、依存関係を持たないほかの単語よりも重きを置かれるべきである。

以上の考察から本研究では、単語の出現回数ではなく単語の重みをベクトルの各要素とした次の文脈ベクトルを用いる。

$$C'(w) =< weight_1, weight_2, \cdots, weight_\omega > \qquad (5)$$

ここで、各単語の重みである $weight_i$ は次式で定義する。

$$weight_i = |w^i| + \delta(w, w^i) \qquad (6)$$

$\delta(w, w^i)$ は $w$ の文脈サイズ内において、$w$ と $w^i$ の間に依存関係がある場合は 1、依存関係がない場合は 0 を値として持つ関数である。また、語と語の依存関係はコーパスに記述されている構文情報などを利用して抽出する。文脈ベクトル $C'(w)$ がコーパスから抽出

---

*A Word-Sense Disambiguation Using Context Vector

†Kouichi Yamashita

‡Keiichi Yoshida

§Graduate School of Science and Engineering, Shizuoka University

[1] Yarowsky は形容詞や複合名詞の語義が、修飾される名詞に大きく依存しており、単語間の距離が非常に近いことを報告している [2]。また、動詞は動作主や目的語と強い依存関係で結ばれていることが知られている。

表 1: クローズド・テストによる実験結果

| 語義の曖昧性解消手法 | | 精度 |
|---|---|---|
| $C(w)$ による手法 | | 76.59% |
| $C'(w)$ による手法 | 重み：式 (10) | 78.28% |
| | 重み：式 (11) | 77.90% |

されたとき、単語 $W$ の語義を求める式は次のように表すことができる。

$$arg\,max_{sense}sim(C(W), C'(W_{sense})) \quad (7)$$

## 3 実験

実験を行うにあたり、曖昧性解消の対象となる単語は名詞、動詞、形容詞、副詞に限定した。また、文脈ベクトルを次のように近似した。

$$C(w) =< |c^1|, |c^2|, \cdots, |c^{11101}| > \quad (8)$$

$$C'(w) =< weight_1, weight_2, \cdots, weight_{11101} > \quad (9)$$

ここで $c^i$ は前述の EDR 概念辞書の主要な概念である。これに伴い、式 (6) は次の式で近似される。

$$weight_i = |c^i| + \delta(w, c^i) \quad (10)$$

実験には EDR 英語コーパスからランダムに抽出した 10000 文を用いた。EDR 英語コーパスでは文はアルファベット順で並べられており、各々の文は独立しているので、文脈サイズは一つの文とする。また、語と語の依存関係はコーパスに記述されている意味フレーム情報から抽出した。

精度の測定には、トレーニングに用いた 10000 文のうちの 1000 文をテストデータとして用いている。実験結果は表 1 に示す。それぞれの精度は、意味の割り当てを正しく行った単語数を曖昧性の解消を試みた単語の総数で割ったものである。

ここで、比較のために次式で定義される重みを用いた多義解消に関する実験結果を付け加えておく。

$$weight_i = |c^i| + \{5 - distance(w, c^i)\} \quad (11)$$

$distance(w, c^i)$ は、$w$ の文脈サイズ内における $w$ と $c^i$ の間の距離を表す。式 (11) は、単語は距離の近いものほど互いに依存しやすいという推定によっている。

表 1 に示されるように、今回提案した式 (10) の重みを用いることにより、従来の手法に比べ 2% 程度の精度向上が見られた。また、式 (11) の重みを用いた多義解消では、精度向上は 1.5% 程度であり、筆者らの手法よりも精度の向上幅が小さい。これは、式 (11) が距離の近い単語であれば依存関係を考慮することなく重みを付与してしまい、ノイズが入りこんだためであると考えられる。

さらに精度を向上させるためには、式 (6) で定義される単語の重みではなく、より妥当な重みを用いる必要がある。特に $\delta(w, w^i)$ に関しては、例えば依存関係をより強く反映するために、$\delta(w, w^i)$ を依存関係がある場合には $k(k \geq 1)$、ない場合には 0 を値として持つ関数とし、妥当な $k$ を求めることが考えられる。また、依存関係の種類により $k$ を変化させることも考えられる。

## 4 おわりに

本稿では、より高い精度で語義の曖昧性解消を行うために、単語の重みを要素とする文脈ベクトルを用いる手法を報告した。ここで、単語の重みとは語と語の依存関係を反映するものとした。式 (6) で示されるように、重みは依存関係による値と出現頻度とを足し合わせたものであり、従って擬似的にトレーニング量を増加させることが可能であると言える。

また、本手法を用いた多義解消に関する実験について報告した。本手法の導入により、従来の手法に比べ 2% 程度の精度向上が見られた。この事実は本手法の有効性を示すものである。

今後の課題としては、さらに精度を向上させるための妥当な重み式の検討、並びにオープンテストによる実験がある。

## 参考文献

[1] Charniak, E. : "Statistical Language Learning," MIT Press, Cambridge, 1993

[2] Yarowsky, D. : "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," In Proceedings of the 14th COLING, pp.454-460, 1992

[3] Miwa, Y., and Nitta Y. : "Co-occurrence Vectors From Corpora vs. Distance Vectors From Dictionaries," In Proceedings of the 15th COLING, pp.304-309, 1994

[4] 福本 文代, 辻井 潤一 : " コーパスに基づく動詞の多義解消," 自然言語処理, Vol.4, No.2, pp.21-39, 1997