新語義発見のための用例クラスタと辞書定義文の対応付け

田中　博貴　　中村　誠　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{h_tanaka, mnakamur, kshirai}@jaist.ac.jp

1 はじめに

語義曖昧性解消は自然言語処理における重要な基礎技術のひとつである．通常の語義曖昧性解消の問題設定では，特定の文脈中に出現した単語の意味を，辞書などによってあらかじめ定義された語義の中から選択する．ところが，単語の意味は日々変化し，新しい意味や用法も生まれている．したがって，辞書による単語の意味の定義は必ずしも完全ではない．そのため，あらかじめ語義を定義せず，単語のインスタンス (用例) をクラスタリングすることで単語の意味を自動的に弁別する研究が行われている [1, 2, 3, 5]．しかしながら，これらの先行研究では，同じ意味を持つ単語の用例をまとめたクラスタを作成することはできるが，作成された用例クラスタがどのような意味を持つのかといった意味の解釈は行われていない．

そこで，本研究では用例クラスタの意味の解釈を行う．具体的には，自動作成された用例クラスタに対し，その単語の意味が既存の辞書の語義に対応するか，あるいはどの意味にも対応しない新語義なのかを判定する手法を提案する [6]．

2 既存語義への対応付け

本節では，自動作成された用例クラスタに含まれる単語の意味が辞書のどの語義に対応するのかを判定する手法について述べる．なお，ここでは用例クラスタは新語義に対応するのではなく，既存の辞書の語義のいずれかに該当すると仮定する．

2.1 概要

対象単語を w とする．コーパスから抽出された w の複数の用例をクラスタリングし，作成された用例クラスタを C_i とする．ここで C_i は，同じ意味を持つとみなせる単語 w の用例の集合である．一方，S_j は，辞書によって定義された w の語義とする．このとき，用例クラスタ C_i に対応する語義 S_j を以下の手法により決定する．

まず，用例クラスタ C_i を特徴ベクトル $\vec{c}_i$ で表現する．同様に，語義 S_j を特徴ベクトル $\vec{s}_j$ で表現する．C_i に対し，ベクトル間の類似度が最大となる語義 S_j を選択する (式 (1))．

$$S_{selected}(C_i) = \arg\max_{S_j} sim(\vec{c}_i, \vec{s}_j) \qquad (1)$$

ここではベクトル間の類似度はコサイン類似度とする．以下，特徴ベクトル $\vec{c}_i$ と $\vec{s}_j$ の構成方法について述べる．

2.2 用例クラスタの特徴ベクトル

用例クラスタの特徴ベクトル $\vec{c}_i$ は，クラスタ C_i に含まれる用例において，対象語 w の周辺に出現する単語を基に作成する．ただし，クラスタリングによって作成された用例クラスタの中には用例数が少ないものもある．実際に我々が用例のクラスタリングを試みたところ，1 個または 2 個の用例のみで 1 つのクラスタが作成されることもあった．そこで，ベクトルがスパースになるのを避けるため，間接的な単語間の共起も考慮する．

まず，図 1 のような単語の共起行列 A を作成する．

$$A = \begin{pmatrix} \vdots & \\ \cdots\ p_{ij} & \cdots \\ \vdots & \end{pmatrix}$$

$A_t = A_f \cup A_d$（横方向），A_f（縦方向），$\vec{o}(t_j)$（列）

図 1: 単語の共起行列

A_f は，A の行に対応し，クラスタの特徴ベクトル $\vec{c}_i$ の素性となる単語から構成される単語集合である．ここでは，BCCWJ [1]の Yahoo!知恵袋コーパスにおいて，出現頻度上位 10,000 の自立語の集合を A_f とした．ただし，式 (2) に定義する $df(t)$ が 0.01 以上の単語 t は一般的すぎるとみなして除外した．

$$df(t) = \frac{\text{単語 } t \text{ が出現する文書数}}{\text{コーパスにおける全文書数}} \qquad (2)$$

一方，A の列に対応する単語集合 A_t は，用例クラスタにおいて対象単語の周辺に出現すると仮定される単語の集合である．基本的には，A_t は A_f と同じ単語集合とする．ただし，実際には A_t は A_f と A_d の和集合としている．この理由ならびに A_d の定義については 2.3 項で述

[1] http://www.tokuteicorpus.jp/

べる．最後に，行列の要素 p_{ij} は単語 t_i と t_j の文書内共起確率である．正確には，p_{ij} は，j 列目の単語 t_j が出現する文書があったとき，同じ文書内に i 行目の単語 t_i が出現する確率 $P(t_i|t_j)$ である．p_{ij} は Yahoo!知恵袋コーパスから学習する．また，j 列目のベクトルを単語 t_j の共起ベクトル $\vec{o}(t_j)$ とする．

クラスタの特徴ベクトル $\vec{c}_i$ を式 (3) と定義する．

$$\vec{c}_i = \frac{1}{N} \sum_{e_{ik} \in C_i} \sum_{t_l \in e_{ik}} \vec{o}(t_l) \tag{3}$$

e_{ik} は用例クラスタ C_i に含まれる用例を表わし，t_l は用例 e_{ik} の文脈に出現する自立語である．また，N は $\vec{c}_i$ の大きさを 1 にするための正規化定数である．用例の文脈に直接出現する単語 t_l ではなく，その共起ベクトル $\vec{o}(t_l)$ の和を特徴ベクトルとすることにより，対象語 w と間接的に共起する単語の特徴が $\vec{c}_i$ に反映される．

2.3 語義の特徴ベクトル

語義の特徴ベクトル $\vec{s}_j$ は辞書の語釈文から作成する．本研究では語義の定義に用いる辞書として岩波国語辞典 [4] を用いた．ただし，岩波国語辞典の語釈文は，その全てが単語の意味を説明した定義文ではない．そこで，岩波国語辞典の語釈文を以下の 4 つのタイプに分類した．

①定義文 ：単語の意味を説明した文．

②例文 ：その語義の典型的な用例．

③参照見出し ：別の見出し語または語義への参照．

④その他 ：上記 3 つに当てはまらない文．見出しの英語表記，注釈などが該当する．

図 2 は岩波国語辞典における「モデル」の 4 つの語義の語釈文と，そのタイプ分けの例である．これら 4 つのうち，「参照見出し」と「その他」は単語の意味を表わしていると言い難いため，特徴ベクトル $\vec{s}_j$ の作成には用いない方がよいと考えられる．我々は，語釈文を上記 4 つのタイプに分類するため，簡単なパタンマッチによるプログラムを実装した．

次に，「定義文」に分類される語釈文のみを用いて $\vec{s}_j$ を式 (4) のように作成する．

$$\vec{s}_j = \frac{1}{N} \sum_{t_k \in d_j} \vec{o}(t_k) \tag{4}$$

d_j は語義 S_j の語釈文における定義文を，t_k は d_j に出現する自立語を表わす．また，N は正規化定数である．$\vec{s}_j$ は，用例クラスタの特徴ベクトル $\vec{c}_i$ と同様に，t_k の共起ベクトル $\vec{o}(t_k)$ の和を正規化することによって得られる．なお，図 1 の共起行列における単語集合 A_d は，

S_1 ①型。模型。 ③▽↓もけい。

S_2 ①手本。模範。②「これをモデルにしてやれば間違いない」

S_3 ①美術製作の対象となるもの・人。文学作品の人物の素材となる人。

S_4 ①「ファッション　モデル」の略。流行の服装をして、客に見せたり写真に撮らせたりするのが職業の（女の）人。 ④▽ｍｏｄｅｌ

図 2: 「モデル」の語釈文

辞書全体における語釈文のうち「定義文」に出現する自立語の集合である．A_f と A_d の和集合を A_t としたのは，コーパスにおける出現頻度が低いために A_f には含まれないが辞書定義文には出現する自立語に対して，共起ベクトル $\vec{o}(t_k)$ を得るためである．

2.3.1 辞書の用例の利用

辞書の語釈文のうち，「定義文」だけでなく「例文」もまた単語の意味を識別する有力な手がかりになると考えられる．そこで，定義文と例文の両方を用いて，$\vec{s}_j$ を式 (5) のように作成する．

$$\vec{s}_j = \frac{1}{N} \left(\sum_{t_k \in d_j} \vec{o}(t_k) + \sum_{t_l \in e_j} w_e \cdot \vec{o}(t_l) \right) \tag{5}$$

e_j は語義 S_j の語釈文中の例文を，t_l は e_j に出現する自立語を表わす．また，w_e は例文に出現する自立語の共起ベクトルに対して与えられる重みである．ここでは，コーパスから作成された用例クラスタと辞書の語義との類似度を測ることを目的としている．意味の説明文である「定義文」よりも，語義の使用例である「例文」の方が，コーパスから作成された用例クラスタの文脈に出現する単語と似ている単語が出現する傾向が強いと予想される．そのため，例文に出現する自立語の共起ベクトルに高い重みを与えるようにした．w_e の値は実験的に決定する．

2.3.2 特徴ベクトルの補正

一般に，辞書における定義文や例文の長さは短いため，作成された語義の特徴ベクトル $\vec{s}_j$ はスパースになりやすい．特徴ベクトルがスパースであるとは，ここでは多くの素性に対する特徴ベクトルの値が 0 となる状態を指す．例えば，図 2 の語義 S_1 においては，定義文に出現する自立語の数は 2 つしかなく，$\vec{s}_j$ におけるほとんどの素性の値が 0 になる．

用例クラスタと辞書の語義との対応付けを試みた結果，定義文が比較的長くスパースではない特徴ベクトルを持

つ語義の方が，定義文が短くスパースな特徴ベクトルを持つ語義と比べて，一貫して用例クラスタの特徴ベクトルとの類似度が高くなる傾向が強いことがわかった．その結果，用例クラスタは常に同じ語義に対応付けられてしまう．この問題を解決するためには語義の特徴ベクトルのスパースネスを緩和する必要がある．

そこで，語義の特徴ベクトルを補正し，値が 0 となる素性の数を減らすことを試みた．具体的には，以下の式 (6) によって補正後の特徴ベクトル $\vec{s}'_j$ を作成した．

$$\vec{s}'_j = \frac{1}{N}\left(\vec{s}_j + \sum_{t_m \in A_f} w_c \cdot \mathrm{s_j}(t_m) \cdot \vec{o}(t_m)\right) \quad (6)$$

ここで，$\vec{s}_j$ は式 (4) や (5) で作成した語義 S_j の元の特徴ベクトル，A_f は $\vec{s}_j$ の素性となる単語集合，$\mathrm{s_j}(t_m)$ は $\vec{s}_j$ における素性 t_m に対するベクトルの値，N は正規化定数である．式 (6) によるベクトルの補正は，元の特徴ベクトルの素性 t_m に対してその共起ベクトル $\vec{o}(t_m)$ をさらに足すことにより，スパースネスを緩和している．なお，w_c は元のベクトル $\vec{s}_j$ に対する $\vec{o}(t_m)$ への重み付けに用いるためのパラメタである．w_c は実験的に決定する．

3 新語義の判定

本節では，用例クラスタが，辞書における既存の語義のいずれにも該当しない新語義の用例を集めたものであるかを判定する手法について述べる．ある対象単語に対し n 個の用例クラスタ C_i が作成されたとする．各 C_i に対し，既存語義近接度 K_i を求める．K_i は，C_i の持つ意味が既存の辞書の意味にどれだけ近いかを表わす指標で，式 (7) のように既存の語義 S_j との類似度の最大値と定義する．

$$K_i \;=\; \max_j \; sim(\vec{c}_i, \vec{s}_j) \quad (7)$$

直観的には，既存語義近接度が小さければ小さいほど，その用例クラスタは新語義である可能性が高い．そこで，用例クラスタを K_i の降順にソートする．以下，用例クラスタ $C_1 \sim C_n$ は，K_i の大きい順に並んでいるものとする．

次に，上記のように並べられた C_i に対し，既存語義と新語義とを分ける境界を発見する (図 3)．まず，相対既存語義近接度 RK_i を，最も大きい既存語義近接度 K_1 に対する K_i の相対値 ($= K_i/K_1$) とする．さらに，隣接する用例クラスタ C_i と C_{i+1} の相対既存語義近接度の差を $DRK_{i,i+1}$ とする (式 (8))．

$$DRK_{i,i+1} \;=\; RK_i - RK_{i+1} \quad (8)$$

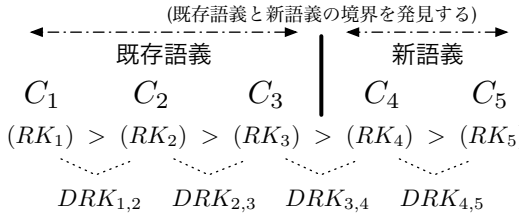

図 3: 新語義の判定

ここでは新語義の判定を行う 2 つの手法を提案する．

- 手法 1
  $1 \le i \le N-1$ のうち，$DRK_{i,i+1}$ が最も大きい i を見つけ，$i+1$ 番目以降の用例クラスタ $C_{i+1} \cdots C_N$ は新語義であると判定する．これは，既存語義近接度の差が大きいところが既存語義と新語義の境界になっているという仮定に基づく．
- 手法 2
  手法 1 の条件に加え，最大の $DRK_{i,i+1}$ の値が閾値 T_k よりも大きいときに限り，用例クラスタ $C_{i+1} \cdots C_N$ を新語義と判定する．T_k 以下の場合は新語義に相当する用例クラスタは存在しないものとする．すなわち，既存語義近接度の差が十分大きくなければ既存語義と新語義との境界とはみなさない．

手法 2 で閾値 T_k を設定する際，用例クラスタと辞書の語義との類似度の大きさは対象単語によってばらつきがあるため，既存語義近接度 K_i に対して閾値を設定することは困難である．そのため，相対化した既存語義近接度 RK_i に対して閾値 T_k を設定した．

4 予備実験

4.1 実験データ

提案手法を評価する予備実験を行った．まず，対象単語として以下の 23 個の単語を用いた．

> モデル，ネタ，カバー，ウイルス，ソース，肉，サービス，地方，アルバム，コード，自分，場合，時間，意味，電話，一緒，目，以前，代，顔，系，郵便，反応

各対象単語に対し，Yahoo!知恵袋コーパスの中から 100 個の用例をランダムに選択した．

本研究では，用例集合に対してクラスタリングを行い，同じ語義を持つ用例を集めたクラスタを自動作成し，各クラスタに対応する辞書の語義を選択することを目的としている．しかし，自動的に作成されたクラスタは，違う語義を持つ用例が 1 つのクラスタにまとめられるという誤りを含む．ここでは，用例クラスタと語義との対

応付け，ならびに新語義の判定手法を評価するために，人手で作成した完全に正しい用例クラスタを実験に用いた．具体的には，コーパスから抽出された用例に対して人手で語義を付与し，同じ語義を持つ用例をまとめてクラスタを作成した．用例に付与する語義は岩波国語辞典の中分類の語義とした．これは比較的荒い意味分類である．また，岩波国語辞典に該当する語義がない場合は新しい語義を定義した．

4.2 実験結果

人手で作成された用例クラスタに対し，2 節で提案した手法によって対応する辞書の語義を選択した．ここでは，語義の特徴ベクトル $\vec{s_j}$ の作成方法の違いに応じて，以下の 4 つの手法を比較した．

M_d ：式 (4) によって $\vec{s_j}$ を作成する．すなわち，辞書の語釈文のうち「定義文」に含まれる単語のみを用いる手法．

$M_{d,e}$ ：式 (5) によって $\vec{s_j}$ を作成する．すなわち，辞書の語釈文のうち「定義文」と「例文」に含まれる単語を用いる手法．

M_d^c ：式 (4) によって $\vec{s_j}$ を作成した後，式 (6) によって特徴ベクトルの補正を行う手法．

$M_{d,e}^c$ ：式 (5) によって $\vec{s_j}$ を作成した後，式 (6) によって特徴ベクトルの補正を行う手法．

実験結果を表 1 に示す．

表 1: 語義の対応付けの実験結果

	M_d	$M_{d,e}$	M_d^c	$M_{d,e}^c$
クラスタ数	63	63	63	63
正解クラスタ数	26	31	31	39
正解率	0.41	0.49	0.49	0.62

表 1 において,「クラスタ数」は対象単語 23 語に対して人手で作成した用例クラスタの総数 (新語義に対応する用例クラスタは除く),「正解クラスタ数」は正しい語義に対応付けられた用例クラスタの数，正解率はその割合を表わす．なお，式 (5) における w_e，式 (6) における w_c はいくつかの値を試し，正解率の一番大きい値を定めた．その結果，$w_e = 2$，$w_c = 5$ となった．ただし，これらのパラメタの調整はテストデータを用いて行っているという点で，今回の実験はクローズドテストである．

表 1 の結果から，辞書の定義文だけでなく例文を用いて語義の特徴ベクトルを作成した方が，また値が 0 となるベクトルの素性の数を減らすための補正を行う方が，辞書の語義との対応付けの正解率が向上することがわかった．ただし，手法 $M_{d,e}^c$ における正解率は 0.62 であり，十分高いとはいえず，改善の必要がある．

次に，4.1 項で作成した全ての用例クラスタに対し，3 節で提案した方法を用いて，用例クラスタが新語義であるか否かを判定する実験を行った．判定の精度，再現率，F 値を表 2 に示す．

表 2: 新語義判定の実験結果

手法	N_1	N_2	N_2	N_2
T_k	–	0.03	0.025	0.02
精度	0.43	0.62	0.63	0.57
再現率	0.65	0.50	0.60	0.65
F 値	0.52	0.55	0.62	0.60

表 2 において，N_1，N_2 はそれぞれ 3 節で述べた「手法 1」「手法 2」を表わす．N_2 については閾値 T_k を変えて実験を行った．新語義判定の F 値は，$T_k = 0.025$ のときの手法 2 が最大で 0.62 であった．

5 おわりに

本論文では，コーパスから作成された用例クラスタに対する意味の解釈，すなわち用例クラスタに対応する辞書の語義を選択する手法と，辞書のどの語義にも対応しない新語義であるかを判定する手法について述べた．今後の課題としては，提案手法の改良・洗練，自動作成された用例クラスタに対する実験，各手法におけるパラメタの最適化方法の検討などがあり，これらに順次取り組む予定である．

参考文献

[1] Stefan Bordag. Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the EACL*, pp. 137–144, 2006.

[2] Fumiyo Fukumoto and Jun'ichi Tsujii. Automatic recognition of verbal polysemy. In *Proceedings of the COLING*, pp. 762–768, 1994.

[3] 九岡佑介, 白井清昭, 中村誠. 複数の特徴ベクトルのクラスタリングに基づく単語の意味の弁別. 第 14 回言語処理学会年次大会, pp. 572–575, 2008.

[4] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 岩波書店, 1994.

[5] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, Vol. 24, No. 1, pp. 97–123, 1998.

[6] 田中博貴. 用例のクラスタリングに基づく単語の新語義の発見. Master's thesis, 北陸先端科学技術大学院大学, 3 2009.