複数の観点から定義された用例間類似度に基づく語義識別

中西　隆一郎　　白井　清昭　　中村　誠
北陸先端科学技術大学院大学 情報科学研究科
{s0910041, kshirai, mnakamur}@jaist.ac.jp

1 はじめに

単語の意味は日々変化し，辞書で定義されていない新しい意味や用法も生まれている．著者らは，辞書にない語の意味を「新語義」と呼び，これをコーパスから自動的に発見する研究に取り組んでいる [3, 9]．その手法の概略は以下の通りである．まず，対象単語の用例をコーパスから収集する．次に，用例集合をクラスタリングし，同じ意味を持つ用例をまとめたクラスタを作成する．最後に，用例クラスタと辞書の語義との類似度を計算し，どの語義とも似ていないクラスタを新語義の用例とみなして検出する．コーパスから新語義を発見することができれば，辞書編纂作業のサポートや自然言語処理用辞書の整備に貢献すると期待される．

本論文では，上記の処理のうち，用例クラスタリングの新しい手法について述べる [5]．提案手法は，同じ意味を持つ用例のクラスタを作成する際に，用例間の類似度を複数の観点から計算することに特徴がある．

2 関連研究

用例のクラスタリングは，辞書を使わずに語義を自動的に推定する語義推定 (Word Sense Induction) もしくは語義識別 (Word Sense Descrimination) と呼ばれるタスクとみなせる．語義識別に関する研究の多くは，用例を特徴ベクトルで表現し，ベクトル間の類似度を基に用例をクラスタリングする．Schütze は，コーパスから単語の共起行列を学習し，それを基に対象語と他の語との二次共起 (間接共起) の情報を反映した特徴ベクトルを作成し，Buckshot と呼ばれるアルゴリズムでクラスタリングを行う手法を提案している [8]．また，意味解析に関する評価型ワークショップ SemEval では，過去2回にわたって英語を対象とした語義識別のタスクが実施され，用例クラスタリングに関するシステムが報告されている [1, 4]．

これらの先行研究では，用例は 1 つの特徴ベクトルで表現される．しかしながら，一般に，語の意味の類似性は様々な観点から認められる．例えば，図 1 に示す「サービス」の用例について考察してみよう．岩波国語辞典によれば,「サービス」には①客に対するもてなし，

(a) 時まで、あとのぶんは サービス 残業 … というわけ
その差約７００時間が サービス 残業。現在過労死が若

(b) ケーキとシャンパンを サービス されたんです。ＣＡか
とりました。飲み物を サービス したり、一緒に写真撮

(c) ファイアーウォールの サービス を開始しようとしたと
う名前でＡｐａｃｈｅ サービス をインストールするに

図 1: 「サービス」の用例

②奉仕，などの意味がある．図 1 (a) の「サービス」は，直後の単語が「残業」であることから②の意味を持つと考えられる．一方，図 1 (b) は「ケーキ」「シャンパン」「飲み物」のような飲食物が周辺に出現していることから①の意味を持つと考えられる．図 1 (c) の「サービス」はコンピュータに関連するテキストに出現することから，岩波国語辞典では定義されていない意味 (ネットワーク上でサーバが提供する「サービス」) であるといえる．すなわち，語の意味は，直前・直後の単語で識別できる場合，文脈に出現する単語で識別できる場合，テキストのトピックによって識別できる場合などがある．

このように，語の意味の類似性は様々な観点で測ることができる．しかし，用例を 1 種類の特徴ベクトルで表現するだけでは，上記のような多様な観点を捉えることは難しい．本研究では，用例を異なる観点から見た複数の特徴ベクトルで表現し，用例クラスタリングの精度を向上させることを目的とする．

著者らは，複数の特徴ベクトルに基づく用例のクラスタリング手法について既に検討している [3]．まず，用例を 4 種類のベクトルで表現し，それぞれの特徴ベクトルでクラスタリングを 4 回実施する．次に，得られたクラスタ集合の良さを，クラスタ内の要素が互いに似ているか，異なるクラスタは互いに似ていないかという観点から評価し，最良のクラスタ集合を選択する．この方式では，対象単語別にみれば，用例クラスタを作成する際に最終的に使用される特徴ベクトルは 1 種類である．しかしながら，上記の考察のように，同じ単語でも語義によって異なる観点から類似性が認められることから，複数の特徴ベクトルを同時に考慮して用例クラスタを作成する方が望ましい．次節ではその一手法を提案する．

3 提案手法

ここでは用例クラスタリングのタスクを以下のように定義する．対象単語を w とする．w を含む用例の集合 $W = \{w_i\}$ が与えられたとき，同じ語義を持つ用例のクラスタに分割し，クラスタの集合 $\mathcal{C} = \{C_k\}$ を得る．

3.1 特徴ベクトル

用例 w_i を以下の 4 種類の特徴ベクトルで表現する [3].

隣接ベクトル　w_i の直前または直後に現われる単語で w_i を特徴付けるベクトル．具体的には，w_i の前後 2 語の単語の出現形ならびに品詞をベクトルの素性とする．

文脈ベクトル　w_i の周辺に現われる単語で w_i を特徴付けるベクトル．また，w_i の周辺に直接現われる単語 x だけではなく，x と同一のトピックを持つ単語もベクトルの素性とすることにより，ベクトルの過疎性を緩和する．単語のトピックは LDA(Latent Dirichlet Allocation) によってコーパスから自動的に推測する．

連想ベクトル　文脈ベクトルと同じく，w_i の周辺に現われる単語で w_i を特徴付けるベクトル．ただし，ベクトルの過疎性を緩和するために，事前にコーパスから作成された単語の共起行列を用いる．単語の共起行列の列を，ある単語が別の単語とどの程度共起しやすいかを表わす共起ベクトルとみなし，w_i の文脈に出現する単語の共起ベクトルの和を文脈ベクトルと定義する．

トピックベクトル　PLSI (Probabilistic Latent Semantic Indexing) によって推定されるトピックによって w_i を特徴付けるベクトル．具体的には，w_i を含む文書を d_i としたとき，$P(z_l|d_i)$ (z_l は PLSI の隠れ変数 (トピック)) を素性とするベクトルを作成する．

これらの特徴ベクトルは用例間の類似度を計算するために用いるが，隣接ベクトルは図 1 (a) の例のように直前・直後に出現する単語が似ているかという観点，文脈ベクトルと連想ベクトルは図 1 (b) のように周辺文脈に出現する単語が似ているかという観点，トピックベクトルは図 1 (c) のようにテキストのトピックが似ているかという観点で語義の類似性を測っている．用例をクラスタリングする際，これら 4 つの特徴ベクトルを併用することで，様々な観点から語義の類似性を捉えることを狙う．

3.2 クラスタリング

図 2 は本手法におけるクラスタリングアルゴリズムの擬似コードである．本手法は凝集型クラスタリングを拡張したアルゴリズムである．まず，初期のクラスタ集合 $\mathcal{C}$ を作成する (1 行目)．次に，全てのクラスタの組についてクラスタ間類似度 $sim(C_i, C_j)$ を計算し，それが最大となる C_i, C_j を求める (3 行目)．両者を併合したクラスタ C_k を作成し (4 行目)，その重心ベクトルと後述するクラスタラベル $L(C_k)$ を更新した後 (5 行目)，$\mathcal{C}$ を更新する (6 行目)．この処理を停止条件を満たすまで繰り返す (2 行目)．

入力=用例集合 W，出力=クラスタ集合 $\mathcal{C}$
1 個々の用例を 1 つのクラスタとみなして初期の $\mathcal{C}$ を作成
2 while (停止条件) do
3 　$sim(C_i, C_j)$ が最大となる C_i,C_j を選択
4 　C_i と C_j を併合したクラスタ C_k を作成
5 　C_k の重心ベクトルと $L(C_k)$ を更新
6 　クラスタ集合 $\mathcal{C}$ を更新 ($\mathcal{C}$ から C_i,C_j を削除し，C_k を追加)
7 done

図 2: クラスタリングアルゴリズムの概要

3.2.1 クラスタ間類似度

クラスタ間類似度は 3.1 項で述べた 4 つの特徴ベクトルを用いて式 (1) のように計算する．

$$sim(C_i, C_j) = \max_{v \in \{\text{ 隣接, 連想, 文脈, トピック }\}} s(v, C_i, C_j) \quad (1)$$

$s(v, C_i, C_j)$ は特徴ベクトル v によって計算されるクラスタ間の類似度である．具体的には，用例を特徴ベクトル v で表現したときのクラスタの重心ベクトル[1]のコサイン類似度と定義する．式 (1) は，クラスタ間の類似度を，隣接，文脈，連想，トピックベクトルで計算される類似度の最大値と定義している．これは，4 つの特徴ベクトルで考慮されている複数の観点のうち，どれか 1 つについてでも類似度が十分高ければ，それらは同じ語義を持つ可能性が高いという考えに基づく．

さらに，クラスタを作成する際には，同一の特徴ベクトルによる類似度が高い用例をまとめるという制約を設ける．例えば，図 2 の 4 行目で最初に類似度が最大となるクラスタの組を併合して新しいクラスタを作成したとき，式 (1) で 4 つの特徴ベクトルのうち隣接ベクトルの類似度が最大であった場合には，以後は隣接ベクトルの類似度が十分高いときのみそのクラスタに新しい要素を併合する．作成されたクラスタは隣接，文脈，連想，ト

[1] クラスタ内の要素の特徴ベクトルを平均したベクトル．

ピックベクトルのいずれかによって計算される類似度が高い要素をまとめたものとなる．これにより，クラスタがどのような観点で似ている用例がまとめられたかを容易に解釈できる．

この制約はクラスタラベル $L(C_k)$ を導入することで実現する．$L(C_k)$ はクラスタ C_k がどの特徴ベクトルの観点から用例をまとめたかを示すラベルである．初期クラスタでの $L(C_k)$ は「未定」とする．また，C_i と C_j が併合されて C_k が作成されたとき，式 (1) の $s(v, C_i, C_j)$ が最大となるベクトルの種類に応じて「隣接」「文脈」「連想」「トピック」のいずれかを $L(C_k)$ とする．さらに用例間類似度 $sim(C_i, C_j)$ を式 (2) のように再定義する．

$$sim(C_i, C_j) = \begin{cases} \text{式 (1)} & \text{if } L(C_i)=L(C_j)=\text{未定} \\ s(L(C_i), C_i, C_j) & \text{if } L(C_i)=L(C_j) \text{ or } L(C_j)=\text{未定} \\ s(L(C_j), C_i, C_j) & \text{if } L(C_i)=L(C_j) \text{ or } L(C_i)=\text{未定} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

式 (2) の 3,4 行目は，2 つのクラスタのラベルが一致しているか，どちらか一方が「未定」のとき,「未定」でないクラスタラベルの特徴ベクトルの類似度をクラスタ間類似度とすることを表わす．また，5 行目は，C_i と C_j のクラスタラベルが異なるときは類似度を 0 とし，両者を併合しないことを表わす．

3.2.2 ベクトル間類似度の正規化

予備実験により，4 つの特徴ベクトルによって計算されるクラスタ間類似度の値には大きな差があることがわかった．式 (1) で 4 つの特徴ベクトルによるコサイン類似度を単に比較するだけでは，ベクトル間類似度が平均的に高い特徴ベクトルのみが常に選択される可能性がある．4 つの特徴ベクトルによる類似度の値を公平に比較するために，ベクトル間類似度を正規化する．

まず，特徴ベクトル v によるベクトル間類似度の標本を X_v とする．X_v は，用例集合 W における全ての用例の組に対する特徴ベクトル v のコサイン類似度の値の集合とする．次に，正規化された類似度 s_R を式 (3) のように定義する．

$$s_R(v, C_i, C_j) = \frac{s(v, C_i, C_j) - min_v}{max_v - min_v} \quad (3)$$

min_v と max_v は，それぞれ標本 X_v における類似度の値の最小値，最大値である．s_R は，C_i と C_j の類似度の大きさを X_v 上で相対的に評価している．

s_R による正規化は，標本 X_v 内における類似度の分布の偏りは考慮されていない．そこで，ベクトル間類似度を正規化する別の方法として式 (4) を考える．

$$s_{SD}(v, C_i, C_j) = \frac{10(\ s(v, C_i, C_j) - \mu_v\ )}{\sigma_v} + 50 \quad (4)$$

μ_v と σ_v は，それぞれ標本 X_v における平均と標準偏差である．ただし，用例間の類似度が 0 になる場合は X_v から除く．s_{SD} は標本 X_v における $s(v, C_i, C_j)$ の偏差値である．4 節の実験では，これら 2 つの正規化の手法について評価する．

3.2.3 停止条件

以下の 2 つの条件を同時に満たすとき，クラスタリングを停止する (図 2 の 2 行目)．

1. クラスタの数が T_n 以下である．
2. 大きさが最大のクラスタの要素数の用例総数に対する割合が T_s $(0 < T_s < 1)$ より大きい．

2. の条件はある程度の数の用例をまとめたクラスタが作成されるまでクラスタリングを継続させるために設定した．4 節の実験では仮に $T_n = 10$，$T_s = 0.2$ とした．

4 実験

評価実験には SemEval-2 日本語タスク [6] の訓練データを利用した．同タスクの 40 語の評価単語に対し，それぞれ 40〜50 語の用例を訓練データから抽出し，用例集合 W を作成する．W をクラスタリングして得られたクラスタ集合 $\mathcal{C}$ を，用例に付与されている語義を正解ラベルとして評価する．一般に，語義識別のタスクでは，同じ語義を持つ用例をまとめてクラスタを作成することと，語義の数を推定する (語義と同じ数だけクラスタを作成する) ことの 2 つが要求される．しかし，本研究は，作成された用例クラスタに対し，それが辞書に定義されている語義か否かを自動判定することで，コーパスから新語義を発見することを想定している．そのため，必ずしも語義の数を推定する必要はなく，同じ語義を持つ用例をまとめたクラスタを作成することが要求される．上記の理由から，今回の実験ではクラスタの評価基準として Purity [2] と Homogeneity [7] を採用した．これらはクラスタを構成する要素のラベルがどれだけ一致するかを評価する指標である．

40 語の評価単語に対する Purity と Homogeneity の平均を表 1 に示す．表の 2,3 行目は提案手法で，ベクトル間類似度を正規化する方法として式 (3) と式 (4) を用いた場合を表わす．4 行目は 4 つの特徴ベクトルを単独で用いたクラスタリング結果から評価単語ごとに最良のものを自動選択する九岡らの手法 [3] を表わす．5〜8 行目

表 1: 実験結果 (1)

	Purity	Homogeneity
提案手法 (s_R)	0.771	0.357
提案手法 (s_{SD})	0.800	0.472
[九岡ら 2008]	0.751	0.294
隣接	0.811	0.487
文脈	0.750	0.282
連想	0.749	0.285
トピック	0.765	0.374
BL	0.745	0.327

は隣接，文脈，連想，トピックベクトルを単独で用いたときの結果である．最後の「BL」はベースラインを表わし，凝集型クラスタリングアルゴリズムで併合する要素の組をランダムに選択する手法である．

提案手法は九岡の手法よりも Purity，Homogeneity ともに上回ることから，複数の特徴ベクトルを利用する手法として適しているといえる．また，正規化の手法としては s_{SD} の方が s_R よりも良かった．しかし，提案手法は隣接ベクトルのみを使用する手法より少し劣る．この要因を調べたところ，単独のベクトルを使用した場合には，どの要素ともマージされずに 1 つの要素だけで構成されるクラスタが多いことがわかった．このようなクラスタは明らかに有用ではない．しかし，Purity や Homogeneity はクラスタ内に同じラベルを持つ要素がどれだけまとめられるかを評価する指標なので，1 要素で構成されるクラスタが多いときには高く見積られる．

表 2: 実験結果 (2)

	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{C}_{\geq 2}\rvert$	AP
提案手法 (s_R)	400	258	0.857
提案手法 (s_{SD})	396	347	0.828
隣接	400	211	0.819
文脈	400	99	0.758
連想	400	103	0.772
トピック	400	233	0.767

表 2 は提案手法を別の観点で評価した結果である．$|C|$ は評価単語 40 語の全てについて作成されたクラスタの総数を，$|C_{\geq 2}|$ はそのうち 2 つ以上の要素から構成されているクラスタの数を表わす．また，AP の定義は式 (5) であり，要素数が 2 以上のクラスタ C_i について，C_i 内で頻度が最大となる語義が占める割合 ($\mathrm{max\_prec}(C_i)$) の平均である．

$$AP = \frac{1}{|\mathcal{C}_{\geq 2}|} \sum_{C_i \in \mathcal{C}_{\geq 2}} \mathrm{max\_prec}(C_i) \qquad (5)$$

提案手法は，単独のベクトルを用いる手法と比べて $|C_{\geq 2}|$ が大きいことから，他のどの用例ともマージされない用例の数が少ないという意味ではクラスタリングに成功しているといえる．また，提案手法の AP も単独のベクトルを用いる手法と比べて高い．すなわち，2 個以上の要素をまとめて作成されたクラスタについては，同じ語義を持つ用例をまとめる傾向が強い．したがって，新語義を発見するための用例クラスタリング手法として，複数の特徴ベクトルを同時に考慮する提案手法は 1 種類の特徴ベクトルのみを用いる手法よりも優れていると言える．類似度の正規化の手法 s_R と s_{SD} を比較すると，AP は s_R の方が大きいが，$|C_{\geq 2}|$ は S_{SD} の方が大きかった．

5 おわりに

本論文では，用例を複数の特徴ベクトルで表現することで異なる観点から語の意味の類似性を定量化し，用例をクラスタリングする手法を示した．今後は，作成された用例クラスタを分析し，我々が狙いとしているように，複数の観点から見た用例クラスタが作成されているのかを調査したい．また，我々は用例クラスタが新語義か否かを判定する手法についても研究を進めており，本研究の成果と合わせて，コーパスから新語義を発見する手法を確立したい．

参考文献

[1] Eneko Agirre and Aitor Soroa. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of SemEval-2007*, pp. 7–12, 2007.

[2] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. *GLDV-Journal for Computational Linguistics and Language Technology*, Vol. 20, No. 1, pp. 19–62, 2005.

[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. 言語処理学会第 14 回年次大会発表論文集, pp. 572–575, 2008.

[4] Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of SemEval-2010*, pp. 63–68, July 2010.

[5] 中西隆一郎. 複数の特徴ベクトルを同時に考慮した語義識別. Master's thesis, 北陸先端科学技術大学院大学, 3 2011.

[6] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. SemEval-2010 task: Japanese WSD. In *Proceedings of SemEval-2010*, pp. 69–74, 2010.

[7] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 EMNLP-CoNLL Joint Conference*, pp. 410–420, 2007.

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

[9] 田中博貴, 中村誠, 白井清昭. 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. 言語処理学会第 15 回年次大会発表論文集, pp. 590–593, 2009.