用例間類似度測定のための属性重みの推定

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1　はじめに

ある単語の用例を集め、それら用例をその単語の語義に基づいて分類するタスクに取り組んでいる。本論文では、このタスクの本質的課題となる用例間距離の測定について論じる。

語義別の用例は本格的な意味解析にとって有用である。例えば、語義別の用例を訓練データとして利用することで語義の曖昧性を解消する分類器を学習することできる [5]。また動詞の格フレームの自動構築 [6] は、動詞の語義別の用例があれば容易に行える。またシソーラスの自動構築 [2] においては、通常、名詞の多義性は無視されるが、語義別の用例があれば、語義を考慮したシソーラスの作成も容易である。辞書の編纂、語学学習においても、語義別の用例は有用であろう。

ある単語 w の語義別の用例を収集するには、単語 w を含む用例をコーパスから抽出し、その用例中の単語 w の語義を識別できればよい。つまり語義別用例の収集は、語義の曖昧性解消のタスクとして扱える。そのために（半）教師有り学習を用いるアプローチによって解決できそうであるが、以下に示す２つの問題から、そのアプローチによる解決は困難である。第１の問題は、訓練データの作成コストである。教師有り学習では大量の訓練データを必要とし、対象となる単語が多い場合、その作成コストが大きすぎる。第２の問題は、語義の設定である。（半）教師有り学習で本タスクに挑む場合、単語 w の語義を予め設定しておかなければならない。単語 w が与えられたときに、w の語義を内省により列挙することは困難である。例えば、語義の粒度を均一に保つことは難しいし、マイナーな語義を見落としてしまう危険性もある。そのためにこのタスクに対しては用例を語義に基づいてクラスタリングするアプローチが妥当である。

ただし用例のクラスタリグを行うためには用例間の類似度（あるいは距離）を測る必要がある。用例間の類似度の設定方法として確立された手法は存在せず、これまでアドホックに対処されていた。そのためこのタスクの精度は、実質、クラスタリング手法ではなく、用例間の類似度の適切さに大きく依存している。

本論文では用例間類似度を線型モデルで表し、そのパラメータ（属性重み）を推定することを試みる。この際、用例間類似度の値は、手作業であっても設定することができないため、推定のための訓練データを構築できないという問題がある。この問題に対して、ここでは以下のような対策をとった。まず線型モデルのパラメータを経験的な値で与え、仮のモデルを作成する。次に語義識別タスクに対する訓練データを用いて、用例対が同じクラスに属する場合は、仮のモデルから類似度を与え、同じクラスに属さない場合は、類似度を 0 とすることで訓練データを作成する。この訓練データをもとにパラメータの推定を行う。

実験では SENSEVAL2 の日本語辞書タスク [9] で用いられた名詞 50 単語を対象とした。SENSEVAL2 で提供されたそれら単語に対する訓練データを用例のセットとし、(1) 単純な線型モデルによる用例間類似度、(2) 経験的なパラメータ値を用いた用例間類似度、(3) 本手法により得られたパラメータ値を用いた用例間類似度の 3 つを用いてクラスタリングを行った。エントロピーの平均値によるクラスタリングの評価を行ったところ、(3)、(2)、(1) の順でよい結果を得ることができた。

2　用例に対する素性リスト

クラスタリグを行う場合、対象のデータが実数値ベクトルで表現されている必要はない。データ間の類似度（あるいは距離）が設定されていれば十分である。多くのクラスタリグ手法は対象のデータが実数値ベクトルで表現されていることを前提としているので、この点について注記しておく。

個々のクラスタリングのアルゴリズムをみれば、一見、実数値ベクトルが必要そうにみえる。例えば k-means ではクラスタの重心を求める処理が入るので、この点だけ見れば対象のデータは実数値ベクトルで表現する必要がある。しかし重心を求めるのは、そのクラスタとデータ間の距離を求めるために行っているのであり、クラスタとデータ間の距離はそのクラスタを構成する各要素との距離が設定されていれば、重心を求めることなしに求めることができる。他のクラスタリングアルゴリズムもほぼ同様である。つまりクラスタリング手法ではクラスタ間の類似度を測ることができればよく、クラスタ間の類似度は各データ間の類似度が与えられれば求めることができる。

ここでは用例間の類似度を定義するために、用例を素性リストで表現する。素性としてここでは以下のも

のを利用する。

ee1	直前の 2 単語（表記）
ee2	直後の 2 単語（表記）
e1	直前の単語
e2	直後の単語
e3	前方と後方の内容語それぞれ 2 つまで
e4	e3 の分類語彙表の番号

例を示す。対象の単語を「記録」として、以下の用例を考える（形態素解析され各単語は原型に戻されているとする）。

過去/最高/を/記録/する/た/。

この場合、「記録」の直前、直後の 2 単語の表記は「最高を」と「した」なので、‘ee1=最高を’，‘ee2=した’となる。また直前、直後の単語は「を」と「する」なので、‘e1=を’，‘e2=する’となる。次に、「記録」の前方の内容語は「過去」、「最高」なので、ここから「記録」に近い順に 2 つとり、‘e3=過去’、‘e3=最高’が作られる。また「記録」の後方の内容語は「する」だけであり、‘e3=する’が作られる。次に「最高」の分類語彙表 [7] の番号を調べると、3.1920_4 である。ここでは分類語彙表の 4 桁目と 5 桁目までの数値をとることにした。つまり‘e3=最高’に対しては、‘e4=3192’ と ‘e4=31920’ が作られる。同様に「過去」の分類語彙表の番号 1.1642_1 から‘e4=1164’ と ‘e4=11642’ が作られる。次は「する」の分類語彙表を調べるはずだが、ここでは平仮名だけで構成される単語の場合、分類語彙表の番号を調べないことにした。これは平仮名だけで構成される単語は多義性が高く、無意味な素性が増えるので，その問題を避けたためである。もしも分類語彙表上で多義になっていた場合には、それぞれの番号に対して並列にすべての素性を作成する。

結果として、上記の用例に対して以下の 11 個の素性を要素とする素性リストが得られる。

```
( ee1=最高を，ee2=した，e1=を，e2=する，
  e3=最高， e3=過去，e3=する，e4=3192，
  e4=31920，e4=1164， e4=11642 )
```

3 用例間の類似度

3.1 類似度のモデル

用例 s_1 と s_2 の類似度 sim を以下で定義した。

$$\begin{aligned} sim(s_1,s_2) \quad = \quad & a\cdot M(ee1)+b\cdot M(ee2) \\ & +c\cdot M(e1)+d\cdot M(e2) \\ & +e\cdot M(e3)+f\cdot M(e4) \end{aligned}$$

ここで $M(x)$ は s_1 と s_2 の素性リスト中の素性 x の一致数を表す。上記の式は類似度測定の一種のモデルであり、これをここでは**線形モデル**と呼ぶことにする。線形モデルでは素性同士がどれくらい一致しているかを調べ、素性毎に重みをつけている形をしている。重みがすべて 1 の場合は、素性リストの全要素を次元で表現し、用例を高次元実数値ベクトルで表現し、それらの余弦尺度によって類似度を測ることに対応している。

問題は線形モデルのパラメータ a,b,c,d,e,f（つまり素性毎に重み）をどのように設定するかである。

3.2 訓練データの作成

線形モデルのパラメータを求めるために、重回帰分析を利用することにする。重回帰分析では線型モデルのパラメータを最小自乗法で求める。しかしここで問題がある。重回帰分析では訓練データとして観測値が必要である。この場合は、訓練データとして用例間の実際の類似度が必要であるが、そのようなものは手作業であっても与えることはできない。

この対処として、ここでは、経験的なパラメータ値から用例間の類似度を与えることにした。用例中の対象単語の語義の類似度は手作業であっても与えることはできないが、異なる語義であることは判定することができる。つまり、異なる語義である場合は類似度を 0 とし、同じ語義である場合は、経験的なパラメータ値を用いて類似度を与える。

これによって訓練データを作成することができる。具体的にここではこの経験的なパラメータ値として、以下を用いた。

```
a = b = 10，  c = d = 5，e = f = 1
```

3.3 最小自乗法によるパラメータ推定

n 個のデータがそれぞれ m 次元の実数値ベクトルで表現されているとする。i 番目のデータ $x^{(i)}$ を以下で表現する。

$$x^{(i)}=(x_1^{(i)},x_2^{(i)},\cdots,x_m^{(i)})$$

$x^{(i)}$ と $x^{(i)}$ の観測値 $y^{(i)}$ に線型モデルを当てはめる。

$$y^{(i)}=\sum_{j=1}^{m}a_jx_j^{(i)}+e_i$$

ここで $e_i\sim N(0,\sigma^2)$ である。ここから以下の残差平方和を最小にするようにパラメータを推定する。

$$\sum_{i=1}^{n}e_i=\sum_{i=1}^{n}\left(y^{(i)}-\sum_{j=1}^{m}a_jx_j^{(i)}\right)^2$$

以下、各パラメータで偏微分を行い、極値問題を解けばパラメータが求まる。これは最小自乗法によりパラメータ推定を行っていることと同じである。

結論だけ述べれば、パラメータは以下で与えられる。

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mm} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{my} \end{bmatrix}$$

ただし

$$S_{ij} = \sum_{k=1}^{n} (x_i^{(k)} - \bar{x_i})(x_j^{(k)} - \bar{x_j})$$

$$S_{iy} = \sum_{k=1}^{n} (x_i^{(k)} - \bar{x_i})(y_j^{(k)} - \bar{y})$$

$\bar{x_i} = 1/n \sum_{k=1}^{n} x_i^{(k)}$、$\bar{y_i} = 1/n \sum_{k=1}^{n} y^{(k)}$。

4 実験

まず訓練データの作成であるが、ここでは SENSEVAL2 の日本語辞書タスク [9] で用いられた名詞 50 単語の訓練データを利用することにした。このデータは語義曖昧性解消のタスクに対する訓練データであり、本質的に各用例に対象単語の語義の id が付与されていると考えて良い。このデータから各用例対が同じ語義がどうかを自動で判定できる。同じ語義の場合は、線形モデルの経験的なパラメータ値から類似度を与え、異なる語義の場合は 0 を与えた。最終的に得られた線形モデルに対する訓練データ数は 846,045 個であった。

ここから最小自乗法により線形モデルのパラメータを推定した。得られたパラメータは以下であった。

```
a = 8.1987    b = 8.0044    c = 3.3696
d = 3.8949    e = 1.0512    f = 0.4125
```

次に以下の類似度の定義に従って用例のクラスタリングを行った。

類似度 (1) 単純な線型モデルによる用例間類似度 (各パラメータ値は 1)

類似度 (2) 経験的なパラメータ値を用いた用例間類似度 (各パラメータ値は前述)

類似度 (3) 本手法により得られたパラメータ値を用例間類似度 (各パラメータ値は上記)

クラスタリングにはクラスタリングツールの CLUTO[1] を利用した[2]。

[1] `http://glaros.dtc.umn.edu/gkhome/views/cluto`

[2] 起動するプログラムはデータ間の類似度からクラスタリングを行う scluster である。そこで使われているアルゴリズムはトップダウンにデータを 2 分割してゆく処理を、目的のクラスタ数が得られるまで再帰的に行う k-way clustering と呼ばれる手法である。

クラスタリングの対象も SENSEVAL2 の日本語辞書タスクで用いられた名詞 50 単語の訓練データとした。これはクラスタリングの正解が付与されている形になっているので、クラスタリングの評価をエントロピーで行うことができる [8]。エントロピーは値が小さいほどよいクラスタリングであることを意味する。結果を表 1 に示す。

単語	データ数	クラスタ数	類似度 (1)	類似度 (2)	類似度 (3)
間	266	9	0.368	0.384	0.383
頭	169	6	0.524	0.505	0.511
一般	267	5	0.635	0.652	0.649
一方	274	4	0.247	0.253	0.253
今	333	5	0.509	0.517	0.512
意味	173	3	0.964	0.920	0.930
疑い	201	2	0.080	0.080	0.080
男	213	4	0.198	0.196	0.192
開発	209	3	0.646	0.652	0.700
核	255	5	0.298	0.292	0.279
関係	414	3	0.650	0.656	0.652
気持ち	256	5	0.470	0.475	0.474
記録	236	3	0.564	0.547	0.547
技術	198	2	0.367	0.365	0.302
現在	341	2	0.362	0.404	0.404
交渉	242	2	0.143	0.145	0.145
国内	277	2	0.886	0.891	0.891
言葉	263	4	0.669	0.651	0.642
子供	354	2	0.987	0.973	0.973
午後	396	3	0.717	0.724	0.715
市場	254	4	0.485	0.483	0.449
市民	207	2	0.953	0.955	0.965
社会	340	6	0.239	0.241	0.232
少年	190	2	0.232	0.237	0.237
時間	283	4	0.537	0.584	0.567
事業	253	2	0.888	0.873	0.870
時代	360	4	0.480	0.483	0.485
自分	362	2	0.302	0.303	0.306
情報	285	3	0.523	0.507	0.504
姿	201	4	0.655	0.567	0.567
精神	157	2	0.677	0.711	0.692
対象	236	2	0.301	0.288	0.288
代表	466	3	0.354	0.343	0.329
近く	238	3	0.681	0.586	0.586
地方	271	2	0.945	0.950	0.947
中心	255	2	0.231	0.252	0.252
手	230	12	0.467	0.420	0.441
程度	202	2	0.137	0.137	0.137
電話	270	3	0.517	0.462	0.465
同日	234	2	0.799	0.796	0.798
花	175	3	0.114	0.111	0.113
反対	241	2	0.197	0.205	0.206
場合	292	2	0.751	0.789	0.789
前	426	4	0.264	0.262	0.267
民間	174	2	0.157	0.157	0.158
娘	203	3	0.362	0.359	0.357
胸	156	5	0.535	0.537	0.549
目	229	9	0.389	0.412	0.414
もの	757	14	0.548	0.522	0.516
問題	636	4	0.114	0.115	0.117
平均値	278.4	3.76	0.4824	0.4786	0.4767

表 1: 実験結果

表 1 より単純な類似度 (1) よりも経験的なパラメータ値を用いた類似度 (2) の方が良い結果が得られている。さらに経験的なパラメータ値を用いた類似度 (2) よりも本手法の類似度 (3) がさらに良い結果を出している。

5 考察

前述した実験において、平均のエントロピーは本手法が最良値を出したが、その差は非常に小さい。また最良値を出した単語数でみると類似度 (1) が 25 単語、類似度 (2) が 15 単語、類似度 (3) が 19 単語 であり[3]、類似度 (1) の最も単純なモデルが優れているという結果になる。これは類似度 (2) の経験的なパラメータ値が類似度 (1) よりも適切でなかったためである。類似度 (3) は類似度 (2) のパラメータを学習により改善したものと捉えることができる。そのため類似度 (3) は類似度 (2) よりも、平均のエントロピーと最良値を出した単語の数の両方において良い値となっている。つまり経験的なパラメータ値をうまく設定できれば、ここで示した学習手法により最良のパラメータを得ることができるであろう。

パラメータを推定するために、ここでは重回帰分析の手法を用いたが、判別分析の手法を用いて線形判別関数を求めることでも同様の推定が可能である。この場合、訓練データの観測値は同じ語義かどうかだけで済むので、より適切な推定が行えるようにも見える。しかし本タスクの場合、訓練データ（用例対）の大部分は異なる語義のクラスに属し、かつ原点となっている。このようなアンバランスな訓練データに対しては適切な推定を行うことが困難である。

用例間の類似度を測る場合、類似度のモデル以上に重要になるのは単語間の類似度である。用例中にある対象単語の周辺の単語は数も少なく、まったく一致しない場合も多い。その際に用例間の類似度を測る手がかりは単語間の類似度が最有力である。通常、単語間の類似度は既存のシソーラスを用いて測るが、既存のシソーラスがその規模や構造の点から本タスクに適しているかどうかは疑問である。今後は本タスクに適した単語間の類似度を大規模に作成していく必要があるだろう。

クラスタリング一般で考えてもデータ間の類似度の設定方法が本質的であり、近年はデータ間の類似度を学習により設定するという研究が盛んであるので（[4] など）、それらの研究成果も取り込みたい。

また用例を対象単語の語義に基づいてクラスタリングするタスクではクラスタの数（つまり語義の数）の推定も重要である [3]。SemEval-2007 の Task-02 ではここでのタスクと本質的に同等のタスクを扱っているが、中心の問題はクラスタ数の推定とその評価方法であった [1]。

最後に、ここでは対象単語が名詞であったが、動詞の場合、用例間の類似度の測定は用例による翻訳で生じる動詞の格フレームの選択問題と同型であることも指摘しておく。用例による翻訳では用例集から入力文と類似の文を検索するが、基本的には文の中心となる動詞を対象単語として、用例間の類似度を測っている。標準的には格の種類と格に入る名詞の類似度から算出される。つまり使われているモデルは、ここで示した線形モデルなので、本手法が応用できる。

[3] 合計が 50 単語にならないのは、同点のものを重複して数えているかである。

6 おわりに

用例をその単語の語義に基づいてクラスタリングするために、本論文では用例間の類似度を測る手法について述べた。用例間類似度を線型モデルで表し、次にパラメータを最小自乗法により推定する。訓練データの構築については経験的なパラメータ値を用いた仮のモデルを使うことを提案した。実験では、本手法の類似度の定義は、単純な定義や経験的な定義よりも、よいパフォーマンスを示すことができた。今後は、本タスクに適した単語間の類似度を大規模に構築し、クラスタリングの精度を更に改善したい。

参考文献

[1] Eneko Agirre and Aitor Soroa. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 7–12, 2007.

[2] Donald Hindle. Noun classification from predicate argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics (ACL-90)*, pp. 268–275, 1990.

[3] Hiroyuki Shinnou and Minoru Sasaki. Division of Example Sentences Based on the Meaning of a Target-Word Using Semi-supervised Clustering. In *Proceedings of LREC-2008*, 2008.

[4] Jason V. Davis and Inderjit S. Dhillon. Structured Metric Learning for High-Dimensional Problems. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 195–203, 2008.

[5] Masaki Murata and Masao Utiyama and Kiyotaka Uchimoto and Qing Ma and Hitoshi Isahara. Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In *Proceedings of the SENSEVAL-2*, pp. 135–138, 2001.

[6] Resnik Philip. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery. In *Proceedings of AAAI-92 Workshop on Statistically-Based NLP Techniques*, pp. 48–56, 1992.

[7] 国立国語研究所. 分類語彙表. 秀英出版, 1994.

[8] 新納浩幸. R で学ぶクラスタ解析. オーム社, 2007.

[9] 白井清昭. SENSEVAL-2 日本語辞書タスク. Vol. 10, No. 3, pp. 3–24, 2003.