EMアルゴリズムによる教師なし学習によって作成された単語意味クラスの評価

諏訪善彦　杉本浩和　鳥澤健太郎
北陸先端科学技術大学院大学情報科学研究科
{y-suwa, sugihiro, torisawa }@jaist.ac.jp

1 はじめに

本論文ではEMアルゴリズム [2] を用いた教師なし学習によりコーパスから作成された単語クラスを, 2名の被験者を用いて評価した. この単語クラスの生成手法は, Rooth ら [6] が提案し, 鳥澤 [3] が拡張したものである. この手法によって生成された単語クラスは, 単語の持つ意味を捉えたクラス分けになっているとされている. 実際, このクラスはすでにいくつかのNLPのタスク [3][4][5] で有用であることがわかっている. しかしNLPタスクでの有用性と, 単語クラスの意味の一貫性が同義であるとはいえない. そこで今回, 2名の被験者を用いてこの単語クラスが本当に意味的な一貫性を持っているか評価を行った. さらに, その評価の結果を用いて単語クラスの再学習を行った. その結果についても述べる.

2 単語クラスの評価

2.1 単語クラスの学習

まず単語クラスの作成に用いた学習アルゴリズムの簡単な解説をする. この手法は Rooth ら [6] が提案したもので, EMアルゴリズムという反復学習法を用いて教師なし学習で単語クラスを学習することができる.

学習では, 学習データの構文解析結果から得られる $\langle v, rel, w\rangle$ の三つ組みを用いる. ここで v は動詞, w は単語, そして rel は動詞 v に係る単語 w の助詞を意味する. この三つ組みは動詞 v と助詞のペア $\langle v, rel\rangle$ と, そのペアに関係する単語 w に分割することができる. したがって, この三つ組みの出現確率を

$$P(\langle v, rel, w\rangle) \stackrel{def}{=}_{a\in A} P(\langle v, rel\rangle|a)P(w|a)P(a)$$

と仮定し, EMアルゴリズムによって $P(\langle v, rel\rangle|a)$, $P(w|a)$ および $P(a)$ の確率の推定を行う. 直感的には, このクラスは $\langle v, rel\rangle$ や w の意味的なクラスである. このクラス a は学習データ中には表記されていない. また, A は k 個のシンボルからなる集合で, 各シンボルは最終的に生成される単語クラスの「名前」として機能する. 要素の数 k は学習前に人が決定する.

学習は $P_j(\cdot)$ から $P_{j+1}(\cdot)$ を反復して計算することで行う. $P_{j+1}(\cdot)$ を求めるには, まず下式より $P_j(a|\langle v, rel\rangle, w)$ を算出する.

$$P_j(a|\langle v, rel\rangle, w) = \frac{P_j(\langle v, rel\rangle|a)P_j(w|a)P_j(a)}{\sum_{a'\in A} P_j(\langle v, rel\rangle|a')P_j(w|a')P_j(a')}$$

そして, 求めた $P_j(a|\langle v, rel\rangle, w)$ を下の式に当てはめて, $P(\langle v, rel\rangle|a)$, $P(w|a)$ および $P(a)$ の確率を推定する.

$$P_{j+1}(a) = \frac{1}{|L|}\sum_{\langle v_i, rel_i, w_i\rangle\in L} P_j(a|\langle v_i, rel_i\rangle, w_i)$$

$$P_{j+1}(\langle v, rel\rangle|a) = \frac{\sum_{\langle v, rel, w_i\rangle\in L} P_j(a|\langle v, rel\rangle, w_i)}{\sum_{\langle v_i, rel_i, w_i\rangle\in L} P_j(a|\langle v_i, rel_i\rangle, w_i)}$$

$$P_{j+1}(w|a) = \frac{\sum_{\langle v_i, rel_i, w\rangle\in L} P_j(a|\langle v_i, rel_i\rangle, w)}{\sum_{\langle v_i, rel_i, w_i\rangle\in L} P_j(a|\langle v_i, rel_i\rangle, w_i)}$$

ここで L は, 学習データ中に現れる三つ組み $\langle v, rel, w\rangle$ のリスト (データの重複を許す) を意味する.

$$L = \langle\langle v_0, rel_0, w_0\rangle, \langle v_1, rel_1, w_1\rangle, \ldots, \langle v_l, rel_l, w_l\rangle\rangle$$

この処理を m 回繰り返して得られる $P_m(\cdot)$ を学習結果とする. これらの式は標準的なEMアルゴリズムの導出過程から導き出すことができる.

最終的に得られる $P(w|a)$ からベイズの定理を用いて $P(a|w)$ を求めることができる. これは単語 w がクラス a の用法で用いられる確率である. 各クラス a について $P(a|w)$ を計算して得られる確率の分布は, 単語 w の用法の多義性を捉えていると考えられる.

また, EMアルゴリズムでは一般に, 初期確率 $P_0(\cdot)$ が結果に大きな影響を及ぼす. にもかかわらず, この確率は多くの場合, 乱数によって生成される. しかし鳥澤 [3] は, この単語クラスの学習において初期確率を設定するヒューリスティックな方法を提案している. 鳥澤の提案している手法では, まず Brown ら [1] が提案した平均相互情報量によるクラスタリングの手法を三つ組みに適用し, 単語 w のクラス分けを行う. このクラス分けは単語の意味的なクラスを生成する. しかし, これは排他的なクラス分けの手法で, ひとつの単語が複数のクラスに属することを許さない. 鳥澤の提案した手法では, このクラス分けの結果にしたがってEMアルゴリズムの初期確率を与える.

平均相互情報量によるクラス分けの結果, k 個の単語クラス $c_1, c_2, \ldots, c_k$ が得られたとする (以下では $\{c_1, c_2, \ldots, c_k\} = A$ とし, ここでの単語クラスと EM アルゴリズムで使われる集合 A を同一視する). このとき, 単語 w と各クラスがどれだけ類似しているかをクラス分けのときと同様に, 平均相互情報量に基づいて計算し, 最も類似しているクラスを n 個選びだす (今回の実験では $n = 10$ とする). これらのクラスを $t_1, t_2, \ldots, t_n$ とすると, 単語 w の初期確率 $P_0(t_i|w)$ を $P_0(t_i|w) = P_i$ と設定する. ただし, P_i は比較的大きな確率であり, 定数で, $P_1 \leq P_2 \leq \cdots \leq P_n$ かつ,$\sum_{i=1}^{n} P_i = P_s$ となるように定められている. ここで P_s は 1.0 未満の確率の定数である (今回は P_s として 0.5 を用いた). また, n 個の類似クラスに選ばれなかったクラス t' には初期確率として $P_0(t'|w) = (1 - P_s)/(k - n)$ を与える.

この方法を用いることで, EM アルゴリズムによる単語クラスタリングは, 各単語 w に関して平均相互情報量によって得られた排他的な単語クラスで w に近いものに属する確率が比較的高い状態でクラスタリングをはじめることができる。このため, 初期確率決定にこの手法を用いることで, 学習結果の収束にかかる時間が短縮されることが確認されている.

2.2 評価基準

これまでに説明した手法に基づいて, 33 年分の新聞記事を入力データとして学習を行った. 55 回の反復学習によって得られた $P(a|n)$ の分布を次のような基準に従って評価を行い, クラス a が単語 n の意味的なクラスになっているのか調査した.

評価の対象は, 以下の条件をすべて満たす単語とした.

1. 各クラスの所属確率の上位 10 単語
2. 動詞, 副詞, 接続詞以外の単語クラスに属する単語
3. クラスへの所属確率 $P(a|w)$ が 0.1 以上の単語

この条件を満たす評価対象の単語を**対象単語**と呼ぶことにする. 対象単語には名詞, 未定義語, 形容詞, 判定詞, 指示詞, 感動詞, 特殊 (JUMAN の品詞体系より) が含まれる.

クラス内で大半の対象単語に共通する意味的な性質を見つけ, その性質をクラスのラベルとした. クラスのうち, 要素が明白に共通の意味を持ち, 文章中にそのクラスの要素が現れたとき置き換え可能な単語が存在するようなクラスを**置き換え可能クラス**と呼ぶ. 一方, 要素が共通の意味を持つが, 置き換え可能な単語が存在しないようなクラスを**置き換え不能クラス**と呼ぶことにした. 表 1 に置き換え可能クラス, 表 2 に置き換え不能クラスの例を挙げる.

また, クラスのラベルの表す性質を持たない対象単語に**削除マーク**を付け, その数をカウントした. 形態素の区切りミスのため単語として意味をなさない要素にも削除マークを付け, 同様に扱った. ここでクラス内の対象単語に占める削除マーク付き単語の割合が 30%以上のクラス (共通した意味を持つ単語が 70%未満のクラスは, 意味的な一貫性が見られないとして, 不適切なクラスであるとした. 同様に, 対象でない品詞を持つ単語 (動詞, 副詞, 接続詞) を 30%より多く, かつ 70%未満の割合で含むクラスも (品詞がばらついているため) 一貫した意味を持たないと考えて, 不適切なクラスとして扱った. 一方, 対象外の品詞の単語が 70%以上を占めるクラスは, その品詞が支配的なクラスとみなして, クラス自体を評価の対象外とした (クラス内の要素を対象単語として扱わない).

さらにクラスの持つ意味の範囲にも注意を払った. 例えば表 3 のクラスにラベルをつけるとき, すべての要素は「動物」という共通の意味を持つが, 「ワニ」に削除マークをつけることで, より狭義の「哺乳類」クラスとすることができる. このような場合にもクラス内の対象単語の 70%をボーダーラインとして, 狭義のラベルを採用した.

表 1: 置き換え可能クラス

CLASS-1612 { 兄弟 }

長男	0.81731	三女	0.74005
二男	0.80862	弟	0.66916
長女	0.79917	妹	0.56536
二女	0.78238	実弟	0.53156
三男	0.74767	実兄	0.42879

表 2: 置き換え不能クラス

CLASS-741 { 野球に関係すること }

打線	0.74682	DF:陣	0.65575
守り	0.74484	攻守	0.63300
守備:陣	0.71322	攻撃:陣	0.58447
守備	0.70983	投手:陣	0.55040
投打	0.68223	堅守	0.53812

表 3: 狭義のラベルをつける例

CLASS-368 { 哺乳類 }

キツネ	0.75483	サル	0.71466	
タヌキ	0.74308	ワニ	0.71250	×
イノシシ	0.74243	ゾウ	0.71163	
犬	0.73710	猫	0.71154	
クマ	0.72909	ネコ	0.69672	

表 4: 不適切なクラス

CLASS-1345 {×}

両方	0.73930	魔法	0.46834
一つ一つ	0.59416	国旗:国歌	0.41620
代	0.56580	+:両方	0.39088
*:ひとつひとつ	0.52894	善意	0.37288
学	0.51509	*:一粒	0.30670

2.3 結果

このような評価基準に沿って, 2 人の被験者 (被験者 A, 被験者 B) が, それぞれ独立に単語クラスの評価を行った. 単語クラスは 2500 個のクラスからなり, 今回はそのうち, 接続詞, 副詞, 動詞のクラスを除いた 2053 個のクラスについて評価した. この 2053 個のクラスを**対象クラス**と呼ぶ. 同一の字面を持つ単語が複数のクラスに所属するとき, それぞれを別の単語とみなした (重複を許した) 場合の対象単語の数は 17650 個であった. 2 人の被験者による評価の結果を表 5 に示す. こ

表 5: 被験者による評価の結果

項目	被験者 A		被験者 B	
	総数	比率	総数	比率
対象クラス	2053	—	2053	—
置き換え可能クラス	1138	55.4%	725	35.3%
置き換え不能クラス	367	17.9%	877	42.7%
適切なクラス	1505	73.3%	1602	78.0%
不適切なクラス	548	26.7%	451	22.0%
対象単語	17660	—	17660	—
削除マーク	1321	7.5%	1317	7.5%
不適切なクラスの要素	4349	24.6%	3513	19.9%
適切な要素	11990	67.9%	12830	72.7%

こで**適切なクラス**とは, 置き換え可能クラスもしくは, 置き換え不能クラスに分類されたクラスのことである. また, **適切な要素**とは, 適切なクラスの中の要素で削除マークのついていないものを指す.

この表から, どちらの被験者の評価においても, 70%以上のクラスが何らかの意味を有していると判定されていることがわかる.

また, 以下の表 6, および表 7 から被験者 2 人の評価のズレを見ることができる. これらの表によると, 置き換え可能クラスと置き換え不能クラスの判定の境界はあまり明白でないことがわかる. また, 置き換え可能クラスと置き換え不能クラスの和である, 適切なクラスに注目すると, 2 人のうちいずれかが適切なクラスであると判定している 1690 個のクラスのうち, 2 人ともが適切なクラスであると判定しているのは, 83.9%にあたる 1418 個のみであることが読み取れる. このことから, 意味の有無の判定は人手を用いた場合でも, かなりのばらつきがあることがわかる.

表 6: 両被験者が適切だと評価したクラスのみを適切なクラスとして扱った場合

項目	総数	比率
対象クラス	2053	—
置き換え可能クラス	597	29.1%
置き換え不能クラス	821	40.0%
適切なクラス	1418	69.1%
不適切なクラス	635	30.9%
対象単語	17660	—
削除マーク	607	3.4%
不適切なクラスの要素	5107	28.9%
適切な要素	11946	67.6%

表 7: 両被験者が不適切だと評価したクラスのみを不適切なクラスとして扱った場合

項目	総数	比率
対象クラス	2053	—
置き換え可能クラス	1267	61.7%
置き換え不能クラス	423	2.1%
適切なクラス	1690	82.3%
不適切なクラス	363	17.7%
対象単語	17660	—
削除マーク	2031	11.5%
不適切なクラスの要素	3270	18.5%
適切な要素	12359	70.0%

3 再学習

3.1 方法

前節での被験者 A の評価を用いて, 単語クラスの再学習を行った. 再学習は, EM アルゴリズムの初期確率データに評価の結果を反映したデータを用いることで実現した.

これまでの学習に用いた初期確率は, 前節で説明したように, 平均相互情報量を用いて作成したクラス $c_1, c_2, \ldots, c_k$ と, 各単語 w との類似度に沿って与えていた. 再学習用の初期確率は, この類似度が誤っていることが前節の被験者 A の評価によって指摘されている場合, その評価に従って, 初期確率の計算への類似度の適用を見直し, より意味的に一貫性のあるクラスが生成されるように配慮して生成した.

単語 w との類似度が上位 n 位までのクラス $t_j (1 \leq j \leq n)$ と単語 w の組み合わせは, 類似度が $n+1$ 位以下のクラスに比べて (定数 P_j により) かなり優先的な確率値から学習をスタートするため, 学習後も単語 w がクラス t_j に属する可能性が高い. いささか逆説的であるが, そのため評価を行った対象の中に, 初期確率で優先的な確率値を割り当てた単語とクラスのペアが含まれていて, その組み合わせについて陽に不適切だと評価されたものが多数存在した. ここで不適切だと評価されたというのは, クラス自体が不適切なクラスだと判断されたものや, クラス自体は意味を持っているが単語 w がそのクラスに属すことが適切でないと評価された (削除マークを付けられた) ものを指す. このように平

均相互情報量によって類似度が高いと判断されたクラスと単語のペアが, 被験者によって陽に否定されていた場合, そのペアに高い初期確率を与えないようにして再学習用の初期確率データを生成した.

具体的には, 平均相互情報量による単語 w との類似度が上位 $j(\leq n)$ 番目のクラス t_j について, 被験者 A が単語 w がクラス t_j(適切なクラス) に属することが望ましくないと評価していた場合, その組み合わせの持つ初期確率を, 類似度が $n+1$ 位以下のクラスと同じように $P_0(t_j|w) = (1-P_s)/(k-n)$ とした. ここで P_s は, 前節でも説明したとおり確率の定数である. このようにすることで, $P_0(t_j|w)$ の値が従来の P_j より大幅に低下するため, 学習後に単語 w がクラス t_j に属する可能性が低くなる. 一方, クラス自体が不適切であると評価されたクラスについては, クラスそのものを削除した. また, 単純に $P_0(t_j|w)$ の値を変更してしまうと, $\sum_{i=1}^{k} P_0(t_i|w) < 1$ となってしまうので, このようにして算出した確率値に正規化を施して初期確率とした. この初期確率を用いて再度 EM アルゴリズムによる学習を行った.

3.2 結果

被験者 A の評価を元に再学習を行って得られた単語クラスの評価を行った. 被験者再学習後のデータからランダムに 200 個のクラスをピックアップし, 評価を行った結果を表 8 に示す. また, ランダムに 200 個の単語を取り出してチェックした結果を表 9 に示す. これ

表 8: 再学習後のクラスの評価

項目	総数	比率
対象クラス	200	—
適切なクラス	164	82.0%
不適切なクラス	36	18.0%

表 9: 再学習後の単語の評価

項目	総数	比率
対象単語	200	—
削除マーク	14	7.0%
不適切なクラスの要素	37	18.5%
適切な要素	149	74.5%

らの表から, 再学習によって不適切なクラスの割合が減少し, 意味的に適切なクラスの比率 (表 5 の被験者 A の適切なクラスの欄を参照) が 73.3%から 82.0%に向上していることがわかる. これによって, 再学習が有効であることが示された.

4 まとめと考察

本論文では, EM アルゴリズムによる教師なし学習により作成された単語意味クラスが, 本当に意味的な一貫性を持っているか, 被験者を用いてチェックを行った. その結果, 70%以上のクラスが, 意味的に適切なクラス分けになっていることが明らかとなった. さらに, その評価を EM アルゴリズムの初期確率に反映して, 再度学習を行ったところ, 適切なクラスの割合が当初のデータより 9%程度上昇し, EM アルゴリズムによる単語意味クラスの作成において, 再学習が有用であることが確認された.

しかし今回の評価実験では, 単語の重複を考慮しないという単純化を行ったため, 単語クラスが単語の多義性をうまく捉えているかどうかの評価を行うことができなかった. 多義性の評価は今後の課題である.

参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer, Class-based n-gram models of natural language, Computational Linguistics, 18(4):31–40, 1992.

[2] A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B, Vol.39, pp.1–38, 1977.

[3] Kentaro Torisawa, An unsuperveised method for canonicalization of Japanese postpositions, In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001), pp. 211–218, 2001.

[4] Kentaro Torisawa, A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases, In Proceedings of the Workshop on Automatic Paraphrasing, pp. 63-72, 2001.

[5] Kentaro Torisawa, An unsupervised learning method for associative relationship between verb phrases, In Proceedings of COLING 2002, 2002.

[6] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil, Inducing a semantically annotated lexicon via em-based clustering, In Proceedings of 37th Annual Meeting of the ACL, 1999.