EM法による単語意味分類のための効果的セッティング

諏訪善彦　　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, torisawa}@jaist.ac.jp

1　はじめに

本論文ではEMアルゴリズム [2] による係り受け情報を用いた単語クラスタリング手法 [3] の拡張についてのべる．より具体的には, EMアルゴリズムに与える初期確率を, 通常行われるようにランダムに決めるのではなく, 平均相互情報量によるハードクラスタリング [1] を用いて決める手法を導入した．性能の評価には, 単語クラスタリングの性能評価に良く使われる pseudo-word disambiguation[3][5] と人手による評価の二つを用いた．その結果, ハードクラスタリングによる初期確率の設定を行った場合とランダムな初期確率を与えた場合では, pseudo-word disambiguation ではほとんど差が見られないものの, すくなくとも繰り返し計算の回数をある程度限定した場合, 人手による評価ではハードクラスタリングによって初期確率の設定を行った場合の方がより意味的に一貫性のあるクラスが得られていることがわかった．

本研究を行うに至った動機は次の通りである．EMアルゴリズムによる単語クラスタリングは, 一つの単語が複数のクラスに属することを許すという点で, 語の持つ意味の曖昧性を捕らえることができる利点があると言われてきた．しかしながら, 別の観点からは, この枠組みの様に一単語が際限なく多数のクラスに (ある確率で) 属せるという仮定も極端であると考えられる．また, 大多数の単語は中心的, あるいは支配的な (一つあるいは少数の) 意味をもつという直感もある．そのような支配的な意味はハードクラスタリングのような, 一単語が一クラスにしか属せないという強い制約を課せられた枠組みでより良くとらえられるかも知れない．

以上は, 単なる推測であり, 根拠のある話ではないが, EMアルゴリズムに与える初期確率をハードクラスタリングの結果を用いて決定することにより, 単語の支配的／中心的意味をより反映した直感的に正しい意味クラスがえられるのではないかと考えたのが本研究を行った動機である．研究の結果も, 現時点では残念ながら以上の推測を裏付けるほど強いものではないが, 有望な結果であると考えている．

2　EM法による単語クラスタリング

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 [3] 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と助詞のペア $\langle v, rel\rangle$ と, 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

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

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．(確率の推定には, 以上の確率モデルから導出される漸化式が使われるが, その漸化式については, [3] を参照のこと．) 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．

3　拡張：初期確率のハードクラスタリングによる設定

EMアルゴリズムによる確率の推定は漸化式による．すなわち, $P(\cdot)$ という確率を推定するのには, 初期確率 $P_0(\cdot)$ から始めて, 確率モデルから導出された漸化式により $P_1(\cdot), P_2(\cdot), \cdots, P_i(\cdot)$ と順次計算していき, ある終了条件を満たすような $P_m(\cdot)$ を $P(\cdot)$ の推定値として出力する．ここで問題は, この繰り返し計算が局所最適解を与えることしか保証されていないことで, 初期確率すなわち $P_0(\cdot)$ の与え方によって, 全く異なる推定値が得られる．

通常, この初期確率はランダムに与えられることが多いが, 我々は平均相互情報量によるハードクラスタリングにより初期確率を決定する手法を導入した．なお, このアルゴリズムに関しては, すでに [4] で言及されているが性能評価は行われてこなかった．動詞と助詞のペ

ア $\langle v, rel\rangle$ と単語クラス c の間の平均相互情報量を次のように与える.

$$I = \sum_{c,\langle v,rel\rangle} P(\langle v, rel, c\rangle) log \frac{P(\langle v, rel, c\rangle)}{P(c)P(\langle v, rel\rangle)}$$

ただし, ここで c は EM アルゴリズムで使われる単語クラスとは異なり, 一つの単語は常に一つのクラスにしか属することができないと仮定する[1]. (つまり, EM アルゴリズムでのクラス a と単語 w に関しては, 単語 w がクラス a に属する確率 $P(a|w)$ は 0 から 1 までの任意の確率値をとれるのに対して, 平均相互情報量でのクラスタリングに用いるクラス c に関しては $P(c|w)$ は 0 か 1 のいずれかの値しかとれない.)

平均相互情報量によるハードクラスタリングの手続きは以上に与えられた平均相互情報量 I を最大化するような単語クラスの集合を計算することを目的とするが, これを厳密に求めるのは困難で, 通常は以下にあるようなボトムアップなグリーディーアルゴリズムで近似を行う.

ボトムアップマージアルゴリズム

ステップ 1 分類の対象である単語集合 $W = \{w_1, w_2, \cdots, w_n\}$ の各々の要素を一クラスとする. すなわち, 単語クラスの集合 $C = \{c_1, c_2, \cdots, c_n\}$ とし, $c_i = \{w_i\}$ $(1 \leq i \leq n)$ とする.

ステップ 2 単語クラスの数が前もって与えられた数になるまで以下を繰り返す.

ステップ 3 すべての可能な単語クラスのペアに対して, そのペアのクラスをマージした後の平均相互情報量を計算し, それがもっとも大きいペアを選びだす. そのペアを一つのクラスにマージし, ステップ 2 に戻る.

実験では, 以上のアルゴリズムでもある程度意味的な一貫性のある単語クラスを生成できることがわかっている. しかしながら, 上述のクラスタリングの過程で, 時々, 各単語を現在属しているクラスとは別のクラスに移動させたときの平均相互情報量を計算し, 移動後の方が平均相互情報量が大きければ, 実際にその移動を行うという手続きを実行することで, より意味的に妥当と思われる単語クラスを生成することができた. (この手続きの詳細については本稿では割愛する.)

ハードクラスタリングの結果から, EM アルゴリズムのための初期確率を得るのは, 以下の手続きによった.

[1]オリジナルのアルゴリズム [1] では bigram を対象としているのにたいして, 本稿では, 係り受けを対象とし, なおかつ, いわゆる「係り元」だけをクラスタリングの対象としているため, 式が異っている.

初期確率決定アルゴリズム

入力 平均相互情報量によるハードクラスタリングで求められた k 個の単語クラスからなる集合 $C = \{c_1, c_2, \cdots, c_k\}$

ステップ 1 分類対象の単語 w すべてに対して以下のステップ 2, 3 を繰り返す.

ステップ 2 C の各々の要素にたいして, w をマージしたときの平均相互情報量を計算し, その値がもっとも大きいもの 10 クラスを選びだし, それらのクラスを平均相互情報量の大きい順に $c'_1, c'_2, c'_3, \cdots, c'_{10}$ とする.

ステップ 3 $P(c'_1|w) = 40/190$, $P(c'_2|w) = 9/190$, $P(c'_3|w) = 8/190$, $P(c'_4|w) = 7/190$, $P(c'_5|w) = 6/190$, $P(c'_6|w) = 6/190$, $P(c'_7|w) = 5/190$, $P(c'_8|w) = 5/190$, $P(c'_9|w) = 5/190$,$P(c'_{10}|w) = 5/190$ となる様に設定し, 残りの確率を $C - \{c'_1, c'_2, c'_3, \cdots, c'_{10}\}$ に等確率で分配する. (これらの確率値は適当にきめたものであり, 平均相互情報量という観点からより「近い」クラスにより大きな確率を与えていること以外に意図はない.)

以上の記述では, 簡単のため, ハードクラスタリングで得られた単語クラス ($c_1, \cdots, c_k$ や, $c'_1, c'_2, c'_3, \cdots, c'_{10}$) を EM アルゴリズムによるクラスタリングで求めるクラスの「名前」として使っていることに注意されたい. ただし, これは, EM アルゴリズムで得られる単語クラスとハードクラスタリングの結果が一致することを意味しておらず, 実験でもかなり異なるクラスが得られた.

4 実験

32 年分の新聞記事 (毎日新聞 92-99, 日経新聞 90-98, 読売新聞 87-01) を既存のツールで構文解析し, これより 12,796 個の単語, 154,356 個の助詞, 単語のペアを含む 138,089,690 個の共起データを抽出した. これらの共起データを 5 等分し, それらのうちの 4 つをマージすることによって, 5 通りの学習データを作成し, 次の二つの設定で, EM アルゴリズムを適用した. ともにクラス数は 1000 とした.

設定 A 初期確率を平均相互情報量に基づいたグラスタリングの結果を使って決定.

設定 R 初期確率を乱数によって決定.

得られた単語クラスの評価は以下にあるように二通りの方法で行った. 一つは, 単語分類の研究で良く使われる psuedo-word disambiguation[3] [5] とよばれるもので, もう一つは被験者による主観的評価である.

4.1 Pseudo-word Disambiguation

pseudo-word disambiguation[3][5] とは, 学習データに含まれないが, 他のデータ (テストデータ) に含まれる共起関係と, その共起関係にふくまれる単語と同じ単語を含むが, 学習データにもテストデータにも含まれない共起データの二つを用意し, 学習の結果得られた単語分類が, テストデータに含まれる共起関係を, 学習データ, テストデータのいずれにも含まれない共起関係よりも, より高い確率で生じると判定できるかどうかをチェックするテストである.

前述したように, 今回の実験では, コーパスから抽出された共起データを 5 等分し, そのうち 4 つをマージすることで, 学習データを作成した. pseudo-word disambiguation で利用するテストデータとしては, 5 つのうちの残り一つを利用した. また, テストで用いられる共起データとしては, テストデータに含まれ, 学習データに含まれない正解データ $\langle v, rel, w\rangle$ をまず用意し, ついで, 単語と助詞のペアで $\langle v, rel\rangle$ と最も近い頻度をもつものを $\langle v', rel'\rangle$ とする. このペアと正解データに含まれる単語 w との共起データ $\langle v', rel', w\rangle$ が正解データにも学習データにも現れていなければ, これらの共起データ $\langle v, rel, w\rangle$ と $\langle v', rel', w\rangle$ をテストで用いる.

テストは, 以上のような手法で共起データを 10000 通りあつめ, EM アルゴリズムで求められた確率によって, $P(\langle rel, v, w\rangle) > P(\langle rel', v', w\rangle)$ となる場合を「正解」としてカウントすることによって行われた. テストの結果は, 図 1 に与えられている. 結論として, EM の繰り返し計算の回数が小さいところでは, ハードクラスタリングによって初期値を設定した方が, ランダムに初期値を設定した場合に比べて良い数値がでているが, 繰り返しが多くなるにつれて両者はほとんど同等の性能がでている.

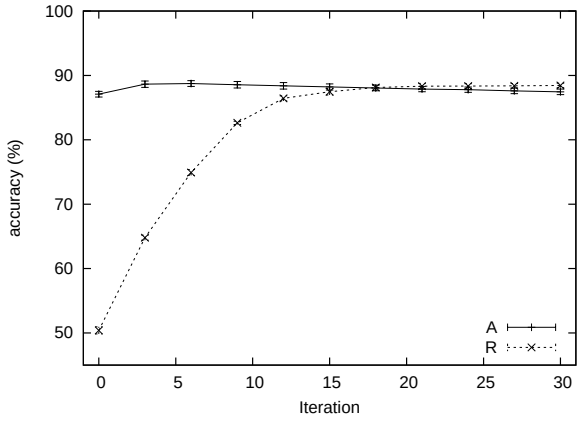

図 1: 学習設定の違いによる pseudo-word disambiguation の精度の違い (各頂点は 5 通りの学習データの平均)

4.2 人手による評価

被験者に異なる設定から生成された 2 つのクラスを提示し, クラスの優劣を判定してもらった. この 2 つのクラスは, 共通する単語が 1 つ以上高い確率で属するものを選んだ. その上で, 次に示す評価基準に従って, どちらのクラスがより優れているか判定してもらった. 具体的な評価手順は以下のように定めた.

1. 2 つのクラスについて, それぞれ意味的な一貫性があるかどうか確認する.
   - クラスからいくつかの単語を取り除くと意味的な一貫性を見つけることができる場合には, 単語の除去を行う. ただしこのとき両クラスに含まれる共通単語は除去せず, 共通単語を許容する意味を見つけるように努める.
   - クラス内に形態素解析などのエラーに起因する, 意味を成さない単語がある場合には除去する. この場合に限り共通単語の除去を認める. ただし, その場合の判定は, 両クラスとも意味的な一貫性を持たない, とする.
   - 意味的な一貫性が見つかる場合でも, 更にいくらかの単語を除去すると一貫性がより強くなる場合には, クラス内の除去単語の割合が 30 %を超えないように注意して単語の除去を行う. このときも共通単語は除去してはいけない.
2. 意味的な一貫性を得るために 30 %以上の単語を除去してしまった場合には, そのクラスはやはり意味を持たないものと判断する.
   - 両クラスとも意味を持たないのであれば, 判定は, 両クラスとも意味的な一貫性を持たない, とする.
   - 意味を持たないのが片側のクラスだけならば, 判定は, 他方のクラスの方が優れた一貫性を持っている, とする.
3. 2 つのクラスの両方に意味的な一貫性がある場合には, 両クラスを比較して, より意味的な一貫性の強いクラスを優れているとする.
4. 両クラスの持つ意味に差がないか, もしくは意味的な一貫性の強さの差が明確でない場合には, どちらのクラスも同程度に優れている, とする.

本来, EM アルゴリズムによって生成される単語クラスは, 全ての単語が全てのクラスに確率的に所属しているが, 今回は, 人手による評価であるため, クラスへの所属確率 (すなわち、単語 w がクラス c に属する確率で、$P(c|w)$ で表される.) が上位 10 位以内でかつ, 所属確率が 0.1 以上の単語のみをクラスの要素と見なし, 簡素化を図った. さらにそれらのクラス内の単語につ

いては確率値の大小を考慮せず, 一様に所属しているものと見なした. また今回の実験において, 意味的な一貫性を持つ単語クラスとは, クラス中の各単語が何らかの共通の上位概念の下位語であること, と定義した. ここで上位概念とは, 通常の「上位語」を単一の単語に限らないように制限を緩めたものを指し, 通常の上位語に加えて,「政令指定都市の県庁所在地」など単語を修飾する言葉を補うことを許す. ただし, 複数の無関係な集合を指す「国と食べ物の名前」のような言葉は上位概念ではない, とした. また, クラスの持つ意味的な一貫性の強さについては, クラス中の単語の共通上位概念が下位語として取りうる単語の数が少ないほど強い, と定義した. 下位語の数は正確に調べることができない場合が多いが, 比較する上位概念が包含関係にある場合や, 明確な差がある場合を想定してこのような基準とした.

このような基準に従って, 5 人の被験者にクラス間の比較評価を行ってもらった. 比較は設定 A と設定 R の EM の繰り返し計算が 30 回のときのクラスの間で行なった. それぞれの比較の組み合わせのデータは, 該当する設定の 5 通りの学習データごとに共通単語を持つクラスのペアを生成し, そのペアのリストの中からランダムに 25 ペアを抽出して使用した. なお, 評価の際, 被験者には提示されているクラスがどのようにして生成されたものか分らないようにした. 結果は表 1 にある. この表によれば, pseudo-word disambiguation ではほとんど差がなかったにもかかわらず, 被験者による判定では設定 A の方が圧倒的に優れているとされていることがわかる. (被検者間での一致の度合を計る kappa 統計量は 0.53.) また, 本稿では省略するが, 同様の方法で, 設定 R で繰り返しが 5 回の場合と 30 回を被験者に比較してもらった場合, 30 回の方が優れているという判定結果がでており, また, 設定 A では繰り返しが 5 回と 30 回では有意な差が見られなかった.

以上の実験結果が示唆するのは, まず, pseudo-word disambiguation で同等の性能が得られる場合でも, 被験者による評価では大きな差が生じ得るということであり, ついで, 本来はより精密な評価をまたなければならないが, 繰り返し計算が 30 回程度より小さいところでは, ハードクラスタリングによる初期確率決定が, 単語分類の意味的一貫性という観点からはおそらく有効であるということである. (設定 A では繰り返し回数が 5 回と 30 回で評価結果に大きな差が見られないことから, 5 回から 30 回程度の繰り返し回数であれば, ほぼ同等の意味的一貫性を持った単語クラスが生成されているものと推測した. また, 設定 R では繰り返し 5 回にくらべて, 繰り返し 30 回がより良い評価結果であることから, 5 回から 30 回までの間では, 少なくとも被験者による評価結果では, 設定 A の方がより強い意味的一貫性をもつであろうと推測した. もちろん, これらの推測は今後さらなる実験によって検証される必要がある.)

表 1: 設定 A(繰り返し 30 回) と設定 R(繰り返し 30 回) の比較結果 (3 名以上が同意した項目の数 / 4 名以上が同意した項目の数)

学習データ	両クラスとも意味的な一貫性を持たない	設定 A のクラスの方が優れている	設定 R のクラスの方が優れている
0	5 / 3	10 / 5	2 / 1
1	5 / 4	15 / 11	0 / 0
2	5 / 3	11 / 8	2 / 1
3	5 / 3	10 / 9	3 / 1
4	4 / 3	16 / 9	0 / 0
合計	24 / 16	62 / 42	7 / 3

5 まとめ

本稿では, 平均相互情報量を用いたハードクラスタリングによって, EM アルゴリズムによる単語クラスタリングで用いる初期確率を設定する方法についてのべ, また, 実験によりその効果を部分的に検証した. 実験は psuedo-word disambiguation と被験者による評価の二通りによっておこない. EM アルゴリズムの繰り返し計算の回数が比較的少ない (30 回程度) ところでは, ハードクラスタリングによって初期確率を決定する方法が有効であることを示唆する結果を得た. 今後はより精密な検討, 検証を加えていく予定である.

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. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[3] 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*, pages 104–111, 1999.

[4] Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 2001.

[5] Bill Gale, Kenneth Church, and David Yarowsky. Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60, Cambridge, MA, 1992.