未定義語義の判別を含む語義曖昧性解消

菊田 篤史　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{a-kikuta, kshirai}@jaist.ac.jp

1 はじめに

テキスト中の単語の意味を判定する語義曖昧性解消 (Word Sense Disambiguation; WSD) は，機械翻訳を始めとする様々な自然言語処理に必要となる基盤技術である [1, 6]．従来の語義曖昧性解消の問題設定としては，単語の語義をあらかじめ定義し，それらの語義の中からテキスト中の単語の意味を選択する場合が多い．しかし，テキストに現われる単語が常に定義済みの語義を持つわけではない．例えば，EDR 概念辞書 [4] には「電話」という単語の語義として (電話機)(電話機を使った通話) の 2 つの意味が定義されている．ところが，以下の文中の「電話」は，EDR 概念辞書で定義された 2 つの意味のいずれでもなく，(電話番号) という意味を持つ．

> A4 判 2 枚ほどの用紙に自分だけの電話帳を作成することができる．

実際に新聞記事を調べてみると，(電話番号) という意味で使われる「電話」という単語は数多く存在する．以下，本論文では，このようなあらかじめ定義されていない語義を未定義語義と呼ぶ．

従来の語義曖昧性解消手法の多くは，あらかじめ定義された語義のいずれかを選択するアプローチを取るため，未定義語義を持つ単語に対しては必ず誤った語義を選択する．これは当然望ましいことではない．そこで，本研究では，テキスト中の単語に対し，その単語の語義が辞書などで定義された語義のいずれかであるか，あるいはそれ以外の未定義語義であるのかを判別する語義曖昧性解消手法を提案する [2]．定義済みの n 個の語義に「未定義語義」というクラスを加え，$n+1$ 個の語義のいずれかを選択する Naive Bayes モデルを EM アルゴリズムを用いて教師なし学習する．また，より良いモデルを学習するために，語義タグ付きコーパスを利用した初期パラメタの設定を行う．

未定義語義を取り扱う関連研究としては Schütze による研究がある [5]．Schütze は，コーパスに出現する個々の単語毎に文脈ベクトルを作成し，クラスタリングによって同じ文脈を持つと考えらえる単語のクラスタを獲得する手法を提案している．このとき，得られた単語クラスタが 1 つの語義を表わす．事前に語義を定義しないため，既存の辞書等に記載されていない未定義語義も検出することが可能である．しかし，単語がいくつ語義 (クラスタ) を持つのかや，自動獲得された語義と辞書などで定義されている語義とをどのように関連付けるかなどについて十分な議論がなされていない．これに対し，本研究では，定義済みの語義または未定義語義のいずれかを選択するモデルを学習するため，システムの出力と定義済み語義の関係は明白である．なお，本研究では未定義語義の数を 1 つと仮定しているが，一般には未定義語義に該当する語義は複数存在すると考えられる．ここでは未定義語義を判別する最初のステップとして，未定義語義は 1 つだけであると仮定して問題を簡単化した．未定義語義の数の推測については今後の課題とする．

2 モデル

本研究では，ある単語 w に対して，以下の中から該当する語義を選択するモデルを学習する．

$$s_1,\ \cdots,\ s_n,\ u\,(=s_{n+1})$$

ここで，s_1〜s_n は辞書などで定義されている語義を表わす．n は定義済みの語義の数を表わす．一方，u は s_k のいずれにも該当しない未定義語義を表わしている．便宜上これを s_{n+1} と表記する．

本研究では，WSD のためのモデルとして Naive Bayes モデル [3] を用いる．すなわち，式 (1) の確率モデルを学習する．

$$P(s_k)\prod_{f_i\in c_j}P(f_i|s_k) \tag{1}$$

c_j は対象語 (曖昧性を解消したい単語) を含む文を，f_i は c_j に含まれる素性を表わす．素性とは文中から得られる WSD にとって有効な情報で，対象語の前後に現われる単語などが相当する．素性の詳細については後述する．$P(s_k)$ は語義 s_k の出現確率，$P(f_i|s_k)$ はある文が語義 s_k を持つ単語を含むとき，その文中に素性 f_i が生起する確率である．語義の曖昧性を解消する際には，未定義語義を含む全ての語義 s_k について式 (1) を推定し，それが最も大きい語義を正しい語義として選択する．

本研究で用いた素性は以下の通りである．いずれも語義曖昧性解消でよく用いられる素性である．

- 対象語の直前/直後の単語の基本形
- 対象語の直前/直後の単語の品詞
- 対象語の 2 つ前/2 つ後の単語の基本形
- 対象語の 2 つ前/2 つ後の単語の品詞
- 対象語から前後 10 単語以内にある自立語の基本形

3 モデルの学習

本節では，式 (1) 中のパラメタ $P(s_k)$, $P(f_i|s_k)$ を学習する手法について述べる．現在の WSD に関する研究は Naive Bayes モデルも含めて教師あり学習による手法が主流である．しかし，教師あり学習には正しい語義が付与された語義タグ付きコーパスが必要である．本研究では未定義語義の判別を目的としているため，そのような正解付きデータの存在は期待できない．そこで，語義タグ付きコーパスを必要としない教師なし学習の手法を用いる．具体的には EM アルゴリズムによってパラメタを学習する．

3.1 EM アルゴリズム

本項では EM アルゴリズムによるパラメタ学習について述べる．なお，このアルゴリズムは基本的に Manning ら [3] によって紹介されたアルゴリズムと同じである．

1. 初期パラメタの設定
   $P(s_k)$, $P(f_i|s_k)$ の初期値を定める．初期値の設定方法については 3.2.1 または 3.2.2 で述べる．
2. E-step
   与えられたパラメタから確率 $P(s_k|c_j)$ を式 (2) で求める．c_j は訓練コーパスに出現する文を表わす．

$$P(s_k|c_j) = \frac{P(s_k)P(c_j|s_k)}{P(c_j)} \quad (2)$$

$$= \frac{P(s_k)P(c_j|s_k)}{\sum_{k=1}^{n+1} P(s_k)P(c_j|s_k)} \quad (3)$$

   ここで $P(c_j|s_k)$ は以下のように近似する．

$$P(c_j|s_k) = \prod_{f_i \in c_j} P(f_i|s_k) \quad (4)$$

3. M-step
   E-step で求めた $P(s_k|c_j)$ を用いてパラメタの値を式 (5)(6) で再推定する．

$$P(f_i|s_k) = \frac{\sum_{c_j : f_i \in c_j} P(s_k|c_j)}{\sum_{f_i} \sum_{c_j : f_i \in c_j} P(s_k|c_j)} \quad (5)$$

$$P(s_k) = \frac{\sum_{c_j} P(s_k|c_j)}{\sum_{k=1}^{n+1} \sum_{c_j} P(s_k|c_j)} \quad (6)$$

4. 収束判定
   E-step と M-step を交互に繰り返し，パラメタを反復推定する．確率モデル全体の尤度を式 (7) と定義し，この値が収束する (変化量が T_l 以下になる) か反復回数が T_t を越えたら学習を終了する．式 (7) における J は学習コーパスにおける文の総数である．

$$\log \prod_{j=1}^{J} \sum_{k=1}^{n+1} P(s_k)P(c_j|s_k) \quad (7)$$

   本研究では $T_l = 1$，$T_t = 10$ とした．

3.2 初期パラメタの設定

EM アルゴリズムは局所最大解しか得られないため，学習の結果は初期パラメタの与え方に大きく依存する．本研究では，定義済みの語義 $s_k(1 \leq k \leq n)$ については語義タグ付きコーパスがあると仮定する．このコーパスには正しい語義として $s_k(k \neq n+1)$ が付与されているが，未定義語義 s_{n+1} は付与されていないものとする．そして，この語義タグ付きコーパスから得られる統計情報をもとに初期パラメタを設定することにより，学習後のモデルの品質向上を図った．

3.2.1 語義タグ付きコーパスから最尤推定する方法

まず，語義の生起確率 $P(s_k)$ の初期パラメタを式 (8), (9) のように設定する．

$$P^0(s_k) = (1 - I_u)\frac{O(s_k)}{\sum_{k=1}^{n} O(s_k)} \quad (1 \leq k \leq n) \quad (8)$$

$$P^0(s_{n+1}) = I_u \quad (9)$$

I_u は未定義語義 s_{n+1} に与えるべき初期パラメタで，定数とする．$P(s_{n+1})$ は未定義語義の出現確率であり，何らかの方法でこれを推測して初期パラメタとしたいところだが，これは困難であるため，低い確率を定数として与えることにした．I_u の適切な設定方法については今後検討したい．本研究では $I_u = 0.1$ としている．一方，定義済みの語義 s_k については，語義タグ付きコーパスから最尤推定している．式 (8) 中の $O(s_k)$ はコーパスにおける s_k の出現回数である．なお，式 (8) で $(1 - I_u)$ という項をかけているのは $\sum_{k=1}^{n+1} P(s_k) = 1$ という制約を満たすためである．

次に，$P(f_i|s_k)$ の初期パラメタを式 (10),(11) のように設定する．

$$P^0(f_i|s_k) = \frac{O(f_i, s_k) + \alpha}{O(s_k) + \alpha|F|} \quad (1 \leq k \leq n) \quad (10)$$

$$P^0(f_i|s_{n+1}) = \frac{1}{|F|} \quad for\ all\ i \quad (11)$$

式 (10) において，$O(f_i, s_k)$ は s_k を含む文中に素性 f_i が出現する回数を，F は素性の集合を表わす．基本的には，定義済みの語義 s_k については，$P^0(f_i|s_k)$ は語義タグ付きコーパスから最尤推定する．ただし，平滑化のため，全ての $O(f_i, s_k)$ に α という小さい値を足している．ここでは $\alpha = 1$ とする．一方，未定義語義 s_{n+1} については，$P^0(f_i|s_{n+1})$ は一様分布であるとする．

上記の方法で初期パラメタを設定し，EM アルゴリズムで Naive Bayes モデルを学習する予備実験を行った．学習したモデルを用いて語義曖昧性解消を行ったところ，ほとんどの場合が最頻出語義が選択され，未定義語義が選択されることはなかった．すなわち，常に最頻出の語義を選択するベースラインモデルとほぼ同じ結果になった．学習後のパラメタを調べてみると，最頻出語義に対する $P(s_k)$ の値が非常に大きく，0.99 を越える場合も多かった．

ここでの初期パラメタの与え方では，語義タグ付きコーパスにおける最頻出語義を s' とすると，$P(s')$ の値は大きく，また s' と共起する素性も多いことから $P(f_i|s')$ の初期値も大きく推定される．EM アルゴリズムによる反復推定によってその傾向がより強く学習され，$P(s')$ や $P(f_i|s')$ の値がどんどん大きくなるように学習が進んだとみられる．すなわち，初期パラメタを語義タグ付きコーパスから最尤推定することによって一種の過学習が生じたと考えられる．結果として，式 (8)〜(11) による初期パラメタの設定はうまくいかなかった．

3.2.2 共起性の強い素性に高い初期値を与える方法

語義タグ付きコーパスを用いた最尤推定による初期パラメタの設定は，初期パラメタの影響が強く，教師なし学習の過程で正解なしの大量のテキストにおける語義の出現傾向を自然に学習することができなかった．そこで，語義タグ付きコーパスから得られる情報をもっと緩やかな制約として初期パラメタに反映させることを考えた．

まず，$P(s_k)$ の初期パラメタについては 3.2.1 と同様に式 (8),(9) で求める．一方，$P(f_i|s_k)$ の初期パラメタは以下のように求める．まず，定義済みの語義 s_k については，$P(f_i|s_k)$ の大きい上位 n 個の素性集合 F_{s_k} と，それ以外の素性集合 $\overline{F_{s_k}}$ に分ける．ここで $P(f_i|s_k)$ は式 (10) のように語義タグ付きコーパスから最尤推定で求める．そして，F_{s_k} 中の素性に対する初期パラメタ $P^0(f_i|s_k)$ の値を $\overline{F_{s_k}}$ 中の素性に対する初期パラメタの r 倍とする．また，F_{s_k} 中，$\overline{F_{s_k}}$ 中の各素性については，$P^0(f_i|s_k)$ の値は全て等しく設定する．$\sum_{f_i} P(f_i|s_k) = 1$ という制約から，具体的な設定式は (12) のようになる．

$$\begin{cases} P^0(f_i|s_k) = \dfrac{r}{|F| + (r-1)n} & f_i \in F_{s_k} \\ P^0(f_i|s_k) = \dfrac{1}{|F| + (r-1)n} & f_i \in \overline{F_{s_k}} \end{cases} \tag{12}$$

一方，未定義語義 s_{n+1} については，$P(f_i|s_k)$ の大きい上位 n' 個の素性集合 F' と，それ以外の素性集合 $\overline{F'}$ に分ける．先ほどと異なるのは，s_k 毎に上位の $P(f_i|s_k)$ を選択するのではなく，全ての f_i,s_k の組について $P(f_i|s_k)$ を比較し，その上位 n' 個の素性を異なりで選別するということである．そして，F' 中の素性に対する初期パラメタ $P^0(f_i|s_k)$ の値を $\overline{F'}$ 中の素性に対する初期パラメタの $1/r'$ 倍とする．F' 中の素性に対する初期パラメタを低く設定したのは，F' は定義済みの語義とよく共起する素性の集合であり，未定義語義とはあまり共起しないと考えたためである．具体的な設定式を以下に示す

$$\begin{cases} P^0(f_i|s_{n+1}) = \dfrac{1}{|F| + (r'-1)n'} & f_i \in F' \\ P^0(f_i|s_{n+1}) = \dfrac{r'}{|F| + (r'-1)n'} & f_i \in \overline{F'} \end{cases} \tag{13}$$

n,r,n',r' は語義タグ付きコーパスから得られる統計情報をどれだけ初期パラメタの設定に反映させるかを調整する役割を持つことに注意していただきたい．n または n' が 0 のとき，あるいは r または r' が 1 のときは，語義タグ付きコーパスを使わずに初期パラメタを全て一様分布に設定することに相当する．一方，n,n' や r,r' の値を増やせば増やすほど，語義タグ付きコーパスから得られる統計情報を信頼し，初期パラメタの設定に反映させることになる．

4 実験

WSD の対象単語として表 1 の 10 語を用いた．表中の括弧内の数値は EDR 概念辞書で定義されている語義の数である．これらの対象単語は，予備調査によって未定義語義で使われる可能性の高い語を調べて選んだ．

表 1: 対象単語

気持ち (3)，教える (2)，決める (5)，情報 (3)，朝 (2)， 世紀 (2)，電話 (2)，非 (3)，与える (3)，条件 (2)

EM アルゴリズムでの学習に用いる正解なしコーパスとして毎日新聞の 12 年分のコーパスを，定義済みの語義が付与されたコーパスとして EDR コーパス [4] を利用した．テストデータとして，学習に用いていなかった毎日新聞の記事の中から対象単語毎に 100 語選択し，正

解語義を人手で付与した．のべ 1000 語のうち未定義語義を割り当てたのは 258 語である．未定義語義の例としては，1 節で挙げた「電話」の他に,「朝」という単語が持つ (北朝鮮) という意味や,「情報」という単語が持つ (情報科学) という意味などがある．

表 2: 未定義語義に対する正解率

	R_u	P_u	F_u	A_d	A_a
EM+init	0.26	0.63	0.37	0.55	0.47
MLE	–	–	–	0.56	0.56
MLE-u	0.61	0.25	0.36	0.27	0.35

実験結果を表 2 に示す．R_u，P_u，F_u は未定義語義判別の評価指標で，それぞれ未定義語義の再現率，精度 (適合率)，F 値 [1]である．一方，A_d は定義済み語義に対する WSD の正解率 [2]，A_a は定義済み，未定義語義全てに対する WSD の正解率である．比較したシステムは以下の 3 つである．

- EM+init
  本研究の提案手法で，初期パラメタを 3.2.2 で述べた方法で設定し，かつ EM アルゴリズムでパラメタ推定する手法である．
- MLE
  語義タグ付きコーパスから最尤推定によって Naive Bayes モデルを学習し，これを用いて語義を判別する手法．常に定義済みの語義を選択し，未定義語義は出力しない．
- MLE-u
  モデルは MLE と同じだが，一位の語義の確率がある一定の閾値以下なら未定義語義と判別する手法である．閾値は F_u が最大となるようにして決めた．

表 2 から，提案手法 EM+init の F_u は 0.37 であることがわかる．これは未定義語義の判別が必ずしもうまくいっていないことを示唆する．特に再現率が低いことから，未定義語義を定義済みの語義として誤って判定することが多いことがわかる．一方，既存の教師あり学習手法を未定義語義の判別に応用した MLE-u では，提案手法と同様の F_u の値が得られたが，定義済み語義の正解率 A_d が大きく低下し，全体の正解率 A_a も提案手法より劣る．このことから，提案手法のアプローチは未定義語義を判別する手法として有望であると言える．

また，全体の正解率 A_a が一番高いのは MLE である．すなわち，未定義語義を考慮せずに定義済みの語義だけを選択するモデルを学習した方が，未定義語義に対する判別は常に誤るが，全体としては正解率が高くなる．これは WSD の分野において教師あり学習の手法が成功を収めていることの顕れかも知れない．そこで，我々は次のような手法を検討している．まず，未定義語義かそうでない語義かを判別する．このモデルは提案手法のような EM アルゴリズムで学習する．定義済みの語義であると判定された語については，定義済み語義の中から語義を選択するモデルによって最終的な語義を判定する．後者のモデルとして既存の教師あり学習の手法が適用できるため，全体の正解率の向上が期待できる．

最後に 3.2.2 で述べたパラメタについて述べる．n' と r' については，$n' = 0$ かつ $r' = 1$ のとき，すなわち $P(f_i|s_{n+1})$ の初期値を一様分布にしたときの方が F_u が高かった．一方，n と r については，$n = 5, 10, 15, 20$, $r = 5, 10, 15, 20$ と変動させたところ，F_u の値が一番大きくなるのは $n = 10$, $r = 5$ のときであった．表 2 にはこのときの実験結果を載せている[3]．また，全般に，n を大きくすればするほど，また r を大きくすればするほど，未定義語義と判別する回数が少なくなり，P_u は上がるが R_u が下がる傾向が見られた．

5 おわりに

未定義語義を判別する語義曖昧性解消の一手法を提案した．今後は，4 節で述べたように，最初は未定義語義か否かを，定義済み語義と判定されたときにはさらに語義の選択を行う二段階の語義曖昧性解消手法を早急に実装し，その有効性を実験で確認したい．また，本研究では未定義語義の数は 1 つとしたが，未定義語義の数を推測する手法についても検討したいと考えている．

参考文献

[1] Nancy Ide and Jean Veronis. Introduction to the special issue on word sense disambiguation. *Computational Linguistics*, Vol. 24, No. 1, pp. 1–40, 1998.

[2] 菊田篤史. 未定義語義の判別を含む語義曖昧性解消. Master's thesis, 北陸先端科学技術大学院大学, 3 2006.

[3] Christopher D. Manning and Hinrich Schütze. *Foundations of Statisitcal Natural Language Processing*. The MIT Press, 1999.

[4] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

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

[6] SENSEVAL-2 日本語タスク. 自然言語処理特集号, Vol. 10, No. 3, 2003.

[1] $F_u = \frac{2P_uR_u}{(P_u+R_u)}$ とした．

[2] システムが判定した語義と正解の語義が一致した割合．

[3] すなわち，今回の実験は，n,r,n',r' の調整に関してはクローズドテストとなっている．