社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

辞書定義文を用いた低頻度語のための語義曖昧性解消モデルの学習

白井　清昭†　　八木　恒和†

† 北陸先端科学技術大学院大学 情報科学研究科　〒923–1292 石川県能美郡辰口町旭台 1-1
E-mail: †{kshirai,t-yagi}@jaist.ac.jp

あらまし　近年，教師ありの機械学習が語義曖昧性解消の手法として主流となっている．しかし，教師あり学習では，コーパスにあまり出現しない単語については語義曖昧性を解消するモデルを学習することができないことが問題となる．本研究は，辞書定義文から語義の上位概念を抽出し，これを予測する確率モデルを学習することにより，低頻度語の語義の曖昧性を解消する手法を提案する．上位概念は複数の語義で共有されるため，語義そのものを予測する確率モデルを学習するときと比べて訓練データを増やす効果が得られる．評価実験の結果，低頻度語についてはベースラインと比べて再現率が大きく向上することを確認した．
キーワード　語義曖昧性解消，辞書定義文，上位概念，低頻度語

Learning Word Sense Disambiguation Model for Low Frequency Words using Dictionary Definitions

Kiyoaki SHIRAI† and Tsunekazu YAGI†

† School of Information Science, Japan Advanced Institute of Science and Technology
1-1, Asahidai, Tatsunokuchi, Ishikawa 923–1292 Japan
E-mail: †{kshirai,t-yagi}@jaist.ac.jp

Abstract　Currently, supervised machine learning techniques are mainly studied for word sense disambiguation, however, disambiguation models could not be learned for low frequency words. In order to disambiguate word senses for low frequency words, this paper proposed the method to extract superior concepts of senses from their dictionary definitions and learn the probabilistic model predicting them. Much training data were available for learning the model, because superior concepts were shared for multiple senses. According to our experiments, the recall of the proposed method remarkably outperformed that of the baseline for low frequency words.
Key words　Word Sense Disambiguation, Dictionary Definition, Superior Concept, Low Frequency Word

1. はじめに

語義曖昧性解消は，文中に現われる単語の意味 (語義) を決める処理であり，機械翻訳をはじめとする様々な自然言語処理に必要とされる技術である．近年では，コーパスを利用して語義の曖昧性を解消する研究が主流であり，中でも教師あり学習による手法が比較的良い成果をあげている [1], [2]．しかし，教師あり学習を行うためには正解付きデータ，すなわち語義タグ付きコーパスを必要とする．そのため，コーパスに現われない単語や出現頻度の低い単語については，語義を判定するモデルを学習することができないという問題点がある．実用的な応用を考えるなら，全ての単語について語義の曖昧性を解消できることが理想的であり，高頻度語しか語義を決めることができないという状況は好ましくない．

この問題に対処するために，本研究では辞書定義文から得られる情報を利用して機械学習を行う手法を提案する．具体的には，辞書定義文から語義の上位概念を抽出し，その上位概念を予測する確率モデルを学習する．辞書定義文から抽出される上位概念は複数の語義で共有されるため，語義そのものを予測する確率モデルと比べて訓練データの量を確保しやすい．このため，低頻度語についても語義の曖昧性を解消するモデルを学習できると考える．

2. 節では提案手法の基本的な考えを説明し，語義曖昧性解消のための確率モデルについて述べる．3. 節では，辞書定義文から語義の上位概念を抽出する手法について述べる．4. 節では提案手法を実験的に評価する．5. 節では関連研究について述べる．最後に 6. 節で本研究のまとめと今後の課題について述べる．

2. 提 案 手 法

2.1 概　　要

以下の例文 (A) を用いて提案手法の基本的な考えを説明する．

(A) 坂野さんは１７歳の時から浅草で修業、活動写真の弁士などを経て漫談の世界に入り、５７年度の芸術祭では大賞を受賞している。

本研究では，語義の定義として EDR 概念辞書 [5] による語義立てを用いる．EDR 概念辞書に記載されている「漫談」の語義を以下に挙げる．6 桁の英数字は概念 ID と呼ばれる語義の識別子である．

3c5631　こっけいな話をしながら, その中で社会批評や風刺をする演芸
1f66e3　とりとめもない話

いま，語義曖昧性を解消するモデルを正解付きデータから機械学習することを考える．しかしながら,「漫談」が訓練コーパス中に一度も現われない単語であるならば,「漫談」の語義が 3c5631 か 1f66e3 を決定するモデルを学習することは不可能である．

そこで，本研究では辞書定義文に着目する．まず，それぞれの辞書定義文の最後に現われる名詞をその語義の上位概念とみなす．例えば，3c5631 の語義の上位概念は「演芸」であり，1f66e3 の語義の上位概念は「話」である．先ほど述べたように，もし「漫談」が訓練コーパスに一度も現われない場合，3c5631 と 1f66e3 のどちらが正しい語義であるかを選択するモデルを学習することはできない．しかし,「漫談」の上位概念が「演芸」か「話」であるかを選択するモデルを学習することは可能である．なぜなら，EDR 概念辞書中には，他にも演芸や話を上位概念とする語義が存在し，これらの語義が訓練コーパス中に存在する可能性があるためである．そのような語義の例を図 1 に挙げる．一般に，機械学習による語義曖昧性解消は，語義と語義を決める単語の周辺に出現する単語 (以下，周辺語と呼ぶ) との共起関係を学習することにより行われる.「漫談」が訓練コーパスに存在しない場合,「漫談」の語義 (3c5631 または 1f66e3) と周辺語との共起性は学習できないが，図 1 に挙げたような単語と語義が訓練コーパスにある場合には，演芸や話といった上位概念と周辺語との共起性は学習可能である．例えば,「落語の世界」「猿楽の世界」のように，“の世界” の前に演芸を上位概念とする語義はよく現われるが，話を上位概念とする語義はあまり現われない，といった傾向が学習できる．そのため，例文 (A) における「漫談」の後にも “の世界” という表現があるので，この「漫談」の語義の上位概念は演芸であると判断でき，正しい語義として 3c5631 を選択できる．

【落語】 10d9a4 こっけいな話を続け, 最後に落ちをつける寄席演芸
【猿楽】 3c3fbb 猿楽という中世の民衆演芸
【紙切】 3c5ab3 紙を切り抜いていろいろの形を作る演芸
【伝説】 3cf737 昔から民間に語り伝えられた話
【実話】 0f73c1 実際にあった本当の話
【裏話】 3c3071 一般には知られていない内輪の話

図 1　演芸または話を上位概念とする語義の例

このように，辞書定義文から語義の上位概念を抽出し，上位概念と周辺語との共起性を学習することにより，訓練コーパスに一度も出現しない単語の語義を決定することができる．また，一般に上位概念は複数の単語の語義で共有されるため，上位概念のコーパスにおける出現頻度は語義そのものの出現頻度に比べて高くなる．したがって，出現頻度が低く，信頼できるモデルを学習することができない単語についても，上位概念を利用することにより訓練データの量を増やす効果があると期待できる．

2.2 モ デ ル

本項では，辞書定義文から抽出された上位概念を用いた語義曖昧性解消モデルについて述べる．まず，ある単語 w の語義を決定するために，以下のような確率モデルを考える．

$$P(s, c|F) \tag{1}$$

式 (1) において，s は w の語義，c は辞書定義文から抽出された s の上位概念である．また，F は w を含む入力文から得られる素性の集合であり，w の周辺語などがその要素となる．本研究で用いた素性の具体的な内容については 2.4 項で述べる．

次に，式 (1) を以下のように近似する．

$$P(s, c|F) = P(s|c, F)P(c|F) \simeq P(s|c)P(c|F) \tag{2}$$

ここでは，式 (2) の第 1 項 $P(s|c, F)$ を $P(s|c)$ として近似している．$P(s|c, F)$ は入力文から得られる素性集合 F と上位概念 c から語義 s を予測するモデルであり，F から s を予測するという点で naive Bayes モデルによる語義曖昧性解消モデル [6] とほぼ等価である．しかし，低頻度語については語義 s の出現頻度が低いため，統計的に信頼できるモデルが学習できないと考える．そのため，語義 s は語義の上位概念 c のみに依存するとみなし，$P(s|c)$ のように近似する．一方，式 (2) の第 2 項 $P(c|F)$ は素性集合 F から語義の上位概念 c を予測するモデルである．2.1 項で述べたように，語義の上位概念は複数の単語で共有されることから，語義 s よりもコーパスにおける出現頻度は高いため，$P(c|F)$ は低頻度語の語義曖昧性解消を行う際でも十分学習可能と考えられる．

さらに，ベイズの定理を用いて以下のような変形を行う．

$$P(s|c)P(c|F) = \frac{P(s)P(c|s)}{P(c)} \frac{P(c)P(F|c)}{P(F)} \tag{3}$$

$$= \frac{P(s)P(F|c)}{P(F)} \tag{4}$$

$$\simeq \frac{P(s)\prod_{f_i \in F} P(f_i|c)}{P(F)} \tag{5}$$

(3) から (4) の変形では $P(c|s) = 1$ とした．これは，3. 節で述べるように，語義 s の辞書定義文から抽出される上位概念 c は常に一意に決まるためである．また，(4) から (5) の変形において，F 中の各素性 f_i の出現は互いに独立であるとして近似している．

本研究では，式 (5) の確率を最大にする語義 s を選択することによって語義曖昧性解消を行う．ここで，全ての語義について F は同じであるので，$P(F)$ の計算は省略できる．

$$\begin{aligned} &\arg\max_{s\in S_w} \frac{P(s)\prod_{f_i\in F}P(f_i|c)}{P(F)} \\ =\ &\arg\max_{s\in S_w} P(s)\prod_{f_i\in F}P(f_i|c) \end{aligned} \tag{6}$$

式 (6) の S_w は辞書に登録されている w の語義の集合である．直観的に言えば，式 (6) の第 1 項 $P(s)$ は語義の出現頻度を学習するモデル，第 2 項 $\prod P(f_i|c)$ は語義の上位概念 c と素性 f_i の共起性を学習するモデルである．

ここで，式 (6) の確率モデルは全ての単語について適用可能であることに注意していただきたい．教師あり学習に基づく手法では，個々の単語毎に語義曖昧性を解消するモデルを学習する場合が多い．しかし，単語毎にモデルを学習する場合は，モデルの学習に時間を要する点や，モデルを格納するために多くのディスクスペースを要する点が問題となる．特に，低頻度語の語義曖昧性を解消する場合には，単語の異なり数も飛躍的に増大するため，上記の問題は無視できない．これに対し，本手法で学習するべきモデルは式 (6) の確率モデル 1 つだけであり，この点からも低頻度語の語義曖昧性解消に適しているといえる．

2.3 パラメタ推定

式 (6) に示したように，本研究で推定するべき確率モデルは $P(s)$ と $P(f_i|c)$ である．まず，$P(s)$ は加算スムージングで推定した (式 (7))．

$$P(s) = \frac{O(s)+\alpha}{\sum_s O(s)+\alpha V} \tag{7}$$

$O(s)$ は語義 s の出現頻度，α は全ての事象に足すべき頻度，V は辞書中の語義の総数である．ここでは $\alpha = 0.5$ とした．一方，$P(f_i|c)$ は線形補間法で推定した．すなわち，式 (8) のように素性の出現確率 $P(f_i)$ との混合モデルとして推定する．両者の重み β は実験的に求める．$P_{MLE}(f_i|c)$ は式 (9) のように最尤推定によって推定した．一方，$P(f_i)$ は式 (10) で推定した．T は訓練データの総数である．また，分子と分母にそれぞれ 1 と 2 を加えているのはスムージングのためである．

$$P(f_i|c) = \beta P_{MLE}(f_i \mid c) + (1-\beta)P(f_i) \tag{8}$$

$$P_{MLE}(f_i|c) = \frac{O(f_i,c)}{\sum_{f_i} O(f_i,c)} \tag{9}$$

$$P(f_i) = \frac{O(f_i)+1}{T+2} \tag{10}$$

2.4 素　　性

式 (6) のモデルに用いる素性 f_i として以下のものを用いた．いずれも語義曖昧性解消においてよく用いられる素性である．

- w の直前または直後の単語
- w の直前または直後の品詞
- w の直前または直後に現われる 2 つの単語の組
- w の前後に現われる 2 つの単語の組
- 同一文中にある自立語の基本形
- w に係る格と格要素の組 (w が用言のとき)
- w の格と係り先用言の組 (w が格要素のとき)
- 係り先文節の主辞 (w が文節の主辞のとき)
- 同一文節の主辞 (w が文節の主辞ではないとき)

3. 辞書定義文からの上位概念の抽出

3.1 上位概念抽出パタン

本項では，辞書定義文から上位概念を抽出する手法について述べる．基本的には，先行研究 [8] と同様に，辞書定義文の末尾にある単語をその語義の上位概念とみなす．また，取り出すべき単語は品詞及び単語表記のパタンマッチにより決める．P1 は抽出パタンの例である．

P1　1:(*)(形容詞) 2:(さ)(名詞) → 1 2

矢印の左辺は辞書定義文の末尾にマッチさせるべきパタンである．パタンの 1 つの要素は (基本形)(品詞) といった組で表現される．“*” は任意の単語または品詞にマッチすることを表わす．一方，右辺は上位概念として返すべき要素を表わす．P1 の場合，左辺の 1 と 2 の部分にマッチした要素を連結して，上位概念として返す．例えば,「海抜」の辞書定義文は「海面からの高さ」であるが，P1 より「高さ」が上位概念として抽出される．

上位概念抽出パタンは全て人手で作成した．最終的に作成したパタンの数は 64 である．以下，主な抽出パタンを品詞別に述べる．

名　　詞

名詞の辞書定義文は多くの場合名詞で終わるので，文末にある名詞を上位概念として取り出す．ただし，複合名詞の場合は最後の名詞のみを上位概念として抽出する．これを行うのが以下の P2 である(注1)．

P2　1:(*)(名詞) → 1
ex. 【IMF】国際通貨基金という国連 機関→ **機関**

また,「〜すること」のように，用言の次に形式名詞が現われる辞書定義文が多く存在する．このときには,「用言 + こと」という形式で上位概念を抽出する (P3).

P3　1:(*)(動詞) 2:(こと | さま | もの)(名詞) → 1 + 2
ex. 【藍染】藍で 染めること→ **染める** + **こと**

名詞の辞書定義文が体言ではなく用言で終わる場合がある．特にサ変名詞の辞書定義文によく見られる．このとき，名詞の上位概念は名詞であるべきと考え,「用言+こと」のように「こと」を補って上位概念として抽出する (P4).

P4　1:(*)(動詞)$^+$ → 1 + こと
ex. 【一顧】ちょっと心にとめて 考える→ **考える** + **こと**

動　　詞

動詞の辞書定義文は多くの場合動詞で終わるので，文末に現われる動詞を上位概念として抽出する．ただし，複合動詞の場合は先頭の動詞のみ取り出す (P5).

(注1)：各パタンの下に，そのパタンによる上位概念の抽出例を挙げた．下線はパタンの左辺にマッチした部分，太字は抽出された上位概念を表わす．

P5　1:(*)(動詞)$^{+}$ → 1
ex. 【洗い切る】事情を 調べ尽くす→ 調べる

また,「〜すること」のように動詞の後に形式名詞が続く場合は,形式名詞を取り除いて上位概念を抽出する (P6).

P6　1:(*)(動詞)$^{+}$ 2:(こと | さま | もの)(名詞) → 1
ex. 【網打する】投網を打って魚を とること→ とる

形　容　詞

形容詞の辞書定義文は「〜するさま」のように用言の後に「さま」が続くものが多い．このとき，用言が形容詞ならその形容詞を (P7)，それ以外なら「用言 + さま」を上位概念として抽出する (P8).

P7　1:(*)(形容詞) 2:(さま)(名詞) → 1
ex. 【細かい】(形が) 非常に 小さいさま→ 小さい

P8　1:(*)(動詞)$^{+}$ 2:(さま)(名詞)→ 1 + さま
ex. 【苦しい】心身に痛みを 感じるさま→ 感じる + さま

接　尾　語

接尾語の上位概念は，基本的には名詞と同じパタンを用いて抽出する．ただし，接尾語の辞書定義文に特有の表現がいくつかあったため，これらについては別にパタンを用意した．例えば,「〜の単位」という定義文に対して，P2 を用いると「単位」が上位概念として抽出される．しかし，その前に現われる名詞も含めた方が上位概念として適していると考え，パタン P9 を作成した．また,「〜を表わす語」という定義文に対応するためにパタン P10 を作成した．

P9　1:(*)(名詞) 2:(の)(助詞) 3:(単位)(名詞)→ 1 の単位
ex. 【ワット】ワットという 電力の単位→ 電力の単位

P10　1:(*)(名詞) 2:(を)(格助詞) 3:(表わす)(動詞)
4:(語)(名詞) → 1
ex. 【掛け】10 分の 1 の 割合を表す語→ 割合

その他の品詞

形容動詞は名詞と同じパタンを用いて上位概念を抽出した．また，副詞については，辞書定義文のパタンが形容詞と類似していたため，形容詞の抽出パタンを用いた．サ変名詞は，文中では名詞としても動詞としても使われる可能性があるが，名詞として使われている場合には名詞の上位概念を，動詞として使われている場合には動詞の上位概念を抽出するべきと考え，名詞と動詞の抽出パタンを使って 2 つの上位概念を抽出した．実際に語義曖昧性解消を行う場合，サ変名詞が名詞として使われている場合には名詞の上位概念を，動詞として使われている場合には動詞の上位概念を用いた．

3.2　EDR 辞書からの上位概念の抽出

3.1 項で述べた抽出パタンを用いて，EDR 日本語単語辞書 [5] にある概念説明 (辞書定義文) から上位概念を抽出した．辞書定義文の形態素解析には茶筌(注2)を用いた．結果を表 1 に示す．

表 1 の「語義数」の行は EDR 日本語単語辞書に含まれる語義の総数,「上位概念」の行は語義の定義文から抽出された上位概念ののべ数と異り数を示している．また，括弧内の数値は上位概念を抽出することのできた語義の割合を表わす．名詞や動詞についてはほとんどの語義の定義文から上位概念を取り出すことができたが，その他の品詞については 7 割から 9 割程度の語義に対してしか上位概念を抽出することができなかった．上位概念の抽出に失敗する主な原因は抽出パタンの不足であり，抽出パタンを追加することによってある程度対処できると思われる．一方，1 つの上位概念は複数の語義の定義文から取り出される可能性がある.「平均語義数」の行は，抽出された上位概念ののべ数を異り数で割った値であり，同じ上位概念が抽出される語義数の平均を示している．2.1 項で述べたように，上位概念ができるだけ多くの語義で共有されていればいるほど，低頻度語に対して訓練データの量を増やす効果が大きいと考えられる．

表 1　上位概念の抽出

	名詞	動詞	形容詞	接尾語	形容動詞	副詞
語義数	170434	30962	1572	1558	5186	3047
上位概念 のべ	169246 (0.99)	30312 (0.98)	1230 (0.78)	1362 (0.87)	4683 (0.90)	2203 (0.72)
異り	24061	8743	638	671	2253	936
平均語義数	7.0	3.5	1.9	2.0	2.1	2.4

表 2　上位概念が抽出された単語数

全て抽出	一部抽出	抽出失敗
87,449 (0.9455)	4,020(0.0435)	1,020(0.0110)

表 3　同じ上位概念が重複して抽出された単語数

重複なし	一部重複	全て重複
74,186 (0.8021)	8,490(0.0918)	8,793(0.0951)

本研究で辞書定義文から上位概念を抽出するのは，語義曖昧性解消を行うモデルを学習するためである．一つの単語が複数の語義を持つとき，それらの全てから上位概念を抽出することができない場合や，全ての語義から同じ上位概念が抽出された場合は，上位概念の抽出が語義曖昧性解消に有効であるとはいえない．表 2, 3 は，語義曖昧性解消にどれだけ有効かという観点から抽出された上位概念を評価したものである．表 2 の「全て抽出」は，単語が複数の語義を持つとき，その全ての語義から上位概念を抽出できた単語数,「一部抽出」は一部の語義からのみ上位概念を抽出できた単語数,「抽出失敗」は全ての語義について上位概念を取り出すことができなかった単語数である．一方，表 3 の「重複なし」は，単語が複数の語義を持ちかつ 1 つ以上の語義から上位概念を取り出すことができたとき，その全ての語義に対して異なる上位概念を抽出できた単語数,「一部重複」は一部の語義について同じ上位概念を抽出した単語数,「全て重複」は全ての語義から同じ上位概念を抽出した単語数を表わす．また，括弧内の数値は EDR 日本語単語辞書に含まれる多義語の総数 (92,489) に対する割合である．

表 3 の「重複なし」に相当する単語は，語義毎に異なる上位概念が抽出されているため，これらを用いることにより語義曖

(注2)：http://chasen.aist-nara.ac.jp/index.html.ja

味性解消の正解率の向上が期待できる．一方，「一部重複」に相当する単語についても，正しいと思われる語義の数を絞り込む効果があると考えれば，上位概念が語義曖昧性解消の正解率向上に寄与するとみなせる．したがって，両者をあわせた約89%の単語について，上位概念を抽出する効果があると期待できる．

4. 評価実験

本節では，提案手法の評価実験について述べる．

4.1 実験手順

実験には EDR コーパス [5] を用いた．EDR コーパスは約 20 万文からなるコーパスであり，各単語に EDR 概念辞書の概念 ID が付与された語義タグ付きコーパスである．EDR コーパスのうち，20,000 文をテストデータ，20,000 文を調整用データ，残りの 161,332 文を訓練データとした．調整用データは，語義曖昧性解消モデルの検討や式 (8) のパラメータ β の最適化に用いた．

まず，訓練コーパスを用いて式 (6) の確率モデルを学習した．2.4 項で述べた確率モデルの素性を得るためには，文の形態素解析や文節の係り受け解析が必要となる．今回の実験では JUMAN (注3) と KNP (注4) を用いて解析を行った．語義曖昧性解消は，式 (6) の値が最大となるような語義を選択することにより行った．ただし，最大確率を持つ語義が複数ある場合には，その全てを答えとして選択した．評価尺度として再現率，精度 (適合率)，F 値(注5) を用いた．また，調整用データを用いて式 (8) のパラメータ β の値を変動させ，F 値が最大となる β を求めた．その結果，$\beta = 0.05$ のときが最大であったので，テストデータの語義曖昧性解消の際にはその値を用いた．

4.2 実験結果

テストデータに対する語義曖昧性解消の評価結果を表 4 に示す．テストデータに含まれる評価単語数は 91,986 である．表 4 において，BL は最頻出語義を常に選択するベースラインモデルを表わす．ただし，最頻出語義が複数ある場合にはその全てを答えとして選択する．一方，NB は naive Bayes モデルに基づく提案手法を表わす．表 4 から，提案手法はベースラインと比べて良い結果が得られていることがわかる．とはいえ，F 値で 4%程度しか向上しておらず，また F 値の値自体も約 63%と決して高いとはいえない．

本研究は低頻度語に対して語義曖昧性解消を行うモデルを学習することを目的とし，提案手法を単独で用いることは想定していない．高頻度語においては教師あり学習で学習したモデルを用い，低頻度語においては本研究の手法を用いることを考えている．我々は，予備実験として，Support Vector Machine を学習して語義曖昧性解消を行い，約 70%の精度が得られたことを確認している．したがって，提案手法と教師あり学習の手法を組み合わせることにより，F 値の値は今回の実験よりもはるかに向上すると予想される．

表 4 実験結果

	再現率	精度	F 値
BL	0.5830	0.5972	0.5900
NB	0.6151	0.6414	0.6280

表 5 実験結果 (品詞別)

品詞	名詞	動詞	形容詞	接尾語	形容動詞	副詞
評価単語数	48,607	29,490	3,021	4,316	2,115	2,439
BL(R)	0.6263	0.5708	0.3774	0.5348	0.6128	0.3678
NB(R)	0.6659	0.6170	0.4429	0.6376	0.6227	0.2509
BL(P)	0.6252	0.5728	0.4556	0.5459	0.6131	0.5091
NB(P)	0.6640	0.6194	0.5441	0.6512	0.6254	0.4700
BL(F)	0.6258	0.5718	0.4128	0.5403	0.6129	0.4270
NB(F)	0.6650	0.6182	0.4883	0.6443	0.6240	0.3272

表 6 低頻度語に対する評価

頻度	$=0$	$\leqq 1$	$\leqq 2$	$\leqq 5$	$\leqq 10$	$\leqq 20$
評価単語数	782	1,432	2,031	3,735	5,998	9,622
BL(R)	0.2826	0.4120	0.4820	0.5365	0.5639	0.5749
NB(R)	0.5307	0.5468	0.5692	0.5922	0.6040	0.6086
BL(P)	0.4529	0.5082	0.5364	0.5585	0.5722	0.5786
NB(P)	0.4689	0.5091	0.5387	0.5731	0.5927	0.6027
BL(F)	0.3480	0.4551	0.5078	0.5473	0.5680	0.5768
NB(F)	0.4979	0.5273	0.5535	0.5825	0.5983	0.6056

表 5 は，語義曖昧性解消の正解率を品詞別に評価した結果である．表 5 中の R,P,F はそれぞれ再現率，精度，F 値を表わす．F 値を比較すると，特に接尾語や形容詞についてはベースラインと比べて改善の度合が大きい．一方，副詞についてはベースラインよりも約 10%程度も劣る．この原因については調査中であるが，副詞の上位概念を抽出する際には副詞のための抽出パタンを作成せず，形容詞の抽出パタンを用いた (3.1 項) ことが一因かも知れない．

低頻度語における正解率を比較するため，訓練コーパスにおける出現頻度が 0, 1 以下，2 以下，5 以下，10 以下，20 以下の単語について再現率，精度，F 値を求め，ベースラインとの比較を行った．結果を表 6 に示す．表 6 中の「評価単語数」の行は，それぞれの頻度の条件を満たす単語の数を表わす．

表 6 から，訓練データにおける頻度が小さい単語ほど，ベースラインと提案手法の F 値の差は大きくなることがわかる．言い換えれば，低頻度の単語ほど提案手法による改善の度合が大きい．精度を比較すると，ベースラインと提案手法の差はそれほど顕著ではなく，頻度 2 以下の単語を評価対象としたときにはほとんど差はない．これに対し，再現率の差は顕著である．ベースラインの場合，頻度 0 の単語のみを評価対象としたときの再現率は，テストデータ全体の再現率と比べて約 30%程度も低下するのに対し，提案手法の場合は約 8.5%程度しか低下しない．これは，2.1 項で述べたように，辞書定義文から抽出された上位概念を確率モデルに組み込んだことにより，訓練データを増加させる効果が生まれたためとみなすことができる．

(注3)：http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
(注4)：http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
(注5)：F 値は $\frac{2RP}{R+P}$ とした．(R は再現率，P は精度)

5. 関 連 研 究

本研究では，訓練コーパスに出現しない単語や出現頻度の低い単語の語義を決定することを主な問題としている．この問題に対するひとつの解決策は，語義タグ付きコーパスを使わない手法を用いることである．例えば，辞書の定義文だけを手がかりとする手法 [3] や，EM アルゴリズムなどの教師なし学習を用いる手法がある．しかし，これらの手法は一般に教師あり学習を用いる手法と比べて正解率が劣る．

訓練データの量を確保するために，小量の正解付きデータと大量の正解なしデータを用いる手法が数多く行われている．語義曖昧性解消を対象にした研究としては Yarowsky の研究 [9] が有名である．Yarowsky は，まず正解付きデータから決定リストを学習し，これを正解なしデータに適用することにより自動的に正解付きデータを作成する．この操作を段階的に繰り返すことにより徐々に正解付きデータを増やし，最終的な決定リストを学習する．このような小量の正解付きデータと大量の正解なしデータを用いる手法は，訓練データを自動的に増やすという効果はあるが，最初に小量の正解付きデータを必要とするため，正解付きデータのない単語についてはやはり語義の曖昧性を解消することができないという問題点がある．もちろん，正解付きデータのない単語については新たに正解付きデータを作成すればよいのだが，実用的な自然言語処理アプリケーションで取り扱うべき単語の数は非常に多く，これら全ての単語に対して小量といえども正解付きデータを新たに作成するのはコストがかかる．これに対し，本研究では小量の正解データを作る必要はない．上位概念の抽出パタンは人手で作成する必要があるが，3.2 項で述べたように，抽出パタンを一度作成すれば辞書中の約 9 割の単語について訓練データの量を増やす効果があると見込まれる．そのため，全ての単語について小量の正解付きデータを用意するよりもはるかに少ないコストで多くの単語に対応できる点が優れている．

本研究は，語義曖昧性解消を行う知識源として，語義タグ付きコーパスと辞書の定義文を用いる研究とみなすことができる．同様の研究は過去にも行われている．Litkowski は，国語辞典を用いた手法と機械学習による手法を組み合わせる方法を提案している [4]．この手法では，辞書定義文を用いた手法が出力する語義 (辞書の語義立て) と機械学習による手法が出力する語義 (WordNet の意味クラス) が異なり，前者の語義を後者の語義に変換してから両者を混合している．ただ，このような語義の変換を精度良く行うことは一般には難しい．玉垣らもまた，国語辞典を用いた手法と機械学習による手法を組み合わせる手法を提案している [7]．彼らの手法は，国語辞典に記載された用例と文法情報を利用しており，辞書定義文そのものを利用していない点が本手法とは異なる．しかし，玉垣らの手法と本手法は相反するものではなく，むしろ両者を組み合わせることにより更なる再現率の向上が期待できる．

6. お わ り に

本研究では，辞書定義文から抽出した語義の上位概念と周辺語の共起関係を反映した語義曖昧性解消モデルを提案した．評価実験の結果，提案手法は訓練データに現われない単語もしくは低頻度でしか現われない単語の語義曖昧性解消に有効であることを確認した．

4.2 項でも述べたが，我々は提案手法を低頻度語の語義曖昧性解消に，教師あり学習による手法を高頻度語の語義曖昧性解消に用い，全体の正解率を向上させることを考えている．今後は，提案手法と機械学習手法の最適な組み合わせ方法について検討したい．また，辞書定義文からの上位概念の抽出方法については改良の余地がある．上位概念は次のように抽出できることが理想的である．

（1） 全ての辞書定義文から抽出できる．

（2） 同じ単語の異なる語義からは異なる上位概念が抽出される．

（3） 辞書全体から抽出される上位概念の異り数が小さい (異り数が小さいと訓練データを増やす効果が向上するため)．

しかし，(2) と (3) は一般には両立しない．例えば，本研究では基本的には一つの単語を上位概念として抽出しているが，複合語や短かい句を上位概念として抽出するようにすれば，(2) の条件を満たすようにすることができる．一方，辞書全体から抽出される上位概念の異り数は増加してしまう．したがって，最適な上位概念の抽出方法を探究する必要がある．

文　　献

[1] *ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 2002.

[2] *Natural Language Engineering – Special Issue on Evaluating Word Sense Disambiguation Systems*, Vol. 8, No. 4, 2002.

[3] Jim Cowie, Joe Guthrie, and Louise Guthrie. Lexical disambiguation using simulated annealing. In *Proceedings of the International Conference on Computational Linguistics*, pp. 359–365, 1992.

[4] Kenneth C. Litkowski. Sense information for disambiguation: Confluence of supervised and unsupervised methods. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Success and Future Direction*, pp. 47–53, 2002.

[5] 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.

[6] Ted Pedersen. A simple approach to building ensembles of naive baysian classifiers for word sense disambiguation. In *Proceedings of the the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 63–69, 2000.

[7] 玉垣隆幸, 白井清昭. 読解支援システムのための語義曖昧性解消に関する研究. 言語処理学会第 9 回年次大会, pp. 481–484, 2003.

[8] 鶴丸弘昭, 兵頭竜二, 松崎功, 日高達, 吉田将. 語義を考慮した単語間の階層構造の抽出について. 情報処理学会情報処理学会自然言語処理研究会, pp. 9–16, 1987.

[9] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.