修 士 論 文

未定義語義の判別を含む語義曖昧性解消

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

菊田 篤史

2006 年 3 月

修 士 論 文

未定義語義の判別を含む語義曖昧性解消

指導教官　白井清昭 助教授

審査委員主査　白井清昭 助教授
審査委員　島津明 教授
審査委員　東条敏 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

410036 菊田 篤史

提出年月: 2006 年 2 月

概 要

文中の単語が，辞書に定義されている複数の意味のうち，どの意味で使用されているかを自動的に判別する処理を語義曖昧性解消 (Word Sense Disambiguation:WSD) という．しかし，辞書に未定義の語義を持つ単語も存在し，従来の辞書に定義されている語義の中から適切なものを選択するという手法では，必ず間違った語義を選択してしまい機械翻訳や情報検索のタスクの誤りの原因となる問題がある．そこで，本研究では，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手法により，辞書に定義されている語義に加え，辞書に未定義であるということも判別できる語義曖昧性解消システムを構築する．

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　研究の背景・目的

文中の単語が，辞書に定義されている複数の意味のうち，どの意味で使用されているかを自動的に判別する処理を語義曖昧性解消 (Word Sense Disambiguation:WSD) という．この処理は，機械翻訳や情報検索などの様々なタスクに幅広く応用することができる．例えば,「株」という単語を含む次の２つの文があったとする．

1. お金を**株**に投資する．(invest a money in **stocks**)

2. あなたの**株**が上がる．(increase in your **estimation**)

1 の文では「株式，株券」，2 の文では「評価」という意味で使用されている．ここで，語義曖昧性解消により適切な語義を判別することができれば，機械翻訳で正しい訳を得ることができる．しかし，辞書に未定義の語義を持つ単語も存在する．例えば,「電話」という単語について岩波国語辞典には,「電話機による通話」,「電話機の略」という２つの語義が定義されている．ところが，文中で電話という単語が「電話番号」という語義で使用されている場合がある．このとき，従来の辞書に定義されている語義の中から適切なものを選択するという手法では，必ず間違った語義を選択してしまい機械翻訳や情報検索のタスクにおいて，間違った訳をユーザに提示したしするなどの問題が起こる．そこで，本研究では，文中の単語の語義を判別する際に，辞書に定義されている語義に加え，辞書に未定義であるということも判別できる語義曖昧性解消システムを構築することによりこの問題の解決を図る．

語義曖昧性解消は，コーパスと呼ばれる例文集を用いて語義を判別するモデルを学習する機械学習の手法が主流である．機械学習の手法の中でも，例文中に単語の正しい語義がタグ付けされている語義タグ付きコーパスを用いて規則を学習する教師あり学習が良い成果を挙げている．しかし，未定義であることを示した語義タグ付きコーパスは存在しないため，この手法をそのまま適用することはできない．よって，機械学習のもう１つの手法である教師なし学習を行う．

教師なし学習は，何も情報が付加されていないプレーンテキストコーパスを用いて学習する手法であり，先行研究として，Manning らの研究がある．Manning らは，語義を判別するモデルである Naive Bayes モデルを EM アルゴリズムと呼ばれる学習アルゴリズムを用いて学習している．また，新納らは，Manning らと同じモデルと学習アルゴリズム

を用いて語義判別を行っているが，プレーンテキストコーパスに少量の語義タグ付きコーパスを加えて学習の精度を向上させている．

本研究でも，EM アルゴリズムにより Naive Bayes モデルを学習する手法により，辞書に定義されている語義であるか未定義語義であるかを判別する．また，語義タグ付きコーパスから得られる統計情報も利用するが，未定義語義も判別の対象となるため新納らとは異なる手法で学習精度の向上を図る．

1.2 本論文の構成

本論文の構成は以下の通りである．

第 2 章では，教師あり学習と教師なし学習による語義曖昧性解消の先行研究と本研究との関連について述べる．第 3 章では，本研究の提案手法の概要について述べる．第 4 章では，語義判別のモデルと学習アルゴリズムの詳細を述べる．第 5 章では，実験とそれによって得られた結果について述べ，考察を行う．第 6 章では，今後の課題について述べる．

第2章　関連研究

語義曖昧性解消は，コーパスなどの知識源から語義を判別するモデルを学習する機械学習により問題を解決する手法が主流である．モデルの学習方法は，大きく分けて教師あり学習と教師なし学習の２つに分けられる．本章では，この２つの学習方法による語義曖昧性解消の先行研究を紹介し，それらの研究と本研究との関連について述べる．

2.1　教師あり学習による WSD

教師あり学習は，コーパスの例文中の単語に正解語義が付加されている語義タグ付きコーパスを用いて語義を判別するモデルを学習する．曖昧性を解消したい単語の語義を $s_1, \ldots, s_k$，対象語が含まれる文中の素性を $f_1, \ldots, f_i$ とすると，語義 s_k と素性 f_i の共起関係を利用して語義を判別するモデルを学習する．語義判別の手がかりとなる素性は，対象語の直前・直後の単語と品詞，同一文中にある自立語，係り受け関係にある単語などが一般的に用いられている．

教師あり学習は，一般的な手法で最も良い成果をあげている．しかし，正解語義が付加されたコーパスが必要であり，そのコーパスを人手で作成しなければならないため，膨大なコストがかかるという問題がある．また，作成コストの高さのため，膨大な量のデータを取得することが困難であるため，低頻度の語義や素性について十分に学習できないという問題もある．

語義を判別するモデルを学習するアルゴリズムの代表として以下が挙げられる．

- 決定木

  決定木 [9] は，学習データから得られた規則を木で表現していき，最終的に作成された木を調べることにより分類されるクラスを決定する．

- 決定リスト

  決定リスト [5] は,「もし，証拠 E なら，クラス c である」というような，$if - then$ の形で表したルールを信頼度の高い順に並べるアルゴリズムである．優先順位付きの規則が学習されるため，結果の考察が容易であるという利点がある．語義曖昧性解消の場合は，証拠 E が素性 f_i，クラスが語義 s_k となる．

- Support Vector Machine(SVM)

SVM は，Vapnik によって考案された分類アルゴリズムである [11]．学習データを分離し，それぞれの学習データの距離の最小値が最大となるように分類する．このアルゴリズムは，過学習を起こしにくいという特徴がある．

- Naive Bayes

  Naive Bayes[12] は，確率モデルにより分類を行う手法である．ある事例 x が与えられたときにクラスが c である条件付き確率 $P(c|x)$ を最大にするクラスを求めることによる分類を行う．

2.2 教師なし学習による WSD

前節で述べたように，教師あり学習を行うには，語義タグ付きコーパスを作成するために膨大なコストがかかる．そこで，正解の語義が付加されていないプレーンテキストコーパスを用いた教師なし学習による語義曖昧性解消の研究も行われている．教師なし学習の代表的な学習アルゴリズムとして Co-training と EM アルゴリズムがある．ここでは，この 2 つのアルゴリズムとそれを用いた先行研究について述べる．

Co-training

Co-training は，少量の正解付きコーパスから分類器を 2 つ作成し，片方の分類器により，プレーンテキストコーパスについて分類を行う．その結果の中で信頼度が高いものを正解付きコーパスに加える．今度は逆に，もう一方の分類器により同様のことを行う．これを繰り返すことによって，学習データとなる正解付きコーパスの量を増やす手法である．Yarowsky は，この手法により語義タグ付きコーパスの量を増やし，それを用いて決定リストを作成して語義曖昧性解消を行っている [5]．しかし，Co-training は，2 つの分類器を作成する際に独立した素性集合を用いる必要があり，それができなければ精度が落ちるという問題がある．この問題を解決するために，素性間の共起性を考慮して Co-training を行う手法を新納が提案している [6]．

EM アルゴリズム

EM アルゴリズムは，Nigam らが文書分類に対して適用したものであり [7]，コーパスから観測されたデータ $x_1, x_2, \ldots, x_N$ から，確率モデル $P_\theta(x)$ の対数尤度を増加させる未知のパラメータ θ を E(Expectation)-step, M(Maximization)-step という 2 つのステップを繰り返すことにより推定するアルゴリズムである．その概要を以下に示す．

1. 適当な初期パラメータ θ を与える

2. 与えられた θ をもとに E-step，M-step を繰り返しモデルの対数尤度が収束するまで θ を更新していく．

- E-step

  与えられたパラメータ θ を用いて，モデル $P(c|x_i)$ のもとでの $\log P_{\hat{\theta}}(x_i, c)$ の期待値を求める．ここでの c はクラスを表す．

- M-step

  E-step で求めた値をもとにモデルの尤度を最大にするような $\hat{\theta}$ を求め θ を更新する．

このアルゴリズムを Manning らが語義曖昧性解消に適用した [1]．Manning らは，パラメータ θ を語義 s_k の出現確率 $P(s_k)$，語義が s_k であるとき素性 f_i が起こる条件付き確率 $P(f_i|s_k)$ に設定する．対象語と同一文中に出現する単語を素性として用い，パラメータの初期値をランダムに与え，語義判別を行う確率モデルである Naive Bayes モデルを学習している．新納らも同様に Naive Bayes モデルを EM アルゴリズムによって学習する手法を用いて曖昧性解消を行っている [2]．Manning らと異なる点は，EM アルゴリズムにおける確率の反復推定に少量の語義タグ付きコーパスの統計情報を反映させることによって，学習されたモデルの精度を向上させていることである．また，EM アルゴリズムは，一般的に反復回数に比例して学習の精度が向上していくが，逆に精度が低下し収束することもある．それを防ぐために，クロスバリデーションにより最適な反復回数を推定することも行っている．一方，Schutze らは，ベクトル空間をもとにした手法を提案している [8]．この手法も EM アルゴリズムにより Naive Bayes モデルを学習するが，素性，文脈，語義をベクトル空間で表現している点が異なる．Schutze らは，パラメータの初期値を設定する際に，文脈間の類似度により K 個のクラスタを作成し，それぞれのクラスタの平均値を求める．それを使用して K-means アルゴリズムにより文脈のベクトルを最も近いクラスタに割り当てることによりベクトル表現された語義のクラスを作成している．

2.3 本研究との関連

本研究では，辞書に定義されている語義に加え，未定義語義の判別も行う語義曖昧性解消システムを構築する．EM アルゴリズムは未知のパラメータを推定する学習アルゴリズムであり，未定義語義も未知であることから，未定義語義の判別に EM アルゴリズムを適用することは適していると考えられる．また，教師なし学習に未定義語義という未知の要素がさらに加わるため，辞書に定義されている語義の判別の正解率が落ちると考えられる．よって，少量の語義タグ付きコーパスを用いて学習の精度を向上させるという手法は効果的であると考えられる．よって，本研究でも，新納らの手法を参考に Naive Bayes モデルを EM アルゴリズムにより学習する手法により語義曖昧性解消をする．語義タグ付きコーパスの統計情報も使用するが，EM アルゴリズムの学習の過程で用いるのではなく，EM アルゴリズムの初期値設定の段階で利用する．

第3章　提案手法の概要

語義曖昧性解消の先行研究では，曖昧性を解消する単語の語義を辞書に定義されている語義の中から選択していた．しかし，実際には辞書に未定義である語義を持つ単語があり，その未定義の意味で使用されている場合が少なくない．本研究では，辞書に定義されている語義のみではなく，未定義の語義であるということも判別する語義曖昧性解消システムを構築する．本章では，このシステムの概要を述べる．

3.1　未定義語義とは

前述したように語義曖昧性解消の先行研究では，あらかじめ辞書に定義されている語義の中から適切な語義を選択している．このため，文中で辞書に未定義の語義で使用されている単語が含まれていた場合，必ず誤った語義を選択してしまい機械翻訳や情報検索の誤りの原因となるという問題がある．例えば,「電話」という単語について EDR 概念辞書には，以下の２つの語義が定義されている．

1. 電話機の略
2. 電話機による通話

しかし，以下の例文のように「電話番号」という語義で使用されている場合がある．

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。

辞書に定義されている語義から適切な語義を選択する手法では，機械翻訳をする際に誤った訳になってしまう．また，情報検索においても，例えば，語義単位でインデックスを作成する際に誤った語義が登録されてしまう．そこで，本研究では，この問題を解決するために辞書に定義されている語義の判別に加えて，辞書に未定義であるということも判別する語義曖昧性解消システムを構築する．未定義であることを判別できれば，機械翻訳では，少なくとも誤った翻訳をユーザーに提示することを防ぐことができ，情報検索では，語義の情報をもとに適切なインデックスを作成することが可能になる．

3.2 語義のクラス

先行研究では，語義のクラスを辞書に定義されている語義数に設定し曖昧性解消を行っている．語義数が k 個の場合，語義のクラスは $< s_1, s_2, \ldots, s_k >$ である．本研究では，文中の単語が辞書に定義されている語義のいずれか，あいるは未定義語義かを判別するにあたり，語義のクラスを $< s_1, s_2, \ldots, s_k,$ 未定義 $>$ に設定し曖昧性を解消する．ここで，辞書に未定義の語義が複数ある場合もあるが，追加する $<$ 未定義 $>$ のクラスは 1 つとする．前節で例に出した「電話」という単語であれば，以下のような語義のクラスで曖昧性を解消することになる．

1. 電話機の略
2. 電話機による通話
3. 未定義

3.3 学習方法

我々が扱う自然言語による表現は多様であり，人手で規則を作成して自然言語を解析することは非現実的である．語義曖昧性解消においても各単語に複数の語義があり，全ての単語について規則を作成することが困難であるため，コーパス等の知識源から機械学習により問題を解決する手法がとられている．その中でも，語義タグ付きコーパスを利用して語義 s_k と素性 f_i の共起関係を学習し語義を判別する教師あり学習が良い成果を挙げている．しかし，辞書に未定義の語義であるということを明示した語義タグ付きコーパスは存在せず，本研究では教師あり学習の手法を用いることはできない．よって，正解語義の情報が付加されていないプレーンテキストコーパスを用いた教師なし学習のアプローチをとる．本研究でも，2 章で述べた Manning ら，新納らの研究を参考に，学習アルゴリズムとして EM アルゴリズム，語義判別を行う確率モデルとして Naive Bayes モデルを採用し，教師なし学習により曖昧性解消を行う．また，学習の精度を向上させるために，新納らと同じく語義タグ付きコーパスから得られる統計情報を利用するが，EM アルゴリズムの繰り返しの過程で利用するのではなく，パラメータの初期値設定のときのみ利用する．

第4章　モデルの学習

本章では，語義を判別する Naive Bayes モデル，語義判別の手がかりとする素性の設定，学習アルゴリズムである EM アルゴリズムとその初期値の設定について述べる．

4.1　Naive Bayes モデル

本節では，本研究で単語の語義を判別するために用いる Naive Bayes モデルについて述べる．

自然言語処理の多くの問題は分類問題であり，ある事例 x が現れたとき，クラスが y である条件付き確率 $P(y|x)$ を推定することで解決できる．語義曖昧性解消の問題も分類問題として捉え，事例を曖昧性を解消する単語が出現する文脈 c_j，クラスを語義 s_k として $P(s_k|c_j)$ を推定することで解決できる．具体的には，$P(s_k|c_j)$ を最大にする s_k を見つければよい．

$$\arg\max_{s_k} P(s_k|c_j) \tag{4.1}$$

式 (4.1) は，直接求めることができないため，以下のように変形する．

$$P(s_k|c_j) = \frac{P(s_k)P(c_j|s_k)}{P(c_j)} \tag{4.2}$$

$$\rightarrow \; P(s_k)P(c_j|s_k) \tag{4.3}$$

$P(s_k)$ は語義の出現確率，$P(c_j|s_k)$ は語義 s_k であるとき文脈 c_j が生起する条件付き確率，$P(c_j)$ は文脈 c_j の出現確率である．$P(c_j)$ は，全ての語義について等しく省略できるため，式 (4.2) から式 (4.3) のようになる．ここで問題となるのは，$P(c_j|s_k)$ の推定であるが，1 つの語義について文脈は無限にあり，同一のものはほとんどないので推定することは現実的に不可能である．そこで，$P(c_j|s_k)$ に以下の仮定を導入する．

$$P(c_j|s_k) = \prod_{f_i \in c_j} P(f_i|s_k) \tag{4.4}$$

f_i は，曖昧性を解消する単語が出現する文脈中の素性，$P(f_i|s_k)$ は語義が s_k であるときに，素性 f_i が起こる条件付き確率である．式 (4.4) は，文脈中の各素性 f_i は互いに独立であると仮定して近似している．よって，語義判別を行う Naive Bayes モデルは最終的に以下のような形になる．

$$\arg\max_{s_k} P(s_k) \prod_{f_i \in c_j} P(f_i|s_k) \tag{4.5}$$

4.2 素性

本節では，語義を判別するための手がかりとして使用する素性について述べる．Naive Bayes モデルでは，$P(s_k|c_j)$ の推定を式 (4.4) のように仮定して近似しているため，この仮定をできる限り満たす素性を選択することは非常に重要なことである．本研究で使用する素性を以下に挙げる．

- 対象語の直前の単語，２つ前の単語 (w_{-1}, w_{-2})

  直前の単語は，より強く対象語の語義を特徴付ける単語となっている場合が多いので，特別なシンボルを付加して他の素性と区別する．単語の前に，w_{-1} には $< PW1 >$，w_{-2} には $< PW2 >$ というシンボルを付ける．なお，直前の単語及び２つ前の単語は自立語，機能語問わず全て素性に加える．

- 対象語の直後の単語，２つ後の単語 (w_{+1}, w_{+2})

  直後の単語も直前の単語と同様である．w_{+1} には $< SW1 >$，w_{+2} には $< SW2 >$ というシンボルを付ける．

- 対象語の直前の単語の品詞，２つ前の単語の品詞 (p_{-1}, p_{-2})

  直前の単語の品詞についても，品詞の種類を問わず素性に加える．p_{-1} には $< PP1 >$，p_{-2} には $< PP2 >$ というシンボルを付ける．

- 対象語の直後の単語の品詞，２つ後の単語の品詞 (p_{+1}, p_{+2})

  直後の単語の品詞も直前の単語の品詞と同様である．p_{+1} には $< SP1 >$，p_{+2} には $< SP2 >$ というシンボルを付ける．

- 対象語の直前の 10 単語内にある自立語

  対象語からの距離が遠くなるほど，曖昧性の解消に有効な素性ではなくなるので，直前の 10 単語の範囲内にある自立語のみを素性とする．

- 対象語の直後の 10 単語内にある自立語

  対象語の直前の 10 単語内にある自立語を素性とする場合と同様である．

- 対象語が文頭，文末に出現したときの特別なシンボル

  対象語が文頭，文末が出現していた場合，< 文頭 >，< 文末 > というシンボルを素性に加える．

なお，数字は自立語ではないが，語義判別に有効である場合が多いため素性に加える．例えば,「間」という単語は「期間」などの時間的なものを表す場合と人などの「関係」を表す場合がある．このとき，数字は語義判別の有効な素性となり得る．数字は $< NUM >$ というシンボルで表現し，複数の数字が連続して出現しても１つのシンボルとする．文の

単語への分割は形態素解析ソフトの Chasen により行い，単語は基本形，品詞は Chasen の解析結果をそのまま用いる．例えば,「電話」という単語を含む,

カメラ付き携帯**電話**や高機能家電など，日本発の「世界のヒット商品」も目立つ．

という文は，以下のように単語に分割される．

カメラ / 付き / 携帯 / **電話** / や / 高 / 機能 / 家電 / など / 日本 / 発 / の / 「 / 世界 / の / ヒット / 商品 / 」 / も / 目立つ / ． /

そして，曖昧性解消に用いる素性は，以下のように設定される．

- w_{-1}, w_{-2}：$< PW1 >$ 携帯，$< PW2 >$ 付き
- w_{+1}, w_{+2}：$< SW1 >$ や，$< SW2 >$ 高
- p_{-1}, p_{-2}：$< PP1 >$ 名詞-サ変接続，$< PP2 >$ 名詞-接尾-一般
- p_{+1}, p_{+2}：$< SP1 >$ 助詞-並立助詞，$< SP2 >$ 接頭詞-名詞接続
- 対象語の直前の 10 単語内にある自立語：カメラ
- 対象語の直後の 10 単語内にある自立語：機能，家電，日本，発

このように，素性は対象語の前後 10 単語の範囲から抽出される．上記の例文中の枠線は素性が取り出される範囲を図示したものである．

4.3 EM アルゴリズムによるパラメータ推定

本節では，EM アルゴリズムによるパラメータの学習方法について詳しく述べる．学習アルゴリズムとして EM アルゴリズムを適用するにあたり，はじめに推定するパラメータの初期値を設定しなければならない．本研究では，Naive Bayes モデルを用いて語義を判別するため，$P(s_k)$，$P(f_i|s_k)$ の初期値を設定する．初期値の設定方法の詳細は 4.4 節で詳しく述べる．EM アルゴリズムは，与えられたパラメータをもとに個々の文脈を表しているモデルのもとでのコーパスの対数尤度を求め，前回の対数尤度と比較しながらその差が収束するまで E-step，M-step を繰り返し，パラメータの値を更新していく．本研究では，語義判別に式 (4.5) の Naive Bayes モデルを用い，そのモデルのもとでのコーパスの対数尤度は式 (4.6) で表される．

$$\log \prod_{j=1}^{J} \sum_{k=1}^{K} P(s_k)P(c_j|s_k) = \sum_{j=1}^{J} \log \sum_{k=1}^{K} P(s_k)P(c_j|s_k) \tag{4.6}$$

この式で，J は学習データ中の文の数，K は語義の数，$P(s_k)$ は語義の出現確率を表す．

- **E-step**

  このステップでは，与えられたパラメータより文脈が c_j のとき語義が s_k である条件付き確率を式 (4.7) で求める．

$$\begin{aligned} P(s_k|c_j) &= \frac{P(s_k)P(c_j|s_k)}{P(c_j)} \\ &= \frac{P(s_k)P(c_j|s_k)}{\sum_{k=1}^{K} P(s_k)P(c_j|s_k)} \end{aligned} \tag{4.7}$$

  右辺の $P(c_j|s_k)$ を直接求めるのは現実的に困難なので，以下の仮定を導入する．

$$P(c_j|s_k) = \prod_{f_i \in c_j} P(f_i|s_k) \tag{4.8}$$

  $P(f_i|s_k)$ は，語義が s_k であるとき素性 f_i が生起する確率である．式 (4.8) では個々の素性 f_i は互いに独立であると仮定した近似を行っている．

- **M-step**

  このステップでは，E-step で求めた値をもとにパラメータの値を再推定する．式 (4.9)，式 (4.10) の分母は，$\sum_{i=1}^{I} P(f_i|s_k) = 1$，$\sum_{k=1}^{K} P(s_k) = 1$ となるように正規化を行っている．

$$P(f_i|s_k) = \frac{\sum_{\{c_j : f_i \in c_j\}} P(s_k|c_j)}{\sum_{f=1}^{F} \sum_{\{c_j : f_i \in c_j\}} P(s_k|c_j)} \tag{4.9}$$

$$P(s_k) = \frac{\sum_{j=1}^{J} P(s_k|c_j)}{\sum_{k=1}^{K} \sum_{j=1}^{J} P(s_k|c_j)} \tag{4.10}$$

  E-step，M-step を繰り返しパラメータを更新していくが，更新時にパラメータの値が 0 になることを防ぐために，もしパラメータの値が 0 になってしまった場合，9.88131291682493e-324 という極小さい値を与え 0 とならないように工夫した．また，EM アルゴリズムが収束したと判断する条件は，繰り返し回数が 10 回に到達するか，パラメータ更新前と更新後の式 (4.6) の値が 1 以下となったら収束したと判断し処理を終了するようにした．収束したときのパラメータの値により式 (4.11) の Naive Bayes モデルで語義判別を行い，値が最大となる s_k をシステムが判別した語義とする．式 (4.11) で対数をとっているが，$P(s_k)\prod_{f_i \in c_j} P(f_i|s_k)$ を直接計算すると桁数が多くなり計算が複雑化して時間がかかり，また，値が極端に小さくなってしまい 0 になるということが起こりうる．そこで，対数をとることにより桁数が少なくなり，さらに乗算を加算で表現できることから計算を効率化できる．また，パラメータの値が 0 になることも少なくなるという効果も得られる．

$$\begin{aligned}
\arg\max_{s_k} P(s_k|c_j) &= \arg\max_{s_k} P(s_k)P(c_j|s_k) \\
&= \arg\max_{s_k} P(s_k) \prod_{f_i \in c_j} P(f_i|s_k) \\
&= \arg\max_{s_k} [\log P(s_k) + \sum_{f_i \in c_j} \log P(f_i|s_k)] \qquad (4.11)
\end{aligned}$$

4.4 初期パラメータの設定

EM アルゴリズムを用いるためには，はじめにパラメータ $P(s_k)$，$P(f_i|s_k)$ に適当な初期値を与えなければならない．EM アルゴリズムの学習精度は初期値の与え方に大きく依存しているため，この初期値の与え方が非常に重要になる．そこで，3つの初期パラメータを設定する手法を提案する．本節では，この3つの手法について述べる．

4.4.1 ランダムな初期値設定

EM アルゴリズムは，一般的にパラメータの初期値を全てランダムに与える．また，WSD の正解率を求める際には，ランダムに初期値を与えて学習する試行を何回か繰り返し，正解率の平均を計算するのが一般的である．

予備実験で，初期パラメータ $P(s_k)$，$P(f_i|s_k)$ を全てランダムに与えて学習を行い語義を判別した．しかし，初期値の与え方によって正解率のばらつきが大きいことがわかった．また，全般に正解率が低かった．この理由として，初期パラメータをランダムに決める手法では，事前に与えられるのは語義の数だけであり，既存の語義がどうのような素性とよく共起するかといった情報は全く与えられていないことが原因と考えられる．すなわち，既存の語義とそれがよく現れる文脈との関係に関する情報が事前に与えられていないため，既存の語義が別の語義に対応付けられたり，未定義語義が既存の語義のどれか1つに対応付けられるように学習されたことが正解率の低下を招いたと予想できる．

4.4.2 最尤推定による初期値設定

語義タグ付きコーパスから辞書に定義されている語義に関するパラメータを最尤推定し，未定義語義に関するパラメータには一様の値を与える手法を提案する．まず，語義の出現確率 $P(s_k)$ の初期パラメータを式 (4.12)，(4.13) のように設定する．

$$P(s_k) = \frac{O(s_k)}{\sum_{k=1}^{K} O(s_k)} \times (1 - P(< \text{未定義} >)) \qquad (4.12)$$

$$P(< \text{未定義} >) = 0.1 \qquad (4.13)$$

$O(s_k)$ は，辞書に定義されている語義の出現頻度である．未定義語義の出現確率 $P(<$ 未定義 $>)$ は未知であるので，定数として 0.1 を与えた．なお，式 (4.12) において $(1 - P(<$ 未

定義 >)) をかけているのは，$\sum_{k=1}^{K+1} P(s_k) = 1$ という制約を満たすためである．ただし，s_{K+1} は未定義語義を表す．

次に，$P(f_i|s_k)$ の初期パラメータを式 (4.14)，(4.15) のように設定する．

$$P(f_i|s_k) = \frac{O(f_i, s_k) + \alpha}{\sum_{k=1}^{K} O(s_k) + |F| \times \alpha} \tag{4.14}$$

$$P(f_i| < \text{未定義} >) = \frac{1}{|F|} \tag{4.15}$$

$O(f_i, s_k)$ は素性 f_i と語義 s_k の共起頻度，$|F|$ は素性の異なり数を表す．$P(f_i|s_k)$ は最尤推定により設定するが，全ての $O(f_i, s_k)$ に α という値を足してスムージングを行っている．ここでは $\alpha = 1$ としている．

この手法で確率モデルの学習を行う予備実験の結果，EM アルゴリズムの反復推定により最頻出語義に関するパラメータが高く学習されることが分かった．最頻出語義に対する $P(s_k)$ の値が 0.99 を超える場合も多かった．よって，ほとんどの事例に対して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く，常に最頻出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．

4.4.3 共起性の強い素性に高い値を与える初期値設定

プレーンテキストコーパスから学習されたモデルの精度を少量の語義タグ付きコーパスを用いて高める手法は，新納らの先行研究で良い成果を収めている．しかし，4.4.2 項で述べた語義タグ付きコーパスを用いて初期パラメータを最尤推定する手法はうまくいかなかった．よって，本研究では，対象語の各語義とよく共起する素性とそれ以外の素性にそれぞれ異なった一様の確率を初期値として与え，語義タグ付きコーパスから得られる情報を緩やかな制約として初期パラメータに反映させる手法を提案する．

辞書に定義されている語義に関する初期値設定

まず，語義の出現確率 $P(s_k)$ の設定であるが，これは語義タグ付きコーパスから最尤推定により式 (4.12) のように求める．

次に，辞書に定義されている語義 s_k のもとで素性 f_i が起こる条件付き確率 $P(f_i|s_k)$ の初期値設定について述べる．まず，語義タグ付きコーパスから $P(f_i|s_k)$ を式 (4.16) より求める．

$$p(f_i|s_k) = \frac{O(f_i, s_k)}{O(s_k)} \tag{4.16}$$

そして，この確率が上位の素性を各語義から n 個ずつ抽出する．その上位 n 個の素性に関する $P(f_i|s_k)$ を α，その他の素性に関する $P(f_i|s_k)$ を β とし，α に β の r 倍の値を与えるように $P(f_i|s_k)$ を設定する．

$$\alpha = r\beta \tag{4.17}$$

ここで，$\sum_{i=1}^{I} P(f_i|s_k) = 1$ なので，α と β には以下の関係が成り立つ．

$$n\alpha + (N-n)\beta = 1 \tag{4.18}$$

N は素性の異なり数を表す．この式に式 (4.17) を代入すると，

$$nr\beta + (N-n)\beta = 1 \tag{4.19}$$

$$(N-n+nr)\beta = 1 \tag{4.20}$$

となり，以下の式が導き出される．

$$\alpha = \frac{r}{N+(r-1)n} \tag{4.21}$$

$$\beta = \frac{1}{N+(r-1)n} \tag{4.22}$$

上位 n 個の素性集合を F_{s_k}，それ以外の素性集合を $\bar{F_{s_k}}$ とすると，最終的に $P(f_i|s_k)$ は式 (4.22)，(4.23) のように設定される．

$$P(f_i|s_k) = \frac{r}{N+(r-1)n} \qquad f_i \in F_{s_k} \tag{4.23}$$

$$P(f_i|s_k) = \frac{1}{N+(r-1)n} \qquad f_i \in \bar{F_{s_k}} \tag{4.24}$$

辞書に未定義の語義に関する初期値設定

ここでは，未定義語義に関する確率の初期値設定について述べる．語義タグ付きコーパスには，単語の語義が未定義であることを示した情報はないため，未定義語義に関するパラメータの初期値は，語義タグ付きコーパスから求めることはできない．そこで，辞書に定義されている語義と素性の条件付き確率の上位の素性は，その語義を特徴付けるものであり，未定義語義を判別するために有効な素性ではないと考えられる．そのため，そのような素性の $P(f_i| <$ 未定義 $>)$ については低い確率を与え，その他の素性については未定義語義の判別に有効である可能性があると考えて高い確率を与える．

まず，辞書に定義されている語義と素性の条件付き確率の上位 m 個の素性を抽出する．ここでは，語義ごとに上位 m 個取り出すのではなく，全ての語義の中での上位 m 個を取り出す．抽出した素性を f_i'，それ以外の素性を f_i'' とし，素性の異なり数を $|F|$ とすると，$P(f_i'| <$ 未定義 $>)$ と $P(f_i''| <$ 未定義 $>)$ は以下の式で求める．

$$P(f_i'| < \text{未定義} >) = \frac{1}{m + m(|F| - m)} \tag{4.25}$$

$$P(f_i''| < \text{未定義} >) = \frac{m}{m + m(|F| - m)} \tag{4.26}$$

本研究では，式 (4.25)，(4.26) により未定義語義に関する初期パラメータを求めたが，この式は適切ではない．この式は，m 個以外の素性の $P(f_i''| <$ 未定義 $>)$ に $P(f_i'| <$ 未定義 $>)$ の m 倍の確率を与えているので，m の値を変えると，低い値に設定する素性の数と，初期パラメータの低い値と高い値の比率の両方が同時に変化する．本来は，式 (4.23)，(4.24) と同様に，低い値を設定する素性の数 m と，$P(f_i'| <$ 未定義 $>)$ と $P(f_i''| <$ 未定義 $>)$ の比率 x は独立に調整できるようにするべきである．このときの初期パラメータの設定は，式 (4.27)，(4.28) のようになる．

$$P(f_i'| < \text{未定義} >) = \frac{1}{m + x(|F| - m)} \tag{4.27}$$

$$P(f_i''| < \text{未定義} >) = \frac{x}{m + x(|F| - m)} \tag{4.28}$$

５章の実験では，式 (4.25)，(4.26) で初期パラメータの設定を行った結果を報告する．m，x の値を独立に調整する実験は今後の課題とする．

第5章　実験

本章では，曖昧性解消に使用する学習データとテストデータ，未定義語義を含む単語，前章で述べた提案手法と比較する異なる 3 つの手法について述べ，実験を行った結果を示す．

5.1　実験方法

5.1.1　学習・評価に用いるデータ

曖昧性解消に用いる学習データとして語義タグ付きコーパスの EDR コーパス，プレーンテキストコーパスの毎日新聞の 91〜02 年まで記事 12 年分を使用した．曖昧性を解消する単語と各単語の学習データに含まれるデータ量を表を表 5.1 に示す．テストデータは，毎日新聞 03 年の記事から単語ごとに 100 文を抽出し，人手で正解語義を付加したものをテストデータとした．語義曖昧性解消をする未定義語義を含む単語の語義について以下に説明する．語義の説明は EDR 概念辞書からの抜粋である．

- 気持ち

  s_1: ほんの少し (a little bit)
  s_2: 体の状態についての感じ (physical condition)
  s_3: 心のうちに抱いている思い (a feeling in one's mind)
  < 未定義 >: 具体的ではっきりした気持ち．
  例文:当分は戦争を嫌う***気持ち***が起ろうから、その間に正しい教育をしなくてはならぬ．

- 教える

  s_1: 物事の道理などを教える (to preach reason)
  s_2: 物事を教授する (to teach a study or a technique)
  < 未定義 >: ただ相手に伝える．
  例文:届けてくれた人の名字と携帯の番号だけ***教え***てもらった．

- 決める

  s_1: (相撲で) 相手の動きがとれないようにする ((in Sumo) to prevent one's opponent from moving)

s_2: (衣服の色柄や着こなしなどを) 型どおり格好よくする (to make something have the proper form)
s_3: ものごとを決める (to make a decision on matters)
s_4: しっかりと考えを決める (to determine)
s_5: 物事をはっきり決める (to decide definitely)
< 未定義 >: 得点する.
例文:右足でポスト際に狙い通りのシュートを***決め***た.

- 情報

  s_1: 電子計算機への入出力のためのデータ (a coded unit or multiple units of data compiled by or entered into a computer)
  s_2: 判断に必要な知識や資料 (knowledge and reference materials necessary in order to make a decision)
  s_3: 事情についての知らせ (knowledge of something)
  < 未定義 >: コンピュータサイエンス.
  例文:***情報***教育、専門教育などのカリキュラムを充実させて成果を上げている.

- 世紀

  s_1: 歴史上の区分 (an era in history)
  s_2: キリスト生誕から 100 年を一期として数える時代区分 (one of the 100-year periods counted forward or backward from the supposed year of Jesus Christ's birth)
  < 未定義 >: 100 年.
  例文:この半***世紀***以上、日本人は自由ということの意味を取り違えてきた.

- 朝

  s_1: 夜明けから正午までの間 (the time before the noon)
  s_2: 夜が開けて間もない頃 (morning time)
  < 未定義 >: 北朝鮮という国.
  例文:米***朝***不可侵条約の締結に応じることを事実上の条件として提示している.

- 電話

  s_1: 電話機を使った通話 (communication by telephone)
  s_2: 電話機 (a telephone set)
  < 未定義 >: 電話番号.
  例文:全国の***電話***帳から２０００世帯の***電話***番号を無作為に選び

- 非

  s_1: 〜ではない (not being something)
  s_2: 間違い (a mistake)

s_3: 欠点 (a defect)
< 未定義 >: 反対する.
例文:米国を***非***とするもの７８%、是とするもの１４%ということになる.

- 与える

  s_1: 物をあてがう (to supply goods)
  s_2: (自分の物を) 他人に渡してその人の物とする (to hand over (one's things) to someone else)
  s_3: (損害を) こうむらせる (to cause a person to suffer harm)
  < 未定義 >: 夢，影響 (良い影響) などを与える.
  例文:子供たちに夢を***与え***てくれたのは野球とそして漫画だろう

- 条件

  s_1: 物事に関して限定する事柄 (a condition limiting something)
  s_2: 物事に関して限定する事柄 (something that limits or restricts a situation)
  < 未定義 >: 単に状態を表す.
  例文:雪という厳しい気象***条件***の中で、足への負担が最も大きい山下り６区を力走した.

未定義語義を含む単語の選別は以下のように行った．まず，EDR コーパスに出現する頻度と語義数を示した単語リストを作成し，その中で語義数が 2〜5 個である単語に候補を絞った．これは，語義数が多くても語義を適切に判別できることが望ましいが，実際 WSD は語義数が多くなるほど判別が難しくなる．本研究において未定義語義の判別を行う最初のステップとして，比較的語義曖昧性が容易な単語について評価を行うため，語義数が 2〜5 個の単語を選択することにした．そして，頻度が上位の単語を含むの例文を人手で調べ，未定義語義を持つ単語を探した．未定義語義を持つか否かを調べた単語のリストを図 5.1 に示す.

調べた単語の総数は 176 単語であり，そのうち未定義語義を持つ単語は 10 単語であった．図 5.1 の太字の単語が未定義語義を持つ単語である．このことから，未定義語義を持つ単語はそれほど多く存在するとは言えないが，今回はドメインを限定しない新聞記事を調査の対象としており，特定のドメインのテキストには未定義語義を持つ単語が多く出現すると予想される.

5.1.2 比較手法

未定義語義を考慮しない Naive Bayes モデルの教師あり学習

本研究の提案手法を評価するために，Naive Bayes モデルを用いた教師あり学習の手法と比較する．EDR コーパスから得られた出現頻度から最尤推定によりパラメータを求め，

関係	政府	用いる	首相	国	歳	進める	増える
計画	国民	当時	始める	違う	最近	家族	語る
決める	必要	含める	超える	政治	台	ソフト	残る
銀行	住む	技術	近く	土地	意見	調べる	環境
戦争	大手	対象	売る	出す	際	今後	かつて
車	**情報**	最後	同社	部屋	伴う	**朝**	部分
構造	聞く	従来	起こる	過去	安い	別	方向
産業	写真	当局	施設	**条件**	大統領	知識	**世紀**
仕事	妻	取り組む	学校	**気持ち**	歩く	学生	会う
文字	終わる	生まれる	感じる	全体	**与える**	見せる	政権
会	期	生徒	山	春	性質	さらに	内容
基本	家庭	大蔵省	会談	機関	単位	責任	会長
者	現れる	高まる	病気	果たす	場合	かなり	調べ
教える	**非**	サービス	半分	文	音楽	続ける	傾向
現地	社長	新しい	理論	現在	絵	新聞	設ける
教師	減る	課題	備える	初	学ぶ	入力	感
ファイル	態度	人事	当初	街	効果	待つ	勢力
ブーム	端末	古い	反	付近	立場	避ける	駅
時代	長官	抱える	見つける	地下	以前	最終	首脳
狙い	議論	位置	警察	差	死	陛下	借りる
センター	帰る	悪い	分かる	制限	伸びる	現実	総裁
紙	案	舞台	農業	要素	役割	着く	**電話**

図 5.1: 未定義語義を持つかを調べた単語のリスト

表 5.1: 学習データ

	語義	語義の分布	文書数		素性の異なり数	
			EDR	毎日新聞	EDR	毎日新聞
気持ち	s_1	0.0408016	392	46496	1666	34248
	s_2	0.0127551				
	s_3	0.9464285				
教える	s_1	0.1360759	316	24429	1474	25989
	s_2	0.8639240				
決める	s_1	0.5976821	1208	107594	3734	45968
	s_2	0.0107615				
	s_3	0.0504966				
	s_4	0.0016556				
	s_5	0.3394039				
情報	s_1	0.0022259	1797	167525	5269	58943
	s_2	0.2526432				
	s_3	0.7451307				
世紀	s_1	0.1666666	480	58985	1735	35994
	s_2	0.8333333				
朝	s_1	0.2301038	578	54610	2477	39676
	s_2	0.7698961				
電話	s_1	0.3010752	651	117894	2664	46774
	s_2	0.6989247				
非	s_1	0.0286738	279	35180	1628	25119
	s_2	0.9641577				
	s_3	0.0071684				
与える	s_1	0.3613744	844	52069	3104	35554
	s_2	0.3933649				
	s_3	0.2452606				
条件	s_1	0.2356215	539	39258	2443	28936
	s_2	0.7643784				

Naive Bayes モデル式 (4.5) の対数をとった形である式 (4.11) により語義を判別する．未定義語義を考慮しないこの手法を BL_1 とする．

未定義語義を考慮した Naive Bayes モデルの教師あり学習

BL_1 では，未定義語義を判別することはできず，常に辞書に定義された語義の 1 つを選択する．そこで，閾値を設定し，各語義について BL_1 の式 (4.11) により求められた値の最大値が閾値より低ければ未定義語義であると判別する．これは，全ての語義についてモデルの対数尤度が低いということは，その文中に出現する素性と辞書に定義されている語義の関係は薄いと考えられ，尤度が最大の語義であっても信頼性が低いと考えられるためである．各語義のモデルの尤度のうち 1 つでも閾値より高いものがあれば，その中で最大値となる s_k を選択する．この手法を BL_2 とする．

初期パラメータを一様に設定する手法

本研究の提案手法では，EM アルゴリズムのパラメータの初期値設定の際に，語義タグ付きコーパスの統計情報から求めた $P(f_i|s_k)$ の高い上位の素性とそれ以外の素性に関するパラメータ $P(f_i|s_k)$ にそれぞれ異なる一様の確率を与えている．提案手法の有効性を確認するため，全ての初期値を等しく設定して曖昧性解消を行った手法と比較を行った．すなわち，$P(f_i|s_k)$ の初期値は一様分布とし，辞書に定義されている語義の出現確率 $P(s_k)$ の初期値は語義タグ付きコーパスの統計情報を用いて式 (4.12) で求め，$P(<$ 未定義 $>)$ の初期値は未知であるので 0.1 の確率を与える手法を試した．この手法を EM_{uni} とする．

5.1.3 評価基準

4 章で述べた提案手法と，5.1.2 節で述べた手法を比較して評価を行う．評価を行う尺度であるが，本研究では，辞書に定義されている語義の正解率をあまり落とさずに未定義語義を判別したいため，辞書に定義されている語義と未定義語義の両方を合わせた全体の評価，未定義語義の評価，辞書に定義されている語義の評価が必要である．未定義語義の判別に関しては，F 値によって評価する．全体の正解率 Cor_{total}，未定義語義に対する F 値，辞書に定義されている語義の正解率 Cor_{s_k} は以下の式で算出する．

$$Cor_{total} = \frac{\text{辞書に定義されている語義の正解数} + \text{未定義語義の正解数}}{\text{対象語の数}} \tag{5.1}$$

$$F\text{ 値} = \frac{2PR}{P+R} \tag{5.2}$$

$$Cor_{s_k} = \frac{\text{辞書に定義されている語義の正解数}}{\text{辞書に定義されている語義にラベル付けされた対象語数}} \tag{5.3}$$

式 (5.2) における P と R は未定義語義判別の再現率と適合率であり，以下の式で求める．

$$\text{再現率}\,(recall) = \frac{\text{未定義語義の正解数}}{\text{未定義語義にラベル付けされた数}} \tag{5.4}$$

$$\text{適合率}\,(precision) = \frac{\text{未定義語義の正解数}}{\text{システムが未定義と判別した数}} \tag{5.5}$$

以上の評価尺度で提案手法と 5.1.2 項で挙げた手法を比較して評価する．

5.1.4 モデルパラメータの調整

本研究では，BL_2，提案手法においてモデルパラメータがあり，この値を変動させ実験を行う．本節では，モデルパラメータについて述べる．各手法のモデルパラメータを以下に示す．

- BL_2- 未定義語義であるか否かを判定する確率の閾値
- 提案手法-n:$P(f_i|s_k)$ の初期値を大きく設定する素性の数
  r:その比率
  m:$P(f_i| <$ 未定義 $>)$ の初期値を小さく設定する素性の数

ここで，r，n，m の値を大きくすることは，語義タグ付きコーパスの統計情報をより初期パラメータに反映させることを意味する．各手法のパラメータを変動させて語義判別を行い，F 値，Cor_{total} が最大となるパラメータを求める．そして，そのパラメータのときの正解率を比較する．

5.2 実験結果

5.2.1 Baseline モデルの実験結果

Baseline では，教師あり学習により Naive Bayes モデルを学習し，10 単語について語義判別を行った．未定義語義を考慮しない BL_1 の結果を表 5.2，未定義語義であるか否かを閾値により判別する BL_2 の結果を表 5.3 に示す．Sys は，システムが未定義であると判別した数である．

表 5.2: BL_1 の結果

	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
BL_1	0.562	0	0(0 / 258)	0(0 / 0)	0	0.76(562 / 742)

表 5.3: BL_2 の結果

閾値	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
-1000	0.268	968	0.972(251 / 258)	0.259(251 / 968)	0.409	0.022(17 / 742)
-1500	0.268	968	0.972(251 / 258)	0.259(251 / 968)	0.409	0.022(17 / 742)
-2000	0.286	879	0.875(226 / 258)	0.257(226 / 879)	0.397	0.080(60 / 742)
-2500	0.313	766	0.748(193 / 258)	0.251(193 / 765)	0.376	0.161(120 / 742)
-3000	0.312	765	0.744(192 / 258)	0.250(192 / 765)	0.375	0.161(120 / 742)
-3500	0.354	616	0.608(157 / 258)	0.254(157 / 616)	0.359	0.265(197 / 742)
-4000	0.406	457	0.492(127 / 258)	0.277(127 / 457)	0.355	0.376(279 / 742)
-4500	0.411	446	0.484(125 / 258)	0.280(125 / 446)	0.355	0.385(286 / 742)
-5000	0.444	326	0.364(94 / 258)	0.288(94 / 326)	0.321	0.471(350 / 742)
-5500	0.466	235	0.275(71 / 258)	0.302(71 / 235)	0.288	0.532(395 / 742)
-6000	0.465	207	0.224(58 / 258)	0.280(58 / 207)	0.249	0.548(407 / 742)
-6500	0.460	177	0.166(43 / 258)	0.242(43 / 177)	0.197	0.561(417 / 742)
-7000	0.457	135	0.104(27 / 258)	0.2(27 / 135)	0.137	0.579(430 / 742)
-7500	0.457	109	0.065(17 / 258)	0.155(17 / 109)	0.092	0.592(440 / 742)
-8000	0.457	109	0.065(17 / 258)	0.155(17 / 109)	0.092	0.592(440 / 742)

BL_1 では未定義を考慮しないため，Cor_{total} は 0.562，Cor_{s_k} は 0.76 という比較的高い正解率が得られた．BL_2 では，閾値を下げていくにつれてシステムが未定義と判別する数が減っていくため，通常の教師あり学習による Naive Bayes モデルの語義判別に近づく．よって，閾値を下げるにつれて Cor_{s_k} は高くなり，未定義語義の再現率，適合率，F 値は低くなっている．この結果の中で，Cor_{total} が最大であるのは，閾値を-5500 に設定したとき，F 値が最大であるのは，閾値が-1500 のときである．この 2 つの閾値のときの結果を手法の比較に用いる．

5.2.2 初期パラメータの値を一様にする手法の結果

EM アルゴリズムの初期パラメータである $P(s_k)$ については，語義タグ付きコーパスの統計情報から求め，$P(f_i|s_k)$ については， $1/|F|$($|F|$ は素性の異なり数) の一様の確率を与えて曖昧性解消を行った．なお，4.3 節で述べたように収束の判断は，繰り返し回数が 10 回に達するか式 (4.6) パラメータ更新前の値と更新後の値の差が 1 以下となったときに設定した．その結果を表 5.4 に示す．この手法では，全ての単語についてシステムが未定義であると判断した数が 0 であり，未定義語義について全く正解を得られなかった．

表 5.4: 初期パラメータの値を一様にする手法の結果

	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
気持ち	0.55	0	0(0 / 45)	0(0 / 0)	0	1(55 / 55)
教える	0.57	0	0(0 / 41)	0(0 / 0)	0	0.96(57 / 59)
決める	0.05	0	0(0 / 33)	0(0 / 0)	0	0.746(5 / 67)
情報	0.73	0	0(0 / 6)	0(0 / 0)	0	0.776(73 / 94)
世紀	0.74	0	0(0 / 21)	0(0 / 0)	0	0.936(74 / 79)
朝	0.07	0	0(0 / 51)	0(0 / 0)	0	0.142(7 / 49)
電話	0.33	0	0(0 / 12)	0(0 / 0)	0	0.375(33 / 88)
非	0.97	0	0(0 / 2)	0(0 / 0)	0	0.989(97 / 98)
与える	0.18	0	0(0 / 41)	0(0 / 0)	0	0.305(18 / 59)
条件	0.93	0	0(0 / 6)	0(0 / 0)	0	0.989(93 / 94)
Total	0.512	0	0(0 / 258)	0(0 / 0)	0	0.69(512 / 742)

5.2.3 提案手法の実験結果

4.4 節で説明した手法により EM アルゴリズムの初期パラメータを設定して曖昧性解消を行った．r, m, n の値を以下のように設定し，全ての組み合わせについて実験した結果を表 5.5，5.6，5.7 に示す．

$r = (5, 10, 15, 20)$

$n = (5, 10, 15, 20)$

$m = (0, 5, 10, 15, 20)$

$m = 0$ は，$P(f_i| <$ 未定義 $>)$ の初期値を全て等しく $1/|F|$ に設定することを表す．全体の正解率 Cor_{total} が最大となる (r, n, m) の組み合わせは，$(15, 20, 15)$ と $(20, 20, 15)$ で 0.485，F 値が最大となる組み合わせは $(5, 10, 0)$ と $(10, 5, 0)$ で 0.368 であった．よって，このパラメータのときの結果を他の手法との比較に用いる．

5.3 考察

5.3.1 提案手法の比較評価

本研究の提案手法，教師あり学習により Naive Bayes モデルを学習し未定義語義を考慮せずに語義を判別する手法 (BL_1)，Naive Bayes モデルに閾値を設けて未定義語義の判別を行う手法 (BL_2)，EM アルゴリズムの初期値を一様に与え Naive Bayes モデルを学習し語義判別を行う手法 (EM_{uni}) の結果を比較し，提案手法を評価する．提案手法は $EM_{rinjmk}(i$，j，k は r，n，m の設定値) と表記する．例えば，$EM_{r10n5m10}$ は $(r$，n，$m)$=(10，5，10) のときを表す．まず，それぞれの手法で Cor_{total} が最大であるものを比較する．その結果

表 5.5: 提案手法の結果 (1)

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
5	5	0	0.449	149	0.275(71 / 258)	0.476(71 / 149)	0.348	0.509(378 / 742)
5	10	0	0.472	106	0.259(67 / 258)	0.632(67 / 106)	0.368	0.545(405 / 742)
5	15	0	0.484	105	0.228(59 / 258)	0.561(59 / 105)	0.325	0.572(425 / 742)
5	20	0	0.456	103	0.197(51 / 258)	0.495(51 / 103)	0.282	0.545(405 / 742)
10	5	0	0.464	138	0.282(73 / 258)	0.528(73 / 138)	0.368	0.526(391 / 742)
10	10	0	0.461	91	0.224(58 / 258)	0.637(58 / 91)	0.332	0.543(403 / 742)
10	15	0	0.478	98	0.205(53 / 258)	0.540(53 / 98)	0.297	0.572(425 / 742)
10	20	0	0.484	65	0.186(48 / 258)	0.738(48 / 65)	0.297	0.587(436 / 742)
15	5	0	0.461	129	0.263(68 / 258)	0.527(68 / 129)	0.351	0.529(393 / 742)
15	10	0	0.459	84	0.205(53 / 258)	0.630(53 / 84)	0.309	0.547(406 / 742)
15	15	0	0.473	93	0.193(50 / 258)	0.537(50 / 93)	0.284	0.570(423 / 742)
15	20	0	0.482	59	0.170(44 / 258)	0.745(44 / 59)	0.277	0.590(438 / 742)
20	5	0	0.45	120	0.220(57 / 258)	0.475(57 / 120)	0.301	0.529(393 / 742)
20	10	0	0.459	76	0.197(51 / 258)	0.671(51 / 76)	0.305	0.549(408 / 742)
20	15	0	0.472	91	0.189(49 / 258)	0.538(49 / 91)	0.280	0.570(423 / 742)
20	20	0	0.483	59	0.170(44 / 258)	0.745(44 / 59)	0.277	0.591(439 / 742)
5	5	5	0.456	125	0.251(65 / 258)	0.52(65 / 125)	0.339	0.526(391 / 742)
5	10	5	0.469	95	0.244(63 / 258)	0.663(63 / 95)	0.356	0.547(406 / 742)
5	15	5	0.481	90	0.213(55 / 258)	0.611(55 / 90)	0.316	0.574(426 / 742)
5	20	5	0.482	64	0.186(48 / 258)	0.75(48 / 64)	0.298	0.584(434 / 742)
10	5	5	0.439	101	0.178(46 / 258)	0.455(46 / 101)	0.256	0.529(393 / 742)
10	10	5	0.457	81	0.201(52 / 258)	0.641(52 / 81)	0.306	0.545(405 / 742)
10	15	5	0.478	91	0.193(50 / 258)	0.549(50 / 91)	0.286	0.576(428 / 742)
10	20	5	0.477	58	0.170(44 / 258)	0.758(44 / 58)	0.278	0.583(433 / 742)
15	5	5	0.429	93	0.139(36 / 258)	0.387(36 / 93)	0.205	0.529(393 / 742)
15	10	5	0.461	76	0.193(50 / 258)	0.657(50 / 76)	0.299	0.553(411 / 742)
15	15	5	0.472	86	0.189(49 / 258)	0.569(49 / 86)	0.284	0.570(423 / 742)
15	20	5	0.48	56	0.162(42 / 258)	0.75(42 / 56)	0.267	0.590(438 / 742)
20	5	5	0.429	91	0.135(35 / 258)	0.384(35 / 91)	0.200	0.530(394 / 742)
20	10	5	0.458	69	0.186(48 / 258)	0.695(48 / 69)	0.293	0.552(410 / 742)
20	15	5	0.469	82	0.186(48 / 258)	0.585(48 / 82)	0.282	0.567(421 / 742)
20	20	5	0.48	54	0.158(41 / 258)	0.759(41 / 54)	0.262	0.591(439 / 742)

表 5.6: 提案手法の結果 (2)

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
5	5	10	0.455	89	0.224(58 / 258)	0.651(58 / 89)	0.334	0.535(397 / 742)
5	10	10	0.47	79	0.209(54 / 258)	0.683(54 / 79)	0.320	0.560(416 / 742)
5	15	10	0.482	80	0.201(52 / 258)	0.65(52 / 80)	0.307	0.579(430 / 742)
5	20	10	0.482	63	0.182(47 / 258)	0.746(47 / 63)	0.292	0.586(435 / 742)
10	5	10	0.449	81	0.201(52 / 258)	0.641(52 / 81)	0.306	0.535(397 / 742)
10	10	10	0.45	65	0.162(42 / 258)	0.646(42 / 65)	0.260	0.549(408 / 742)
10	15	10	0.474	76	0.170(44 / 258)	0.578(44 / 76)	0.263	0.579(430 / 742)
10	20	10	0.483	54	0.162(42 / 258)	0.777(42 / 54)	0.269	0.594(441 / 742)
15	5	10	0.453	81	0.205(53 / 258)	0.654(53 / 81)	0.312	0.539(400 / 742)
15	10	10	0.431	43	0.085(22 / 258)	0.511(22 / 43)	0.146	0.551(409 / 742)
15	15	10	0.468	72	0.166(43 / 258)	0.597(43 / 72)	0.260	0.572(425 / 742)
15	20	10	0.484	53	0.158(41 / 258)	0.773(41 / 53)	0.263	0.597(443 / 742)
20	5	10	0.455	81	0.209(54 / 258)	0.666(54 / 81)	0.318	0.540(401 / 742)
20	10	10	0.431	40	0.077(20 / 258)	0.5(20 / 40)	0.134	0.553(411 / 742)
20	15	10	0.467	70	0.162(42 / 258)	0.6(42 / 70)	0.256	0.572(425 / 742)
20	20	10	0.483	49	0.155(40 / 258)	0.816(40 / 49)	0.260	0.597(443 / 742)
5	5	15	0.446	86	0.193(50 / 258)	0.581(50 / 86)	0.290	0.533(396 / 742)
5	10	15	0.452	77	0.182(47 / 258)	0.610(47 / 77)	0.280	0.545(405 / 742)
5	15	15	0.48	80	0.186(48 / 258)	0.6(48 / 80)	0.284	0.582(432 / 742)
5	20	15	0.481	56	0.166(43 / 258)	0.767(43 / 56)	0.273	0.590(438 / 742)
10	5	15	0.45	82	0.182(47 / 258)	0.573(47 / 82)	0.276	0.543(403 / 742)
10	10	15	0.444	71	0.166(43 / 258)	0.605(43 / 71)	0.261	0.540(401 / 742)
10	15	15	0.471	68	0.158(41 / 258)	0.602(41 / 68)	0.251	0.579(430 / 742)
10	20	15	0.482	53	0.158(41 / 258)	0.773(41 / 53)	0.263	0.594(441 / 742)
15	5	15	0.446	80	0.174(45 / 258)	0.562(45 / 80)	0.266	0.540(401 / 742)
15	10	15	0.443	70	0.162(42 / 258)	0.6(42 / 70)	0.256	0.540(401 / 742)
15	15	15	0.467	65	0.155(40 / 258)	0.615(40 / 65)	0.247	0.575(427 / 742)
15	20	15	0.485	51	0.155(40 / 258)	0.784(40 / 51)	0.258	0.599(445 / 742)
20	5	15	0.447	81	0.178(46 / 258)	0.567(46 / 81)	0.271	0.540(401 / 742)
20	10	15	0.439	68	0.158(41 / 258)	0.602(41 / 68)	0.251	0.536(398 / 742)
20	15	15	0.465	65	0.155(40 / 258)	0.615(40 / 65)	0.247	0.572(425 / 742)
20	20	15	0.485	49	0.155(40 / 258)	0.816(40 / 49)	0.260	0.599(445 / 742)

表 5.7: 提案手法の結果 (3)

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
5	5	20	0.441	55	0.158(41 / 258)	0.745(41 / 55)	0.261	0.539(400 / 742)
5	10	20	0.428	39	0.100(26 / 258)	0.666(26 / 39)	0.175	0.541(402 / 742)
5	15	20	0.464	39	0.100(26 / 258)	0.666(26 / 39)	0.175	0.590(438 / 742)
5	20	20	0.477	52	0.155(40 / 258)	0.769(40 / 52)	0.258	0.588(437 / 742)
10	5	20	0.443	54	0.151(39 / 258)	0.722(39 / 54)	0.25	0.544(404 / 742)
10	10	20	0.422	37	0.089(23 / 258)	0.621(23 / 37)	0.155	0.537(399 / 742)
10	15	20	0.456	35	0.081(21 / 258)	0.6(21 / 35)	0.143	0.586(435 / 742)
10	20	20	0.479	45	0.139(36 / 258)	0.8(36 / 45)	0.237	0.597(443 / 742)
15	5	20	0.441	53	0.143(37 / 258)	0.698(37 / 53)	0.237	0.544(404 / 742)
15	10	20	0.421	36	0.085(22 / 258)	0.611(22 / 36)	0.149	0.537(399 / 742)
15	15	20	0.446	33	0.073(19 / 258)	0.575(19 / 33)	0.130	0.575(427 / 742)
15	20	20	0.481	44	0.135(35 / 258)	0.795(35 / 44)	0.231	0.601(446 / 742)
20	5	20	0.438	53	0.143(37 / 258)	0.698(37 / 53)	0.237	0.540(401 / 742)
20	10	20	0.421	36	0.085(22 / 258)	0.611(22 / 36)	0.149	0.537(399 / 742)
20	15	20	0.444	33	0.073(19 / 258)	0.575(19 / 33)	0.130	0.572(425 / 742)
20	20	20	0.481	40	0.127(33 / 258)	0.825(33 / 40)	0.221	0.603(448 / 742)

表 5.8: 提案手法との比較 (Cor_{total} について)

	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
BL_1	0.562	0	0(0 / 258)	0(0 / 0)	0	0.76(562 / 742)
BL_2	0.466	235	0.275(71 / 258)	0.302(71 / 235)	0.288	0.532(395 / 742)
EM_{uni}	0.512	0	0(0 / 258)	0(0 / 0)	0	0.69(512 / 742)
$EM_{r15n20m15}$	0.485	51	0.155(40 / 258)	0.784(40 / 51)	0.258	0.599(445 / 742)
$EM_{r20n20m15}$	0.485	49	0.155(40 / 258)	0.816(40 / 49)	0.260	0.599(445 / 742)

を表 5.8 に示す．提案手法は，BL_1 より Cor_{total}，Cor_{s_k} 共に低くなっている．BL_2 と比較すると，Cor_{total}，precision，Cor_{s_k} で BL_2 より高くなっている．EM_{uni} と比較すると，Cor_{total}，Cor_{s_k} ともに劣っているが，EM_{uni} は，未定義語義を全く判別できていないので，未定義語義の判別に関しては勝っている．次にそれぞれの手法で F 値が最大であるものを比較する．結果を表 5.9 に示す．ここでも，BL_1，BL_2 との比較においては，Cor_{total} が最大のときと同様の結果が得られた．本研究の提案手法は，未定義語義の再現率が低いことから，未定義語義を辞書に定義されている語義として誤って判定することが多いことがわかる．しかし，Cor_{total} が最大のとき，F 値が最大のとき共に適合率が高く，未定義語義であると判別する数は少ないが，未定義語義と判別したものに対しては正解となっているものが多いことがわかる．一方，BL_2 は，未定義語義を多く判別することで，辞書に定義されている語義の正解率 Cor_{s_k} が大きく低下し，全体の正解率 Cor_{total} も提案手法より大きく劣る．このことから，提案手法は BL_2 より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，全体の正解率が最も高いのは BL_1 であり，未定義語義を考慮せずに辞書に定義されている語義のみを選択する手法の方が，未定義語義の判別は常に誤るが全体の正解率は高くなることがわかる．そこで，まず,「辞書に定義されているか」,「未定義」であるかの 2 つのクラスで判別し,「辞書に定義されている」と判別されたものに対して精度がよい教師あり学習を行い語義を判別する 2 段階の手法を検討している．この手法により，全体の正解率の向上が期待できる．

5.3.2 提案手法の単語別評価

ここでは，提案手法を単語別に見た場合の評価をする．Cor_{total} が最大のときの単語別の結果を表 5.10，F 値が最大のときの単語別の結果を表 5.11 に示す．Cor_{total} が最大のとき，F 値が最大のとき共に「朝」,「電話」という単語について再現率，適合率，F 値で良い結果を得ている．Cor_{total} が最大のときに関しては，適合率は 100%である．この理由として，辞書に定義されている語義と未定義語義の違いが明確であることが挙げられる.「朝」という単語は，EDR 概念辞書に定義されている語義は,「夜が明けて間もない頃」,「夜明

表 5.9: 提案手法との比較 (F 値について)

	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
BL_1	0.562	0	0(0 / 258)	0(0 / 0)	0	0.76(562 / 742)
BL_2	0.268	968	0.972(251 / 258)	0.259(251 / 968)	0.409	0.022(17 / 742)
EM_{uni}	0.512	0	0(0 / 258)	0(0 / 0)	0	0.69(512 / 742)
$EM_{r5n10m0}$	0.472	106	0.259(67 / 258)	0.632(67 / 106)	0.368	0.545(405 / 742)
$EM_{r10n5m0}$	0.464	138	0.282(73 / 258)	0.528(73 / 138)	0.368	0.526(391 / 742)

けから正午までの間」，未定義の語義は「北朝鮮という国」である．未定義語義については，他国の名前と共起することが多く，辞書に定義されている語義とはほとんど共起しない．また，辞書に定義されている語義の正解率は低くなっているが，これは,「夜が明けて間もない頃」と「夜が明けてから正午までの間」の語義は，語義を特徴付ける単語が明確ではなく，対象語の前後の品詞にもあまり違いがないためと考えられる.「電話」という単語については，辞書に定義されている語義は,「電話機を使った通話」,「電話機」，未定義語義は「電話番号」である．未定義語義は，数字と共起することが多く，他の 2 つの語義とはあまり共起していなかったことが高い精度で判別できた理由であると考えられる．また，もう 1 つの理由として，学習データの毎日新聞記事に多く出現していたことも挙げられる．一方，学習データに多く出現していたにもかかわらず，上手く未定義語義を判別できなかった単語として「決める」がある．この単語の未定義語義は「得点する」であり，この未定義語義とよく共起する「ゴール」,「トライ」とは辞書に定義されている語義とはあまり共起しない．この単語は，辞書に定義されている語義に関しても極端に正解率が低いため，判別できない原因として辞書に定義されている語義数が多いことが考えられる．

5.3.3 r, n, m の値による結果の考察

4.4 節で述べた手法で EM アルゴリズムの初期値を設定し実験を行ったが，ここでは，そこで用いた r, n, m の個々の値の変化により結果がどのように変化するのかを全体の正解率が最大であったときと，F 値が最大になったときの (r, n, m) の値 $(15, 20, 15), (5, 10, 0)$ を基準に値を変えて考察する．

r 値の変化による結果の違い

m, n の値を固定し，r の値を変動させたときの結果を表 5.12 に示す．$n = 20, m = 15$ に固定して r の値を変動させたとき，r の値が増加するにつれ，Cor_{total}，適合率，Cor_{s_k} は上がり，Sys，再現率，F 値は下がる．しかし，上がったときと下がったときともに値の差はわずかであった．次に，$n = 10, m = 0$ に固定して r の値を変動させたとき，Cor_{total}，

表 5.10: 提案手法の各単語についての結果 (Cor_{total} が最大)

	r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
気持ち	15	20	15	0.38	2	0.044(2 / 45)	1(2 / 2)	0.085	0.654(36 / 55)
	20	20	15	0.39	2	0.044(2 / 45)	1(2 / 2)	0.085	0.672(37 / 55)
教える	15	20	15	0.5	1	0(0 / 41)	0(0 / 1)	0	0.847(50 / 59)
	20	20	15	0.5	1	0(0 / 41)	0(0 / 1)	0	0.847(50 / 59)
決める	15	20	15	0.13	2	0(0 / 33)	0(0 / 2)	0	0.194(13 / 67)
	20	20	15	0.13	0	0(0 / 33)	0(0 / 0)	0	0.194(13 / 67)
情報	15	20	15	0.51	1	0(0 / 6)	0(0 / 1)	0	0.542(51 / 94)
	20	20	15	0.51	1	0(0 / 6)	0(0 / 1)	0	0.542(51 / 94)
世紀	15	20	15	0.44	3	0.047(1 / 21)	0.333(1 / 3)	0.083	0.544(43 / 79)
	20	20	15	0.44	3	0.047(1 / 21)	0.333(1 / 3)	0.083	0.544(43 / 79)
朝	15	20	15	0.48	29	0.568(29 / 51)	1(29 / 29)	0.725	0.387(19 / 49)
	20	20	15	0.47	29	0.568(29 / 51)	1(29 / 29)	0.725	0.367(18 / 49)
電話	15	20	15	0.63	8	0.666(8 / 12)	1(8 / 8)	0.8	0.625(55 / 88)
	20	20	15	0.62	8	0.666(8 / 12)	1(8 / 8)	0.8	0.613(54 / 88)
非	15	20	15	0.95	2	0(0 / 2)	0(0 / 2)	0	0.969(95 / 98)
	20	20	15	0.95	2	0(0 / 2)	0(0 / 2)	0	0.969(95 / 98)
与える	15	20	15	0.25	0	0(0 / 41)	0(0 / 0)	0	0.423(25 / 59)
	20	20	15	0.25	0	0(0 / 41)	0(0 / 0)	0	0.423(25 / 59)
条件	15	20	15	0.58	3	0(0 / 6)	0(0 / 3)	0	0.617(58 / 94)
	20	20	15	0.59	3	0(0 / 6)	0(0 / 3)	0	0.627(59 / 94)

表 5.11: 提案手法の各単語についての結果 (F 値が最大)

	r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
気持ち	5	10	0	0.48	11	0.2(9 / 45)	0.818(9 / 11)	0.321	0.709(39 / 55)
	10	5	0	0.45	13	0.244(11 / 45)	0.846(11 / 13)	0.379	0.618(34 / 55)
教える	5	10	0	0.5	6	0.073(3 / 41)	0.5(3 / 6)	0.127	0.796(47 / 59)
	10	5	0	0.39	9	0.073(3 / 41)	0.333(3 / 9)	0.12	0.610(36 / 59)
決める	5	10	0	0.07	5	0.030(1 / 33)	0.2(1 / 5)	0.052	0.089(6 / 67)
	10	5	0	0.07	8	0(0 / 33)	0(0 / 8)	0	0.104(7 / 67)
情報	5	10	0	0.43	2	0(0 / 6)	0(0 / 2)	0	0.457(43 / 94)
	10	5	0	0.43	6	0.333(2 / 6)	0.333(2 / 6)	0.333	0.436(41 / 94)
世紀	5	10	0	0.55	11	0.238(5 / 21)	0.454(5 / 11)	0.312	0.632(50 / 79)
	10	5	0	0.65	11	0.333(7 / 21)	0.636(7 / 11)	0.437	0.734(58 / 79)
朝	5	10	0	0.49	42	0.745(38 / 51)	0.904(38 / 42)	0.817	0.224(11 / 49)
	10	5	0	0.61	49	0.725(37 / 51)	0.755(37 / 49)	0.74	0.489(24 / 49)
電話	5	10	0	0.6	10	0.666(8 / 12)	0.8(8 / 10)	0.727	0.590(52 / 88)
	10	5	0	0.4	10	0.666(8 / 12)	0.8(8 / 10)	0.727	0.363(32 / 88)
非	5	10	0	0.9	4	0.5(1 / 2)	0.25(1 / 4)	0.333	0.908(89 / 98)
	10	5	0	0.95	4	0.5(1 / 2)	0.25(1 / 4)	0.333	0.959(94 / 98)
与える	5	10	0	0.18	1	0.024(1 / 41)	1(1 / 1)	0.047	0.288(17 / 59)
	10	5	0	0.23	2	0.024(1 / 41)	0.5(1 / 2)	0.046	0.372(22 / 59)
条件	5	10	0	0.52	14	0.166(1 / 6)	0.071(1 / 14)	0.1	0.542(51 / 94)
	10	5	0	0.46	26	0.5(3 / 6)	0.115(3 / 26)	0.187	0.457(43 / 94)

Sys，再現率，F 値は下がっていった．Cor_{s_k} は，$r = 10$ のとき下がっているが，$r = 15$, $r = 20$ と上がっているので，上がると考えて良いと思われる．$(n, m) = (20, 15), (10, 0)$ のときともに Cor_{s_k} は上がり，Sys が下がっているので，r の値を増加させると，システムが未定義であると判別する数は減るが，辞書に定義されている語義の正解率 Cor_{s_k} をわずかではあるが上げることができると考えられる．

表 5.12: r 値の変化による結果の違い

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
5	20	15	0.481	56	0.166(43 / 258)	0.767(43 / 56)	0.273	0.590(438 / 742)
10	20	15	0.482	53	0.158(41 / 258)	0.773(41 / 53)	0.263	0.594(441 / 742)
15	20	15	0.485	51	0.155(40 / 258)	0.784(40 / 51)	0.258	0.599(445 / 742)
20	20	15	0.485	49	0.155(40 / 258)	0.816(40 / 49)	0.260	0.599(445 / 742)
5	10	0	0.472	106	0.259(67 / 258)	0.632(67 / 106)	0.368	0.545(405 / 742)
10	10	0	0.461	91	0.224(58 / 258)	0.637(58 / 91)	0.332	0.543(403 / 742)
15	10	0	0.459	84	0.205(53 / 258)	0.630(53 / 84)	0.309	0.547(406 / 742)
20	10	0	0.459	76	0.197(51 / 258)	0.671(51 / 76)	0.305	0.549(408 / 742)

n 値の変化による結果の違い

ここでは，r，m の値を固定し，n の値を変動させたときの結果について考察する．その結果を表 5.12 に示す．$r = 15, m = 15$ に固定して n の値を増加させたとき，Cor_{total}，適合率，Cor_{s_k} は大きく上がり，Sys，再現率，F 値は下がっていった．次に，$r = 5, m = 0$ に固定したとき，Cor_{total}，Cor_{s_k} は $n = 15$ まで正解率が大きく上がったが，$n = 20$ で大きく下がっている．Sys，再現率，F 値は下がっている．$r = 15, m = 15$ のときは，Cor_{total}，Cor_{s_k} ともに n を増加させるに従い上がっていくのに対し，Cor_{total}，Cor_{s_k} が $n = 20$ で大きく下がっているが，n の値を増加させることにより，全体の正解率と辞書に定義されている語義の正解率を大きく上げることができると考えられる．

m 値の変化による結果の違い

m 値を変化させたときの結果を表 5.13 に示す．$r = 15, n = 20$ を固定して m の値を変化させていった場合，適合率，Cor_{s_k} が上がり，Sys，再現率，F 値が下がった．$r = 5$, $n = 10$ を固定した場合，Sys，再現率，F 値が下がっていくだけで，他ははっきりとした特徴が見られなかった．m は未定義語義に関するパラメータの初期値を設定する際に用いる値であるので m を増加させるにつれ，Sys，再現率，適合率，F 値のいづれかが上がると予想していたが逆に下がってしまった．ただし，これは，式 (4.25)，式 (4.26) の誤っ

表 5.13: n 値の変化による結果の違い

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
15	5	15	0.446	80	0.174(45 / 258)	0.562(45 / 80)	0.266	0.540(401 / 742)
15	10	15	0.443	70	0.162(42 / 258)	0.6(42 / 70)	0.256	0.540(401 / 742)
15	15	15	0.467	65	0.155(40 / 258)	0.615(40 / 65)	0.247	0.575(427 / 742)
15	20	15	0.485	51	0.155(40 / 258)	0.784(40 / 51)	0.258	0.599(445 / 742)
5	5	0	0.449	149	0.275(71 / 258)	0.476(71 / 149)	0.348	0.509(378 / 742)
5	10	0	0.472	106	0.259(67 / 258)	0.632(67 / 106)	0.368	0.545(405 / 742)
5	15	0	0.484	105	0.228(59 / 258)	0.561(59 / 105)	0.325	0.572(425 / 742)
5	20	0	0.456	103	0.197(51 / 258)	0.495(51 / 103)	0.282	0.545(405 / 742)

た式で初期値を設定した結果であり，正当な評価とは言えない．$P(f_i| <$ 未定義 $>)$ の初期パラメータについては式 (4.27)，式 (4.28) で初期値を設定し，実験し直す必要がある．

表 5.14: m 値の変化による結果の違い

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
15	20	0	0.482	59	0.170(44 / 258)	0.745(44 / 59)	0.277	0.590(438 / 742)
15	20	5	0.48	56	0.162(42 / 258)	0.75(42 / 56)	0.267	0.590(438 / 742)
15	20	10	0.484	53	0.158(41 / 258)	0.773(41 / 53)	0.263	0.597(443 / 742)
15	20	15	0.485	51	0.155(40 / 258)	0.784(40 / 51)	0.258	0.599(445 / 742)
15	20	20	0.481	44	0.135(35 / 258)	0.795(35 / 44)	0.231	0.601(446 / 742)
5	10	0	0.472	106	0.259(67 / 258)	0.632(67 / 106)	0.368	0.545(405 / 742)
5	10	5	0.469	95	0.244(63 / 258)	0.663(63 / 95)	0.356	0.547(406 / 742)
5	10	10	0.47	79	0.209(54 / 258)	0.683(54 / 79)	0.320	0.560(416 / 742)
5	10	15	0.452	77	0.182(47 / 258)	0.610(47 / 77)	0.280	0.545(405 / 742)
5	10	20	0.428	39	0.100(26 / 258)	0.666(26 / 39)	0.175	0.541(402 / 742)

r, n, m の 2 つの値を固定し，残り 1 つの値を変動させていった結果，r, n, m のどの値を上げていくにしても，未定義であると判別した数，再現率，F 値は下がり，全体の正解率，適合率，辞書に定義されている語義の正解率が上がる傾向がみられた．r, n, m の値を全て上げていった結果を表 5.15 に示す．表 5.15 より，r, n, m の値を上げていくと，全体の正解率，適合率，辞書に定義されている語義の正解率は上がり，未定義であると判別した数，再現率，F 値は下がる傾向にあることが分かる．

本研究の提案手法は，未定義語義であるとラベル付けされた文の数が 258 に対して，シ

ステムが未定義であると判別した数の最高値は 149 と少ないが，それが正解である数が多く適合率が良いシステムであると言える．

表 5.15: r,n,m の値を同時に上げていったときの結果

r	n	m	Cor_{total}	Sys	recall	precision	F 値	Cor_{s_k}
5	5	0	0.449	149	0.275(71 / 258)	0.476(71 / 149)	0.348	0.509(378 / 742)
10	10	5	0.457	81	0.201(52 / 258)	0.641(52 / 81)	0.306	0.545(405 / 742)
15	15	10	0.468	72	0.166(43 / 258)	0.597(43 / 72)	0.260	0.572(425 / 742)
20	20	15	0.485	49	0.155(40 / 258)	0.816(40 / 49)	0.260	0.599(445 / 742)
20	20	20	0.481	40	0.127(33 / 258)	0.825(33 / 40)	0.221	0.603(448 / 742)

第6章　おわりに

従来の語義曖昧性解消は，辞書に定義されている語義のいづれであるかを判別していたため，辞書に定義されていない意味で使用されている単語に対しては必ず間違った語義を付けてしまうという問題があった．本研究では，そのような問題に対処するために辞書に未定義の語義も判別の対象とする語義曖昧性解消システムを構築し，実験・評価を行った．最後に今後の課題を挙げる．

- 再現率の改善

  提案手法の結果では，未定義語義であると判別されたものに対しては正解であるものが多く，適合率は良かった．しかし，システムが未定義語義であると判別した単語数は，最も多いときで，未定義語義にラベル付けされた数 258 に対して 149 であり，ほとんどが 100 以下であった．したがって，提案手法の EM アルゴリズムの初期値設定法を改良するなどして，未定義語義の判別の再現率を向上させる必要がある．

- ２段階の語義曖昧性解消

  10 個の単語について曖昧性解消を行ったが，語義数が多い単語については極端に正解率が下がっていた．これを解決するために，まずはじめに「辞書に定義されている」か「未定義である」かの２値分類問題にし，そこで「辞書に定義されている」と判別されたものに対して，定義された語義の中から最適な語義を選択する．この際，後者の手法として，良い成果を挙げている教師あり学習の手法が適用できるため WSD の正解率の向上が期待できる．

- 正確な初期値設定による評価

  本研究では，EM アルゴリズムの初期値を設定する際に，誤った式 (4.25)，(4.26) により初期値を求めてしまった．正確な式 (4.27)，(4.28) により初期値を設定し実験し直す必要がある．

謝辞

本研究を進めるにあたり，丁寧な御指導，御教示を賜りました白井清昭助教授，島津明教授に心より深く感謝致します．また，多くの貴重な御意見を頂きました，山田寛康助手，中村誠助手，多くの討論をして頂きました自然言語処理講座の皆様に深く感謝致します．

参考文献

[1] Christopher D Manning, Hinrich Schutze. Foundations of statistical natural language processing, pp.252-256(1999).

[2] Hiroyuki Shinnou, Minoru Sasaki. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm, Seventh Conference on Natural Language Learning, pp.41-48(2003).

[3] 小川千隼, 白井清昭. 辞書定義文中の上位概念を用いた頑健な語義曖昧性解消, 北陸先端科学技術大学院大学 (2005).

[4] Fumiyo Fukumoto, Yoshimi Suzuki. Word Sense Disambiguation in Untagged Text based on Term Weight Learning, EACL'99:European Chapter of the Association for Computational Linguistics.

[5] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods, Proceeding on the Annual Meeting of the Association for Computational Linguistics, pp.189-196(1995).

[6] 新納浩幸 素性間の共起性を検査する Co-training による語義判別規則の学習, 情報処理学会自然言語処理研究会, 145-5, pp.29-36 2001.

[7] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Michell. Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2/3):103-134 1991.

[8] schutze, H. Automatic Word Sense Discrimination, Computational Linguistics,pp.97-124 1991.

[9] J.Ross Quinlan. Programs for Machine Learning, Morgan Kaufmann Publishers, 1993

[10] Ronald L.Rivest. Learning decision list, Machine Learning,Vol.2,pp.229-246, 1987

[11] Vladimir N. Vapnik. Statistical Learning Theory, A Wiley-Interscience Publication, 1998.

[12] Mitchell, T. Machine Learning, Mc-Graw Hill, 1997.