翻訳知識獲得のための多言語インターネットディレクトリィの自動統合

研究代表者　福　本　文　代　山梨大学大学院医学工学総合研究部・教授
共同研究者　鈴　木　良　弥　山梨大学大学院医学工学総合研究部・准教授

1 はじめに

本研究は，異なる言語で記述された複数のインターネットディレクトリィを統合することで得られる多言語関連文書から，機械翻訳システムに必要となる翻訳知識を自動的に獲得することを目的とする．機械翻訳は自然言語処理の成果の一つであり，コンピュータを介したユニバーサルコミュニケーションを実現するためのコア技術として注目されている．高品質な翻訳を生成するためには，対訳語に関する大量の語彙知識が必要である．近年，Web をはじめとする大規模データが手軽に入手できるようになったことを背景に，大規模データから互いに類似した内容を持つ多言語関連文書を自動的に抽出し，そこから対訳語を獲得する研究が盛んに行われている．この手法における対訳語の精度は，意味的に類似した多言語関連文書を高精度で抽出できるかに依存する．これまで統計手法や機械学習をはじめとする様々な手法が提案されているが，混沌としたWeb データが抽出対象であるために，いずれも質の高い対訳語を得るまでには至っていない．本研究はインターネットディレクトリィの階層構造が人手で構築されていることに注目し，これを利用することで質の高い大量の対訳語を獲得することを目指す．具体的には，日本語と英語で記述された2 つの階層構造を統合することで，互いに類似した内容を持つ文書対を抽出し，得られた文書対から対訳語を抽出する手法を提案する．

日本語インターネットディレクトリィの各分野と英語インターネットディレクトリィの各分野との対応付けを行うため，図 1 に示すように，各分野に属する英語文書（日本語文書）を辞書引きにより翻訳することで，日本語文書（英語文書）を作成し，これらの文書を日本語 ODP 階層（英語 ODP 階層）へ分類する．しかし，一般に辞書は，語義が複数登録されているため，単純な辞書引きによる方法では,多義の問題が生じる．また，辞書の語彙は限られているため未知語の問題，すなわち辞書に記載されていない語が存在するため，辞書引きができないという問題も生じる．そこで本研究では，グラフベースの半教師付きアルゴリズムを提案することで，多義語を含む未知語の意味推定を行う手法を提案する．

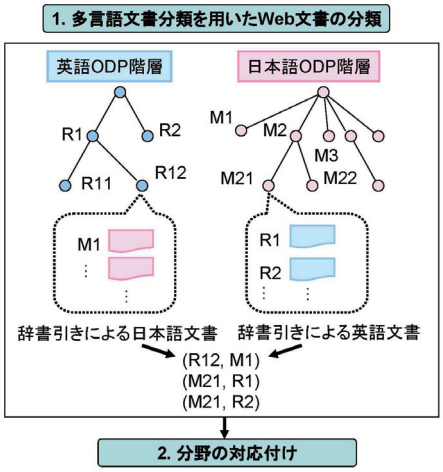

図1　インターネットディレクトリィの統合による翻訳知識の自動獲得

2 関連研究

機械翻訳において知的な知識処理を実現するため，自然言語の意味処理技術に関心が寄せられ，実際に利用されつつある．自然言語の意味処理を行うためには言語知識を大量にもつことが必要となるが，特に語彙知識は各種の意味処理において必須の知識源となる．統計的手法に基づく単語分類は，既存の辞書である語彙知識源に定義された意味クラスに分類するというタスクとしてこれまでに多くの手法が提案されている．語彙知識源として利用されている英語辞書の一つに　Levin の English Verb Classes と呼ばれる語彙辞書がある．Levin の語彙辞書は，文の構造に基づき語彙の分類体系である意味クラスが設定されているため，構文から意味へと処理する自然言語の機械処理との親和性が良い．従って文書データ中の単語を Levin の分類

体系に分類する手法が確立できれば，次々と創り出され使用される新語に対しても機械処理が可能となり，情報爆発時代に真に対処可能な意味処理のための知識源が構築可能となる．Schulte らは，iterative クラスタリングと unsupervised latent class analysis の２つのクラスタリングアルゴリズムを用いて動詞単語を Levin の意味クラスに分類する手法を提案している(Schulte, 2000)．Stevenson らは教師付き分類手法を提案している(Stevenson, 2003)．Brew らはスペクトラルクラスタリングを用いることでドイツ語動詞を Levin の意味クラスに分類手法を提案している(Brew, 2002)．

グラフベースのクラスタリング手法として，Widdows らは教師なしクラスタリング手法を提案した(Widdows, 2002)．グラフは単語をノードとし，単語間の関係を示すエッジには単語間の構文関係が用いられている．20 単語を WordNet で示されている 10 クラスに分類した結果，82%の精度が報告されている．松尾らは Web データから単語の出現回数を求めた統計値に基づき単語を分類する手法を提案している(Matsuo, 2006)．分類のためのクラスタリング手法として Newman 法を用いて分類を行った．しかし，これらの研究のいずれもハードクラスタリング手法を用いているため，単語が複数の語義を持つという多義性の問題に対処する枠組みにはなっていない．この多義性の問題を解決するためにソフトクラスタリング，すなわち入力要素が複数のクラスタに属することを許すクラスタリング手法を用いて単語の意味分類を行う手法も提案されている.Pereira らは分布クラスタリングと呼ばれる手法を用いて名詞単語を分類している(Pereira, 1993)．Schulte らは半教師付き EM アルゴリズムを用いて動詞単語を分類する手法を提案した．彼らの手法は，動詞の格フレーム情報と格を取る名詞の意味を WordNet から抽出することで動詞単語を分類した．本稿で提案するアルゴリズムは Schulte らと同様，半教師付きクラスタリング手法である．彼らとの違いは，多義を解消することができる点，及び未知語を既知の意味クラスに分類している点である．

Korhonen らは多義性に注目し，これを解消する手法を提案している(Korhonen, 2003)．しかし彼らの手法はクラスタリングを複数回適用し，その結果によりある単語が複数のクラスに分類されていることで多義性を判断しているため，分類対象となる単語を１回のクラスタリングで複数クラスに属するという方法にはなっていない．したがって精度面で課題が残されている．本研究は，多義性と未知語に対応可能な半教師付きクラスタリングを提案する．

3 単語間の類似度計算

3-1 格フレーム抽出

クラスタリングによる単語の意味分類では，クラスタリングの入力単語集合における各単語の類似度を求める必要がある．本研究では英語の動詞単語に注目し多義語を含む未知語の意味推定を行う．各単語の類似度を求めるために，動詞を動詞が取る格要素で表現する．各動詞は格要素を次元とするベクトルで表現される(Schulte im Walde, 2000; Brew and Walde, 2002)．我々は動詞データとして Korhonen らが作成した動詞格フレームデータを用いた(Korhonen, 2003)．Korhonen らのデータは統計的構文解析ツール(Briscoe and Carroll, 2002)により解析した結果から動詞とその格フレームを抽出したデータであり 163 の格パターンから成る．より具体的には，COMLEX と British National Corpus(Leech, 1992)から抽出された 6,433 の動詞に対して，動詞ごとに 10,000 文を抽出し格フレームのパターンを抽出している．文抽出で用いたコーパスは BNC を含む 5 種類のテストデータである(Korhonen, 2006)．我々はこのデータを用いて単語間の類似度を求めた．

3-2 分布類似度

統計的手法に基づく分布類似度は，これまでに多数提案されている．本研究では以下で示す 8 種類の分布類似度を用いて単語間の類似度を求めた．以下，x と y は動詞単語ベクトルを示しベクトルの各次元は 163 の格要素パターンを示す．

(1) バイナリー余弦尺度 (bCos)

バイナリー余弦尺度は，163 次元で構成される動詞単語ベクトル x と y の余弦尺度値を示す．ベクトルの各次元は格要素パターンの頻度を示す．バイナリー余弦尺度はパターンが存在する場合は 1，存在しない場合はゼロとなる．

(2) 相対頻度の確率に基づく余弦尺度 (rfCos)

相対頻度の確率に基づく余弦尺度は，バイナリー余弦尺度の各次元の要素が 1，0 をとるのに対し，頻度を正規化した値となる．

(3) Dice Coefficient (Dice)

Dice Coefficient は，以下の式で示される．

$$Dice(x,y)=\frac{2\cdot|F(x)\cap F(y)|}{|F(x)|+|F(y)|}$$

F(x)は動詞単語ベクトルの格パターンの種類を示す.

(4) Jaccard's Coefficient (Jacc)

Jaccard's Coefficient は，以下の式で示される.

$$Jacc(x,y)=\frac{|F(x)\cap F(y)|}{|F(x)\cup F(y)|}$$

(5) L1 Norm (L1)

L1 Norm は Minkowski 距離の 1 種であり 2 点間の距離を示す.

$$L1(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

(6) Kullback-Leibler 距離 (KL)

Kullback-Leibler 距離は以下で示される.

$$D(x\,||\,y)=\sum_{i=1}^{n}x_i*\log\frac{x_i}{y_i}$$

Kullback-Leibler 距離は yi がゼロのときは定義されない．そこでスムージング手法を用いることでゼロの場合を回避した．語のスムージング手法はこれまで多くの手法が提案されている．我々は Add-one スムージング手法を用いた (Witten and Bell , 1991).

(7) α-skew divergence (α div.)

α-skew divergence を以下に示す.

$$\alpha div(x,y)=D(y\,||\,\alpha\cdot x+(1-\alpha)\cdot y)$$

α-skew divergence を用いた Lee らの先行研究において α=0.9 が高精度であることが報告されている (Lee, 1999)．そこで本研究においても α を 0.9 に設定した.

(8) Jensen-Shannon (JS)

Jensen-Shannon を以下に示す.

$$JS(x,y)=\frac{1}{2}\left[D(x\,||\,\frac{x+y}{2})+D(y\,||\,\frac{x+y}{2})\right]$$

上記 8 種類の類似度尺度において，バイナリー余弦尺度，相対頻度の確率に基づく余弦尺度，Dice coefficient，Jaccard's coefficient を除く全ての式はその値が小さいほど 2 つの動詞単語は類似していることを示す．そこでそれらの式については，逆数をとることで類似度の値と定めた.

4 クラスタリング手法

本節では，クラスタリング手法について紹介する．本研究で提案するクラスタリング手法は，Reichardt らにより提案されたグラフベースの教師なしクラスタリング手法(RB アルゴリズム)に基づく(Reichard, 2006)．RB アルゴリズムは磁性体分類のためのソフトクラスタリング手法であり，エネルギーが最小になるように磁性体を分類する手法である．我々はこのアルゴリズムを半教師付きクラスタリングに拡張することで，多義を判定すると同時に未知語を既知のクラスに分類した．以下では，クラスタリングで用いる制約とクラスタリングの具体的なアルゴリズムを示す.

4-1 制約

半教師付きクラスタリングは，意味クラスが付与された少数のデータを用いて意味クラスが付与されていないデータを分類するクラスタリング手法である．これまでに提案されている半教師付きクラスタリング手法と同様，本手法では must-link と cannot-link という 2 つのノードの間の制約を用いる．Must-link とは 2 つのノード(本研究では動詞単語)が同じクラスタに属さなければならないということを表す制約である．一方，cannot-link とは 2 つのノードが同じクラスタには属さないという制約である．これらの制約は，少

量のラベルが付与されたノードから導くことができる．本研究では，クラスタリングの入力となる動詞単語の任意の組に対して，この制約を付与することでクラスタリングアルゴリズムを適用した．

4-2 クラスタリングアルゴリズム

本研究で提案するクラスタリング手法は，Reichardt らにより提案されたグラフベースの教師なしクラスタリング手法(RB アルゴリズム)に基づく(Reichard, 2006)．RB アルゴリズムは磁性体分類のためのソフトクラスタリング手法であり，磁性体が持つエネルギーが最小になるように磁性体を分類するために開発された手法である．Hamiltonian と呼ばれるエネルギー関数は，リンクが張られているノードと張られていないノード間のエネルギーを求める関数である．クラスタリングはノード間の局所的な情報のみ用いて作成されるため，大規模なデータに対しても適用可能である．さらに，エネルギーの極小と極大値を比較することでノードが複数のクラスタに属することを可能としている．Reichardt らは大学のフットボールチーム，たんぱく質データなど複数のデータに適用することで手法の有効性を示した．我々はこのアルゴリズムを半教師付きクラスタリングに拡張することで，多義を判定すると同時に未知語を既知のクラスに分類した．

vi(1≦i≦n)をクラスタリングの入力である動詞単語とする．N は単語数を示す．Σiは動詞 vi に付与された意味クラスのラベルとする．Hamiltonian は以下の式で定義される．

$$H(\{\sigma_i\}) = -\sum_{i<j}(A_{ij}(\theta) - \gamma p_{ij})\delta_{\sigma_i\sigma_j}. \qquad (1)$$

ここで δ は Kronecker の δ を示す．関数 $Aij(\theta)$はグラフの隣接行列を示し，仮に動詞単語 vi と vjがラベル付けされたデータである場合には式(2)で定義される．一方，それ以外の場合には式(3)で定義される．

$$A_{ij}(\theta) = \begin{cases} 1 & v_i と v_j が mult-link である \\ 0 & v_i と v_j が cannot-link である \end{cases} \qquad (2)$$

$$A_{ij}(\theta) = \begin{cases} 1 & sim(v_i, v_j) \geq 0 \\ 0 & それ以外 \end{cases} \qquad (3)$$

式(3)における $sim(vi, vj)$は動詞単語 vi と vjの類似度尺度を示す．我々は 3.2 節で示した 8 種類の類似度尺度のいずれかを用いて(3)を求めた．式(3)の θ は閾値を示す．例えば θ が 0.9 のとき，クラスタリングの入力となる任意の動詞の対の上位 10%に相当する単語対の値を 1 とし，残りの 90%に相当する単語対の値はゼロとした．

式(1)における行列 pijは動詞単語 vi と vjの間に存在するリンクの確率を示し，式(4)で示される．

$$p_{ij} = \sum_{i<j} \frac{A_{ij}(\theta)}{N(N-1)/2} \qquad (4)$$

式(4)において N は入力となる動詞単語の総数を示し，$N(N-1)/2$ は動詞対の総数を示す．式(1)のパラメータ γ が大きな値であるほど，各動詞は多くのクラスタに属する．すなわち，式(1)は実施にクラスタ内とクラスタの外のエッジが持つ値と全てのリンクが等しい値を持つ（期待値）との比較を示しており， この値が極小となるようなクラスタが求めるべきクラスタとなる．式(1)である Hamiltonian H の極小値は simulated annealing を用いて求められる． Simulated annealing 法による極小値の求め方を図 2 に示す．

入力 { $Ai\ j(\theta)$ // 隣接行列
γ
}
出力 {
(Hmin, {σi}min) // Hamiltonian H の極小値とそれに対応したクラスタの組
}
初期化 {
1. 各 {σi}に対してランダムに初期値を割り当てる.
2. Hin := H({σi}), H := Hin, Hmin := Hin をそれぞれ求める. .
4. Tmax := 10×|Hin|, Tmin := 0.1×|Hin|. // Tはパラメータを示し, Tmax と Tminはそれぞれ最大, 最小値を示す.
}
Simulated annealing {
For t = 1, 2, ・・・, N { // N は繰り返し数を示す.
T := Tmax−(Tmax− Tmin)(t− 1)/(N-1) // Tは線形に減少するとする.
For i = 1, 2, ・・・, n {
1. σiがsに変化したときの△H(σi →s)を以下の式で求める.

$$\Delta H(\sigma_i \to s) = -\sum_{m(m\neq i)} (A_{im}(e) - \gamma p_{im})\delta_{s\sigma_m}$$

2. q≠ σiとなるようなクラスタラベルqをランダムに選ぶ.
3. σiのクラスタラベルがqである確率p(σi=q)を以下の式で求める.

$$p(\sigma_i = q) = \frac{\exp(-\frac{1}{T}\Delta H(\sigma_i \to q))}{\sum_s \exp(-\frac{1}{T}\Delta H(\sigma_i \to s))}$$

4. 仮に r < p(σi = q)ならば, σi とH を以下のように更新する. rは0≦r≦1のランダムな値.
H := H +_H(σi = q), σi := q
5. 仮に H < HminならばHmin := H, {σi}min := {σi}
}
}
}

図 2： Simulated annealing 法による極小値の算出方法

図 2 で示される Simulated annealing を M 回繰り返すことで Hamiltonian 関数の極小値を求める. 各回数で得られるクラスタ結果に対して m 回以上, 同じ値が存在するとき, その値を極小値とみなした. その結果, 動詞単語 v_i が 2 つ以上の{σi}min に属しているとき, vi は多義語とみなした. RB アルゴリズムを用いた動詞単語の分類手法の流れは以下で示す 4 つのステップから成る.

1. 入力
入力同士単語の集合を {v_1, v_2, … v_n} (n は入力単語の総数)とする.

2. 類似度計算
3.3 節で紹介した 8 種類の類似度尺度のいずれかを用いて, 任意の動詞対の類似度を求める.

3. 隣接行列の作成
式(2)と(3)を用いて隣接行列 $A_{ij}(\theta)$ を作成する.

4. RB アルゴリズムの適用

3 で求めた隣接行列に対して，図 2 で示した RB アルゴリズムを適用し，動詞クラスタを得る．

入力データに対して，図 2 で示したアルゴリズムを M回適用することで Hamiltonian の極小値を求める．そこで我々は，MPI(Message Passing Interface)を用いることで並列化を行った．実装では統計数理研究所のスーパーコンピュータ SPARC Enterprise M9000, 64CPU, 1TB memory を用いてクラスタを求めた．

5 実験

5-1 実験データ及び評価方法

実験では，Korhonen らが作成した 110 からなる動詞単語を用いて未知語と多義に関する本手法の有効性を検証した．このデータを用いた理由は，先行研究との比較が容易であるためである．データは 2 種類，すなわち多義を含まないデータと含むデータが用意されている．我々は 110 の動詞単語から無作為に 10%の動詞を抽出し，これらを意味クラスのラベルが付与されたデータとして用いた．残りの 90%のデータはラベルなし（未知語）のデータとして使用した．無作為抽出は 10 回行い，それらの平均精度を求めた．動詞の格フレームと格要素は(Korhonen, 2003)で作成されたデータを用い，動詞間の類似度を求めた．クラスタリングの実験では，simulated annealing の繰り返し数 M を 1,000，極小値の重複回数 m を 3 とした．

評価尺度は(Schulte im Walde, 2000)らの precision, recall, F-score を用いた．Precision は正しく判定できた動詞語義の数を動詞語義の総数で除した値であり，recall は，正しく判定できた動詞語義の個数を入力動詞単語の正しい語義の総数で除した値である．

本手法の有効性を検証するため，我々は半教師付き学習として最もよく用いられている EM アルゴリズムとの比較を行った．EM アルゴリズム の入力は，動詞の格フレームを用いて求めた確率分布を用い，繰り返し数は 30 回とした．

5-2 実験結果

実験結果を表 1 に示す．表 1 において γ，及び θ は RB アルゴリズムで用いられているパラメータを示す．Sim は本手法で用いた 8 種類の類似度尺度を示す．表 1 の結果は，それぞれのパラメータに対してパラメータ推定を行った結果，最も高い精度(F-score)を示す．C は得られたクラスタ数を示し，62 は正解クラスタ数を示す．EM は EM アルゴリズムの結果を示す．

	θ	γ	Sim	C(62)	Prec	Rec	F-score
Semi-supervised RB	0.2	1.0	Dice	48	0.563	0.626	0.557
EM	--	--	--	59	0.301	0.512	0.387

表 1： 分類結果

表 1 より，本手法の F-score 0.557 であり EM よりも優れていることがわかる．クラスタ数を比較すると正解クラスタ数が 62 であるのに対し，本手法で得られたクラスタ数は 48 であり EM は 59 であった．しかし，EM アルゴリズムでは、テストデータは全てラベル付けされた意味クラスのどれかに分類される．すなわち，テストデータがラベル付けされた意味クラス以外のクラスである場合，これを判定することはできない．一方，本手法では，新たに 3 つのクラスを正しく認識することができていることが確認できた．このことから，テストデータが既存の意味クラスに属さない場合にもこれを認識することができるため，新規の知識を獲得できるという点において辞書拡張に有効な手法であるといえる

本手法では類似度尺度として 8 種類の尺度を用いた．そこでそれらが分類精度に与える影響を調査した．実験結果を表 2 に示す．表 2 の結果は，θ と γ に対してパラメータ推定を行った結果，最も高い F-score が得られた際の結果を示す．

Sim	θ	γ	*C*	Prec	Rec	F-score
Cos	0.1	1.1	39	0.402	0.583	0.476
rfCos	0.1	1.0	39	0.396	0.565	0.466
Dice	0.2	1.0	48	0.536	0.626	0.577
Jacc	0.1	0.7	36	0.314	0.785	0.449
L1	0.2	0.7	46	0.378	0.724	0.497
KL	0.1	0.9	38	0.411	0.630	0.497
α div.	0.1	1.2	39	0.421	0.634	0.506
JS	0.4	1.0	37	0.380	0.539	0.446
EM	--	--	59	0.301	0.512	0.387

表 2: 類似度尺度の違いによる分類精度

表 2 より最も高い F-score が得られた尺度は Dice Coefficient であり精度が低かった尺度は Jensen –Shannon であったが，各種尺度に大きな精度差はみられなかった．このことから，RB アルゴリズムは，類似度尺度に対して汎用性があること，また Hamiltonian におけるエネルギー極小化が単語の多義性と一致するという我々の仮説を実証していると言える．

半教師付き RB アルゴリズムは γ と θ のパラメータを用いている．そこでこれらパラメータ値の精度への影響を調査した．図 3 は，γ 値に対する 110 動詞単語の精度を示し，図 4 は θ 値に対する 110 動詞単語の精度を示す．類似度尺度は Dice Coefficient を用いた．

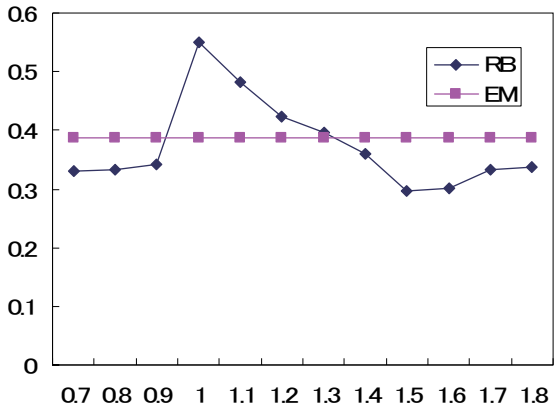

図 3: γ 値の違いによる分類精度

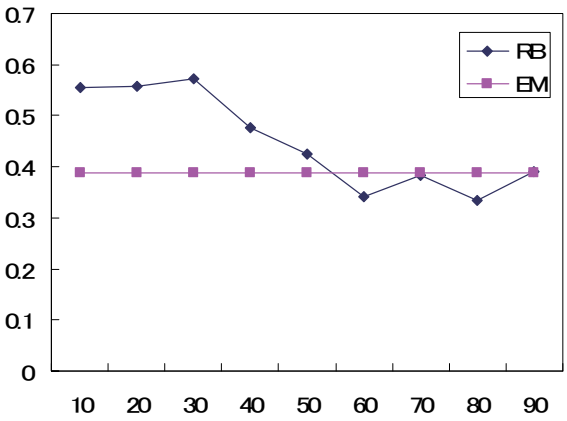

図 4: θ 値の違いによる分類精度

図 3 より，γ値が 0.9 よりも小さくなると精度は極端に下がることがわかる．これは，ほとんどの入力動詞単語が多義であるにもかかわらず，一つのクラスタのみに属してしまうためである．同様にγ値が 1.3 以上になるとほとんどの動詞単語が多数のクラスタに属してしまうため，精度が低下する．このことからγ値は極端に大きすぎても逆に小さすぎてもよい精度が得られないということが言える．

図 4 において，θの値が 30%以上になると精度低下に影響を与えていることがわかる．これはθの値が大きいほど抽出される動詞対の総数が少なくなるためであり，入力組が少なくなるとクラスタリングがうまく機能しないためである．実験の結果からθの値 30%以下が適切であるということが言える．

5-3 未知語に対する手法の有効性

未知語に対して本手法がどの程度正しく分類することができるかを検証するため，ラベル付けされたデータの量に対する本手法の有効性を調査した．実験結果を図 5 に示す．図 5 において横軸は全データ数に対するラベル付けされたデータ数の割合を示し，縦軸は F-score 値を示す．

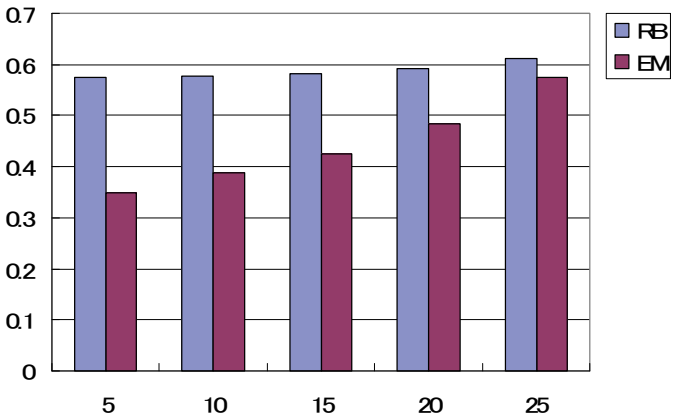

図 5: ラベル付けされたデータ量に対する分類精度

図 5 より，本手法及び EM アルゴリズム共に 5%と 10%の場合を除き，ラベル付けされたデータ量が多いほど高い精度が得られていることがわかる．本手法は 5%から 25%までのいずれにおいても EM よりも優れた精度が得られていること，また Korhonen らの報告から本実験で用いた 110 の動詞単語の多義分類は非常に難しいということが指摘されていることからも，本手法は有効であると言える．

5-4 多義語に関するエラー解析

本手法は、未知語を意味クラスに分類すると同時に，多義を判定することが可能なアルゴリズムである．そこで多義語に対して，どの程度正しく判定できているかを人手により検証した．その結果，71 の多義語のうち，13 の多義語がただしく判定できていることがわかった．例えば``drop”という単語は表 3 で示す 4 つの意味クラスに正しく判定できていた．

意味クラス	クラスタ
Putting	{**drop**, fill}
Change of State	{**drop**, dry, build}
Existence	{**drop**, hang, hit}
Motion	{**drop**, walk, travel}

表 3: drop の分類結果

その他の動詞単語は誤って分類された．誤りは２種類のタイプに分類することができる．第 1 の誤りのタイプは，部分的な誤り，すなわち多義語の一部のみが正しく判定できている場合である．表 4 に例を示す．表 4 において，``Target sense(s)”は，多義語が持つ意味を示し，``Clustered sense(s)”は，クラスタリングアルゴリズムにより分類された意味を示す．　``#times”は，誤りの総多義語数を示す．

#times	verb	Target sense(s)	Clustered sense(s)
23	hang	verbs of existence, **verbs of putting** verbs of killing **verbs involving the body**	**verbs of putting** **verbs involving the body**
31	remove	**verbs of removing,** **verbs of killing** **verbs of sending and carrying**	**verbs of removing** **verbs of killing** verbs of change of possession **verbs of sending and carrying**

表 4: 多義語に関する誤り解析（部分一致）

表 4 において，第 1 行目の例は，hang は 4 つの意味，すなわち verbs of existence，verbs of putting，verbs of killing，verbs involving the body を持つにもかかわらず，verbs of putting と verbs involving the body しか正しく判定できていないことを示す．第 2 行目の例は，remove は 4 つの意味をもつにもかかわらず，3 つの意味しか正しく判定されず，かつ verbs of change of possession という誤った意味に分類されている例である．

2 つめの誤りタイプの例は，多義語が一つの意味しか持たないと判定された誤りである．表 5 に誤り例を示す．表 5 において多義動詞 sit は verbs of existence と verbs of putting の 2 つの意味があるにもかかわらず，verbs of existence のみに分類されている例である．この種のあやまりタイプには，総計 4 つの動詞単語が存在した．

#times	verb	Target sense(s)	Clustered sense(s)
4	sit	**verbs of existence** verbs of putting	**verbs of existence**

多義語に関するエラー解析から今後の課題として少なくとも 3 点挙げることができる．1 点目は，半教師付き RB アルゴリズムへの階層構造の導入である．すなわち，入力単語集合に対して本手法を複数回適用することにより，大まかな意味の分類から粒度の細かい分類が可能になるのではないかということである．2 点目は，クラスタリングの情報として格フレームと格要素に加え，格要素になりえる語の意味を用いることで動詞間の高精度な類似度を求めるということである．3 点目は，定量的な評価を実施するために Levin's classes 以外のシソーラス辞書，例えば WordNet への適用が考えられる．

6 まとめ

日本語インターネットディレクトリィの各分野と英語インターネットディレクトリィの各分野との対応付けを行う際に生じる多義語と未知語の問題に対処するため，グラフベースの半教師付きアルゴリズムを提案することで，多義語を含む未知語の意味推定を行う手法を提案した．

本研究で提案するクラスタリング手法は，Reichardt らにより提案された磁性体分類のためのグラフベースによる教師なしクラスタリング手法(RB アルゴリズム)に基づくアルゴリズムであり，我々はこのアルゴリ

ズムにおいて，(1) エネルギー最小を多クラス分類と捉える，(2) 教師なしを半教師付きに拡張することで未知語を分類するようアルゴリズムを拡張することで，多義語を含む未知語の意味推定を行った．実験の結果，既存の半教師付き学習よりも優れていること，また既存の意味クラスにないテストデータも判定できることを確認した．今後の課題は，5 章で述べたことのほか，1. 日英インターネットディレクトリィから収集した日英文書の単語変換，2. 機械学習を用いた文書分類によるインターネットディレクトリィの分野対応，3. 日英対応文書抽出，4. 対応文書からの対訳語抽出を実施する予定である．

【参考文献】

[Brew, 2002] C. Brew and S. S. Walde, “Spectral Clustering for German Verbs”, Proc. of 2002 Conference on Empirical Methods in Natural Language Processing, 2002, pp. 117-123.

[Briscoe, 2002] E. J. Briscoe and J. Carroll, “Robust Accurate Statistical Annotation of General Text”, Proc. of 3rd International Conference on Language Resources and Evaluation”, 2002, pp. 1499-1504.

[Korhonen, 2003] A. Korhonen and Y. Krymolowski and Z. Marx, “Clustering Polysemic Subcategorization Frame Distributions Semantically”, Proc. of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 64-71.

[Korhonen2006], Korhonen and Y. Krymolowski and T. Briscoe, “A Large Subcategorization Lexicon for Natural Language Processing Applications”, “Proc. of the 5th International Conference on Language Resources and Evaluation”, 2006.

[Lee1999] L. Lee, “Measures of Distributional Similarity”, Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 25-32.

[Leech, 1992] G. Leech, “100 Million Words of English: the British National Corpus”, Language Research, 1992, Vol. 28, no. 1, pp. 1-13.

[Levin, 1993] B. Levin, “English Verb Classes and Alternations”, Chicago University Press, 1993.

[Matsuo, 2006] Y. Matsuo and T. Sakaki and K. Uchiyama and M. Ishizuka, “Graph-based Word Clustering using a Web Search Engine”, Proc. of 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 542-550.

[Mihalcea, 2005] R. Mihalcea, “Unsupervised Large Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling”, Proc. of the Human Language Technology / Empirical Methods in Natural Language Processing Conference”, 2005. pp. 411-418.

[Pereira, 1993] F. Pereira and N. Tishby and L. Lee, “Distributional Clustering of English Words, “Proc. of the 31st Annual Meeting of the Association for Computational Linguistics”, 1993, pp. 183-190.

[Reichard, 2006] J. Reichardt and S. Bornholdt, “Statistical Mechanics of Community Detection”, PHYSICAL REVIEW E, 2006, vol. 74.

[Schulte, 2000] S. Schulte im Walde, “Clustering Verbs Semantically according to their Alternation Behaviour”, Proc. of the 18th International Conference on Computational Linguistics, 2000, pp. 747-753.

[Stevenson, 2003] S. Stevenson and E. Joanis, “Semi-Supervised Verb-Classes Discovery using Noisy Features”, Proc. of the 7th Conference on Natural Language Learning at HLT-NAACL, 2003,
pp. 71-78.

[Widdows, 2003] D. Widdows and B. Dorow, “A Graph Model for Unsupervised Lexical Acquisition”, Proc. of 19th International Conference on Computational Linguistics, 2002, pp. 1093-1099.

[Witten, 1991] I. H. Witten and T. C. Bell, “The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression”, IEEE Transactions on Information Theory, 1991, vol. 37, No. 4, pp. 1085-1094.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Identification of Domain-Specific Senses in a Machine-Readable Dictionary	Proc. of the 29th Annual Meeting of the Association for	2011, 6

	Computational Linguistics: Human Language Technologies, pp. 552-557.	
Semantic Classification of Unknown Words based on Graph-based Semi-supervised Clustering	Proc. of the International Conference on Knowledge Engineering and Ontology Development, pp. 37-45.	2011, 10