自然言語処理 116－5
（1996．11．18）

シソーラス上での共起頻度を利用した動詞の多義解消

内山将夫　板橋秀一

筑波大学 電子・情報工学系

本稿では語義の共起関係に基づいた尤度パラメータの標本空間をシソーラスの構造に沿って動的に拡張することにより動詞の多義を解消する手法を提案する．尤度が1位と2位の語義の尤度差が閾値よりも大ならば語義の判定をし，小ならば一段拡張された標本空間で多義の解消を試みる．本稿の実験では，EDR日本語コーパスから頻度500以上の異なり動詞74語を抽出し，延べで約89,000の動詞について多義性解消の実験を行った．提案手法とクラスベースの手法との判定率を比べると，提案手法の判定率の方が統計的に有意に高く，その有効性が示された．

Verb Sense Disambiguation Using Co-occurrence on a Thesaurus

Masao Utiyama　Shuichi Itahashi

Institute of Information Sciences and Electronics, University of Tsukuba

This paper proposes a method disambiguating verb senses using co-occurrence-based plausible parameters whose sample spaces are extended according to a thesaurus. The method selects the most plausible sense if its plausibility is significantly greater than that of the second most plausible one. If the difference is not significant, the sample space is extended and the significance test is tried again. The method was applied to 74 polysemous verbs (about 89,000 instances) extracted from the EDR Japanese Corpus. The applicability of the proposed method is significantly higher than that of a class-based method, which shows the plausibility of the proposed method.

1　はじめに

単語の多義性の解消を，与えられた単語の語義の集合の中から適切な語義を選択することであると定義する．この要素技術は，機械翻訳における訳語の選択，仮名漢字変換における変換候補の選択，文献検索における文献の検索，などの応用がある重要な技術であり，様々な研究が行なわれている [1]．最近の傾向ではコーパスに基づいて多義性を解消するというものが多い．本稿でも，統計的な方法により単語の多義性を解消することを試みる．対象は動詞である．語義間の共起関係をEDR日本語コーパス[1] [2]から抽出し，それを利用して多義を解消する．共起頻度が足りない場合にはシソーラス (分類語彙表 [3]

[1]本稿で用いたEDR日本語コーパス，日本語単語辞書，概念体系はVersion1.5である．

／EDR概念体系) により補完する．

従来の研究でシソーラスを用いて共起頻度の補完をするものとしては，クラスベースの手法と事例ベースの手法とがある．クラスベースの手法は単語の代りに単語の上位にある節点を利用する．そのための節点は，前もって決めておいたり [4]，コーパスにおける単語間の統計的な情報を利用して決めたりする [5, 6, 7]．一方，事例ベースの手法では このような抽象化は行わない．入力単語がコーパスに出現していない場合は，出現している単語のうちで入力単語とシソーラス上での距離が最短の単語を利用して多義性を解消する [8, 9, 10]．二つの手法のうち，クラスベースの手法ではクラス内にある単語同士の差異を記述できない [11]，また，事例ベースの手法では最短距離にある節点の振舞いが入力単語の振舞いと異なる場合には多義の解消に失敗することになる．これは，一方では平均化により情報が失なわれ [11]，他方では個別化によりノイズに弱くなる [6]，という二律排反な状況である．

本稿ではノイズを避けて適当な抽象度の節点により多義の解消をする手法について述べる．従来の研究では固定的に選ばれていた上位節点を，入力に応じて統計的に動的に選択するという点が提案手法の要点である．以下，2章では多義の解消法について述べ，3章では本稿で提案する手法の有効性を実験により示す．4章は結びである．

2 多義の解消法

本稿で提案する手法は，語義の尤度となるパラメータ (確率変数) の標本空間をシソーラスに沿って段階的に拡張することによりノイズを避けて多義を解消する手法であるが，まず，標本空間を固定した場合における多義の解消法について述べ，そのあとで，可変な標本空間における多義の解消法について述べる．

2.1 固定された標本空間

関係 r にある単語 W と動詞 V とが与えられ，それらの語義が $W = \{w_1, w_2, \ldots\}$，$V = \{v_1, v_2, \ldots\}$ であるとする．語義 w_h と v_i とが関係 r で共起することを $r(w_h, v_i)$ で表す．その共起頻度 $n(r(w_h, v_i))$ は，標本空間を $\{r(w_1, v_1), r(w_2, v_2), \ldots\}$ とする確率変数 $N(r(w_h, v_i))$ の観測値として表される．このとき，$n_i = n(r(W, v_i)) = \sum_{w_h \in W} n(r(w_h, v_i))$ に基づいて V の語義を決める．$n(r(W, v_i))$ は $\{r(W, v_1), r(W, v_2), \ldots\}$ を標本空間とする確率変数 $N(r(W, v_i))$ の観測値である．

本稿では，この標本空間における共起頻度の分布が一般化超幾何分布[2]に従うと仮定する．つまり，標本空間を $\{r(W, v_1), r(W, v_2), \ldots, r(W, v_k)\}$ としたとき，W と v_i との共起頻度を表す確率変数 $N(r(W, v_i))$ の値には $0 \leq N(r(W, v_i)) \leq N_i$ という制限があり，$N_1 + N_2 + \cdots + N_k = N$ であるとする．このとき次のような確率変数 F_i を考える．

$$F_i = \frac{N_i - n_i}{N - n} \tag{1}$$

ただし，$n = n_1 + n_2 + \cdots + n_k$．すると，$F_i$ の期待値，分散，および，F_i と F_j の共分散は，N_i の事前確率として一様分布を仮定すると，以下の通りである．

$$\langle F_i \rangle = (n_i + 1)/(n + k) \tag{2}$$

$$var(F_i) = p_i(1 - p_i)/(n + k + 1) \tag{3}$$

$$cov(F_i, F_j) = -p_i p_j/(n + k + 1) \tag{4}$$

ただし，$p_i = \langle F_i \rangle$ である．

動詞の語義を適当に並べかえて，$r(W, v_1), r(W, v_2), \ldots,$ について $i \geq j$ ならば $n_i \geq n_j$ であるようにする．このとき，$D = F_1 - F_2$ の推定値の下側信頼限界 (Pl) に基づいて語義の判定をするかしないかを決める．D の期待値と分散は，それぞれ，$\langle D \rangle = \langle F_1 \rangle - \langle F_2 \rangle$，$var(D) = var(F_1) + var(F_2) - 2cov(F_1, F_2)$ である．α と θ を適当に選んで，$Pl = \langle D \rangle - \alpha\sqrt{var(D)} > \theta$ である場合には v_1 を語義とする．そうでない場合には関係 r にお

[2] ある母集団が k 種類の個体からなるとき，それぞれの種類の個体数を $N_1, N_2, \ldots N_k$ とする ($N = N_1 + \cdots + N_k$)．n 個の個体を非復元抽出したとき，それぞれの種類の個体が $n_1, n_2, .., n_k (n = n_1 + \cdots + n_k)$ だけ選ばれる確率は，一般化超幾何分布 $\binom{N_1}{n_1} \cdots \binom{N_k}{n_k} \binom{N}{n}^{-1}$ で表される．

いては語義の判定を行わない．α と θ の値は3章で述べる．もし，複数の関係がある場合には最大の Pl である関係に基づいて語義を決める．

2.2 可変の標本空間

前節で述べた手法は，標本空間 $\{r\}\times\{w_1,w_2,\ldots\}\times\{v_1,v_2,\ldots\}$ を縮小した標本空間 $\{r(W,v_1),r(W,v_2),\ldots\}$ における共起頻度の分布についての手法である．ここでは，標本空間をシソーラスに沿って拡張することを考える．標本空間を段階的に拡張し，各段階において統計的な判定を行い，判定が下された時点で語義の判定のプロセスを終える．まず，標本空間の拡張の仕方について述べ，次に，推定値の下側信頼限界の求め方について述べる．

2.2.1 標本空間の拡張の仕方

ここで考える標本空間は $\{r\}\times U_i\times\{v_1,v_2,\ldots\}$ である．U_i はシソーラスの構造に従って段階的に拡張される．本稿ではシソーラスとは一つの根を有する DAG(Directed Acyclic Graph) であるとする．シソーラスの節点は語義または語義を一般化した概念を表わしている．ある節点の支配下の節点とは，その節点から到達できる節点である．

$W=\{w_1,w_2,\ldots\}$ であるとき，根から w_j までの道上の節点において，根からの距離が i にある節点が支配する葉の集合を U_{ij} とする[3]．このとき，根から w_j までの距離を l とすると，

$$U_{0j}\supseteq U_{1j}\supseteq\cdots\supseteq U_{lj}=\{w_j\} \qquad (5)$$

である．U_i を以下のように定義する．

$$U_i=\bigcup_{w_j\in W}U_{ij} \qquad (6)$$

[3]任意の単語の任意の語義は葉で表現されると仮定する．この場合には各々の語義は互いに支配関係にない．分類語彙表の場合には，これが成立する．しかし EDR 概念体系の場合には，語義にあたる概念が葉であるとは限らないため，その語義にあたる節点が別の語義にあたる節点を支配している場合がある．その場合には，ある節点における葉の数が，その節点が支配する語義の数と一致しない．そのため (8) 式や (9) 式において考慮されない語義がでる．本稿ではこの問題は無視し，全ての語義が葉に相当するとして尤度を計算した．

(5) 式と同様に，$i\le j$ のときには $U_i\supseteq U_j$ が成立する．図1の例で $W=\{w_4,w_5\}$ とすると，$U_0=\{w_1,w_2,w_3,w_4,w_5,w_6\}$, $U_1=U_2=\{w_4,w_5,w_6\}$, $U_3=\{w_5\}$ である．

複数の親を持つ節点の場合には，根からの距離として複数の道の中で最長のものを選択すれば，$i\le j$ のときに $U_i\supseteq U_j$ となる．

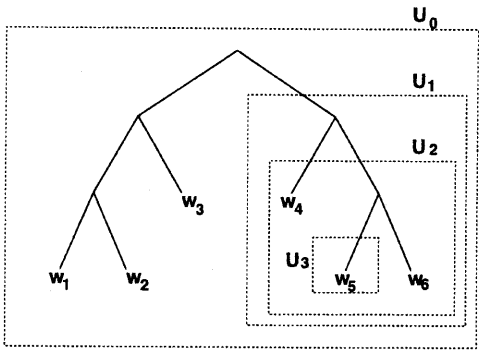

図 1: U_i の例

2.2.2 推定値の下側信頼限界の求め方

関係 r にある単語 W と動詞 V について，語義を $W=\{w_1,w_2,\ldots\}$, $V=\{v_1,v_2,\ldots\}$ とする．ここでの標本空間は $I=\{r\}\times U_i\times V$ である．また，$W'=W\cap U_i$ とする．

語義 v_j の尤度は次のように定義された確率変数 $F(W',v_j|I)$ により表される．

$$\begin{aligned}F(W',v_j|I) &= F(v_j|I)F(W'|v_j,I)\\ &= F(v_j|I)\sum_{w\in W'}F(w|v_j,I) \qquad (7)\end{aligned}$$

$F(W',v_j|I)$ は，まず，動詞の語義を選び，次に，その語義のもとで単語の語義を選ぶことを表す．

$F(v_j|I)$ や $F(w|v_j,I)$ は，前と同様に一般化超幾何分布に従う確率変数であり，期待値は以下の通りである．$n(r(u,v))$ は関係 r における共起頻度を表す．分散や共分散も (3)，(4) 式と同様である．

$$\langle F(v_j|I)\rangle = \frac{\sum_{u\in U_i}n(r(u,v_j))+1}{\sum_{u\in U_i,v\in V}n(r(u,v))+|V|} \qquad (8)$$

$$\langle F(w|v_j,I)\rangle = \frac{n(r(w,v_j))+1}{\sum_{u\in U_i} n(r(u,v_j))+|U_i|} \quad (9)$$

$\sum_{u\in U_i} n(r(u,v_j))$ は，葉における動詞語義との共起頻度を根に向って再帰的に伝播することで各節点に記録する．また，$|U_i|$ はシソーラスの構造から決まる値であるため，あらかじめ計算しておくことができる．そのため，$\sum_{u\in U_i} n(r(u,v_j))$ や $|U_i|$ の値を支配下の葉の集合から実行時に計算する必要はない．また，(8) 式や (9) 式の値は，U_i ごとに計算され，U_i は最大でシソーラスの高さだけの数しかないので，これらの値を計算することは計算量の面で困難ではない．

$F(W',v_j|I)$ の期待値は，$F(v_j|I)$ と $F(W'|v_j,I)$ とが独立であるとみなすと，以下の通りである．

$$\langle F(W',v_j|I)\rangle = \langle F(v_j|I)\rangle \sum_{w\in W'} \langle F(w|v_j,I)\rangle \quad (10)$$

動詞の語義を適当に並べかえて，$\langle F(W',v_1|I)\rangle \geq \langle F(W',v_2|I)\rangle \geq, \ldots$ であるようにし，$D = F(W',v_1|I) - F(W',v_2|I)$ の推定値の下側信頼限界 (Pl) に基づいて語義の判定をするかどうかを決める．D の期待値は $\langle D\rangle = \langle F(W',v_1|I)\rangle - \langle F(W',v_2|I)\rangle$ である．α と θ を適当に選んで，$Pl = \langle D\rangle - \alpha\sqrt{var(D)} > \theta$ である場合には v_1 を語義とする[4]．そうでない場合には，U_i の段階では語義の判定をせずに，標本空間をシソーラスに沿って拡張した U_{i-1} で再び語義の判定をする．U_0 においても判定ができないときには関係 r においては語義は判定しない．α と θ の値は3章で述べる．もし，複数の関係がある場合には最大の Pl に基づいて語義を決める．

[4]分散 $\sqrt{var(D)}$ は (7) 式から，(11) 式と (12) 式を利用することにより導かれる．ただし，(10) 式のときと同様な確率的独立性を仮定する．

$$var(\sum_{i=1}^{k} a_i x_i) = \sum a_i^2 var(x_i) + \sum_{i\neq j} a_i a_j cov(x_i,x_j). \quad (11)$$

また，確率変数 x と u とが独立なとき，

$$cov(xy-uv) = \langle x\rangle\langle u\rangle cov(y,v). \quad (12)$$

3 実験

EDR 日本語コーパス[5]から，多義性解消の対象となる動詞を主辞とする文節と それと係り受け関係にある文節の集合とを抽出した．これをセットと呼ぶ．抽出されたセットの集合から頻度 500 以上の動詞 74 語を選び，それらの動詞を含むセットを実験の対象とした．実験では頻度が 5000 を超える 4 語 (ある，いう，する，なる) については，無作為に抽出した 5000 のセットについて実験を行った．その他の動詞については，抽出された全てのセットについて実験を行った．実験に利用したセットの数は約 89,000 である．この 74 語には異なり語義数が 1 の動詞は含まれていない．また，頻度が 10 未満の語義を語義としてもつ動詞を多義性解消の対象とするようなセットは実験のデータから除いた．シソーラスとしては分類語彙表と EDR 概念体系の二つを用いた．

各動詞について，抽出したセットに対して 10 分割のクロスバリデーション法で多義性解消の実験を行った．すなわち，抽出したセットの集合を 10 個の均等な大きさの部分集合に分け，9 個の部分集合を訓練データとして共起頻度を得て，残りの部分集合をテストデータとして多義の解消をするということを 10 回繰り返した．その結果を表1に示す．

訓練データから語義の共起頻度を得るとき，分類語彙表を使うときには，単語に付与されている分類番号が n 個あるとする全ての分類番号との共起頻度を $1/n$ とした．EDR 概念体系を使うときには，EDR 日本語コーパスにおける単語に概念識別子が付いているときには，それとの共起頻度を 1 とした．概念識別子が付いていないときには分類語彙表のときと同様に均等に頻度を割当てた．なお，テストのときには，コーパスに付与されている概念識別子は用いていない．

多義性解消の実験は，次の要因を組み合わせて行なわれた．その正確な組合せ方は表1にある．ま

[5]EDR 日本語コーパスは新聞・雑誌・辞典などの流通文書から 1 文単位で抽出された 22 万文からなるコーパスであり，各文は形態素・構文・意味解析がされており，多くの形態素には語義として概念識別子が付与されている．語義の係り受け関係は構文解析の結果から抽出できる．

標本空間	固定								可変		最頻
共起対象	語義 ($\alpha = 1.55$)				クラス ($\alpha = 1.95$)				語義 ($\alpha = 1$)		
Thesaurus	分類		EDR		分類		EDR		分類	EDR	
手法	Dagan	差	Dagan	差	Dagan	差	Dagan	差	差		
判定率	.267	.332	.302	.363	.688	.682	.765	.764	.726	.865	1.000
適合率	.750	.752	.753	.751	.710	.707	.696	.696	.713	.695	0.652

表 1: 判定率・適合率の平均値

ず，標本空間は2章で述べた意味において固定または可変である．次に，多義性解消の手法は，2章で述べたものと Dagan[12] の手法を変形した手法である．提案手法は表1では「差」として記載されている．また，最大頻度の語義を常に選ぶ手法とも比較した．それは「最頻」として示されている．シソーラスには EDR 概念体系か分類語彙表かを用いた．共起の相手は標本空間が可変の場合には語義であり，固定の場合には語義かクラスである．クラスというのは語義のクラスであり，分類語彙表の場合には分類番号の上位 3 桁を共有する語義が同一のクラスである．EDR 概念体系の場合には，根からの距離が 4 である概念が一つのクラスを代表する概念であり，それに支配される概念の集合が一つのクラスである．2章で述べた手法をクラスに拡張するには，動詞の語義とクラスとの共起頻度を そのクラスに属する語義との共起頻度の和とする．表で，$\alpha = 1.55$ などとあるのは，2章で述べた，差に基づく手法においてパラメータ α の値を 1.55 としたということである．θ は 0 に固定した．Dagan の手法のパラメータは文献 [12] に従った．これらの値は，比較の便宜のために適合率の差が 1 %未満になるように調整した場合の値である．

表1には判定率と適合率の平均を示した．その定義を次に示す．正解は EDR 日本語コーパスでの語義に従って判定した．

$$\text{判定率} = \frac{\text{語義の判定がされたセットの数}}{\text{その動詞を含むセットの数}} \quad (13)$$

$$\text{適合率} = \frac{\text{正解の数}}{\text{語義の判定がされたセットの数}} \quad (14)$$

表1で，固定された標本空間において共起の対象として語義とクラスとを比較すると (Dagan の手法における閾値は同じだが) 前者の判定率が低い．適合率は高いが，その差は 5 %ほどなので，40 %程度の判定率の差を埋めるほどではない．Dagan の方法と提案手法とを比べると，語義を共起の相手としたときには 6 %ほど提案手法の方が判定率が高い．これは提案手法の方が共起頻度の違いに敏感なことを示している．クラスを共起の相手とした場合には Dagan の手法の方が判定率が高いが，判定率の差は 1 %未満である．手法の違いが，共起の相手を語義にした場合に比べて効かないのは，クラスを共起の相手とした場合には語義ごとの共起頻度の情報が無視されるためであると考えられる．

次に，固定された標本空間におけるクラスの結果と可変な標本空間の結果を，差に基づく手法について比べると，可変な標本空間における方が，分類語彙表の場合には 4 %程度，EDR 概念体系の場合には 10 %程度，判定率が高い．クラスを共起の相手としたとき，Dagan の手法と提案手法との判定率の違いが 1 %未満であるのと比べて，4 あるいは 10 %程度の違いは大きいと考える．なお，本稿程度の規模の実験では，1 %程度の差があれば，EDR 日本語コーパスが日本語全体を良く代表していると仮定して，日本語全体でも同様な傾向があるといえる．

提案手法とクラスベースの手法とを比較すると，クラスベースの手法では，単語や語義をまとめてクラスとするときに情報 (本稿の場合には共起頻度の情報) を損失する可能性がある．しかし，可変な標本空間の場合には共起頻度の情報は損失されない．この違いが，判定率の差として表1に現われている．また，クラスは，本稿での場合のように，先験的に決めるか，あるいは，データに基づいて決める [5, 6, 7]

必要がある．データに基づく場合には，必要が生じた時点でクラスを変更する必要がある．しかし可変の場合にはクラスの設定自体が不要である．

提案手法と事例ベースの手法とを比較すると，提案手法では，動詞の語義の尤度は，シソーラスの構造と動詞語義のシソーラス上での頻度分布と入力単語により決まる．これはシソーラス上での距離をコーパスでの共起情報と入力単語とを利用して尤度に変換しているとみなすこともできる．同様なことは文献 [13] で行なわれている．しかし，文献 [13] は単語間の類似性を定義することを目的としていて，多義の解消は直接の目的とはしていない．

複数の係り受け関係の依存関係を取扱うことは今後の課題である．依存関係を考慮した研究には，既存の格フレームを利用したもの [8, 10]，格フレームあるいは決定木を獲得するもの [14]，対数線型モデルにより依存関係を推定するもの [15] などがある．

4 おわりに

標本空間をシソーラスの構造に従って動的に拡張し，動詞多義の解消をする手法を提案した．EDR 日本語コーパスから頻度 500 以上の動詞 74 語を抽出し，延べで約 89,000 の動詞について多義性解消の実験を行った．分類語彙表をシソーラスとして利用した場合には，73 %の判定率で 71 %の適合率であった．EDR 概念体系の場合には，87 %の判定率で 70 %の適合率であった．最頻の語義を常に選ぶ場合の判定率は 100 %，適合率は 65 %であった．提案手法とクラスベースの手法との判定率を比べると，提案手法の判定率の方が統計的に有意に高く，提案手法の有効性が示された．

動詞と係り受け関係にある複数の単語間の関係をモデル化すること，仮名漢字変換などに提案手法を応用すること，などが今後の課題である．

参考文献

[1] 長尾真, 佐藤理史, 黒橋禎夫, 角田達彦. 自然言語処理, 岩波講座 ソフトウェア科学, 第 15 巻, 第 5 章. 岩波書店, 1996.

[2] 日本電子化辞書研究所. ＥＤＲ電子化辞書マニュアル. http://www.iijnet.or.jp/edr, 1995.

[3] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[4] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories. In *Proceedings of COLING-92*, pp. 454–460, 1992.

[5] Philip Resnik. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *Proceedings of AAAI Workshop on Statistically-based NLP Techniques*, pp. 48–56, 1992.

[6] 野美山浩. 事例の一般化による機械翻訳. 情報処理学会論文誌, Vol. 34, No. 5, pp. 905–912, 1993.

[7] 田中英輝. シソーラスを利用した言語データ最適一般化アルゴリズム. 自然言語処理研究会 NL108-14, 情報処理学会, 1995.

[8] 黒橋禎夫, 長尾真. 格フレーム選択における意味マーカと例文の有効性について. 自然言語処理研究会 NL91-11, 情報処理学会, 1992.

[9] 飯田仁. 人工知能におけるスーパーコンピューティング — 言語表現の類似性を利用する自然言語処理技術 —. 情報処理, Vol. 36, No. 2, pp. 164–168, 1995.

[10] 藤井敦, 乾健太郎, 徳永健伸, 田中穂積. 動詞多義性解消における格要素の貢献度について. 自然言語処理研究会 NL111-9, 情報処理学会, 1996.

[11] Ido Dagan, Shaul Marcus, and Shaul Markovitch. Contextual word similarity and estimation from sparse data. In *Proceedings of the Annual Meeting of ACL*, pp. 164–171, 1993.

[12] Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, Vol. 20, No. 4, pp. 563–596, 1994.

[13] Hiroyuki Shinnou. Redefining similarity in a thesaurus by using corpora. In *Proceedings of COLING-96*, pp. 1131–1134, 1996.

[14] 田中英輝. 動詞訳語選択のための「格フレーム木」の統計的な学習. 自然言語処理, Vol. 2, No. 3, pp. 49–72, 1995.

[15] Rebecca Bruce and Janyce Wiebe. Word-sense disambiguation using decomposable models. In *Proceedings of the Annual Meeting of ACL*, pp. 139–145, 1994.