辞書に基づく、学習を用いない語義曖昧性解消

鈴木 敏 (satoshi@cslab.kecl.ntt.co.jp)
NTT コミュニケーション科学基礎研究所

語義曖昧性解消は学習に基づいて行われるのが一般的だが，このための学習データの作成は負荷が大きく，できればこの負担は回避したい．本稿では，学習データの作成を伴わない語義曖昧性解消手法を提案する．提案する手法は辞書から取り出したネットワーク型オントロジーを基に語義曖昧性を判断する手法である．ここでいうネットワーク型オントロジーとは，単語語義間の上位下位関係が双方向に重みをもって与えられている構造である．提案手法では，評価対象文中の多義語に関して，近傍の単語と共通する上位語を多く持つ語義ほど確からしいと考える．すなわち，近傍には意味的に近い語義が集まるという仮定に基づいている．この意味処理をネットワーク型オントロジーを用いて高い精度で行う点が本手法の特徴である．計算機実験を通して，Senseval2 日本語辞書タスクの結果と比較し，学習を用いなくても高いレベルの語義曖昧性解消が可能であることを示す．

1　はじめに

テキストの意味理解のための主要な課題の一つに語義曖昧性解消がある．この課題に対して，様々なアプローチが為されているが，その多くは学習に基づいて行われるものである．学習に基づく語義曖昧性解消は精度が高いが，それらの手法を実現するための学習用データの作成は負荷が大きく，できれば避けたい作業である．また，学習用データとしては評価対象に近いコーパスを利用することが好ましいため，任意の文の解析を目的とするのであれば，大量の学習用データが必要になるという問題もある．

本稿では，学習データの作成を伴わない語義曖昧性解消手法を提案する．提案する手法は辞書から取り出したネットワーク型オントロジーを基に語義曖昧性を判断する手法である．ここでいうネットワーク型オントロジーとは，単語語義間の上位下位関係が双方向に重みをもって与えられているデータそのものを指す．本稿では，このネットワーク型オントロジーを辞書の再帰的展開を用いて導出し，語義判定の基礎データとして利用する．再帰的展開により，辞書中の定義文の共起情報からは直接取り出せない単語間の関係を広く取り出すことが可能になり，高い精度での意味処理を行うことができる．

語義の選択は，評価対象の文中の各単語と共通する上位語を多く持つ語義ほど確からしいと考える．すなわち，同じ文中には意味的に近い語義が集まるという仮定に基づいている．以上をまとめると，再帰的展開により，辞書のスパースな単語間関係から非常に多くの単語間の上位下位関係を取り出し，この関係を利用して単語間の意味的な距離から語義を推定するという手法である．

以下，前半で辞書からネットワーク型オントロジーを生成する方法を概説し，後半にはそれを用いた語義曖昧性解消手法の詳細を記す．まず次節では，語義文の再帰的展開を用いて間接的な単語間の関係を取り出す手法について概要を述べる．続く節では，展開した語義文と上位語との関係を示し，上位語に成り易さと相関のある数値により，上位下位関係を表すネットワーク型オントロジーを生成できることを示す．4 節では，ネットワーク型オントロジーを用いた確率モデルによる語義曖昧性解消手法の詳細を記す．5 節では実際に語義曖昧性解消を行い，Senseval2 日本語辞書タスクの結果と比較しながら提案手法の有効性を評価する．

2　定義文の再帰的展開

まずはじめに，辞書における単語間の間接的な関係を取り出す手法について述べる．この手法によれば，辞書の定義文を再帰的に展開することにより，直接的な単語間の関係以外の情報を取り込むことが可能となり，データスパースネスの解消を行うことができる (鈴木 2005)．以下は単語の多義性を考慮しない場合の再帰的展開手法の概説である．

辞書（国語辞典）は見出語と定義文の組合せから成り立っている．定義文は単語の集合であり，これを見出語の集合とみなせば，一つの定義文を複数の定義文の集合として再定義することができる．ところが，このような展開は無限に続いてしまう．従って，展開された定義文中の単語数も無限になり，頻度計算は一般に不可能になる．しかしながら，定義文の展開を行なう毎に一定の割合でその影響が小さくなるとすれば，無限に展開された定義文の影響は元の定義文に比べて微小になる．このとき，定義文の影響力は語義展開の回数に従う等比数列として表すことができる．同時に，様々な深さまで展開した定義文の集合体を考え，その総和を無限級数として計算すると必ず有限な値となる．これにより，定義文の集合体中の単語頻度も有限になり，計算可能となる．これらを確率モデルに置き換えることで，無限の展開を含めた定義文の集合体から単語の出現確率を取り出すことが可能になり，拡張された定義文として再定義することができる．

計算の概要を以下にまとめる．以下，n 回展開された定義文を n 次の定義文と呼ぶ．また，辞書中の定義文を 0 次定義文とする．

まず，見出語 w_i と 0 次定義文中の単語 w_j の関係は $P(w_j^{(0)}|w_i)$ と表すものとする．ここで，$w^{(n)}$ は n 次の定義文中の単語 w を表す．従って，確率 $P(w_j^{(0)}|w_i)$ は，見出語 w_i の 0 次定義文中に現れる w_j の出現確率である．

この表記を用いると，各見出語に関する 0 次定義文中の単語の出現確率は

$$A = \begin{bmatrix} P\left(w_1^{(0)}|w_1\right) & \cdots & \cdots & P\left(w_1^{(0)}|w_m\right) \\ P\left(w_2^{(0)}|w_1\right) & \ddots & & \\ \vdots & & \ddots & \\ P\left(w_m^{(0)}|w_1\right) & & & P\left(w_m^{(0)}|w_m\right) \end{bmatrix} \quad (1)$$

の列ベクトルとして表される．ここで，m は辞書中の見出語の数である．行列 A の各要素 $P(w_j^{(0)}|w_i)$ は見出語 w_i の定義文中の単語頻度 $N_i(w)$ を用いて

$$P(w_j^{(0)}|w_i) = \frac{N_i(w_j^{(0)})}{\sum_{all\ k} N_i(w_k^{(0)})} \tag{2}$$

と書ける．このとき，全ての列ベクトルは，要素の合計が 1 であり，確率表現となっている．さらに，この表記に従うと，n 次定義文は A^{n+1} により表される．

目的とする定義文の集合体 C は，語義展開の度に定義文の影響が一定の割合 a で減少すると仮定すると，

$$C = (1-a)(A + aA^2 + \cdots + a^{n-1}A^n + \cdots) \tag{3}$$

と書ける．ここで，係数 $1-a$ は正規化のための定数である．

式 (3) は無限級数の計算から

$$(I - aA)C = (1-a)A \tag{4}$$

と書け，線型計算により解を求めることができる．

計算により得られる行列 C は，列ベクトルの各要素の合計が必ず 1 となり，確率として扱うことが出来る．C の (j,i) 要素を $P(w_j^*|w_i)$ と書くと，w^* は定義文の集合体の中の単語を意味することになる．この定義文の集合体を拡張定義文と呼ぶことにすると，$P(w_j^*|w_i)$ は拡張定義文中の単語の出現確率頻度である．

3　拡張語義文と上位語の相関

上記は単語の多義性を考慮しない場合についての説明であった．しかし，本稿では語義曖昧性解消を目的とするため，上記の手法を語義毎に適用する必要がある．即ち，複数の語義をまとめた定義文から 1 語義の語義文へと読みかえることになる．ここで問題となるのは，語義の再帰的展開のためには，語義文中の多義語の語義判定が必要になることである．語義が判定できなければ，再帰的展開も不可能となる．

この問題を解決するために，語義文中の語義曖昧性を排除した辞書である Lexeed(笠原，佐藤，田中，藤田，金杉，天野 2004) を利用した．Lexeed に再帰的展開を適用することにより，語義文を拡張した拡張語義文を生成することができ，拡張語義文中の語義の出現確率を計算できる．

次に，単語の上位語は定義文中に現れる傾向があるという仮説に基づき, 上位語と拡張語義文中の語義との関係を，出現頻度の観点から調べてみた．ここでは，上位語の正解データとして，Lexeed 定義文の HPSG 構文解析による主辞を仮の上位語とし，これを人手で修正したものを用いた．多義性を考慮しない場合の拡張定義文と上位語の相関については既に報告されているが ((鈴木敏 2009)，今回は多義性を考慮した拡張語義文と上位語との関係について検証する．

図 1 は，拡張語義文中での語義の出現率と上位語との関係を表す図である．横軸は語義の出現頻度を高い順に並べたときの順位を，縦軸は上位語正解データとの一致度を表している．例えば図からは，拡張語義文中での出現頻度が最も高い語義は約 22%の割合で上位語と一致することがわかる．拡張語義文中の出現頻度が高い程，上位語に一致する確率が高くなっていることが図から明らかである．従って，この結果から，拡張語義文中の出現頻度と上位語との間に相関があることが読み取れる．即ち，拡張語義文中での出現頻度が高い程，その語義は上位語である可能性が高いということになる．

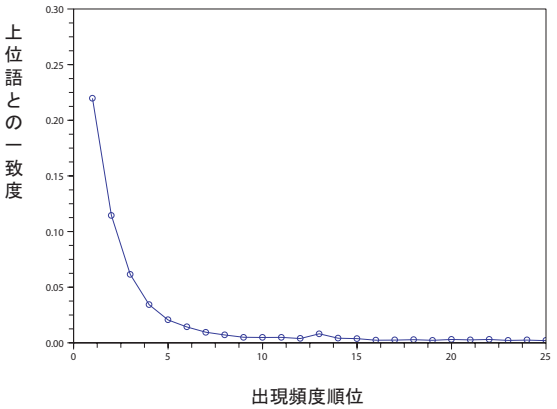

図 1: 拡張語義文中の語義出現頻度と上位語との相関

これにより，拡張語義文中の語義の確率頻度 $P(w_j^*|w_i)$ は，語義 w_j^* が語義 w_i の上位語になるための成り易さの指標として利用できることがわかった．同様に確率頻度行列 C 全体を考えると，これら語義間の関係は上位語を示すオントロジーとみなすことができ，全体としては双方向に結合を持つネットワーク型オントロジーとして扱うことができる．

4　ネットワーク型オントロジーと語義曖昧性解消

本手法では，近傍の単語と共通の上位語が多い程，意味的な距離は近くなるので，意味的に近い語義が近傍に集まるはずであるとして語義曖昧性解消をおこなう．木構造型オントロジーの場合には上位語を探すには木構造を上にたどれば良いが，ネットワーク型オントロジーでは上位語候補はほぼ全単語 (語義) に及ぶことになり，上位語に成り易さを計算することにより，共通の上位語を確率的に求めることができる．

語義曖昧性解消の具体的な計算方法は以下の通りである．

前述の確率頻度を用いて，単語間の意味的関連性を以下のように定義する．ただし，見出語 w の拡張語義文中の単語 w^* の出現確率を $P(w^*|w)$ と表す．

二つの単語 w_i, w_j があるとき，これらの単語が同時に文中に現れる確率は，

$$\begin{aligned}
&P(w_i, w_j) \\
&= \sum_{w^*} P(w^*, w_i, w_j) \\
&= \sum_{w^*} P(w_i|w^*)P(w_j|w^*)P(w^*) \\
&= \sum_{w^*} \frac{P(w^*|w_i)P(w_i)}{\sum_w P(w^*|w)P(w)} \frac{P(w^*|w_j)P(w_j)}{\sum_w P(w^*|w)P(w)} P(w^*) \\
&= \sum_{w^*} \frac{P(w^*|w_i)P(w^*|w_j)}{\sum_w P(w^*|w)P(w)} P(w_i)P(w_j)
\end{aligned} \tag{5}$$

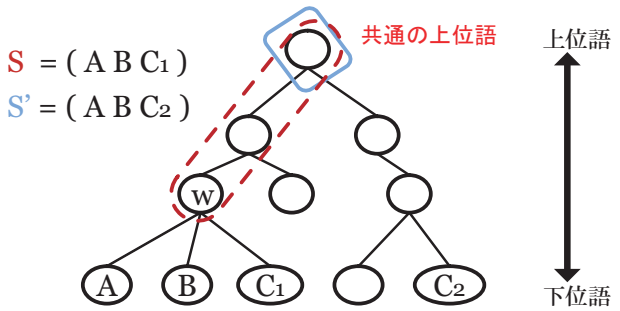

図 2: 共通の上位語の例 (木構造型オントロジーの場合)

であるとする．ここで，w は見出語であり，w^* は語義文中に出てくる単語 (語義) である．

図 1 によれば，$P(w^*|w)$ が大きい程，語義 w^* は見出語 w の上位語になりやすい．従って，式 (5) の値が大きい程，単語 w_i, w_j は共通の上位語 w^* を多く持っていることになる．即ち，単語の意味的な上下関係を表す木構造型オントロジーがあるとすれば，式 (5) の値が大きい程，単語 w_i, w_j はオントロジー上で近い位置に存在していることを意味する．

図 2 はこれを表したもので，多義語 C を含む文 (ABC) がある場合，C の語義を C_1 とするか C_2 とするかで共通の上位語は異なる．式 (5) の値が大きくなるのは，図で C_1 を選んだ時のように，共通の上位語が多く存在する場合である．ただし，提案手法ではネットワーク型オントロジーを利用するため，木構造型オントロジーのように，決定論的に上位語を取り出すのではなく，確率的に取り出すことになる．

式 (5) を拡張し，n 個の単語 $w_i, w_j, \cdots, w_k$ が同時に文中に現れる確率を，

$$
\begin{aligned}
& P(w_i, w_j, \cdots, w_k) \\
&= \sum_{w^*} P(w^*, w_i, w_j, \cdots, w_k) \\
&= \sum_{w^*} P(w_i|w^*, w_j, \cdots, w_k) P(w^*, w_j, \cdots, w_k) \\
&= \sum_{w^*} P(w_i|w^*) P(w^*, w_k, \cdots, w_k) \\
&= \sum_{w^*} P(w_i|w^*) P(w_j|w^*) \cdots P(w_k|w^*) P(w^*) \\
&= \sum_{w^*} \frac{P(w^*|w_i)P(w_i)}{\sum_w P(w^*|w)P(w)} \frac{P(w^*|w_j)P(w_j)}{\sum_w P(w^*|w)P(w)} \\
&\quad \cdots \frac{P(w^*|w_k)P(w_k)}{\sum_w P(w^*|w)P(w)} P(w^*) \\
&= \sum_{w^*} \frac{P(w^*|w_i)P(w^*|w_j)\cdots P(w^*|w_k)}{\left(\sum_w P(w^*|w)P(w)\right)^{n-1}} \\
&\quad \times P(w_i)P(w_j)\cdots P(w_k) \qquad (6)
\end{aligned}
$$

であるとする．ここで，見出語の事前生起確率 $P(w)$ を一定とすれば，式 (6) は

$$
\begin{aligned}
& P(w_i, w_j, \cdots, w_k) \\
&\propto \sum_{w^*} \frac{P(w^*|w_i)P(w^*|w_j)\cdots P(w^*|w_k)}{\left(\sum_w P(w^*|w)\right)^{n-1}} \qquad (7)
\end{aligned}
$$

表 1: 曖昧語義推定結果

正誤	曖昧度	単語	ID
	1	官邸	00020140
	1	内閣	00024423
	1	記者	00011765
	1	会見	00001315
	1	民主	00012770
	1	連合	00010768
	1	所属	00015344
	1	議員	00012023
o	4	問題	00026221
	1	政権	00009729
	1	影響	00015281
o	3	見通し	00017341

表 2: 語義候補と計算値

単語	ID	値
問題	6838	6.37313e-49
	15904	1.00014e-48
	19170	7.3893e-49
	26221	9.13399e-48
見通し	9857	2.57811e-48
	12148	2.09646e-48
	17341	9.13399e-48

と書ける．

これを用いれば，文中の単語が $w_i, w_j, \cdots, w_k$ であるような文に対して，式 (6) あるいは式 (7) の値が最大になるような語義の組合せを計算すれば，文中の各単語の語義を同時に推定できる．

5 計算機実験

上記の手法を実際のテキストに適用した結果を示す．本稿では名詞のみを対象に語義曖昧性解消を行った．適用の手順は次のとおりである．

まず，Lexeed の名詞に対して，語義ごとに再帰的展開手法を適用した．見出語約 2.2 万語に対し，語義は約 3.4 万語義であった．解析対象の入力文は，まず始めに形態素解析器 (茶筌)(松本，北内，山下，平野，松田，高岡，浅原 2000) により単語分割され，続いて，分割された単語の内，Lexeed に載っている名詞に対して式 (7) を適用し，多義語の語義を推定した．

表 1,2 に式 (7) を用いて文中の各単語の語義を同時に推定した結果の一例を示す．入力として用いた文を以下に示す．

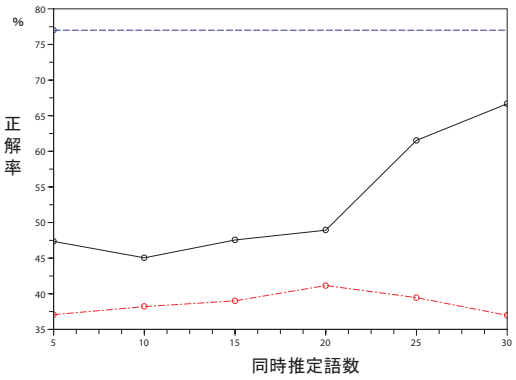

図 3: 語義曖昧性解消の結果と Senseval2 日本語タスクでの結果の比較

表 3: 同時推定語数 (上段) とサンプル数 (下段)

1-5	6-10	11-15	16-20	21-25	26-30
490	1565	715	139	26	21

村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し，社会党の新民主連合所属議員の離党問題について「政権に影響を及ぼすことにはならない．離党者がいても，その範囲にとどまると思う」と述べ，大量離党には至らないとの見通しを示した．

表 1 は語義推定の結果であり，各コラムは左から順に「正解／不正解」「語義曖昧度」「単語表記」「語義 ID」を表している．この例では，曖昧性のある単語は「問題」と「見通し」の二つのみであった．これらの単語はそれぞれ 4 語義，3 語義の曖昧性があり，この文は全部で 12 通りの語義の組合せがあることがわかる．

表 2 は曖昧性のある単語の語義候補とその語義を用いたときの式 (7) の値である．実際の計算では，12 の組合せ全てを計算しているが，表に示しているのは当該語義を用いたときの最大値のみである．式 (7) の値の最大値は $9.13399e-48$ であり，このときの語義の組合せ (語義 ID:26221,17341) が推定された語義である．用いる語義により，値に差が表れており，上記の手法が機能していることがわかる．

この例題は Senseval2 日本語辞書タスク (白井清昭 2003) の学習データの一部であり，付与されている正解語義により，正解／不正解の判断が可能である[1]．

さらに，語義曖昧性解消の性能を評価するため，Senseval2 日本語辞書タスクの結果と比較した．比較結果を図 3 に示す．図の縦軸は正解率，横軸は同時推定した単語数である．実線が提案手法の結果，点線が語義を任意に選んだときの正解率の期待値，最上部の破線は Senseval2 日本語辞書タスクでもっとも成績がよかった手法 (村田真樹，内山将夫，内元清貴，馬，井佐原，白井清昭 2003) の値である．また，表 3 は同時推定の単語数と評価サンプル数の対応を表している．結果は，学習を用いた手法には及ばないものの，任意に選ぶ場合の期待値を明らかに上回る正解率を出すものであった．これにより，学習を用いない場合でも，高い精度での語義曖昧性推定が可能であることが示されたことになる．

提案手法が有効に働く理由の一つとして，ネットワーク型オントロジーによるスパースネスの解消が挙げられる．例えば，木構造型オントロジーである日本語語彙大系を学習型語義曖昧性解消手法に組み込んだ手法も提案されているが (藤田早苗，Francis，藤野昭典 2008)，この場合，単語間の関係を十分に取り出すためには木構造の上位階層を使わざるを得ず，詳細な意味関係を取り込むことは難しい．一方，提案手法では，ネットワーク型オントロジーを用いており，単語間の関係のスパースネスが解消され，情報量の小さい意味関係までも利用できているものと考えられる．

6 おわりに

本稿では，学習を用いず，辞書の再帰的展開による意味的な単語間の距離を用いることで語義曖昧性解消を行った．提案した手法は，学習を用いた手法の精度には及ばないものの，語義を任意に選ぶ場合に比べて明らかに高い精度を示した．

従って，本手法は単独で語義曖昧性解消を行うための手法考えてもよいが，学習型語義曖昧性解消手法へ組み込むためのパーツとしても意義があると考えられる．或いは，学習用正解データ作成のための補助ツールとして利用することを考えてもよい．

一方で，本手法は単語オントロジーの応用例としての意味もあり，ネットワーク型オントロジーの有用性を示すものでもある．

参考文献

村田真樹，内山将夫，内元清貴，馬青，井佐原均，白井清昭 (2003)．“SENSEVAL-2J 辞書タスクでの CRL の取り組み - 日本語単語多義性解消における種々の機械学習手法と素性の比較．” 自然言語処理, **10** (3), pp. 115–134.

藤田早苗，FrancisBond，藤野昭典 (2008)．“上位意味クラス推定を用いた語義曖昧性解消．” 言語処理学会第 14 回年次大会 (NLP2008), pp. 568–571 Tokyo. 言語処理学会.

白井清昭 (2003)．“SENSEVAL-2 日本語辞書タスク．” 自然言語処理, **10** (3), pp. 3–24.

鈴木敏 (2005)．“辞書に基づく単語の再帰的語義展開．” 情報処理学会論文誌, **46** (2), pp. 624–630.

松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸 (2000)．“日本語形態素解析システム『茶筌』version 2.2.1 使用説明書．” *http://chasen.aist-nara.ac.jp/*.

笠原要，佐藤浩史，田中貴秋，藤田早苗，金杉友子，天野成昭 (2004)．“「基本語意味データベース:Lexeed」の構築 (辞書, コーパス)．” 情報処理学会研究報告. 自然言語処理研究会報告, **2004** (1), pp. 75–82.

鈴木敏 (2009)．“辞書からの上位語情報抽出とオントロジー自動生成．” 自然言語処理, (accepted).

[1] Senseval2 では岩波国語辞典の語義に基づいて語義を付与しているため，Lexeed の語義とは完全に一致するわけではないが，Lexeed には岩波国語辞典への語義タグが人手により付与されているため，語義の判断が可能となっている．