自 然 言 語 処 理 109－5
（1995．9．14）

コーパスから二重確率的シソーラスを自動的に獲得する手法の提案

本田 岳夫, 奥村 学
Email:{honda,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]
本稿では、シソーラスをコーパスから自動的に構築する手法を提案する。本手法のモデルは、確率的分類の手法と確率文脈自由文法の学習の知見を元に、シソーラスのカテゴリ間が決定的ではなく、確率的に結合しているモデルである。実際の解析にこのシソーラスを適用する場合、状況に応じてシソーラスの形態を変化させることが可能となる。
[キーワード] コーパス, シソーラス, 確率文脈自由文法

Automatomatic Construction of the Hidden Hierarchical Thesaurus from Corpora

HONDA Takeo,OKUMURA Manabu
School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract
This paper proposes a method to construct hidden hierarchical thesauri from cooccurrence relation in corpora. The proposed model is a model that each concept class connects other classes probabilistically. When this thesaurus is applied to NLP, the hierarchical structure is determined as evidence demands.
Key Words corpus, thesaurus, stochastic context free grammar

1 はじめに

多くの自然言語処理システムでは、シソーラス (概念階層) が有用な言語情報の一つとして用いられている。例えば、次のようなものが挙げられる。

- 格フレームの選択制限に用いられる。
- 事例ベースの枠組では、不十分な事例の補償に用いられる。
- 語彙的結束性を計算する指標として用いられる [3, 4]。

従来の自然言語処理でよく用いられるシソーラスに、*Roget's International Thesaurus*、分類語彙表、*WordNet*などがある。これらは人手で作られ、機械可読な形式で手に入り、カバレジも広い。しかし、これらのシソーラスはもともと人間のために作られており、計算機による自然言語処理には必ずしも適切であるとはいえず、単語の分類が荒すぎたり、単語間の十分な区別を提供していない場合がある。これらの理由として、次のようなことが挙げられる。

- これらのシソーラスが広いカバレジを目的としており、特定のドメインを狙っていない。
- 人手で構築したシソーラスは、辞書編集者の直観で語彙分類を行なっており、分類の基準が常に明らかであるわけではない。

また、シソーラスを人手で構築することは高コストであり、近年、コーパスから自動的にシソーラスを構築することが注目されてきている。

これらの手法は、

1. 共起データ (頻度; 確率) をコーパスから取り出し、
2. 共起データに基づいて単語間の類似度 (距離) を計算し、
3. 単語のクラスタ (概念ノード) を作る

というステップを踏む。取り出す共起データは、動詞–名詞の共起データであることが多い。類似度計算のステップでは、相互情報量を用いる手法 [11]、単語があるクラスタに属する条件つき確率を用いる手法 [7] などがある。クラスタを作る段階でこれらの手法は、一度分類した単語に関して再分類を行なっていないため、より適切なクラスタがあると思われる単語が、異なったクラスタに分類される場合がある。Pereira[5] らは、単語がクラスに属することをクラスに属する確率分布で表す "soft cluster" を用いたクラスタリング手法を提案している。この手法のクラスタは、各単語 w に対するメンバーシップ確率 $p(c|w)$ として定義される。各単語は確率的にクラスに属すため、不適切な分類に決定されることはない。しかし、この手法でも、クラス間のリンクは静的である。

つぎに、既存のシソーラスを用いた解析の問題点を、動詞格フレームを用いた意味解析を例に説明する。動詞の格フレームは統語的、意味的曖昧性解消や、機械翻訳における訳語選択などに利用され、自然言語処理における基本的な情報である。選択制限は、意味的妥当性を規定するため、辞書中に記述された単語間の共起関係に対する意味的制約である。名詞と動詞の共起関係に対しては、動詞の格フレームの格スロットを満たす名詞がどのような意味的性質を持つかを記述する。例えば、「食べる」という動作の行為者は '生き物' でなければならない。意味指標 (semantic marker, 意味マーカ) やシソーラス中の概念名 (クラスタ位置) を用いることが多い。しかし、格フレームを人手でかき分けることには限界があり、自動的に格フレームを獲得する手法の確立が必要となってくる。近年、コーパスと、シソーラスを用いて動詞格フレームの選択制限を実現しようとする試みが多く行なわれている [1, 6, 9, 8, 10]。しかし、data sparseness の問題、反例の獲得が不十分であるために過般化が起こるなど

の問題がある。また、例えば、「ワープロで論文を書く」「鉛筆で論文を書く」の場合の、「ワープロ」と「鉛筆」は、「書くための道具」という視点で類似しているが、「ワープロを電器店で買う」「テレビを電器店で買う」の場合には、「電化製品」という視点で類似している。この時、「書くための道具」や、「電化製品」といった視点が欠落していると、解析精度が下がってしまうという問題もある。

本研究では、上記の問題点を踏まえ、次のようなシソーラスを提案する。

1. 単語が属すクラスタが確率的であるだけでなく、クラスタ間のアークも確率的なシソーラス (二重確率的シソーラス)
2. 階層 (クラスタ) 間のアークが決定的でなく確率的で、状況 (証拠) が得られることに応じてシソーラスが変化するようなモデル

このモデルでは、格フレームの意味マーカをシソーラス上のクラスタを振る時、クラスタ間のアークが確率的であることで、その格フレームが持つ視点を表現できる可能性がある。また、ある単語が属するシソーラス上のクラスタがその単語の持つ意味であると見なせるので、単語の意味属性をあらかじめ振る必要がなくなる。あるいは、あらかじめ振られた意味属性の詳細化とまとめ上げの尺度が与えられる可能性がある。

2 二重確率的シソーラスの獲得

二重確率的シソーラスを獲得する手続きについて説明する。本手法は、確率文脈自由文法の推定アルゴリズムである、Inside-Outside アルゴリズム [2] を元にしている。Inside-Outside アルゴリズムは、最尤推定法の一つである EM(Expectation Maximization) アルゴリズムの一つである。確率文脈自由文法では、文法規則に付けられた確率 (規則確率) を推定する。導出過程である文法規則が用いられる頻度の期待値を推定するのに現在の規則確率と訓練集合を用い、これらの期待値の適当な比として新しい規則確率を計算する。本手法では、文法カテゴリをシソーラスの概念クラスと見ることで、概念クラスがシソーラス上に現れ、下位の概念クラスにわかれる確率を計算する。また、クラスがシソーラス上に存在する期待値を推定するのに語彙の頻度情報と、現在の規則確率から計算する。

2.1 記号の定義

シソーラスの概念クラスの集合を $C = \{c_1, \ldots, c_{NT}\}$、語彙の集合を $D = \{d_1, \ldots, d_T\}$ とする。それぞれの語彙は、特徴ベクトルであり、例えば、コーパス中で名詞がある動詞にかかる頻度を持つ。

クラス c_i がクラス c_j, c_k からなる確率を A_{ijk}、クラス c_i が $\{d_m\}$ からなる確率を B_{im} とする。このとき、

$$\sum_{j,k} A_{ijk} + \sum_{m} B_{im} = 1 \tag{1}$$

を満たす。A_{ijk}, B_{im} の初期値を与えるには語彙の部分集合 D_s が与えられた時、そこに任意のクラス c がある確率 $p(c|D_s)$[7] を用いて学習前処理を行なう (2.2節)。

次に内側確率と外側確率を定義する。内側確率 $e(c_i|D_s)$ は、クラス c_i がデータの部分集合 D_s を持つ確率であり、外側確率 $f(c_i|D_s)$ は、クラス c_i がデータの部分集合 D_s を持たない確率である。

2.2 学習前処理

まず、$p(c|D_s)$ から内側確率 $e(c_i|D_s)$ を計算する。$p(c|D_s)$ は概念クラスの名前が任意なので $e(c_i|D_s)$ のそれぞれに $1/NT$ を与える。

$$e(c_i|D_s) = \frac{1}{NT} p(c|D_s) \tag{2}$$

外側確率は、c_pが開始記号である確率を初期値とし、内側確率と、より広いスパンの外側確率から再帰的に計算できる。

$$f(c_0|D) = 1 \tag{3}$$

$$f(c_i|D_s) = \sum_{jk}\sum f(c_j|D_t)A_{jik}e(c_k|D_u) \tag{4}$$

ここで、

$$D_s \cap D_u = \emptyset, D_s \cup D_u = D_t$$

である。また、入力が系列ではないため、確率文脈自由文法とは異なり、右側部分木と左側部分木の順序を区別する必要がない。

内側確率と外側確率が得られると、A_{ijk}, B_{im}は (5)(6) を用いて推定できる。

$$\hat{A}_{ijk} = \frac{\sum_{D_s \subset D}\sum A_{ijk}e(c_j|D_t)e(c_k|D_u)f(c_i|D_s)}{\sum_{D_s \subset D} e(c_i|D_s)f(c_i|D_s)} \tag{5}$$

$$\hat{B}_{im} = \frac{\sum_{D_s=\{d_m\}} e(c_i|D_s)f(c_i|D_s)}{\sum_{D_s \subset D} e(c_i|D_s)f(c_i|D_s)} \tag{6}$$

分母はシソーラス中にクラス c_iが出現する期待値である。

ここで、$p(c|D)$ は、

$$p(c|D) = \prod_{d_i \in c} p(c|d_i) \tag{7}$$

$$= \prod_{d_i \in c} \frac{p(d_i|c)p(c)}{p(d_i)} \tag{8}$$

$p(d_i|c)$ は、

$$p(d_i|c) = \sum p(d_i|c, V = v)p(V = v|c) \tag{9}$$

c, d_iが独立と仮定すると、

$$p(d_i|c) = \sum p(d_i|V = v)p(V = v|c) \tag{10}$$

ベイズ規則を適用して、

$$p(d_i|c) = p(d_i)\sum \frac{p(V = v|d_i)p(V = v|c)}{p(V = v)} \tag{11}$$

ここで、$p(V = v|d), p(V = v|c), p(V = v)$ は次の通りである。

- $p(V = v|d)$: d と共起する動詞 v の相対頻度
- $p(V = v|c)$: c 中の名詞と共起する動詞 v の相対頻度
- $p(V = v)$: データ全体で v が出現する相対頻度

よって、$p(c|D)$ は、

$$p(c|D) = \prod_{d_i \in c}\sum \frac{p(V = v|d_i)p(V = v|c)}{p(V = v)}p(c) \tag{12}$$

$p(c)$ は確率の制約を満たすための定数であるとみなせば、(5)(6) で A_{ijk}, B_{im}を推定する時、これらの確率の制約を満たしているので、ここでは無視できる。

2.3 本学習

A_{ijk}, B_{im}の初期値が決まったら、続いて本学習には入る。本学習は、学習前処理とほぼ同様だが、内側確率の定義が異なる。内側確率は、B_{im}を初期値とし、より狭いスパンの内側確率から再帰的に計算する。

$$e(c_i|D_m) = B_{im} \quad (D_m = \{d_m\} \text{ のとき}) \tag{13}$$

$$e(c_i|D_s) = \sum_{jk} \sum A_{ijk} e(c_j|D_t) e(c_k|D_u) \tag{14}$$

外側確率の計算は、学習前処理の時の計算 (3),(4) と同じである。A_{ijk}, B_{im}の推定も学習前処理と同様、(5,6) で行なう。

3 実際の利用法について

2節で獲得された二重確率的シソーラスは、木構造に関しても確率的である。このシソーラスに何も情報が与えられない時には、Viterbi アルゴリズムの拡張などにより最尤な木を求めることができる。また、あるノードが存在するなどの情報が与えられれば、シソーラスの木構造の尤度が変化するので、情報が与えられた時それぞれに異なる木構造が得られる。

二重確率的シソーラスの利用法には次のようなものが考えられる。

- 格フレームの選択制限として使う

  名詞の動詞との共起頻度のベクトルを葉ノードにして二重確率的シソーラスを獲得する。格フレームの意味マーカとしてシソーラス上のクラスを取り、そのクラスとの確率を計算する。

  例えば、「レコードを聴く」「ラジオを聴く」「テレビを直す」「ラジオを直す」の場合には、シソーラスの構造が確率的であるために、二重確率的シソーラスでは、「レコード」と「ラジオ」、「テレビ」と「ラジオ」が同じクラスになるシソーラスの構造があり、それぞれ適切なクラスに意味マーカをはることができる。

- 単語の類似度の計算

  単語の類似度を計算する場合、その使われている文脈に依存して動的なものとして扱う必要がある [12]。

  例えば、「ワープロ」と「ビデオデッキ」、「ワープロ」と「万年筆」の類似度を計算する時には、「テレビ」という語が与えられた時には、「電化製品」という文脈で「ワープロ」と「ビデオデッキ」の類似度が、「鉛筆」が与えられた時では、「書くための道具」という文脈で「ワープロ」と「万年筆」の類似度が高くなる。これらを二重確率的シソーラスを用いて実現するには、「ワープロ」と「テレビ」が与えられたことで、それらの語が近くにあるシソーラスが決定される。そのシソーラス上で「ワープロ」と「ビデオデッキ」の類似度を計算する。

- 語彙的結束性の情報として使う

  談話中のある位置での近傍の単語間の類似度を計算することで、談話の段落分割をする研究が行なわれている [13]。ここでは、談話の段落中の単語を葉ノードにし、その単語の近傍に現れる単語をベクトルにしたものを用いてシソーラスを獲得し、実際に段落分割を行なう時には、対象とする談話中のある範囲の単語の集合 Dを与えて、シソーラスの部分木が生成される確率$\sum e(c_i|D)$ を計算する。この確率が大きければ、単語が木のより葉に近い部分でまとまっていると考えられるので、段落に分割されにくく、確率が小さいものは、意味的に分散していると考えられるので、その近傍で段落に分割されやすいと考えられる。

4 おわりに

本稿では、二重確率的シソーラスをコーパスから自動的に獲得する手法について述べた。このシソーラスは、語彙がクラスに属す確率と、クラスがその上位のクラスに属す確率を持つという点で二重に確率的である。このシソーラスは、確率文脈自由文法の推定アルゴリズムである Inside-Outside アルゴリズムを元に、語彙があるクラスに属す確率の計算を付与したアルゴリズムによって計算される。

現在、本アルゴリズムをインプリメントしている最中である。今後、実際に得られたシソーラスの有効性を、3節で述べたような場面に適用する実験を行なう予定である。

参考文献

[1] Ralph Grishman and John Sterling. Generalizing Automatically Generated Selectional Patterns. In *COLING 94*, pp. 742–747, 1994.

[2] K. Lari and S.J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, Vol. 4, pp. 35–56, 1990.

[3] Jame Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaurual Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, pp. 21–48, 1991.

[4] Manabu Okumura and Takeo Honda. Word sense disambiguation and text segmentation based on lexical cohesion. In *COLING94*, pp. 755–761, 1994.

[5] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional Clustering of English Words. In *ACL 93*, pp. 183–190, 1993.

[6] Philip Resnik. Semantic Classes And Syntactic Ambiguity. In *Human Language Technology*, pp. 278–283, 1993.

[7] Takenobu Tokunaga, Makoto Iwayama, and Hozumi Tanaka. Automatic Thesauri Construction Based on Grammatical Relations. In *IJCAI 95*, 1995. to appear.

[8] 春野雅彦. 最小汎化を用いたコーパスからの動詞格フレーム学習. 「自然言語処理における学習」シンポジウム, pp. 9–17, 1994.

[9] 平岡冠二, 松本裕治. コーパスからの動詞格フレーム獲得と名詞のクラスタリング. 情報処理学会自然言語処理研究会資料, No. 104-11, pp. 79–86, 1994.

[10] 李航. 一般化された実例と確率を用いた曖昧性解消. 情報処理学会自然言語処理研究会資料, No. 98-7, pp. 49–56, 1993.

[11] 柏岡秀紀, Ezra W.Black. 相互情報量を用いた単語の分類手法. 「自然言語処理における学習」シンポジウム, pp. 104–111, 1994.

[12] 小嶋秀樹, 伊藤昭. 意味空間のスケール変換による動的シソーラスの実現. 情報処理学会研究会資料 NL108-13, pp. 81–88, 1995.

[13] 小嶋秀樹, 古郡廷治. 単語の結束性にもとづいてテキストを場面に分割する試み. 情報処理学会研究会資料 NL95-7, pp. 49–56, 1993.