MDL原理を用いたシソーラスの自動学習

李 航　安倍 直樹
RWCP　理論NEC研究室

1 はじめに

近年、コーパスにおける単語間の共起関係をもとにシソーラスを自動的に学習する(単語の自動分類を行なう)研究が非常に盛んであり、数多くの方法が提案されている(例えば、[1][2][3][6][9][11])。本稿では、これら従来方法とは異なる新しいシソーラスの学習方法を提案する。

シソーラスの自動学習には、(a)知識(特に応用、或はドメイン依存の知識)のコンパクトかつ正確な記述、(b)人間がシソーラスを作成する時の労力の削減、(c)人間の作成するシソーラスにおける恣意性の排除等のメリットがあると考えられている。本研究では、特に、シソーラスを学習し、次いで格フレームを学習し、その知識を自然言語処理(曖昧性解消)に利用するという処理の流れの中で(a)、(b)、(c)のメリットを確かめることにする。本研究のシソーラスの学習方法は我々が[5]で提案した格フレームの学習方法と同じ枠組のものである。

本稿では、まずシソーラスの学習問題を確率モデル[1]の推定問題として定式化する。それからMDL原理(記述長最小原理)[7][8]を用いたその確率モデルの推定方法、つまりシソーラスの学習方法を提案する。最後に実験結果について述べる。

2 問題の定式化

単語の共起関係を利用したシソーラスの学習は通常以下の過程からなる。(i)単語の共起関係(例えば、格フレームにおける単語間の関係)をコーパスから抽出する、(ii)単語間の類似度(或は距離)を定義する、(iii)類似度に基づいて単語のクラスタリングを行なう。最終的に(iii)のプロセスを再帰的に繰り返すことによってシソーラスを作成する。一方、シソーラスを用いた格フレームの学習の場合、(i)の過程を経て、シソーラスで定義された単語間の類似度に基づいて格フレームを一般化するのが普通である。ここで、格フレームの一般化とは既知の格フレームの知識から未知の格フレームの知識を推定することをさす。この二つの学習問題は、異なる条件の下での同じ確率モデルの推定問題として定式化することができる。

例えば、動詞とその目的格となる名詞のペアをデータから抽出したとする。本研究の立場では、このような動詞名詞ペアは確率分布(確率モデル) $P_{obj}(v,n)$ によって生成されたものであると仮定する。ここでは、確率変数 v は動詞集合 $\mathcal{V}$ の値をとるとし、n は名詞集合 $\mathcal{N}$ の値をとるとする(一般に任意の共起関係に関してこのような確率モデルを定義することができる)。確率的なアプローチによる自然言語処理では、動詞名詞ペアの出現頻度(共起頻度)を基に $P(v,n)$(或は、条件付き確率分布 $P(n|v)$)の値を推定し、曖昧性解消等を行なうのが一般的である[4] (以下では共起関係を表す添字 obj 等を省略する)。しかし、実際にこのような(単語ベースの)確率モデルを推定しようとすると、データが不足するいわゆる「Data Sparseness 問題」が生じるし、また類似した単語は同じ生成確率をもつとしたほうが知識をコンパクトに記述できる。従って、単語の集合をいくつかのクラスに分類して、それら分類の各々に生成確率を与え、個々の分類中の単語はそれぞれ一様分布により生成されると考えたほうがよい。すなわち、以下のような(クラスベースの)確率モデルを考えるわけである。

$$v \in C_v, n \in C_n, P(v,n) = \frac{1}{|C_v \times C_n|} \cdot P(C_v, C_n) \tag{1}$$

ここで、確率変数 C_n は名詞集合 $\mathcal{N}$ のある分割中の値(部分集合)をとるとし、確率変数 C_v は動詞集合 $\mathcal{V}$ のある分割中の値(部分集合)をとるとする。名詞集合の分割 $\mathbf{\Gamma}_n$ とは、$\mathbf{\Gamma}_n \subseteq 2^{\mathcal{N}}$、$\cup_{C_i \in \mathbf{\Gamma}_n} C_i = \mathcal{N}, \forall C_i, C_j, C_i \in \mathbf{\Gamma}_n \cap C_j = \emptyset$ 満足する集合のことである。動詞集合の分割 $\mathbf{\Gamma}_v$ も同様に定義される。また、以下では、動詞集合、名詞集合の分割の直積 $\mathbf{\Gamma}_v \times \mathbf{\Gamma}_n$ の要素を「クラスタ」と呼び、式1に示される確率値をクラスタの「Centroid 値」とする。また、動詞集合、名詞集合の分割、さらにその直積における各クラスタのCentroid値を決めることを「クラスタリング」と

[1]数理統計学では、確率分布のクラスのことを確率モデルと呼ぶこともある。しかし、本稿では、確率分布のクラス中の個々の分布のことを確率モデルと呼ぶことにする。

よぶことにする。本研究では、シソーラスの学習問題を上の確率モデルを推定する問題として定式化する。具体的には、観測された動詞名詞ペアが式1で表される確率モデルのクラスの中の一つによって生成されたものであると仮定し、複数の確率モデルからそのデータをもっともよく説明できるモデルを選びクラスタリングの結果とし、単語のシソーラスを作成していく。一方、格フレームの学習問題は、シソーラスが与えられた上での、以上の確率モデルの推定問題に定式化することができる(詳しくは[5]を参照されたい)。

近年MDL原理を用いた確率モデルの推定法(MDL推定法)に関する研究が盛んである[2] [7][8]。MDL推定法が多くの良い性質をもつことも明かにされてきている。本研究では、MDL原理を確率モデル推定の基準として採用する。MDL推定法においては、(全)記述長と呼ばれる量が仮説の評価規準として用いられる。歴史的にはMDL原理は情報理論の分野で二段階符号化やデータ圧縮の定式化として考え出されたもので、MDL原理の記述長は通信における符号化の「符号語長」に当たる。全記述長Lは、モデル記述長L_{mod}と(モデル中の)パラメータ記述長L_{par}とデータ記述長L_{dat}の和として計算される。即ち、$L = L_{mod} + L_{par} + L_{dat}$である。MDL原理は、全記述長$L$が最小のモデルが最適なモデルであるとする。直観的にいうと、これは以下のような考えによるものである。$L_{mod}+L_{par}$はモデルの複雑さの指標で、L_{dat}はモデルとデータのフィットの悪さの指標である。普通、モデルが単純であれば($L_{mod} + L_{par}$が小さければ)、モデルとデータのフィットが悪くなる(L_{dat}が大きくなる)。一方、モデルが複雑であれば($L_{mod} + L_{par}$が大きければ)、モデルとデータのフィットがよくなる(L_{dat}が小さくなる)。つまり、モデルの単純さ($L_{mod} + L_{par}$)と、モデルとデータのフィットのよさ(L_{dat})の間には、トレードオフの関係があるのである。このトレードオフを$L = L_{mod} + L_{par} + L_{dat}$の最小化によって解決するのがMDL原理である。実際、この原理に従って選択したモデルが未知のデータをより正確に予測できるという意味で、ほぼ最適なモデルであることが理論的に証明されている。

次に、本研究で扱う問題における全記述長の計算方法について述べる。観測データ(動詞名詞ペア)が与えられ、さらに動詞集合、名詞集合の分割が定まると、特定のクラスタリングが決まり、そのクラスタリングの全記述長を計算することができる。可能な名詞の部分集合の数は$2^{|\mathcal{N}|}$であるので、各々の名詞の部分集合の記述長は$\log_2 2^{|\mathcal{N}|} = |\mathcal{N}|$となる。従って、名詞集合を$k_n$個の部分集合に分割する記述長は$k_n \times |\mathcal{N}|$である。同様に、動詞集合を$k_v$個の部分集合に分割する記述長は$k_v \times |\mathcal{V}|$である。よって、ある特定のクラスタリングのモデル記述長はL_{mod}は、

$$L_{mod} = k_n \times |\mathcal{N}| + k_v \times |\mathcal{V}| \tag{2}$$

として計算される[3]。つまり、動詞集合、名詞集合の分割を伝えるのにL_{mod}だけの符号語長が必要なのである。次に、クラスタリングのパラメータ記述長L_{par}は、

$$L_{par} = \frac{(k_v \times k_n - 1)}{2} \times \log_2 N \tag{3}$$

として計算される。ここで、Nは動詞名詞ペアの総頻度であるとする。L_{par}は$k_v \times k_n - 1$個の自由パラメータを$O(\frac{1}{\sqrt{N}})$の精度で符号化して送るための符号語長である。最尤推定を行う際の標準偏差も$O(\frac{1}{\sqrt{N}})$であることに注目されたい。一方、クラスタリングのデータ記述長L_{dat}は、

$$L_{dat} = -\sum_{i=1}^{k_v}\sum_{j=1}^{k_n} N(i,j) \times \log_2 \hat{P}(i,j) \tag{4}$$

として計算される。ここで、$N(i,j)$はi番目の動詞部分集合とj番目の名詞部分集合によるクラスタ(i,j)における動詞名詞ペアのデータ中の出現頻度で、$\hat{P}(i,j)$はそのクラスタ中の動詞名詞ペアの生成確率の最尤推定値であるとする。当然、$N = \sum_i^{k_v}\sum_j^{k_n} N(i,j)$が成り立つ。さらに、$\hat{P}(i,j)$は、

$$\hat{P}(i,j) = \frac{N(i,j)}{n(i,j) \times N} \tag{5}$$

として推定される。ここで、$n(i,j)$はクラスタ(i,j)における動詞名詞ペアの総数であるとする。符号論的に言えば、L_{dat}は$\hat{P}(i,j)$を生成モデルとして観測データを送る時の符号語長である。

[2] 日本語の解説としては[10]を参照されたい。

[3] モデル記述長は推定問題に対する人間の事前知識を反映するものであり、恣意性がある。別の符号化を用いれば記述長の計算式も当然変わってくる。

3 アルゴリズム

名詞集合だけを考えても、それに対する可能な分割の数が $\sum_{i=1}^{|\mathcal{N}|}\sum_{j=1}^{i}\frac{j^{|\mathcal{N}|}\cdot(-1)^{i-j}}{(i-j)!\cdot j!}$ にもなる (動詞集合の場合も同様である) ので、クラスタリングは組合せ最適化問題である。本研究では、組合せ最適化問題の手法としてよく知られる Simulated Annealing を用いてクラスタリングを行なう。クラスタリングは動詞集合、名詞集合の両方の分割によって行なってもよい。しかし、本研究では現在、片方の集合の分割によるクラスタリングだけを行なっている。さらに分割を二分割としている。例えば、名詞集合を分割する場合、$k_v = |\mathcal{V}|$ と $k_n = 1, 2$ としている。次にそのアルゴリズムについて述べる。

1 名詞集合 $\mathcal{N}$ を (任意の) 二つの部分集合に分割する。

2 ランダムに名詞を一つ選び、その名詞の属する部分集合から削除し、もう一方の部分集合に入れる。そして、この名詞の移動前と移動後の二つのクラスタリングの全記述長 L_1, L_2 を計算する。それぞれの全記述長を Annealing のエネルギーとし、その差 $\Delta L = L_2 - L_1$ に着目する。$\Delta L < 0$ であれば、その変換を確定し、そうでなければ、確率 $P = \exp(-\Delta L/T)$ で変換を確定する。ここで、T は Annealing の「温度」と呼ばれるパラメータである。実際の処理では、例えば、T の初期値が 1 であるとし、ステップ 2 を 1000 回実行した後 T の値を 0.9 倍に下げる。ステップ 2 を指定された回数実行しても単語集合の分割が変わらなければ、次へ。

3 ステップ 2 で得られた分割の両部分集合に対し、再帰的に以上の処理を行なう。

以上の手順によって名詞のシソーラスを作成することができる。動詞のシソーラスの学習も同様に行なうことができる。

4 実験結果

Wall Street Journal コーパス 12 万文 (ACL CD-ROM 1) から、動詞とその目的格となる名詞の共起データをパターンマッチングの手法を用いて抽出した。さらに、共起データを基に、動詞、名詞のシソーラスの自動学習を行なった。図 1に、本手法による自動学習で得られた２０動詞のシソーラスの例を示す。学習できたシソーラスの中に人間の直観と一致しないものも少なくなかった。トレーニングデータの不足がその一因であると思われる。今後は、200 万文の Penn Tree Bank のデータを用いてシソーラスの学習を行なう予定である。さらに、学習できたシソーラスと人間の定義したシソーラスの両方を用いて、格フレームの一般化を行ない [5]、その両者の精度等の比較を行なっていく予定である。シソーラスの他の学習方法との比較も課題である。

謝辞

本研究の機会を与えてくださった NEC C&C 研究所の中村勝洋統括部長、及び藤田友之部長に深く感謝いたします。

参考文献

[1] P. F. Brown, V.J. Della Pietra, P. V. deSouza, J. C. Lai, R.L.Mercer, *Class-Based n-gram Models of Natural Language, Computational Linguistics, Vol.18, No.4, 1992.*

[2] 淵武志, 係り受けデータのみを用いた単語のグルーピング、「自然言語処理における学習」シンポジウム、情報通信学会、日本ソフトウェア科学会合同, *1994.*

[3] D. Hindle, *Noun Classification from Predicate-argument Structures, ACL90, 1990.*

[4] D. Hindle, M. Rooth, *Structural Ambiguity and Lexical Relations, ACL91, 1991.*

[5] 李航, 安倍直樹, シソーラスと *MDL* 原理を用いた格フレームの一般化, 「自然言語処理における学習」シンポジウム、情報通信学会、日本ソフトウェア科学会合同, *1994.*

[6] F. Pereira, N. Tishby, L. Lee, *Distributional Clustering of English Words, ACL92, 1992.*

[7] J., Rissanen, *Universal Coding, Information, Prediction, and Estimation, IEEE Trans. on IT, Vol. IT-30, 1984.*

[8] J., Rissanen, *Stochastic Compexity and Modeling, The Annals of Statistics, Vol.14, No.3, 1986.*

[9] 田本真詞, 川端豪, 漸次的精緻化を用いた単語共起のクラスタリング、「自然言語処理における学習」シンポジウム、情報通信学会、日本ソフトウェア科学会合同, *1994.*

[10] 山西健司, 韓太舜, *MDL* 入門：情報理論の立場から, 人工知能学会誌, *Vol.7, No.3, 1992.*

[11] T. Tokunaga, M. Iwayama, H. Tanaka, *Automatic Thesaurus Construction Based on Grammatical Relations, 95TR-0002, Tokyo Institute of Technology January 1995.*

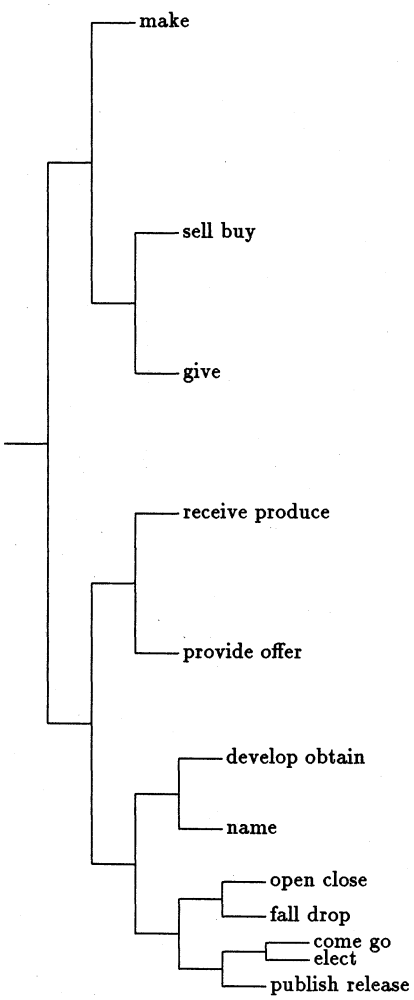

図 1: 学習できた動詞シソーラスの例