Machine Learning of Probabilistic Generative Model Based on Bayesian Nonparametrics

(ベイジアンノンパラメトリクスに基づく確率的生成モデルの機械学習)

東京大学大学院　情報理工学系研究科　数理情報学専攻　佐藤一誠 (学籍番号 : 48-66211)

指導教員　中川裕志

概要

言語学には，「同じ文脈を持つ単語は，同じ意味を持つ傾向にある」という *Distributional Hypothesis* と呼ばれる仮説がある．本研究では，この仮説を基にした単語と文脈の関係をモデル化するために，*Bayesian Nonparametrics* を用いた Sequence と Graph の確率的生成モデルを提案する．単語と文脈の関係をモデル化することで，意味の近い単語をクラスタリングすることができる．*WordNet* を用いた評価実験によって提案モデルの有効性を確認した．

1　はじめに

近年，教師なし学習の分野で，確率的生成モデルの研究が重要な役割を果たしている．確率的生成モデルは，データの生成過程をモデル化することで，データに特有の情報を抽出することができる．例えば，文書の潜在的なトピックなどの情報を抽出することにより，文書のクラスタリングが可能となる．本研究では，確率的生成モデルに対する以下の 2 つの研究を扱う．

1. 対象とするデータ (現象) をいかにモデル化 (表現) するか (モデリングの研究)
2. 既存の，もしくは構築したモデルをいかに学習するか (学習手法の研究)

本研究では，「1. モデリングの研究」として，Sequence と Graph の新たな確率的生成モデルを提案する．特に，*Bayesian Nonparametrics* に基づく確率的生成モデルを構築する．この目的は，言語学における *Distributional Hypothesis* をもとにした言語の意味に関する現象をモデル化することである．*Distributional Hypothesis* とは，「同じ文脈を持つ単語は，同じ意味を持つ傾向にある」というものである．ここで，文脈に関しては，図 1 に示すとおり，2 通りの文脈を考慮する．1 つ目は，単語の文脈として，その単語より n 単語前の単語列 (n-gram) を文脈とするものである．図 1 の例では，$n = 2$ のとき”study”と”fields”の文脈は”the scientific”となる．このような場合は Sequence として扱うことができるので，単語と文脈を扱える Sequence の確率的生成モデルを構築する．2 つ目は，単語の文脈として，構文情報を用いて取り出した，単語の関係から文脈を取り出すものである．図 1 の例では，”study”と”fields”の文脈として”linguistics”が考えられる．本研究では，主語述語構造をコーパス中から抽出し，主語 (名詞) の文脈として述語 (動詞) を，述語 (動詞) の文脈として主語 (名詞) を考える．この関係は，*Disassortative Graph* と呼ばれる構造となるので，このようなグラフ構造の確率的生成モデルを構築する．以上のように，単語とその文脈の関係をモデル化することにより，意味的に近い単語のクラスターを抽出し，*WordNet* を用いた単語の類似度による評価を行う．

さらに，本研究では，「2. 学習手法の研究」として，*Bayesian Nonparametrics* の要素技術の 1 つである *Hierarchical Dirichlet Process* を用いた階層ベイズモデルに対する変分ベイズ法を導出する．ただし，本稿では紙面の都合上割愛する[*1]．本稿の構成は以下の通りである．第 2 章で，Sequence の確率的生成モデルの説明をする．第 3 章で，Graph の確率的生成モデルの説明をする．第 4 章で，まとめを行う．

2　Hierarchical Pitman-Yor Process を用いた Sequence の確率的生成モデル

Distributional Hypothesis を，次のように解釈する：「文脈には意味を示すトピックの分布が存在し，同じトピックに属する単語は意味が似ている」．図 2 を用いて説明する．”is based on”という文脈の後には，複数の単語が観測される．これらの単語は，いくつかの意味のトピックに分けることができる．これをデータ (Sequence) の生成の観点から捉えると，文脈 u は各々固有にトピックの出現分布 G_u を持っている．トピックは，それぞれ意味の近い単語が高い確率で生成されるような単語の出現分布を持っている．この単語分布に従って文脈に続く単語が生成されるとモデル化する．ここで問題なのは，トピックの数が予めわからないことである．このような問題に対し，近年，Bayesian Nonparametrics が有効であることが確認されている．Bayesian Nonparametrics を用いることによりこのような潜在的なトピックの数を推定することができる．本研究では，2006 年に Y.WTeh によって提案された Bayesian Nonparametrics の要素技術の 1 つ Hierarchical Pitman-Yor Process(HPY) を用いる．したがって，文脈 u におけるトピックの分布 G_u は，以下のように定式化される．

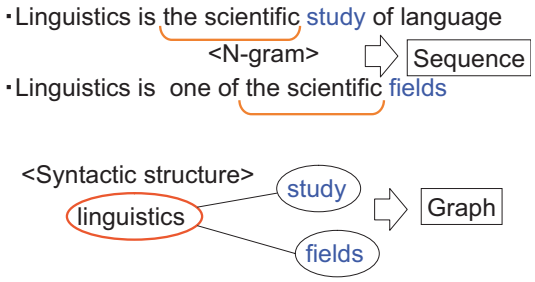

図 1　文脈の例

$$G_u \sim PY(d_{|u|}, \alpha_{|u|}, G_{\pi(u)}) \tag{1}$$

$PY(\cdot)$ は，Pitman-Yor process を示す．$d_{|u|}$，$\alpha_{|u|}$ は，Pitman-Yor process のパラメータでそれぞれ，ディスカウントパラメータ，集中度パラメータと呼ばれており，文脈 u に固有の値を持つ．$G_{\pi(u)}$ は，G_u の基底測度である．ここで，$\pi(u)$ は，u の先頭の単語を削除した単語列である．例えば π(is based on) = based on となる．つまり，G_u は，$\pi(u)$ に依存した分布となっており，それぞれ再帰的・階層的にモデル化されている．図 3 に示すように，これは文脈によって構成された Suffix Tree の各ノードに，トピックの分布が存在するようなモデル化を行っており，各ノードのトピックの分布は，親ノードのトピックの分布を基に生成されている．このような文脈 u のトピックの分布 G_u から，文脈 u に続く単語の w が生成される:$w \sim G_u$.

本研究では，コーパス中のすべての単語のトピックを MAP 推定し，単語のクラスタリングをおこなった．同じ単語でも，文脈に応じてトピックが異なるため，このクラスタリングはソフトクラスタリングになっている．この単語のクラスタリングを *WordNet* を用いた単語類似度の計算手法によって評価を行った．コーパスは *Reuters* を用いた．

評価指標は，単語毎に，*WordNet* における類似度の高い単語が同じクラスタにあるか、逆に、類似度の低い単語と同じクラスタに入っていないかを考慮して accuracy を計算した[*2]．いくつかのベースライン[*3]の中で最高の accuracy は，53.265% であったが，本提案モデルでは，60.055% とベースラインを上回る結果であった．

[*1] 修士論文を参照

[*2] 計算方法は修士論文を参照

[*3] 具体的なベースラインは修士論文を参照

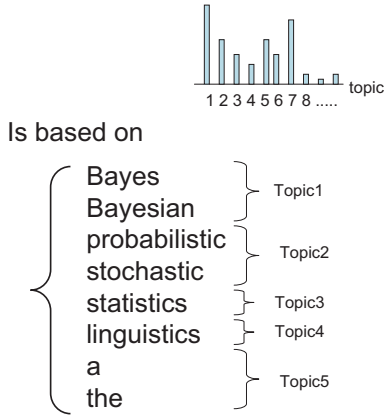

図 2　トピックを介した文脈と単語の例

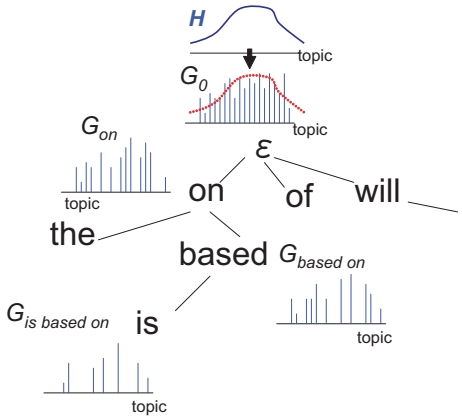

図 3　各ノードにトピック分布を持つ文脈の Suffix Tree

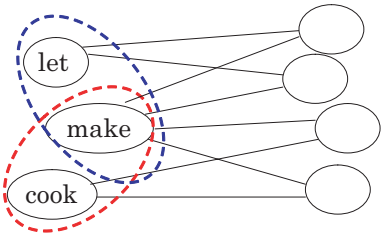

図 4　ノードの多重クラスタリングの例

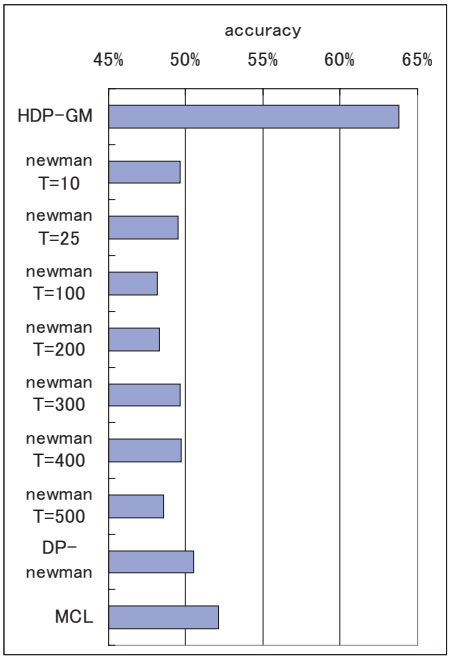

図 5　WordNet 類似度を用いた評価

3　Hierarchical Dirichlet Process を用いた Graph の確率的生成モデル

主語述語構造をコーパス中から抽出し，主語 (名詞) の文脈として述語 (動詞) を，述語 (動詞) の文脈として主語 (名詞) を考える．この関係は，*Disassortative Graph* と呼ばれる構造となるので，このようなグラフ構造の確率的生成モデルを構築する．

グラフの確率的生成モデルを用いることにより，グラフのノードのクラスタリングを行うことができる．特に，本研究におけるノードのクラスタリングは，複数のクラスに属することを許すソフトクラスタリングになっている．この理由は，単語が複数の意味を持っており，複数のクラスに属しているものと考えられるからである．例えば，図 4 のように，let, make, cook という単語ノードをクラスタリングする場合，排他的なクラスタリングだと，{{let, make}, cook}},{{let}, {make, cook}}, {{let, make, cook}} などが考えられるが，これらは好ましくない．むしろ {{let, make },{make, cook }} のような多重性を考慮したクラスタリングが有用であると考えられる．

グラフの生成過程を以下のように考える．U を一様分布とする．各ノードがリンクを貼る確率測度 G_0 は Pitman-Yor Process に従うとする：$G_0 \sim PY(\gamma, d, U)$．$G_0$ からリンクを貼るノードをサンプリングする：$v \sim G_0$．H を Dirichlet 分布とする．各ノード v は，複数のクラスを保持しており，各クラスの出現確率の確率測度を G_v とし．Dirichlet Process に従うものとする: $G_v \sim DP(\alpha, H)$．トピックは，ノードへのリンクを張る確率測度により特徴付けられている．このクラスの確率測度に応じて，別のノード i にリンクを張る:$l_i \sim G_v$．これは，リンク毎に v のクラスが存在すると考えられる．各単語ノード毎に (複数の) クラスを推定し，単語ノードのクラスタリングを行った．

この単語ノードのクラスタリングを，第 2 章と同様に *WordNet* を用いた単語類似度の計算手法によって評価を行った．コーパスは *Reuters* を用いた．図 5 は，いくつかの従来のグラフクラスタリング手法とクラスタリング結果を比較したものである．**HDP-GM** は提案手法である．**newman** は，2007 年に Newman によって提案されたグラフの確率的生成モデルで，T はクラス数を示す．Newman のモデルは，クラス数を予め指定しなければならないという問題があるが，2007 年に桑田らによって *Dirichlet Process* を用いてこのような問題が解決された (**DP-newman** とする)．**MCL(Markov CLuster algorithm)** は，2000 年に，Dongen によって提案されたグラフクラスタリングアルゴリズムである．他の手法と比べ，提案モデルによるクラスタリングの効果が確認できる．

4　まとめ

本研究では，確率的生成モデルの研究として，*Bayesian Nonparametrics* を用いた Sequence と Graph の生成モデルを提案した．さらに，*Hierarchical Dirichlet Process* を用いた階層ベイズモデルに対する変分ベイズ法を導出した．特に本稿では，Sequence と Graph の生成モデルについて説明し，単語クラスタリングタスクにおいてその有効性を示した．