汎用的データにおける
確率的言語モデルの抽出とその利用
―データマイニングの新たな可能性―

1.　背景

現在情報社会の進展とともに膨大なデータが生成される一方で、その膨大なデータを解析し、真に有益な情報を得るデータマイニング技術が多くの分野で必要とされている。しかし、従来の解析では、データのフォーマット、形式、意味などを把握しなければならないため、汎用的な解析手法を構成するのは難しい。また、教師付き学習データを用いてそのような情報を得る場合にも、教師付き学習データを作成するコスト、学習データには存在しない情報がテストデータに現れた場合などを考慮しなければならない。こうした問題を解決するため、あらゆるデータに対し、柔軟に解析が行えるように、前提となる知識なしにデータ自体の特徴のみから解析を行う手法が必要である。

2.　目的

データの種類、フォーマットによらずにデータ解析を行えるように、データ自体から得られる情報だけを利用して解析するアルゴリズムを開発する。この解析には自然言語処理で用いられていた確率的言語モデルを利用する。これにより、今までは解析をすることが困難だった様々なデータに対しても解析を行うことができるようにする他、従来の辞書を用いた自然言語処理解析においても、従来にはなかった解析を行えるようにする。

3.　開発の内容

本プロジェクトのシステムは主に次の４つのモジュールから成っている。なお(i)については 2002 年度未踏 Youth プロジェクトから継続開発したものである。

(i)　WX 法

WX 法は与えられたデータを最小構成要素（Unit）に分解する。分解する際には、与えられたデータの分布状況の情報のみを用いる。WX 法は Suffix Arrays を用いて全ての部分列を列挙した後にそれぞれに評価値を与え、ME（期待値最大化法）に基づく準最適化法を用いてデータを Unit に分解する。

報道陣代表のカメラマンが公邸正門前まで接近して撮影することを許可、31日午前10時(日本時間1日午前0時)過ぎ、カメラマンが事件発生後初めて公邸正面玄関前で撮影を行った。左翼ゲリラ「トゥパク・アマル革命運動」(MRTA)の指揮者、セルパ容疑者らはカメラマンに、「最終決断するまでは後退しない。政府の交渉態度はごう慢である」などと強硬な姿勢を伝えるメッセージを伝えた。〈2、7面に関連記事〉 セルパ容疑者らはペルーの放送局「CANAL5」のカメラマンに対し「勝利者も敗北者もなく、血の流れることのない解決策が求められる。今まで血は流されていない。一方で、私たちの仲間は服役中であり、その仲間を釈放せよ」と、拡声機で伝えた。 ゲリラ側は31日午前8時半(同31日午後10時半)ごろ、共同通信社に対し公邸内への立ち入りを許可するとの掲示を公邸の窓に出した。張り出された紙には日本語で「共同通信へ　進入可　MRTA」「謹賀新年」と掲示。共同通信記者が公邸に向かったが、公邸前で警備当局に阻止された。「合法政党化」に重点　ゲリラ側、要求を縮小――ペルー日本大使公邸占

Fig1. WX 結果　CD-毎日新聞 2002 年度版

Butsomethoughtfullpersons,whohadseenhim
walkingacrossoneofhisfieldsonacertainDecember morning--
sunnyandexceedinglymild--mighthave
regardedGabrielOakinotheraspectsthanthese.In
hisfaceonemightnoticethatmanyofthehuesand
curvesofyouthhadtarriedontomanhood:thereeven
remainedinhisremotercranniessomerelicsoftheboy.
Hisheightandbreadthwouldhavebeensufficientto
makehispresenceimposing,hadtheybeenexhibited
withdueconsideration.Butthereisawaysomemen
have,ruralandurbanalike,forwhichthemindismore

Fig.2 WX 結果 Calgary Corpus , book1
（Space 除去）

(ii) Class Model

Class Model は与えられた Unit 列の並び方から Unit をグループに分類する。分類基準には、同じ Class に属する Unit は前後の Unit の出現状況が似ていることを用いる。Class Model は前後の Unit の出現状況を調べる範囲によって Bi-gram Class Model（前後１Unit ずつ調べる)、Tri-gram Class Model（前後 2Unit ずつ調べる）に分けられる。さらに本プロジェクトでは近似的に周辺の出現状況を調べることにより計算量、領域量を大幅に削減した Approximate Class Model(AppCM)も開発し、大規模なデータに対するスケーラリビティも備えた。AppCM において計算量は Unit 数、Class 数、それぞれ比例する程度である。

Tab1. Class Model 結果

- **用いたテストデータ：**Calgary Corpus, book1 を WX 法を用いて分解したもの
- **64 個の Class に Bi-gram Class Model で分類した中で三つの Class を抜粋。**

1)（主に動詞からなる Class）

not hear light mean close direct like long act seem sent pass present mind than great strange certain should appear reason watch consider live usual good turn breath open clear utter promise natural notice express respect old hope dress walk fell remark wonder strong whisper fall high follow latter change begin trouble faint answer suggest apparent return laugh occasion reach sens perfect enter distinct listen suffer proper general doubt silent human small quiet experi want fear

2) （主に名詞からなる Class）

time hand side way night name thing face other moment voice horse woman door day mistress person tone eye place arm morning thought matter back husband gate fire house farm church ground sound head eyes position world farmer manner point field own life shepherd evening wind waggon yard money case shape year window friend spirit words colour water flock fair road feeling outside hill girl right stone power grave mouth foot subject corner

3)（主に数詞からなる Class）

1 6 5 7 0 9 ? 8 20 4 43 40 18 24 44 41 27 25 45 42 21 38 29 36 11 23 31 28 33 35 30 26 39 37 34 22 32

(iii) Trigger Model

離れている Unit 間の相関情報を調べ、特に有意な今日関係にある Unit のペアを抽出する。抽出は Trigger を用いたことによる対数尤度の変化を計算し、Unit の遠距離の共起、依存関係の有意度を検出する。Trigger 対の抽出には Unit 同士だけではなく、Class 間も調べることができる。計算量は Unit 数の二乗に比例する。

Tab2. Trigger Model 結果

1	日本大使公邸占拠	ペルー
2	ＮＴＴ	電話
3	ペルー	トゥパク・アマル
4	北京	香港
5	セルビア	ベオグラード
6	行政改革	行革
7	第	回
8	駅	鉄道
9	ヘブロン	イスラエル
10	要求	人質

CD 毎日新聞 2002 年度 1 月分記事を WX 法によって分解したものを Trigger Model によって求められた共起関係の有意度が高い順に表示。但し同じ単語の対からなるペアは除く

(iv) WXC compression

WX 法、ClassModel を用いて、データを圧縮する。従来のデータ圧縮法に比べ、本プロジェクトで開発した圧縮アルゴリズム（WXC）は、復元時の処理量、領域量が少ない、また任意の位置から復元ができるランダム復元の性質を有しつつ圧縮率を維持できるなど、従来のデータ圧縮にはなかった性質を持っている。

Tab3. WXC 法と従来のデータ圧縮法との復元時の特徴の比較（元データが 1MB)

	復元処理	ストリーム復元	ランダム復元	使用メモリ
LZ	コピー（高速）	○	×	数 10kB – 数 100kB
PPM	圧縮と同じ（低速）	○	×	数 MB – 数 10MB
BWT	逆変換（高速）	×	×	数 MB
WXC	コピー（高速）	○	○	数 10kB – 数 100kB

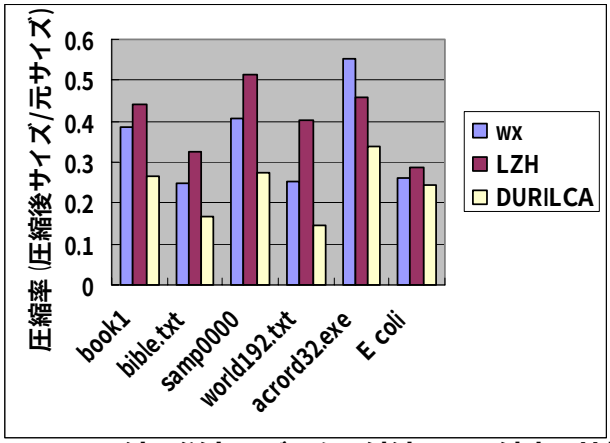

Fig2. WXC 法と従来のデータ圧縮法との圧縮率の比較

wx ： 今回開発したデータ圧縮

LZH： LZ(lh5) + Huffman

DURILCA[3] : PPMⅡ + filter

・book1 (768771) Fiction book (Calgary corpus)

・bible.txt (4047392) The King James version of the bible (The Large Corpus)

・e.coli (4638690) Complete genome of the E. Coli bacterium (The Large Corpus)

・samp0000 (104856) CD 毎日新聞 2002 年度、先頭 1MB

・acrord32.exe(7671876) Acrobat Reader 実行ファイル

4. 従来の技術との相違

今回開発した手法は、あらゆるデータに対応できるという点、つまり、学習情報、辞書などを用いずに解析を行う点が従来手法と異なる点である。

形態素解析などにおいては、教師付き学習と比べると、本手法は精度、計算量ともに劣るが、従来手法では困難であった未知の固有単語分類、解析をはじめとして、本来必要とされる部分の解析を行うことが可能である点が大きな特徴である。例えば、専門用語の種類毎に名詞を分けることも可能である。

また、自然言語情報以外にも適用できるので、XML、HTML といったメタ言語情報、ゲノム情報、ログ情報などもそのまま解析することが可能である。最後に計算量、領域量についてだが、数十 MB 程度のデータであれば、通常のパソコンにおいても数分で処理することが可能である。

5. 期待される効果

本プロジェクトによる効果は、二つ考えている。

一つ目は、自然言語処理、及びデータマイニングにおいて従来の教師付き学習とはまったく違う新しい道が示せたことである。現実的な視点から見ると、本手法がそのまま従来の手法にとってかわるのではなく、従来手法と相補的な立場に立つといえる。

二つ目は、自然言語処理分野、データマイニング分野、情報理論、そして学習理論などの諸分野のより緊密な協調を促すことである。本プロジェクトでは、自然言語処理分野、データマイニング、情報理論（データ圧縮）で用いられていた技術がそれぞれ別の分野においても応用が可能なことが示せた。

6. 普及の見通し

本プロジェクトは現段階では基礎的な研究段階であるため、そのまま一般への普及は期待できない。普及のためには、これらの技術を用いた応用アプリケーションの開発などが必須となる。本プロジェクトの成果を、論文発表、web 公開（7.参照）することにより、周辺研究を促すとともに、私自身も今後本プロジェクトの成果を元にした様々な開発を行う予定である。

7. 開発者について

名前　岡野原　大輔
所属　東京大学　理学部　情報科学科４年（2004 年４月より）

連絡先
e-mail VZV05226@nifty.com （自宅）
hillbig@is.s.u-tokyo.ac.jp（学校）
＊本プロジェクトに関するメールは、自宅の方にお願いします。
website http://member.nifty.ne.jp/DO/index.htm
プロジェクトの成果などを公開