日経記事の計算機処理における日本語 WordNetの有効性

吉　武　春　光

1．まえがき

最近、ビッグデータ（big data）という言葉が頻繁に使われている。ビッグデータの定義はないが、従来のデータベース管理ツールでは処理するのが困難なほど巨大で複雑なデータ集合のことを指す場合が多いようだ。具体的なビッグデータとしてはSNSなどの掲示板への書き込みやウェブショップのアクセス記録などがある。しかし、それらは外部に提供されておらず、入手は難しい。外部に提供されているデータとして有名なものに新聞記事がある。筆者は、日本経済新聞社の新聞記事（本紙、産業新聞、流通新聞）を18年分（1995年１月～2012年12月）入手した。容量は約８GBである。18年分の量をビッグデータと言えるかどうかという指摘はあるだろうが、ビッグデータに近い量を処理して実験を行うのには妥当であると判断した。

さて、筆者は，コンピュータを用いた自然言語処理に長年、取り組んできた。今まで使ってきた手法は、語の意味を記述し、語の構文に従って、文の意味を合成し、更に、談話の意味も合成する、というものである。本質的にたいへんなのは、語の意味を１つ１つ記述している作業が膨大であることである。更に、ビッグデータに対して意味の合成を続けていくと、データ量が爆発してしまうという現象に遭遇する。そこで、一般に行われているビッグデータ処理では、意味の合成は行わずに、語の出現頻度を利用している。

しかし、処理対象によっては、語の出現頻度だけを利用した処理では上

手く行かないことがある。そこで、本研究では、既に多量の意味の記述が行われている既存の意味記述辞書を使い、更に、処理可能な範囲内のデータ量に収まることを目指し、日本語 WordNet を使用した意味処理を行うことにした。

２．日経新聞記事の特徴と前処理

まず、データを計算機に読み込む必要があるので、日本経済新聞社の本紙（以下､日経データと略す）18 年分（容量約５GB）を分析した。日経データは、月ごとに１つの CSV 形式のファイルとなっている。例えば、次が１文である。

"NIRKDB20000101NKM0410","20000101","NKM","日本経済新聞　朝刊","＄社会","2000 年その時､「安全確認」安どの新年――病院、救命医総動員、停電や断水なく。","　一九〇〇年代に別れを告げ、二十世紀最後の年へ。コンピューターの誤作動が心配される…(省略)…と語った。”

カンマ区切りの各欄はダブルクオーテーションで囲まれた値になっており、先頭から次の通りになっている：

articleid、date、mediacode、media、bunrui、headline、htmlsource

ここで、mediacode 欄は、日経本紙の場合 NKM、日経産業新聞の場合 NSS、日経流通新聞の場合 NRS となっている。media 欄は、メディアコードの下位分類になっており、日経本紙の場合だけが、”日本経済新聞　朝刊”と”日本経済新聞　夕刊”という異なる値を持ち、日経産業新聞の場合は”日経産業新聞”、日経流通新聞の場合は”日経流通新聞”となっている。bunrui 欄は、俗に言う内容分類タグがカンマ区切りで付与されている。htmlsource 欄は記事の本文であるが、
 タグが段落の区切りとして入れてある。

分類欄は、18 年間における一貫性が保証されない上に、分類漏れや分類ミスの発生が懸念されるので、本研究では使用しないことにした。html ソース中の
 タグであるが、記事における段落が何らかの意味まとまりであると考えられるので、以後の処理では、
 タグで段落の区切りとし、各々の段落に段落番号を付与することにした。つまり、１つの記事は、複数の段落から構成され、各段落は複数の文から構成されているとした。

３．日本語 WordNet

WordNet とは、英語の単語についての意味辞書である。プリンストン大学のジョージ・ミラー (George A. Miller) 教授が 1985 年より開発を行っている。なお、この WordNet に影響を受け、色々な国用の WordNet が開発されているので、このプリンストン大学版の WordNet を指し示す場合は、Princeton WordNet と表記することにする。WordNet の構造は、単語を意味的に分類し、階層構造としてデータベース化したものである。データベース内のテーブルには次がある。

表１　WordNet の構造

テーブル名	説　　明
word	単語情報
synset	単語の持つ概念をまとめたもの
sense	単語と概念の紐付けを管理する
synlink	synset 間の関係性を管理する
ancestor	synset 間の関係性の深さ（世代数）を管理する
xlink	synset と上位オントロジー SUMO との関係性を管理する
variant	単語の特殊な読み方を管理する 日本語 Wordnet ではデータ無し。

意味は概念として synset と呼ばれる文字コードで管理されている。約 11 万 7000 個の synset に分類された約 15 万語が収録されている。synset の意味は、独自の構成要素を用いずに英語を用いて記述されているので、

WordNet は、意味辞書というよりシソーラスと捉えるほうが妥当であろう。

さて、WordNet に影響を受け、情報通信研究機構 (NICT) が 2006 年より日本語 WordNet を開発している。日本語 WordNet は基本的に Princeton WordNet の synset に対応して日本語が付与されている。日本語 WordNet の現在のバージョンは Wn-Ja 1.1 で、全ての関係（上位語、全体部分関係 など）は Princeton WordNet 3.0 に準じている。日本語ワードネットに収録された synset 数や単語数、語義数は次の通りである。

57,238 概念 (synset 数)
93,834 words 語
158058 語義 (synset と単語のペア)
135,692 定義文
48,276 例文

3.1 日本語 WordNet の利用可能性

日本語 WordNet は、語の概念を分類し階層構造としてデータベース化してある。では、いったい、どういう意味処理に使えるのであろうか？

野間 (2013) は、数年分の新聞記事からキーワードで絞り込んだ文章に対して TTM を使って解析し、SPSS を使って統計的に処理することで特徴が見えていると報告した。しかし、第２章で述べた新聞記事を数年分に渡ってビッグデータ的に解析してみると、記事の特徴を見いだせなかったと述べている。これは、新聞記事からキーワードで絞り込む際に、恣意的な要因が働いたのではないかと思われるからである。この恣意性を排除するためには、新聞記事に対して、客観的な意味処理を行う必要がある。そこで、予め、辞書中の語に対して情報を付けておいてから、新聞記事の段落や文を処理する際に、その情報を用いて重み付けを行い、重みが大きいものが重要語だと客観的に判断するということが必要になる。

3.2 WordNet の語の間の距離を用いた文の意味処理

WordNet は語の意味を絶対的に記述したものではないので、自然言語の処理で具体的に利用できるのは、語の synset 同士の距離を使った処理であろう。つまり、文を解析して得られた語の依存構造に対して、語の synset 同士の距離を計算し、語の synset の多義を解消する処理である。

４．日経新聞記事の解析実験と考察

本章では，実際に、コンピュータを用いて、日経記事の形態素解析、構文解析、日本語 WordNet を用いた synset 付与、語の synset 同士の距離を計算し、語の多義を解消する実験を行った。処理の全体像を図１に示す。

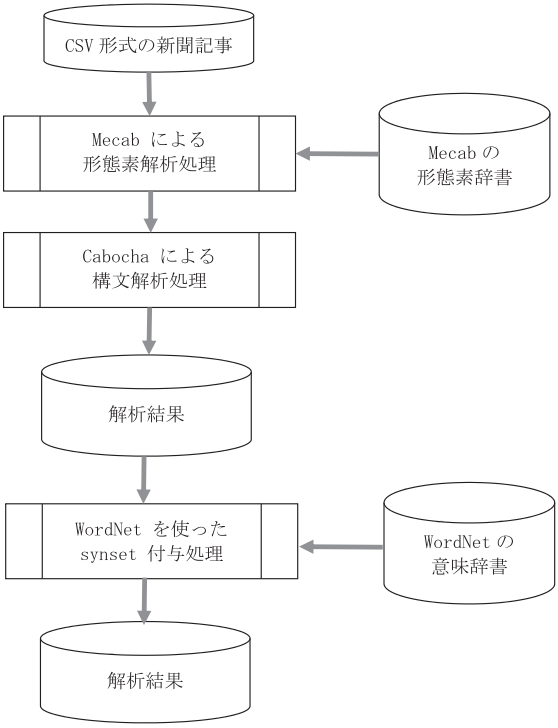

図１　新聞記事解析実験処理の流れ

日経データの中の htmlsource を取り出し入力原文としたが、句点「。」の存在によって、１文の最後であると判断した。その入力原文を、形態素解析器 Mecab そして、構文解析器 Cabocha で処理して、語の係り受けの依存木をリストの形で出力した。そして、日本語 WordNet を使って、語の意味 (synset) を取得した。なお、入力原文に句点「。」が付いていない箇条書きみたいな行が存在していても、後の構文解析器 Cabocha において、複数行の間の係り受け関係を調べることができるので、問題は生じない。

4.1 Cabocha による構文解析処理

Cabocha は、工藤拓氏（現 Google）が奈良先端科学技術大学院大学に在学中に開発したもので、現在は、フリーソフトウェアとして配布されている。構文解析処理の前処理としては、必ず、形態素解析処理が必要になるが、Cabocha では同じく工藤拓氏が作成した Mecab という形態素解析プログラムを呼んでいる。Mecab は、標準では IPA 辞書 (ipadic) と呼ばれる、形態素情報が載った辞書を使っている。IPA 辞書とは、情報処理振興事業協会 (IPA) で設定された IPA 品詞体系 (THiMCO97) に基づいて一部修正を加えたものである。

Cabocha は実行時に -f １ というオプション（係り受け解析レイヤ）を指定することで、計算機処理に適したフォーマットで出力される。下記は、Cabocha -f １ の入力として
『太郎は花子が読んでいる本を次郎に渡した』
を入力した場合の処理結果である。

```
% cabocha -f1
太郎は花子が読んでいる本を次郎に渡した
* 0 5D 0/1 1.062087
太郎　　名詞,固有名詞,人名,名,*,*,太郎,タロウ,タロー
は　　助詞,係助詞,*,*,*,*,は,ハ,ワ
* 1 2D 0/1 1.821210
花子　　名詞,固有名詞,人名,名,*,*,花子,ハナコ,ハナコ
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
* 2 3D 0/2 0.000000
読ん　　動詞,自立,*,*,五段・マ行,連用タ接続,読む,ヨン,ヨン
で　　助詞,接続助詞,*,*,*,*,で,デ,デ
いる　　動詞,非自立,*,*,一段,基本形,いる,イル,イル
* 3 5D 0/1 0.000000
本　　名詞,一般,*,*,*,*,本,ホン,ホン
を　　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
* 4 5D 1/2 0.000000
次　　名詞,一般,*,*,*,*,次,ツギ,ツギ
郎　　名詞,一般,*,*,*,*,郎,ロウ,ロー
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
* 5 -1D 0/1 0.000000
渡し　　動詞,自立,*,*,五段・サ行,連用形,渡す,ワタシ,ワタシ
た　　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
EOS
```

ここで、第１カラム目がアスタリスク＊である場合と、そうでない場合がある。第１カラム目がアスタリスク＊で始まる行は Cabocha の構文解析結果を示している。第２カラム目が文節番号、第３カラム目が係り先の文節番号である。文節番号は 0 から始まる整数で、- 1 という文節番号は係り先がないことを示している。第４カラム目は主辞 / 機能語の位置である。主辞は文節内で最も機能を果たす語のことで、機能語は文節内で語彙的意味を持たない助詞などの語のことである。第５カラム目は係り関係のスコアである。係り関係のスコアは , 係りやすさの度合を示しており、一般に大きな値ほど係りやすいことを表す。但し、係り関係のスコアの意味付けに関しては不明となっている。なお、語の位置番号も、0 から始まる整数となっている。

次に、第１カラム目がアスタリスク＊でない行は、Cabocha 内部で呼ばれた Mecab の解析結果を示している。第１カラム目は処理対象になった語そのもので、第２カラム目が、カンマ区切りで示された品詞などの Mecab

の解析結果である。第２カラム目の第７番目には語幹が入っている。

4.2 日本語 WordNet を使った意味情報の付与

日本語 WordNet は２つの形式で提供されている。１つは単なる英語と日本語の対応関係の表であり、もう１つは英単語も含んだ形の sqlite ３ データベースの形式である。そして、プログラミング言語 Perl や Python から sqlite ３ データベースをアクセスするためのフロントエンド（API）が提供されている。本研究では、プログラミング言語 Perl を用いたが、この提供されているフロントエンドを一部修正したものを使った。

Cabocha による構文解析の結果として得られた語幹（原語ではない）に対して、日本語 WordNet 辞書を引いて、意味情報 synset を付与した。なお、日本語処理で問題になる同義語であるが、今回は処理対象の文量が多いので、同義語処理を行わず、Mecab に内蔵されている辞書を、そのまま使用した。解析結果の一部を次に示す。

NIRKDB19950101NKM0180　１　１　日本（名詞,固有名詞,地域,国,*,*,日本,2,３）　{(08921850-n)}
NIRKDB19950101NKM0180　１　１　作曲（名詞,サ変接続,*,*,*,*,作曲,2,３）{(03081660-n),(00939452-n),(04933544-n),(01705494-v),(01706014-v)}
NIRKDB19950101NKM0180　１　１　家（名詞,接尾,一般,*,*,*,家,2,３）{(03259505-n),(13812607-n),(08559508-n),(08078020-n),(03544360-n)}
NIRKDB19950101NKM0180　１　１　協会（名詞,一般,*,*,*,*,協会,2,３）{(08049401-n),(08305114-n)}

各欄は [Tab] コードで区切り、各欄の意味は、次の通りとした。

記事 ID [Tab] 文番号 [Tab] 段落番号 [Tab] 語 [Tab] (Cabocha から得られた情報) [Tab] {(synset １),(synset ２)…}

ここで、Cabocha から得られた情報は、次の通りとした。
(品詞、品詞詳細１、…、品詞詳細５、語幹、文節番号、係り先文節番号)
なお、文節番号は最初を１として、係り先文節番号が ０ の場合は、係り先がないことを意味している。

なお、語は変化する（活用する）ものがあるので、日本語 WordNet 辞書を引く際に使用したのは、原文の語ではなく、Cabocha 解析の結果、得られた語幹とした。

なお、１年分の語の数は約５千万語であった。２年分では約１億語になり、18 年分では、約９億語になった。

さて、Mecab 辞書と日本語 WordNet 辞書の登録語が、必ずしも同じではないので、名詞が連続する場合と、名詞と接尾語がつながっている場合には、その合成語が日本語 WordNet 辞書に登録されているかどうか、つまり synset が得られるかどうかを調べ、もし得られた場合は、その合成語を使うようにした。例えば、上記の例の場合、次が得られている。

NIRKDB19950101NKM0180　１　１　日本 (名詞 , 固有名詞 , 地域 , 国 ,*,*, 日本 ,2, ３)　{(08921850-n)}
NIRKDB19950101NKM0180　１　１　作曲家 (名詞 , 名詞＋接尾 ,*,*,*,*, 作曲家 ,2, ３) {(10624540-n),(09947232-n)}
NIRKDB19950101NKM0180　１　１　協会 (名詞 , 一般 ,*,*,*,*, 協会 ,2, ３) {(08049401-n),(08305114-n)}

ここでは、「作曲」が５つの synset を持っており、「家」も５つの synset を持っているために、５ x ５ = 25 もの意味の組合せの可能性が生じるが、「作曲家」の synset は２つしかない。つまり、25 もの多義が２に減ったということである。

５．あとがき

本研究は，ビッグデータとして日経の新聞記事を取り上げ、日本語 WordNet を使った簡単な意味処理の有効性を確認した。4.2 節で述べたように、synset を利用すれば、意味の多義を減らすことが可能になる。現在は、Cabocha の処理で得られた係り受け関係を利用していないが、係り受け関係の情報を用いて意味処理すれば、更に、多義を減らすことが可能であると予想される。そこで、次は、係り受け関係に対して、同様の処理を行う予定である。ビッグデータに対する簡単な意味処理で、どこまで効果を発揮できるのかを確認していきたい。その上で、吉武 (2007) で述べたような本格的な意味記述を使った場合の有効性の確認などを進めたい.

なお，本研究で使用した日経データは、本学商学部の 2013 年度の共通図書費を使って購入したものである。購入を認めてくださった商学部の先生方と図書館に感謝します。

参考文献

工藤 拓、松本 裕治 (2002). " チャンキングの段階適用による日本語係り受け解析 ” 情報処理学会論文誌 43 巻６号 p.1834-1842.

"Princeton WordNet" http://wordnet.princeton.edu/wordnet/

“ 日本語 WordNet” http://nlpwww.nict.go.jp/wn-ja/

野間利博 (2013). “ 新しい市場創出の意思決定における情報通信技術活用の考察 (２)” 西南学院大学大学院 経営学研究論文集　57 号　2013 年１月

吉武春光 (2007): “ SDRT による談話の意味記述 ”, 西南学院大学商学論集 , Vol. 53, No.3 & 4, pp. 211-238, 2007.2 月