日本語連語データの整備

渡辺 耕平*　田辺 利文**　小山 泰男***　吉村 賢治**　首藤 公昭**

*,**福岡大学工学部電子情報工学科
***(株)セイコーエプソン

*td052013@cis.fukuoka-u.ac.jp
**{ tanabe, yosimura, shudo }@tl.fukuoka-u.ac.jp
*** Koyama.Yasuo@exc.epson.co.jp

1　はじめに

最近、自然言語処理において、複数の単語からなる慣用的、成句的な表現に対処することが不可欠であることが広く認識されるようになっている(Sag et al., 2002)。筆者らは日本語に関して機械処理で問題となるであろう連語候補を収集、整理する作業を従来から行っている(shudo et al., 1980, 首藤ら, 1988, 首藤, 1989, 安武ら, 1997)。連語候補による考察や予備的実験は(koyama et al., 1998,岩瀬ら, 2001, shudo et al., 2004)等に既に報告している。表現の収集は、確率的束縛性（要素単語相互の確率的な共起しやすさ）、語彙的一体性（要素単語の間への他の単語の割り込みにくさ）、熟語性（構成性原理の成り立ちにくさ）の３つの性質に注目して行っているが、これらの情報をいかに辞書中に記載するかについて、現状を報告する。

2　表現と３つの属性

我々は広範な領域の大規模日本語データに基づき 1970 年代から人手によって意味上の単位と考えるべき表現の収集・整理を行ってきた。(shudo et al., 1980, 首藤ら, 1988, 首藤, 1989, 安武ら, 1997) 我々が収集してきた表現は確率的束縛性、語彙的一体性、熟語性の 3 つの性質のうち少なくとも 1 つを持つと考えられる長単位表現（単語列）と言うことが出来る。ここで、確率的束縛性とは、要素単語相互の確率的な共起しやすさを意味する。語彙的一体性とは、分離しにくさ（要素単語の間への他の単語の割り込みにくさ）を、熟語性とは、構成性原理の成り立ちにくさを意味しており、構成している単語の通常の意味から全体の意味を構成するのが難しいことを指す。収集した各表現は基本的にこれらの性質の有無や程度を表す 3 つ組によって性格付けされるが、これらの性質の有無の判断は収集者の内省によっている。

3　辞書の見出し記述と展開処理

連語辞書の網羅性向上のためには、表現のゆれも考慮する必要がある。表現の辞書中への記載に対しては表記の「ゆれ」情報を全て辞書に入れると、データ量の増大と処理速度の低下を招く恐れがある。例えば「争いを引き起こす」という表現は、「あらそい」や「ひきおこす」などの平仮名による表記を除いただけでも、他に５つの表現として現れる可能性がある。

争いを引起こす、争いを惹き起こす、
争いを引き起す、争いを引起す、
争いを惹き起す

このような「ゆれ」の情報を含めてコンパクトに辞書に記載するため、この場合には、

争い-を-(引(き)/惹き)-起(こ)す

のように、「ゆれ」に伴う表記をまとめて１つの見出しとして記載することにしている。ここで、ハイフン「-」は、(1)文節の境界、(2)漢字表記とひらがな表記の境界、にそれぞれ挿入している。また、辞書の見出しから元の表現を抽出することを展開処理とよぶことにする。現在の日本語連語辞書の見出し総数は約 66500 個、漢字部分のみを考慮して展開処理を行った場合約 89400 個、ひらがなをも考慮して展開処理を行った場合には、約 413500 個となっている。

4　確率的束縛性について

4.1　確率的束縛性の算出

確率的束縛性とは、要素単語相互の確率的な共起しやすさを表わすものであり、この性質を有する表現として、例えば“悪夢・に・うなされる”などがある。ここで、表現例中の記号‘・’は通常の単語境界を表す。日本語には慣用表現や定型表現が数多くあり、このような表現は単語間の連接確率が大きいと思われる。ここで、n個の単語からなる単語列$w_1w_2...w_n$を連語とした場合、連語$w_1w_2...w_n$の生起確率$P(w_1w_2...w_n)$は

$$P(w_1w_2...w_n) = P(w_1)P(w_2|w_1) \cdots P(w_n|w_1...w_{n-1})$$

で表わされる。上式における右辺の条件付き確率$P(w_n|w_1...w_{n-1})$を$P(w_n|w_{n-N+1}...w_{n-1})$で近似したN-gram確率を用いることも多いが、精度の点で問題がある。そのためN-gramモデルの局所性を補う手法が数多く提案され、その一つとして、頻出する単語連鎖や定型表現をまとめて一単位として扱う試みもされている。本稿では、上式における右辺の条件付き確率を直接テキストコーパスから実測し、確率的束縛性を客観的に捉えることを考える。

4.2　展開処理およびテキストコーパス

テキストコーパスとのマッチング処理を行うため、連語辞書の見出しに対して展開処理を行った。展開処理は、辞書中の「名詞-格助詞に-動詞」の形の表現に制限した。今回、テキストコーパスとして毎日新聞の記事6年分のデータを用いた。テキストコーパスは、動詞が活用しているものもマッチングの対象とするため、日本語形態素解析システムchasen を用いて形態素解析した。

ここで、展開処理を行う前の見出し数は2483個[1]で、展開処理を行った後の、展開された見出し語数は5616個となった[2]。

4.3　算出結果

「名詞-格助詞に-動詞」の形の表現において、実測した条件付き確率（連接確率）を求めた。それぞれの表現において、連接確率の最大値が大きかった10個を値が大きいものから順に表１に示す。連接確率が大きかった箇所を記号=で示すことにする。また、連接確率の最大値の横に、コーパスにおける生起頻度も記載した。ただし、生起頻度が一桁であるものは信頼性に欠けるものとして、記載していない。

表現	連接確率の最大値
双肩=に	1.0 （28/28）
小耳=に	1.0 （14/14）
冥利-に=(尽きる/つきる)	1.0 （12/12）
腑-に=(落ちる/おちる)	1.0 （11/11）
口々=に	0.995 （400/402）
明るみ=に	0.994 （1159/1165）
大目-に=(見る/みる)	0.989 （92/93）
念頭=に	0.989 （2354/2380）
快方-に=(向かう/むかう)	0.984 （64/65）
一堂=に	0.981 （1026/1045）

表１　連接確率の大きい表現

また、「名詞-格助詞に-動詞」の形の表現において、コーパス中の「名詞-格助詞に」の後に、辞書中の「動詞」がどの程度現れているかの調査を行った。例えば「学校に」の後の単語としては「行く」や「上がる」などの数種類しかないものと考え、連語として収集している。それで、実際に連語辞書に収集されているものの中で、これら辞書中の動詞が来る割合がどの程度になっているかを「収集度」として求めた。但しこの場合も名詞のひらがな表記の場合を考慮していない。表２に、収集度が高い「名詞-格助詞に」の表現のうち動詞のバリエーションが多い表現、およびその収集度（確率）を示す。

表現	収集度(確率)
眠り-に	0.752
落ちる、つく、入る、etc	
恩-に	0.750
着せる、着る、報いる、etc	
口-に	0.705
合う、入れる、する、出す、運ぶ、etc	
気-に	0.634
かける、障る、留める、なる、etc	

表２　収集度が高い「名詞-格助詞に」の表現

[1] 「名詞-格助詞に-動詞」の動詞部分にハイフンが含まれない見出し数であり動詞部分が「成功する」「注意する」などの表現は含まない。

[2] 但し、ここでの展開では名詞のひらがなに対しては行っていない。その理由として、展開処理はテキストコーパスとのマッチングを目的としており、(1)テキストコーパス中には、漢字で表現できる文字はひらがなではあまり現れないこと、(2)名詞のひらがなの展開処理を含めるとマッチングの時間が爆発的に増加すると予想されること、と考えたためである。精度を高めるには、当然ながらひらがなにおいても展開処理をする必要があり、これは今後の課題としたい。

5　語彙的一体性について

語彙的一体性とは、構成している単語の分離しにくさを表わす。言い換えると、要素単語の間への他の単語の割り込みにくさを指す。例えば、“赤・の・他人”、“鶴・の・一声”などがある[3]。

5.1　語彙的一体性の算出

語彙的一体性も、テキストコーパスから求めることができる。「名詞-格助詞に-動詞」の形の表現は、「名詞-格助詞に」と「動詞」の間に、単語が割り込めるものとして辞書に記載している。本稿では「名詞-格助詞に-動詞」の見出し2483個に対して、「名詞-格助詞に」と「動詞」との間に、割り込む単語が6個以下の場合に限定して、テキストコーパスとのマッチングを行った[4]。このマッチングにより、「名詞-格助詞に」と「動詞」が直接接続する場合に比べ、表現によっては生起頻度が増えることになる。「割り込み」を考慮しても生起頻度が増えない表現の割合は全体の約39%であり、そのような表現は「完全」に語彙的一体性があるということができる[5]。一方、語彙的一体性が「完全」でない場合には、割り込む単語の数だけでなく、品詞などの情報をも含めて語彙的一体性を記述することが必要かと思われる。

6　熟語性について

熟語性とは、構成性原理の成り立ちにくさを表わす。言い換えると、構成している単語の通常の意味から全体の意味を構成するのが難しい性質を指す。例えば、この性質を有する表現として、“血祭り・に・上げる”、“油・を・売る”などがある。

6.1　熟語性の記述

熟語性を有する表現は、(1)熟語的な意味しか持たない表現、(2)熟語的な意味と文字通りの意味が用いられる表現、の2種類に分類できる。(1)の例として“血祭り・に・上げる”、(2)の例として“油・を・売る”などがある。正しい意味をシステムが認識するためには、(1)のような表現は、その表現の組み込みが必要条件であり、また、(2)のような表現は、表現の組み込みだけでなく、文脈情報を用いる必要がある。辞書に組み込む場合、どのように文脈情報を記述すべきかが非常に重要な問題となってくる。

熟語性のコンパクトな表現として、(1)のような表現では場合は値を1とし、(2)のような表現では、文脈情報を用いることが基本かつ必要ではあるが、0から1の間の値を「熟語的な意味として生起する確率」としてテキストコーパスから求めることなども考えられる。それでもなお、その値を推測するには膨大なコストがかかると思われる。

6.2　翻訳システムを使った実験

熟語的な意味しか持たないような表現は、システムに登録しておかないと意味を正しく認識することができないものと考え、このような表現を翻訳システムの入力とし、どの程度正しく翻訳されるかを実験により求めることを考える。日英機械翻訳システムは 2006 年 3 月において最新の市販の高精度のものを使った。

6.2.1 対象表現の抽出および展開処理

まず、「名詞-格助詞に-動詞」に属している 3693 表現のうち、熟語的な意味しか持たないと思われる約 19%、717 表現を人手で抽出した。次に人手で抽出された 717 表現に対して、展開処理を行い 979 表現を得た[6]。

6.2.2 英訳実験とその結果

展開処理を行って得られた 979 表現に対し、市販されている翻訳ソフトを用いて実験を行い、どの程度正しく英訳されたかを算出した[7]。その結果、見出し表現に対して表記のゆれを含んだもののうち、１つでも英訳が正解であった見出し表現が 223 表現であり見出し表現全体の 31%(=223/717)であった。また、表記のゆれ全体では、正解であったものは 270 表現であり、正解率は 27%(=270/979)であった[8]。

6.3　考察

熟語的な意味しか持たないような「名詞-格助詞に-動詞」の形の表現に対する、市販の翻訳システムの網羅性は、実験の結果、30%程度であることが分かった。しかし、実際には、熟語的な意味しか持たない表現の使用頻度は表現によりかなりずれがあると思われるため、使用頻度を重みとした正解率を算出する必要がある。この点に関しては今後の課題としたい。

---

[3] 例で挙げたこれらの表現は、熟語性も有している。

[4] 単語は、chasenで認定した区切り単位を採用した。また、入り込む単語数を6個までに制限した理由として、3文節までは「名詞-格助詞に」と「動詞」が係り受け関係にあるものと想定した。

[5] ここでの全体とは、表現そのものの生起頻度が1以上であった表現であり、1815個で「名詞-格助詞に-動詞」の約73%であった。

[6] 例えば、「愛情-に-(飢/餓)える」の場合には「愛情に飢える」と「愛情に餓える」の２つの表現に展開される。

[7] 翻訳を行う際のオプションは、入力は「名詞-格助詞に-動詞」で、動詞句に相当するものであるため「３人称単数動詞句化」を選択した。

[8] 誤訳例としては、見出し表現「血祭(り)に上げる」は、「血祭りに上げる」が"Raises to a blood festival."、「血祭に上げる」は"Raises to 血祭."などが観測された。

7　おわりに

本稿では、連語候補表現のうち、「名詞-格助詞に-動詞」に対し、実測で確率的束縛性、語彙的一体性を求めたこと,および熟語性の意味しか持たない表現を抽出し市販のシステムを用いた実験結果について述べ、現在の日本語連語辞書に追加すべき情報の記述などを考察した。確率的束縛性を有するデータの応用として、ケータイなどの予測変換として使うことなどが考えられる。しかし、「名詞-格助詞に-動詞」の形の見出し数2483個に対し、一度もコーパスに生起しなかった表現が702個で、全体の約28%であった。また、語彙的一体性の算出において、「名詞-格助詞に」と「動詞」の間への単語の割り込みを許すことによって、表現の生起が観測されたケースもあった。これらの結果から、依然としてデータのスパースネスが問題として根強く残っている。今後の課題としては、(1)サンプルとして用いるテキストコーパスのサイズを増やす、(2)「名詞-格助詞に-動詞」以外の形以外の表現に対する確率的束縛性の推定、(3)熟語性の記述法、(4)語彙的一体性の推定などが挙げることができる。また、連語辞書の見出し記述で、漢字表記のゆれを(α_1/α_2/ ... /α_n)と記載する場合には、αの生起頻度をC(α)とすれば、i<jのときにC(α_i)≧C(α_j)と定めることも考えられ、ゆれがある場合の標準的な表記を優先して取り扱うことも考えられる。なお、現在も引き続き辞書の編纂は継続して行っている。

謝辞

生起頻度を求めるにあたり、毎日新聞データ CD-ROM'91,92,95,96,97,98 版を利用しています。利用を許可していただいた毎日新聞社に深く感謝します。

参考文献

岩瀬修, 森元逞, 首藤公昭. 2000. 連語を組み込んだ統計言語モデル. 電子情報通信学会第34回音声言語情報処理研究会: SP2000-113: pp.109-114.

首藤公昭, 吉村賢治, 武内美津乃, 津田健蔵. 1988. 日本語の慣用的表現について– 語の非標準的用法からのアプローチ – 自然言語処理研究会 NL-66-1: pp.1-7.

首藤公昭. 1989. 日本語における固定的複合表現. 文部省科学研究費補助金特定研究(Ｉ), 課題番号63101005.

安武満佐子, 小山泰男, 吉村賢治, 首藤公昭. 1997. 固定的共起表現とその変化形. 言語処理学会第3回年次大会発表論文集: pp449-452.

chasen, http://chasen.naist.jp/hiki/ChaSen/

Iwan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP*. The Proc. of the 3rd CICLING: pp.1-15.

Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi and Kenji Yoshimura. 2004. *MWEs as Non-propositional Content Indicators.* The Proc. of the Workshop on Multiword Expressions at 42nd Annual Meeting of the ACL: pp.32-39.

Kosho Shudo, Toshiko Narahara and Sho Yoshida. 1980. *Morphological Aspect of Japanese Language Processing*. The Proc. of the 8th COLING: pp.1-8.

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo.1998. *Large-Scale Collocation Data and Their Application to Japanese Word Processor Technology*. The Proc. of the 17th COLING: pp.694-698.