

連想情報に基づいた省略語推定システム Ellipses Estimating System Based on Associative Information

寺岡 丈博[†], 岡本 潤[‡], 石崎 俊[†]
Takehiro Teraoka, Jun Okamoto, Shun Ishizaki

[†] 慶應義塾大学大学院政策・メディア研究科, [‡] 慶應義塾大学SFC研究所
Keio University Graduate School of Media and Governance, Keio Research Institute at SFC
{teraoka, juno, ishizaki}@sfc.keio.ac.jp

Abstract

Natural language processing technology has developed remarkably, but it is still difficult for computers to understand contextual meanings like humans. The purpose of our work has been to construct an associative concept dictionary for Japanese verbs and make computers understand contextual meaning accurately. As one of its applications, we constructed an automatic system which can estimate elliptical words. This article presents the result of comparing estimated words in some sentences from weblogs by the system to the associated ones by humans.

Keywords — Associative Concept Dictionary, Association Experiment, Verb

1. はじめに

人間は言葉話す或いは書く際に言葉の背景にある膨大な情報を一般的な知識として利用するため、コンピュータの言語理解機能を人間に近づけるには、人間が持つ複雑で膨大な言語関連情報を体系化したものが必要である。既存の連想概念辞書(以下、名詞連想概念辞書)[3]は、そのような言語情報を名詞について体系化したもので重要文の抽出[4]や多義性の解消モデル[5]、創造的デザインの分析[2]など多岐に渡って応用されている。しかし日常の文脈で動詞が意味理解の重要な役割を果たすため、本研究では動詞連想概念辞書を構築し、省略語を推定するシステムを試作し、有効性を確認した[6]。本稿では、試作段階と比べてより人間に近い意味処理の過程を加味したシステムを用いてブログコーパスの文に対して省略語を推定し、人間の推定した内容と比較していく。

2. 動詞連想概念辞書

動詞連想概念辞書は、動詞を刺激語した連想実験から得た情報を用いて構築しており、232語の動詞に対して連想語数が約57,000語、異なり語数が約15,000語となっている。連想実験では、刺激語を小学校の国語の教科書で扱われている動詞[1]とし、「動作主」「対象」「始点」「終点」「時点」「場所」「手段」「様態」「理由」「目的」の深層格からそ

の一つが課題として一緒に提示されるため、被験者は刺激語動詞の深層格に入る単語を連想する。被験者は湘南藤沢キャンパスの大学生と大学院生を対象にしている。さらに名詞連想概念辞書[3]と同様に刺激語 x と連想語 y の間の連想距離 $D(x, y)$ を式(1)のように定量化している[6]。

$$D(x, y) = \frac{7}{10}F(x, y) + \frac{1}{3}S(x, y) \quad (1)$$

$$F(x, y) = \frac{N}{n + \delta} \quad (2)$$

$$\delta = \frac{N}{10} - 1(N \geq 10) \quad (3)$$

$$S(x, y) = \frac{1}{n} \sum_{i=1}^n s_i \quad (4)$$

$F(x, y)$ は刺激語 x に対して連想語 y が連想された頻度の逆数、 $S(x, y)$ は連想語 y が連想された順位 s_i の相加平均を表している。また $F(x, y)$ は補正值 δ を分母に加えることで正規化しており、被験者数 N を大幅に増加させた時に連想した人数 n が少ない場合、 $F(x, y)$ が極端に大きくなるのを防いでいる。

3. 省略語推定システム

構築した特定の深層格に関する省略語を推定するシステムの概要を図1に沿って述べる。

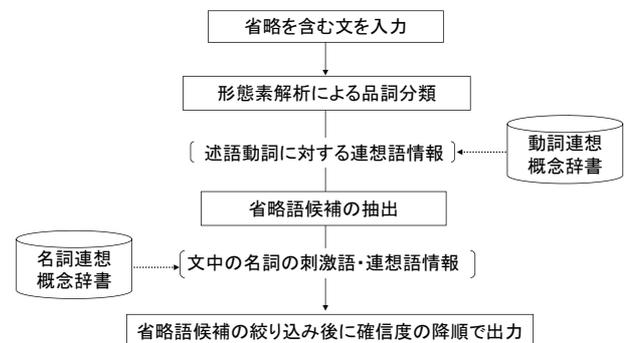


図1 システムの概要

まず省略語を含む文に対して形態素解析を行い、述語動詞に対して動詞連想概念辞書を用いて指定した深層格に対応する連想語を抽出して省略語の候補とする。次に、文中の全ての名詞に対して、

名詞連想概念辞書[3]を用いて深層格と対応させた概念(ここでは「環境概念」と「部分・材料概念」)の連想語や逆引きとして得られる刺激語を抽出する。この際に抽出する連想語や刺激語の名詞に対して「上位概念」の連想語や「下位概念」の逆引きで得られる刺激語つまり上位語に対して連想距離が小さい単語も全て抽出する。これらの内、省略語候補と共通する語が最終的に省略語として確信度付きで出力される。出力の際に、省略語に対する動詞連想概念辞書の連想距離 $L_V(L_V \leq 1)$ [6]と名詞連想概念辞書の連想距離 $L_N(L_N \leq 1)$ [3]を用いて式(5)のように設定した確信度 $c(0 < c \leq 1)$ の降順で出力される。

$$c = \frac{1}{L_V L_N} \quad (L_V L_N \neq 0) \quad (5)$$

出力される語で最も高い確信度 c_{max} を用いて確信度の閾値を $c_{max}/3$ と定め、確信度の低い単語の出力を防いでいる。例えば、「財布を忘れたから友達から借りた。」という文に対して「借りる」の対象を推定した際は、「金(かね)」と「手」が省略語として出力されようとするが、確信度がそれぞれ0.255と0.015のため、閾値により「金(かね)」のみが出力される。

4. システムの評価

省略語推定システムの評価を行うために、ブログで実際に用いられている文を課題文として省略語の推定を行った。推定する深層格は「対象」「手段」「場所」であり、各々10文ずつ計30文の課題文をブログコーパスから抽出した。そしてシステムが推定した語と被験者が省略語として推定した語を比較した。

表1 「調べる」の「手段」に対する推定結果

人間による推定		システムの出力	
単語	L_S	単語	c
インターネット	1.34	インターネット	0.055
本	2.75	本	0.046
雑誌	5.33	雑誌	0.023
辞書	5.67	アンケート	0.019
新聞	5.67		
アンケート	6.00		

表1は、課題文「売るためだけに興味のないことや商品を調べる」に対して「調べる」の「手段」を推定した結果である。評価する上で被験者が推定した単語の頻度と順位から式(1)を用いて文と単語間の距離 L_s を算出した。この距離 L_s が小さい程連想され易いことを表している。システムの推定した単語と被験者が省略語として推定した内容

を比較した結果、表2のようになった。本研究ではシステムが出力した単語を被験者が省略語として推定していたものを正解語としており、正解率は全出力語数のうち正解語数が占める割合である。

表2 システムの正解率

深層格	出力語数	正解語数	正解率(%)
対象	34	21	61.8
手段	29	20	69.0
場所	20	18	90.0
全体	83	59	71.1

5. 考察

全体の正解率より、本システムの有効性が示されたとともに両連想概念辞書を組み合わせることで、動詞と名詞の意味ネットワークがシステム内で実現できていると考えられる。しかし、深層格によって正解率に隔たりが生じており、特に「対象」に関する出力は正解語の他に誤った語が多かったことが分かる。このシステムにニューラルネットワークを組み込み、活性化拡散モデルを実装することで、このような不要語の出力を抑えられる可能性がある。そのため、より人間の意味理解過程に近いシステムの構築に取り組み、いずれば人間の意味理解過程の同定に繋げていきたい。

6. 謝辞

本研究で使用したブログコーパスはデータセクション株式会社から提供していただいたものである。代表取締役社長の池上俊介氏に心から感謝の意を表する。

参考文献

- [1] 甲斐睦朗, 松川利広, (2001)“語彙指導の方法 語彙表編”, 光村図書
- [2] 森田純哉, 永井由佳里, 田浦俊春, 岡田亮土, (2008)“概念合成によるコンセプトのデザインと連想: 概念の連想数と動作概念の役割”, 認知科学, Vol.15, No.4, pp.599-614.
- [3] 岡本潤, 石崎俊, (2001)“概念間距離の定式化と既存電子化辞書との比較”, 自然言語処理, Vol.8, No.4, pp.37-54.
- [4] 岡本潤, 石崎俊, (2003)“連想概念辞書の距離情報を用いた重要文の抽出”, 自然言語処理, Vol.10, No.5, pp.131-151.
- [5] Okamoto, J., Uchiyama, K. and Ishizaki, S., (2008)“A Contextual Dynamic Network Model for WSD Using Associative Concept Dictionary”, Proceeding of the Sixth International Language Resources and Evaluation Conference, pp.1595-1599.
- [6] 寺岡文博, 岡本潤, 石崎俊, (2009)“動詞連想概念辞書を用いる省略語の推定と評価 動詞連想概念辞書の構築と応用”, 言語処理学会第15回年次大会発表論文集, pp.845-848.