概念間の相互情報量による語彙的曖昧性の解消*

2L－4

永山崇　伊藤毅志　古郡廷治†

電気通信大学大学院 電気通信学研究科 情報工学専攻‡

1 はじめに

計算言語学では経験的、統計的に言語現象を捉えようとする傾向が強くなってきている[2]。このような傾向は1950年代にも見られたが、当時の計算機の記憶容量や処理速度の制約から望むような成果が得られなかった。しかし、計算機の能力が向上し、また大規模コーパスが利用しやすくなったことから、近年コーパスから言語知識を抽出し言語処理に利用しようという研究が盛んに行われている。

本稿は、タグつきコーパスから抽出した概念間の相互情報量を用いて日本語文中の語彙的曖昧性を解消する手法を提案し、その手法に基づいて曖昧性の解消実験を行った結果を報告する。なお実験にはEDR電子化辞書[1]とEDR日本語コーパスを用いた。

2 語彙的曖昧性

語彙的曖昧性は自然言語処理における基本問題の一つである。例えば次の文を見てみる。

例1. 日が昇る。

この文には「日」、「昇る」という単語が現れるが、このような一般的な単語も多くの意味を持っている。我々はこのような単語に対し、複数ある意味の中から一つの意味を特定し、それを用いて文の解釈を行う。EDR電子化辞書は例1の各単語の意味として次のものを持つ。

日　：(1) 日にち,(2) 太陽,(3) 日曜,(4) 日本,...
昇る　：(1) 月や日が昇る,(2) 高い所へ移動する,...

実際にはこの中から(2)太陽、(1)月や日が昇る、の意味を選択し、それを用いて文を解釈する。

3 曖昧性の解消法

単語が複数の意味をもつとき、その同定は語彙間の連想関係を使って行うことが出来る。

例1の場合を考えてみると、単語「日」の意味(2)「太陽」と、単語「昇る」の意味(1)「月や日が昇る」との間に連想関係が存在することによって、この2つの意味が特定される。

単語間の連想性を捉える方法にはヒューリスティックによる方法、シソーラスによる方法、辞書による方法、コーパスによる方法等がある。中でもコーパスによる方法には次のような利点がある。

- 自動化が可能である。
- 関係の有無だけでなく強さも扱うことができる。
- 言語の多様性を捉えることができる。

本研究ではコーパスから自動的に概念間の連想性を抽出し、語彙の曖昧性の解消に用いる。コーパスから連想性を捉える方法にはChurch-Hanks(1989)[3]によって提案された相互情報量(mutual information)を用いることにする。

4 相互情報量

相互情報量は元来情報理論の考えであり、2つの要素が共起して現れやすい度合いを統計的に数値化したものである。

今、2つの要素x,yの出現確率をそれぞれ$P(x),P(y)$、xとyが共に出現する確率を$P(x,y)$とすると、相互情報量$I(x,y)$は式1のようになる。

$$I(x,y)=\log\frac{P(x,y)}{P(x)P(y)} \quad (1)$$

もしxとyとの間に依存関係がある場合、共起確率$P(x,y)$は確率$P(x)P(y)$より大きくなり、その結果$I(x,y)\gg 0$となる。逆にxとyとの間に特別な関係がない場合$P(x,y)$は$P(x)P(y)$より小さくなり、その結果$I(x,y)\ll 0$となる。

Churchらは相互情報量により単語間の連想を捉えたが、本研究では各単語に概念(語義)を付加したタグ付きコーパスを用い、単語のもつ概念間の連想を捉える。概念x、y間の連想性は$P(x)$、$P(y)$、$P(x,y)$を次式から計算し、求める。

$$P(x)=\frac{\text{コーパスでの概念}x\text{の出現頻度}}{\text{コーパスでの概念の総出現頻度}} \quad (2)$$

*Word Sense Disambiguation with Conceptual Mutual Information

†Takashi Nagayama, Takeshi Ito, Teiji Furugori
‡University of Electro-Communications

$$P(y) = \frac{\text{コーパスでの概念}y\text{の出現頻度}}{\text{コーパスでの概念の総出現頻度}} \quad (3)$$

$$P(x,y) = \frac{\text{概念}x\text{、}y\text{の同一文中での出現頻度}}{\text{同一文中での概念の共起頻度の総数}} \quad (4)$$

この式により計算された語義間の連想度は、各語義の意味的な繋がりの強さを表す。

例1に現れた「日」の各意味と連想度が大きい語義を次に示す。

単語	日			
語義	太陽	日にち	日曜	...
連想される概念	水星	見ごろ	祝日	
	翳す	紅葉	土曜日	
	庇	消印	人出	
	真っ赤	低気圧	連休	
	地平線	週	待ち合わせる	

5 相互情報量を用いた曖昧性の解消

次に相互情報量を用いた実際の曖昧性の解消法を示す。式1によって求めた連想性は1つの概念間の連想性を表したものである。そのため、ある単語のある概念が他の単語の語義全体からどの程度想起されるかを調べるには他の単語の概念との連想性の総和をとる必要がある。

語義の連想性の総和は、文が単語$W1W2...Wn$から成り、単語Wiが語義$Si_1, Si_2, ...Si_{m_i}$を持つとき式5によって求める。

$$A_{kl} = \sum_{i}^{n, i\neq k} \frac{1}{m_i} \sum_{j}^{m_i} I'(Si_j, Sk_l) \quad (5)$$

$$I'(Si_j, Sk_l) = \begin{cases} I(Si_j, Sk_l) & \text{if } I \geq 0 \\ 0 & \text{if } I < 0 \end{cases} \quad (6)$$

ここで$I < 0$の時$I' = 0$としているのは、負の相互情報量は本来「通常より共起しにくい」を意味するが、実際には偶然出現した関係を拾っていることが多く、信頼性が低いと考えられるからである。

このようにして単語の各語義について想起度を計算し、その中で最も大きい想起度を持つものを選択することによって語義を一意に同定する。

6 実験

EDRコーパスから抽出した100文に対して語彙曖昧性解消実験を行った。文中で曖昧性をもつ単語は438個であった。この各単語の各語義に対し式5から連想性を計算し、そのもっとも高いものを選択し、タグ付きの結果と比較を行った。

実験を行った結果、71%の単語についてコーパスにタグ付けされている意味と同じ意味を選択することが出来た。残りの約30%についてはタグと同じ意味を選択出来なかったが、その主な原因は次のようなものである。

(a) 単語が非常に類似した語義を複数もっていたため。

(b) 文が短く必要な情報が得られなかったため。

(c) 学習コーパス中で単語がその意味で用いられなかったため。

(d) 誤答を強く連想させる概念が存在したため。

(a)は人間でも判断出来ないような意味の違いであり、これは辞書の構成に起因している。(b)は実験が1文中の関係のみを扱っているために起こる失敗であり、前後の文の利用によって解消できる。(c)はより大きなサイズのコーパスを用いれば解消できるが、現在はタグ付きコーパスを利用しているのでそのサイズにはおのずと限界がある。そのため、タグのないコーパスからの概念間の連想性の抽出も考慮する必要がある。(d)の失敗は本手法では捉えられないもので、意味素性による制約手法、ニューラルネットワークなどの他の手法との併用を行う必要がある。

7 終わりに

コーパスから抽出した概念間の相互情報量を用いて曖昧性の解消を行った。今後、さらに精度を上げるためには、他の手法との併用や、概念間の連想性の抽出の工夫などを行うことが必要である。

参考文献

[1] EDR電子化辞書: 日本電子化辞書協会

[2] K. W. Church, R. L. Mercer: Introduction to the special Issue on Computational Linguistics Using Large Corpora. Association for Computational Linguistics, 1993.

[3] K. W. Charch and P. Hanks: Word Association Norms, Mutual Information,and Lexicography: Computational Linguistics, Volume16, Number1, pp22-29, March1990