係り受け関係と相互情報量を用いた単語の意味獲得[1]

2R－1

前川篤志　伊藤毅志　古郡廷治[2]
電気通信大学 情報工学科[3]

1．はじめに

近年、計算機の高速化、記憶容量の大型化に伴い、言語現象を経験的、統計的に捉えようとする傾向にある[1]。特に、単語の意味的曖昧性の解消においては、大規模コーパスから抽出した相互情報量、機械可読の辞書やシソーラスをもとに、その文脈に沿った単語の意味を同定するという手法が主流となっている[2][3][4]。

しかし、次々と新出単語や新出概念が現れる今日において、既存の辞書やシソーラスに載っていない単語や概念が文脈中に現れることは、しばしば起こり得ることである。本研究では、既存の辞書やシソーラスを使うことなく、コーパスのみから得られる情報をもとに、文脈に沿った単語の意味を獲得する手法を提案する。例として動詞「開く」を用いて実験を行い、得られた結果に考察を加えた。

2．単語の意味表現方法

現在、単語の意味を表現する方法としては、以下の3つが主流である。

a. 機械可読辞書の語義定義文を用いて単語の意味ベクトルを作成する方法[2]。

b. 単語の意味が既知の文の用例を集め、これを用いて単語の意味ベクトルを作成する方法[3]。

c. 曖昧性を持つ単語と係り受けの関係にある単語の用例を集め、それらをシソーラスなどから求めた単語間類似度を用いてクラスタリングする方法[4]。

本研究における手法は、上記のc.に属するが、シソーラスは使わず、単語の意味を文脈に沿ったその単語の類義語（同品詞）の順序集合で表現する。例えば、「日を決める」という文においては、この文脈における「日」の類義語の集合｛日時、日にち…｝を提示し、これを意味として捉える。

3．文脈に沿った類義語の抽出

文脈に沿った類義語の抽出には、係り受け関係と相互情報量を用いる。

3.1. 係り受け関係

係り受け関係を用いる理由は、「ある名詞と同じ格関係を持つ動詞どうしは意味的に類似している」という仮説に基づいている。この仮説を検証するために行った予備実験を以下に記す。

①ＥＤＲ日本語共起辞書から、「NをV」の関係にあるもの（『を』格）を取り出す。

②動詞V_1とV_2の類似度を以下の式で表し、類似度の高い順に提示する。ただし、共通共起単語数は、１２以上のものを対象とした。（全ての動詞の組み合わせによる平均共通共起単語数は11.3であった。）

$$\text{単語間類似度}=\frac{N}{N_1+N_2-N} \qquad (1)$$

N　：共通共起単語（名詞）数

N_1：V_1の共起単語（名詞）数

N_2：V_2の共起単語（名詞）数

V_1	V_2	類似度
開	開催	0.0653
開	開け	0.0539
開	広げ	0.0453
開	通	0.0388
開	設け	0.0367

表１．「開」と類似度が高い動詞

表1．は、動詞「開く」と類似度の高い上位５単語を表したものであるが、この結果は前述の仮説に反映していると言えるだろう。

---

[1] Determining Word Senses from Syntactical Relation and Mutual Information

[2] Atsushi Maekawa, Takeshi Ito and Teiji Furugori

[3] Department of Computer Science, The University of Electro-Communications

3. 2. 相互情報量

相互情報量は、Church-Hanks (1990) [1]により「意味的に類似した単語は同じ文脈に現れる」という仮説に基づき、単語間の連想性を捉える一手法として提案されたものである[4]。

3. 3. 類義語の抽出方法

文脈に沿った単語の類義語の抽出は、以下の手順で行う。

①係り受け関係を用いて、単語の類義語の候補を複数抽出する。

②抽出された候補の中から、相互情報量を用いて、文中の各単語と連想度の高い類義語を順に抽出する。ただし、相互情報量の計算は自立語に限定する。

例えば、単語Wからなる、文 $W_1W_2\cdots W_n$があり、W_kの類義語の候補として係り受け関係により抽出された語が$S_1\cdots S_j\cdots S_m$であったとする。このとき、次式によりそれぞれの類義語を評価し、Fs_jの値の高いｔ個の類義語を文脈に沿った単語 Ｗｋの類義語の集合（意味）とする。

$$Fs_j = \sum_{i=1}^{n(i\neq k)} I(W_i, S_j) \qquad (3)$$

4．実験

実験は、動詞「開く」の『を』格を例にとって行った。係り受け関係による類義語の候補は、データとしてＥＤＲ日本語共起辞書を用い、上位７語を抽出した。相互情報量の計算にはＥＤＲコーパスを用いた。また、「開く」の『を』格を持つ文をＥＤＲコーパスから無作為に５文抽出して実験に用いた。

以下は、その文である。

ア．１．Ｇ１１諸国の中央銀行総裁は本日午後、通常の月例会議を開いた。

イ．100 メートルほどの距離を隔てて２つのビール会社が一風変わったビアホールを開いている。

ウ．ベンチにゆっくりと腰をおろした彼は、おもむろに本を開いた。

エ．６・３制の改革や教育基本法の見直しを主張してきた自民党文教族は、臨教審に突破口を開いてほしい、と期待していた。

オ．この大学は政治上の理由から外国人には門戸を開いていない。

前述の手法を用いて、それぞれの文から得られた「開く」の意味（類義語の順序集合）は次のとおりである。ただし、ここではFs_jの値の高い上位３語を提示している。括弧内の数値がFs_jの値である。

ア．{開催(6.654)，開け(2.271)，設け(2.068)}

イ．{通(1.077)，設け(0.794)，見つけ(0.485)}

ウ．{広げ(5.764)，つく(2.824)，開け(1.743)}

エ．{広げ(3.273)，見つけ(3.238)，開け(3.056)}

オ．{広げ(3.995)，設け(2.331)，開催(1.660)}

5．考察

実験結果は、それぞれの文脈に沿った「開く」の意味を的確に捉えている。ここでは、同じ集合内におけるFs_jの差が大きく、その値が高いほど、文脈における「開く」の意味がよく捉えられているようである。

6．おわりに

動詞「開く」を例にあげて、コーパスのみから得られる情報をもとに文脈に沿った単語の意味を獲得する手法を提案した。今後、他の動詞についても実験し、この手法の妥当性を人間の評価と比較し検討したい。

参考文献

[1] K. W. Church and P. Hanks: Word Association Norms, Mutual Information, and Lexicography: Computational Linguistics, Vol. 6, No. 1, pp. 22-29, (1990)

[2] Y. Wilks, D. Fass, C. Guo, J. MacDonald, T. Platre and B. Slator: Providing Machine Tractable Dictionary Tools, In J. Pustejovsky (ed.), Semantics and the Lexicon, pp. 341-401, Kluwer Academic Pub (1993).

[3] Y. Niwa and Y. Nitta: Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, In Proceedings of the 15th International Conference on Computational Linguistics (CLING-94), pp. 304-309, (1994).

[4] 内元，宇都呂，長尾：動詞の語彙的知識獲得における類義語の用例を用いた多義性の類別，情報処理学会研究報告，Vol. 94, No. 47 (94-NL-101), pp. 105-112 (1994).

---

[4] $I(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$

このＩ(x, y)の値が大きいほど、ｘとｙの連想度が高いことを表す。