共起情報を用いた多義動詞の類別と名詞のクラスタリング

平岡 冠二　　松本 裕治
{kanji-h,matsu}@is.aist-nara.ac.jp
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

コーパス内の共起情報に基づいて、語彙的知識を獲得する試みや、意味的に類似した語をクラスタリングする手法は、これまで数多く提案されてきた[3, 5]。一般には、

仮説 1 意味的に類似した語は同じ文脈に現れる

という仮説がたてられ、その有効性が認められるような実験結果も数多く見受けられる。しかし、そうした研究において語の多義性の問題は無視されることが多いのもまた事実である。特に、動詞に多義性がある場合、名詞を意味的にクラスタリングすることは困難となる。名詞の類似性はそれを支配する動詞を基準としているため、一つの動詞が複数の意味を持ちさまざまな名詞を支配するような現象は誤った語の類似性を導いてしまうことがある。

こうした問題に焦点を当てた研究も幾つかなされている。単言語コーパスの共起情報だけを用いて多義動詞の類別を行なうものに、文献[4]がある。この手法では、

仮説 2：多義は類似した語の集合で区別される

という仮説がたてられ、動詞を、共起する名詞を軸とする n次元ベクトル によって表した後、overlapping clustering(一つのエントリが複数のクラスタに属することを許す) によって多義動詞の意味を区別した動詞の分類を行なっている。

本稿で提案する多義動詞類別の手法も、以上に挙げた二つの仮説に基づくものであるが、多義動詞の類別と同時に名詞間類似度の精度を向上させる方法を用いている。以下の章では、その基本的アイデアと、処理の概要について述べる。

2 動詞の意味分割

多義語の意味は文脈に依存しており、特に多義動詞の場合、それが支配する名詞 (格要素) の意味によって決定される。

例 1
- 椅子に 腰を 掛ける。
- 椅子に 服を 掛ける。
- 彼は 証人に 立つ。
- 彼は 教壇に 立つ。

よって、格要素となる名詞を意味的に分類することにより、多義動詞の持つ複数の意味を類別することが可能であると考える。

上に挙げた例のようにどの格要素が多義動詞の意味を決定するかは一概にいうことはできないが、本研究では、動詞の「を格」に共起する名詞だけを対象に意味分類を行なう。この理由は、(1) 現時点では、形態素・構文解析における曖昧性を回避できない (2)「を格」の格要素については比較的正確に、かつ容易にテキストから抽出できる (3)「を格」は動詞の意味を決定する最も大きな要因と考えられるからである。

以上で述べた動詞分割の考え方をまとめると以下のようになる。

Step 1：動詞と その「を格」に出現する名詞との共起傾向から、名詞間の類似度を定義する。

Step 2：ある動詞 **v** が名詞の集合 **N** をその 「を格」に取った時、名詞類似度に基づいて **N** を意味的なクラスターに分類する。

Step 3：Step 2 の結果 m 個 のクラスターが得られた場合、仮説 2 に基づいて動詞 **v** を m 個 に分割し、それぞれを別の動詞と見なして Step 1 からを繰り返す。

3 名詞の意味分類

名詞の意味分類は名詞間の類似度を基に類似した名詞同士を集まり (クラスター) にすることで行なうが、本研究ではそのアルゴリズムとして代表的な非階層分類法である ISODATA(Iterative Self Organizing Data Analysis Techneques A)[9] を用いた。紙数の都合上、そのアルゴリズムの詳細については述べないが、クラスタリングの条件 (語のまとまり易さ) をコントロール可能であることを除けば、一般的な融合法や分裂法などと大差ないものである。

$$(2)\qquad sim(v,n_i,n_j)=\begin{cases} min(|\,MI_{を}(v,n_i)\,|,|\,MI_{を}(v,n_j)\,|) & : \begin{pmatrix} MI_{を}(v,n_i), MI_{を}(v,n_j) \text{ が} \\ \text{同じ符合である時} \end{pmatrix} \\ 0 & : \text{上記以外} \end{cases}$$

名詞の類似度は、動詞 v とその「を格」に出現する名詞 n の相互情報量 $MI_{を}(v,n)$ を基に導かれる [1, 6]。

$$(1)\qquad MI_{を}(v,n)=\log_2 \frac{\frac{f(v,n)}{N}}{\frac{f(v)}{N}\frac{f(n)}{N}}$$

N ：コーパス中の総文数
$f(v), f(n)$ ：v および n の出現頻度
$f(v,n)$ ：v の「を格」に n が出現する頻度

(1) 式 の値は、名詞の共起傾向を示すものと考え、同じ共起傾向を持つ名詞対 n_i, n_j に動詞 v から見た類似度 $sim_{を}(v,n_i,n_j)$ を (2) 式 のように、また、動詞全体から見た類似度 $SIM_{を}(n_i,n_j)$ を (3) 式 のように定義する。

$$(3)\qquad SIM_{を}(n_i,n_j)=\sum_v sim_{を}(v,n_i,n_j)$$

本手法では分割された動詞をそれぞれ別の動詞として扱うため、分割の処理が行なわれる度に $f(v)$ の値が小さくなり、結果として (3) 式 の値が急増してしまう。よって実際には、(3) 式 の値を [0-1] に単調変換し類似度全体の整合性を保っている。

4 多義動詞類別の手法

動詞分割に使用される名詞クラスターは、**仮説 1** に基づいているため、多義動詞の影響を受けている。よって、十分な類別精度を得るには、名詞間類似度から、多義動詞の影響を取り除かなければならない。

本手法では、動詞分割が行なわれる度に名詞間類似度を再計算することで、多義動詞の影響を類似度から徐々に取り除く方法を用いている。具体的に説明すると、まず、緩い (語がまとまり易い) 条件で名詞をクラスタリングし動詞分割を行なう (大まかな分割)。分割による共起情報の変更にともない名詞間類似度を再計算した後、前回よりクラスタリングの条件を厳しくしていくことで、名詞間類似度と動詞類別の精度を段階的に向上させる。以上の処理の概要を 図 1 に示す。

クラスタリングの条件を際限なく厳しくすれば、過剰な動詞分割を行なってしまうため、条件の上限を設定し

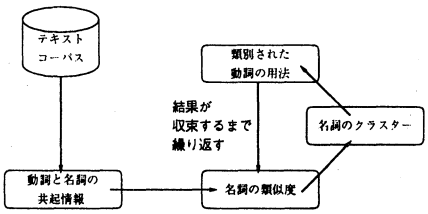

図 1: 類別処理の概要

ておかなければならないが、その設定に一般的指針があるわけではなく、経験的に得るしか方法はない。そこで本研究では、「分類語彙表」[8] を参照することにより動詞の過剰分割を抑制する方法を、ヒューリスティックとして用いる。

「分類語彙表」は、日本語の語彙を概念の階層構造によって分類したものであり、その体の類 (名詞シソーラス) には、約 45,000 語の名詞がその葉の位置に登録されている。ある名詞対 n_i, n_j に付けられた分類コードが一致するレベル (複数のコードがある場合は最大一致レベルのもの) に応じて、「分類語彙表」での名詞間距離 $D_{bgh}(n_i,n_j)$ を次のように定義する。

一致レベル	1	2	3	4	5	6
$D_{bgh}(n_i,n_j)$	16	8	4	2	1	0

過剰分割を抑制する基本的アイデアは極めて単純であり、「分類語彙表」の中で類似した名詞で動詞の分割を行なった場合、これを無視するだけである。以下に、その手順を示す。

「分類語彙表」での名詞間距離 $D_{bgh}(n_i,n_j)$ を用いて、名詞クラスター C の意味的集密度 d_C を以下のように定義する。

$$d_C=\frac{1}{\sharp C(\sharp C-1)}\sum_{n_i\in C}\sum_{n_j\in C} D_{bgh}(n_i,n_j)$$

ここで、$\sharp C$ は、クラスター C の要素の数であるが、「分類語彙表」 にない名詞に関しては あらかじめ除外しておく。

いま i 回目のループで、ある名詞クラスター C^{i-1} が m 個のクラスター $\{C_1^i, C_2^i, \cdots, C_m^i\}$

となり動詞分割に使われたとする。このとき、クラスター分割距離差分 Δ_{C^i} を以下のように定義する。

$$\Delta_{C^i} = d_{C^{i-1}} - \frac{1}{m}\sum_{k=1}^{m} d_{C_k^i}$$

$\Delta_{C^i} < 0$ であれば適切な動詞分割であるとしその結果を採用する。そうでなければ過剰な動詞分割とみなしてその時点での動詞分割結果を無効にする。

5 実験および結果

本研究で行なった実験は二種類で、どちらも同じデータ・条件で行なった。一つは、図 1 で表したものであり、もう一つは、これに先ほど述べたヒューリスティックを付加したものである (→ 図 2)。これ以後、前者を実験-1、後者を 実験-2 と呼ぶことにする。

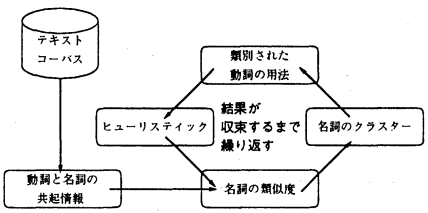

図 2: 実験-2 の概要

5.1 実験データ

基礎データとなる動詞と名詞の共起情報は、EDR 共起辞書 [2] と、朝日新聞の記事データ (総計 約６４万文) から約１８万組を得た。名詞間類似度はコーパス中に１０回以上出現した名詞 (6175 語) を対象に約１４０万組を獲得した。また類別実験は、コーパス中に３０回以上出現した動詞 (1768 語) を対象に行なった。

5.2 結果

実験結果の一例として、動詞 “引く” の類別結果を 表 5.2 に示す。表5.2 は、実験-1 と 実験-2 の結果を並べたものであり、2列目にある名詞は、動詞分割に使われた名詞クラスターの要素である。実際には、動詞 “引く” は 実験-1 で９つ、実験-2 では８つに分割された。

表 1: 「引く」 の類別結果

実験前		実験-1		実験-2	
引く	所得税	引く-1	所得税	引く-1	所得税
	年金		年金		年金
	経費		経費		経費
	人目	引く-2	人目	引く-2	人目
	注目		注目		注目
	血	引く-3	血		血
	水道		水道		水道
	注意		注意		注意
	⋮	⋮	⋮	⋮	⋮

6 評価および考察 (IPAL との比較)

類別結果を評価するために、情報処理振興事業協会の「IPAL 日本語基本動詞辞書」[7](以後 IPAL) に記載されている動詞の用法と、実験で得られた結果との比較を行なった。この評価結果の一例を 表 6 に示す。

表 2: IPAL との比較

I : IPAL 内で「を格」 を持つ用法数
E_i : 実験-i で得た動詞の多義性の数
$I \wedge E_i$: I の用法の一つが、E_i での一つの動詞と対応すると判断された数

	I	E_1	E_2	$I \wedge E_1$	$I \wedge E_2$
引く	10	9	8	4	4
走る	1	10	8	1	1
呼ぶ	5	18	18	2	2
置く	7	17	15	4	3

表 6 が示すように、一部の用法については類別に成功しているが、IPAL との一致率 ($I \wedge E_i / I$) は、それほど高いものではなく、最大で ５７%、平均では、実験-1・実験-2 共に ４５% であった (内容は異なる)。

ヒューリスティックを用いた 実験-2 の場合、過剰な動詞分割を押えることができた (一つの動詞につき約２つの分割を押えた) が、一致率を向上させるまでには至っていない。不一致の原因としては、二つの事が考えられる。一つは名詞間類似度が不足していたためであり、実際には類似した名詞であっても、その類似性が観測され

ず動詞の分割に使われてしまった例がほとんどであった。また、実験-2 では「分類語彙表」にない名詞が全体の 7 割強を占めており、十分な効果を発揮することができなかったこともある。この問題に関しては、コーパス量を増やすことで、ある程度対処できると考えられる。

もう一つは、名詞類似度に含まれるさまざまな影響を取り除けなかったことが考えられる。つまり、誤った類似度を導く原因は、動詞の多義によるものだけでなく、名詞の多義も関連しているということである。本研究では共起情報から名詞の類似度を導く際、名詞の意味を一つに固定しているが、一般に、名詞の意味は一つではなく、複数の局面(概念)を合わせ持っている。そして、名詞の意味はそれを支配する動詞によって決まるものであり、どのような動詞が、名詞のどの局面を引き出すかということは表層の情報から判断できず、結果として誤った距離を導いてしまう可能性がある。よって、今後名詞の多義を類別することが必要となるが、名詞の意味の違いをとらえるために動詞を固定したのでは全く解決にならない。こうした、動詞・名詞どちらか一方から他方を一元的に見る弊害を避けるには、双方の問題を段階的に解消していくしかないと考えられる。この点に関しては今後の課題である。

次に、用法の数だけを比較するために、IPAL から 20 の動詞を、用法の数が少ないものから多いものまでを選んだ。この 20 の動詞を横軸に並べ、IPAL ならびに 実験-1 での用法の数(E_1 の値)を縦軸に取ったグラフを 図 6 に示す。図 6 の中で、実線で書かれているのが IPAL

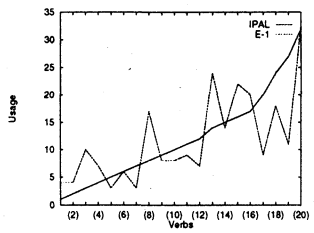

図 3: IPAL と 実験-1 での用例数の比較

での用法の数を表しており、破線は、実験-1 での用法の数を表している。実験-1 のグラフはかなりばらつきがあるものの、IPAL での用法の数が増えるにつれ、同様の増加傾向を示している。よって、多義動詞の用法の数に関しては、おおよその推定が可能であると考えられる。今後、このばらつきを押えより精度を高める必要があるが、これも先ほどと同様に、名詞の多義性の問題を避けて通ることはできない。

7 おわりに

コーパスを用いた語彙知識獲得の研究における問題の一つに、動詞が持つ多義性の問題がある。本論文では、動詞の「を格」に現れる名詞を意味的な集合にすることで、多義動詞の持つ複数の意味を分割する手法について述べた。この手法を用いた実験では、多義動詞の一部の意味については類別が可能であること、多義性の数に関してはおおよその推定が可能であることを示したが、大量のコーパスを必要とする性質上、十分な精度を示すことはできなかった。現在、今回使用したテキストデータの数倍の規模を持つコーパスが利用可能であるため、より大規模な実験が可能であるが、我々は本稿の手法を拡張することで、十分な多義動詞類別が可能であるとは考えていない。動詞の多義の問題は、名詞の多義を無視して解消できるものではないため、双方を段階的に解消していくべきであろう。

謝辞

EDR 電子化辞書データの使用を許可して下さった株式会社 日本電子化辞書研究所に感謝致します。

参考文献

[1] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):235–242, 1990.

[2] EDR. 日本語共起辞書評価版 *(第 2.1 版)*. (株) 日本電子化辞書研究所, 1994.

[3] Francesc Ribas Framis. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proc. of COLING-94*, pages 769–7748, 1994.

[4] Fumiyo Fukumoto and Jun'ichi Tsujii. Automatic recognition of verbal polysemy. In *Proc. of COLING-94*, pages 762–768, 1994.

[5] Ralph Grishman. Generalizing automatically generated selectional patterns. In *Proc. of COLING-94*, pages 742–747, 1994.

[6] D. Hindle. Noun classification from predicate argument structures. In *Proc. of the 28th Annual Meeting of ACL*, pages 268–275, 1990.

[7] IPA. 計算機用日本語基本動詞辞書 *IPAL(Basic Verbs)*. 情報処理振興事業協会, 1987.

[8] 国研. 分類語彙表. 国立国語研究所 秀英出版, 1964,1993.

[9] 奥野 忠一, 久米 均, 芳賀 敏郎, and 吉澤 正. 多変量解析法. 日科技連, 1974.