

動詞の多義性解消における格の弁別能力と集中度について

藤井 敦*

秋山 典丈

徳永 健伸

田中 穂積

東京工業大学 工学部

本論文は、システムに蓄えられた例文と入力文との間で対応する格が取る名詞の類似度に基づいて動詞の多義性を解消する従来の手法に、「格の弁別能力」と「格の集中度」という二つの尺度を導入した新しい方法を提案する。「格の弁別能力」とは格が動詞の語義を弁別する能力を表す尺度であり、取り得る名詞の例(事例)が語義によって異なる格ほど弁別能力は大きくなる。動詞の多義性解消において弁別能力の大きい格をより重視する。「格の集中度」とは、格が取る事例の意味的な広がりを表す。広がり大きい格は様々な名詞を取りやすいことから、その格が持つ事例と入力中の名詞との間に、より大きな類似度を割り当てる。いくつかの動詞について新聞記事を中心とするコーパスを用いて実験を行い、二つの尺度を考慮した場合に解析精度が向上することを確認した。

1 はじめに

動詞の多義性解消は自然言語解析における重要な処理の一つである。機械翻訳における訳語選択や、係り受け解析における名詞句の係り先の決定においても動詞の多義性解消が重要な役割を担う。

動詞の語義ごとに格フレームを記述した IPAL 動詞辞書(以下 IPAL と略す) [1] の例文とシソーラスを用いて動詞の多義性を解消する手法が黒橋らによって提案されている [2]。これは、入力文と IPAL に登録された例文の間で対応する格が取る名詞の類似関係を計算して最も類似性の高い格フレームを選択する方法である。本研究では、この方法に二つの新しい考えを加えて解析精度の向上を試みた。

第一に、ガ格やヲ格などの各々の格が動詞の語義を弁別する能力を表す尺度として「格の弁別能力」を導入する。同様の概念として「かかりの広さ」が提案されている [3, 4]。これは、かかり要素がどの程度述部を限定するかという概念で、述部を限定するほどかかりの広さは狭くなる。格の弁別能力とは、この概念をさらに動詞の語義にまで拡張したものであり、語義を限定する力が強い格ほど弁別能力は大きくなる。

直観的には、格が取り得る名詞が語義によって差

が無い格ほど弁別能力は小さくなる。例えば IPAL において動詞「かける」には 34 個のサブエントリがある¹。「彼が眼鏡をかける」という文の「かける」は、そのうちの一つ「ひも状のような物を身体に付ける」のサブエントリに属する。この例文の動詞の語義を決めるのは「彼」よりも「眼鏡」であると考えられる。これは、「彼」という単語が「かける」の別の語義(水をかける、橋をかけるなど)のガ格にも現れることができるので、それだけでは動詞の語義を決めるのが困難であるのに対して、「眼鏡」は他の語義のヲ格には現れにくいので動詞の語義を決定しやすいためだと考えられる。つまり、この例では動詞「かける」のヲ格はガ格よりも弁別能力が高いと考える。

第二に、格が取り得る名詞の意味的な広がりを表す尺度として「格の集中度」を導入する。意味的な類似性が高いほど集中度が大きくなり、逆に類似性が低い、つまり広がり大きいほど集中度は小さくなる。集中度が小さい格は様々な意味の名詞を取る傾向があると考えられる。

黒橋らの手法では、システムが持っている格ごとの名詞の例(以下、事例と記述する)の数が不十分である場合に、入力文中の名詞が、複数の語義の(対応する格の)事例との間で同じ類似度となり、多義性が解消できないという問題が生じる。例え

*fujii@cs.titech.ac.jp

¹漢字の表記の違いを含む。

ば、動詞「くわえる」について IPAL では5つのサブエントリが用意されている。そのうちの二つ「ある事に対して何らかの見解を付け足す」(語義1とする)と「これまであった物にさらに新しく何かを付け足す」(語義2とする)のヲ格について考えてみる。IPAL ではヲ格の事例として、それぞれ「配慮・解釈・検討・批判・説明」、「塩・新人・彼・ゴッホの絵」が挙げられている。今「材料をくわえる」(語義1に属する)という入力があったとき、「材料」と先に挙げた語義1、語義2の事例との類似度を計算すると、どちらの語義に対しても類似度が0となる²。

格の集中度を考慮すると、語義1の事例は互いに意味的な類似性が高いのに対して、語義2の事例には意味的な類似性がほとんど無く³、集中度が小さい。そこで、語義2のヲ格は様々な意味の名詞を取る傾向があると考え、「材料」と語義2の事例との間に、より大きい類似度を割り当てる。

本稿では、2節で黒橋らのアルゴリズムについて概説し [2]、3節で格の弁別能力について、4節では格の集中度について説明する。さらに5節ではコーパスを用いた評価実験について述べる。

2 基本アルゴリズム

動詞の多義性解消に関して、黒橋らは、格フレーム辞書として IPAL、シソーラスとして分類語彙表 [5] を用いた方法を提案している [2]。そのアルゴリズムを概説する。

- (1) 入力文の格要素と格フレームの事例の対応付けを行う。
- (2) 対応付けられたそれぞれの格要素について、入力文の名詞と格フレームの事例との類似度を計算する。類似度の値は分類語彙表における二つの名詞の分類コードの一致するレベルによって決まる。一致するレベルと類似度との関係を表1に示す。
- (3) 式(1)に従って格フレームと入力文との対応の評価値 (score) を計算し評価値の最も高い格フレームを選択する。

$$score = \begin{cases} 0 & (\text{if } l > n) \\ sum \times \sqrt{\frac{1}{n}} \times \sqrt{\frac{n}{m}} & (\text{otherwise}) \end{cases} \quad (1)$$

n : 対応付けされた格要素数

²黒橋らの手法による
³分類語彙表による

l : $n +$ (入力文側の対応付けられていない格要素数の中で必須的であるものの数)

m : $n +$ (格フレーム側の対応付けられていない格要素数の中で必須的であるものの数)

sum : 対応付けされた格要素数の類似度の和

以上を、図1に示す。長方形は名詞を表す。またガ・ニ・ヲ格は必須的、デ格は任意的としている。

表1: 分類コードと類似度

レベル	0	1	2	3	4	5	6	一致
類似度	0	0	5	7	8	9	10	11

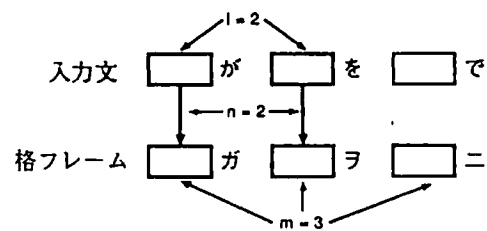


図1: 入力文と格フレームの対応付け

3 格の弁別能力

はじめに、本論文で扱う記号について説明する。

- $N_{f,k}$: 語義 f , 格 k の事例の集合
- $N_{f,k}^i$: $N_{f,k}$ の i 番目の要素
- $|N_{f,k}|$: $N_{f,k}$ の要素数
- I_k : 入力中の格 k が取る名詞
- $sim(X, Y)$: 名詞 X と Y の間の表1に基づく類似度
- $d(k)$: 格 k の弁別能力
- $c(f, k)$: 語義 f , 格 k の集中度

格の弁別能力とは格が動詞の語義をどの程度限定するかを表す尺度であり、格が取る事例が語義によって差が無いほど弁別能力は小さくなる。格の弁別能力は各動詞の格ごとに計算する。

n 個の語義を持つ動詞の格 k の弁別能力 $d(k)$ を式(2)で与える。

$$d(k) = \frac{1}{nC_2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{|N_{i,k}| + |N_{j,k}| - |N_{i,k} \cap N_{j,k}|}{|N_{i,k}| + |N_{j,k}|} \quad (2)$$

$N_{f,k}$ の中で I_k との間の類似度が最も大きい事例を $E_{f,k}$ とする。入力文に現れた格 i ($i = 1 \dots K$)について類似度の加重和を計算し、これを式(1)における sum とする。

$$sum = \sum_{i=1}^K (d(i)^\alpha \times sim(I_i, E_{f,i})) \quad (3)$$

α は定数であり、現在は $\alpha = 1$ としている。評価値 $score$ の計算は式(1)と同じである⁴。

4 格の集中度

格の集中度とは格が取る複数の事例が意味的どの程度類似しているかを表す尺度であり、各語義の格ごとに計算する⁵。

$N_{f,k}$ 中の事例の全ての組について類似度を計算し、その平均を語義 f , 格 k の集中度と考え、式(4)で与える。 $m = |N_{f,k}|$ とする。

$$c(f, k) = \begin{cases} \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^m sim(N_{f,k}^i, N_{f,k}^j)}{mC_2} & (\text{if } m > 1) \\ \text{最大値} & (\text{if } m = 1) \end{cases} \quad (4)$$

$m = 1$ すなわち事例が一つしかない場合、集中度は類似度の最大値(表1では11)とする。

集中度が低い格は、より多くの名詞を取りやすいので、分類語彙表の分類コードを用いて計算した類似度(2節参照)を格の集中度によって補正し、集中度の小さい格との間により大きな類似度を与える。 n を語義数とするとき、 I_k と $E_{f,k}$ の間の補正後の類似度 $sim'(I_k, E_{f,k})$ を式(5)で計算する。 β は定数であり、現在 $\beta = 1$ としている。

$$sim'(I_k, E_{f,k}) = sim(I_k, E_{f,k}) + \beta \frac{1/c(f, k)}{\sum_{i=1}^n 1/c(i, k)} \quad (5)$$

5 評価

5.1 実験

本研究で提案した手法を評価するために新聞記事を中心とするコーパスを解析する実験を行った。

⁴ただし、入力文中の全ての格の名詞が例文の名詞と完全に一致した場合は格の弁別能力や集中度は考慮しない。

⁵弁別能力は語義ごとではなく動詞ごとに計算することに注意。

コーパスはあらかじめ人手によって係り受けと語義を決定し、データとして用意した。また、弁別能力の効果を調べるために、複数の格が充足されているもののみを使用した。実験に用いたデータを表2に示す⁶。

表2: 実験に用いたデータ

動詞	語義数	データ数
やめる(止める・辞める)	2	54文
くわえる(加える)	4	168文

これらのデータを6等分し、そのうちの一つをテストデータ、残りの5つを訓練データ群とする。初期状態ではIPALの例文を与えてテストデータの解析精度を調べる⁷。以降、訓練データ群をシステムに一つずつ与えて、事例・格の弁別能力・集中度を更新しながらテストデータの解析精度を調べる。図2にシステム及びデータの流れを図示する。

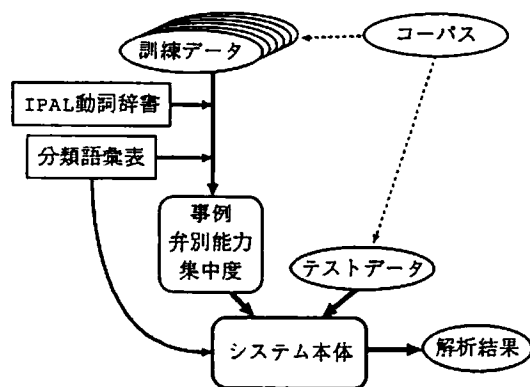


図2: システムの構成

テストデータを変えて同様の操作を6回繰り返し、訓練データの数ごとに平均を計算した。結果を図3、4に示す。グラフの横軸はシステムに与えた訓練データ数、縦軸は解析精度(%)を表す。またグラフの番号と手法との対応を以下に示す。

1. 格の弁別能力・集中度の両方を考慮
2. 格の集中度のみを考慮
3. 格の弁別能力のみを考慮
4. どちらも考慮しない

⁶動詞「くわえる」について、IPALには5種類のサブエントリ(語義)が用意されているが、実験に用いたコーパスには、そのうちの4種類しか出現しなかった。

⁷語義を一意に特定でき、尚かつそれが正解であるものを正解とし、(解析精度) = (正解数) ÷ (テストデータ数) で計算する。

5.2 考察

今回の実験によって以下のことが分かった。

- 格の集中度は、事例データが少ない初期の段階から解析精度を上げることができる。事例ベースのシステムにおいて事例データが十分得られないような場合に、この手法は特に有効と考えられる。
- 事例データが少ない段階では、類似度が0となることが多いため、格の弁別能力による重み付けの効果は期待できないが、データが増えるに従って効果が現れる。
- 格の弁別能力・集中度ともに解析精度を上げるが、格の集中度の方が効果的である。

6 おわりに

本研究では、従来の動詞多義性解消の手法に格の弁別能力と集中度を導入し、評価実験を行うことでその有効性を示した。

今後の研究課題として、動詞によっては格の弁別能力や集中度の違いが顕著に現れるものもあると考えられるため、さらに多くの動詞について実験を行う予定である。本論文中的 α 、 β 等の定数の選び方も含め、弁別能力や集中度の計算についてもさらに検討が必要であると考えられる。

格の弁別能力の有効性を評価するためには、複数の格が充足されたデータを用意しなければならないが、コーパスでは格の省略が頻繁に起こる。格の省略をどのように扱うかについても検討の余地がある。

また、分類語彙表は名詞の多義性についてあまり考慮されておらず、規模としても大きくないことから、より大規模なシソーラスを利用して名詞の類似度をより精密に計算することなども今後の課題である。

謝辞

本稿作成にあたり有益なコメントを頂いたJAISTの奥村学氏、東工大の乾健太郎氏に深く感謝致します。

参考文献

- [1] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL, 1987.
- [2] Sadao Kurohashi and Makoto Nagao. A Method of Case Structure Analysis for Japanese Sentences Based on Examples in Case Frame Dictionary. *IEICE TRANSACTIONS on Information and Systems*, Vol. E77-D, No. 2, pp. 227-239, February 1994.
- [3] 佐伯哲夫. 現代日本語の語順. 笠間書院, 1975.
- [4] 徳永健伸, 岩山真, 乾健太郎, 田中穂積. 日本語語順の推定モデルとその応用. 情報処理学会自然言語処理研究会, Vol. 81, No. 2, pp. 9-16, 1991.
- [5] 国立国語研究所. 分類語彙表, 1964.

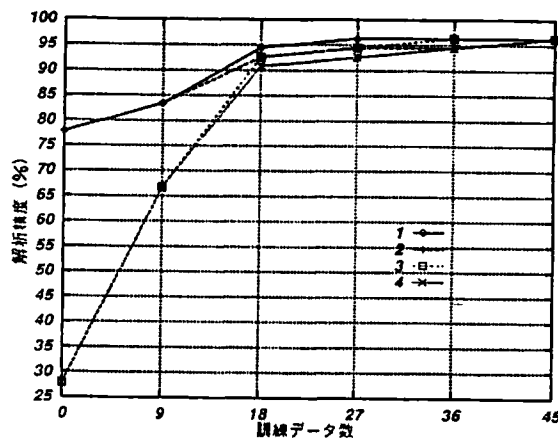


図 3: やめる

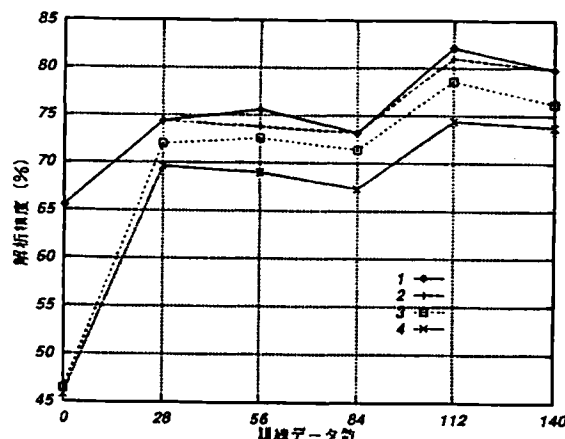


図 4: くわえる