既存の電子化辞書から獲得した格フレームによる構文的曖昧さ解消

吉田 真也　　峯 恒憲　　雨宮 真人
九州大学大学院システム情報科学研究科知能システム学専攻

1 はじめに

計算機による自然言語処理において重要な問題の一つに曖昧さ解消問題がある。生じる曖昧さのうち構文解析時に生じるものは構文的曖昧さと呼ばれるものである。我々は格フレームがこの構文的曖昧さ解消に有効ではないかと考えている。構文解析時に格フレームを用いることにより、格構造に合わない構造、つまり文法的には正しいが意味的には間違っている構造のものは不適格なものだと判断できる。従って、係り受けの曖昧さを解消することができ、曖昧な構文木を削減することができる。格フレームを用いた解析の解析精度は格フレームの精度に依存する部分が大きい。したがって、より精度が高く規模が大きい実用的な格フレームを利用することが重要となる。

われわれはは、実用的な規模、精度の格フレームを自動的に獲得する研究を行なっており、昨年作成した格フレームが動詞概念の推定に有効であることを示した [7]。しかし、実際に構文解析で利用した場合有効かどうか評価を行なっていなかった。そこで本稿では、獲得した格フレームを実際に構文解析内で利用し、構文的曖昧さ解消にどの程度有効か予備実験した結果について報告する。

2 格フレームと類似度計算方法

2.1 格フレームの構造

今回利用した格フレーム [6] は深層格まで考慮した格フレームであり、以下のような構造になっている。

動詞表記		
動詞概念		
表層格	深層格	名詞概念集合
	⋮	
	深層格	名詞概念集合
⋮		
表層格	深層格	名詞概念集合
	⋮	
	深層格	名詞概念集合

2.2 EDR 概念体系辞書

類似度計算時に利用する EDR 概念体系辞書 [3] は概念の上下関係を表したもので、各々のノードが概念を表した木構造をしている。最上位の概念を表すルートノードがただ一つ存在し、このルートノード以下に階層的に深さ 16 に渡って、直接に単語と結び付かない擬似的な概念を含めて約 40 万の概念が登録されている。

2.3 類似度計算方法

ある名詞がある動詞の格フレームに記述されている名詞概念集合部分に入り得るものなのかどうか判定するための尺度として、類似度を用いる。構文解析で格フレームを利用する際に必要な類似度は概念間と概念集合間の類似度である。ここで 2 つの概念間の類似度を以下のように定義する。

概念 N_i, N_j の類似度

$$\mathrm{sim}(N_i, N_j) = \frac{1}{k} \times Dmax_{ijk} \times \left(\alpha \frac{CS_{ijk}}{S_{ik}} + \beta \frac{CS_{ijk}}{S_{jk}} \right)$$

$$0 \leq \mathrm{sim}(N_i, N_j) \leq 1$$

$$\alpha + \beta = 1$$

k : 採用する概念体系のレベル

α, β：2概念間の重み。今回は、$\alpha = \beta = 0.5$

S_{ik}：N_iの上位概念のうち，深さがk以下のものの数

CS_{ijk}：N_iとN_jの共通上位概念のうち，深さがk以下のものの数

$Dmax_{ijk}$：CS_{ijk}のうちの最大の深さ

また、2概念集合間の類似度は、集合内に含まれる概念間の類似度の最大値とし、以下のように定義する。

概念集合 Set_i,Set_jの類似度

$\mathrm{sim}(Set_i, Set_j) = \max_{N_i \in Set_i, N_j \in Set_j} \mathrm{sim}(N_i, N_j)$

3 曖昧さ解消手法

構文的曖昧さの一つとして、名詞句の係り先と成り得る動詞が複数あることが挙げられる。この曖昧さは格フレームを用いることにより解消可能ではないかと考えられる。

3.1 格フレームの利用

ここでは構文解析において実際どのように格フレームを利用するのか述べる。

格フレームの利用法

1. 2つの規則

   1. VP $\longrightarrow$ NP VP
   2. NP $\longrightarrow$ VP NP (NP:名詞句, VP:動詞句)

   を用いて還元動作を行なう場合に以下の格解析を行なう。

2. 該当する動詞句、名詞句の中から、格フレーム解析に利用する動詞、名詞、助詞を抽出する。この3つのうち、一つでも欠けていたら格解析を行わずに、構文解析を続行する。

3. 該当する動詞の格フレームを検索する。

4. 入力文側の助詞と格フレーム側の助詞とを比較して、同じものを探す。もし同じものが見つからない場合は、名詞句がその助詞を伴って動詞に係ることはない、即ち誤った構造であると判断し解析を終了する。

5. 名詞の概念集合と、3. でマッチした格フレームのエントリの名詞概念集合との間の類似度を計算する。集合間の類似度は、2節で述べた集合間の類似度の計算方法にしたがって行う。

6. 局所的な格解析では判定不可能の場合は、その動詞に係る他の名詞の係り受け解析の結果により判定する。

上記手法の補足説明を以下に述べる。

助詞の比較

利用法 3. で行う助詞の比較は、還元に使用するルールによって若干異なる。規則 **1.** の場合は上記の手法通りに行うが、規則 **2.** の場合は共通の助詞を探す作業は行わず、以前使った助詞を除外する作業のみを行う。規則 **2.** のような場合、規則の右辺の NP に含まれる助詞は、この規則内の VP に係るためのものではなく、解析対象部分以降の NP、VP などに係るためのものだからである。

類似度計算について

4. で行う類似度計算の詳細について説明する。類似度を計算した結果が、ある閾値以上だった場合のみ、その係り受けが正しい可能性があると判断する。閾値は、

$$\text{閾値} = \frac{\text{最大深さの共通概念の許容レベル}}{\text{概念探索レベル}}$$

とする。2概念間の共通上位概念の内で最も深いレベルの概念が、どの程度抽象的な場合まで、2つの概念が類似していると見做すか判断するための値である。

局所的な解析では係り先を決定できない場合

5. で行なう作業は以下の通りである。

例) 風に乗った子供たちの声が聞こえた。

名詞	動詞	類似度	意味 (動詞概念)
風に	乗った	0.75	$v_{乗\,1}$, $v_{乗\,2}$, $v_{乗\,3}$
風に	聞こえた	$\times$ 0.4	$v_{聞\,1}$, $v_{聞\,2}$
子供たちの	乗った	0.8	$v_{乗\,4}$, $v_{乗\,5}$
声が	乗った	0.75	$v_{乗\,1}$, $v_{乗\,6}$
声が	聞こえた	0.9	$v_{聞\,3}$, $v_{聞\,4}$

この場合、“乗る”の係り先は“子供たち”と“声”の2通りあり、どちらも類似度計算の結果は閾値以上だが、“乗る”のもう一つの係り受け (風に乗る) と同じ意味 ($v_{乗\,1}$) を持っている方を採用し、“乗る”は“声”に係ると判定する。

4 実験

毎日新聞 (92年版) より抽出した複文50文に対して、格解析あり／無しの場合について構文解析を行なった。ただし今回は局所的な名詞・動詞の係り受け解析である (3.2 の **5.** の手法は実装中)。抽出した文は、格フレーム作成時に共起データが200以上あった動詞のうち共起データ数上位5つのものを含む文である。文法規則は、規則数110の係り受け文脈自由文法である。実験結果を表1に示す。ここで、類似度計算時のパラメータは、概念探索レベル = 4、類似度の閾値 = 0.5 として行なった。

5 実験結果に対する考察

実験の結果、構文木が半減もしくはそれ以下に削減できているものもあり、局所的な解析だけを行なったものとしては満足いく結果である。しかし、曖昧さが解消できない場合、正しい構造のものまで削減する場合も存在した。その理由としては以下のようなことが考えられる。

- 各係り受けの類似度がほぼ等しい場合。概念探索レベル、閾値を変化させても削減率を上げるのは難しい。これは局所的な係り受け解析の限界を示している。
- 閾値が高すぎる場合。閾値が高すぎて、正しい係り受けであるにも関わらず、適さないと判定してしまう。

動詞	格解析なし	格解析あり
出る	4	2
	1	-
	8	8
	44	24
	1	-
	28	8
使う	45	28
	8	4
	32	22
	10	8
	25	6
行う	54	9
	23	13
	14	6
	6	6
	4	3
	2	2
	9	9
	66	45
	29	11
持つ	6	6
	56	4
	7	2
	4	2
	5	5
	44	24
	1	-
	9	5
受ける	51	11
	3	3
	42	42
	87	58
	9	5

表 1: 実験結果

- EDR 電子化辞書に登録されていない名詞がある場合。格解析そのものが実行できない。
- 格フレームに記述している名詞概念集合内の情報が不足している場合。
- 入力文側の助詞と一致するものが格フレーム側に無い場合。
- 閾値・探索レベルを、全ての文で同じ値に設定したため。

6 おわりに

今回は、格フレームを用いてある一つの名詞がある一つの動詞に係り得るかという局所的な係り受け解析のみを行なった。その結果局所的な解析だけを行なったものとしては満足できる結果であった。だがうまく削減できない場合も存在した。そ

の原因としては格フレームの不備や、全ての文章で類似度の閾値・概念探索レベルを一定にしたことではないかと考えられる。今後は、閾値・探索レベルを変化させ、どのような値が解析対象文に適切なのかを調べる必要がある。また、局所的な名詞・動詞間の係り受けをだけを見て構造が正しいかどうか判断するのではなく、格フレームの利用手法 **6.** で述べたように、文全体の係り受けも見て判断するようにする予定である。

謝辞 本研究では、毎日新聞92年度版並びにEDR電子化辞書第1.5版を利用した。研究利用の便宜を計って頂いた関係者の方々並びに開発者の方々に感謝します。

参考文献

[1] 黒橋 禎夫 , 長尾 眞. “格フレーム選択における意味マーカと例文の有効性について”. 情報処理学会研究報告 NL 91–11 , pp.79–86, 1992.

[2] 黒橋 禎夫, 長尾 眞. “格構造解析への評価関数の導入による統語的曖昧性の解消”. 情報処理学会研究報告 NL 92-9, pp65–72, 1992.

[3] 日本電子化辞書研究所. “EDR 電子化辞書製品版 (第 1.5 版)”. 1996.

[4] 堤 豊, 堤 泰治郎. “統計データに基づいた構文解析のあいまいさ解消方式”. 電子情報通信学会論文誌, D–II, Vol.J72–D–II, No.9, pp.1448–1458, 1989.

[5] 松本裕治, 黒橋禎夫, 山地 治, 妙木 裕, 長尾 眞. “日本語形態素解析システム JUMAN Version 3.3”. 1997.

[6] 東 優. “既存の電子化辞書を用いた格フレーム獲得”. 九州大学大学院修士論文.1997.

[7] Tsunenori Mine , Masaru Higashi , Makoto Amamiya. “Case Frame Acquisition and Verb Sense Disambiguation on a Large Scale Electronic Dictionary”. Proc. of NLPRS97 , pp221–226 , 1997.