動詞語義例文と概念辞書を用いた動詞意味選択

3R－4

村木　一至　　落合　尚良
(株)日本電子化辞書研究所

1.はじめに

近年、自然言語処理の新しい手法として「事例（用例）に基づく自然言語処理」が数多く提案されている[1,2,3]。これらの多くは用例とシソーラスを用いて統語的な曖昧性や意味的な曖昧性を排除しようとするものである。その一つに語義例文とシソーラスを用いた動詞語義選択法の有用さが報告されている[2]。従来の格フレームと制約を用いた動詞語義選択だけでは不十分な語義選択能力を、動詞語義例文で補完しようとする方法である。

EDRの日本語単語辞書中に定義された高頻出動詞の語義数は数十を越えることがある。特定文中の解釈を下位範疇化属性中の選択制約のみで行うことは非常に困難といえる。そこで、EDR辞書の動詞語義にその語義を典型的に表わす例文を付与する事を考えた。

本稿では、先ず、動詞語義例文の作成方法を説明する。また、語義例文と既存のシソーラスを用いて動詞の語義選択実験を行なったので、その結果を報告する。

2.語義例文

EDR日本語単語辞書に登録されている動詞の中で、語義を２個以上持っているものは約5000語存在する。そこで、この語義に曖昧さのある動詞の語義毎（約15000語義）にその語義を典型的に表すような例文を付与した。例文は次のような記述指針の下、３名の機械翻訳用辞書作成経験者が作成した。

- ・該当する概念を典型的に表す単文であること
- ・用例文中の語彙は使用頻度の高いと思われる一般名詞を用いること
- ・単語辞書に付与されている文形情報に記述されている格要素をなるべく多く含むこと
- ・固有名詞を使用する必要がある時は国名のみとし、人名、企業名等は用いないこと
- ・否定を用いた例文を記述しないこと

実際に”固まる”に付与した語義例文を図１に示す。語義欄は、EDR辞書の"概念説明"、語義例文欄には、単語分割された文例が付与されている。

語義	語義例文
物事の状態などが確かなものになる	方針/が/固ま/る
一つの物事に夢中になって，他を顧みない	男/が/宗教/に/固ま/る
軟らかいものや液状のものなどがかたくなる	卵/が/固ま/る
人が一か所に集まる	兵隊/が/基地/に/固ま/る

図１　「固まる」の語義例文

3.語義例文を用いた動詞語義選択実験

語義例文を用いた語義選択の方法は既にいくつか報告[2,3]がある。

実験では参考文献[2]で与えられた方法を用いた。つまり、その方法は同一動詞表記を述語とする語義例文と入力文との間で、対応する格のヘッドの類似度をシソーラスを用いて計算し、その格要素毎の類似度の総和を入力文と語義例文との類似度とする。入力文と最も類似度が高い語義例文が付与されている語義を入力文に用いられている動詞の語義とする。

3.1 シソーラス

今回、類似度計算のためのシソーラスは以下の４種を用いた。

（１）EDR概念辞書Ver0.1
1989年にEDRで概念分類を行なったものである。６桁の分類番号を持つ。

（２）EDR概念辞書Ver0.5（1994年3月）
現在改良中の概念体系である。収録語彙数は約20万語。分類されている概念数は約45万。中間の分離ノード数は約6000である。

（３）分類語彙表[5]
国立国語研究所から提供されているシソーラス。5桁の分類番号、2桁の段落番号、2桁の段落内番号を持っている。今回は分類番号と段落番号の計７桁を分類番号とした。収録語彙数は約6万語。

（４）角川類語新辞典[6]
角川書店から発行されているシソーラス。大分類、中分類、小分類の3桁の分類番号と必要に応じてさらに細分化された4桁目を設けている。今回はこの4桁を分類番号とした。収録語彙数は約6万語。

Similarity based VerbSense Disambiguation using Verb Sense Example Sentences in EDR Japanese Word Dictionary
MURAKI Kazunori and OCHIAI Takayoshi
k-muraki@hum.cl.nec.co.jp, Japan EDR Institute Ltd.

実験(引用含む)	1	2	3	4	5[1]
語義例文	EDR	EDR	EDR	EDR	用法辞典[4]
対象シソーラス	EDR概念辞書Ver0.1	EDR概念辞書Ver0.5	分類語彙表	角川類語辞典	分類語彙表
実験文数	46	46	200	200	200
正しい語義が一つに選択できたもの	57%	63%	71%	65%	44%
複数選択された中に正しい語義があったもの	23%	4%	4%	15%	16%
選択された語義の中に正解がなかったもの	20%	33%	25%	20%	40%
シソーラスの分類の深さ語義例文	6	14	7	4	7

表１　実験結果

3.2類似度計算

各々の単語間の類似度は以下の距離原理によって得られた距離の逆数である。ここで用いるシソーラスはルートを持つループの無い有向グラフであるので、距離を以下のように考える。

- ２単語（語義）への有向パスがより多く共有されるとき、距離が小さい。
- 上記ルートから共有される有向パス上の最もルートから遠いノード（Ｐ）とそのＰより２単語（語義）へのパスの長さの和が小さいとき、より距離が近い。
- EDR辞書においては２単語の距離は各々の単語の語義（一般に複数）の間の最小距離とする。
- 入力文と語義例文の距離は各対応する各要素の距離の和である。

4.実験結果

EDR日本語単語辞書の中から５個以上の語義を持つ動詞１０単語を選び実験(表1)を行った。表１は分類語彙表[5]、角川類語新辞典[6]など既存のシソーラスとEDRの概念辞書を用い、EDR語義を選択した４つの実験結果と、シーソラスに分類語彙表、選択すべき語義を日本語基本動詞用法辞典[4]の（同辞典は一つの語義に例文を複数記載しているので入力文として選んだ語義例文とは異なる）語義例文とした文献[1]の結果５の引用を示す。EDR語義例文と用例辞典の語義例文の語義選択度は実験３と実験５の比較よりEDR例文の方が好ましい結果となっている。他方、４種のシソーラスを用いた同類語の選別性は実験２以外は概ね等しいレベルにあるといえる。実験２の選択語義中に正解がなかった例が増えた原因を分析した結果、以下のような事由があることが判明した。

EDR辞書Ver0.5の語義（この例では格要素のヘッドである名詞の語義）が多視点で分類されているがために、視点を無視した類似度計算法では、無闇に同類と判定する傾向がある。これによって、語義選択の類似度計算としては不都合が生じた。V0.1は、他のシソーラスと同様画一的な分類である。

5.まとめ

我々が付与した語義例文の付与基準と付与した語義例文の利用可能性を測定するために、簡単な語義選択実験を行ない、その結果を報告した。この結果、我々が付与した語義例文は、既に提案されている語義選択アルゴリズムに役立つ見通しが得られた。

また、我々の概念辞書中の概念分類２種（単一視点、多視）を語義選択の類似度計算に使用するとき、単一視点分類V0.1は他の市販シソーラスとほぼ同様な機能が期待できることが確認された。他方、より高度な応用が期待される多視点分類V0.5は、視点によって類似度計算を制御することにより、更に語義選択性の強化に寄与するものと考えられる。

[参考文献]

[1]村木一至、土井伸一、桧山努 "辞書中の語義例文に基づく事例ベース動詞意味選択" 人工知能学会第7回全国大会 17-2 1993

[2] 土井伸一、村木一至 "辞書に事例を付記することによる訳語選択・意味選択の強化" 情報処理学会第44回全国大会 1P-2 1992

[3] 黒橋禎夫、長尾眞 "格フレームの選択における意味マーカと例文の有効性" 情報処理学会自然言語処理研究会報告 91-11 1992

[4] 小泉保　他編　日本語基本動詞用法辞典　大修館書店(1989)

[5] 国立国語研究所　分類語彙表 秀英出版(1964)

[6] 大野晋、浜西正人　類語新辞典 角川書店(1981)