テキストからの共起関係自動抽出の試み

2E-6

中島 弘之 ・ 梶 博行
日立製作所システム開発研究所

1．はじめに

自然言語の解析や生成において、共起関係知識が有効であることが知られている。共起関係は、蓄積するデータ量を少なくするため、意味マーカーを用いて抽象的に表現されることが多い。共起関係を抽象化するためには、その基礎データとして、単語レベルの共起関係知識を蓄積しておく必要がある。また、抽象化できない、語に固有の共起関係も多く、単語レベルの共起関係の蓄積は重要な課題である。しかし、膨大な量の共起関係知識を効率良く獲得し、蓄積していく有効な方法は確立されていない。

本報告では、共起関係を自動的に抽出する方式を提案する。本方式は、テキストを解析し、係り受け関係に曖昧性のない共起関係のみを抽出することにより、誤った共起関係知識の獲得を回避するという特徴を持つ。この方式に基づき、翻訳システムが誤って解析する可能性の高いパターンの係り受け関係を検出する曖昧性検出文法を利用して、日本語共起関係自動抽出システムを試作し、抽出実験を行なった。

2．共起関係の利用法

「どのような語とどのような語が、ある意味的関係を持って同一文中に現われやすいか」という知識を共起関係知識と呼ぶ。共起関係知識は、自然言語処理システム、たとえば機械翻訳システムの、係り受け解析深層格解析や訳語選択などに有効である[1)]。日英翻訳における日本語の係り受け解析を例にとって、共起関係の利用法を説明する。

ここでは、共起関係は［述語、格要素、深層格］の3項関係で表現する。

仕様書を　設計した　人が　書く。（1）

機械を　　設計した　人が　書く。（2）

上の例文（1）、（2）は、それぞれ、連用修飾句「仕様書を」と「機械を」の係り先が、「設計した」か「書く」か曖昧である。このとき、

［書く、仕様書、対象格］

［設計する、機械、対象格］

という共起関係知識があれば、文（1）では、「仕様書を」の係り先は「書く」であり、文（2）では、「機械を」の係り先は「設計する」であると決定できる。

3．共起関係利用に関する問題点

共起関係知識は、2．で述べたように、翻訳システムにとって有効である。しかし、共起関係知識は、まとまった形で整理されておらず、計算機処理できる形に知識を蓄積・整理しておく必要がある。すなわち、共起関係辞書を作成する作業が必要である。共起関係は2つの単語と、その間の深層格関係の3項関係であるから、単語の数をm、深層格の数をnとすると、およそ$m^2 \cdot n$個のオーダーの数の共起関係があることになり、知識を効率良く獲得しないと、辞書作成の工数が膨大になることは明らかである。

共起関係知識を効率良く獲得する方法としては、助詞をキーワードとしたＫＷＩＣリストを作成することにより、テキストから共起関係を人手によって抽出する方法が提案されている[2)]。しかし、助詞を挟んで隣接する2語の間にどのような深層格関係があるか（あるいは、深層格関係がないか）は、人間が判断しなければならず、知識の自動抽出には至っていない。

4．共起関係自動抽出方式

共起関係知識を効率良く獲得するため、翻訳システムが誤って解析する可能性の高いパターンを持つ係り受け関係を検出する曖昧性検出文法を利用した、共起関係自動抽出システムを試作した。本システムは、日本文を解析し、解析結果である語と語の関係の集合の中から、曖昧性のない関係のみを抽出することにより、共起関係を自動的に獲得することを可能にしている。

本システムの共起関係抽出の過程を、例を挙げて説明する。

増幅器で　変調した　信号を　増幅する。（3）

例文（3）を解析する。解析結果は、図1に示した

Automatic Acquisition of Knowledge of Word Co-occurrences from Text
Hiroyuki NAKAJIMA and Hiroyuki KAJI
Systems Development Laboratory, HITACHI Ltd.

ような、内容語をノードとし、その間の意味的関係をアークで表したグラフ（概念依存図式[3]）で表現する。これを、語と語の共起関係の集合と考える。１つの語の係り先が複数考えられる場合（係り受け関係が曖昧な場合）は、係り先の候補をすべて求める[4]。例文（３）では、連用修飾句「増幅器で」の係り先は「変調した」と「増幅する」の２つあり、図１に示したように、両方の係り受け関係を求める。曖昧な共起関係の除去は、「複数のアークを出していないノードを格要素とする共起関係」のみを、解析結果である共起関係の集合の中から抽出することによって行なう。また、連体修飾の共起関係は、深層格が曖昧である場合が多いので、抽出の対象外とする（連体修飾の曖昧性は、連用修飾の共起関係を用いて解消できる）。図１の例では、共起関係、

［増幅する、信号、対象格］

のみが抽出される。

５．実験結果と検討

例文として、英語の文法書の日本語例文１３００文を用いて、共起関係の抽出実験を行なった。共起関係は、格要素が名詞、述語が動詞・形容詞・形容動詞のいずれかのもの、深層格が表１に示した、文を構成する上で重要な８つのものに限定した。連用修飾の係り受けの曖昧性と埋め込み文による連体修飾の深層格関係の曖昧性解消を、共起関係の主な利用目的に考えているからである。また、①補助動詞との区別が付きにくいため形態素解析で誤りが生じ易い、②ほとんどすべての語と共起するか、共起する語の意味属性が比較的明確であるので単語レベルの共起関係としてあまり有効でないという理由から、「有る」、「居る」、「する」、「作る」の４つの動詞を含む共起関係は除外した。抽出したデータのうち、正しく抽出できたもの、誤って抽出したもの（誤りの原因別に分類した）の比率を求め、表２にまとめた。また、共起関係抽出率（例文中に含まれる共起関係で上記の制限を満たすもののうち、何％が抽出されたか）を求め、表３にまとめた。

この結果、表２に示すように、正解率７９．５％で共起関係の自動抽出が行なえることがわかった。誤って抽出した共起関係のうち、係り受けの曖昧性の検出漏れによるものは４．２％であった。これは共起関係を蓄積し、日本語解析部にフィードバックすることにより、改善されていくものと考えられる。しかし、形態素解析と深層格解析で曖昧性を検出していないために生ずる誤り率が高く、これらの解析での曖昧性検出を行なうことが今後の課題である。また、誤って抽出したデータを、検出し除去することも必要である。共起関係の抽出率は表３に示すように、４７％であった。抽出率の低さは、構文的に曖昧性のない共起関係のみを抽出するという本方式の性質上やむを得ないと言えるが、これも蓄積した共起関係を曖昧性検出に反映させれば、次第に改善されていくと考えられる。

６．おわりに

係り受けに曖昧性のない共起関係のみを選んで共起関係を自動抽出するシステムを試作し、抽出実験を行なった。その結果、７９．５％の正解率、４７％の抽出率で共起関係を自動抽出できることがわかった。今後の課題は、①形態素解析、深層格解析での曖昧性検出を行ない、正解率を向上させること、②抽出した知識を翻訳システムの日本語解析で利用し、その効果を明らかにすること、③蓄積した共起関係知識を曖昧性検出にも反映させて、誤データを除去し、抽出能力を自動的に向上させる方式を確立することである。

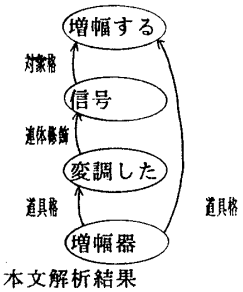

図１．日本文解析結果

表１．抽出対象の深層格

動作主格	対象格
場所格	起点格
終点格	受益者格
要素格	経験者格

表２．抽出した共起関係の分類

正解		643個	79.5%
誤り	形態素	48個	5.9%
	係り受け	34個	4.2%
	深層格	84個	10.4%
合計		809個	100.0%

表３．共起関係の抽出率

例文中の共起関係	1368個	
正しく抽出した共起関係	643個	抽出率　47.0%

参考文献

[1] 冨永ほか：英日機械翻訳における英文解析過程での多義の扱いについて、情報処理学会第３６回全国大会論文集、 pp. 1235~1236 (1988)

[2] 田中ほか：自然言語の分析における知識データ、情報処理学会自然言語処理研究会、54-3 (1986)

[3] 梶ほか：日立における機械翻訳システム、情報処理、Vol. 26、 No. 10、 pp. 1214~1216 (1988)

[4] 平井ほか：日英機械翻訳用前編集支援システム（１）、情報処理学会第３６回全国大会論文集、pp. 1229~1230 (1988)