共起データを用いた係り受け解析の学習効果

安原 宏
沖電気工業株式会社 マルチメディア研究所
yasuhara@okilab.oki.co.jp

1．はじめに

電子化辞書などの大規模な辞書資源の整備が進んでいる。一方、文法規則の大規模な開発は機械翻訳などで進んでいるが、保守性などで課題が多い。保守の困難性は文法の記述レベルがＮＰやＶＰといった抽象的な文法カテゴリで表現される事による。抽象の背後にある言語パターンの動きを理解した専門家でないと修正は困難である。抽象度が高まればそれがカバーする言語表現は多様なものになる。さらに自然言語には規則で記述することを拒むような例外現象が多発する。これも保守性を困難にしている要因である。もし文法規則が辞書データのように言語表現と直接的に対応し、レコードが相互に干渉しない形で独立に記述出来れば開発や保守が容易になる。

本論文においては、上記のような観点に立って日本語の係り受け解析を文節間の単純な共起データのみを用いて実現することを述べ、この解析を効果的に実現するために導入した学習機能とその評価について重点的に述べる。本アプローチは、いわゆる文脈自由文法的な深い入れ子構造の構文解析規則の類を用いた規則主導の解析方式ではなくデータ主導の方式ともいえる。また実テキストから大量のデータを収集するため統計的アプローチの1つと見なすこともできる。

2．縮退型共起関係

2．1共起データの位置付け

自然言語処理の言語資源として辞書と文法規則があるが、共起データは単語辞書の延長にあたる句構造辞書と見ることもできるし、単純なCFG規則と見ることもできる。つまり辞書と規則の中間に位置する言語資源である。この両面性は共起データの特徴にもなる。良い面としては作成や保守が容易になる反面、量が多く収束するかどうか保証できない点が挙げられる。

2．2 データ構造

係り受けの関係は、係りの文節と受けの文節から構成される。係りと受けの間の関係名を具体的に指示することも可能であるが、表層的な関係に徹すれば係り側の付属語なり、活用形で代用することも可能である。この方が関係名の揺らぎをなくする点で安定しているといえる。ここでは、山上,安原(1993)に示すような関係名を使用する。

係り受け関係の基本要素を、文節を品詞ＰＯＳ、付属語ｆ、係り受け関係Ｒｅｌ及び連続性Cを用いて、

（ＰＯＳi+ ｆi）＋（ＰＯＳj+ ｆj）＋Ｒｅｌ+C

例　解決すべき課題も　->　サ変+べき　+名詞+も　+連体格　+1
　　商品に取り入れる。->　名詞+に　+動詞+。　+ニ格　+1

のように表現する。ＰＯＳは記号を、小文字ｆはリテラルを表現する。連続性とは係り受けが連続して成立しているときに1、そうでないときは0とする。頻度も付与するがここでは省く。連続性や頻度は係り受けの結合度の情報として使用する。以下、縮退型共起関係あるいは単に共起関係と呼ぶ。「縮退」とは自立語をリテラルではなく品詞記号で代表させたからである。

これまで一般的に共起関係と呼ばれているものとは、受け側の文節が終止形や体言ではなく文中に出現した

生のｆｊを付与している点及び記号等も含んでいる点で異なる。これは、文節の分類を細かくするのに有効に作用する。

２．３規模

文節を構成するのは、自立語品詞と付属語列である。縮退文節のレコード規模は、自立語品詞数を１０、付属語列の数を１０００と仮定すると、約１０×１０００（1万）である。従って共起関係は関係子の違いを無視して最大１億に達する。付属語列は複合を許すから理論的には無限のパターンを生じる。しかし実際の用例はそれ程拡散しないと予想される。特にテキストを特定の分野に限定すると類似のパターンが出現しやすくなる。実際、付属語パターンは、システムの辞書には約７００種の付属語相当語が入っているが、新納、井佐原(1995)によると新聞１年分からは、２９６種の付属語相当語が抽出されたにすぎない。

３．係り受け解析方式

３．１概要

図１に示すように形態素解析には単語辞書を用い、係り受け解析には縮退型共起関係データベースを利用する。係り受け解析の結果は学習更新部に送られ、必要ならば縮退型共起関係データベースを更新する。

縮退型共起関係データベースは標準共起関係データベースを核として分野毎に共起関係データベースを差し替えたり、追加することにすると、基本語辞書と専門用語辞書の考え方に非常に良く似た取扱いが可能である。

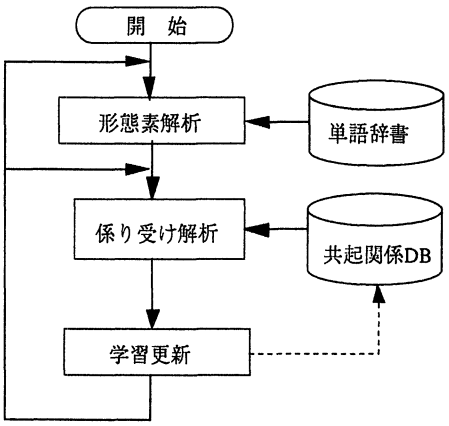

図１　係り受け解析システム構成

３．２学習メカニズム

あらかじめ初期状態として縮退型共起関係の集合ＣＲＧiniが与えられているとする。共起関係を多く含んでいるほうが解析率が高くなる。今、文Ｓに含まれる係り受け関係の集合をＣＲＧsとすると、ＣＲＧsがＣＲＧiniに全て含まれていれば、原則的には係り受け解析が成功する。もちろん係り受けの曖昧性を解消しないと正解を与えるとは限らない。もしＣＲＧsの係り受け関係ＣＲjがＣＲＧiniに含まれていなければその係り受け解析は導出不能である。

学習は係り受け解析が結果を出力した直後に画面上で行う。既存の共起関係集合に該当するものが無い場合に新たに共起関係を加えることと、係り受け関係で複数の曖昧な係り受け関係が生じたときに陽に正しい共起関係を指示することにより実行される。前者は新規の規則学習（登録）であり、後者は優先順位の変更で仮名漢字変換における学習機能に相当するものである。

形式的に書けば、文節を左からα、β、γで表わし、αがβにｒの関係で係るのを（α、β、ｒ）で表現すると、次の２通りである。（α、β、γは連続している必要はない。）

（１）α、β間に係り受け関係が共起データベースに無いとき：
　　（α、β、無）＝＞（α、β、ｒ0）を新規登録する。
（２）一方の係り受けを別のものに変更するとき：
　　（α、γ、ｒ0）＝＞（α、β、ｒ1）　あるいは
　　（α、β、ｒ2）＝＞（α、γ、ｒ3）

（２）では、左辺の古い関係の優先順位が右辺の関係より下がる。従って次回からは右辺が適用される。

4．学習効果

評価実験の初期状態として、実際のテキストを解析して収集した約８５００の共起関係からなる共起関係データベースを利用した。これまでにない新規のパターンが出現すると追加して行く。ここで興味があるのは、新規パターンの発生状況である。これまでと同一パターンが多いと新規登録の回数は減る。予想される事は、ジャンルが異なれば、新規パターンの蓄積が増大する事である。本システムでこれまで解析した文章は主として新聞記事、特に社説が多かった。従って、書き言葉という特徴を持ち用法も偏っている。新規に選んだ分野は、話し言葉が出てくるインタビュー記事１４３文である。特に文末の付属語表現に特徴が見られる。

4．1 インタビュー記事

（1）表現上の特徴

・会話文の特徴がある。
・質問部では疑問文や付加疑問文が出てくる。
・見出し部分でサ変名詞の体言止めや助詞省略による非文の表現がある。
・１文当たりの平均文節数は７．２であり、社説の平均８～９とそれ程変わらない。

（2）学習の具体例

以下のようなインタビュー特有の表現が未登録として出現した。

・～ています。～ていますね。～でもありますね。～ました。～ますが、～ますね。
・～されたんです。～なんです。～ですね。～でもあるのですね。～はずです。～かでしょう。～からです。～ですから。
・～ませんか。～ことはないでしょうか。～ですか。
・～た。～ていきたい。～のでは。　～てもらった。

これらは一般に１つの新規登録ではすまない。複数のパターンによる係りがあればその数だけ新規登録を必要とする。表１にはこのテキストを解析したときの文末のパターンとその出現回数を示している。例えば、「～ています。」は１４文、「～ですね。」は５文であるがそれらの新規共起関係のパターンを以下に示す。

（1）例１：「ています。」

１４文中、パターン総数２４個で、
／名詞を～ています。／４例　／名詞に～ています。／２例　／名詞は～ています。／２例
／動詞て～ています。／２例　（計１０例）

残り１４例は１回だけ出現したのパターンである。従って新規登録数は１８である。再利用は６例である。

（2）例２：「ですね。」

５文中パターン総数７個で、すべて異なっている。従って新規登録数は７である。再利用は０である。

他のパターンも総じて新規登録である。この程度の例文数では新規登録が多いのは当然であろう。

4．2　異なる方式の係り受け解析との比較結果

山上,安原(1993)による方式で、同じ文章に対して係り受け解析を試みた。一般的な係り受け規則で解析しているため、「です」「ます」調に対して新規に文法を追加する必要はなかった。係り受け総数８９３個中、６２個が失敗であった。７割の正解率である。

この方式で失敗した箇所から、本方式が効果を持っている点を以下に挙げる。（●は係り先の失敗箇所、○は正しい係り先き。）

・名詞について●連体詞～○名詞ですか。
・動詞た○名詞として～●。
・名詞については●名詞であると同時に、～○名詞であると～

表１　インタビュー記事１４３文中の文末パターン

受け側付属語パターン	出現数	受け側付属語パターン	出現数	受け側付属語パターン	出現数
用言終止形。	１６	～ない。	2	～かでしょう。	1
～ています。	１４	～ていました。	1	～ですけれども。	1
～です。	１３	～ませんか。	1	～てしまうほどでした。	1
～ます。	１１	～ません。	1	～てもらいたいものです。	1
～ました。	１１	～ませんわ。	1	～へ。	1
～ている。	6	～ていますね。	1	～だ。	1
～ですか。	6	～ましたが。	1	～のでは。	1
サ変名詞。	5	～ませんね。	1	～ていなかった。	1
～ですね。	5	～たんです。	1	～てもらった。	1
～ますが。	3	～んです。	1	～ことができる。	1
～ません。	3	～はずです。	1		
～でしょう。	3	～されたんです。	1		
～てくる。	3	～でした。	1		
～ますね。	2	～でもあるのですね。	1		
～であります。	2	～くらいです。	1		
～ていますか。	2	～なんです。	1		
～だけではありません。	2	～からです。	1		
～たい。	2	～ですから。	1		
～た。	2	～ことはないでしょうか。	1		
～ていきたい。	2	～そうです。	1	合計	１４３

・（並列句）Ａか○Ｂかで、Ａなど●Ｂの○Ｃから、Ａでの●Ｂの○Ｃは
・（連体形）名詞でも●動詞連体形 名詞を○動詞やすい。
　　名詞などで名詞と●動詞連体形 名詞が○動詞てくる。
　　動詞て●形容詞連体形 名詞を○動詞。
・動詞ほど●形容詞○動詞ている。
・名詞に○形容詞名詞が～●動詞そうです。

いずれも文法的には正しい係り受けであるが付属語のパターンまで考慮すると係り受けになりにくいものを避けることが可能になっている。

５．おわりに

本論文では、係り受け関係の共起データを収集することにより、それを係り受け関係の解析規則の集合と見なして解析する方式の概要を述べ、学習機能を導入したシステムを中心に記述した。与えられた分野のテキストを解析して蓄積した共起関係集合に対して、同一分野のテキストを解析するときの新規規則の学習量より、新たな分野のテキストを解析することによる新規規則の学習量の方が多いことは予想されることであるが、実際に新聞の社説等の文章を中心にしたものから、インタビュー記事を解析してみたところ予想通り新たな規則が学習でき、解析の後半においてはそれらのパターンが再現すれば学習規則が解析に作用していることが確認できた。しかしながら、１４３文程度の解析では新規登録が中心で効果の最終確認には至っていない。

参考文献

新納、井佐原(1995).コーパスからの付属語的表現の自動抽出.人工知能学会誌,Vol.10,No.3,pp429-435.
山上, 安原(1993). 形態素情報による日本語の係り受け解析. 情処学会自然言語処理研究会資料, 98-2, 9-16.