社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－170 (10)
2005／11／21

読点に頼らない統計的構文解析

金山 博
日本アイ・ビー・エム株式会社 東京基礎研究所
242-8502 神奈川県大和市下鶴間 1623-14
hkana@jp.ibm.com

概 要

日本語の統計的構文解析において、自立語の語彙の違いが統計モデル上で充分に反映されず、語彙選択を必要とする係り受けの解析誤りの原因となっている。本稿では、「既存の統計的構文解析器は、読点に過剰に依存している」という仮定に基づき、読点を無視して学習を行う統計モデルを構築して、用言に係る助詞句の係り受けの改良を図る。提案手法により、語彙を区別する素性の効用が増すとともに、不自然な読点が打たれている文に対しての頑健性が高まった。

Statistical Parsing without Commas

KANAYAMA Hiroshi
Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
hkana@jp.ibm.com

Abstract

In Japanese statistical syntactic parsing, the selection of content words does not have much effect on dependency decision between bunsetsus mainly because of the data sparseness. To overcome parsing errors caused by this lack of lexical information, this paper proposes a statistical learning method that ignores commas in sentences, drawing on the observation that the existing statistical parsers rely too much on such punctuation. This method increases the effect of features that distinguish among content words, and the model is robust for sentences where commas are not used properly.

1 はじめに

(1) **コンピュータが** 不安定で○ 困る。×

現実の文を統計的手法で構文解析したところ、このような誤解析があった。太字で書かれた文節の係り先が、○の文節であるべきであるのに、×の文節であると判定されてしまう。「コンピュータが」と「不安定だ」の親和性の高さが考慮されていないのだ。

各種の統計的構文解析器は、品詞や活用形などの各文節の属性と、文節間の距離を手掛かりとして、係り受けの傾向をコーパスから学習し、高い精度を実現している。属性の中でも特に「読点の有無」の影響は大きく、例(2)のように「不安定で」の文節に読点が打たれていれば、係り先が◎の文節であることが正しく判定される。

(2) **コンピュータが** 不安定で、◎ 困る。

読点一つで結果が変わるほど、学習に用いられる構文構造付きコーパスの中では、読点の位置と係り

受け構造との間に強い関連がある。これは、コーパス中の文の多くは新聞記事などの整った書き言葉であるため、文(2)のように、読みやすくするために適切な読点が用いられているからだと考えられる。

このようなコーパスを用いて学習・テストをしている限りは問題とならない。しかし、現実の文には必ずしも綺麗な読点が打たれていることが期待できない。特に、webの掲示板への書き込みや電子メールといった、充分な推敲を伴わないテキストでは、文(3)(4)のように、不自然な読点が打たれた文が散見される。そのため、用言を中心とした句構造を扱う「評判分析」[3]などのアプリケーションにおいて、構文解析が原因の句構造の誤認識が大きな障害となっている。

(3) **コンピュータが、** 不安定で○ 困る。×
(4) **部屋が、** 小さいけど○ そこに 決めた。×

文(1)や文(3)を、文(2)と同様に正しく構文解析するには、各文節の読点に着目するのでなく、格助詞と用言の組など、語彙による違いを考慮する必要がある。しかし、学習に用いるコーパスにおいて、読点が構文構造に強い影響を及ぼしている限り、語彙情報に重みを与えるのは難しい。もちろん、読点の使われ方の傾向が異なる文からなる構文構造付きコーパスを別途用意するというのも現実的ではなかろう。

本稿では、「既存の統計的構文解析器は、読点に過剰に依存している」という仮定に基づき、読点の有無を考慮しない統計的構文解析を試みる。そして、アプリケーションの性能を大きく左右する「助詞句→用言」の係り受けの改良を図る。これにより、通常の手法と比較して、統計モデルと整合性を持つ形で取り入れられた語彙情報の効用が大きくなり、さらに、不自然な読点が打たれている文に対する頑健性が高まることを示す。

まず、2節で、統計的構文解析、大規模語彙情報の利用に関する関連研究について説明する。3節では、文中で読点が使われる割合、文書タイプによる使われ方の違いを観察した結果を記す。4節にて、本稿で提案するモデルにおける読点の扱い方、語彙情報の取り入れ方について述べる。5節で、解析精度の測定を行い、本手法の有効性について議論する。

2 関連研究

本節では、既存の統計的構文解析の諸手法とその中での語彙情報や読点の効用、また大量の語彙情報を構文解析に利用する研究について述べる。

これまで考案されてきた各種の統計的構文解析の手法[11, 2, 12, 8, 9]では、品詞・活用形・助詞の種類・文節間の距離・読点の有無等の属性と、少量の高頻度の語彙情報のみを用いて、高い係り受け精度を実現している。

概して、読点の効用は大きい。内元ら[4]の実験では、係り元や係り先の読点の有無を素性として学習に用いない場合に、それぞれ1.7%、2.6%の精度低下が見られた。春野ら[2]の場合はそれぞれ1.2%〜1.6%、藤尾ら[11]の場合は1.4%と、読点の効用の大きさは共通している。

一方で、個々の自立語の区別は、統計モデルの中で重要な要素とはなっていない。内元ら[4]の実験結果では、自立語の語彙に関する2000個以上の素性を使用しない時にも、精度は0.1%〜0.5%しか低下していない。データスパースネスに対処するために、シソーラス等を導入することも考えられるが、春野らの実験[2]では、分類語彙表[10]の1桁・2桁を素性として加えた場合に、逆に精度が低下したことが報告されている。

また、構文構造付きコーパスとは別の大規模なリソースから獲得した語彙情報を構文解析に適用させる試みがなされている。河原ら[6]は、自動構築した格フレームを、ルールベースの構文解析器の格解析部分に適用させ、精度向上を図っている。また、阿辺川ら[5]は、大規模コーパスから獲得した用言と格との共起情報を用いて、統計的構文解析の結果をre-rankingするモデルを提案している。いずれも、大量の語彙を扱うことができ、的確に言語特性を捉えられる語彙情報を取り入れているものの、係り受けの選択に用いられる数値と独立であることもあって、構文解析の精度向上の度合いは限定的であった。

一方、このように外部リソースの語彙情報を統計モデルに取り入れる手法は、データスパースネスを防ぐ意味でも非常に重要である。本稿では、筆者が以前に提案した3つ組／4つ組モデルの拡張版[7]と同様に、共通の特徴を持つ語彙をまとめて素性化したものを学習に用いる。詳しくは4.2節で述べる。

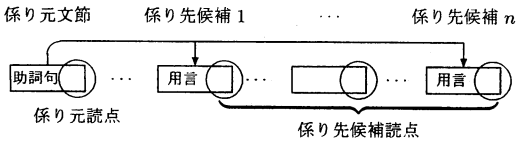

図 1: 読点が使われているか否かの判定方法の概念図。「係り元読点」は、係り元の文節に読点があるか、「係り先候補読点」は、最も近い係り先候補の文節と最も遠い係り先候補文節との間の一つ以上の文節に読点があるかを示す。なお、3つ組／4つ組モデル [8] の手法により係り先候補が絞り込まれていることを前提としている。

3 読点に関する調査

1節の例で見たような、用言に係る「名詞＋助詞」からなる文節の係り受けが、どれくらいの割合で読点によって制御されているかを、EDR コーパス [1] のうち 3,390 文を用いて調査した。ここで扱う用言に係る助詞句のことを、以後は**連用格助詞文節**と呼ぶことにする。これは、用言を修飾する格助詞ないし副助詞・係助詞が文節の末尾（読点を除く）に付く句である。

表 1 は、連用格助詞文節が文法的に係り得る文節が 2 つ以上ある場合に、「係り元読点」または「係り先候補読点」のいずれかが存在する割合である。読点の有無の判定方法を図 1 に示した。

連用格助詞文節全体で見ると、65.0% の場合において読点が用いられている。特に、係り先候補がすべて動詞の場合は、その割合が 67.7% に上がり、同種の語に係り得る場合に読点を用いて係り先を明らかにする傾向を示唆している。また、提題を表す「は」にも読点が用いられることが多く、係り先候補がすべて動詞の場合の読点の使用率は 80% を超えている。

さらに、読点の打ち方にどれだけの揺れがあるかについてを知るために、5 つの連用格助詞文節が含まれる文を新聞記事から 10 文、web 掲示板への投稿から 10 文抽出し、連用格助詞文節の読点を消したものに対して、被験者 10 人に読点を振らせる実験を行った。なお、原文中では、それぞれ 50 の連用格助詞文節のうち、新聞は 10 箇所、web 掲示板は 7 箇所に読点が振ってあった。すると、原文と読点の有無が原文と一致する割合は、新聞の文では 87.4%、web 掲示板の文では 86.8% であり、大きな差は無かった。しかし、被験者のうち 9 人以上が同一の判定をしたものと、原文の読点が一致していないケースは、新聞の文で 50 件中 1 件だけなのに対して、web 掲示板では 5 件あった。これは、web 掲示板では直感に反する読点の用法が多いことを裏付けている。

係り元	係り先候補	読点ありの割合	
が	動詞	606/872	(69.5%)
が	用言	686/1046	(65.6%)
を	動詞	890/1475	(60.3%)
を	用言	932/1610	(57.9%)
に	動詞	782/1259	(62.1%)
に	用言	857/1420	(60.4%)
は	動詞	930/1147	(81.1%)
は	用言	1117/1444	(77.4%)
全体	動詞	3918/5784	(67.7%)
全体	用言	4378/6735	(65.0%)

表 1: 連用格助詞文節の係り先候補が複数ある場合に、図 1 に示される係り元読点または係り先候補読点が存在する割合。「係り元」は連用格助詞文節の助詞、「係り先候補」の「動詞」「用言」は、最大 3 つに絞り込んだ係り先候補が、それぞれ「すべて動詞の場合」「すべて用言（動詞・形容詞・形容動詞）」の場合を表す。

4 読点の無視と語彙の導入

ここでは、本稿で提案する、連用格助詞文節の係り受けの問題を解決するための手法について述べる。

4.1 読点の寄与の緩和

表 2 に示す、3 つ組／4 つ組モデルの学習に用いている素性群のうち、読点が関係するものは、素性番号 6・13・18・27・30 である。これらの素性の寄与を緩和する手段として、以下の 3 つを考えた。

A 名詞類以外に付く読点をすべて無視する

B 係り元が連用格助詞文節の時に限って、名詞類以外に付く読点をすべて無視する

C 係り元が連用格助詞文節の時に限って、連用格助詞文節に付く読点を無視する

いずれの場合も、素性自体を削除するのではなく、対象となる読点を無いものとみなして素性の値を求める。例 (5) に見られるような並列を表す読点の用法は本稿の議論の対象外であるため、名詞類に付く読点は常に無視しないようにしている。

素性番号	素性の種類	異なり数
1	係り元係り属性	8
2	係り元主辞品詞	26
3	係り元語形品詞	45
4	係り元活用形	6
5	係り元助詞	41
6	係り元読点の有無	2
7	係り元副詞語彙	51
8	係り先受け属性	7
9	係り先主辞品詞	26
10	係り先語形品詞	45
11	係り先活用形	6
12	係り先助詞	41
13	係り先読点の有無	2
14	係り先引用「と」の有無	2
15	係り先類似主辞の有無	2
16	係り先文末	2
17	係り先主辞語彙	51
18	文節間読点の数	3
19	文節間「は」の数	3
20	文節間係り元同一語形	2
21	文節間係り先同一主辞	2
22	誤解析に基づいて追加した素性	25
23	1・8 の組み合わせ	
24	3・9 の組み合わせ	
25	5・8 の組み合わせ	
26	3・10 の組み合わせ	
27	3・6・13 の組み合わせ	
28	3・12 の組み合わせ	
29	5・12 の組み合わせ	
30	3・6・9・16 の組み合わせ	
31	3・4・10・11・16 の組み合わせ	
32	5・17 の組み合わせ	

表 2: 3つ組／4つ組モデルにおいて用いている素性群。8番〜21番の素性は、係り先に関する素性なので、2つまたは3つの係り先候補に対して別々に考える。22番は、文献 [7] で追加した、誤解析主導で追加した素性。

(5) 彼はフランス、ドイツを旅した。

B、C では、係り元文節が連用格助詞文節である場合のみ読点を無視するようにしているが、実行時に各係り受けの確率を独立に求めるため、係り元文節の属性に応じて素性の定義を切り替えても問題はない。また、文献 [7] にて提案した「モデルの分割」により、係り元文節の種類に応じて別の統計モデルを用いることによって、他のモデルに影響を与えることなく一部のモデルを変更することが可能である。

表 3 は、通常の手法（読点有）と、上記 3 つの方法で読点を無視した場合の解析精度を比較したものである。学習には EDR コーパス [1] の 192,725 文を、テストには同コーパス中の別の 3,390 文[1]を用いた。

[1]3 節での観察に用いた文集合とは異なる。

	全体	連用格助詞文節
読点有	88.87%	90.24%
読点無 A	87.78% (−1.09%)	89.04% (−1.20%)
読点無 B	88.33% (−0.54%)	89.14% (−1.10%)
読点無 C	88.45% (−0.42%)	89.33% (−0.91%)

表 3: 通常と同様に読点を利用した場合（読点有）と、読点の寄与を緩和する 3 種類の方法（読点無 A〜C）による解析精度の比較。括弧内は読点有の場合との精度の差。

手法 A、B は精度の低下が著しい。次節で導入する語彙素性は係り元が連用格助詞文節の場合のみに影響を与え、用言連用形や副詞など、格助詞句以外に付与される読点が関係する事象の解決に寄与しないことから、以降では手法 C を用いることにする。

4.2 語彙素性の追加

連用格助詞文節の係り先を、用言の語彙に応じて制御するために、「簡易格フレーム」を用いる。これは、助詞 P と用言 V に対して、「V が常に必須格 P を取る（○）」「V が P を必須格に取る格フレームがある（△）」「V が P を必須格に取らない（×）」のいずれかの値を持つものである。必須格の有無で傾向が分かれやすいヲ格・ニ格・ト格（引用助詞を除く）のエントリを作成した。さらに、ガ格の体言部分が「人間」に限定したエントリを加えた。

表 4 がエントリの例である。4 種の格と 16,129 の用言の組み合わせの、合計 64,156 エントリからなるリソースを、日英機械翻訳における用言句の翻訳パターンから変換することによって作成した。

簡易格フレームの情報を用いて、係り受けを特徴付ける以下の語彙素性 F_1 を導入する。

F_1 係り元文節が連用格助詞文節で、
助詞が【を | に | と】であり、
係り先候補文節が 【動詞 | 形容詞 | 形容動詞】で、
係り元の助詞・係り先候補の用言に対応する
簡易格フレームの値が【○ | △ | ×】である

【A|B|C】と表記された部分は、選択によって別の素性となる要素を表す。例えば「係り元文節が “を” の助詞句、係り先候補文節が形容詞、対応する簡易格フレームの値が△である」が一つの素性の条件であり、このような 27 通りと「F_1 に該当しない」ことを示す 1 通り、合計 28 通りの素性関数がある。同

助詞	用言	値
を	取り扱う	○
を	取り込む	△
を	就労する	×
に	程近い	○
に	間違う	△
に	監督する	×
と	折衝する	○
と	矛盾する	△
と	引く	×
HUM-が	踏み外す	○
HUM-が	いい加減だ	△
HUM-が	梅雨明けする	×

表 4: 簡易格フレームのエントリ例。助詞と用言に対して○・△・×の値を持つ。

じ傾向を持つ語彙をまとめて素性として、個々の語彙をすべて区別しているわけではないので、同一の語が学習コーパスに出現していなくても良い。

さらに、ガ格の名詞を区別するために、以下の F_2 を導入する。上記同様に考えて、55 通りの素性関数が生成される。

F_2 係り元が連用格助詞文節で、助詞が【が | は】で、
係り元文節の主辞の名詞の意味クラスが
【人間 | 組織 | その他】であり、
係り先候補文節が 【動詞 | 形容詞 | 形容動詞】で、
「が-HUM」・係り先候補の用言に対応する
簡易格フレームの値が【○ | △ | ×】である

なお、いずれの場合も、表層格に影響を与える「れる」などの助動詞が用言に付いている場合を除外するといった、副作用を防ぐ処理を行っている。

5 実験

5.1 語彙素性の効用

4.2 節で導入した語彙素性を導入することの効果について、通常の読点を用いて学習した場合と、4.1 節の C の方法で読点の寄与を緩和した場合について実験を行った。学習・テストには 4.1 節と同じコーパスを使用し、係り元が連用格助詞文節の場合の精度を測定した結果を表 5 に示す。

読点の効用を緩和することによって、語彙素性の効用が 2.5〜2.8 倍になっていることがわかる。テストコーパスを 10 分割して t 検定をしたところ、「読点有」で、「語彙無」と「語彙 F_1」の差の場合を除いては、5% の有意水準による精度の差が認められた。

	読点有	読点無 C
語彙無	90.24%	89.33%
語彙 F_1	90.30% (+0.06%)	89.50% (+0.17%)
語彙 F_1+F_2	90.37% (+0.13%)	89.66% (+0.33%)

表 5: 語彙素性・読点の有無による、連用格助詞文節の係り先の精度の変化。括弧内は、語彙情報を用いない場合との精度の差を表す。

読点を無視することによって語彙素性の効用が現れた例を表 6 に挙げる[2]。1 の文では、「結び、」の文節の読点が節の区切りを強調してしまい、「に→分け」の共起の強さが埋もれてしまっていたのが、読点を無視することによって正しく求まるようになった。2 や 3 の例は、読点が用いられていないケースである。それぞれ「限定して」「苦情に」の後に読点が打たれていた方が自然といえる文である。読点有の場合は、語彙素性の重みが小さくなったため、統計値を逆転させるに至らなかったが、3 では「に→過ぎる（簡易格フレームで「△」）」と「に→反論する（簡易格フレームで「○」）」の事象の差が反映されており、通常では区別されない微妙な語彙選択の現象がモデルに取り込まれていることがわかる。

5.2 不自然な読点が打たれた文への頑健性

不自然な読点が用いられている文に対しての、提案手法の有効性を検証する。そのために、EDR コーパスの文中の連用格助詞文節の読点の有無を様々な割合で入れ替えた文を生成して実験を行った。

結果を図 2 に示す。p の値が増すにつれて精度が下がるが、読点無の場合は精度の低下が僅かになっており[3]、語彙無の場合は $p \geq 0.08$、語彙有の場合は $p \geq 0.06$ の時に、読点無の精度が読点有の精度を上回っている。このことから、読点に頼らないモデルは、読点の打ち方の変化に対して耐性があること、語彙素性の効果を高めることがわかる。

なお、3 節の実験では、web 掲示板の文の中には、連用格助詞文節の読点の有無の 10% が多くの被験者の直感と異なっていた。直接の比較はできないものの、解析対象の文の特徴によっては、読点無の方が実用的な解析ができると考えられる。

[2]これらは、EDR コーパス上での実験結果から抽出した文を基にした、同様の事象となる作例である。

[3]連用格助詞文節以外のモデルは読点の影響を受けるので、読点無の場合も精度は一定とはならない。

1	昨年、 **2回に** 分け○ 契約を 結び、× １０万ドルを 支払った。
2	規制の **対象を** 駅に 近い 店舗に 限定して○ 一見 解決したように 見える。×
3	しかし、 読者の **苦情に** 出版社は あくまで 思いこみに 過ぎないと× 反論している。○

表 6: 読点を無視することによって語彙素性の効用が現れた例。読点有では語彙素性の有無にかかわらず上記の誤りが見られたが、読点無の場合は語彙素性を導入すればこれらが正しく解析される。

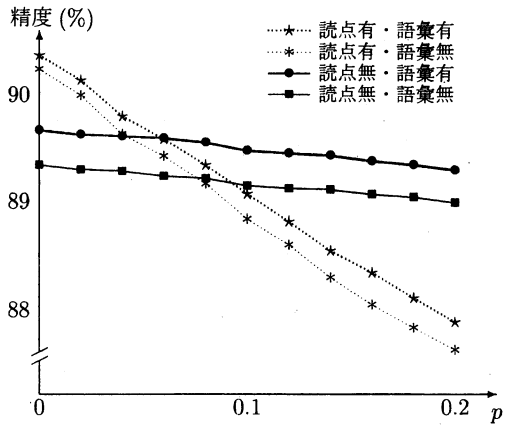

図 2: 連用格助詞文節の読点の有無を確率 p で入れ替えた時の、連用格助詞文節の係り先の精度の変化。

6 まとめと今後の展望

実験により、連用格助詞文節に付く読点を考慮しない統計的構文解析器では、語彙素性を導入した時の効果が高まることが確認できた。特に、今回用いた語彙素性は、単純で直感的なものであり、実用上問題となるような構文解析の誤りを防止するべく人手で修正することができるという利点がある。

通常の構文構造付きコーパスを用いて学習・テストを行うと、読点を無視したモデルでは解析精度が低下する。しかし、6% 以上の該当文節における読点の有無をランダムに入れ替えた文に対しては、読点を無視した場合の精度が従来手法を上回り、提案手法の読点の違いに対しての頑健性が示された。

今後はより詳細な格フレーム情報を取り入れて、本稿で扱っていない格の相互関係や名詞の種類による傾向の違いを反映させられるようにしたい。また、読点の無視によって、外部リソースから得た語彙情報の効用が増すため、格フレームや意味的情報を構文解析に適用させる研究が加速されると考える。

参考文献

[1] EDR. EDR (Japan Electronic Dictionary Research Institute, Ltd.) electronic dictionary version 1.5 technical guide, 1996. Second edition is available via `http://www.iijnet.or.jp/edr/E_TG.html`.

[2] Masahiko Haruno, Satoshi Shirai, and Yoshifumi Ooyama. Using decision trees to construct a practical parser. In *Proc. COLING-ACL '98*, pp. 505–511, 1998.

[3] Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe. Deeper sentiment analysis using machine translation technology. pp. 494–500, 8 2004.

[4] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. Japanese dependency structure analysis based on maximum entropy models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 196–203, 1999.

[5] 阿辺川武, 奥村学. 大規模統計情報を用いた日本語係り受け解析の精度向上. 言語処理学会第 11 回年次大会発表論文集, pp. 919–922, March 2005.

[6] 河原大輔, 黒橋禎夫. 大規模格フレームに基づく構文・格解析の統合的確率モデル. 言語処理学会第 11 回年次大会発表論文集, pp. 923–926, March 2005.

[7] 金山博. 統計的日本語構文解析器の部分的修正. 情報処理学会第 160 回自然言語処理研究会, pp. 1–8, 2004.

[8] 金山博, 鳥澤健太郎, 光石豊, 辻井潤一. 3つ以下の候補から係り先を選択する係り受け解析モデル. 自然言語処理, Vol. 7, No. 5, pp. 71–91, 2000.

[9] 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[10] 国立国語研究所. 分類語彙表, 1964.

[11] 藤尾正和, 松本裕治. 統計的手法を用いた係り受け解析. 情報処理学会第 117 回自然言語処理研究会, pp. 83–90, 1997.

[12] 内元清貴, 村田真樹, 関根聡, 井佐原均. 日本語係り受け解析に用いるMEモデルと解析精度. 言語処理学会第 5 回年次大会ワークショップ論文集, pp. 41–48. 言語処理学会, March 1999.