節内と節間の整合性をとる構文・格解析

河原 大輔
独立行政法人 情報通信研究機構
dk@nict.go.jp

黒橋 禎夫
京都大学大学院情報学研究科
kuro@i.kyoto-u.ac.jp

1 はじめに

係り受け解析の手法は、基本的には二つの文節間に係り受けが成立するかどうかを判定しており、文の係り受け構造はその判定結果を集めたものとなっている。対象となっている文節の周囲の情報を可能なかぎり考慮することが重要であるが、節や文などの広い範囲で言語的に正しいかどうかを判定することは難しい。

我々は、述語項構造を単位として評価を行うことによって、この問題に取り組んでいる [9]。述語項構造は「誰がどこで何をした」のような事態を表す基本単位である。述語項構造の正しさの評価は、大規模コーパスから述語項構造を収集・集約した「格フレーム」に照らし合わせることによって行っている。これにより一定の精度向上は達成しているが、述語項構造内の整合性は考慮しているものの、述語項構造間の整合性を考慮しているわけではない。たとえば、次の文において「ポイントは、」の係り先は「ことです」が正解であるが、「まとめて」に係るように誤って解析される。

(1) ポイントは、 一つに まとめて× 宅配便で送る ことです○

これは、「ポイントは (を) 一つに まとめる」という述語項構造が高頻度に表れる表現だからである。しかし、「一つにまとめて宅配便で送る」という二つの節を考えると、「まとめる」の対象は荷物などにしかなりえず、「ポイント」にはならないことがわかる。また、「ポイントは (が) ことだ」という述語項構造はまったく問題なく言えるため、「ポイントは、」の係り先は「ことです」になることがわかる。このように、節 (述語項構造) 内と節間の整合性を考慮することができれば、係り受け解析の精度向上が期待できる。

本稿では、節内と節間の整合性を考慮した確率的構文・格解析手法を提案する。本手法は、節内は従来の格フレームに基づく格解析、節間は格フレーム間関係知識に基づいて評価を行う。格フレーム間関係知識とは、大規模コーパスから係り受け関係をもつ格フレームペアを集めたものである。本手法は、節内における格フレームの選択と、節間における格フレーム間の選好がうまく整合するような構文・格構造を選ぶモデルとなっている。

2 格フレーム間関係知識の自動獲得

本研究では、大規模コーパスから連用修飾関係をもつ節ペアを収集し、それらを格フレーム ID のペアで表すことによって、格フレーム間関係知識とする。たとえば、コーパス中に「荷物をまとめて、宅配便で送った」という文があれば、(まとめる:5, 送る:6) のような格フレームペアを抽出する。ここで「まとめる:5」や「送る:6」は格フレーム ID を表しており、「まとめる」「送る」のそれぞれについて複数個存在する格フレーム (表 1) の中から「荷物をまとめる」「宅配便で送る」にもっとも近い格フレームを選択した結果となっている。この格フレームの選択は、既存の確率的構文・格解析 [9] を適用することによって行う。抽出した格フレームペアは、3 節における解析で用いるために、格フレームペアごとに頻度を計数しておく。

実際に、ウェブテキスト 1 億文の構文・格解析結果から格フレーム間関係知識を獲得した。用いた格フレームは、[8] の手法によりウェブテキスト 16 億文から自動構築したものである。この格フレームは 4.3 万個の述語からなり、一つの述語あたり平均 22.2 個の格フレームをもつ。獲得した格フレーム間関係知識は、1.1 億異なりの格フレームペアとなった。その例を表 2 に挙げる。

これまでにも、このような推論的な知識をコーパスから自動獲得する研究が行われている。たとえば、含意・同義知識 [4, 6, 5, 7]、動詞関係知識 [2]、因果関係知識 [10]、事態間関係知識 [1] などの自動獲得である。獲得した知識は、質問応答、情報抽出、情報検索、

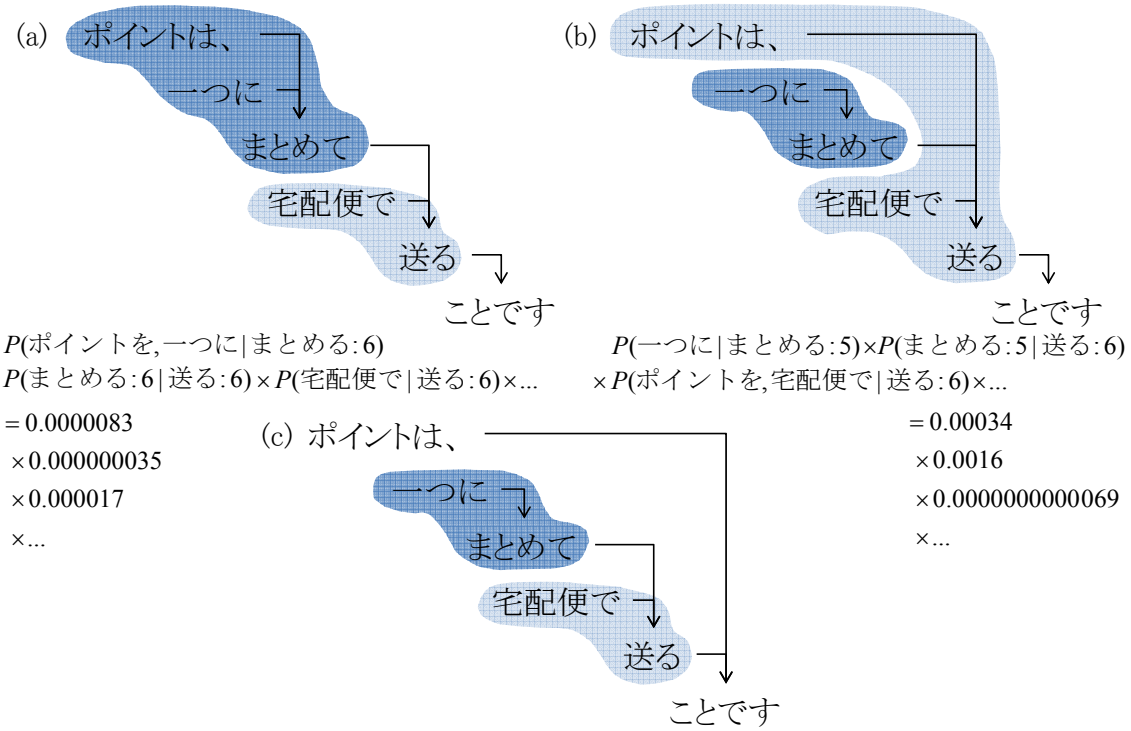

図 1: 例文 (1) の係り受け構造の曖昧性と確率計算例

表 1: 「まとめる」「送る」の格フレームの例

格フレーム ID	格	用例
⋮	⋮	⋮
まとめる:5	ガ ヲ ニ	私, 人, ... 荷物, 荷, 貴重品, ... <数量>つ, コンパクト, ...
まとめる:6	ガ ヲ ニ	博士, ... ポイント, 見分け方, 重要点, ... <数量>つ, 以下, ...
⋮	⋮	⋮
送る:1	ガ ヲ ニ デ	人, 私, ... メール, メッセージ, 情報, ... 友達, アドレス, 方, ... メール, 郵送, 郵便, ...
⋮	⋮	⋮
送る:6	ガ ヲ ニ デ	女性, ... 荷物, 物資, 品物, ... 人, 日本, 実家, ... メール, 郵便, 宅配便, ...
⋮	⋮	⋮

表 2: 自動獲得した格フレーム間関係知識の例

格フレーム ID のペア	頻度
(一緒だ:10, 送る:6)	26
(送る:1, 送る:6)	22
(ある:1, 送る:6)	21
⋮	⋮
(まとめる:5, 送る:6)	3
⋮	⋮

含意認識などのシステムで利用され始めている。たとえば VerbOcean[2] は、検索エンジンを用いることによって、数個の動詞間関係について 2.9 万個の動詞ペアを獲得している。それと比べると、本研究で獲得した格フレーム間関係知識は、その数千倍であり、かなり大規模に獲得できたことがわかる。

これらに対して、本研究では、コーパスに対して格解析を適用し格フレームを同定している。これは、述語の曖昧性解消を行っていることを意味しており、従来研究では単位として動詞または動詞句を対象としていたのに比べて、曖昧性を解消している点で異なる。また、動詞で扱うよりも、格フレームという汎化され

た形で扱っているために、データスパースネスに強いと考えられる。

3 節内と節間の整合性をとる構文・格解析

本研究では、節内と節間の整合性をとる確率的生成モデルを提案する。図 1 に、例文 (1) における「ポイントは、」の係り先の曖昧性と、それぞれの構造に対する生成確率の例を示す。図中の (a) において、P(まとめる:6| 送る:6) は格フレーム間関係の生成確率を表しているが、コーパス中に「{ ポイント, 見分け方, $\cdots$} を まとめる」と「{ 荷物, 物資, $\cdots$} を 送る」が関係をもつことがほとんどないため、生成確率が非常に低くなる。また、(b) において、「ポイントを宅配便で 送る」という述語項構造はほとんど言われない表現であるため生成確率が低い。この結果、構成する生成確率のすべてが高い値をもつ係り受け構造 (c) が解として選択される。

提案モデルは、入力文 S が与えられたときの構文構造 T と格構造 L の同時確率 $P(T, L|S)$ を最大にするような構文構造 T_{best} と格構造 L_{best} を出力する。次のように、$P(S)$ は一定であるので、本モデルは $P(T, L, S)$ を最大にすることを考える。

$$\begin{aligned}(T_{best}, L_{best}) &= \operatorname*{argmax}_{(T,L)} P(T, L|S) \\ &= \operatorname*{argmax}_{(T,L)} \frac{P(T, L, S)}{P(S)} \\ &= \operatorname*{argmax}_{(T,L)} P(T, L, S)\end{aligned}$$

本モデルは「節」を基本単位とし、主節 (文末の節) から順次生成していく。本論文における節とは、述語を一つ含みそれに関係する格要素群を含む部分 (述語項構造) および連体修飾句の 2 種類と考える。$P(T, L, S)$ は、文 S に含まれる節 C_i を生成する確率の積として次のように定義する。

$$P(T, L, S) = \prod_{C_i \in S} P(C_i|C_h)$$

ここにおいて、C_h は節 C_i の係り先の節である。主節は係り先をもたないが、仮想的な係り先 EOS をもつとする。

確率 $P(C_i|C_h)$ は、[9] をベースに定義するが、異なる点は格フレーム生成確率、つまり格フレーム CF_h から格フレーム CF_i を生成する確率である。従来は、格フレーム生成確率を次のように、動詞 v_h から動詞 v_i を生成し、動詞 v_i から格フレーム CF_i を生成する確率に近似していた。

$$P(CF_i|CF_h) \approx P(v_i|v_h) \times P(CF_i|v_i)$$

提案手法では、確率 $P(CF_i|CF_h)$ を直接計算し、格フレーム間の整合性を考慮する。この確率は、格フレーム間関係知識から計算することができる。

実際には、データスパースネスに対処するために、述語生成確率 $P(v_i|v_h)$ を利用して補間を行う。

$$P'(CF_i|CF_h) \approx \lambda P(CF_i|CF_h) + (1 - \lambda)P(v_i|v_h)$$

λ は、[3] と同様に、格フレームペア (CF_i, CF_h) の頻度に基づいて決定する。

4 実験

4.1 係り受け解析の精度評価

提案手法による解析実験を行った。格フレームは、ウェブテキスト約 16 億文から構築したものを用いた。

本実験は、ウェブテキスト 759 文*を形態素解析器 JUMAN†に通した結果を提案システムに入力することによって行う。その 759 文には、京大コーパス‡と同じ基準でタグ付けを行っており、これを用いて評価を行う。評価の対象としては、文末から二つ目までの文節以外の係り先すべてとした。

ベースラインとしては、構文解析器 KNP§と、確率的構文・格解析 [9] の二つを用いた。

表 3 に評価結果を示す。表において、「構」は構文解析器 KNP、「構+格」は確率的構文・格解析 [9]、「構+並+整」は提案手法を表している。提案手法の精度は、「構」「構+格」のそれぞれに対して、すべての係り受けでは 0.8%、0.4%、体言から用言に係る文節のみでは 1.7%、0.6%向上した。

4.2 議論

表 4 に、「構+格」では誤りになるが、提案手法によって正解になった例を挙げる。四角形で囲まれた文節の係り先が×下線部から○下線部に変化したことを示している。たとえば (a) の例では、「ぐねぐね道を」が正しく「揺られる」に係るように解析されるように

*この文は格フレーム構築とパラメータ推定には用いていない。
†http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
‡http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
§http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 3: 係り受け構造の精度

評価対象	構	構+格	構+格+整
すべて	3,832/4,400 (87.1%)	3,852/4,400 (87.5%)	3,866/4,400 (87.9%)
体言 → 用言	1,644/1,914 (85.9%)	1,666/1,914 (87.0%)	1,676/1,914 (87.6%)
体言 → 体言	1,012/1,112 (91.0%)	1,012/1,112 (91.0%)	1,012/1,112 (91.0%)
用言 → 用言	658/806 (81.6%)	655/806 (81.3%)	656/806 (81.4%)
用言 → 体言	509/554 (91.9%)	509/554 (91.9%)	512/554 (92.4%)

表 4: 解析が正しくなった例

(a) ぐねぐね道を 立ったまま$_{\times}$ バスに 揺られる$_{\bigcirc}$ ことを覚悟しました。
(b) たいてい、人は強欲だから 何かを 得る$_{\bigcirc}$ ために 捨てた$_{\times}$ ものにも未練を残している。
(c) せんべいの箱はデパートみたいな山積みではなく、間隔を あけて 陳列されているのが$_{\bigcirc}$ また 良い。$_{\times}$

なった。これは、「バスに揺られる」の格フレームからはヲ格をとらない「立つ」の格フレームが生成されやすいためと考えられる。このように、格フレーム間関係知識を考慮することにより格フレームの選択が正確になり、そのため述語項構造内の係り受け (体言→用言) 精度がよくなったことがわかる。

しかし、格フレーム間関係知識による連用節間の係り受け精度の改善は見られなかった。連用節の係り先はそもそも曖昧な場合が多いことが大きな原因と考えられる。

(2) 結論から 言ってしまえば、 書いてあることはそれほど 間違っていないし、$_{\times}$ しっかりと読み込めば彼の言っていることが重要なことであることは 分かる。$_{\bigcirc}$

たとえば、この文においては「言ってしまえば、」の係り先の正解は「分かる。」であるが、「間違っていないし、」に誤って解析している。このような例では意味的にどちらにも関係しているので、係り受け解析の評価方法を再考する必要がある。

5 おわりに

本稿では、自動獲得した格フレーム関係知識を用いて確率的構文・格解析を行う手法を提案した。提案手法により構文解析の精度が向上することを確認した。本手法のポイントは、節内と節間の整合性がもっともよい構文・格構造を選択する点である。自動獲得した格フレーム間関係知識は、今後、含意関係認識などの言語処理システムに適用していきたいと考えている。

参考文献

[1] Shuya Abe, Kentaro Inui, and Yuji Matsumoto. Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples with verbal nouns. In *Proceedings of IJCNLP2008*, pp. 497–504, 2008.

[2] Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP2004*, pp. 33–40, 2004.

[3] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[4] Dekang Lin and Patrick Pantel. DIRT - discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 323–328, 2001.

[5] Viktor Pekar. Acquisition of verb entailment from text. In *Proceedings of HLT-NAACL2006*, pp. 49–56, 2006.

[6] Kentaro Torisawa. Acquiring inference rules with temporal constraints by using Japanese coordinated sentences and noun-verb co-occurrences. In *Proceedings of HLT-NAACL2006*, pp. 57–64, 2006.

[7] Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of COLING-ACL2006*, pp. 849–856, 2006.

[8] 河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol. 12, No. 2, pp. 109–131, 2005.

[9] 河原大輔, 黒橋禎夫. 自動構築した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. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[10] 乾孝司, 乾健太郎, 松本裕治. 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933, 2004.