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

阿辺川 武
東京工業大学大学院 総合理工学研究科
abekawa@lr.pi.titech.ac.jp

奥村 学
東京工業大学 精密工学研究所
oku@pi.titech.ac.jp

1 はじめに

係り受け解析は日本語処理の基本技術として認識されており，これまでに多くの研究が行われている．初期の研究では，2 文節間の係りやすさを決定する規則を人手で作成していたが，網羅性や一貫性といった点から問題が多い．近年では，係り受け情報が付与された大規模なコーパスが利用可能になったことから，機械学習アルゴリズムを用いた統計的な係り受け解析技術が提案されるようになった [3, 5, 7, 8].

これらの統計的日本語係り受け解析では，文に含まれる全ての係り関係は互いに独立であるという仮定を置いており，2 文節間の係り関係の有無を個々に判定している．また係り関係の判定には，2 文節の持つ情報を素性という形で表現し基本的なものとして 2 文節の主辞や語形の形態素の情報を使用している [7]．しかし基本素性だけでは 2 文節以外の文節の情報が反映されないため，2 文節間に存在する文節など周辺文脈の情報を素性として追加しなければならない．しかしこれらの素性を利用しても，1 つの用言に同じ格が係らないといった日本語の特徴までは表現できないのが現状である．

一方，ある名詞がある動詞と共起しやすいといった情報は係り受け解析に有用であると考えられる．機械学習を用いた統計的係り受け解析では，構文構造付きコーパスを訓練データとテストデータに分け，訓練データのみから係り関係を特徴づける素性を抽出する．しかし係り関係タグが付与された訓練データを大量に用意することはコストの面から難しく，少量のデータからでは共起情報を直接学習することは難しい．

近年，新聞コーパスといった大規模な電子データが利用可能になってきており，既存の係り受け解析器を用いることで，人手によりタグ情報が付与されたコーパスほど正確ではないが，ある程度精度の高い共起情報が大量に入手できるようになった．本稿では以上のことをふまえ，従来の機械学習による統計的係り受け解析で有効に活用されていない一文一格の制約や共起情報を用いて，係り受け解析の精度向上を実現する手法を提案する．

2 共起情報を利用した係り受け解析

日本語では，同じ深層格を持つ格要素が 1 つの用言に係ることはないという一文一格の制約がある．従来の係り受け解析では係り関係同士が独立であるという仮定を置いているため，周辺文脈を表現する素性を追加してこの制約を組み入れる必要があった．例えば，2 文節間に特定の形を持つ文節 [7] や，既に係り先に係っていると判断した文節 [3]，係り元や係り先の前後の文節 [5] といった情報を素性として追加している．しかし，このような素性だけでは厳しい制約を課すことができない場合がある．

a) 会見で閣議で決定した政令を発表した．

上記の例文では,「会見で」は「発表する」に係ると考えるのが妥当である．しかし従来の統計的係り受け解析アルゴリズム [3, 8] では,「会見で」は「決定する」に係るような解析結果を出力する．一文一格は本来では深層格のレベルで考えるべきだが，例文においては簡単に表層格による制約が考慮できれば正しい係り受け解析結果が得られると考えられる．

また，もう 1 つ解析に有効な情報として，格要素が用言に係りやすいことを表す共起情報がある．例文では「会見で」の係り先候補は「決定する」もしくは「発表する」であるが，共起関係を考えれば「発表する」に係りやすい．機械学習アルゴリズムを利用する従来の係り受け解析では，訓練データに共起対が頻出すれば素性として考慮されることもあるが，低頻度であったり，そもそも訓練データ中に存在しなければ，いくら一般に共起関係があると思われる対であっても素性に反映されることはない．

2.1 提案モデル

前節をふまえ，本節では，一文一格の制約および格要素と用言の共起関係を考慮した統計モデルを提案する．最初に提案モデルでは格要素が用言に係る関係に限定し，それ以外の関係である用言が用言に係るといった関係は考慮しないことを述べておく．用言 v に $c(v)$ 個からなる格要素集合 $\{e_{1..c(v)}\}$ が係っているとき，この係り関係の確率 $P(\{e_{1..c(v)}\} \rightarrow v)$ を v に対

する $\{e_{1..c(v)}\}$ の条件付確率として定義する．

$$P(\{e_{1..c(v)}\} \rightarrow v) = P(\{e_{1..c(v)}\}|v)$$

格要素は名詞と助詞から構成されていると考えると，$\{e_{1..c(v)}\}$ は名詞集合 $\{n_{1..c(v)}\}$ と助詞集合 $\{r_{1..c(v)}\}$ で表わせる．

$$\begin{aligned} P(\{e_{1..c(v)}\}|v) &= P(\{n_{1..c(v)}\},\{r_{1..c(v)}\}|v) \\ &= P(\{r_{1..c(v)}\}|v) \\ &\quad \times P(\{n_{1..c(v)}\}|\{r_{1..c(v)}\},v) \end{aligned}$$

ここで，名詞集合 $\{n_{1..c(v)}\}$ は用言 v に対して独立であると仮定すると，

$$P(\{e_{1..c(v)}\}|v) = P(\{r_{1..c(v)}\}|v)\prod_{j=1}^{c(v)} P(n_j|r_j,v)$$

本稿では $P(\{r_{1..c(v)}\}|v)$ を用言に対する助詞集合共起確率，$P(n_j|r_j,v)$ を格要素共起確率と呼ぶ．

最後に文中に含まれるすべての用言 $v_{1..m}$ と，それらに係る格要素を考える．文全体の格要素と用言の係り関係の確率を $P(\hat{T})$ として，用言ごとに係り確率が独立であると仮定すると次のようになる．

$$P(\hat{T}) = \prod_{i=1}^{m} P(\{r_{1..c(v_i)}\}|v_i)\prod_{j=1}^{c(v_i)} P(n_j|r_j,v_i) \quad (1)$$

本提案モデルが従来モデルと異なるところは，用言に係る格要素集合のうち，助詞間の非独立性を考慮に入れていることである．これにより同じ用言に係る格要素間に助詞の組合せの観点から制約を設けることができる．助詞間の独立を仮定すると式 (1) は次のように表される．

$$P(\hat{T}) = \prod_{i=1}^{m}\prod_{j=1}^{c(v_i)} P(n_j,r_j|v_i) \quad (2)$$

この独立性を仮定したモデルは，確率値の計算法や全体を用言数で割った平均値をスコアとして用いている点などが異なるが八木ら [9] のモデルに近い．

2.2 モデルの適用例

ここでは先程の例文 a) に対して本提案モデルの適用を試みる．例文 a) において「会見で」と「閣議で」の係り先の曖昧性は 3 通りある．すべての組合せにおいて助詞集合共起確率と格要素共起確率を計算し，文の確率 $P(\hat{T})$ を計算すると表 1 のようになる．その結果，すべての共起確率の積が一番大きい候補 2 の係り受けが出力となる．

2.3 被修飾名詞の考慮

日本語係り受け解析では係り先方向は後方のみであることを仮定している．2.1 節のモデルでもこの仮定を置いているので，格要素集合は用言の前方から構成されている．しかし，用言が連体修飾をしているときは，用言の係り先である被修飾名詞を考慮に入れる必要が生じる．

b) 太郎が泳いでいる姿を見た．
c) 太郎が泳いでいる少年を見た．

例文 b) では「太郎が」は「泳ぐ」「見る」のどちらの用言に係るかは文脈依存であり，この 1 文からは決定できない．一方，例文 c) の場合は「泳ぐ」は「少年」を連体修飾しており，意味的には「少年が泳ぐ」と被修飾名詞は用言に対しガ格の格要素となっている．そのため同様にガ格をとる「太郎が」は「泳ぐ」ではなく「見る」に係ることが一意に決定できる．

このように，用言 v が連体修飾しているときには，被修飾名詞 $\tilde{n}$ が格要素となり得るか，すなわち内の関係であるか外の関係であるかを決定し，さらに内の関係である場合はどの格の格要素になるかを決定する必要がある．本提案モデルでは，内／外の関係の判別には文献 [1] の手法を用いることにする．内の関係だと判定された場合，格助詞集合 R から式 (3) を最大にする格助詞 $\tilde{r}^*$ を求め，用言 v に対する格とする．

$$\tilde{r}^* = \underset{\tilde{r}\in R}{\operatorname{argmax}}\ P(r_{\{1..c(v),\tilde{r}\}}|v)P(\tilde{n}|\tilde{r},v) \quad (3)$$

3 評価実験

提案モデルの有効性を確認するために実験を行なう．しかし提案モデルは格要素と用言の係り受け関係のみを考慮しているため，すべての文節の係り関係を計算することはできない．そこで既存の係り受け解析モデルから複数の係り受け候補を出力し，その結果に対して提案モデルを用いて re-ranking を適用する手法をとることにする．

3.1 係り受け解析モデル

re-ranking の入力となる係り受け解析モデルには複数の候補を出力できるモデルとして，内元らにより提案された最大エントロピー法に基づく後方文脈モデル [8] を使用した．使用した素性は，文献 [7] で述べられている素性とその組合せを基本素性として，文献 [5] などを参考に独自の素性と組合せを追加した．なお訓練データ中で頻度 7 以下の素性を削除している．

3.2 re-ranking 方法

提案モデルでは格要素が用言に係る関係のみに着目している．したがって格要素の係り先が用言の候補と

表 1: 「会見で閣議で決定した政令を発表した.」の計算例

	$\log(P_1)$	$\log(P_1)$	$\log(P_2)$		$\log(P_1)$	$\log(P_1)$	$\log(P_1)$	$\log(P_2)$		$\log(P(\hat{T}))$
候補 1	会見で -4.61	閣議で -4.75	{で, で} -8.18	決定する	–	–	政令を -6.98	{を} -2.10	発表する	-26.6
候補 2	–	閣議で -4.75	{で} -4.01	決定する	会見で -3.74	–	政令を -6.98	{で, を} -5.09	発表する	**-24.6**
候補 3	–	–	{無} -2.73	決定する	会見で -3.74	閣議で -6.64	政令を -6.98	{で, で, を} -9.45	発表する	-29.5

$$P_1 = P(n|r, v), \quad P_2 = P(\{r_{1..c(v)}\}|v)$$

名詞の候補では確率を比較することができない．そこで re-ranking を行なう候補を図 1 のように制限する．係り受け解析モデルで 1 位の候補に対し，格要素の係り先が複数の用言間で異なる候補のみを re-ranking する対象文とし，それ以外の文は対象外とする．1 位候補を含めた全ての対象文に対し，提案モデルで確率を計算し，最も高い確率を持つ候補を最終的な出力とする．今回の実験では係り受け解析の出力候補数は 10 とし，候補内に 1 位以外の対象文が 1 つもないときは re-ranking は適用されない．

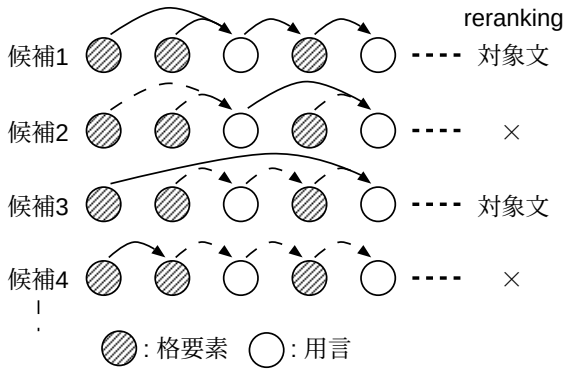

図 1: re-ranking 対象文の選択

3.3 共起確率の計算

提案モデルで使用する助詞集合共起確率と格要素共起確率は，大規模コーパスを解析して得られた共起対から算出した．まず新聞 26 年分の文章を JUMAN[1] を用いて形態素解析し，独自に実装した後方文脈モデル[2] で係り受け解析を行ない，その結果から共起対を収集した．その際，信頼性の高い共起対のみを収集するため，読点が付いている格要素は除外した[3]．

続いて，名詞-助詞-用言の共起確率の計算には，鳥澤ら [6] の手法と同様に PLSI を使用した[4]．名詞と助詞-用言からなる行列を構成し，PLSI を用いて行列を圧縮した．隠れ変数 z の次元は 500 としている．助詞は格助詞の他に係・副助詞も対象としており，ハ格だけは単独で扱い，それ以外の係・副助詞はまとめて 1 つの助詞として扱っている．低頻度語への対応として頻度 5 未満は 1 つの未知語クラスとした．

また，助詞集合共起確率も共起情報の収集に用いた解析結果から収集した．ハ格の格要素は文末の用言に係りやすいなど用言の位置により共起する助詞集合が異なると考えられるため，用言が文中にあるか文末にあるかで助詞集合を区別して収集している．確率の計算は，用言と共起した助詞集合の 2 次行列に PLSI を適用して求めている．

なお今回の実験では，re-ranking に用いる値は，後方文脈モデルにより算出された確率と提案モデルにより算出された確率の積を利用している．

3.4 実験設定，結果

実験には京大コーパスを用いた．訓練データの量による比較のために，大小 2 つのデータセットを用意した．

- 京大コーパス 2.0 small (訓練 7,615 文 77,775 文節, テスト 1,246 文 12,509 文節) [5]
- 京大コーパス 3.0 large (訓練 24,263 文 234,474 文節, テスト 9,287 文 89,982 文節) [6]

文節の係り受けの正解率は文末の 1 文節を除いた文節ごとに正しく係り先が決定できた割合，文正解率は文全体の解析が全て正しいものの割合を示す．

最初に re-ranking が有効であるかを実証する実験を行なった．re-ranking の入力となる後方文脈モデルによる正解率と提案モデルによる re-ranking 後の正解率を表 2 に示す．re-ranking 対象文が複数存在し，実際に re-ranking が適用された文のみの正解率も示す．

次に提案モデルに関連する種々のモデルに対して同様な実験を行った．比較するモデルは，a) 格要素共起確率を後方文脈モデルの素性として使用したモデル，b) 格要素集合が独立であるモデル [式 (2)]，c) 連体修飾する用言において被修飾名詞を考慮しないモデル，

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
[2] Maximum Entropy の計算には次のパッケージを利用した http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html
[3] 後方文脈モデルで京大コーパスを解析したところ，読点無の格要素の正解率 90.6%に対して，読点付きの格要素は 76.4% であった
[4] PLSI の計算には次のパッケージを利用した http://chasen.org/~taku/software/plsi/
[5] 訓練:一般記事 95/1/1〜8，テスト:一般記事 95/1/9
[6] 訓練:一般記事 95/1/1〜11 社説 95/1〜8，テスト:一般記事 95/1/14〜17，社説 95/10〜12

表 2: 実験結果

		京大コーパス 2.0 small		京大コーパス 3.0 large	
		文節正解率	文正解率	文節正解率	文正解率
すべての文	後方文脈モデル	88.86%	46.63% (581/1,246)	90.75%	54.04% (5,019/9,287)
	re-ranking 後	88.99%	46.71% (582/1,246)	90.87%	54.17% (5,031/9,287)
re-ranking が適用	後方文脈モデル	89.68%	41.35% (232/561)	91.10%	47.73% (2,125/4,452)
された文のみ	re-ranking 後	89.94%	41.53% (233/561)	91.32%	48.00% (2,137/4,452)

表 3: 各モデルの比較 (京大コーパス 3.0 large)

後方文脈モデル	90.75%
提案モデル	90.87%
a) 格要素共起確率のみを素性に追加	90.54% (-0.33%)
b) 格要素独立 [式 (2)]	90.60% (-0.27%)
c) 連体修飾の解析なし	90.82% (-0.05%)
d) 提案モデルの確率のみ	90.10% (-0.77%)

d)re-ranking に使用する確率に提案モデルの確率だけを用いたモデルである．提案モデルの正解率との差を表 3 に示す．

3.5 考察

実験結果より，訓練データの量によらず後方文脈モデルよりも正解率が高いことから提案モデルの有効性が確認された．また b) のモデルが提案モデルよりも正解率が低いことから，格要素間の関係を考慮することの有用性も確認された．

a) の格要素共起確率をそのまま素性として使用したモデル，および b) の格要素を独立として扱ったモデルでは，元々の後方文脈モデルの正解率より低いものとなった．これは照応解析というタスクこそ異なるが文献 [2] と同様の結果となった．解析で有効であると思われる共起情報であるが，単純な利用方法では，精度は向上しないものと思われる．

d) 提案モデルの確率値だけを使用したモデルは，後方文脈モデルの確率値との積を使用した場合よりも精度が悪かった．提案モデルは格要素に関して名詞と助詞の情報だけを用いているのに対して，後方文脈モデルでは係り受け関係に大きく影響を与える読点や距離の情報などを用いている．そのため双方のモデルの確率値の積を利用することにより，お互いの利点を有効に統合できたからと考えられる．

今回の実験では，文献 [4, 5] などの最新の係り受け解析アルゴリズムの正解率に比べて低いものになっている．原因はアルゴリズムの違いにあるとともに，用いた素性の差も大きい．有効な素性を用いて元となる係り受け解析の結果が向上できれば，re-ranking 後の正解率も向上する見込みがある．

4 おわりに

本稿では，大規模コーパスから収集した統計情報を用いて，日本語係り受け解析の精度を向上させる手法を提案した．従来手法の全ての係り受け関係は独立であるという仮定を利用せず，同じ用言を係り先とする格要素間は非独立であるという性質を使用した．その結果，既存の機械学習アルゴリズムによる係り受け解析の正解率をさらに向上させることができた．

今回の実験では，表層格として格助詞の他にも係・副助詞をそのまま使用している．本稿で被修飾名詞が用言のどの格になるかを 2 つの共起確率を用いて推定しているが，係・副助詞に対しても同じ手法で格を推定できるのではないかと考えているので，今後の課題としたい．

参考文献

[1] 阿辺川武, 奥村学. 日本語連体修飾節と被修飾名詞間の関係の解析. 自然言語処理, Vol. 12, No. 1, pp. 107–123, 2005.

[2] Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of HLT/NAACL-04*, pp. 289–296, 2004.

[3] 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.

[4] 工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析モデル. 情報処理学会研究報告 162-NL-29, 2004.

[5] Manabu Sassano. Linear-time dependency analysis for japanese. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pp. 8–14, 2004.

[6] Kentaro Torisawa. An unsupervised method for canonicalization of japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 211–218, 2001.

[7] 内元清貴, 関根聡, 井佐原均. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3397–3407, 1999.

[8] 内元清貴, 村田真樹, 関根聡, 井佐原均. 後方文脈を考慮した係り受けモデル. 自然言語処理, Vol. 7, No. 5, pp. 3–17, 2000.

[9] 八木豊, 野呂智哉, 橋本泰一, 徳永健伸, 田中穂積. 単語の共起情報を利用した文法主導の係り受け解析. 情報処理学会研究報告 157-NL-3, 2003.