3つ組/4つ組モデルによる韓国語構文解析

金山 博
日本アイ・ビー・エム株式会社
東京基礎研究所
hkana@jp.ibm.com

李 東黙
韓国アイ・ビー・エム株式会社
ソフトウェアソリューション研究所
dmyi@kr.ibm.com

1　はじめに

日本語と韓国語の間には、主辞後置性、助詞を介した用言の修飾、格の省略、自由な語順など、構文構造の共通点が非常に多い。そこで本稿では、日本語の係り受け解析の考え方を韓国語に適用する試みについて述べる。用いる手法は、文法知識をもとに係り先の探索空間を絞りこむことによって文脈情報を捉えながら係り受けの傾向を学習する「3つ組/4つ組モデル」[6]である。日本語の文法知識を参考にして、3つ組/4つ組モデルと親和性の高い韓国語の文法を設計し、韓国語の構造に即したヒューリスティクスを加えることにより、二文節間の係りやすさを計算する統計モデルに比べて高い正解率を得た。また、文法の被覆率や係り受けの正解率を向上させる際の手順の中で見出された両言語の文法構造の類似点と相違点について論じる。

2　関連研究

韓国語の構文解析は、日本語に比べると報告が少ないものの、形態素解析に続く処理として様々な試みがある。用言の格フレームに相当する文法パターンを定義して正解率を上げる方法 [7] や、生のコーパスから獲得した格フレームや「N_1 의 N_2 (‘N_1 の N_2’)」の情報を用いる方法 [8] は、語彙情報の活用という点で共通している。また、語節内の語やその前後の形態素の情報から係り受けに影響する特徴を適切に捉える手法 [1] や、統計的機械翻訳のモデル上で英語・韓国語の構文解析と語の対応付けを同時に行う方法 [9] などがある。

日本語構文解析向けに考案された3つ組/4つ組モデルについては4節で述べる。なお、その手法を韓国語構文解析に適用した報告がある [2]。そこでは、遠くの文節に係る際には3つ組/4つ組モデルが他の手法よりも高い正解率を示した。その際に用いた文法規則は、コーパスから自動生成したもので、3つの係り先に係る割合が91.5%と低いことから、係り先候補の絞り込みが十分にできていない可能性がある。本稿では、正解率を高めるために文法規則とヒューリスティクスを改良していく。

先行研究において係り受け解析の正解率は82〜86%程度の値が報告されているが、コーパスの違いや、対象外とする文の存在など、条件が異なるため、本稿では先行研究との性能の比較は焦点としないこととする。

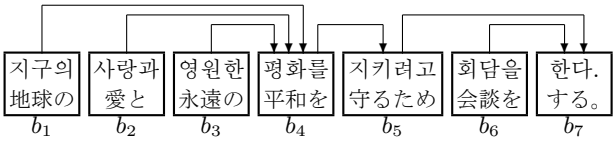

図1　韓国語の語節間の係り受け構造の例。

表1　韓国語・日本語のコーパスの比較。

	KTB (韓国語)	EDR (日本語)
文の平均文字数（空白除く）	73.7	28.0
文の平均語節/文節数	25.5	8.53
語節/文節の平均形態素数	1.83	2.86
右隣の語節/文節に係る割合	70.0%	61.8%

3　韓国語の構文解析

韓国語の依存構造の分析は一般的に、空白で区切られた単位である語節（어절）ごとに行う。語節は、名詞・動詞などの内容語に、助詞・語尾などの機能語が付与されたもので、概ね日本語の文節に相当するが、韓国語では複合名詞や助動詞相当句が複数の語節に分割されるなど、日本語よりも細かい単位になる傾向がある。文中における空白の挿入には一般的な規則があるものの、現実の文ではしばしば揺れが見られるため、語節に含まれる形態素のパターンもまちまちである。

図1に韓国語の係り受け解析の例を示す。係り元語節の係り先を、それより右側の文節から選択するという点において、日本語の解析とほぼ同じように考えることができる。

問題の複雑さを知るために、今回用いる韓国語の構文構造付きコーパス Penn Korean Treebank [5]（以下KTB）と、日本語の EDR コーパス [4] との間で、両言語の特徴を比較してみた。表1に、文あたりの語節/文節数、語節/文節あたりの形態素数、全ての係り先が右隣の語節/文節であった時の係り受けの正解率を示す。両コーパスはそれぞれ独立の文なので単純な比較は難しいが、KTB の方が3倍程度文長・語節数が大きい。一方で語節/文節あたりの形態素数は日本語のほうが多く、韓国語で語節が細分化されやすい傾向、日本語では複数の機能語が使われやすい傾向がわかる。

表 2　簡素化した日本語の文法規則の例。

係り属性の定義
・語形が格助詞「の」: 連用・連体
・語形が格助詞「を」: 連用
・語形が副詞: 連用・副詞
受け属性の定義
・名詞を含む文節: 連体
・用言・判定詞を含む文節: 連用
・副詞を含む文節: 副詞
修飾可能性の追加
・語形が格助詞「と」→ 「一緒に」の文節
・語形が格助詞「を」 →主辞が「条件 | 中心 | きっかけ |… + に」の文節
・文節が「全部で」→ 数詞を含む文節

4　3 つ組/4 つ組モデル

日本語係り受け解析のための 3 つ組/4 つ組モデル [6, 12] とは、文法及びヒューリスティクスを用いて係り先候補を高々 3 つに絞り、係り元文節とすべての係り先候補文節の属性を用いて、それぞれの候補に係る確率を求める手法である。以下にその概要を示す。

4.1　係り先候補の絞り込み

まず、各文節が、同一文内でその文節より右側にあるそれぞれの文節を修飾し得るか否かを、文法を用いて決定する。その際には、表 2 に示すような文法規則を考え、係り元文節の係り属性と重なりを持つような受け属性の文節を、修飾可能であるとして列挙する [14, 10]。なお、語形とは文節内で句読点を除く最右の形態素、主辞とは文節内の最右の自立語である。

修飾可能であるとされた文節集合のうち、係り元文節から最も近い文節、2 番目に近い文節、最も遠い文節に係る場合が 98.6 %を占めるという観測結果 [6] に基づき、係り先の候補が 4 つ以上ある場合、上記の 3 文節のみを考えて、他の文節は無視するというヒューリスティクスを用いる。以降では、このように 3 つ以下に制限された文節集合を、単に係り先候補と呼ぶ。

4.2　係り受けの計算とモデルの特徴

文節 b が、左から n 番目の係り先候補文節 c_{bn} に係る確率 $P(b \to c_{bn})$ を、文節 b の属性 Φ_b 及び c_{bn} の属性 $\Psi_{c_{bn}}$[*1]を用いて、係り先候補が 2 つの場合は 3 つ組の式 (1)、係り先候補が 3 つの場合は 4 つ組の式 (2) で計算する。

$$P(b \to c_{bn}) = P(n \mid \Phi_b, \Psi_{c_{b1}}, \Psi_{c_{b2}}) \quad (1)$$

$$P(b \to c_{bn}) = P(n \mid \Phi_b, \Psi_{c_{b1}}, \Psi_{c_{b2}}, \Psi_{c_{b3}}) \quad (2)$$

文全体の構造の確率 $P(T)$ は、上記の係り確率が独立であるという仮定に基づいて、文中の各係り受け確率の積

$$P(T) \simeq \prod_b P(b \to c_{bn}) \quad (3)$$

として[*2]、$P(T)$ が最大となるような係り受けを後方からのビームサーチにより探索する。

式 (1), (2) の特徴は、「係り元文節とその係り先候補の全ての属性を同時に考慮する」こと、そして「それぞれの係り先候補への係りやすさを求めるのではなく、各候補が選ばれる確率を直接求める」ことである。これにより、係り元文節から見た各候補の相対的な位置、他の候補の属性との相互関係（文脈情報）などを自然に反映させることができる。

5　3 つ組/4 つ組モデルの韓国語への適用

韓国語と日本語の類似性をもとに、3 つ組/4 つ組モデルを用いた韓国語の係り受け解析を試みる。まずは係り先の候補を高々 3 つに絞り込む準備をする。

5.1　語節間の修飾可能性の列挙

各語節が修飾し得る語節を、表 2 に示した日本語の文法に倣って定義した。まず、語節の係り属性と受け属性を定める。下記は KTB の品詞体系に基づく規則の一部である。

■係り属性

- 連用属性を持つ語節：語形が AD*(副詞), PCA(格助詞), PAD(副助詞), PCJ(接続助詞), PAU(補助詞), ENM(名詞形転成語尾), ECS(連結語尾), XSF(接尾辞), N*(名詞類) のいずれか
- 連体属性を持つ語節：語形が N*(名詞類), PCJ(接続副詞), DAN(冠形詞), PAN(冠形格助詞), EAN(冠形転成語尾), XSF(接尾辞), 副詞「특히 (‘特に’)」のいずれか
- 副詞属性を持つ語節：語形が AD*(副詞)

■受け属性

- 連用属性を持つ語節：語節の中に V*(動詞・形容詞・補助用言), 副詞「없이 (‘無く’)」を含む
- 連体属性を持つ語節：語節の中に N*(名詞類) を含む
- 副詞属性を持つ語節：語節の中に AD*(副詞) を含む

語節 e_0 と、それより右側の語節 e_1 について、e_0 の係り属性と e_1 の受け属性に共通するものがある時、e_0 は e_1 を修飾可能であるとする。さらに、これらの規則だけでは記述できない特殊な係り受けについて、次のような規則を追加する。

■修飾可能性の追加

- 語形が副助詞「과 (‘と’)」→副詞「함께 (’一緒に’)」
- 語形が格助詞「을 (‘を’)」　→
名詞「조건 (‘条件’)| 중심 (‘中心’)|
계기 (‘きっかけ’)|...」＋助詞「으로 (‘に’)」

このような特殊な係り受けについても、表 2 に示したように日本語にほぼ対応する現象があるのが興味深い。

[*1] $\Psi_{c_{bn}}$ には係り元文節と係り先候補の間の文節の属性も含まれているため、厳密には c_{bn} だけでなく b にも依存する。

[*2] 係り受けが交差しない制約を加えるため、この式は厳密ではない。

表 3　韓国語において文法規則で修飾可能な語節を絞った後の正しい係り先の分布。「合計」の列は第 1・第 2・最遠の 3 語節に限定した時の被覆率を示す（単位は %）。

候補数	比率	第 1	第 2	最遠	合計
1	10.5	100.0	–	–	100.0
2	11.4	85.9	14.1	–	100.0
3	10.4	76.2	13.4	10.4	100.0
4	9.3	74.7	11.3	8.0	93.9
5	8.6	72.1	11.2	7.7	91.0
6 以上	49.8	76.2	9.8	4.4	90.4
合計	100	–	–	–	93.9

これらの規則を追加することにより、文法の被覆率[*3]は 96.6% となった。対応できていない係り受けは、非常に稀な表現や、コーパスの誤りと思われるものが含まれるため、今回はこれ以上の被覆率を求めない。

5.2　3 つの係り先への絞り込み

5.1 節の文法規則によると、日本語に比べて係り先候補の数が非常に多くなった。その原因として、韓国語ではそもそも文中の語節の数が多いこと、名詞が語形の場合に体言・用言の両方に係れる[*4]ことなどがある。3 節で見たように、韓国語のほうが隣の文節に係る割合が多いとはいえ、当然ながら係り先候補が多いほど問題が難しくなる。特に、日本語では複合名詞が一つの文節とみなされるため、その内部構造の分析は係り受け解析の中では問われない。一方で、韓国語では語節単位の解析の中に複合名詞の分析が含まれることになり、多くの名詞から係り先を選択する必要が生じる。

そこで、複合名詞の末尾以外の名詞の語節に対して隣接以外の修飾ができないようにするなどのヒューリスティクスを加え、修飾可能な文節の数を減らすようにした。その結果、候補中の正しい係り先の分布は表 3 のようになった。正解の係り受けが文法で被覆されている場合、係り元語節から最も近い語節、2 番目に近い語節、最も遠い語節の 3 つの中に正しい係り先が含まれる割合は 93.9% である。依然として日本語での同様の調査 [12] の 98.5% とは差があるが、3 つ組/4 つ組モデルを適用するための準備は整ったといえる。

5.3　係り受け事象の抽出

前項までの処理によって語節 b の係り先候補を 3 つ以内に絞り込んだので、式 (1) と式 (2) によって、それぞれの候補を修飾する確率を推定する。そのために、係り元と係り先候補の語節を表現する Φ_b と $\Psi_{c_{bn}}$ を設計する。日本語の構文解析器に倣って導入した素性を表 4 に示す。例えば図 1 の b_1「지구-의」の係り先を求める時には、係り元語節と係り先候補 b_2, b_4, b_6 との関係を図 2 のように素性で表現する。これらの素性と、表 5 に示す組み合わせ素性を用いて、係り元に対して係り先候補がある状況を 1 つの事象として捉える。

[*3] コーパス中にある正解の係り先語節が文法規則によって修飾可能であるとされた割合。

[*4] 用言の格要素において助詞が省かれることが多いため。

表 4　係り元と係り先候補の関係を表現する素性。

素性番号	素性の記述
(1)	係り元主辞品詞
(2)	係り元語形品詞
(3)	係り元助詞・語尾
(4)	係り元副詞語彙
(5)	係り先主辞品詞
(6)	係り先主辞語彙
(7)	係り先語形品詞
(8)	係り先助詞・語尾
(9)	二文節間の은 (‘は’) の数
(10)	二文節間の読点の数

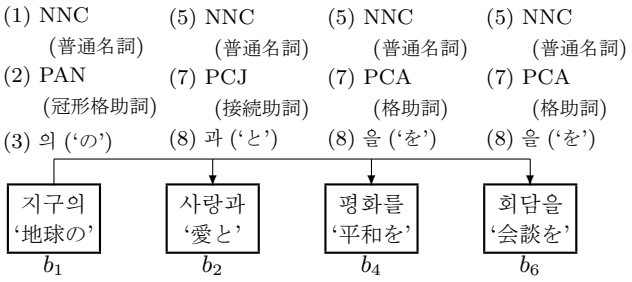

図 2　図 1 の語節 b_1 の係り先を、係り先候補 b_2, b_4, b_6 から選択するモデルの素性値。括弧内の数字は素性番号に対応する。

表 5　韓国語構文解析で用いる組み合わせ素性。

素性番号	組み合わせ素性
(1) × (5)	係り元主辞品詞・係り先候補主辞品詞
(2) × (5)	係り元語形品詞・係り先候補主辞品詞
(2) × (7)	係り元語形品詞・係り先候補語形品詞
(3) × (5)	係り元助詞/語尾・係り先候補主辞品詞
(3) × (8)	係り元助詞/語尾・係り先候補助詞/語尾

6　評価実験

5 節で設計した韓国語向けの 3 つ組/4 つ組モデルを用いた係り受け解析の実験を行う。Penn Korean Treebank [5] の 5,000 文のうち 3,006 文を学習に、1,043 文をテストに用いた。ここではコーパスに人手で付与された正しい形態素解析の結果（語節・形態素・品詞）を入力として[*5]、語節単位の係り受けの正解率を測定する。対照実験として以下の 2 つの方法と比較した。

■ベースライン　係り先を常に右隣の語節とした場合。

■距離モデル　統計的構文解析の先駆である Collins の手法 [3] に倣い、2 文節間の係りやすさを個別に計算する。係り元・係り先の文節を表すために表 4 と同じ素性セットに加えて、係り元と係り先の距離を考慮するため、1 (隣接する語節)、2（距離が 2〜5)、6（距離が 6 以上）の 3 値の素性を導入する。

距離モデル・3 つ組/4 つ組モデルの学習には最大エントロピー法を用い、各語節へ係る確率を推定して、文末からのビームサーチで最適な構文木を決定する。

実験の結果を表 6 に示す。全体の正解率を比較する

[*5] 引用符、括弧や標準的な空白が欠けている場合など、語節よりも細かい単位がコーパスに付与されている場合もある。

表 6　韓国語係り受け解析の正解率 (単位は %)。「名詞」「格助詞」「副詞」は、係り元語節の語形を限定した時の結果。

	ベースライン (隣を修飾)	距離モデル	3 つ組/ 4 つ組モデル
全体	**70.0** (18226/26035)	**81.2** (21131/26035)	**84.0** (21859/26035)
名詞	86.5	89.3	89.3
格助詞	57.3	75.2	79.4
副詞	49.8	65.7	71.3

と、それぞれの間で統計的に有意な差が出ており、距離モデルに比した 3 つ組/4 つ組モデルの有効性が確認された。今回は学習コーパスの量が少ないこともあり、係り元と係り先の語彙の組み合わせ素性、すなわち特定の語彙相互の係りやすさを示す値は学習に用いていないが、それでも係り受けの大局的な情報をもとにして先行研究と同等の正解率が得られた。係り元の語形が名詞である語節に注目した場合、係り先候補の数が多くなる傾向が特に強いせいもあって、距離モデルと同等の正解率しか得られなかった。一方、語形が格助詞や副詞の語節は、複数の候補（主に用言）の位置や文法的特徴を捉えて、距離モデルに比して優れた解析結果を出すことができた。

図 1 の最初の語節 b_1「지구-의」の係り先の計算を見ると、手法の差がわかりやすい。距離モデルの場合は b_1「지구-의」が b_2「사랑-과」に係る確率が、b_4「평화-를」に係る確率を上回ってしまう。これは、b_1 と b_2 の距離が 1 であるため $b_1 \rightarrow b_2$ の係り受けが優先されるからである。一方、3 つ組/4 つ組モデルでは、「N-의」に 3 つの係り先候補があり、一番目が「N-과」で二番目が「N-을」の状況で、二番目の側に係りやすいという傾向が学習されるため、正しい係り受けの $b_1 \rightarrow b_4$（「평화-를」）の確率が高くなる。また、候補間の相対的な係りやすさを、b_1 が係り得ない形容詞句 b_3「영원-한」に影響されることなく計算できる点も有効に働いている。

7　日本語 ・ 韓国語の比較

日本語の構文解析の際に、助詞「は」が遠くに係る傾向が強いことはよく知られている。それに対応する韓国語の助詞「은」に対しても、素性 (5) のパラメータの中に 2 番目・3 番目の候補への係りやすさが反映されており、日本語との類似性が表れている。

日本語の構文解析の際に効果があった素性のうち、韓国語において利用できなかったものとして、用言活用形と読点がある。日本語でいう「連用形」「テ形」等は、韓国語においては連結語尾や冠形転成語尾で表現されるため、助詞に加えて語尾の語彙を素性化すれば日本語と同等の現象を捉えられる。

日本語の読点は、遠くへの修飾を示す時に用いられる。日本語の 3 つ組/4 つ組モデルで学習されたパラメータを見ると、係り元文節に読点がある場合に第一候補の優先度が極端に下がる。一方で、韓国語の読点は、基本的に名詞の並列構造を示す時にのみ使われる。従って連用修飾などの長距離の修飾関係は、係り元助詞と係り先動詞など、語彙の組み合わせの影響が高まることが予測される。すなわち、読点の打ち方の揺れに対して頑健な日本語構文解析 [11] の考え方が、韓国語の構文解析の改良に繋がると思われる。

8　おわりに

本稿では、日本語向けに設計された係り受け解析の手法を韓国語に適用した。係り先候補を限定することによって事象を精細に捉える手法が、両言語で同じように効果があることを確認した。今回は韓国語の形態素解析の正解を用いているが、誤りを含む自動処理の結果を入力とするならば、それを前提として柔軟な規則に書き換えるなどの改良が必要になる。また、構文解析やその後の処理にとって最適になるように、形態素・語節の単位や品詞の定義を再設計すること [13] も重要になってくるであろう。

参考文献

[1] Jinho D. Choi and Martha Palmer. Statistical dependency parsing in Korean: From corpus generation to automatic parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pp. 1–11, 2011.

[2] Hoojung Chung and Heechang Rim. A new probabilistic dependency parsing model for head-final, free word order languages. *IEICE TRANSACTIONS on Information and Systems*, Vol. E86-1, No. 11, pp. 2490–2493, 2003.

[3] Michael Collins. Three generative, lexicalised models for statistical parsing. In *Proc. of the 35th ACL*, 1997.

[4] EDR. EDR (Japan Electronic Dictionary Research Institute, Ltd.) electronic dictionary version 1.5 technical guide, 1996.

[5] Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. Penn Korean treebank: Development and evaluation. In *Proc. Pacific Asian Conf. Language and Comp*, 2002.

[6] Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 411–417, 2000.

[7] Hyeon-Yeong Lee, Yi-Gyu Hwang, and Yong-Seok Lee. Parsing of Korean based on CFG using sentence pattern information. *International Journal of Computer Science and Network Security*, Vol. 7, No. 7, 2007.

[8] Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, and Key-Sun Choi. Towards fully lexicalized dependency parsing for Korean. In *Proceedings of The 13th International Conference on Parsing Technologies*, 2013.

[9] David A. Smith and Noah A. Smith. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 49–56, 2004.

[10] 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.

[11] 金山博. 読点に頼らない統計的構文解析. 情報処理学会第 170 回自然言語処理研究会, 2005.

[12] 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.

[13] 山本和英. 計算機処理のための韓国語言語体系と形態素処理. 自然言語処理, Vol. 7, No. 4, pp. 25–62, 2000.

[14] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 統計的構文解析における構文的統計情報と語彙的統計情報の統合について. 自然言語処理, Vol. 5, No. 3, 1998.