類似性を用いない並列構造解析

河原 大輔　　黒橋 禎夫
情報通信研究機構　京都大学大学院情報学研究科
dk@nict.go.jp　　kuro@i.kyoto-u.ac.jp

1　はじめに

並列構造の曖昧性は、文章の意味理解に大きく影響を与えるため、並列構造の解析は非常に重要である。並列構造を構成する句や節の間には、構文、意味的な類似性があることが多いと考えられるため、これらの類似性を用いた並列構造解析手法が提案されてきた。類似性としては、文字列の (部分) 一致、品詞および単語の一致や意味的な類似性などが用いられる。意味的な類似性は、シソーラスに基づく手法 [5, 4][1]や、コーパスに基づく分布類似度を用いる手法 [1] などによって計算される。

これに対して、本研究では、並列構造は、その周りの係り受け関係によってサポートされるため、 その解析に類似性は必要ないという仮説を立てる。この仮説に基づき、大規模な選択選好知識を用いた確率的構文・格・並列解析システムを提案する。ウェブテキストを用いて実験を行い、構文・並列構造の解析に本モデルが有効であることを示す。

2　並列構造解析の考え方

本研究では、並列構造の存在を示す表現を並列キーと呼ぶ。並列キーは、読点や「と」「や」「および」などの付属語をもつ文節である。

例えば、次の文においては「健康と」が並列キーである。

(1)　法王の健康とチベットの平和を祈った

この文は、以下に示すように、4 つの並列構造の曖昧性をもつ[2]。

a. (法王の 健康と) (チベットの 平和)
b. 法王の (健康と (チベットの 平和))
c. ((法王の 健康と) チベットの) 平和
d. 法王の (健康と チベットの) 平和

以下では、類似性に基づく並列構造解析の手法と、類似性を用いない手法により、この例を解析する場合について説明する。

2.1　類似性に基づく並列構造解析

類似性に基づく並列構造解析の手法は、シソーラスや分布類似度により「健康」と「平和」が類似していることがわかるため、「健康」と「平和」が並列構造を形成していることを認識する。その結果、「健康」と「平和」が並列構造になっている (a) と (b) に絞り込むことができる。さらに、並列構造の前部と後部のバランスを考慮することにより、正解の並列構造 (a) を選択することができる。

2.2　類似性を用いない並列構造解析

例文 (1) に対して格解析を行うことを考える。この例文には、すでに示したように、4 つの並列構造の曖昧性があるが、これを構文構造とともに示すと、図 1 のようになる。図において、矩形は並列構造の範囲を示しており、曲線は、我々の確率的生成モデルの生成過程を表している。そのうち、点線は、並列構造を構成している複数の語から生成されることを示しており、その場合にはそれらの生成確率の平均をとることになる。

図 1 のそれぞれの構造の下には、その構造の生成確率を示している。ただし、文末の「祈る」の生成確率は省略している。各確率の後の + は、その表現がコーパスに存在し、その生成確率がある程度高いことを示

[1] Resnik は、シソーラスとコーパスにおける頻度を組み合わせる手法を提案している。

[2] 「〜と」が並列構造を導くかどうかという曖昧性もあるが、ここでは考慮しない。実際には、この曖昧性は、並列構造の範囲の曖昧性とともに、確率モデルの中で同時に解消される。

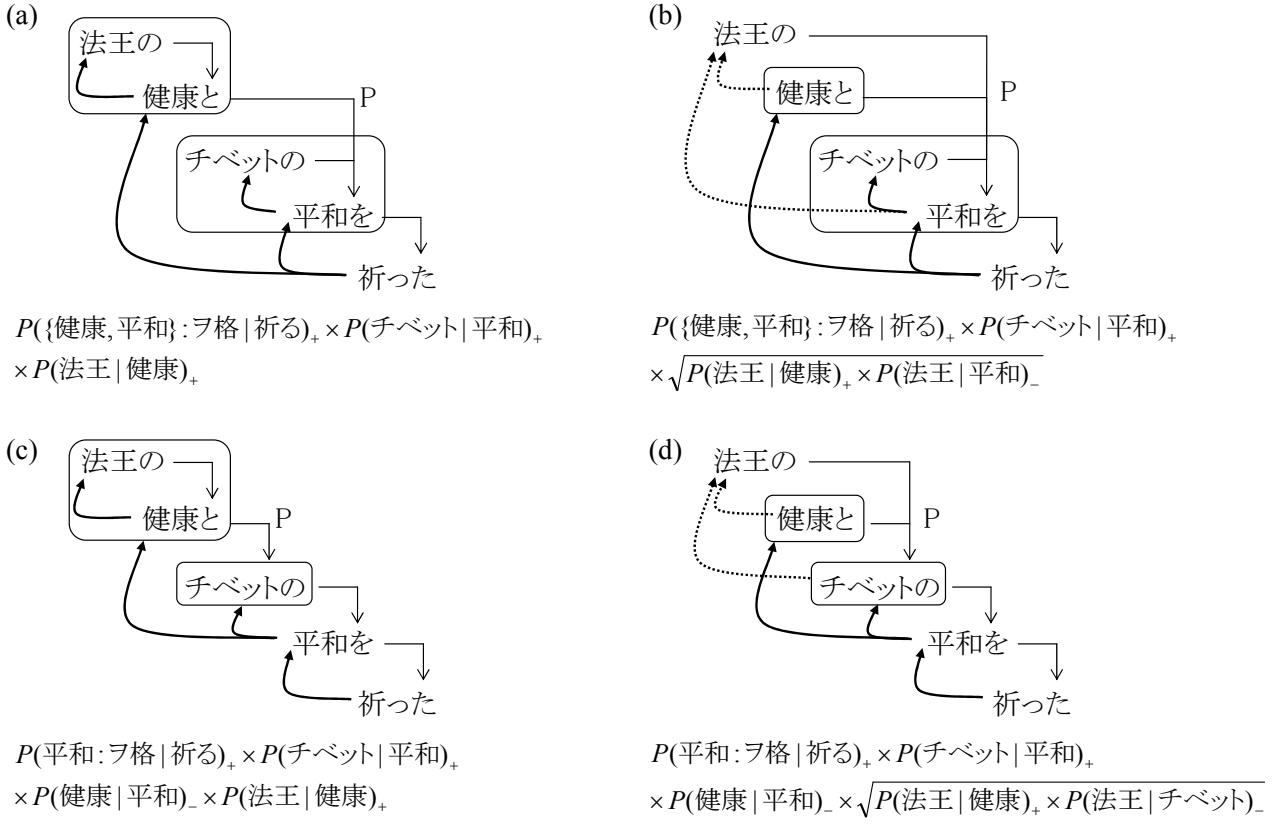

図 1: 並列構造の曖昧性と選択選好に基づく解釈の例

している。– は、逆に、その表現が意味をなさないものであり、生成確率が低いことを示している。例えば、「法王の平和」は意味をなさないので、P(法王 | 平和) は低くなり、構造 (b) の生成確率も低くなる。また、(c) と (d) では、P(平和:ヲ格 | 祈る) に関しては問題ないが、「健康」を生成する P(健康 | 平和) が「健康の平和」とは言えないために低くなる。この結果、構成するすべての生成確率が + となっている構造 (a) がもっとも高い確率値をもち、解として選択される。これらの生成確率は、用言と名詞の選択選好に基づいて推定されるものであり、選択選好が並列構造解析の大きな手がかりとなっている。

本論文では、並列構造解析の一手法として、類似性を用いない手法を提案する。本手法は、構文・格解析の確率的生成モデルの枠組みの中で行う。並列構造を認識するための手がかりとして選択選好を用いるが、これには、大規模ウェブテキストから抽出した格フレームおよび名詞間の共起情報を用いる。

3 構文・並列・格構造解析の統合的確率モデル

本研究では、依存構造に基づく確率的生成モデルを提案する。本モデルは、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。

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

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 1 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

表 1: 構文構造の精度

評価対象	構	構+並+格 (w/類似性)	構+並+格 (wo/類似性)
すべて	3,829/4,404 (87.0%)	3,863/4,404 (87.7%)	3,873/4,404 (87.9%)
並列キー	878/1,108 (79.2%)	881/1,108 (79.5%)	894/1,108 (80.7%)

次のように定義する。

$$P(T, L, S) = \prod_{C_i \in S} P(C_i, rel_{ih_i} | C_{h_i})$$

ここにおいて、C_{h_i} 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。rel_{ih_i} は C_i と C_{h_i} の係り受け関係を表し、通常の係り受け (D) または並列 (P) の 2 値をとるものとする。

$P(C_i, rel_{ih_i} | C_{h_i})$ は、[2] をベースに定義するが、異なる点は以下のとおりである。

- ベースモデルでは、名詞並列に対して、並列前部は並列後部から生成していたが、本モデルでは、並列前部と並列後部ともに、それらを支配している述語から生成する。
- ベースモデルでは、並列の係り受け関係として類似度スコアを結合したものを用いていたが、本モデルでは類似度スコアを用いない。そのため、前述したように、係り受け関係は D または P の 2 値となる。
- 並列構造に係る連体修飾句を生成するときには、並列構造のそれぞれの主辞からその連体修飾句を生成する確率を計算し、その幾何平均をとることにする。

確率モデルの詳細は [2] を参照されたい。

4 実験

提案モデルによる解析実験を行った。格フレームと名詞間の共起情報は、ウェブテキスト約 5 億文から抽出したものを用いた。

4.1 構文構造の評価

並列構造の範囲同定の評価は、構文構造の評価に含まれると考え、構文構造の評価実験を行った。本実験は、ウェブテキスト 759 文[3]を形態素解析器 JUMAN[4]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[5]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 2 つ目までの文節以外の係り先すべてと、並列キー文節の係り先の 2 種類とした。

ベースラインとしては、構文解析器 KNP[6]と、類似性に基づく確率的構文・並列・格解析 [2] の 2 つを用いた。前者のシステムにおける並列構造解析は、構文解析を行う前に、類似性のもっとも高い並列構造を一意に決定する手法を用いている。後者のシステムは、並列構造の類似度を確率モデルの中で直接的に用いている。

表 1 に評価結果を示す。表において、「構」は構文解析器 KNP、「構+並+格 (w/類似性)」は確率的構文・並列・格解析 [2]、「構+並+格 (wo/類似性)」は提案手法を表している。提案手法の精度は、「構」「構+並+格 (w/類似性)」のそれぞれに対して、すべての係り受けでは 0.9%、0.2%、並列キー文節のみで 1.5%、1.2%向上した。すべての係り受けに対してマクネマー検定を行った結果、「構」に対する精度向上は有意 ($p < 0.01$) であったが、「構+並+格 (w/類似性)」に対しては有意ではなかった。

表 2 に、「構+並+格 (w/類似性)」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。例えば (1) の例では、「電車の発射合図や、」と「携帯電話の着信音までが」が並列構造を作ることが正しく解析できるようになった。これは、「(発射) 合図」と「(着信) 音」が「音楽になる」の格フレームから生成されやすいこととが考慮されたためと考えられる。

4.2 議論

本論文では、類似性を用いずに並列構造解析を行う手法を提案した。提案手法は、従来の類似性に基づく手法と同等の構文精度を達成することができた。

[3] これらの文は格フレーム構築とモデル学習には用いていない。
[4] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
[6] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 2: 解析が正しくなった例

(1)また、以前まで駅員が警告音としてけたたましく鳴らしていた電車の 発射合図や、 携帯電話の$_{\times}$ 着信音までが$_{\bigcirc}$ 音楽になるほどに、音楽の利用法が変わって来ています。

(2)鍋にだし汁 3カップと、 残りの$_{\times}$ しょうゆ、みりん、酒を$_{\bigcirc}$ いれて、イカを並べ入れ、落とし蓋をして、沸騰する迄は強火で、その後は弱火から中火で1時間程煮込んで下さい、たまに蓋をとって煮汁をかける事。

(3)その頃には練習スタジオはなく、 学校の音楽室、$_{\bigcirc}$ 昼間の スナックなどで$_{\times}$ 練習していました。

我々の手法は類似性を用いていないが、並列要素が、格フレームの同じ格スロットから生成されることを利用しているため、それらは本質的に類似していると考えられる。つまり、選択選好が類似度を生み出していると考えられる。これは、Lin の分布類似度 [3] の考え方に似ている。分布類似度は、文脈が似ている語は似ているという考えである。Lin は、コーパスの構文解析結果から係り受け関係 (係り受け関係にある 2 語とその間の関係) を抽出し、同じような係り受け関係をとる語は似ているという考えに基づいて類似度を学習している。Chantree らは、この選択選好に基づく分布類似度を用いて、類似性をチェックすることによって並列構造を同定する手法を提案している [1]。これに対して、提案手法は、より直接的に選択選好を並列構造解析に適用している。

提案手法の問題点として、入力文中に、選択選好として学習されていない語がある場合には解析に失敗することが多いということが挙げられる。

- 挫折を経験した３６歳の男・ 吉村浩輔と ドロップアウトした１８歳の少女・島崎未来の$_{\bigcirc}$ 夢と希望を描く ドラマで、$_{\times}$ 音楽が一方の “主役”。

この文において「吉村浩輔と」の正解係り先は「島崎未来の」であるが、提案手法は係り先を「ドラマで、」と解析し、誤りとなる。これは、「吉村浩輔の夢」という名詞句が学習されていないためである。このような固有表現は、語彙レベルで学習するのはほとんど不可能であるので、ある種の汎化が必要となる。例えば、「吉村浩輔の夢」を「<人名>の夢」のように汎化して扱うとすると、「<人名>の夢」はコーパス中に出現している可能性が高いと思われる。出現していないとしても、入力文中の「島崎未来の夢」から動的に「<人名>の夢」を学習できる可能性もある。

また、上記の例では「挫折を経験した３６歳の男・吉村浩輔」と「ドロップアウトした１８歳の少女・島崎未来」という文節列が並列構造をとっており、各構成文節もそれぞれ類似しているが、本研究のモデルでは考慮していないという問題がある。このような文節列の類似性は、類似性に基づく従来の解析手法では考慮されており、提案手法におけるオープンな問題である。

5 おわりに

本稿では、確率的構文・格・並列解析の枠組みにおいて、類似性を用いずに並列構造解析を行う手法を提案した。並列構造を同定する手がかりとして、大規模コーパスから抽出した大規模な選択選好選好を用いた。ウェブテキストを用いて解析実験を行った結果、構文・並列構造の解析に提案手法が有効であることがわかった。

参考文献

[1] Francis Chantree, Adam Kilgarriff, Anne de Roeck, and Alistair Wills. Disambiguating coordinations using word distribution information. In *Proceedings of RANLP2005*, 2005.

[2] Daisuke Kawahara and Sadao Kurohashi. Probabilistic coordination disambiguation in a fully-lexicalized Japanese parser. In *Proceedings of EMNLP-CoNLL2007*, pp. 306–314, 2007.

[3] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*, pp. 768–774, 1998.

[4] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, Vol. 11, pp. 95–130, 1999.

[5] 黒橋禎夫, 長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol. 1, No. 1, pp. 35–57, 1994.