樹状文脈に対する文脈木の構文解析への応用

森 信介
日本アイ・ビー・エム東京基礎研究所
〒242-8502 神奈川県大和市下鶴間1623-14

1 はじめに

本論文では、まず、確率に基づく構造的言語モデルにおいて、参照される履歴の範囲を柔軟に選択することを実現する樹状文脈木について述べる。単語 n-gram モデル等に利用される通常の文脈木は、ノードのラベルが単語列であり、木構造をなす履歴を扱うことができない。樹状文脈木は、ノードのラベルを木とすることにより木構造をなす履歴の参照範囲を可変とすることを実現する。さらに、このような柔軟な履歴参照機構をもつ構造的言語モデルの構文解析への応用について述べる。日本経済新聞からならるコーパスからパラメータを推定し、形態素列を入力とする構文解析実験を行なった結果、形態素単位の係り受け単位で92.8%の解析精度を得た。これは、樹状文脈木を用いない履歴の参照範囲が固定のモデルによる精度(89.8%)より有意に高く、履歴を可変とすることの優位性が実験的にも確認された。

2 係り受けを記述する言語モデル

この節では、我々が提案する係り受けに基づく構造的確率的言語モデルについて述べる。文は形態素の列とみなされ、モデルは文が与えられると、これを先頭から順に読む。このとき、各形態素の予測とその時点での係り受け構造を更新が交互に行なわれる。最後の形態素の予測とそれを含む文の構造が更新されると、モデルは入力文に対する複数の文構造とそれぞれの生成確率を返す。

2.1 構造的確率的言語モデル

構造的確率的言語モデルは、各形態素を、先行する形態素列ではなく、それを覆う部分解析木から予測する。したがって、文 $m = m_1 m_2 \cdots m_n$ とその構文木

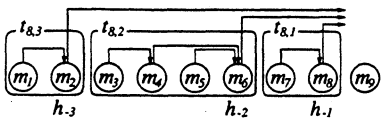

図 1: 部分解析木の例

T の組の生成確率 P は、以下の式で計算される。

$$P(m,T) = \prod_{i=1}^{n} P(m_i|t_{i-1})P(t_i|m_i,t_{i-1}), \qquad (1)$$

ここで t_i は i 番目の部分解析木の列を表す。図1は、9番目の形態素が予測されるときの状況を例示している。この図から、まず9番目の形態素 m_9 が8番目の部分解析木の列 $t_8 = t_{8,3}t_{8,2}t_{8,1}$ から予測され、次にこの部分解析木の列 t_8 と9番目の形態素 m_9 から9番目の部分解析木の列 t_9 が予測され、10番目の形態素を予測する直前の状態になる。

ここで問題となるのは、式(1)の2つの条件付確率の条件部分の分類方法である。英語の構造的言語モデル[1]では、以下の式のように2つの最右の部分解析木の主辞(図1の例では m_6 と m_8)を用いる。

$$P(m_i|t_{i-1}) \approx P(m_i|root(t_{i-1,2}), root(t_{i-1,1})),$$

ここで、$root(t)$ は木 t の根のラベル(形態素)を返す関数である。同様の近似が、構造の予測にも適用される。日本語の構造的言語モデル[2]では、部分木の主辞とそれに係る形態素で履歴を区別する。このように、従来の方法では参照される履歴の範囲は固定であるが、どの範囲が最適かは、学習コーパスのサイズや予測の際の部分解析木そのものに依存し、これを柔軟に選択する機構があれば、予測精度が有意に改善すると考えられる。

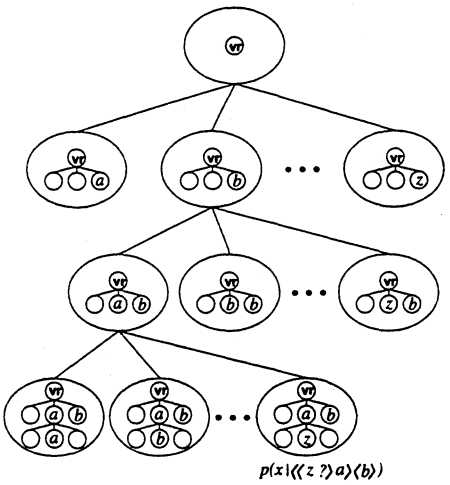

図 2: 樹状文脈木

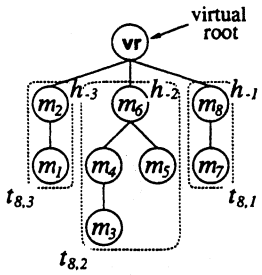

図 3: 履歴木

2.2 樹状文脈木

木構造の履歴の参照範囲を柔軟に選択する機構として、樹状文脈木 (Arbori-context Tree) [3] が提案されている。これは、履歴が記号列である場合の n-gram モデルにおいて可変長の履歴を扱う文脈木 [4] の自然な拡張である。文脈木のノードのラベルが記号列であるのに対して、樹状文脈木のノードのラベルは木になっている。各ノードには、履歴がラベルに一致する条件下での事象の確率分布が記述されている (図 2参照)。

構造的言語モデルでは、履歴は先行する形態素列を覆う部分解析木の列であるが、仮想の根を導入し、部分解析木の根をその子ノードとすることで、単一の木と見なすことができる。これを履歴木と呼ぶ。図 3は図 1の履歴木である。部分木の列 $t_1 t_2 \cdots t_n$ とそれを直下に持つ根ノード r がなす木を $\langle t_1 t_2 \cdots t_n \rangle r$ と表記するとすれば、図の履歴木は $\langle t_8 \rangle \mathrm{vr}$ となる。ここで、vr は仮想の根であり、この表記法での履歴木の最右に必ず存在するので、省略可能とする。樹状文脈木の各ノードのラベルは履歴木の根を含む部分木であり、その直下のノード列はラベルの木の任意のノードに各アルファベットに対応する葉を加えること (特殊化) により得られる一連のノードである。各ノードには、ラベルが履歴の部分木になっている場合の事象の確率分布が付与されている。図 2の右下のノードは、部分解析木の列の最後の木の根のラベルが b で、部分解析木の列の最後から 2 番目の木の根のラベルが a で、さらに、その右から 2 番目の子ノードが z である場合の次の事象 x の生起確率 $p(x|\langle\langle z? \rangle a \rangle \langle b \rangle)$ を保持している。ここで ? は任意のアルファベットを表すとする。図 1の状態では、$m_4 = z$, $m_6 = a$, $m_8 = b$ のとき、図 2の右下のノードのラベルが履歴木にマッチするので、このノードに付与されている確率分布を用いて、次の形態素を予測する。構造の予測に際してもほぼ同じであり、部分解析木の列の最右に、予測されたばかりの形態素のみからなる木を加えた履歴木を樹状文脈木で分類する。形態素予測と構造予測の樹状文脈木をそれぞれ ACT_m と ACT_s とし、その返値がラベルであるとすると、これらを用いる構造的言語モデルは、以下の式で表される。

$$P(\boldsymbol{m}, T) = \prod_{i=1}^{n} P(m_i | ACT_m(\langle t_{i-1} \rangle)) \times P(t_i | ACT_s(\langle t_{i-1} m_i \rangle)),$$

固定の履歴を用いるモデルは、ACT_s と ACT_m が常に履歴木の一定の部分を返す場合であり、樹状文脈木を用いるモデルは、従来の構造的言語モデルの自然な一般化になっている。

3 構文解析

確率的言語モデルに基づく構文解析器は、形態素列 $\boldsymbol{m}$ を与えられると、確率が最大となる構造を以下の式に従って出力する。

$$\hat{T} = \operatorname*{argmax}_{T} P(T|\boldsymbol{m})$$

$$= \operatorname*{argmax}_T P(T|\boldsymbol{m})P(\boldsymbol{m})$$
$$= \operatorname*{argmax}_T P(\boldsymbol{m}|T)P(T) \quad (\because \text{Bayes' formula})$$
$$= \operatorname*{argmax}_T P(T) \quad (\because P(\boldsymbol{m}|T) = 1)$$

最後の行の $P(T)$ は、葉の形態素列が $\boldsymbol{m}$ に等しい統語構造の生成確率であり、本論文で述べる構文解析器では、2節で述べた構造的言語モデルによって計算される。

4 評価

2節で述べた、参照する履歴が固定あるいは可変となる構造的言語モデルを構成し、3節で説明した構文解析手法を用いる構文解析器を作成した。それぞれを同一の学習コーパスから推定し、同一のテストコーパスに対する構文解析を実験を行なった。この節では、この結果を提示し、解析精度を比較検討する。また、文節単位での係り受け解析のみを行なう手法との比較のため、文節を仮定した場合の解析精度を算出し、既存手法との比較について述べる。

4.1 実験の条件

実験には、日本経済新聞の記事からなるコーパス (表1参照) を用いた。各文は、形態素に分割され、構文構造が付与されている。形態素は比較的短く、平均文字数は 1.54 である。また、品詞の分類は 16 種類と比較的粗い (品詞情報が構文解析にあまり寄与しない)。

モデルの構成と構文解析に際しては、機能語に分類される品詞 (助詞、助動詞、語尾、記号) のみ表記を区別し、内容語は品詞で代表させることとした。この結果、言語モデルのアルファベットは、192 個の機能語と 4 個の機能語の未知語を表す記号と 12 個の内容語の品詞から構成されることとなる。内容語の表記を縮退させた理由は、データスパースネス問題を避けることである。なお、内容語の既知形態素数は 9,216 であった。

4.2 評価

構造的言語モデルによる構文解析器を作成し、精度の評価を行なった。文字列に対する構文解析も可能であるが、形態素への分割の誤りが生じ、結果の評価が困難であるため、構文解析の精度の評価には形態素列を入力とした場合の出力を用いた。精度は、以下の式で示されるように、推定した係り受け関係の数に対する、正しい係り受け関係の割合である。ここで、正しい係り受け関係とは、ある形態素の係り先がコーパスに付与された係り先に一致していることを意味する。

$$\text{解析精度} = \frac{\text{係り先が正しい形態素の数}}{\text{形態素の数}}$$

係り先がない最後の形態素と係り先が明白な最後から 2 番目の形態素は、評価の対象にしない。

表 1: コーパス

	文数	形態素数	文字数
学習	9,108	260,054	400,318
テスト	1,011	28,825	44,667

表 2: 形態素単位の係り受けの精度

言語モデル	解析精度
可変履歴のモデル	92.8% (24,867/26,803)
固定履歴のモデル	89.8% (24,060/26,803)
ベースライン *	79.4% (21,278/26,803)

* それぞれの形態素は次の形態素に係るとする

表 2は、予測に際して可変履歴を用いるモデル、固定履歴を用いる従来のモデル、および、それぞれの形態素は次の形態素に係るとした場合のベースラインの解析精度である。可変履歴を用いることで、従来モデルの誤りの約 30% が削減されている。このことから、参照履歴を可変にすることが構文解析において有効 (危険率 0.01 で有意) であることが分かる。

図 4は、学習コーパスの大きさと解析精度の関係を示すグラフである。右上部分の傾きから、学習コーパスを 4 倍にすることで解析精度の約 1.8% の上昇が見込めることを示している。現在の学習コーパスは 9,108 文であり、これを 4 倍にするのは十分可能なので、学習コーパスの増量という単純な戦略で 95% 程度を達成できるであろう。

表 3は、文節を単位とする従来の係り受け解析との

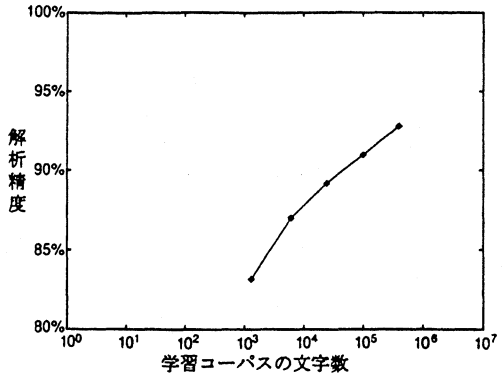

図 4: 学習コーパスの大きさと解析精度の関係

比較である。テストコーパスの 100 文を形態素解析器 (JUMAN-3.61) と構文解析器 (KNP-2.0b6)[1]で解析し、各文節の係り先が正しいか否かを人手でチェックした。構造的確率的言語モデルによる解析精度は、文節の定義を KNP の出力と同一にし、各文節の係り先を、文節末の形態素の係り先の形態素を含む文節とすることで算出した。実験の結果、提案手法による精度は従来手法を上回った。しかし、検定の結果、精度の差は有意水準 10% でも棄却された。また、以下に列挙する点から精度の差は有意とはいえず、同程度の精度と見なすのが適切であろう。

- テスト文が少ない。
- 対象とする曖昧性解消の範囲が大きく異なる。
  KNP: 品詞細分類から文節単位の係り受け
  (助詞の統語的分類等は形態素解析による)
  本手法: 品詞大分類から形態素単位の係り受け
  (複合語の構造解析等も含む)
- 本論文の実験環境の KNP は有償の辞書を利用していない。

すでに述べたように、データスパースネス問題が生じるため、現状では内容語の表記を利用していない。学習コーパスの一部をヘルドアウトデータとし、表記の情報を利用する形態素を選択すると、解析精度はわずかながら改善し 92.9% となった。しかし、この結果から、この方法では内容語の表記の情報を適切に利用

[1] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/

表 3: 文節単位の係り受けの精度

言語モデル	解析精度
可変履歴のモデル	87.8% (674/768)
JUMAN+KNP	85.3% (655/768)
ベースライン*	62.4% (479/768)

* それぞれの文節は次の文節に係るとする

していると は考えられず、この点の改善による精度向上の余地があろう。

5 結論

本論文では、形態素単位の係り受け構造に基づく確率的言語モデルに樹状文脈木を導入し、参照される履歴の範囲を柔軟に選択することについて述べた。次に、このモデルに基づく構文解析器を実装し、日本経済新聞に対する実験を行なった。精度評価の結果、可変履歴を用いることで、固定履歴のモデルより解析精度が高いことが確認された。

参考文献

[1] Ciprian Chelba and Frederic Jelinek. Structured Language Modeling. *Computer Speech and Language*, Vol. 14, pp. 283–332, 2000.

[2] 森信介, 西村雅史, 伊東伸泰, 荻野紫穂, 渡辺日出雄. 形態素係り受けモデルによる構文解析. 情報処理学会研究報告, 第 2000-NL-140 巻, 2000.

[3] Shinsuke MORI, Masafumi NISHIMURA, and Nobuyasu ITOH. Improvement of a Structured Language Model: Arbori-context Tree. In *Proceedings of the Seventh European Conference on Speech Communication and Technology*, 2001.

[4] Dana Ron, Yoram Singer, and Naftali Tishby. The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length. *Machine Learning*, Vol. 25, pp. 117–149, 1996.