確率付き項構造による曖昧性解消

宮尾 祐介　辻井 潤一
東京大学大学院理学系研究科情報科学専攻
{yusuke,tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

本稿では意味表現 [9] に確率値を割り当てることを最終的な目標として，それに必要なデータ構造である確率的素性構造を提案する．これは，型付き素性構造 [6] の上に最大エントロピーモデル [2] で確率モデルを構築したものである．型付き素性構造は単一化ベースの文法理論 (HPSG [11], LFG [4] など) や意味表現の記述に利用されているデバイスである．型付き素性構造に対して確率値を割り当てることで，素性構造で表現される意味表現がどの程度もっともらしいかを定量化することができると考えられる．本論文は意味表現の曖昧性解消に有利な確率モデルを提案することを目的としており，意味表現の具体的な形式については素性構造で表現されているということ以外はなんら言及しない．

現在までに曖昧性解消の研究は数多くされており，語彙的確率文脈自由文法に基づく統計的パーザ [8] が高い精度を挙げているが，これらのパーザの出力は括弧付けや構文木にとどまっており，様々なアプリケーションで解析結果を利用するには不十分であると考えられる．3節で説明するように，本稿で提案するモデルは語彙的確率文脈自由文法の一般化になっており，さらに木構造だけでなく複雑なグラフ構造で表現された関係を捉えることができる．

型付き素性構造の上に確率モデルを構築するために，本研究では最大エントロピーモデル [2] を適用した．現在までに相対頻度モデルに基づき素性構造に確率を割り当てる研究 [5] がされてきたが，素性構造を独立な事象に分割することが難しいため (Abney [1] を参照)，有用なモデルは提案されなかった．最大エントロピーモデルは確率事象を任意の副事象 (素性) に分割し，その副事象間の独立性を仮定しないため，素性構造に対して一貫性があるモデルを構築できる．

確率的素性構造の有効性を評価するため，本モデルを単一化ベースの構文解析システムを用いて得られた項構造に適用し，曖昧性解消の実験を試みた．その結果，単純に括弧付けや構文木の精度を測定した場合は他の統計的パーザほどの性能は出なかった．しかしその誤りはほとんどの場合システムと正解との解釈の違いによるものであり，それを除くと高い精度で正解の項構造が得られた．本稿ではまだ小規模な実験にとどまっているため，さらに大規模な実験で評価する必要があると考えられる．

2節では確率的素性構造を形式的に定義する．3節では確率的素性構造を項構造に適用することによる利点について議論する．4節では，確率付き項構造を用いた曖昧性解消実験の結果について分析する．

2 確率的素性構造

最大エントロピーモデル [2] では，履歴事象 h のもとでの目標事象 t の生起確率 $p(t|h)$ を，訓練データにおける事象 (t,h) の相対出現頻度 $\tilde{p}(t,h)$ から推定する．具体的には，$p(t|h)$ は以下のような式でモデル化される．

$$p(t|h) = \frac{1}{Z_h}\prod_i \alpha_i^{f_i(t,h)}$$

$$Z_h = \sum_t \prod_i \alpha_i^{f_i(t,h)}$$

ここで，f_i は事象を副事象に分割するための関数で，素性関数[1]と呼ばれる．事象 (t,h) がある特性を持つとき，$f_i(t,h)$ は1以上の整数値をとり，それに対応する重み α_i が加えられる．α_i は，訓練データ $\tilde{p}(t,h)$ の尤度を最大化するように推定される．

本研究では，素性構造 s が与えられたときそれがある文の意味表現となっている確率をモデル化する．具体的には，素性構造 s が与えられたとき，それが選択すべき意味表現であるという事象 $B = \{True, False\}$ を考え，確率 $p(B|s)$ を以下のような最大エントロピーモデルで与える．

$$p(B|s) = \frac{1}{Z_s}\prod_i \alpha_i^{f_i(B,s)}$$

$$Z_s = \sum_B \prod_i \alpha_i^{f_i(B,s)}$$

ここで問題となるのは，素性関数 f_i の形式である．最大エントロピーモデルを適用した他の研究で明らかなように，システムの性能は素性関数の良し悪しでほぼ決定される．ここでは最終目標が意味表現の曖昧性解消であるため，以下のような素性構造指標に基づく素性関数を提案する．

$$f_i(s) = \sum_{\pi \in \Pi_s} Unif(t_i, \pi(s))$$

[1] まぎらわしいが，「素性構造」の素性と「素性関数」の素性は関係がない．

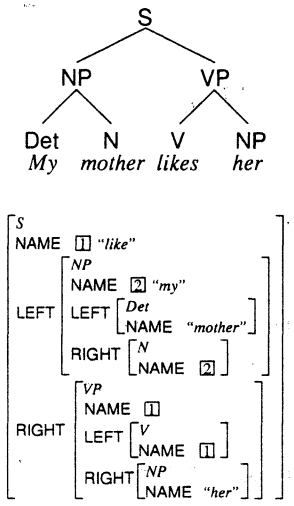


図 1: 文 "My mother likes her" に対する構文木とその素性構造による表現.

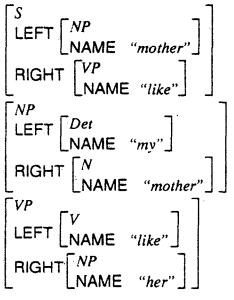


図 2: 図 1 の構文木の分岐に対応する素性構造指標.

ただし, t_i はモデルの設計者が与える素性構造で, 本稿では素性構造指標と呼ぶ. Π_s は素性構造 s 中の非循環のパスの集合, $\pi(s)$ は s のパス π の値を表す. $Unif(s_1, s_2)$ は素性構造 s_1, s_2 が単一化可能なときに 1, 不可能なときに 0 を返す関数である. 直感的には, 素性構造 s の中にある特徴的な素性構造 t_i が現れたとき, 素性関数はインクリメントされる. すなわち, 素性構造指標 t_i によって素性構造 s の特徴を捉えそれに対応した α_i で重みづけをする.

このような素性構造指標に基づく素性関数を用いると, 語彙的確率文脈自由文法で捉えている事象を本モデルでも捉えることができる. 例えば, 図 1 は "My mother likes her" という文に対する構文木とその素性構造による表記である. 図 2 にあるような素性構造指標を用いると, 構文木における各分岐の生成確率を捉えることになる. このように, 本モデルは語彙的確率文脈自由文法を一般化したものになっている. ここでは, 主辞と非終端

want(1 you, drink(1, what)) .

図 3: 文 "What do you want to drink" に対する項構造.

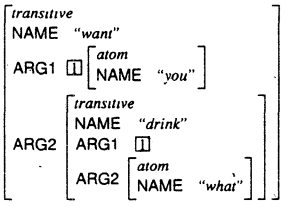


図 4: 図 3 の項構造の素性構造による表現.

記号の情報しか記述されていないが, 他の確率的パーザ [8] のように単語間距離などを素性構造に追加することもできる. さらに次節で述べるように, このモデルを意味表現に適用することにより, 構文木中の各分岐の直接の子の二項関係だけでなく任意の多項関係や非有界依存の関係も捉えることができる.

3 項構造への適用

本研究は, 確率的素性構造を意味表現に適用し, 確率付き意味表現を用いて曖昧性解消を行うことを目的としている. したがって, まず意味表現は素性構造で表現されているという仮定を置く. また, 2 節で定義した確率的素性構造を用いて曖昧性解消を行うためには, 曖昧性解消に役立つ情報が素性構造指標で捉えられなくてはならない. したがって, 意味表現について以下のような直感的な仮定を置く.

- 直接的に意味の関連がある実体どうしは直接素性で結ばれている.

意味表現に関してはこれ以外にはなんら仮定を置かない. 語彙的構文木や項構造はこの条件を満たしているため, ここで言う意味表現に含まれる.

本稿では, 確率的素性構造の評価のために意味表現として単純な項構造を採用する. 項構造は構文木より多くの情報を記述しており, また, より複雑な意味表現の実装にはまだ揺れがあり一貫性のある意味表現を実装するのが難しい. より複雑な意味表現における確率的素性構造の有効性の評価は今後の研究課題である.

図 3 は文 "What do you want to drink" に対する項構造を表している. また, 図 4 はその素性構造による表現である. 図 5 はそれに対応する構文木である. 構文木では drink と you のコントロールの関係, drink と what の非有界依存の関係が表現されていないが, 項構造では表現されている. したがって, 図 6 のような素性構造指標を用いることにより, 本モデルではこれらの関係を捉えて確率値を割り当てることができる. また, 構文木で

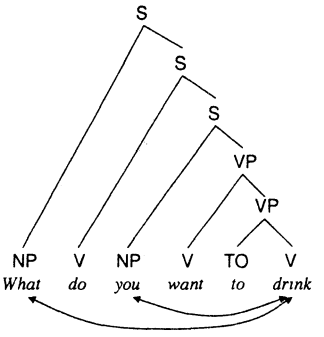


図 5: 図 3 の項構造に対応する構文木.

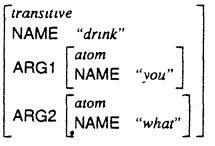


図 6: 素性構造指標の例.

は 2 つの娘の間の関係しか捉えられないが，本モデルでは素性構造指標は任意の素性構造なので，図 6 で表現されているような “drink”, “you”, “what” の 3 項関係や，“want”, “drink”, “what” のような再帰的な構造も捉えることができる.

4 実験結果

本節では，HPSG に基づく大規模な英文法である XHPSG 文法 [12] を用いた曖昧性解消の簡単な実験を行い，2 節で提案した確率的素性構造の有効性を評価する．XHPSG 文法は可能な全ての項構造をし，それにコーパスから学習した確率的素性構造モデルで確率値を割り当てる．学習データとテストデータには，Penn Treebank [10] を利用した．Penn Treebank には項構造を復元するためのタグや非有界依存の関係を表すための指標付けがされている．Penn Treebank の ATIS コーパスから非文を取り除き，337 文について項構造を獲得した．それを学習データ 307 文，テストデータ 30 文にランダムに分割した．最大エントロピーモデルのパラメータ学習には ChoiceMaker Maximum Entropy Estimator (CMEE) [3] を使用した．素性構造指標は，述語とそれの直接の項からなる素性構造のみを使用している．また，述語のラベル (図 4 における NAME 素性) には品詞を用いたモデル (品詞モデル)，表層文字列を用いたモデル (表層モデル)，その両方を用いたモデル (品詞 + 表層モデル) を用いて実験を行った．今回は実装の容易さから品詞や表層文字列を利用したが，さらに意味範疇などを利用することにより，モデルを改善することが出来ると考えられる.

表 1: 確率的素性構造を利用した曖昧性解消実験の結果.

モデル	括弧付け	構文木
品詞	71.35	64.04
表層	73.03	65.73
品詞 + 表層	73.60	66.29

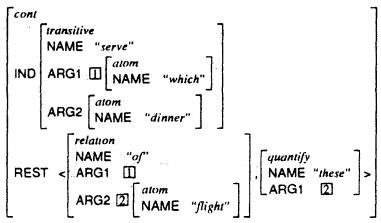
表 2: 表 1 の実験の文ごとの分析.

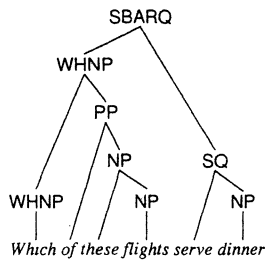
誤りの原因	文数
正解	3
解釈の違いのみ	8
XHPSG が正解を出さない	5
不正解＋解釈の違い	2

本システムは項構造を出力するので本来は項構造の精度を測定すべきであるが，項構造の精度を測定するには素性構造間の類似度を計算しなければならないため，非常に難しい．そこで，本稿では出力した項構造に対応する構文木の括弧付けの精度と構文木の精度を測定した．より詳細な文法的関連の精度を測定する研究 [7] を元に，今後項構造の精度を評価することが必要であると考えている.

括弧付けと構文木の精度は表 1 のようになった．現在のところ最先端の統計的パーザ [8] ほどの性能は出ていない．主な理由は，XHPSG システムが Penn Treebank と同じ構文木を出力することを目的として開発されていないため，XHPSG システムと Penn Treebank での構文構造の解釈の違いが誤りとなってしまうからである．そのため，本来は正解とすべきであるが解釈の違いから誤りとなってしまう場合が多い．例えば，図 7 は “Which of these flights serve dinner” という文に対する XHPSG システムの出力 (上) とそれに対応する構文木 (中)，同じ文に対する Penn Treebank の構文木 (下) を表している．XHPSG が出力した項構造は正解であると考えられる．しかしそれに対応する構文木では “flights” 単体に非終端記号 NP が付与されているが，Penn Treebank の正解では “these” と結び付いてからはじめて NP が付与されている．XHPSG 文法では名詞と名詞句との明確な区別がないため，このような違いが生まれている．また，XHPSG では “serve dinner” の句に直接 SQ(疑問文) が付与されているが，Penn Treebank では一度 VP が付与されてからさらに SQ が付与されている．これは，XHPSG 文法では “serve” の主語 (“Which”) が埋められていない状態を SQ としているためである．以上のような違いは言語現象の捉え方の違いに起因するもので，本来誤りとすべきものではないと考えられる.

表 2 は，そのような解釈の違いによる誤りがどの程度現れているかを調べた結果である．「正解」は構文木が完





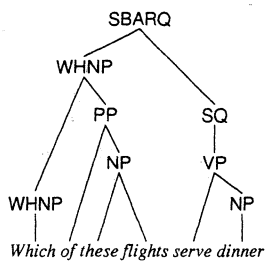


図 7: 文 "Which of these flights serve dinner" に対する XHPSG システムが出力した項構造 (上), それに対する構文木 (中) と Penn Treebank の正解 (下).

全に一致しているもの,「解釈の違いのみ」は前述した違いのみで本来は正解とすべきものである.「不正解+解釈の違い」は, 解釈の違いと不正解を両方含んでいるものである. このように, 多くの場合について正解の項構造を出力していることが分かる. 一方, XHPSG システムが曖昧性解消以前にそもそも正解の項構造を出力できていない場合も無視できない. これについては, XHPSG システムのさらなる改良が必要である.

5 おわりに

本稿では, 構文解析器の出力として意味表現に確率値を付与したものを想定し, それに必要なデータ構造として確率的素性構造を提案した. それを利用して, 単一化ベースの文法を用いて得られた項構造の曖昧性解消の実験を行い, その有効性を評価した.

現在のところ, 本モデルを実装した XHPSG システムは構文解析器として実用的なレベルにはまだ達していない. 実用的なシステムにするため, 以下の研究が今後必要であると考えている.

- XHPSG 文法の拡張によりカバレッジを広げる.
- 構成的な計算により確率値計算を高速化する.
- 大きなコーパスを用いた学習と大規模な実験を行う.
- 曖昧性解消を項構造の精度で評価する.

参考文献

[1] Steven P. Abney. Stochastic attribute-value grammars. *Computational Linguistics*, 23(4), 1997.

[2] Adam L. Berger, Stephen A. Della Pietra, and Vincent. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[3] Andrew Borthwick. Choicemaker maximum entropy estimator, 1999. ChoiceMaker Technologies, Inc. Email `borthwic@cs.nyu.edu` for information.

[4] Joan Bresnan, editor. *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge, MA, 1982.

[5] Chris Brew. Stochastic HPSG. In *Proc. 7th EACL*, pages 83–89, 1995.

[6] Bob Carpenter. *The Logic of Typed Feature Structures.* Cambridge University Press, 1992.

[7] John Carroll and Guido Minnen. Can subcategorisation probabilities help a statistical parser? In *Proc. WVLC-6*, pages 118–126, 1998.

[8] Michael Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL*, 1997.

[9] Ray Jackendoff. *Semantic Structures.* The MIT Press, 1990.

[10] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Proc. AAI'94*, 1994.

[11] C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, 1994.

[12] Yuka Tateisi, Kentaro Torisawa, Yusuke Miyao, and Jun'ichi Tsujii. Translating the XTAG English grammar to HPSG. In *Proc. TAG+4 Workshop*, pages 172–175, 1998.