語彙的・構造的意味情報を用いたパーズランキング

藤田 早苗,♡ Francis Bond,♠ Stephan Oepen,♣ 田中 貴秋♡
♡ {sanae,takaaki}@cslab.kecl.ntt.co.jp, ♠ bond@ieee.org, ♣ oe@csli.stanford.edu
♡ NTT コミュニケーション科学基礎研究所 ♠ NiCT 情報通信研究機構
♣ University of Oslo and CSLI Stanford

1 はじめに

構文解析システム、意味解析システムといった解析システムを利用する上で、 問題となるのが、膨大な解析曖昧性である。 一般に、解析結果は膨大な量になるため、これらの解析システムの結果を人手 で修正せずに利用するためには、膨大な解析結果の中から、より確からしい解析結 果を上位にランキングすることが求められる。

Riezler et al. (2002)や Oepen et al. (2004) は、高度な言語知識を含む文法による解析に統計モデルを導入することにより、 パーズランキングの精度を向上をさせることに成功している。 そこで、本稿では、更に語の意味クラスや上位概念などを定義した外部言語資 源を用いて、さらなる精度向上を目指す。

これまで、ほとんどの統計モデルは、訓練データそのものから得られる情報か らのみ作成した素性を用いて学習している。 しかし、語の意味クラスや上位概念などの意味情報を利用すれば、 表層的な字面ベースの素性ではスパース過ぎる場合でも、 統計モデルをスムージングして精度をあげることができると考えられる(Korhonen, 2002)。 しかしながら、これまで、パーズランキングの訓練データ(ツリーバンク) に意味情報が付与された言語資源がなかったため、訓練に意味情報を用いる実験がほとんどされてきていない。

そこで、本稿では、ツリーバンク(HPSGに基づく解析)と同じ文に 意味情報(辞書に基づく語義タグ)が付与された 檜コーパス(Bond et al., 2007)を用い て、統語情報だけでなく、意味情報を用いたパーズランキングの統計モデルを作成し、 パーズランキングにおける意味情報の有効性を調査する。

また、知識ベースの機械翻訳など、意味処理を行なう自然言語処理では、 用言の結合価や選択制限などの情報は非常に効果的であることがわかっ ている(Ikehara et al., 1997)。 そこで、既存の結合価辞書との対応関係を自動的にとり、 結合価や選択制限の情報を統計モデルに反映させて、それらの情報の有効性を調査する。

その結果、統語情報のみを利用した場合のモデル(55.3%)より、意味情報や結 合価情報を同時に用いたモデルの方が精度が高い(62.9%)事を示す。

2 檜コーパス

本稿では、統語情報と意味情報の両方を持つ日本語の檜コーパス(Bond et al., 2007) を用いて実験を行なう。檜コーパスは、辞書(Lexeed)の定義文、例文、新聞文に対するツリーバンク、および、 センスバンクから構成される。また、この辞書には、語義毎に 日本語シソーラスである日本語語彙大系(Ikehara et al., 1997)の意味 属性が付与されている。語彙大系は、2,710の意味属性からなり、深さ0から11階層ま での階層に分けられている。 図 1に、簡略化したLexeedのエントリの例を示す。

こうした、統語・意味情報を両方持つコーパスは、他に、 英語では、Penn treebankに意味タグを付与したOntoNotes (Hovy et al., 2006)などがある。

2.1 統語情報

檜ツリーバンクは、Jacy (Siegel and Bender, 2002)を用いて解析されている。 Jacyは、 統語解析と意味解析が密接に関連した解析が可能 な、主辞駆動句構造文法(HPSG: Head-driven Phrase Structure Grammar) (Pollard and Sag, 1994)に基づいた文法である。

運転手$_1$の定義文(図 1)の解析結果のうち、 正しい導出木を図 2に示す。ここで、右下の添字は語 義番号を示している。 解析候補は4候補あるが、檜コーパスでは、 正しい解析結果を人手で選択してある。

また、意味解析の結果を簡単な構造的意味情報に 修正したものが図 3である。 図 3の左側が主となる述部、右側の[ ]の部分は関係する名詞句とラベル、( )の部 分は各事象のインデックスである。 これらには、述語-項構造や並列構造などの情報が含まれる。 また、< >は、入力文中の文字位置を示している。 例えば、densha_n_1は、文中の0から2の位置、すなわち「電車」に対 応する。

```
proposition_m(h1)<0:12>[MARG unknown(e2)]
unknown(e2)<0:12>[ARG _hito_n(x5)]
densha_n_1(x7)<0:2>
_jidousha_n(x12)<3:6>
_ya_p_conj(x13)<0:3>[L-INDEX _densha_n_1(x7),
                     R-INDEX _jidousha_n(x12)]
_unten_s(e23)<7:9>[ARG1 _hito_n(x5),
                   ARG2 _ya_p_conj(x13)]
proposition_m(h2)<0:12>[MARG _unten_s(e23)]
```

図 3: 運転手$_1$の定義文:簡略化した構造的意味情報

2.2 意味情報

檜センスバンクでは、檜ツリーバンクの対象文(定義文、例文、京大 コーパスなど)と同じ文に対し、 各語が辞

[見出し語　運転手 (品詞 : 名詞)
語義 1 [定義文 [電車$_1$ や 自動車$_1$ を 運転$_1$ する 人$_4$](図 2,3)
例文 [大きく$_5$ なっ たら 電車$_1$ の 運 転手$_1$ に 成る$_6$ の が 夢$_3$ です 。]
意味属性 ⟨292:**運転手**⟩ (⊂ ⟨4:**人**⟩)]]

図 1: Lexeedの例:運転手$_1$の情報の一部

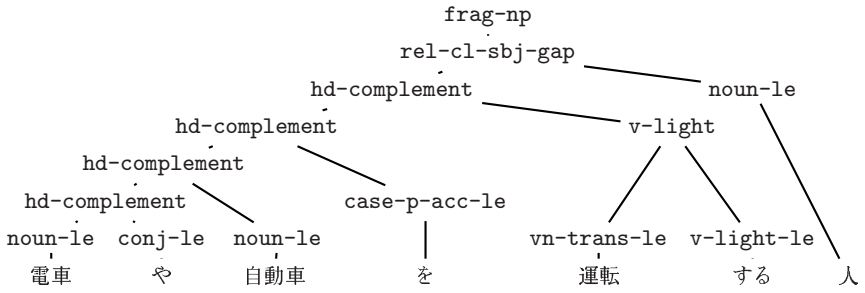

図 2: 運転手$_1$の定義文:導出木
ノードに付与されたラベルは、語の語彙タイプおよび、適用された文法規則の識別子である。

書Lexeedのどの語義に当たるかをタグ付けしてある。 例えば、運転手$_1$の定義文では、「運転」は全4語義中、一番目の語義(運転$_1$)、つまり、「大きな機械、乗り物などをあやつり動かすこと。」という語義でタグ付けされている。また、運転$_1$は、Lexeedで語彙大系の意味属性⟨2003:**操縦**⟩にリンクされている。

3 パーズランキング

解析候補のうち、最もよい解析結果を選択するためのモデルを 構築する。最終的には、解析候補に対し、確率によるランキングを行ないたい。 そのため本稿では、様々な統語的、意味的素性を定義し、 それらの素性の効果を、統計的機械学習によって検証する。

3.1 統語素性

まず、HPSGの導出木(図 2)から統語素性を作成する。 作成方法は、英語のツリーバンクであるRedwoodsを対象にToutanova et al. (2005)が用いた素性と同様であ り、深さ1の導出木の部分木から作る素性(以下、SYN-1)、先祖の情報を追加した素性(SYN-GP)、語彙タイプの*n*-grams の素性(SYN-ALL)を利用する。 本稿の実験では、実験的に良かったunigram, bigramを用いている。

3.2 意味素姓

図 3のような構造的意味情報を用いて 意味素姓を作成する。

まず、述部とその項の組合せから、Toutanova et al. (2005) と同様に、意味素性(SEM-BS)を作成する(表1)。これを基本の 意味素性とする。 ここで、図 3には 、_unten_s は 、ARG1_hito_n_1、ARG2 ”(_densha_n_1) _ya_p_conj (_jidousha_n_1)”と いう2つの項をとるという情報が含まれる。 これらの情報をすべてもつ素性(# 20)、述語と各項の組合せに展開した素性(# 21)、ARG1などの素性を取り除いた 素性(# 22,23)を作成する。

#	sample features
20	⟨0 _unten_s ARG1 _hito_n_1 ARG2 _ya_p_conj⟩
20	⟨0 _ya_p_conj LIDX _densha_n_1 RIDX _jidousha_n_1⟩
21	⟨1 _unten_s ARG1 _hito_n_1⟩
21	⟨1 _unten_s ARG2 _jidousha_n_1⟩
21	⟨1 _ya_p_conj LIDX _densha_n_1⟩
21	⟨1 _ya_p_conj RIDX _jidousha_n_1⟩
22	⟨2 _unten_s _hito_n_1 _jidousha_n_1⟩
23	⟨3 _unten_s _hito_n_1⟩
23	⟨3 _unten_s _jidousha_n_1⟩

表 1: 意味素性の例(SEM-BS): 図 3から作成
1列目の番号は、素性を作成方法によって区別するために付与したものである。

次に、檜センスバンクで、各語に付与されている語義を用いて、SEM-BS の内容語を、語義(SEM-WS)、および、各語義に付与された意味属性に置き換える(SEM-Class)。例えば、電車$_1$(_densha_n_1)と自動車$_1$(_jidousha_n_1)の意味属性は、両方⟨988:**乗り物(本体(移動(陸圏)))**⟩、運転$_1$(_unten_s)の意味属性は⟨2003:**操縦**⟩、人$_4$(_hito_n_1)の意味属性は⟨4:**人**⟩である。これらを用いて、表 1の内容語の部分を意味属性に置き換えた素性(SEM-Class)の一部を、表 2に示す。ここで、項の内容語のみ意味属性で置き換える場合(#40,41-0)と、すべての内容語を置き換える場合の両方の素性(#40,41-1)を作成している。

語義情報を用いる場合、素性には、語ベースよりも詳細な情報が含まれ、意味属性を用いる場合には、語を2,700の意味クラスでスムージングすることになる。

更に、本稿では、各意味属性を一定のレベル以上の上位意味属性へと置き換えることで、さらなるスムージングをはかる(SEM-L) 。本稿では、レベル2から5までで実験した。意味属性は、レベル2の場合9ク

#	sample features
40	⟨0 _unten_s ARG1 C4 ARG2 C988⟩
40	⟨1 C2003 ARG1 C4 ARG2 C988⟩
40	⟨1 C2003 ARG1 C4 ARG2 C988⟩
40	⟨0 _ya_p_conj LIDX C988 RIDX C988⟩
41	⟨0 _unten_s ARG1 C4⟩
41	⟨0 _unten_s ARG2 C988⟩
41	⟨1 C2003 ARG1 C4⟩
41	⟨1 C2003 ARG2 C988⟩

表 2: 意味素性の例(SEM-Class): 意味属性を利用

ラス、レベル3の場合30クラス、レベル4の場合136クラス、レベル5の場合392クラスに集約される。例えば、レベル3の場合、⟨988:**乗り物(本体(移動(陸圏)))**⟩の上位カテゴリは⟨706:**無生物**⟩, ⟨2003:**操縦**⟩は⟨1236:**人間活動**⟩で、⟨4:**人**⟩はもともとレベル3なのでかわらない。

最後に、日本語語彙大系の結合価辞書[1]の日本語側の情報を用いて素性を作成する(SP)。Jacyには、選択制限や選択嗜好の情報は組み込まれていないが、結合価辞書を利用することで、選択制限の情報を反映できる。

この結合価辞書は、動詞や形容詞の選択制限や下位範疇化構造を含んでいる。図 4に「運転する」の簡略化したエントリを示す。ここで、N1,N2等は主格や目的格を表す変数である。また、⟨ ⟩で示したのは格の選択制限であり、語彙大系の意味属性のリストで与えられる。この結合価辞書には、18,512エントリ、10,146 種類の動詞、2,618エントリ、1,723種類の形容詞の情報が登録されている。

```
┌  N1   ⟨4:人⟩          "が"
├  N2   ⟨986:乗り物⟩    "を"
└  運転する
```

図 4: 「運転する」の結合価辞書エントリ(ID:300513)

結合価辞書の情報を利用するため、構造的意味素性(図 3)から抽出できる述語-項構造と、述語の一致する結合価辞書エントリの選択制限との一致度を計算する。この一致度の高低(High/Med/Low/Zero)などから素性を作成する。

4 評価と結果

各素性毎に統計モデルを作成し、各素性の有効性を評価する。以下、SYN-1, SYN-GP, SYN-ALLを用いた統計モデルを統語モデル、SEM-?, SPを用いた統計モデルを意味モデルと呼ぶ。

ここで、評価実験には、檜コーパスのうち定義文を用いた。檜ツリーバンクは、出力された解析候補から、作業者が正解を選択して作成している。そこで、作業者が選択した解析結果を正例、選択しなかった結果を負例として、学習データを作成する。このため、正しい解析結果しか出力されなかった場合は学習データとして利用できない。また、複数の正解を許している。

実験では、定義文の全セットを、訓練データ(30,345文)とテストデータ(2,790 文)に分け、訓練データの一部(Train-sub)を予備実験に利用している。これらの文の平均長さは、9.4単語/文、平均解析候補数は101.0/文である。

学習器は、対数線形モデルに基づく*maximum entropy / minimum divergence* (MEMD) を利用した。具体的には、**Toolkit for Advanced Discriminative Modeling** (TADM:[2] Malouf, 2002) を用いた。

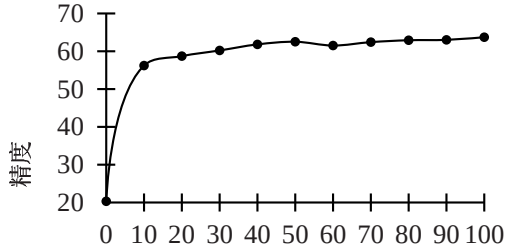

図 5: 学習曲線(SEM-ALL)

4.1 結果

表 3に各素性を用いて学習した結果を示す。精度は、文全体が正しいかどうかで評価している。つまり、1位にランクされた解析結果が檜コーパスの正解と完全に一致した場合のみ、正解としている。これは、部分木が正しくても全体が正しくないと正解にならないため、厳しい評価方法といえる。

表 3より、最も単純な統語モデル(SYN-1)は、最も単純な意味モデル(SEM-BS)より精度がよい。統語モデルに先祖の情報を追加したり(SYN-GP), *n*-grams データを追加した場合(SYN-ALL)、結果はSYN-1より改良されている。

意味素性の方では、SEM-BSを基本に、他の意味素性を追加する形で実験を行なっている。+SEM-WSでは、若干下がっているが、これは、データがスパースになったためと思われる。+SEM-Classでは、若干の上昇が見られる。しかし特に有効な素性は、上位意味属性(+SEM-L)と、結合価情報(+SP)であり、これらを用いた結果では、統語素性と同程度の精度を出す事ができた。上位意味属性では、階層の深さは2-5まで実験しているが、効果にあまり差が見られなかった。また、意味素性をすべて組合せる(SEM-ALL:上位意味属性はレベル3を利用)と、さらに結果が改良された。最後に、統語素性(SYN-ALL)と意味素性(SEM-ALL)を一緒に利用すると、最もよい結果が得られた。

また、図 5に、SEM-ALLの学習曲線を示す。学習曲線は、まだ上昇を続けているので、例文や京大コーパスなどを訓練データに追加することで、さらに結果を改良できると考えられる。

5 議論

Xiong et al. (2005)も、意味情報を用いてパーズランキングを行なっているが、結果の改良はわずかだった。ここで使われた素性は、述部とその項の間の相関に基づく素性と、述部とその項のHowNetを利用して獲得した上位語に基づく素性だった。しかしながら、彼らは語の直接の上位

[1] Fujita and Bond (2004)により拡張された結合価辞書を利用。

[2] http://tadm.sourceforge.net

素性タイプ	Train-sub (4,984文)		全訓練データ(30,345文)	
	精度(%)	素性数	精度(%)	素性数
SYN-1	52.8	3,502		
SYN-GP	55.2	115,771		
SYN-ALL	55.3	133,140		
SEM-BS	49.3	250,213	57.3	1,188,593
+SEM-WS	49.1	342,767	56.2	1,903,790
+SEM-Class	50.9	491,466	57.5	2,017,766
+SEM-L2	52.8	189,501	60.3	808,320
+SEM-L3	52.7	209,728	59.8	875,809
+SEM-L4	53.0	242,886	59.9	999,680
+SEM-L5	52.2	303,178	60.4	1,239,808
+SP	52.2	166,320	59.1	1,218,111
+SEM-ALL	55.2	512,110	62.7	3,384,318
SYN-SEM	62.1	878,885		
Baseline	20.3	random		

表 3: パーズランキングの結果

語を利用しており、更にレベルを修正して一般化を行なった実験は行なっていない。

Toutanova et al. (2005)とBaldridge and Osborne. (2007)による英語のHPSGのツリーバンクに対するパーズランキングの先駆的な研究では、基本の意味素性(本稿のSEM-BSに対応)を用いているが、統語、意味の情報を融合した場合に、結果があまり改良されていない。

本稿では、主に上位の意味属性に集約することで、効果を得ている。語彙大系のような人手で作成された辞書資源は作成が難しく、多くは存在しないが、大きなサイズのツリーバンクが存在するような言語では、すでにそうした言語資源（WordNetなど)が存在するため、他の言語でも十分利用が可能である。

今後の課題として、人手で付与された語義ではなく、自動的な語義判定によって得られた語義を用いた場合でも、パーズランキングの結果が改良されるかどうか実験したい。その場合、上位意味属性ならば、粗いレベルでの判定ですむので、高い精度で獲得できると予想できる。また、檜コーパスの他の分野（例文、新聞）でも同様の実験を行ないたい。

6 まとめ

本稿では、意味に基づく素性を示し、パーズランキングに対して有効であることを示した。檜コーパスでの訓練とテストにより、統語、意味素性を一緒にしたモデルは、統語素性のみを用いたモデルより、パーズランキングの精度が7.6% (55.3% から62.9%)向上した。

参照文献

Jason Baldridge and Miles Osborne. 2007. Active learning and logarithmic opinion pools for HPSG parse selection. *Natural Language Engineering*, 13(1):1–32.

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2007. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation (Special issue on Asian language technology)*.

Sanae Fujita and Francis Bond. 2004. A method of creating new bilingual valency entries using alternations. In *COLING 2004 Multilingual Linguistic Resources*, pages 41–48. Geneva.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, New York City, USA. URL `http://www.aclweb.org/anthology/N/N06/N06-2015`.

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*. Philadelphia, USA.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *CONLL-2002*. Taipei, Taiwan.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2004. LinGO redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 2(4):575–596.

Carl Pollard and Ivan A. Sag. 1994. *Head Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *41st Annual Meeting of the Association for Computational Linguistics: ACL-2003*.

Melanie Siegel and Emily M. Bender. 2002. Efficient deep processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*. Taipei.

Kristina Toutanova, Christopher D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic hpsg parse disambiguation using the redwoods corpus. *Research on Language and Computation*, 3(1):83–105.

Deyi Xiong, Qun Liu Shuanglong Li and, Shouxun Lin, and Yueliang Qian. 2005. Parsing the Penn Chinese treebank with semantic knowledge. In Robert Dale, Jian Su Kam-Fai Wong and, and Oi Yee Kwong, editors, *Natural Language Processing — IJCNLP 005: Second International Joint Conference Proceedings*, pages 70–81. Springer-Verlag.

池原悟, 宮崎雅弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林 良彦. 1997. 日本語語彙大系. 岩波書店.