確信度つき委員会方式による部分係り受け解析

乾 孝司 *1　乾 健太郎 *1 *2
*1 九州工業大学大学院情報工学研究科
*2 科学技術振興事業団さきがけ研究21
{t_inui,inui}@pluto.ai.kyutech.ac.jp

1 はじめに

近年，大規模なコーパスが利用可能となり，統計的手法を用いた構文係り受け解析に関する多くの研究が報告されている [2, 3, 4, 10, 12, 13, 14, 17]．しかしながら，統計的手法だけに頼る方法には明らかに限界があり，実際に報告されている精度も伸び悩んでいるのが現状である．このような背景から我々は，委員会方式 (committee-based method) を取り込んだ統計的部分係り受け解析 (Probabilistic Partial Parsing) の有効性ついて検証を進めてきた．本稿ではその結果を報告する．

部分解析という考え方は古くから提案されているが (e.g. Jensen ら [9])，我々はこの考え方をさらに進めた統計的部分解析の調査を進めてきた [15, 16]．この中で，n-best の解の重みつき多数決によって見積った確信度は正解率と強い正の相関関係を示すこと，これにより被覆率を犠牲にすれば正解率の向上が見込めることがすでに明かになっており，藤尾ら [14] も同様の結果を報告している．

委員会方式は，あるタスクについて複数の異なるシステムの結果を考慮することで，タスクへの問題解決能力を向上させる手法である．これまでに英文を対象とし，品詞タグづけ [1, 7]，構文解析 [8]，機械翻訳 [6]，音声認識 [5] 等の分野でその有効性が報告されており，統計的部分係り受け解析においてもその効果が期待できる．

本稿の構成は以下の通りである．まず，2節で統計的部分係り受け解析の概要を説明する．つぎに，3節で委員会方式を取り込んだ統計的部分係り受け解析手法について論じ，4節で実験およびその結果を報告する．

2 統計的部分係り受け解析

2.1 係り受け確率

入力文を s，その文節列を $b_1\ b_2\ \ldots\ b_n$ とし，b_i が b_j に係ることを $r(b_i, b_j)$ であらわすとする．s が係り受け関係 $r(b_i, b_j)$ をもつ確信度を $P(r(b_i, b_j)|s) = \frac{P(s, r(b_i, b_j))}{P(s)}$ ($\forall i.\ \sum_j P(r(b_i, b_j)|s) = 1$) と定式化し，この確率を係り受け確率と呼ぶことにする．

個々の係り受け確率は，言語モデルに従って n-best の係り受け構造候補を求めたあと，各係り受け関係を含む候補の確率を出現係り受けごとに足し合わせ，n-best の候補の確率和で正規化することによって推定できる．ボトムアップな言語モデル (e.g.[3]) から推定する場合は，モデルが推定する部分分布 $P(r(b_i, b_j)|s)$ を直接利用することもできる[1]．

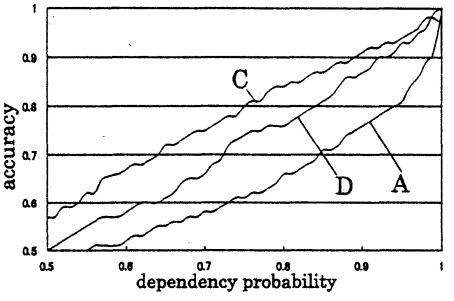

図 1: P-A 曲線

ここで，各 b_i について，係り受け確率を最大にする係り受け関係を $r^*(b_i, b_j)$ ($\Leftrightarrow \arg\max_{b_j} P(r(b_i, b_j)|s)$) とすると，閾値 σ よりも高い係り受け確率 $P(r^*(b_i, b_j)|s)$ をもつ係り受け関係 $r^*(b_i, b_j)$ を選択的に決定することで，部分解析が実現できる．

2.2 P-A 曲線

係り受け確率が $P(r^*(b_i, b_j)|s)$ となる係り受け関係の正解率を $A(P(r^*(b_i, b_j)|s))$ とする．もし，言語モデルが十分に洗練されていて，係り受け確率を正確に推定していれば，$A(P(r^*(b_i, b_j)|s)) \simeq P(r^*(b_i, b_j)|s)$ が成り立つ．さらに理想的には，任意の係り受け確率において $A(P(r^*(b_i, b_j)|s)) = P(r^*(b_i, b_j)|s)$ の関係が成立する．

ここで係り受け確率 (dependency probability) と係り受け正解率 (accuracy) の関係をプロットして得られる曲線を P-A 曲線 と呼ぶ．図1の3本の曲線は実験 (4節) から実際に得た P-A 曲線である．各言語モデルが出力する確信度 (係り受け確率) は，いずれもある程度正確に正解率を予測している．

2.3 C-A 曲線

次に，$r^*(b_i, b_j)$ の中で，閾値 σ 以上の確率をもつ係り受け関係だけを選択的に決定する作業を考える．

σ を変化させた場合の，被覆率 (coverage)[2] と正解率 (accuracy)[3] の関係をプロットして得られる曲線を C-A 曲線

[1] 詳細は文献 [14, 15, 16] に譲る．

[2] 被覆率 = 係り先が決定された文節数 / テストセット中の文節数

[3] 正解率 = 係り先が正解である文節数 / 係り先が決定された文節数

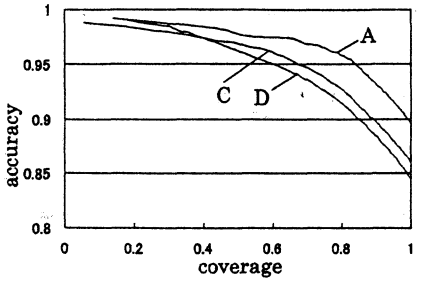

図 2: C-A 曲線

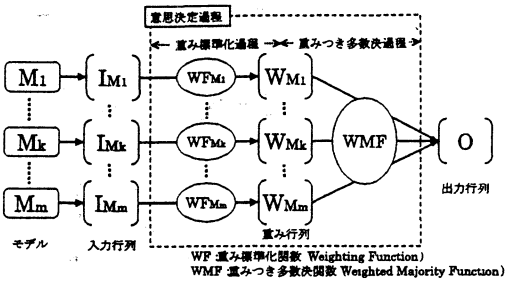

図 3: 委員会方式を取り込んだ部分係り受け解析 (概略)

と呼ぶ.

部分解析では，このC-A 曲線が言語モデルの評価尺度となる．たとえば図2の例では，モデルA，C，Dの順に性能が良いことを示している．

3 委員会方式の導入

本節では，前述の部分係り受け解析に委員会方式を取り込むための枠組を提案する．図3にその概要を示す．

委員会は，各言語モデルから出力される，確信度を要素とする係り受け行列を入力行列として受け取り，重み標準化過程，重みつき多数決過程を経て意思決定をおこなう．そして，最終的に確信度を要素とする係り受け行列 (出力行列) を出力する．入力として係り受け行列を仮定するのは制約として強すぎるように見えるかもしれない．しかしながら，既存の言語モデルを見る限り，多くの場合に2.1節で述べた方法で係り受け確率が推定できることから，この仮定は自然であると考える．

3.1 重み標準化過程

図1のP-A 曲線からわかるように，確信度と実際の正解率の関係はモデルによってばらつきがある．たとえば，図1のモデルA が確信度0.9 と推定した場合と，モデルC が同じく確信度0.9 と推定した場合では，実際の正解率に大きな開きがある．したがって，ある問題に対し両者がともに確信度0.9 と推定した場合にはC を信じることが有利なことがわかる．このように，各モデルが推定する確信度を混合する際には，何らかの補正をおこなうことが望ましい．この補正作業を重みの標準化と呼ぶ．標準化の戦略 (重み標準化関数) には少なくとも次の3つが考えられる．

Simple 入力行列をそのまま重み行列として用いる．

$$w_{ij}^{M_k} = P_{M_k}(r(b_i,b_j)|s) \qquad (1)$$

ただし，M_k は委員会を構成する k 番目のモデル，$P_{M_k}(r(b_i,b_j)|s)$ は M_k の入力行列 I_{M_k} の i行j列 要素となる係り受け確率，$w_{ij}^{M_k}$ は M_k の重み行列 W_{M_k} の i行j列 要素をあらわす．

Normal P-A 曲線が示す関数 A_{M_k} を標準化に利用する．

$$w_{ij}^{M_k} = A_{M_k}(P_{M_k}(r(b_i,b_j)|s)) \qquad (2)$$

ただし，$A_{M_k}(\cdot)$ は，確率 $P_{M_k}(r(b_i,b_j)|s)$ をもつ係り受け関係の正解率をあらわす．また，M_k のP-A 曲線は，共通の訓練データから獲得する[4]．

Class 言語モデルには問題クラスごとにその解決能力にばらつきがあるため，“Normal” で考慮するP-A 曲線では十分に正確な標準化ができない可能性がある．そこで，問題クラスごとのP-A 曲線を標準化に利用する．

$$w_{ij}^{M_k} = A^{C_{b_i}}{}_{M_k}(P_{M_k}(r(b_i,b_j)|s)) \qquad (3)$$

$A^{C_{b_i}}{}_{M_k}(\cdot)$ は訓練データ中に出現する問題クラス C_{b_i} に該当する係り受け関係における係り受け正解率をあらわす．クラス分類には係り文節 b_i の係り属性などを利用する．

3.2 重みつき多数決過程

重み標準化過程より得られる重み行列は，重みつき多数決過程で重みとして利用され，各モデルの解が混合される．今回は，委員会方式で一般的に適用されている2つの重み混合戦略 (重みつき多数決関数) を採用した．

Switching 最大の重みを選択する．

$$o_{ij} = \arg\max_{M_k} \tilde{w}_{ij}^{M_k} \qquad (4)$$

ただし，o_{ij} は出力行列 O の i行j列 要素をあらわす．

Voting 重みつき多数決をおこなう．

$$o_{ij} = \frac{1}{m}\sum_{M_k} w_{ij}^{M_k} \qquad (5)$$

ただし，m は委員会を構成するモデル数である．

このような重み標準化関数と重みつき多数決関数から構成される確信度つき委員会方式は，既存の委員会方式の一般形と見なすことができる．たとえばHalteren ら [7] は，英文の品詞タグづけに委員会方式を適用している．この方式では，各モデル (tagger) について品詞タグの種類ごとの平均適合率・再現率から重みを計算し，“voting” による多数決をおこなう．これは，入力行列のすべての要素を1 あるいは0 とし，重み標準化関数に問題クラスごとの平均正解率を用いたものと見なせる．

このように考えると，既存の委員会方式のバリエーションの多くは重み標準化過程の違いに帰着できる．我々も，

[4]実際には，過学習を避けるために，獲得したP-A 曲線に何らかのスムージングを施して用いる．

表 1: 個々のモデルにおける total/11-point 正解率

model	closed				open			
	total	Simple	Normal	Class	total	Simple	Normal	Class
A	0.8896	0.9537	0.9564	0.9580	0.8964	0.9579	0.9544	0.9537
B	0.8681	0.9307	0.9317	0.9382	0.8661	0.9292	0.9293	0.9315
C	0.8600	0.9291	0.9293	0.9332	0.8605	0.9287	0.9286	0.9288
D	0.8464	0.9188	0.9193	0.9238	0.8455	0.9183	0.9182	0.9173
E	0.8017	0.8779	0.8792	0.8923	0.7982	0.8756	0.8758	0.8822

委員会方式の有効性は重みをどの程度精密に標準化できるかに大きく依存すると考えるが，この仮説は実験を通して経験的に検証する必要がある．

4 実験

4.1 セッティング

参加モデル

以下の 5 つの言語モデル (システム) を用いて委員会を構成した．それぞれの言語モデルは互いに独立に作成されたモデルであり，その特徴は大きく異なる．

- KANA (江原 [13])
- 茶掛 (藤尾ら [14])
- SLUNG+QUADRUPLET (金山ら [10])
- PGLR+LEX (白井ら [17])
- Peach Pie Parser (内元ら [12])

なお今回の実験では，実装上の都合のため，各入力行列の各列要素について値が最大の要素を残し，他のすべての要素の値を 0 と見なした．

以下では，それぞれのモデルを匿名で参照する．

テストセット

実験データには，京大コーパス (ver.2.0) [11] を用いた．まず，京大コーパスの全文 19,956 文を 5 つのモデルで解析し，いずれかが解析に失敗した 3,427 文を取り除いた．

次に，残りの 16,529 文を対象として，細分されている文節をより大きな文節に合わせることでモデル間の文節区切りを統一化した．たとえば，ある単語列 (xyz) に対してモデル α が 1 文節 (b^{α}_{xyz}) と認定し，モデル β が 3 文節 (b^{β}_{x}, b^{β}_{y}, b^{β}_{z}) と認定した場合は，3 文節を統合して 1 文節と見なすことにした (b^{β}_{x}, b^{β}_{y}, $b^{\beta}_{z} \rightarrow b^{\beta}_{xyz}$)．

さらに，統合する前に末尾以外に位置していたどの文節 (b^{β}_{x}, b^{β}_{y}) も末尾文節 (b^{β}_{z}) より後方には係らないという条件を課し，この条件に当てはまらない文節列を含む文は実験データから取り除いた．また，統合した後の文節 (b^{β}_{xyz}) の係り先および確信度は，統合する前に末尾に位置していた文節 (b^{β}_{z}) の情報をそのまま利用した．

この処理によって得られた 14,440 文 (平均 8.0 文節) を実験データとし，5 分割の cross validation をおこなった．分割したデータのうち，4 つのデータブロックを重み標準化戦略 “Normal”, “Class” における訓練データとして用い，残りをテストデータに利用した．また，訓練データによるクローズドテストも実施した．

問題のクラス分類

標準化戦略 “Class” については，今回の実験では試験的に以下に示す 12 の問題クラスを用いた．

1. 主題のハを含む文節．2. ノ格となる文節．3. ガ格となる文節．4. ヲ格となる文節．5. ニ格となる文節．6. デ格となる文節．7. 上記以外の助詞を含む文節．8. 動詞 (連体形) を含む文節．9. 動詞 (連体形以外) を含む文節．10. 副詞からなる文節．11. 形容詞からなる文節．12. 上記以外の文節．

4.2 実験結果

まず，重み標準化の有効性を検証するために，個々のモデルが単独で委員会を構成する場合の性能を調べた．表 1の “Simple”,“Normal”,“Class” は，被覆率 = 0.5, 0.55, $\cdots$, 1.0 の 11 点の平均係り受け正解率 (11-point 正解率) をあらわしている．“total” は被覆率 1.0 での係り受け正解率である．下線を付した数値は，各モデルの 11-point 正解率の最大値であり，複数のモデルで委員会を構成する場合における評価のベースラインとなる．

表から明らかなように，クローズドテストでは，“Class” による標準化をおこなうと一貫して精度が向上した．一方，オープンテストでは，“Normal”, “Class” で標準化したとしても必ずしも精度が向上しなかった．このことから，今回用いた標準化戦略はある程度効果があるものの，まだ工夫の余地があることがわかる．

次に，複数のモデルで委員会を構成し，“Voting” によって混合した場合の結果を表 2に示す．表中の各数値は，11-point 評価における誤り削減率をあらわしている．ただし，誤り削減率は次式によって求めた値であり，最大値は 1.0 で，ベースラインよりも精度が落ちれば負値をとる．

$$(A_c - A_l)/(1 - A_l)$$

A_c は委員会による意思決定の結果から得られた 11-point 正解率，A_l は委員会の構成メンバ中のリーダの 11-point 正解率 (ベースライン) のことである．ここでリーダとは，構成メンバの中で 11-point 正解率が最大値をとるモデルを指す．たとえば，構成メンバが AB の場合は A がリーダとなり，そのときのベースラインは，表 1よりクローズドテストで 0.9580 (Class)，オープンテストで 0.9579 (Simple) となる．

まず，クロードズテストの結果に注目すると，いずれのメンバ構成でも一貫して “Simple”,“Normal”,“Class” の順に精度が向上することがわかる．このことから，重みの標準化の精度が委員会の性能の向上に寄与することが確認された．

次に，委員会のメンバ構成の違いに注目すると，委員会を構成したからといって，必ずしも精度が向上するわけではなく，逆に精度の低下も見られる．この結果は，委員会方式の適用により性能向上を達成した先行研究の報告 (e.g.[8]) とはやや異なるように見える．しかしなが

表 2: 複数のモデルで委員会を構成した場合の誤り削減率 (誤り削減率 >0.2 を太字で表示)

model	closed			open			model	closed			open		
	Simple	Normal	Class	Simple	Normal	Class		Simple	Normal	Class	Simple	Normal	Class
AB	-0.0643	0.0000	0.0619	-0.0143	0.0238	-0.0119	ABC	-0.0167	0.0238	0.0786	0.0166	0.0451	0.0214
AC	-0.0381	0.0000	0.0524	0.0190	0.0451	0.0024	ACD	0.0262	0.0476	0.0857	0.0594	0.0736	0.0356
AD	-0.0524	-0.0262	0.0167	-0.0024	0.0190	-0.0333	ACE	-0.0667	-0.0452	0.0071	-0.0285	-0.0214	-0.0594
AE	-0.2476	-0.1929	-0.1143	-0.1948	-0.1401	-0.1805	ADE	-0.0929	-0.0690	-0.0357	-0.0641	-0.0523	-0.1116
BC	0.1246	0.1456	**0.2087**	**0.2029**	**0.2117**	**0.2234**	BCD	**0.2362**	**0.2476**	**0.2832**	**0.2993**	**0.3051**	**0.2964**
BD	0.1003	0.1294	0.1796	0.1737	0.1839	0.1869	BDE	0.1456	0.1489	0.1828	**0.2131**	**0.2102**	0.1883
BE	-0.0405	-0.0275	0.0502	0.0380	0.0409	0.0642	CDE	**0.2260**	**0.2275**	**0.2560**	**0.2626**	**0.2570**	**0.2416**
CD	0.1931	**0.2081**	**0.2470**	**0.2346**	**0.2402**	**0.2331**	ABCD	0.0429	0.0667	0.1095	0.0618	0.0760	0.0475
CE	0.0269	0.0449	0.1198	0.0744	0.0871	0.1067	BCDE	**0.2460**	**0.2476**	**0.2816**	**0.3109**	**0.3051**	**0.2905**
DE	0.0748	0.0958	0.1457	0.1224	0.1371	0.1297	ABCDE	0.0357	0.0452	0.0833	0.0499	0.0499	0.0095

ら，BCD や CDE のようにメンバの構成によっては委員会方式が有効な場合も少なくない．

また，A が構成メンバに含まれる場合は誤り削減率が低いが，これは A の精度がもともと他に比べて顕著に高いためだと考えられる．BC や CD といった比較的同等の精度をもつメンバが委員会を構成する場合は高い誤り削減率を示していることから，A と同等の精度をもつモデルがメンバに加われば，より効果的に誤りを削減できる可能性がある．

5 つのモデルの中では E がもっとも精度の低いモデルであるが，BCD と BCDE，あるいは CD と CDE を比較すればわかるように，E が構成メンバに加わったとしても削減率が低下するわけではない．これより，性能の低いモデルがメンバに加わっても，必ずしもそのモデルが精度の向上を阻害するわけではないことがわかる．

一方，A と E が共に構成メンバとなる場合に，一貫して顕著に精度が低下した．このことは構成メンバ間の相互作用が委員会としての意思決定に影響を及ぼすことを示唆している．今後，実験データを詳細に検討し，構成メンバ間の影響を調査する必要がある．

最後に，混合法として "Voting" と "Switching" を比較したところ，"Switching" においても精度の向上が見られたが，総じて "Voting" による混合の方がよい精度が得られた．"Switching" では構成メンバ間で意見が割れた場合に誤って意思決定をする可能性が高いのに対し，"Voting" では均衡した重み全てが意思決定に反映される．この違いが両者の精度差にあらわれていると考えられる．

5 おわりに

本稿では，委員会方式を取り込んだ部分係り受け解析手法について，既存の 5 つの確率言語モデルを用いて評価実験をおこない，その有効性を検証した．今後は，結果の詳細分析をおこない，さらに委員会方式の性質を調査していく予定である．

謝辞

委員会方式については，奈良先端科学技術大学院大学の松本裕治氏に示唆に富む多くの助言を頂きました．同氏に深く感謝いたします．また，快く実験に協力してくださった NHK 放送技術研究所の江原暉将氏，奈良先端科学技術大学院大学の藤尾正和氏，東京大学の金山博氏，東京工業大学の白井清昭氏および郵政省通信総合研究所の内元清貴氏の諸氏に感謝いたします．

参考文献

[1] E. Brill, and J. Wu. Classifier Combination for Improved Lexical Disambiguation. In *Proc. of the 17th COLING*, 1998.

[2] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proc. of the AAAI*, pp.598–603, 1997.

[3] M. J. Collins. A new statistical parser based on bigram lexical dependencies. In *Proc. of the 34th ACL*, 1996.

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

[5] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *EuroSpeech 1997 Proc.*.

[6] R. Frederking, and S. Nirenburg. Three heads are better than one. In *Proc. of the 4th Applied NLP*, 1994.

[7] H. van Halteren, J. Zavrel, and W. Daelemans. Improving data driven wordclass tagging by system combination. In *Proc. of the 17th COLING*, 1998.

[8] J. C. Henderson, and E. Brill. Exploiting Diversity in Natural Language Processing: Combining Parsers. In *Proc. of the 1999 Joint SIGDAT Conference on Empirical Methords in Natural Language Processing and Very Large Corpora.*

[9] K. Jensen, G. E. Heidorn, and S. D. Richardson, editors. *natural language processing: The PLNLP Approach.* Kluwer Academic Publishers, 1993.

[10] H. Kanayama, K. Torisawa, Y. Mitsuisi, and J. Tsujii. Statistical Dependency Analysis with an HPSG-based Japanese Grammar. In *Proc. of the NLPRS*, 1999.

[11] S. Kurohashi, and M. Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of NLPRS*, 1997.

[12] K. Uchimoto, S. Sekine, and H. Isahara. Japanese dependency structure analysis based on maximum entopy models. In *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 1999.

[13] 江原暉将. 最大エントロピー法を用いた日本語文節間係り受け整合度の計算. 言語処理学会第 4 会年次大会予稿集, pp.382–385, 1999.

[14] 藤尾正和, 松本裕治. 語の共起確率に基づく係り受け解析とその評価. 情報処理学会論文誌, Vol.40, No.12, pp.4201–4212, 1999.

[15] 乾健太郎, 白井清昭, 田中穂積, 徳永健伸. 統計に基づく部分係り受け解析. 言語処理学会第 4 会年次大会予稿集, pp. 386–389, 1998.

[16] 乾孝司, 木村啓, 乾健太郎. 統計的部分構文解析器のふるまいについて. 言語処理学会第 5 会年次大会「構文解析－現状の分析と今後の展望－」ワークショップ論文集, 1999.

[17] 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, pp.85–106, 1998.