D-7 確率的に一般化された委員会方式による統計的構文解析

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 PREST,JST

自然言語文の統計的部分係り受け解析に委員会方式を導入することにより，既存の解析器の性能を向上させることに成功した．本稿では，従来の委員会方式を確率的に一般化した新しい委員会方式を提案し，日本語コーパスを用いた解析実験の結果を報告する．

1 統計的部分係り受け解析と委員会方式

統計的部分係り受け解析は，確率言語モデルによって推定された確率を利用して各係り受け関係の確信度を見積もり，解析器が『正解である』と確信している部分だけを出力する解析手法である．これにより，被覆率をある程度犠牲にするだけで，係り受け正解率を大幅に向上させることができる [1]．この方式では被覆率 - 正解率の間のトレードオフを利用者が自由に選択することができるので，現状の不完全な解析技術でも，ブートストラッピングや自動要約，テキスト簡単化 [2] といったこれまで以上に幅広い応用に利用することができる．

一方，委員会方式とは，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によって，タスクへの問題解決能力を向上させる意思決定方式であり，これまでに音声認識や機械翻訳などの多くの分野でその有効性が報告されている．しかしながら，部分解析にこれを適用するためには，いくつかの拡張が必要となる．

2 委員会方式の確率的一般化

図 1に新しい委員会方式の概要を示す [3]．委員会の入出力はともに係り受け確率行列[1]である．従来の委員会方式に対する拡張点は次の 3 点である．

1) 各委員は，各文節についてその係り先に重みつきの票を投じる．票の重みは対応する係り先に対する確信度で与える**確率的投票**.

2) 重み標準化関数 (WF) を導入し，委員間での票の重みの信頼性を標準化する**重み標準化**.

3) 各委員は，2 位以下の係り先候補にも重みつきの票を投じる**多重投票**．この拡張は付加的ではあるが，精度向上への貢献が期待できるため採用した．

3 解析実験

提案した枠組みの有効性を検証するため，5 つの代表的な統計的日本語係り受け解析器 (例えば [4]) [2]の各組合せを構成委員とする計 26 組の委員会を作成して係り受け解析実験をおこない，解析精度の変化を調査した．

実験データには京大コーパス (ver2.0) 中のテキスト約 14,000 文を用い[3]，5 分割の交差検定をおこなった．WF には 3 つのオプションを，CF には単純な重みつき多数決をおこなうオプションのみを用意した．解析精度は，被覆率区間 [0.5,1.0] の 11 点の平均係り受け正解率 (11 点正解率) および総係り受け正解率で評価した．

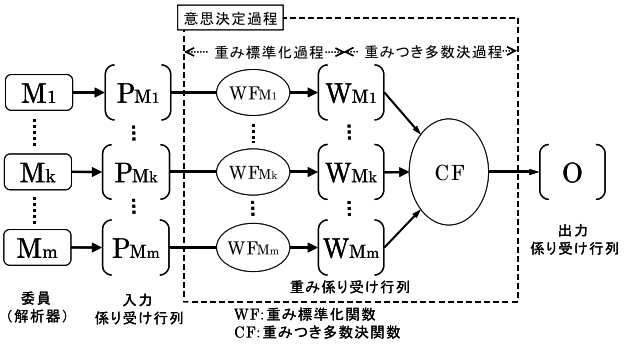

図 1: 確率的に一般化された委員会方式

実験より，提案した枠組みの有効性を支持する結果を得た．**重み標準化**については，いずれの委員会においても一貫して，標準化しない場合に比べ 11 点正解率で 0.3 ポイント程度精度が向上した．**多重投票**については，特に構成委員数が少ない場合に効果が高く，多重投票を行なわない場合に比べ 11 点正解率で，最大 47% の誤り削減率が得られた．

総合的な結果としては，ほぼすべての委員会で精度の向上が確認できたが，顕著に精度が上がる委員会もあれば (11 点正解率で最大 31% の誤り削減率)，ほとんど精度が変化しない委員会もあり，効果にはばらつきがあった．しかし現状では，このばらつきの原因が明らかでなく，詳細な調査をおこなっていく必要がある．

最後に，既存の係り受け解析器の中で最もよい性能をもつ解析器 KNP(総係り受け正解率で 91.25%)[5] を委員として加えた追実験をおこなった．KNP は規則ベースのノンパラメトリックな解析器なので，総係り受け正解率を票の重みと見なした．KNP を含む委員会は総係り受け正解率で 92.00% を示し，本実験で用いたベンチマークによる評価に関する限り，既存のどの係り受け解析器よりも高い精度が得られた．

今後は，委員会の効果のばらつきの要因を詳細に分析するとともに，応用システムへの適用による実用性の検証をおこなう予定である．

参 考 文 献

[1] 乾, 木村, 乾. 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 WS 論文集, 1999.

[2] 乾, 山本, 野上, 藤田, 乾. 文章読解支援における構文的言い換えの効果について. 信学会 (WIT-99-2),1999.

[3] Inui, T. and K. Inui. Committee-based Decision Making in Probabilistic Partial Parsing. *The 18th COLING*, 2000.

[4] 内元, 関根, 井佐原. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, Val.40, No.9, 1999.

[5] 黒橋. 日本語構文解析システム KNP 使用説明書 version 2.0b6. 京都大学大学院情報学研究科, 1998.

[1] 文 s 中のある文節 b_i が文節 b_j に係ることを $r(b_i, b_j)$，その確率を $P(r(b_i, b_j)|s)$ としたときに，確率 $P(r(b_i, b_j)|s)$ を ij 要素にもつ行列のこと (本稿では確率 $P(r(b_i, b_j)|s)$ を確信度と呼ぶ).

[2] 各解析器はそれぞれ独立に作成されており，それらの特徴は大きく異なる．実験データによる各解析器の単独での精度は，総係り受け正解率で 80% ～ 89%，11 点正解率で 85% ～ 96% であった．

[3] ただし，コーパスとは異なる文節区切りを採用している．各文節がすべて隣に係るとしたときの正解率は 57.4% であった．