# Web 上のテキストデータを用いた非タスク指向型対話システムのための係り受け解析による連想メカニズムの構築

若原 基　Rafał Rzepka　荒木 健治
北海道大学大学院情報科学研究科

## 1. はじめに

近年，非タスク指向型対話システムを用いた様々なソフトウェアが実装され，一定の評価を得ている．本研究は，非タスク指向型対話システムが発話を行う際に用いる連想メカニズムの構築を目的としている．本稿では，2 章でテキストデータから係り受け解析を用いて情報を抽出する手法を，3 章で連想メカニズムを構築する指針を述べる．

## 2. Web 上のテキストデータからの係り受け解析を用いた情報抽出

本研究では，対話システムにおける連想メカニズムを構築するために，Web 上のテキストデータをコーパスとし，得られた日本語文に係り受け解析を適用して情報抽出を行っている．係り受け解析には CaboCha[1] を用いている．解析結果から，名詞・形容詞・動詞・付属語を単位として，述語・短文・係り受け関係・係り受け関係の出現頻度を得る．得られる情報の定義を以下に示す．

- **述語**：形容詞・動詞・「名詞＋助動詞」の組
- **短文**：「名詞＋述語」の組
- **係り受け関係**：係り受け関係にある 2 つの短文の組
- **係り受け関係の出現頻度**：係り受け関係の異なり数

以上の情報をデータベースに格納する．

## 3. 連想メカニズムの構築

まず，入力された発話文から抜き出した名詞をクエリとしてデータベースを検索し，その名詞を含む短文との係り受け関係を有する短文のうち，出現確率が閾値以上であるものを抽出する．これにより，対話システムが発話を行う際，入力された発話文と関連性の高い情報を選択して使用することができる．

Web から関連語を抽出する手法としては，Higuchi ら [2] のものが挙げられる．Higuchi らの手法では，本来「名詞＋述語」の関係にない短文も出力されるが，本手法ではそのような出力を比較的少なく抑えることができると考えられる．

## 4. おわりに

本稿では，Web 上のテキストデータから係り受け解析を用いて入力語についての関連情報を抽出し，その情報から連想メカニズムを構築する手法について述べた．今後，対話システムに連想メカニズムを組み込んで実験を行い，抽出されたデータの特徴を詳しく分析する．

## 参考文献

[1] 工藤拓，松本裕治：“チャンキングの段階適用による係り受け解析”，情報処理学会論文誌 Vol.43, No.6, pp.1834-1842 (2002)

[2] Shinsuke Higuchi, Rafał Rzepka and Kenji Araki: “A Casual Conversation System Using Modality and Word Associations Retrieved from the Web”, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp.382-390 (2008)