NINJAL シンポジウム「データに基づく日本語研究」　2018.12.15–16　東京証券会館
ワークショップ「コーパスとモダリティ」

モダリティアノテーションとその統計分析

浅原正幸（国立国語研究所）

本発表では松吉 (2013) が整備した『現代日本語書き言葉均衡コーパス』 (以下 BCCWJ) に対するモダリティアノテーション BCCWJ-EME の特徴的な語彙素 n-gram を検討する。BCCWJ-EME は、自然言語処理の情報抽出タスクのために、 BCCWJ のコアデータに出現する情報発信者の主観的な態度を事象（行為・出来事・状態）単位に付与したものである。文が表出する事象の認識や事実性解析を目的とする。現在までに Yahoo! 知恵袋・白書・新聞・書籍サンプル中の約 40,000 件の事象に対して、態度表明者の情報、相対時、事象の事実性・仮想性、表現類型のモダリティ（態度)、真偽判断のモダリティ（事態の成立・不成立・成立確率)、価値判断のモダリティ（事態成立の望ましさ）の情報が付与されている。このデータに対して、前後 5 語彙素まで n-gram を特徴量として用いた決定株を学習器としたブースティング (bact-0.13) を用いて特定のモダリティに特徴的な語彙素 n-gram を抽出した。本発表ではアノテーションの実例とともに抽出された語彙素 n-gram について紹介する。