自然言語処理における最適化

高村大也　　東京工業大学
takamura@pi.titech.ac.jp

本講演では，自然言語処理において最適化の理論や手法がどのように利用されているか，その特徴や問題点は何かについて，最先端の実例とともに説明する．

自然言語処理における様々な研究課題は，大きく二つに分けることができる．一つは，言語解析(言語理解とも言われる)であり，もう一つは言語生成である．前者の言語解析は，人間が生成した言語表現の構造や意味などを推定する課題である．形態素解析や構文解析などの伝統的な研究課題から，意見抽出などのような比較的新しい課題まで様々な研究課題がここに含まれる．多くの言語解析課題は単純な分類問題として定式化できるが，複数の分類問題を同時に考慮することで性能向上が期待できることがしばしばある．このようなケースについて最適化という観点から考察をする．学習時にも複数の問題を考慮する手法と，推論時のみに複数の問題を考慮する手法があり[1]，それぞれについて説明を行う．

一方，後者の言語生成は，機械に言語表現を生成させる課題である．言語生成においては，可能な出力の種類数が膨大であり，最適化が重要な役割を果たすことが多い．言語生成を含む研究課題としては，対話生成，機械翻訳，文書要約などがあり，ここでは主に文書要約について説明する．特に，複数の文書から要約を生成する複数文書要約課題のためのモデルとして，最大被覆モデル，施設配置モデルなどを紹介する[2]．これらの二つのモデルは入力文書集合から文をそのまま抽出して要約を生成するアプローチにもとづいているが，それぞれの文を圧縮しつつ文を選択する要約モデルも紹介する[3]．また，要約モデルで用いられる目的関数の多くは劣モジュラ性を持つ．本講演では，自然言語処理で提案されている手法について，劣モジュラ性という観点からも整理して紹介する[4]．

[1] Dan Roth and Wen-tau Yih, “A linear programming formulation for global inference in natural language tasks”, CoNLL, 2004.

[2] Hiroya Takamura and Manabu Okumura, “Text summarization model based on maximum coverage problem and its variant”, EACL, 2009.

[3] Hajime Morita, Ryohei Sasano, Hiroya Takamura and Manabu Okumura, “Subtree extractive summarization via submodular maximization”, ACL, 2013.

[4] Hui Lin and Jeff Bilmes, “A class of submodular functions for document summarization”, ACL, 2011.