JSLS 2024 Handbook

The 25th Annual International Conference of the Japanese Society for Language Sciences

言語科学会第 25 回年次国際大会

July 12-14, 2024

Azarea, Shizuoka Prefectural Gender Equality Center

静岡県男女共同参画センター「あざれあ」

英語における形容詞の細密な品詞分類法

田村 一馬（会津大学）・ブレイクジョン（会津大学）

Finer-grained part-of-speech classification of adjectives

Kazuma TAMURA (University of Aizu), John BLAKE (University of Aizu)

1. はじめに

本研究では，細分化した英語の形容詞分類体系を用い，単語トークンを自動的にラベル付けする新しい POS タガーを提案する．POS（品詞）は文中の単語の文法的役割を示すカテゴリーであり，タガーはこれらのタグセットに基づいて単語のラベル付けを行う．しかし，既存のタグセットは，単語の意味論的，形態論的，および構文的特性を考慮した場合に，まだ細分化の余地がある．計量文献学および法廷著者分析においては，単語トークンと POS タグの特有なパターンが著者の特定の手掛かりとなることがある．例えば，創作文では物語の描写を生き生きとさせ，読者を引き込むために様々な形容詞が用いられる．このため，詳細な分類体系を用いることで，著者の推定やその説明の精度を向上させる可能性がある．

2. 先行研究

英語の主要な単語クラスには多くのタガーやタグセットが存在するが，本研究では主要な 8 つのクラスに焦点を当てる．標準的な POS タグセットである Penn Treebank（Santori, 1991; Taylor et al., 2003）には 33 種類の文法タグが含まれているが，形容詞は普通形（JJ），比較級（JJR），最上級（JJS）の 3 つにしか分類されていない．これは例えば，“An empty stream, a great silence, an impenetrable forest. The air was warm, thick, heavy, sluggish”（Conrad, 2007, pp.104-5）という文において，すべての形容詞が JJ として分類され，文中の特徴的な形容表現を十分に捉えることができないことを意味する．Atwell (2008)は，文法上の区別をより細かく捉えるために，より詳細なタグの必要性を提起している．本研究では，このような要望に応え，文法をより深く分析する能力を研究者に提供することを目指す．

3. 方法

既存の文献を検討した結果，形容詞には極性，述語，後置，分詞などを含む 28 の潜在的な下位分類があることが判明した．さらに，各下位分類に対してはルールベースおよび確率的な識別手法が既に確立されている．本研究で使用した下位分類は，Natural Language Toolkit（Bird et al., 2009）内の nltk.tag パッケージを使用して検証された．最も高い品詞注釈基準を満たしたアルゴリズムは NLTK タガーに組み入れられた．また，文中におけるこれらの形容詞の使用傾向と著者推定の関係を捉えるために，説明可能な AI モジュールである SHAP（Lundberg et al, 2017）を使用した．

4. 結果

今回開発した細分化された形容詞タガーとタグセットには，現在分詞形容詞，過去分詞形容詞，限定形容詞，外置形容詞を加えた計 7 つのサブカテゴリが含まれている．これらのタガーとタグセットは，Gutenberg corpus（Lebert, 2008）を使用して検証され，その結果，形容詞の特異な使用法が著者特定とその説明に寄与することが示された．図 1 は，SHAP を用いて著者の分類と各特徴量の寄与の関係をプロットした結果である．

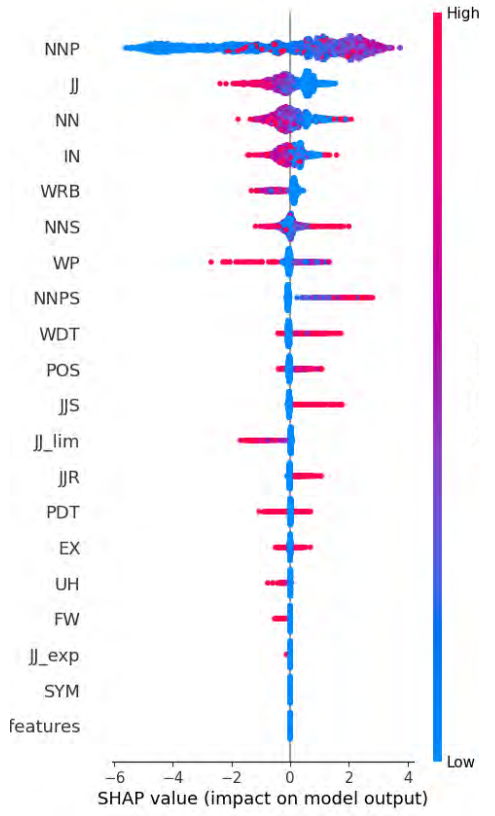

図 1.細分化された POS タグセット使用時の SHAP beeswarm plot 出力結果

5. 考察

本研究では，NLTK タガーと細分化された形容詞タグを用いて品詞注釈を行うことができた．新たに追加した下位分類においては，誤検出を最小限に抑えつつ識別能力を最大化することが重要な基準であった．しかし，NLTK タガーにはいくつかの制限が存在する．例えば，単一の単語トークンに複数のタグを割り当てることが困難である点や，単語間の依存関係を明示することが難しい点が挙げられる．これらの制約は，一部の形容詞サブカテゴリの分類において障壁となる可能性がある．今後の研究では，spaCy や TreeTagger などの代替タガーも用いて，さらに詳細な検証を行う計画である．

参考文献

Atwell, E. S. (2008). Development of tag sets for part-of-speech tagging.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Conrad, J. (2007). *Heart of Darkness*. Penguin.

Lebert, M. (2008). *Project Gutenberg (1971-2008).* Project Gutenberg.

Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank Project*. https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: an overview. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora (pp.5-22).* Springer Science.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems*. 30.