ニュース記事の地理的分類法：辞書拡張法を用いた時系列テキストデータの国別分類システム

Dictionary Expansion Technique
for Geographical Classification of Very Short Longitudinal Texts

H26助人20

代表研究者　渡　辺　耕　平　ロンドン政治経済学院　方法論学部　博士課程
Kohei Watanabe　*Doctoral Candidate, Department of Methodology, London School of Economics and Political Science*

In this project, a new geographical classifier was created to overcome shortcomings of the widely-used simple keyword matching approach: simple keyword matching can achieve high precision in classification, but its recall tends to be very low due to the small size its vocabulary. Classification by the new system is performed by automatically constructing a large dictionary using a type of lexicon expansion technique based on a pre-defined dictionary of place names. Owing to the larger vocabulary of expanded dictionary, the system is able to classify subunits of documents (paragraphs and sentences) precisely without compromising recall. The lack of human supervision also allows the dictionary to be updated frequently to adapt to temporal changes in text content.

研究目的

本研究の目的は、社会科学における大規模な国際ニュース研究を可能にするような、従来のキーワード一致に基づいた手法よりも精度が高い、文書の国別の分類法を開発すること、および、その精度の厳密な測定である。

概　　要

本研究では、従来から広く用いられてきた手法よりも大幅に高い精度を実現するニュースの地理的な分類法が開発された。これまで使われてきたキーワード一致による分類は、高い適合率（precision）を発揮する一方で、再現率（recall）が犠牲になっている[1]。この原因は、キーワードとして用いられている地名および国名が手作業で選択されており、自ずとその数が限られているからである。本研究で開発された手法は、これらの地理的な語彙を辞書拡張法によって、自動的に増大させることであり、大幅に増加した語彙によって、文節や段落などの非常に短い単位も、高い再現率を維持しながら正確に分類でき、なおかつ、人が関与しないことによる高い頻度での辞書の更新によって、データの時間的な変化にも柔軟に対応できるようになった。

－以下割愛－