人文学DXを指向する情報基盤の構築

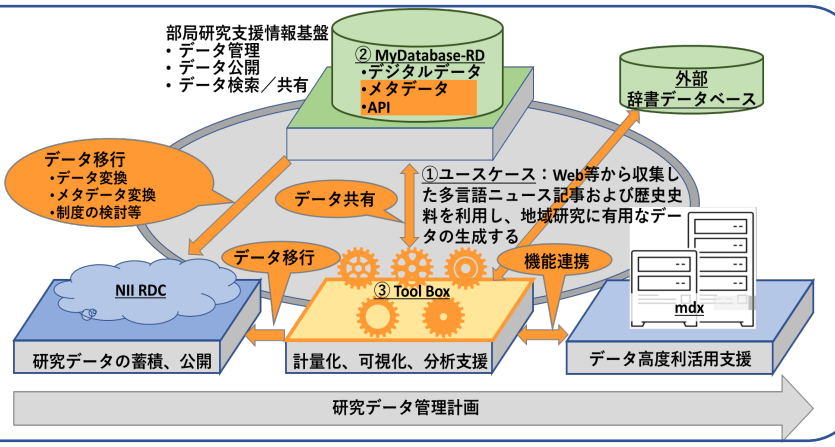
Construction of an Information Infrastructure for the Humanities DX

原 正一郎[*1]、馬場 弘樹[*2]、森 信介[*1]、村上 勇介[*1]、関野 樹[*3]、山田 太造[*4]、後藤 真[*5]

Shoichiro Hara[*1], Hiroki Baba[*2], Shinsuke Mori[*1], Yusuke Murakami[*1], Tatsuki Sekino[*3], Taizo Yamada[*4], Makoto Goto[*5]

1.京都大学(Kyoto University) 2.中央大学(Chuo University) 3.国際日本文化研究センター(International Research Center for Japanese Studies)
4.東京大学(University of Tokyo) 5.国立歴史民俗博物館(National Museum of Japanese History)

課題の概要：人文学DXの実現には人文学データの計量的処理が不可欠である。本課題ではテキストデータの計量的処理をユースケースとして、①Webニュースデータから地域変動を検出する計量的手法の開発、②データ分析を支援するツールボックスの構築、③関連する大量・多様な人文学研究データの蓄積・共有・利活用を支援するデータベースシステムの構築、④研究データを適切に管理する手法の確立を試みる。この問いに挑みながら、定性的手法が主流である地域研究において、情報学の計量性を活かした地域研究手法の構築を目的とする（本発表は①と②について）。



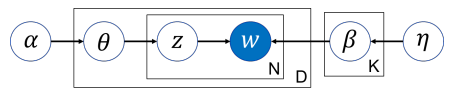
データ分析① Webニュースデータの分類と可視化：大量のWebニュースデータを内容（話題：Topic）に応じて自動的に分類する。本研究ではLatent Dirichlet Allocation（LDA）によるtopic modelを用いた。毎日新聞CD-ROM版（2010-01-01～2019-12-31の全国版および46地方版、約200万記事、約9億文字）を対象として、東日本大震災に関連したトピックの抽出を試みた（各トピック数に対するPerplexityとCoherenceおよび各トピックを構成する語彙群の定性的評価より、以下ではトピック数を184とした）。

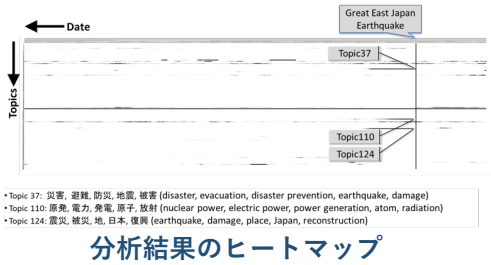
トピックモデル: 教師なし学習手法

$p(\beta,\theta,z,w) = p(\beta)p(\theta)p(z|\theta)p(w|\beta,z)$ から

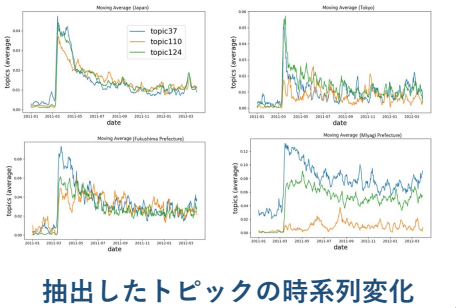
$p(\beta,\theta,z|w) = \frac{p(\beta,\theta,z,w)}{p(w)}$ を求める問題

計算コスト（時間）が高い(mdxに期待)



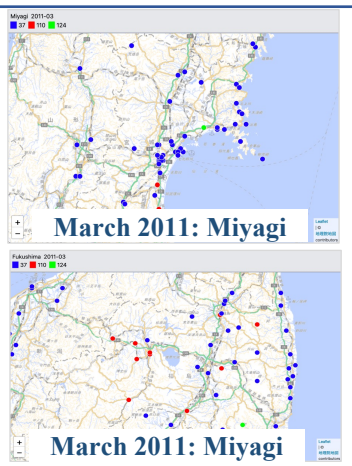


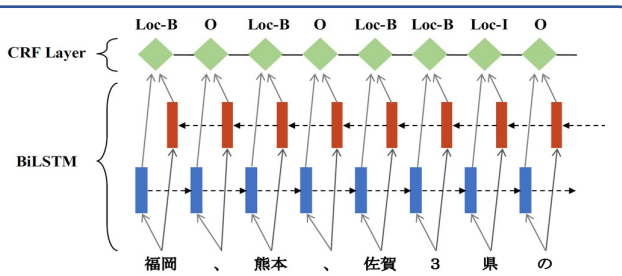
分析結果のヒートマップ



抽出したトピックの時系列変化

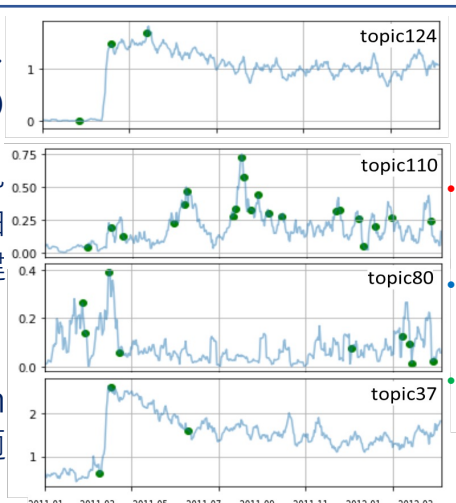
データ分析② Webニュースデータからの地名抽出と緯度・経度推定：抽出したトピックを空間分布として可視化するため、①地名に関わる語彙を自然言語処理技術を使って自動抽出した。②抽出した語彙で地名辞書を検索し、緯度・経度を取得した。③抽出した地名には同名異地名などの曖昧性が存在する。そこで「同一記事中に現れる地名は相互に近傍に存在する」という前提のもとで，各地名間の総距離が最小になるように地名を選択した。





① BiLSTM-CRF：固有表現抽出。学習データから地名抽出法を学習する（上図）。
② 地名辞書を検索し、地名に対応する緯度・経度の候補集合を取得する。
③ 単純な戦略を用いて緯度・経度を一つに絞り込む。

データ分析③ Webニュースデータによる地域変容抽出の試み：地域社会に何らかの変容が見られる瞬間（Tipping Point）を自動抽出する。ここでは、震災後の保健・健康活動状況の変化を事例として、フィールド調査や文献調査の結果と、トピックの時系列データにBayesian Online Change Point Detection を適用した変化点抽出の結果を比較した。



抽出したトピックの変化点：

- **第1ティッピング・ポイント**（1～2週間後）：頻度の高い単語が日々変動しており解釈が困難。
- **第2ティッピング・ポイント**（3週間後）：救命から健康管理へと活動内容が変化している。
- **第3ティッピング・ポイント**（約3ヵ月後）：家屋の仮復旧、仮設住宅の設置、ボランティアによる瓦礫撤去作業が活発化など。

先行研究：

- **フェーズ0**：初動体制を確立するための期間。災害発生後24時間以内。
- **フェーズ1**：人命と安全を確保するための緊急対策の期間。72時間以内
- **フェーズ2**：避難所での生活安定のため一時的対策期間。4日～1ヶ月以内。
- **フェーズ3**：避難所から仮設住宅への移動などのための応急対策期間。1ヶ月以降。
- **フェーズ4**：仮設住宅対策や新たなコミュニティづくりを含むコミュニティ再建のための復旧・復興期間。2ヶ月以降。

抽出したトピック変化点と先行研究との比較

ツールボックスの開発：テキスト分析に必要な機能をツールボックスとして研究者の利用に供する。大量テキストデータに対する学習・予測等を実運用に耐えられる時間内で実行するためmdxを利用した。

- テキストの語彙分割、時空間語彙認識および計量化（緯度・経度および時間の推定）、主題分析用LDA（Latent Dirichlet Allocation）、基本的な分析結果の可視化に関するツール群を実装した（図）。
- 時系列データ分析や異常点検出等のアルゴリズムの充実を図る。

ツールボックスの利用例