ノート　**Note**

Wikipedia を用いた授業動画における学問推定

伊藤雪乃 [1]　工藤佑一郎 [1]　内堀雄真 [1]
一島力男 [2]　高橋幸雄 [1]

Automatic Academic Areas Classification for University Lecture Movie

Yukino ITOH[1]　Yuichiro KUDO[1]　Yuma UCHIBORI[1]
Rikio ICHISHIMA[2]　Yukio TAKAHASHI[1]

キーワード: 学問分野推定，授業動画，Wikipedia
Keywords: Academic Area Classification, Lecture Movie, Wikipedia

1. はじめに

2020 年新型コロナウィルス感染症 (COVID-19) の流行により，多くの大学がオンラインでの遠隔授業を行った．遠隔授業の普及により，自分の好きなタイミングで学び，時間を有効活用できるようになった．しかし，数多くの講義の中から自分の学びたい内容の授業をシラバスのみで判断するのは困難であると考えられる．本研究は，オンライン授業等で利用された授業動画から学生が必要な情報に効率的にアクセスすることを支援する講義動画アーカイブシステムの評価を行なう．本システムは講義動画の情報を学問分野毎に整理して利用することで，学生が所望する内容へアクセスを支援することを目的としている．

2. 関連研究

ブログ記事集合から抽出した Wikipedia エントリタイトルに対して，分野に対応するトピックモデルを推定し，その特性を分析する研究がある [1][2]．近年，ブログサービスやブログツールの普及により，さまざまな情報がブログに記載され，それらの情報を検索サービスから取得できるようになった．しかし，特定のトピックについて検索した場合でも検索結果には様々な観点が混在しており，検索結果を単なるリストとして提示するだけでは，検索結果にどのような観点が含まれているか知ることができない．これら研究では，Wikipedia を知識源とする分野トピックモデルを，ブログ記事集合から推定した通常のトピックモデルと比較して，両者の特性の違いを分析し，ブログ記事集合中の話題の広がりを俯瞰する目的において両者が相補的な関係にあることを示している．

3. 提案手法

本研究は Wikipedia から取得した上位語下位語を利用し授業動画の科目の分野推定を行った．検索用語における上位語と下位語の例を図 1 に示す．

上位下位関係となる用語ペアの抽出処理は Wikipedia のページから以下の 3 種類を情報源として上位下位関係候補を抽出し，各候補が上位下位関係であるか否かを統計的に判定している．

(1) hierarchy: 箇条書きなどの階層構造から上位下位関係の候補を抽出する．

(2) definition: 最初の文 (定義文) から上位下位関係の候補を抽出する (「〜とは，….」などを利用)

(3) category: category tag にある単語から上位下位関係の候補を抽出する．

[1]国士舘大学理工学部
[2]国士舘大学体育学部

上位語	検索用語	下位語
作品	SF 作品	E.T
楽器	ギター	エレクトリックギター
都市	日本の都市	東京
食材	カレーの食材	にんじん
製品	企業の製品	iPhone
食品	チョコレート	ミルクチョコレート

図 1: 上位下位関係の語の例

4. 評価実験

4.1. 実験データ

実験対象の授業動画として Microsoft365 Stream (図 2) に 2021 年 6 月 15 日時点で登録された 6123 個の授業動画 (タイトル，投稿者，投稿日，説明文，再生回数，コメント回数) に関する情報を取得した．また各授業動画に字幕情報が付与されている場合にはその字幕データもテキストデータで取得した．

図 2: Microsoft365 Stream

各授業動画の投稿された日付の分布を図 3 に示す．

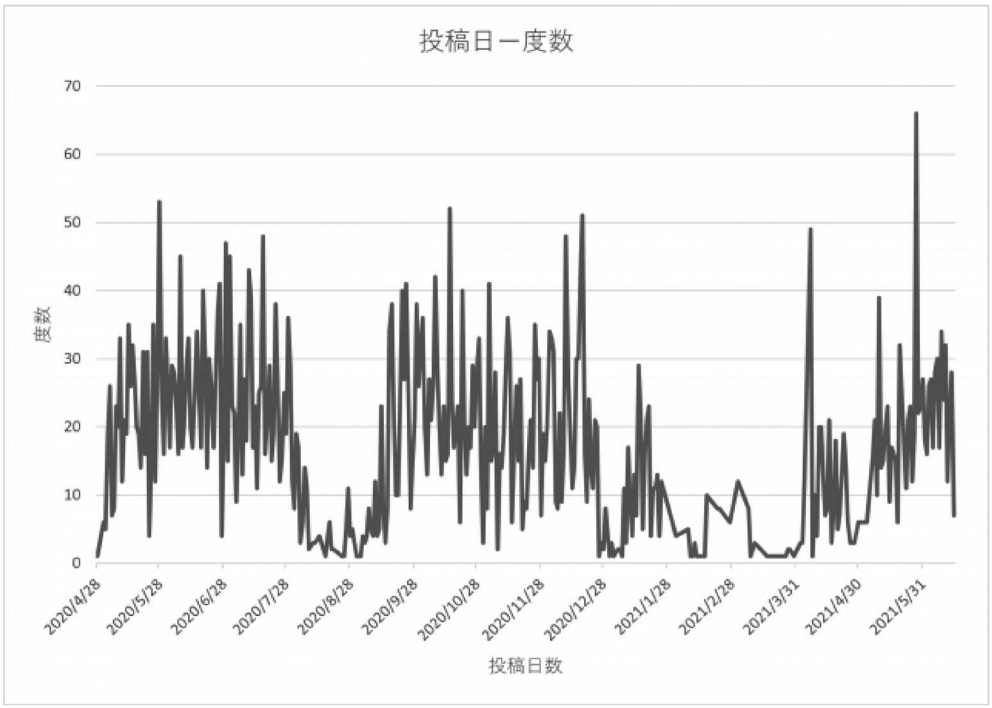

図 3: 授業動画の投稿日分布

4.2. 実験手順

収集した授業動画の中から 7 学部の科目と総合教育科目ごとに 5 つずつ，計 40 個の授業動画を抜粋した．授業タイトルから推定した分野とプログラムで字幕データから推定した分野を比較し，一致していれば 1 点，一致していなければ 0 点とし各学部 5 点を満点，全体で 40 点を満点として正答率を求めた．分野推定で用いられる学問分野については，日本学術振興会が定めている科学研究費助成事業審査区分表の小区分 [3] を学問分野の一覧とする．

4.3. 実験結果

実験結果を表 1 に示す．政経学部，法学部，総合科目の正答率は高く，80%であった．これは授業内で授業タイトルに関連した言葉を多く使っていたため精度が高くなったと考えられる．一方で体育学部と 21 世紀アジア学部の正答率が低く，40%となった．体育学部に関しては，プログラムで推定した分野の中に授業には関係のない寄生虫学や細菌学等の結果が現れた．これは授業内で怪我への配慮や，治療の方法の言葉に反応したと考えられる．全体での正答率は 65.0%であった．

5. おわりに

本研究は学生が所望する内容へアクセスを支援することを目的として評価実験を行った．学部により精度には差があることがわかったが全体では良好な結果が得られた．この結果の懸念点として，精度が低い学部科目に関しては，授業内で授業タイトルに関連していない言葉が多く使われていたためであるのではないかと考えられる．今後の課題として上位語下位語以外の情報を用いること，分野推定の結果を用いて授業動画アーカイブを作成することが挙げられる．

表 1: 実験結果

学部	スコア	正答率
政経学部	4	80.0%
体育学部	2	40.0%
理工学部	3	60.0%
法学部	4	80.0%
文学部	3	40.0%
21 世紀アジア学部	2	40.0%
経営学部	3	60.0%
総合教育科目	4	80.0%
全体	25	65.0%

参考文献

[1] 牧田 健作，鈴木 浩子， 小池 大地， 宇津呂 武仁， 河田 容英: Wikipedia を知識源とする分野トピックモデルの推定と分析，情報処理学会研究報告，Vol.2012-DBS-155, No.11, pp.1-11, 2012.

[2] Jong-Hoon Oh, Ichiro Yamada, Kentaro Torisawa and Stijn De Saeger, ”Co-STAR: A Co-training Style Algorithm for Hyponymy Relation Acquisition from Structured and Unstructured Text,” In Proceedings of COLING-2010, pp.842-850, 2010.

[3] 科学研究費助成事業, 審査区分表, 小区分一覧: 日本学術振興会, https://www.jsps.go.jp/j-grantsinaid/02_koubo/shinsakubun.html (2021-12-12 参照)