4ZC-01

シラバス中の組織間での単語意味揺らぎの分析

瀬端 賢人 †　　中島 克也 ‡　　小林 亜樹 †

† 工学院大学工学部情報通信工学科　　‡ 工学院大学大学院工学研究科 電気・電子工学専攻

1 はじめに

オープンエデュケーションの進展でシラバス等メタデータの重要性も増しているが，同一単語でも文書中での意味や用法が異なる場合があり，著者，文書間での違いの程度を知る必要がある．

本研究では，シラバスにおける異文書間に出現する同一単語の機械的な意味の差を分析する．

具体的には，単語を前後の出現語の構成によりベクトル表現とする手法を用いて，異なる大学のシラバス中で共通に用いられる単語を対象として，それぞれベクトルの異なり方を分析する．

単語，文などの分析単位毎に，組織，分野の違いなどがベクトルの異なり方に与える影響や，ベクトルの差異の意味や概念などについて議論し，メタデータ活用時の課題について考察する．

2 分析手順

2.1 準備

分析するシラバスの提供元大学をA Bとする．各大学のWebサイトなどから，シラバスデータを取得する．

各大学毎の分析対象テキストをそれぞれ形態素解析器により単語に分割し，単語に半角で空白区切りを入れる．この結果を利用して，元テキストを単語単位に区分された単語列に変換する．これを分析対象テキストと呼び，T^A, T^B とする．また，両者を連結したテキストを

$$T^C = T^A \mid T^B \tag{1}$$

とする

2.2 分析対象語の抽出

両大学それぞれのテキストデータに用いられている単語の名詞の中で使用回数が多い順にソートする．ソートされた二つの単語群から共通する単語 $X_k(k = 1 \sim N)$ を N 個選定する．

2.3 各語の分析

分析対象語 X_k それぞれについて，A B各大学内テキストにおける語ベクトルを求める．このとき，X_k は，X_k^A, X_k^B のような置換処理を行った各テキストを連結し，処理テキスト T_k^C を得る．

$$T_k^C = T^A|_{\text{replace}}(X_k \rightarrow X_k^A) \mid T^B|_{\text{replace}}(X_k \rightarrow X_k^B) \tag{2}$$

得られた分析対象語ベクトルをそれぞれ $V(X_k^A)$, $V(X_k^B)$ とする．その後，両者の差分 $D(A-B) = V(X_k^A) - V(X_k^B)$, $D(B-A) = V(X_k^B) - V(X_k^A)$ を計算する．この差分ベクトルは，同一文字列である分析対象語の大学間での使用状況の差異を示す可能性がある．そこで，それぞれについて，T^C 内の各語のベクトルとの類似度上位を求める．最後に，類似度上位のうち，低頻出語を除外した類似語リストを得る．

3 実験

3.1 対象テキスト

大学Aとして，工学院大学，Bとして芝浦工業大学の公開Webシラバステキストを処理対象とした．工学院大学の2014年度シラバス * から1, 2, 3, 4学年の工学部(1部)，建築学部，情報学部，グローバルエンジニアリング部の各学科(表2)の「授業のねらい」「受講にあたっての前提条件」「授業計画及び準備学習」「具体的な到達目標」からテキストをデータを取得した．

芝浦工業大学の2015年度シラバス † からも同様に1, 2, 3, 4学年の工学部，システム理工学部，デザイン工学部の各学科(表3)の「授業の概要」「達成目的」「授業計画」「履修登録前の準備」からテキストデータを取得した．

テキストを単語単位に分割する．

各大学毎の分析対象テキストをそれぞれ形態素解析器により単語に分割する

ベクトル化する際に単語単体として意味を持たないと考えられる英数字，記号を取り除く事前処理を行った．

事前処理を行った後の T^A の総単語数は2699832個

T^B の総単語数は3666276個である．

3.2 分析対象語

大学A Bそれぞれの頻出上位単語を基に大学A Bに共通する単語から上位 $N = 100$ を分析対象として選定する候補語とする．

Analysis of word meanings on the difference between authors in syllabuses.

†Kento Sebata　‡Katsuya Nakajima　†Aki Kobayashi

†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

‡Electrical Engineering and Electronics, Kogakuin University Graduate School

*http://syllabus.sc.kogakuin.ac.jp/syllabus/daigaku/2014/1bu.html

†http://syllabus.sic.shibaura-it.ac.jp/

3.3 ベクトル化

T_k^C について word2vec[1][2][3] を使用し出現語のベクトル表現を得る．本研究ではテキストデータをベクトル化する際の次元数は 50，読み込む前後の単語は 5 とした．

4 結果

工学院大学で用いられている単語ベクトル「教育」から芝浦工業大学で用いられている単語ベクトル「教育」を引き，算出したベクトルから近い単語ベクトルの類似度を示した．分析対象単語の中で $D(A-B)$ に対する類似度が最も高い語を含む分析対象語「教育」を示す．

表 1:「教育」の差分に関する上位類似語

単語	類似度	出現回数 (大学 A)	出現回数 (大学 B)
教育		4442	6810
実習	0.505	206	5544
者	0.424	6192	3846
学校	0.418	2549	2478
成果	0.360	3359	864

5 考察

表 1 に現れた各単語について実テキスト上での使われ方の違いを調査した．その結果, 大学 A の「実習」では機械工学科や機械システム工学科の「加工」や建築学科の「製図」などの製作に関する単語の近傍に多く使われており,「コンピュータ」や「ソフトウェア」,「プログラミング」などの単語の近傍には「実習」ではなく「演習」または「実験」などの単語が使われていることがわかった．

一方で大学 B は「プログラミング」や機械の「加工」，建築の「製図」など区別なく「実習」という単語が使われていた．

このように，工学院大学と芝浦工業大学での「実習」の使い方の違いが「教育」の差を表したと考えられる．

表 2: 工学院大学の全学科

機械工学科	機械システム工学科
応用化学科	環境システム工学科
情報通信工学科	まちづくり学科
建築学科	建築デザイン学科
コンピュータ科学科	

表 3: 芝浦工業大学の全学科

工学部	
機械工学科	機械機能工学科
材料工学科	応用化学科
電気工学科	通信工学科
電子工学科	土木工学科 社会基盤コース
土木工学科社会 システムデザインコース	建築学科
建築工学科	情報工学科
教職課程	
システム理工学部	
総合科目	教職課程
電子情報システム学科	機械制御システム学科
環境システム学科	生命科学科生命科学コース
生命科学科生命医工学コース	数理科学科
デザイン工学部	
共通教養科目	教職課程
共通基礎科目	共通専門科目
専門科目	

6 おわりに

本研究は異文書における同一単語の機械的な意味の差を分析した．具体的には 2 つの大学の異なるシラバスを異文書として扱い, 同一単語のベクトルの差について分析した．

単語の使われ方の差を抽出することができた．今後は端的な意味を引き出していきたい．

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”, ICLR Workshop, 2013

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality”, In Proc. of NIPS2013, pp.3111–3119, 2013

[3] Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig, “Linguistic Regularities in Continuous Space Word Representations”, In Proc. of NAACL-HLT-2013, pp.746–751, 2013