1ZH-04

情報分野における教育講座と学習基準の自動対応付け

三浦行揮† 野寄祐樹† 齋藤大輔† 鷲崎弘宜† 深澤良彰†

早稲田大学大学院基幹理工学研究科情報理工・情報通信専攻†

1. はじめに

社会には大学での講義や社会人向けの講座など様々な教育講座が存在する。教育機関では、これらの講座が扱う内容を把握するために、教育講座を学習基準と対応付けた表の作成が行われている[1]。学習基準とは SFIA(Skills Framework for an Information Age)のような、学習者が学ぶべきスキルを示すものを指す[2]。対応付けの例を図 1 に示す。しかし、この対応付けは人手で作成されるためコストが大きいという課題がある。特に情報分野では技術の進歩に合わせて講義で扱う内容が変わるため、対応付けを行う頻度は高くなり、コストはより大きくなる。

この課題を解決するために、我々は教育講座と学習基準の自動対応付けを提案する。提案手法により対応付け作成のコストを削減でき、講義内容が変わった際にも同じシステムを通して対応付けを作成することができる。

本研究では、分散表現のコサイン類似度を用いた手法、Bag-of-Words を用いた教師あり学習、分散表現を用いた教師あり学習という 3 つの手法で自動対応付けを行った。また、手動で作成した対応付けと比較することで自動対応付けを評価した。

実験の結果、教材の内容把握として学習済み分散表現モデルを使用し、対応付け作成として教師あり学習を用いた手法において最も F 値が高い値を示した。

評価基準項目 ＼ 教育コンテンツ	ソフトウェア工学A	プログラミング基礎	オペレーティングシステムA
ソフトウェアの標準化	○		
ソフトウェアエンジニアリングツール・開発技術	○		
ソフトウェア構築の基礎知識	○	○	
ソフトウェア設計の基礎知識	○	○	
プログラミング基礎技術		○	
システム開発の概念と方法論	○		
システム開発のアプローチ	○		
ソフトウェア要件定義	○		
アプリケーション設計		○	

図 1 対応付けの例

Automated Educational program mapping on learning standard in computer science.

† Waseda University

2. 手法

本研究で提案するシステムは、教育講座の教材(Power Point および PDF)を入力とし、指定の学習基準に自動で対応付けされた表を出力する。本研究では以下の 3 つの手法で対応付けを行った。ここで、学習基準に含まれる単語は教材中の単語に比べ抽象的であるため、単語の存在有無ではなく特徴を計算できる手法を選択した。

(1) コサイン類似度を用いた類似度比較
(2) 教師あり学習(Bag-of-Words を使用)
(3) 教師あり学習(学習済み分散表現を使用)

2.1 コサイン類似度

教材が扱う学習基準項目を判断するために、教材中に存在する単語と学習基準の各項目が含んでいる単語の類似度を計算した。類似度を計算する際にはあらかじめ Wikipedia の文章を学習した分散表現モデル(Github 上で公開されたもの[3])を使用した。教材や学習基準項目に存在する単語を分散表現モデルにより 300 次元のベクトルとし、コサイン類似度を計算することで類似した単語の存在を判断する。分散表現には Python の genism ライブラリから word2vec を使用した[4]。

2.2 教師あり学習(Bag-of-Words)

自動で対応付けを作成するために、教材の特徴を説明変数、対応付く学習基準の項目を目的変数として、ランダムフォレストによるマルチラベル分類を行った。教師あり学習とシステムの全体像を図 2 に示す。説明変数とした教材の特徴は、入力に用いた全ての教材中に存在する単語から辞書を作成し、教材 1 つが含む単語から Bag-of-Words により多次元のベクトルで表現した。正解データは教材の作成に関わった人を中心に作成した手動対応付けを使用した。機械学習には Python の Scikit-learn ライブラリを使用した。

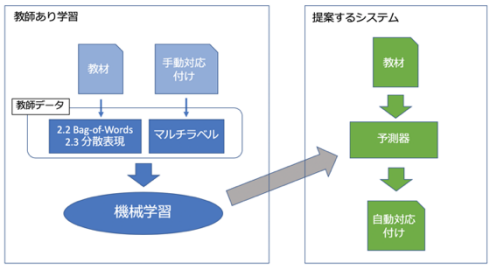

図 2 提案するシステム

2.3 教師あり学習(学習済み分散表現)

2.1 節と 2.2 節の処理を組み合わせて教師あり学習を行う。学習のプロセスは 2.2 節と同様に行い、教材の特徴としての説明変数を Bag-of-Words ではなく分散表現を用いて表す。1 つの教材に含まれる全ての単語をそれぞれ 300 次元のベクトルで表現し、その平均を説明変数とした。

3. 評価

実際に使用されている教育講座と学習基準に提案手法を適用し、3 つの手法による自動対応付けの F 値を比較する。

3.1 データセット

本研究では教育講座として Smart SE、学習基準として ITSS+を使用する。Smart SE とは、早稲田大学を中心とした AI・IoT・BD 技術分野の社会人学び直しプログラムである[5]。ITSS+とは、IPA(情報処理推進機構)が社会人向けに学び直しの指針として作成した、IT に関するスキルをまとめた表である。Smart SE の教材は 92 個用意し、それぞれの手動対応付けを事前に作成する。そのうちランダムな 82 個の教材と手動対応付けを訓練データ、残りの 10 個をテストデータとした。コサイン類似度による自動対応付けも同様の 10 個の教材に適用する。

3.2 自動対応付け手法の評価

自動対応付け手法の評価として、手動で作成した対応付けと比較し、Precision、Recall、F 値を計算した。結果を表 1 に示す。

表 1　自動対応付けの結果

	コサイン類似度	教師あり学習 (BoW)	教師あり学習 (分散表現)
Precision	0.313	0.545	0.706
Recall	0.417	0.240	0.480
F 値	0.357	0.333	0.571

表 1 から、学習済み分散表現を用いた教師あり学習が最も高い F 値 0.571 を示すことがわかる。F 値が 0.571 に留まった原因として、教師データの多くが対応有りとする学習基準項目が存在し、予測の際に説明変数の値に関わらず対応有りとされたことが考えられる(対応無しも同様)。

また、教師あり学習では説明変数として BoW ではなく分散表現を用いた手法において高い F 値を示した。これは、BoW は単語の有無から特徴ベクトルを作成するのに対し、分散表現では前後の単語から目的の単語の意味を数値化するため、教材の特徴を抽象的に表現できたためだと考えられる。学習基準中の単語は教材中の単語に比べ抽象的なので、BoW よりも分散表現が適していた。

さらに、コサイン類似度を用いた手法が低い F 値を示したのは、教材や学習基準中の抽象的な単語(IT, IoT)と、より具体的な単語(IoT システム、IoT サービス、IoT データ)をコサイン類似度で区別できなかったためだと考えられる。これにより本来対応付かない学習基準項目も対応有りとして予測するケースが多くなる。これは、コサイン類似度を用いた手法において Recall に対して Precision が低いことからも分かる。

4. おわりに

本研究では 3 つの手法を用いて教育講座と学習基準の対応付けを自動で行った。手動で作成した対応付けと比較したところ、教材の内容把握として学習済み分散表現モデルを使用し、対応付け作成として教師あり学習を用いた手法において最も F 値が高い値を示した。

分散表現を用いた教師あり学習においてさらに F 値を向上させるためには、正解データを増やすことに加え、教材の特徴より正しく数値化することが必要と考えられる。このためにドメイン知識をカバーした学習済み分散表現モデルの作成や、教材の重要な部分に重みをつけることを進めている。

5. 参考文献

[1] Tetsuro Kakeshita, Mika Ohtsuki, “A Relationship Analysis Tool among J07, JITEE and Job Type Utilizing i-Competency Dictionary”, IIAI-AAI, 273-278, 2015
[2] C. Nuangjamnong, S. P. Maj, D. Veal, “Resource Redundancy - A Staffing Factor using SFIA”, Innovations and Advances in Computer Sciences and Engineering, 31-55, 2009
[3] Hiroki Nakayama, “A curated list of awesome embedding models tutorials, projects and communities.” https://github.com/Hironsan/awesome-embedding-models, 2019
[4] Joseph Lilleberg, Yun Zhu, Yanqing Zhang, “Support vector machines and Word2vec for text classification with semantic features”, IEEE ICCI*CC, 14th, 2015
[5] 鷲崎弘宜, D-DATa & スマートエスイー: 早稲田大学における大学院や社会人対象の高度データ人材育成の取り組み, 大学教育と情報, 2018 年度, No.1 10-11, 私立大学情報教育協会, 2018