A Research Framework for Understanding Education-Occupation Alignment with NLP Techniques

Renzhe Yu
University of California, Irvine
Irvine, CA, USA
renzhey@uci.edu

Subhro Das
MIT-IBM Watson AI Lab, IBM Research
Cambridge, MA, USA
subhro.das@ibm.com

Sairam Gurajada
IBM Research – Almaden
San Jose, CA, USA
sairam.gurajada@ibm.com

Kush R. Varshney
IBM Research – T. J. Watson Research Center
Yorktown Heights, NY, USA
krvarshn@us.ibm.com

Hari Raghavan
IBM Corporate Social Responsibility
New York, NY, USA
hraghav@us.ibm.com

Carlos X. Lastra-Anadon
IE University
Madrid, Spain
clastra@faculty.ie.edu

Abstract

Understanding the gaps between job requirements and university curricula is crucial for improving student success and institutional effectiveness in higher education. In this context, natural language processing (NLP) can be leveraged to generate granular insights into where the gaps are and how they change. This paper proposes a three-dimensional research framework that combines NLP techniques with economic and educational research to quantify the alignment between course syllabi and job postings. We elaborate on key technical details of the framework and further discuss its potential positive impacts on practice, including unveiling the inequalities in and long-term consequences of education-occupation alignment to inform policymakers, and fostering information systems to support students, institutions and employers in the school-to-work pipeline.

1 Introduction

One important role of higher education is to prepare students for the workforce, but not all college graduates benefit equally from their degrees: more than 40% of recent college graduates are either unemployed or work in jobs not requiring a degree (Federal Reserve Bank of New York, 2020). On the other side of the equation, 45% employers worldwide report having difficulty "finding the right skills or talent" (Manpower Group, 2018). Are there significant gaps between what higher education offers and what employers look for? What are the sources of these gaps? Addressing these questions can bring substantial policy implications and positive societal impacts, such as mitigating inequalities in labor market outcomes across student groups and major areas.

Given that college education is delivered predominantly through structured coursework (Kuh et al., 2007), we assume that curricular content and its correspondence with employers' demand may be an important driver of differences in student outcomes in the labor market. Nonetheless, there has been little consensus on the definition of this correspondence and the understanding of how it contributes to the observed gaps (Cleary et al., 2017). One challenge behind this void is the lack of data that can capture the dynamics of labor market demands and the details of curricular content on a large scale. With the recent availability of digitized records of course content and job requirements as well as advances in computational methods, granular and scalable analysis of the correspondence between the two becomes possible (Börner et al., 2018).

In this paper, we present a novel research framework to measure the alignment between curricular content and labor market demands. We leverage neural network-based language models and other NLP techniques to learn representations of relevant documents. Based on these representations and theoretical insights, we incorporate three lenses through which to measure the alignment: skill overlap (economic), instructional design features (educational), and semantic text similarity (technical). This framework represents the first comprehensive and scalable approach for connecting the content of

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 100–106
Bangkok, Thailand (online), August 5, 2021. ©2021 Association for Computational Linguistics

education and workforce, which was either treated as a black box or investigated on a small scale in prior research (Walker, 2020; Hora, 2019). Moreover, the computational capacity of this framework can empower system-wide policy research as well as local practices regarding curricular alignment and workforce preparation, thereby bringing positive societal impacts. For example, university stakeholders can track the downstream consequences of ill-aligned curricula especially for students from marginalized groups.

In Section 2, we briefly summarize prior research related to the technical and substantive aspects of our work. Section 3 details our three-dimensional framework that measures education-occupation alignment. Section 4 envisions the societal benefits of our framework through assisting downstream policy research and field practice of different stakeholders in the school-workforce pipeline. Finally, we conclude with a summary and next steps in Section 5.

2 Related Work

2.1 Natural Language Processing

Recent advances in language models have shown promising results in representing texts for different downstream NLP applications. Pre-trained language models such as GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019) encode a sentence (or a document) into multi-dimensional vectors (a.k.a. embeddings). More recently, specialized long-text document encoders (Adhikari et al., 2019; Beltagy et al., 2020) have emerged and achieved state-of-the-art performance in benchmark tasks.

Based on these embeddings, one can learn the alignment, or similarity, between different documents with sentence-pair regression models (Reimers and Gurevych, 2019) or twin networks architectures such as Siamese networks (Bromley et al., 1993). Alternatively, document alignment can be evaluated via labels. Each document can be attached to one or multiple pre-defined labels, and two documents align well if they share a decent proportion of labels. Under this formulation, the core task becomes attaching labels which is essentially a text classification problem. In our context, for example, the "labels" that curricular and job content share are skills. Due to the sheer volume of possible skills, this becomes an extreme multi-label classification task (XMC) as recently pointed out by Bhola et al. (2020). In their closely relevant work, BERT models are employed to learn an embedding for a job description and XMC models then classify each embedding into a subset of skills over a large pool of predetermined skills set.

2.2 Content Analysis of Curricula, Jobs, and Their Relationships

Curricular content is key in teaching and learning research, but most prior research is built upon a small sample and/or requires extensive human coding (Hong and Hodge, 2009). In recent years, large-scale computational analyses of digitized curricular documents (e.g., textbooks, syllabi) have emerged to inform both instructors and policymakers (Lucy et al., 2020; Jiang and Pardos, 2020). The majority of these pioneering works employ bag-of-words representations of the documents and there remains abundant scope for deeper dives into intellectual and pedagogical features beyond the surface level.

There is a large literature documenting changes in job market demands in advanced economies. Only until the recent decade real-time data on job vacancies has enabled detailed assessment of the evolving skills required by jobs (Deming and Noray, 2020). For instance, Das et al. (2020) document the increase in the demand for jobs in the fields of big data and artificial intelligence (AI), Fleming et al. (2019) show that high- and low-wage jobs are gaining tasks and earning more.

The microscopic relationship between curricular content and job requirements is a novel topic, although it is essentially built upon the literature on labor market returns to education (Walker, 2020) with the recent availability of big data. To our knowledge, Börner et al. (2018) presented the first study on this relationship, using textual content of syllabi and job postings. Another relevant work examines the interplay between curricular content and academic research in a similar manner (Biasi and Ma, 2021). Our work expands on these attempts and incorporates more disciplinary perspectives to create a holistic research framework.

3 Measuring Education-Occupation Alignment

3.1 Problem Framing

As described in Section 2.2, the textual content of curricular offerings and job requirements have been increasingly available in machine-readable formats in the digital era. In general, different types of curricular documents include information such as subject matter content, learning objectives, instructional design, etc. Job-related documents, on the other hand, commonly describe required skills, responsibilities, qualifications, etc. Our framework is intended to be largely agnostic of specific document types and datasets as long as they include most of the aforementioned information and each document represents an individual course or professional position. In the descriptions below, we use course syllabi and job postings as examples of such documents.

The focus of our framework is measuring education-occupation alignment. Specifically, given a syllabus S_i, we want to learn an alignment metric $align\,(i;j)$ to capture how much it aligns with job posting P_j, and then use this metric to derive macroscopic alignment measures depending on the scope of analysis. Note that the metric is not symmetric and anchored to course syllabi, because educational providers (programs, institutions, etc.) in practice have more motivation and power to accommodate the labor market than the opposite. Building upon the existing literature, our framework incorporates the following three disciplinary dimensions along which to operationalize $align\,(i;j)$.

3.2 Economic Dimension: Skill Overlap

First, we treat skills as the bond between jobs and courses, because economists have highlighted skills as organizing units of the labor market (Acemoglu and Autor, 2011). In this sense, education-occupation alignment can be conceptualized as the extent to which a course syllabus covers skills required by the labor market. Intuitively, we have:

$$align\,(i;j) = \frac{|D_{ij}|}{|D_j|} \quad (1)$$

where $D_{ij} = \{s|s \in S_i, s \in P_j\}$, $D_j = \{s|s \in P_j\}$, and s is a specific skill in a finite skill pool.

Most job documents include expected skills, but curricular documents are not necessarily skill-focused. Therefore, computing Equation (1) translates into the task of predicting skills from the content of syllabi. There can be multiple NLP approaches for this task, and here we present an example inspired by Bhola et al. (2020) who frame skill identification as a multilabel classification problem. Specifically, we describe a BERT-LSTM architecture as illustrated in Figure 1. The lower part of the graph serves to learn document representations. It takes in a curricular or job document, leverages a pretrained BERT (Devlin et al., 2019) to learn a vector representation (`[CLS]` token) for each sentence, and feeds these sentence vectors through an LSTM (Hochreiter and Schmidhuber, 1997) model in sequential order to get the document-level representation (the last hidden state). This two-level stacked architecture is used because BERT is typically used to handle sentence-level tasks and both syllabi and job postings are usually longer than the recommended maximum sequence length. The top part of Figure 1 is a multilabel classifier constructed as a feed-forward neural network where the prediction targets are skill labels. Additional tweaks such as Correlation Aware Bootstrapping (Bhola et al., 2020) can be simply added for the sake of performance.

In the application scenario mentioned above, this skill prediction architecture can be trained and validated (except for the pre-trained BERT) on the job posting data and used to map course syllabi to the same skill space.

3.3 Educational Dimension: Instructional Design Features

Second, we focus on identifying the extent to which courses equip students with general social and cognitive skills, such as problem solving and communication, as research has validated their long-term economic returns (Deming, 2017). This dimension complements the last one because the focus on skill overlap is better at differentiating specialized skills that are concentrated in a smaller cluster of jobs, compared to general competencies that appear in almost every single posting (Coffey et al., 2020) and therefore are harder to predict in the multilabel classification framework (Figure 1).

In the educational literature, most of these general skills are aligned with the target competencies in a variety of teaching and learning frameworks (Fink, 2013; Krathwohl, 2002), which in most cases further connect to specific learning activities and instructional design. While not all curricular documents include detailed descriptions of course design, it is worth exploring the possibility of NLP-assisted coding of course design features. Table 1 presents an example of research-informed rubric, where each item captures a design feature which is associated with one (or more) competency. Some of the features are simply occurrences of

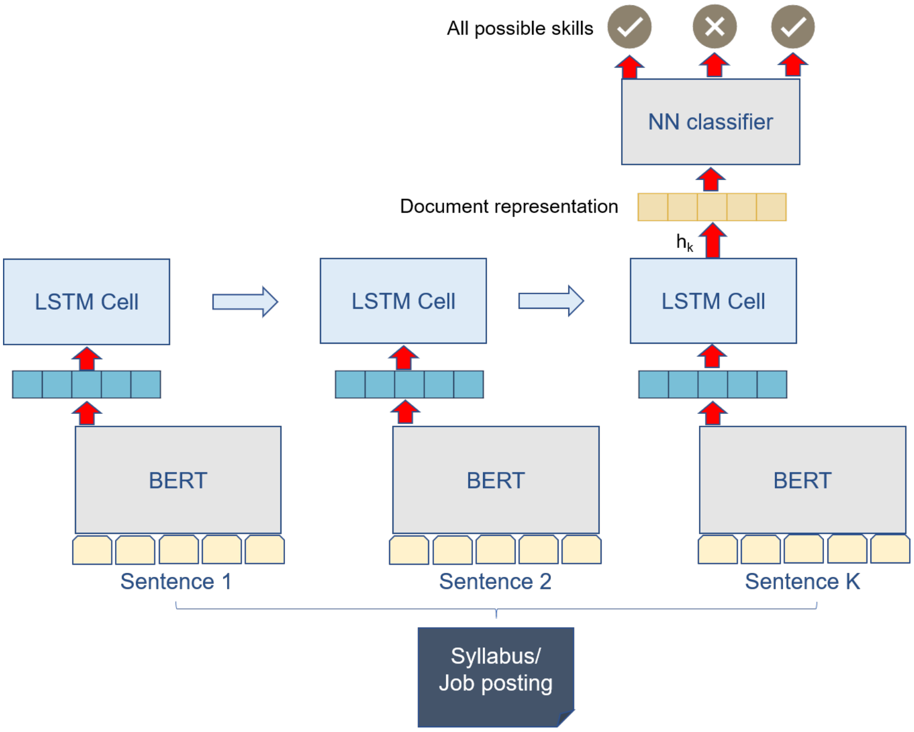

Figure 1: BERT-LSTM architecture to predict skills from course syllabi or job postings

Course design	Competency
Require group project	Collaboration
Require in-class presentation	Communication
Space out assignments	Time management
Encourage reflections	Critical thinking

Table 1: Example rubric of course design features mapped to general competencies

certain learning activities while others need more holistic examination of the course structure. To automatically code these items from a syllabus, a model architecture similar to that in Figure 1 might be useful, where the output skill labels are replaced by course design items. Admittedly, neither a comprehensive rubric connecting course design to higher-order competencies nor an NLP-assisted item coding pipeline is well researched, but both are promising directions.

With this setup, education-occupation alignment is not directly captured by $align\,(i; j)$ for individual pairs of documents; instead, a simple count of predicted course design items in syllabus S_i that are associated with any of predetermined general competencies will serve as the overall alignment measure for S_i.

3.4 Technical Dimension: Semantic Text Similarity

The last dimension is holistic and purely data-driven. Because scholarly understanding of the detailed language in curricular versus job documents is still limited, we assume that the overall semantic text similarity between them might reflect latent aspects of education-occupation alignment such as culture or values. As a prerequisite, we still wish to learn a vector representation for each document such that the similarity between S_i and P_j can be expressed as a simple function such as cosine similarity:

$$align\,(i; j) = \frac{\boldsymbol{v_i} \cdot \boldsymbol{v_j}}{\|\boldsymbol{v_i}\|\|\boldsymbol{v_j}\|} \tag{2}$$

where $\boldsymbol{v_i}$ and $\boldsymbol{v_j}$ are the document vectors of S_i and P_j, respectively. To learn these vectors, predictive architecture like in Figure 1 is not applicable because there is no target available to train the LSTM component for. Instead, we suggest feeding each document into pretrained Longformer (Beltagy et al., 2020) for the output representation (`[CLS]` token with global attention). To ensure that the document vectors are comparable through cosine similarity, siamese and triplet networks need to be created to update model weights following Reimers and Gurevych (2019). Alternatively, the earlier “doc2vec” model (Le and Mikolov, 2014) can also be used, as the resulting vectors are ready to use for cosine similarity.

4 Positive Impact

4.1 For Policy Research

With an established alignment metric for individual pairs of educational and job documents, we can aggregate them within different temporal, geographical or disciplinary boundaries to answer policy-relevant questions, such as the two examples below.

Inequalities in Education-Occupation Alignment. Traditional educational statistics were unable to collect information about curricular content. With the education-occupation alignment metric, we are able to systematically evaluate the differences across major areas and institutional characteristics in how well they prepare students for the workforce. From an equity perspective, if institutions with more students from underrepresented groups exhibit lower levels of labor market alignment, it suggests that the current landscape of higher education might exacerbate inequities in economic mobility. In other words, students’ socio-economic gaps that originate from their family background might propagate into their professional career.

Education-Occupation Alignment and Student Outcomes. Does better education-occupation alignment contribute to better student outcomes? The established alignment metric and the nationwide standard for administrative data[1] together make it possible to provide an empirical answer to this kind of question on a large scale, where important student outcome metrics may include graduation and earnings. If this alignment is an important driving force of student outcomes, institutional effort to prepare students for workforce should be directed more towards curricular reforms. Additionally, the fine-grained alignment metric enables us to examine if it forecasts longer-term student outcomes with different levels of confidence at different types of institutions and different time points.

4.2 For Practitioners

From a practical perspective, the capacity to quantify education-occupation alignment at scale can provide actionable insights to various stakeholders. Such insights might well be incorporated into some standard information system to directly facilitate the decision-making of these people.

College students. Existing research has shown that students at different institutions have limited knowledge about the returns associated with different degree majors (Baker et al., 2018); and that simple interventions such as providing earning prospects have the potential to help individual students make optimal choices. In a similar vein, our computational framework could help a student understand the possible alignment or lack of alignment between a prospective degree program and local labor market demands, based on the summary of detailed course-level analysis within that program or a similar program at another institution, assuming a certain degree of generalizability.

Higher education administrators. Our framework could help administrators identify potential curricular “hidden gems” or “problem areas” at their institution that might align well or not well with skills demanded by the labor market. Motivated by a desire to improve institutional effectiveness and students’ labor market outcomes, they could on one hand pursue curricular reforms and/or industry partnerships for the “problematic areas”, while sustaining resources for student recruitment, employer engagement and other operational aspects of the “hidden gems.”

Employers. Our framework could help employers refine or update their student recruitment strategies, if necessary, based on the alignment levels across major fields and institution types in their target area(s). The goal of this practice is to hire graduate talents whose skills are better suited for the employer’s needs. In some scenarios, such efficiency-oriented decisions might ultimately extend opportunities to students who previously were not as likely to be considered for certain roles with the employer due to the lack of granular analysis of course content. In this case, the use of our framework might eventually contribute to the diversity, equity, and inclusion (DEI) goals of the employer and of the local community in general.

5 Summary

We propose a research framework for measuring the alignment between curricular content and job requirements by leveraging NLP techniques. Based

[1] https://nces.ed.gov/ipeds/

on neural representations of curricular and job documents, our framework includes three dimensions for quantifying education-occupation alignment: 1) amount of specialized knowledge and skills shared by the two types of documents, 2) quantity of instructional design features associated with general social and cognitive competencies, and 3) overall semantic text similarity between the two corpora. We discuss how the framework can help researchers answer education and economic policy questions, and empower stakeholders in practice to make more informed decisions around recruitment, course development, major/course choice, etc.

We focus on sketching the high-level picture of the proposed framework through examples, while leaving plenty of space for technical details and future work by ourselves and others. Given the importance of education-occupation alignment especially in the post-pandemic era, and the widely available yet underutilized corpora data of curricular and job content, we call for more cross-disciplinary collaborations on the topic to contribute to healthier education-occupation dynamics of the future.

Acknowledgments

This work was conducted under the auspices of the IBM Science for Social Good initiative.

References

Daron Acemoglu and David Autor. 2011. Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification.

Rachel Baker, Eric Bettinger, Brian Jacob, and Ioana Marinescu. 2018. The effect of labor market information on community college students' major choice. *Economics of Education Review*, 65:18 – 30.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5832–5842.

Barbara Biasi and Song Ma. 2021. The Education-Innovation Gap.

Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. 2018. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, 115(50):12630–12637.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jennifer Lenahan Cleary, Monica Reid Kerrigan, and Michelle Van Noy. 2017. Towards a New Understanding of Labor Market Alignment. In Michael B. Paulsen, editor, *Higher Education: Handbook of Theory and Research*, volume 32, pages 577–629.

Clare Coffey, Gwen Burrow, Rob Sentz, Kevin Kirschner, and Yustina Saleh. 2020. Resilient skills: The survivor skills that the class of covid-19 should pursue. Technical report, Emsi.

Subhro Das, Sebastian Steffen, Wyatt Clarke, Prabhat Reddy, Erik Brynjolfsson, and Martin Fleming. 2020. Learning Occupational Task-Shares Dynamics for the Future of Work. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

David J. Deming. 2017. The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

David J Deming and Kadeem Noray. 2020. Earnings Dynamics, Changing Job Skills, and STEM Careers. *The Quarterly Journal of Economics*, 135(4):1965–2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federal Reserve Bank of New York. 2020. The Labor Market for Recent College Graduates.

L Dee Fink. 2013. *Creating significant learning experiences: An integrated approach to designing college courses*. John Wiley & Sons.

Martin Fleming, Wyatt Clarke, Subhro Das, Phai Phongthiengtham, and Prabhat Reddy. 2019. The future of work: How new technologies are transforming tasks. Technical report, MIT-IBM Watson AI Lab.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Philip Young P Hong and David R Hodge. 2009. Understanding social justice in social work: A content analysis of course syllabi. *Families in Society*, 90(2):212–219.

Matthew T Hora. 2019. *Beyond the skills gap: Preparing college students for life and work*. Harvard Education Press.

Weijie Jiang and Zachary A Pardos. 2020. Evaluating sources of course information and models of representation on a variety of institutional prediction tasks. In *Proceedings of the 13th International Conference on Educational Data Mining*.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

George D. Kuh, Jillian Kinzie, Jennifer A. Buckley, Brian K. Bridges, and John C. Hayek. 2007. Piecing Together the Student success puzzle: Research, Propositions, and Recommendations. *ASHE Higher Education Report*, 32(5):1–182.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312.

Manpower Group. 2018. Solving the Talent Shortage: Build, Buy, Borrow and Bridge. Technical report, Manpower Group.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ian Walker. 2020. Heterogeneity in the returns to higher education. In *The Economics of Education*, pages 75–90. Elsevier.