Social work in the classroom? A tool to evaluate topical relevance in student writing

Heeryung Choi
School of Information
University of Michigan
heeryung@umich.edu

Zijian Wang
Department of EECS
College of Engineering
University of Michigan
zijwang@umich.edu

Christopher Brooks
School of Information
University of Michigan
brooksch@umich.edu

Kevyn Collins-Thompson
School of Information
University of Michigan
kevynct@umich.edu

Beth Glover Reed
Social Work and Women's Studies
University of Michigan
bgr@umich.edu

Dale Fitch
School of Social Work
University of Missouri
fitchd@missouri.edu

ABSTRACT

In a climate where higher education institutions are actively aiming to increase inclusivity [2], we explore how a deep learning-based tool focused on text analysis is able to help assess how students think about issues of privilege, oppression, diversity and social justice (PODS). We created a vocabulary boosting and matching tool augmented with domain-specific corpora and relevance information. We find that the adoption of domain-specific corpora enhances model performance when identifying PODS-related words in short student-written responses to writing prompts, by building a more highly focused PODS vocabulary.

1. INTRODUCTION AND RELATED WORK

Universities are expanding their efforts toward creating more inclusive institutions of higher education [2]. One specific example is the principled blending of curricula with social justice and diversity issues in order to encourage PODS thinking (Privilege, Oppression, Diversity, Social justice) in the School of Social Work at the University of Michigan. PODS principles have been emphasized not only in individual courses but throughout the whole Social Work curriculum. Such a move naturally raises the question of scaled evaluation, both of individual students (e.g. formative or summative assessment) and programmatic evaluation.

In previous work, we explored mechanisms to detect elements of PODS thinking in student writing through semi-supervised machine learning [1]. We adopted the Empath tool [3] to generate an expanded vocabulary from a few seed words for PODS thinking detection, but were extremely limited in our ability to achieve accurate results. The first issue stems from the selection of large but general corpora which, while large in size and topic coverage, were not effective when we attempted to learn domain-specific bigrams. The other issue is how to filter less relevant words while boosting the size of the relevant lexicon. While generating a lexicon for Social Justice on Empath, we found that semantically irrelevant words like "therefore" and "yet" were in the output lexicon [1]. Thus, we expand on previous results and demonstrate a more robust and thorough treatment of the issues of detecting PODS thinking in student writing.

In this work, we consider the specific case of short student writings given in response to a writing prompt. Our goal is to build a technology solution that gives accurately coded responses and that enables instructors to identify quickly which students need elaborated feedback. The system will allow the instructors to focus remediation efforts on those who are of the highest need and to assess how well the overall curricula could increase PODS competency of students. Here we demonstrate the feasibility of using deep learning methods to detect evidence of PODS and apply these methods to a particular writing activity, innovating on the process used by others [3] to improve accuracy and reliability.

2. INSTRUMENTS

We created **Metapath**, a text analysis tool that allows users to use not only general corpora but also domain-specific corpora. Metapath is built on the ability of the Word2Vec model to calculate the similarity of concepts by mapping words and phrases to a vector space via a skip-gram model, and computing the cosine similarity of the corresponding vectors [4]. Given a word, the model gives users a 'most similar' word list ordered by the similarity score. In a preprocessing step, short words ($length \leq 2$), non-English terms, and most stopwords are considered as noise and removed from the corpora. After data cleaning, all words are stemmed using Porter stemming. Common phrases, i.e., multiword expressions, can be detected automatically by calculating mutual information gain within a threshold and minimum count. For example, the words 'Los Angeles' will become the phrase `los_angeles` after phrase detection while the model will return a list of high similarity words like `san_francisco` and `santa_barbara`. The judgment of whether the words are common phrases is based on the formula

$$\frac{cnt(a,b) - min\_count}{cnt(a) \cdot cnt(b)} \cdot N > threshold$$

where $cnt(a,b)$ means the frequency of word a and word b located together and N is the total vocabulary size.

We chose to use domain-specific corpora, i.e., MICUSP (Michigan Corpus of Upper-level Student Papers) and BAWE (British Academic Written English) [5], for detecting common phrases. The general Wikipedia corpus is used to train the model. In addition, considering the contextual nature of the PODS words, existing student responses gathered from courses were included as a corpus. The domain-specific corpora are able to detect more related phrases on the topics of interest. For example, the proportions ($10^{-3}\%$) of stemmed words like 'prejudic' and 'social_justic' in domain-specific corpora were relatively high (respectively 0.079 and 0.015), compared to the proportions of the same words in the general corpora, which were much lower (0.012 and 0).

3. EVALUATION

We conducted an evaluation to assess how well Metapath can assess PODS-related writing, using our domain-specific corpora, along two dimensions: comparing (1) inter-rater reliability (IRR) for PODS word annotation between human raters and Metapath and (2) IRR for quality evaluation between human raters and Metapath. The latter method is to include percentage of relevance of PODS words, which shows how semantically related each word is to seed words.

3.1 Data

The students' short written responses on PODS topic were used to evaluate Metapath, collected from four sections of a course offered in the School of Social Work ($n = 100$, word counts; $\bar{x} = 695.52$, $\sigma = 434.08$, $min = 115$, $max. = 2747$).

3.2 Approaches

For the evaluation, two expert human coders annotated PODS-related words in the student responses and evaluated overall PODS-relevance of each writing piece with three different marks: high, medium, and low. Their annotations and quality evaluation on student responses were compared with result of Metapath. To build a lexicon to evaluate PODS relevance of student writing, Metapath was boosted by essential PODS words, i.e., privilege, oppression, diversity, and social justice. Furthermore, two keywords from the writing prompt, i.e., "issues" and "actions", were also used to boost the PODS lexicon. After we boosted a lexicon (dim=500), the lexicon was used to calculate the IRR on annotations among two human raters and Metapath. The lexicon and its percentage of relevance were used to assess the overall PODS relevance of each response. After all the responses were ranked based on their percentage of relevance, they were categorized into high, medium, and low. The threshold of the each category was based on the proportion of each category decided by the human raters.

4. RESULTS AND DISCUSSION

We calculated group agreement among the two human raters and Metapath using Krippendorff's alpha (α). For the annotation comparison, IRR among two human raters alone is $\alpha = 0.4480$ ($n = 100$). When we added Metapath the overall group agreement dropped to $\alpha = 0.3804$ ($responses = 100$, $boosted\ words = 4300$, the maximum and minimum possible agreement the 3-rater scenario: $-0.4056 \leq \alpha \leq 0.6324$). IRRs between each human rater individual and Metapath were $\alpha = 0.1622$ and $\alpha = 0.1822$. For the quality evaluation, we achieved $\alpha = 0.3441$ ($responses = 100$, $boosted\ words = 660$) as the level of agreement between human raters and Metapath, which is close to the IRR between the two human raters ($\alpha = 0.4393$, the maximum and minimum possible agreement among 3-rater scenario: $-0.1875 \leq \alpha \leq 0.6223$). IRRs between each human rater individually and Metapath were $\alpha = 0.3702$ and $\alpha = 0.2234$. Overall, the evaluation showed that Metapath could identify PODS-related words and overall PODS relevance. The IRR that Metapath reached was close to those of human raters and not too low, considering the possible minimum and maximum agreement range.

It is worth pointing out that higher agreements in PODS word detection do not align with higher agreements in overall PODS relevance. We varied the size of Metapath's vocabulary by 500 words through setting *the number of boosted words* parameter. Even quite large vocabularies boosted the effectiveness of Metapath in the first task, declining only when values reached $n \approx 4000$. However, the IRR for quality analysis was the highest when $n = 660$.

Further research is needed to explore and improve the performance of Metapath. While identifying PODS-related words, there are still words and phrases in the field of social work that are not detected by Metapath, as noted by the experts. One way to address this is to focus on improved corpora, such as increasing the amount of response data generated by social work students and articles or books curated by PODS experts, or by using corpora based on accumulated Social Work student's writing. Finally, we note that this task is highly multifaceted, and here we have taken just a first pass at addressing it. Issues of personally-lived experiences, intersectionality of topics, and the nature of the writing prompt itself may require more traditional natural language processing techniques in order to capture deeper relationships in the text more fully.

5. ACKNOWLEDGEMENTS

The data used come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes, with funding from the ESRC (RES-000-23-0800). This study was funded in part with support from the Michigan Institute for Data Science (MIDAS).

6. REFERENCES

[1] H. Choi, C. Brooks, and K. Collins-Thompson. What does student writing tell us about their thinking on social justice? In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 594–595. ACM, 2017.

[2] E. DeRuy. The complicated process of adding diversity to the college syllabus. *The Atlantic*, Jul 2016.

[3] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in Large-Scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[5] M. B. O'Donnell and U. Römer. From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 7(1):1–18, 2012.