The Remembered

A Global Study of Literature Dissertations' Bibliography

Gutiérrez De la Torre, Silvia Eunice
silviaegt@gmail.com
Leipzig Universität, Deutschland

My doctoral project aims to build a map of contemporary literary research practices in German universities. To do so, I will apply reference mining techniques on a corpus of ca. 1,000 full-text electronic doctoral dissertations (ETDs) from Literature Departments in Germany (2000-2020) compiled via a research agreement with the National Library. In the following lines I will explain why ETDs are a necessary source to gain a much-needed strategic bird sight of the field, but also, what the potential problems of data heterogeneity in this dataset are, and how they will be addressed.

Dissertations are the default "rite of passage" through which students become researchers. Doctoral research typically passes through several phases of approval by an institutional community: the proposal acceptance, the internal upgrade/review, and the external panel or examination. Thus, they are a good compass of what different institutions consider "satisfactory knowledge" to become an academic. Moreover, other than briefer genres of academic writing such as journal articles, dissertations tend to require in Germany, on average, 5 years (Jakisztat et al., 2012). Hence, it is safe to say that the reference lists contained in 1,000 Literature ETDs represent the readings of at least 5,000 years of human research that has not been analyzed until now.

Citation Analysis (CA) is the area of bibliometrics that studies the relationship between a cited and a citing document, and it has been increasingly used to "study, map, and evaluate academic research" (Hammarfelt, 2012). The basic input for CA is a citation database and Reference Mining (RM), a Natural Language Processing (NLP) task focused on the "detection, extraction and classification of bibliographic references and their constituent components" (Rodrigues Alves et al., 2018), has been a useful computational method to obtain this information as data.

However, as noted by Rodrigues Alves et al. (2018), unlike scientific publications (which have been the target of most RM methods), Humanities texts are significantly less structured. The following three reasons should be considered: 1) Humanities research uses both primary and secondary sources, and the former are, by definition, more varied; 2) references can happen anywhere in the text (footnotes, image captions, etc.); and lastly 3) "the variety of publication venues, languages, scholarly communities in the arts and humanities are broader, making reference practices and styles less uniform" (Rodrigues Alves et al., 2018). To this list we should add the fact due to the lack of strict editorial guidelines, dissertations tend to have a less structured reference list than in established publications.

Available methods (Tkaczyk et al., 2018) rely on a painstaking tagging process which nonetheless only works with texts that are close to the original training dataset (Grennan & Beel, 2020). Moreover, these methods require complex, black-box-like, and highly carbon-emitting computer power. In this presentation I will showcase a method that combines exploratory data analysis with readable machine learning results to automatically detect pages with reference lists on which NLP techniques such as Name Entity Recognition (NER) and fuzzy matching will be used to pair reference strings with richer metadata.

My hypothesis is that by paying attention to the citation patterns in a nation-wide corpus of literary dissertations, it is possible to reveal patterns of literary research in at least three dimensions: broad regional or institutional trends; interdisciplinary connections; and genre defined behaviors. With this in mind, this research proposes to answer the following questions:

1. What are the regional or institutional trends of literary research? (i.e. who are the most cited authors in Bavarian institutions, and how are they different or similar from those cited in other parts of Germany?)
2. How intercultural are these approaches? How often is Jorge Luis Borges cited along his German influences: Hölderlin, Silesius, Goethe?
3. Is it true that multidisciplinary approaches are becoming more popular and if so, what are their characteristics? For example, do medieval studies have a canon that includes publications from different sciences?
4. Is it possible to create topology of the citation networks of different genres and subgenres? For instance, which patterns of co-citation emerge in feminist fiction?

Reference mining Humanities dissertations is a challenging task that will require much more coordinated effort. Yet, this proposal offers, on the one hand, a comprehensive technique to extract, at least coarsely, the bibliographic "bricks" upon which PhD students build "new knowledge"; and on the other, a map of what and how (in which bibliographic interconnections) students in Germany have been analyzing literature in the last 20 years.

Bibliography

Hammarfelt, B. (2012). "Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies." *Journal of Documentation* , 68 (4): 536–58.

Jaksztat, S., Preßler, N., & Briedis, K. (2012). *Promotionen im Fokus: Promotions- und Arbeitsbedingungen Promovierender im Vergleich*. https://www.dzhw.eu/pdf/pub_fh/fh-201215.pdf

Rodrigues Alves, D., Colavizza, G. and Kaplan, F. (2018). "Deep Reference Mining From Scholarly Literature in the Arts and Humanities". *Frontiers in Research Metrics and Analytics*, 3 . Frontiers doi:10.3389/frma.2018.00021. https://www.frontiersin.org/articles/10.3389/frma.2018.00021/full (accessed 23 October 2020).

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). "Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers". *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108. https://doi.org/10.1145/3197026.3197048