500–1000 words (Excluding references)

Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora (web-as-corpus, digital cultural heritage, natural language processing, machine learning)

Web-scale corpora, automatically collected from the web and encompassing billions of words, present significant opportunities for diverse fields of research. These corpora play a pivotal role in the advancement of large language models, such as the one underpinning ChatGPT. Moreover, they include masses of texts produced in different situations with different objectives, and they host new forms of digital cultural heritage that are constantly emerging and evolving. Therefore, they open up new avenues for research within the humanities, and social sciences, but also require multidisciplinary collaboration to guarantee their usability. (Laippala et al. 2021; Välimäki and Aali 2022)

A notable challenge associated with web-scale corpora is the absence of metadata detailing their contents. Typically, they lack information regarding the origin and the content of the documents. Documents featuring different text varieties, ranging from legal notices to advertisements, news articles, fiction, and song lyrics, all have an equal status in the corpora. This study aims to address this challenge by exploring various approaches to classify web corpora to specific subsections. In particular, we focus on *registers*, typically applied in corpus linguistics, defined as situationally defined text varieties (Biber and Conrad 2019), and *genres*, often utilized in literary studies when examining various forms of literary work (e.g., Goyal and Prakash 2022; Zhang et al. 2022).

In recent years, web register identification has taken leaps forward, with web register classifiers achieving nearly human-level performances (Laippala et al. 2023; Kuzman et al. 2023). However, in practice, when applied to web-scale corpora, the predicted register classes are still very broad, including a wide range of linguistic variation. Therefore, in this study, we examine if and how the available information can be deepened by combining two approaches: registers and genres.

We apply machine learning to train two text classifiers: one targeting registers and one focusing on genres. We utilize these classifiers to predict the classes for one million documents in the widely applied, web-scale Oscar dataset (Suarez et al. 2019; Laippala et al. 2022). Then, we 1) evaluate the distributions of the text classes predicted using the two classifiers; 2) analyze the intersections of the classes, and 3) examine how the combination of the two approaches extends the metadata available for the corpus.

The register classifier is trained using the CORE corpora (Biber and Egbert 2018; Laippala et al. 2022, 2023). The scheme is hierarchical and covers eight main categories with broad, functional labels such as *narrative, informational explanation* and *opinion,* and more detailed subcategories such as *news report, research article* and *review*. The training data for our genre classification model consists of books from Kindle US. The genre categories are assigned by the authors and selected from the possible categories of Kindle US, which include categories such as *Children's books*, *Science & Math*, and *Action & Adventure*. The original dataset is available at https://huggingface.co/datasets/marianna13/the-eye and a cleaned version used in training is available at our Huggingface page at [deleted-for-review].

As some of the genre classes in the dataset are overlapping, we perform some pre-processing steps to improve the data quality. We choose a subset of genres that maximize the performance in two ways: firstly, the chosen genres need to be present in our corpus, Oscar. As a web corpus, genres most suitable for our task include *Medicine & Health*; *Cookbooks, Food & Wine*; *Engineering & Transportation*; and *Politics & Social Sciences*, which are common topics in online sources. Secondly, as categories are partially overlapping and some contain very few examples, we choose other categories based on their support in the dataset and by testing the performance with different candidate subsets.

The register model is implemented by finetuning XLM-RoBERTa-Large (Conneau et al. 2020) using the CORE corpus with the task of register identification modeled as multilabel classification. For training, we use the Huggingface Transformers library. Preliminary results show the register classifier is able to reach an F1-score of 0.77. We also use XLM-RoBERTa-Large for the basis of our genre classifier. Experiments are done on multiple genre subsets as described above. Similarly to the register model, the task is framed as multilabel classification and again, we use the Huggingface Transformers library. We select the best prediction threshold based on the F1-score. Our preliminary results show an F1-score of 0.70.

We use the two classifiers to label one million documents of the Oscar corpus. In our experiments, we see some expected combinations between certain registers and genres, such as the *Lyrical* register and the *Literature & Fiction* genre often coinciding, but equally registers such as *Interactive Discussion* being divided into multiple genres, like *Engineering & Transportation* and *Politics & Social Sciences* based on the topic of the discussion. Our preliminary qualitative evaluation shows that the predicted genre and register labels provide valuable auxiliary information which facilitates the use of the corpus in new ways in the study of digital cultural heritage. We will be analyzing the intersection and the different combinations of genre-register pairs using topic modeling

and study of keywords as well as evaluating the benefits of cross-labeling a corpus as a tool for creating additional metadata.

References

Biber, Douglas, and Conrad, Susan. 2019. “Register, Genre, and Style”. 2nd ed. of Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, Douglas, and Egbert, Jesse. 2018. “Register variation online”. Register Studies. 2. 166-171. 10.1075/rs.19018.smi.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. “Unsupervised Cross-lingual Representation Learning at Scale”. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Goyal, Anshaj, and Prakash, V. Prem. 2022. “Statistical and Deep Learning Approaches for Literary Genre Classification”. In Advances in Data and Information Sciences, 297–305. Lecture Notes in Networks and Systems. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-5689-7_26

Kuzman, Taja, Rupnik, Peter, and Ljubešić, Nikol. 2023. “Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora”. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics

Laippala, Veronika, Kyröläinen, Aki-Juhani, Kanerva, Jenna and Ginter, Filip. 2021. “Dependency profiles in the large-scale analysis of discourse connectives”. Corpus Linguistics and Linguistic Theory, Vol. 17 (Issue 1), pp. 143-175. https://doi.org/10.1515/cllt-2017-0031

Laippala, Veronika, Rönnqvist, Samuel, Oinonen, Miika, Kyröläinen, Aki-Juhani, Salmela, Anna, Biber, Douglas, Egbert, Jesse, Pyysalo, Sampo. 2023. “Register identification from the unrestricted open Web using the Corpus of Online Registers of English”. Lang

Resources & Evaluation, 57, 1045–1079.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. “Towards better structured and less noisy Web data: Oscar with Register annotations”. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. “Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures”. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019 (pp. 9 – 16). Leibniz-Institut für Deutsche Sprache.

Välimäki, Reima and Aali, Heta. 2022. “The Ancient Finnish Kings and their Swedish Archenemy: Nationalism, Conspiracy Theories, and Alt-Right Memes in Finnish Online Medievalism”. In Studies in Medievalism XXXI: Politics and Medievalism (Studies) III edited by Karl Fugelso, Elizabeth Emery, Valerie B. Johnson, M. J. Toswell, Kevin J Harty, Jacob Doss, Helen Young, Heta Aali, Reima Välimäki, Grace Khuri, Whitney Whitney Leeson, Gregory Halfond, Richard Utz, Ana Maria Machado and Angélica Varandas, 55-78. Boydell and Brewer: Boydell and Brewer. https://doi.org/10.1515/9781800105744-010

Zhang, Jinbin, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen and Rohit Babbar. 2022. “Detecting Sequential Genre Change in Eighteenth-Century Texts”. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 243-255). (CEUR Workshop Proceedings; Vol. 3290). CEUR-WS.org.

Vanhat referenssit

Esimerkki:
Douglass, Bruce P., David Harel, and Mark B. Trakhtenbrot. 1998. “Statecarts in use: structured analysis and object-orientation.” In Lectures on Embedded Systems, edited by Grzegorz Rozenberg and Frits W. Vaandrager, 1494:368–394. Lecture Notes in Computer Science. London: Springer-Verlag.

References

Laippala, V., Kyröläinen, A., Kanerva, J. and Ginter, F. (2021) Dependency profiles in the large-scale analysis of discourse connectives . Corpus Linguistics and Linguistic Theory, Vol. 17 (Issue 1), pp. 143-175. https://doi.org/10.1515/cllt-2017-0031

Välimäki, R. & Aali, H. (2022). The Ancient Finnish Kings and their Swedish Archenemy: Nationalism, Conspiracy Theories, and Alt-Right Memes in Finnish Online Medievalism. In K. Fugelso, E. Emery, V. Johnson, M. Toswell, K. Harty, J. Doss, H. Young, H. Aali, R. Välimäki, G. Khuri, W. Whitney Leeson, G. Halfond, R. Utz, A. Machado & A. Varandas (Ed.), Studies in Medievalism XXXI: Politics and Medievalism (Studies) III (pp. 55-78).

Biber, D., & Conrad, S. (2019). Register, Genre, and Style (2nd ed., Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. doi:10.1017/9781108686136

A. Goyal and V. Prem Prakash. “Statistical and Deep Learning Approaches for Literary Genre Classification”. In: Advances in Data and Information Sciences. Ed. by S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. Mishra, and B. K. Singh. Vol. 318. Singapore: Springer Singapore, 2022, pp. 297–305. doi: 10.1007/978-981-16-5689-7_26. url: https://link.springer.com/10.1007/978-981-16-5689-7%5C%5F26.

Zhang, J., Ryan, Y. C., Rastas, I., Ginter, F., Tolonen, M., & Babbar, R. (2022). Detecting Sequential Genre Change in Eighteenth-Century Texts. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 243-255). (CEUR Workshop Proceedings; Vol. 3290). CEUR-WS.org.

Biber, D. & Egbert, J. (2018). Register variation online. Register Studies. 2. 166-171. 10.1075/rs.19018.smi.

Laippala, V., Rönnqvist, S., Oinonen, M., Kyröläinen, A.- J., Salmela, A., Biber, D., Egbert, J., Pyysalo, S. .(2023). Register identification from the unrestricted open Web using the Corpus of Online Registers of English. Lang Resources & Evaluation, 57, 1045–1079.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić (2023). Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo (2022). Towards better structured and less noisy Web data: Oscar with Register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019 (pp. 9 – 16). Leibniz-Institut für Deutsche Sprache.