Modeling Prototypicality and Uncertainty in Genre Concepts

Schroeter, Julian
julian.schroeter@lmu.de
Ludwig-Maximilians-Universität München

Perspectival Modeling (Underwood 2016, 2019a, 2019b, 2020) is among the most powerful methods for investigating the literary change of genres based on machine learning. Concerning the semantic change of loosely ordered genres, further challenges for perspectival modeling arise. This poster provides a new technique for modeling the prototypicality (Rosch 1973, 1975, 1978, Taylor 2007; Hempfer 2010) of genre concepts. Modeling prototypicality in this way is a principled strategy for accessing the vagueness of conceptual boundaries.

The model uses a so-called c@1 score (Peñas / Rodrigo 2011), a variant of the accuracy score accounting for undecidability. Similar to the implementation from the Authorship verification task 2021[1] and based on grid search, the boundaries of undecidability that lead to optimal c@1-accuracy for prediction are calculated. This strategy is integrated into the extraction of predictive probabilities in logistic regression models for genre classification within perspectival modeling as it has been developed by Underwood (2019b). These results are presented within different types of established and new visualizations. Furthermore, the results can be used for a more principled way of expressing conceptual vagueness in quantitative terms. While it uses two existing methods with perspectival modeling and c@1-score, the model presented in the poster is new to computational literary studies. Its strength lies in its dual function of making conceptual looseness interpretable in the way it has been conceived in literary theory and semantics from the 1960s on.

The analyses and visualizations are based on a large corpus of more than 700 German mid-length prose fiction from the 19th century. This corpus has been constructed since 2018 and it can be accessed on GitHub.[2] The field of German novellas includes different genres such as the › Novelle ‹ or › Erzählung ‹ is a field of genres with conceptual boundaries that are rather controversial (Meyer 1987, 1998, Lukas 1998). Hence, the novella corpus provides an appropriate field for developing a model of prototypicality. Fig. 1 shows the core results to be presented on the poster.

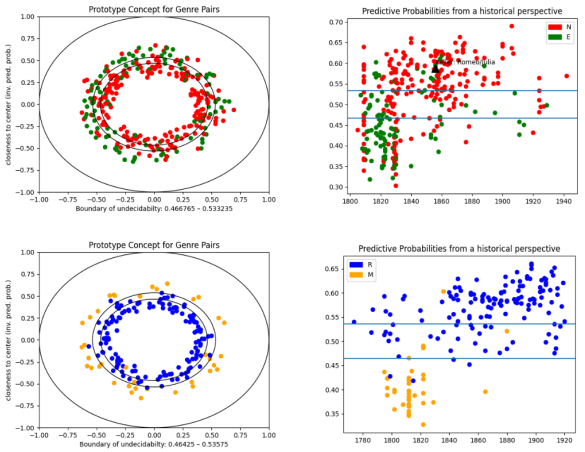

Figure 1: For ›Novellen‹ (N) versus ›Erzählungen‹ (E): Prototypicality based on inverse predictive probability including c@1-boundary of undecidability (upper left) and including the year of first publication (upper right) showing also the position of G. Kellers novella »Romeo und Julia auf dem Dorfe«, published as a ›Erzählung‹ but canonized later as a ›Novelle‹; the c@1-score for the optimal boundary of undecidability is 0.70 for ›Novellen‹ versus ›Erzählungen‹. Novels (R: ›Romane‹) versus fairy tales (M: ›Märchen‹) Prototypicality (lower left) and predictive probability covering temporal index (lower right). The predictive probability is calculated as the average predictive probability for 1000 iterated predictions after resampling. The c@1-score for the optimal boundary of undecidability is 0.89 for novels versus fairy tales.

The upper and lower left plots approximate a popular visualization of prototypicality according to the prototype-as-exemplar approach (Taylor 2007, Hempfer 2010). The boundary of undecidability provides an approximation to the conceptual undecidability that can be assessed by comparing different genre pairs. The visualizations on the right include a temporal index and provide additional information on decreasing or increasing consolidation of genre semantics over time. Although both genre pairs have similar optimal boundaries, the boundary for novels vs. fairy tales generates clearcut zones (novels, fairy tales, and a region of undecidability), which is reflected by a very good c@1-accuracy score of .89 for novels vs. tales. In contrast, the picture remains more ambiguous for the distinctiveness between Novellen and Erzählungen , which is reflected by a c@1 score of .70 for the optimal boundary of undecidability. Further visualizations on the poster will illustrate in detail the methods behind the model: A technique of bootstrapping based on randomized resampling (n= 1000) was used. To guarantee robustness, the location of each text regarding its predictive probability for genre classification averages over 1000 predictions independently of the training samples. The classification is based on 500 words selected by RFECV.[3] However, the best techniques of feature selection, bootstrapping, resampling, choice of algorithms, and parameter optimization may be controversial and can be discussed based on the information to be complimented. Finally, different possible criteria for assigning genre labels such as first or later publication or categorization in current literary studies (Schröter 2019, Underwood 2019b, Calvo Tello 2021) can be discussed based on additional tabular information. The poster is intended to encourage scholars interested in the historicity of literary cultures to discuss the ›next generation‹ of modeling semantic change of literary genres.

Notes

1. https://github.com/pan-webis-de/pan-code/tree/master/clef20/authorship-verification
2. https://github.com/julianschroeter/19CproseCorpus
3. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html . The pipeline is implemented at https://github.com/julianschroeter/PyNovellaHistory .

Bibliography

Calvo Tello, José (2021): *The Novel in the Spanish Silver Age. A Digital Analysis of Genre Using Machine Learning*. Bielefeld: Bielefeld University Press.

Hempfer, Klaus W. (2010): Zum begrifflichen Status der Gattungsbegriffe: Von ›Klassen‹ zu ›Familienähnlichkeiten‹ und ›Prototypen‹. *Zeitschrift für französische Sprache und Literatur* 120: 14–32.

Lukas, Wolfgang (1998): Novellistik. In *Zwischen Restauration und Revolution 1815–1848*. In Gerd Sautermeister (ed.): Hansers Sozialgeschichte der deutschen Literatur 5. München: 251–80.

Meyer, Reinhart (1987): *Novelle und Journal, I: Titel und Normen: Untersuchungen zur Terminologie der Journalprosa, zu ihren Tendenzen, Verhältnissen und Bedingungen*. Stuttgart: Steiner.

Meyer, Reinhart 1998. Novelle und Journal. In *Zwischen Restauration und Revolution 1815–1848*. In Gerd Sautermeister (ed.): Hansers Sozialgeschichte der deutschen Literatur 5. München: Hanser: 234–50.

Peñas, Anselmo / **Alvaro Rodrigo** (2011): A Simple Measure to Assess Non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1415–24.

Rosch, Eleanor (1973): On the internal structure of perceptual and semantic categories. In Timothy E. Moore (ed): *Cognitive Development and the Acquisition of Language*: New York: Academic Press: 111–44.

Meyer, Reinhart (1975): Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104 (3): 192–233.

Meyer, Reinhart (1978): Principles of Categorization. In *Cognition and Categorization*, 27–48. Hillsdale (NJ).

Schröter, Julian (2019): Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen. *Journal of Literary Theory* 13 (2): 227–57.

Taylor, John R. (2007): *Linguistic categorization*. 3. ed., repr. Oxford textbooks in linguistics. Oxford: Oxford Univ. Press.

Underwood, Ted (2016): The Life-Cycle of Genres. *Journal of Cultural Analytics* 1.

Underwood, Ted (2019a): Algorithmic Modeling. In *Fotis Jannidis, J. Flanders (eds.): The Shape of Data in Digital Humanities: Modeling Texts and Text Based Resources*, 250–63. London: Routledge.

Underwood, Ted (2019b): *Distant Horizons. Digital Evidence and Literary Change*. Chicago, London: The University of Chicago Press.

Underwood, Ted (2020): Machine Learning and Human Perspective. *PMLA* 135 (1): 92–109.