Classification of Text-Types in German Novels

Daniel Schlör (daniel.schloer@informatik.uni-wuerzburg.de), University of Wuerzburg, Germany und Christof Schöch (schoech@uni-trier.de), University of Trier, Germany und Andreas Hotho (hotho@informatik.uni-wuerzburg.de), University of Wuerzburg, Germany

1. Introduction

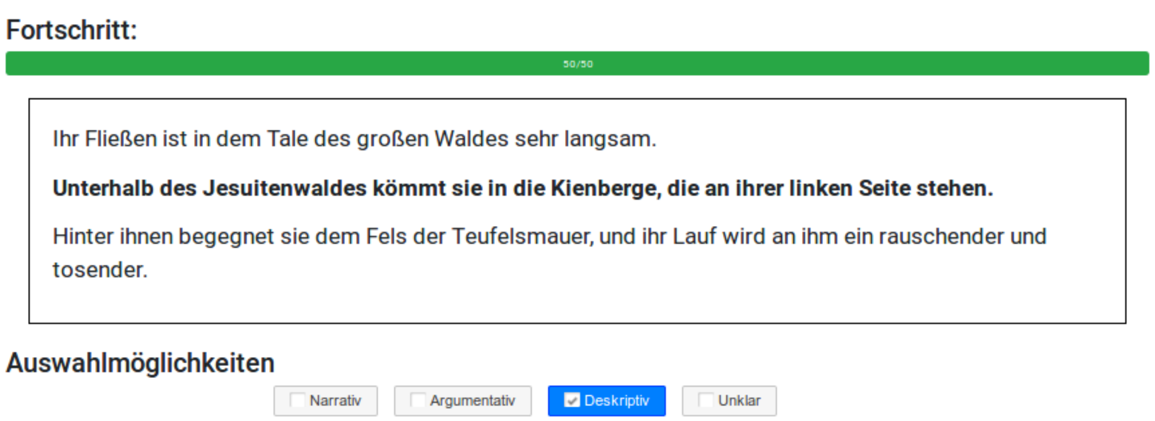
When working with literary texts, a problem for linguists, literary scholars and for machine-based text understanding is the classification of text-types. The term "text-type" refers to a variety of different phenomena reaching from a superordinate view of genre (Chatman 1990) to functionally motivated text-types as aggregations of structural or linguistic features (Biber 1988, 1989). While the taxonomy, textual layer and functionality of different theories behind text-types may differ widely, a common concept behind these theories is the understanding of textual surface structures varying in their respective text-type (Fludernik 2000). The text-types "descriptive", "narrative" and "argumentative" emerge very frequently in many theories (Werlich 1975, Adam 1985, Chatman 1990). Being able to automatically assign sentences to these text-types is therefore highly desirable when aiming to support quantitative literary studies. In this work we present our text-type dataset, a feature based machine-learning model and a deep-learning based model and show that both are able to classify text-types.

2. Annotation Scheme

Functionalizing the existing theories to a respective abstraction of surface phenomena, we came to our annotation guidelines:

- descriptive: the description of a physical object in its dimensions, parts, and/or properties
- narrative: the representation of a chain of events, actions or activities in its temporal progress driven by persons or other "actors"
- argumentative: the presentation, explanation and justification of an abstract idea in its logical context

To consider possible shortcomings of these guidelines or the underlying model and to give the annotators an opportunity to indicate their uncertainty, we introduced additionally the label "unknown". In our annotation tool (see figure 1), sentences were presented within a small contextual framing.



Fortschritt:

50/50

Ihr Fließen ist in dem Tale des großen Waldes sehr langsam.

Unterhalb des Jesuitenwaldes kömmt sie in die Kienberge, die an ihrer linken Seite stehen.

Hinter ihnen begegnet sie dem Fels der Teufelsmauer, und ihr Lauf wird an ihm ein rauschender und tosender.

Auswahlmöglichkeiten

Figure 1: Contextual frame for the sentence (bold) to be annotated in our annotation tool

3. Dataset

We chose a random subset of 30 novels of the DROC corpus (Krug et al. 2017) for our experiments. From each novel we extracted a continuous segment of about 1% of the length. Each of the 1773 sentence obtained in this way was annotated by 3 annotators in order to judge the complexity of the annotation task and to identify a subset of sentences annotated with highly reliable text-type labels. Especially the argumentative text-type seemed to be difficult to annotate since the annotators often disagreed or showed uncertainty on the text-type e.g. for exclamations or other speech acts.

When demanding full agreement among the annotators, the number of instances is reduced from 1773 to 830 sentences denoted as D S (218 descriptive, 352 narrative, 260 argumentative). A majority vote (i.e. the consent of at least two of three annotators) leads to 1503 labeled instances denoted as D M (366 descriptive, 540 narrative, 597 argumentative).

4. Experimental Setup

4.1. Feature Construction

For the task of classifying text-types, we modeled several features targeting different surface levels:

Bag-of-Words . As a baseline, we computed a simple count-based feature vector, representing how often which words occur per instance.
Indicator-Words. Especially the *argumentative* text-type has frequent discourse-related indicator words. We used discourse particles, modal particles, interjections and punctuation marks.
Word-Vectors. In contrast to sparse bag-of-words vectors, dense word-vectors like Word2Vec (Mikolov et al. 2013) comprise semantic information. We used FastText (Mikolov et al. 2017) since the character-based model is able to model compound words implicitly, which are very common in German. After unsupervised training of 100 dimensional [1] word-vectors on the complete DROC corpus, a TSNE-based visualization of the vectors (Maaten & Hinton 2008) revealed a noticeable cluster of words of the *descriptive* text-type ranked via SD2-Zeta (Schöch et al. 2018) (see figure 2). Therefore, we modeled FastText based vector-semantics in two ways: averaging the word-vectors over all words within an instance and counting cluster-membership for words using a previously trained k-means (k=5) model.

Germa-Net . We used GermaNet (German WordNet) to generalize words in two ways: First, we followed each path towards to the most general hypernym and selected different hypernym levels as degrees of generalization. Second, we used GermaNets categorical structure for abstraction. Besides its taxonomic structure, GermaNet classifies words into 54 categories, e.g. the verb ‘sagen’ (engl. to say) belongs to the category ‘verb / communication’. We use these categories as generalizations for each word to reflect e.g. descriptions of places (‘adjective / place’) or people (‘adjective / body’).

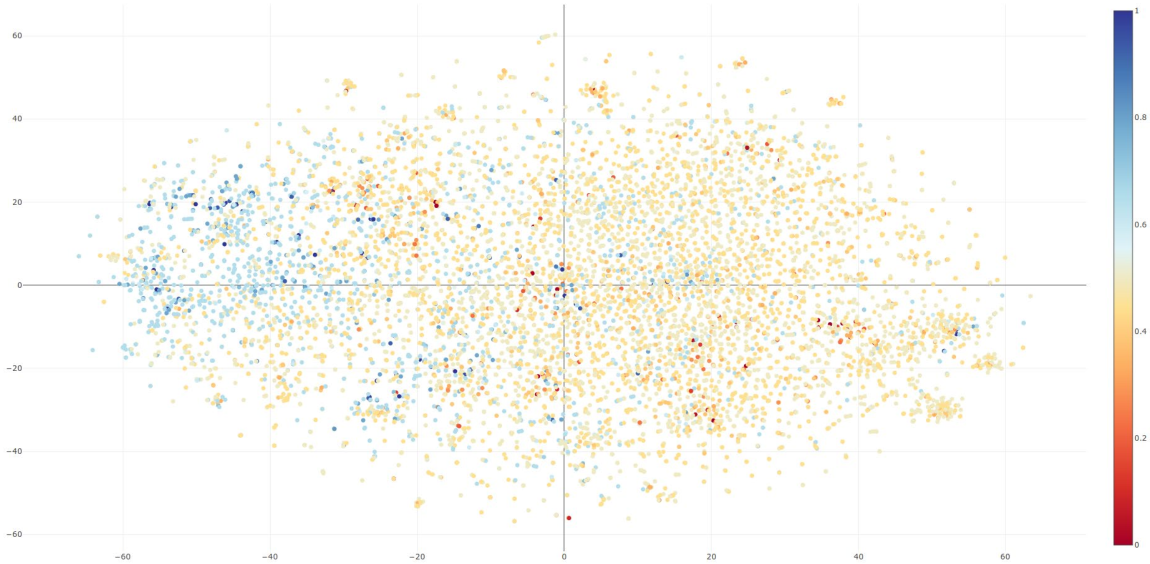


Figure 2: TSNE-Visualization of FastText word-vectors. Blue color indicates that a word is more prominent for the descriptive text-type. The indicativeness was judged opposing the descriptive and other text-types via SD2 Zeta. Other text-types show similar results.

4.2. Classification Task

To examine the contribution of each feature, we conducted an extensive feature analysis using a Support Vector Machine (SVM) [2] as a classifier and compared its performance to a deep-learning based Recurrent Neural Network (RNN) Model.

The SVM was used with linear and RBF kernels, varying the hyperparameter C [3] . Additionally we also evaluated the influence of the degree of generalization using GermaNet on the classification performance.

For the RNN, we adopted the BiGRU model from (Song et al. 2017) but introduced additional loss functions and model parameters as follows: We varied the number of GRU layers between 1 and 4 and the dimensions of the hidden layers between 100 and 400. We used the pretrained FastText model as initial embedding weights.

We conducted a grid-search based parameter study for both classifiers to find the best model-parameter configuration. The model architecture is depicted in figure 3.

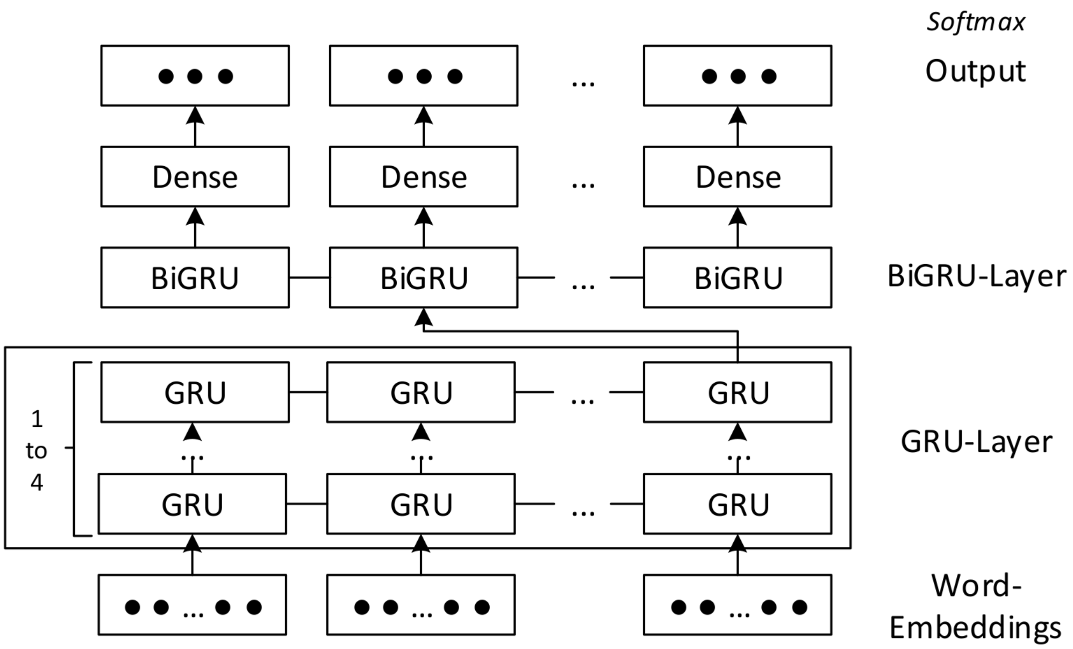
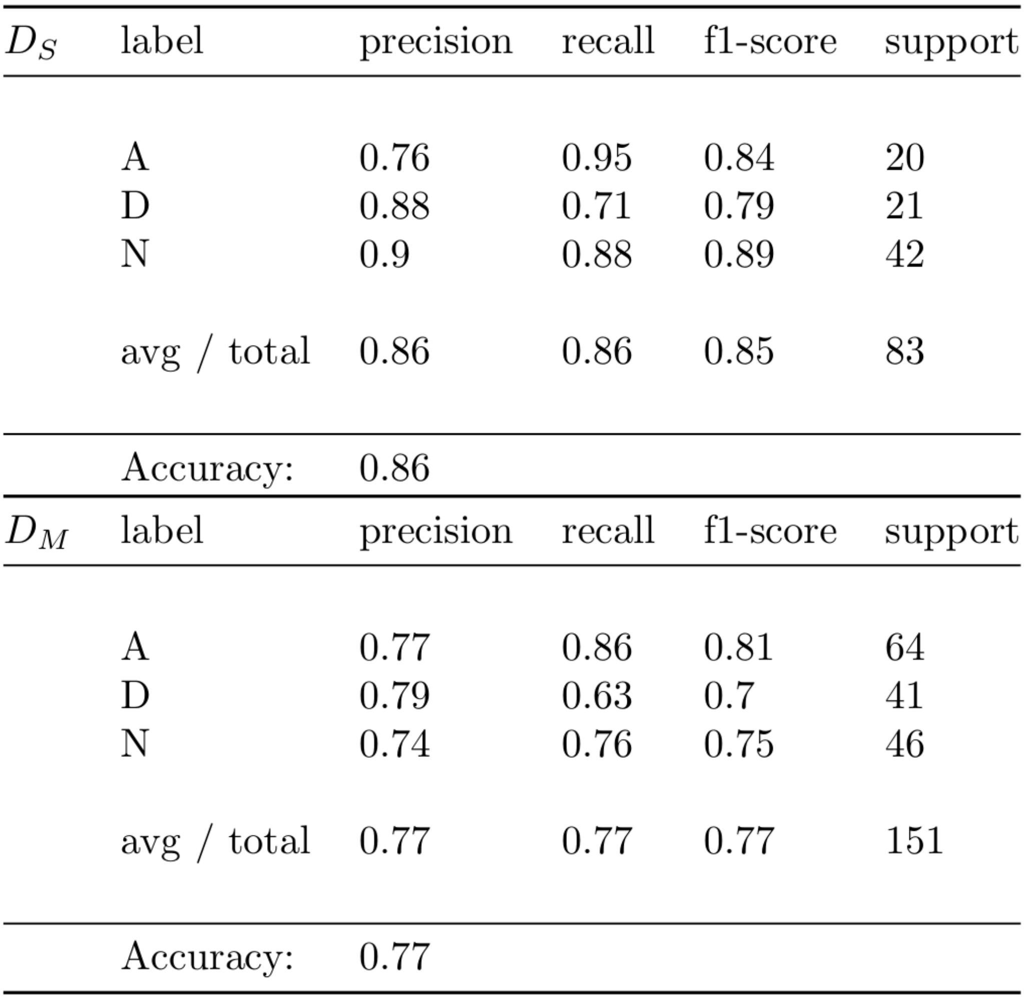


Figure 3: BiRNN Model. Number of GRU Layers were varied between one and four in a parameter-study. Categorical-cross-entropy as loss and softmax as activation achieved best results.

Table 1: Precision, recall, f1-score and accuracy for the SVM classifier on a random hold-out dataset from D S and D M



D_S	label	precision	recall	f1-score	support
	A	0.76	0.95	0.84	20
	D	0.88	0.71	0.79	21
	N	0.9	0.88	0.89	42
	avg / total	0.86	0.86	0.85	83
	Accuracy:	0.86			
D_M	label	precision	recall	f1-score	support
	A	0.77	0.86	0.81	64
	D	0.79	0.63	0.7	41
	N	0.74	0.76	0.75	46
	avg / total	0.77	0.77	0.77	151
	Accuracy:	0.77			

5. Results

We evaluate the models on sampled hold-out dataset fractions of 10% and report accuracy for both datasets, D S and D M . The majority baseline (always predicting the most frequent class) yields an accuracy of 0. 457 for D M and 0.398 for D S .

The best performance of 0.864 mean [4] accuracy ± 0.047 standard deviation for D s resp. 0.776 ± 0.022 for D M was achieved by the linear SVM classifier, C = 1, only using the average FastText word vectors as feature (see table 1 for a label-wise evaluation). This result supports our finding of rather distinct text-type word-vector clusters in our explorative analysis. However, we surprisingly find that additional features do not improve the result. The best feature combinations without averaged FastText vectors yield significantly [5] lower accuracies (0.813 resp. 0.735) and use the FastText cluster features. The best features without any FastText use bag-of-words and part-of-speech features for D M (0.696) and additionally GermaNet hypernym and category features for D S (0.751).

The generalization-depth study aimed at finding the optimal degree of lexical generalization between the word itself and a very abstract and non-indicative root-hypernym. We therefore followed the hypernym path from each word to each root-hypernym and selected the highest and the lowest three hypernym-levels in a classification setup only using this feature. Our results indicate that the second hypernym (i.e. the hypernym of the hypernym of the word) is the most promising level of abstraction for our task. However, we are aware of the problem that hypernym relations for different words might each represent different degrees of abstraction, depending on the level of detail at which the taxonomy is modeled.

The RNN model achieved significantly [6] lower accuracies for all model variations in comparison to the SVM. We believe that this is mainly a problem of too few training data since in theory, the RNN should be able to model the best performing feature, the averaged FastText vectors. In contrast to (Song et al. 2017), our best performing model used categorical-cross-entropy as loss function and one GRU layer with a dimensionality of 400 on both datasets reaching a mean [7] accuracy of 0.801 ± 0.049 standard deviation for D S and 0.702 ± 0.029 for D M .

[8] .

All detailed results, including the results for our feature study are also available on GitHub

6. Discussion and Future Work

Our results show that an SVM as well as a deep-learning approach are able to classify text-types with an accuracy far beyond the baseline. To some extent, handcrafted features are able to compensate the small amount of training data for the D S dataset with an SVM and outperformed the best deep-learning based model. Since the performance of deep-learning based models heavily relies on a sufficient amount of training data, this outcome isn't very surprising and might be revised if more data becomes available. In comparison, the feature driven linear SVM classifier might also have an advantage when it comes to interpretability: The coefficient-weights of an SVM classifier can be interpreted as a whitebox model (Zehe et al. 2017) and reveal interpretable insights into the decision process, whereas the decisions made by a neural network cannot easily be inspected, which is crucial for theory construction and deconstruction in the (digital) humanities.

For future work, we plan to examine in detail why the handcrafted features such as indicator-words or WordNet abstractions don't seem to be as useful as expected. We also plan to incorporate data augmentation methods and finally do a consolidating annotation run to have a bigger and cleaner text-type dataset.

Appendix A

Bibliography

1. Adam, Jean-Michel, (1985). "Quels types de textes?" Le français dans le Monde 192, 39-43.
2. Biber, Douglas, (1988). Variation Across Speech and Writing. Cambridge University Press, Cambridge.
   Biber, Douglas, (1989). A typology of English texts. Linguistics 27, 3–43.
   Chatman, S. (1990). Coming to Terms: The Rhetoric of Narrative in Fiction and Film. Ithaca, NY: Cornell University Press.
3. Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1383-1392).
4. Fludernik, Monika. (2000). "Genres, text types, or discourse modes? Narrative modalities and generic categorization." *Style* 34.2, 274-292.
5. Krug M., Puppe F., Reger I., Weimer L., Macharowsky L., Feldhaus S. & Jannidis F. (2018) Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers Nr. 27. Göttingen: DARIAH-DE
6. Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research* , *9* (Nov), 2579-2605.
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
8. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed wordrepresentations. *arXiv preprint arXiv:1712.09405* .
9. Schöch C., Schlör D., Zehe A., Gebhard H., Becker M &, Hotho A. (2018). Burrows' Zeta: Exploring and Evaluating Variants and Parameters. DH 2018: 274-277.
10. Song, W., Wang, D., Fu, R., Liu, L., Liu, T., & Hu, G. (2017). Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 112-122).
11. Werlich, Egon. (1975). Typologie der Texte: Entwurf eines textlinguistischen Modells zur
12. Grundlegung einer Textgrammatik. Heidelberg: Quelle & Meyer.
13. Zehe A., Schlör D., Henny-Krahmer U., Becker M. & Hotho A. (2018). A White-Box Model for Detecting Author Nationality by Linguistic Differences in Spanish Novels. DH 2018: 519-521

Notes

[1] Using 100 dimensions proved suitable in a initial experiment.

[2] Random Forest classifier achieved slightly worse to similar results.

[3] C ∈{1,10,100,1000}

[4] For 20 repetitions on different hold out-sets

[5] For 20 repetitions on different hold out-sets, α = 0.05 using Pitman's permutation test as suggested by (Dror et al. 2018)

[6] For 10 repetitions on different hold out-sets, α = 0.05 using Pitman's permutation test.

[7] For 10 repetitions on different hold out-sets.

[8] https://github.com/cligs/projects2019/DH_TextTypes