Comparison of Methods for the Identification of Main Characters in German Novels

Markus Krug, Fotis Jannidis, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Frank Puppe
University of Würzburg

1. Motivation

Digital literary studies have embraced social network analysis as a powerful tool to understand, to formalize and to analyse social networks in literary texts [Elson et al. 2010b, Hettinger et al. 2015]. Extracting networks automatically from texts is still a challenging task with the following steps: identification of all references to characters (which is not the same as named entity recognition), coreference resolution followed by a final step in which the amount of interaction between the characters is defined, for example by the amount of verbal exchanges or the co-occurrence in a text segment. In the following we will discuss different ways to solve this task using an annotated corpus of German novels. One of the related problems is the definition of an evaluation metric which connects the computational problem to literary concepts like “main characters” and “character constellation”. Our goal is to find the best way to capture the intuition behind these literary concepts in a formalized procedure.

2. Related Work

Social Network Analysis (SNA) is a well-established discipline, e.g. in the social sciences, which literary studies can apply for the analysis of character networks [Trilcke 2013]. Approaches to automatic extraction of social networks from literary text using Natural Language Processing techniques have been manifold.

Most works start by identifying entities in the text and connect them via Coreference Resolution. Park et al. [Park et al. 2013] extract social networks on the basis of proximity of names in the text and define a kernel function to distinguish protagonists from less important characters. Celikyilmaz et al. [Celikyilmaz et al. 2010] use an actor-topic-model trained by unsupervised learning to create social networks from narratives. Elson, Dames and McKeown associate speakers with direct speech passages in novels [Elson et al. 2010a] and create social networks from the dialogues to validate literary hypotheses like whether the amount of dialogues is inversely proportional to the amount of characters that appear in the novel [Elson et al. 2010b].

Moreover, two end-to-end systems for the extraction and visualization of social networks from English literary texts already exist: PLOS [Waumans et al. 2015] works similarly to the approach by Elson, Dames and McKeown by creating networks from dialogue interactions. In subsequent analysis they showed, for example, that the node degree of a character follows an exponential distribution. He et al. use their own speaker identification system to detect family connections between entities [He et al. 2013]. SINNET by Agarwal et al. [Agarwal et al. 2013b] finds different types of directed events in a text which either both entities or only one entity can be aware of. On the basis of these events, a directed social network can be created.

3. Data

For this work, we have a data basis of 452 German novels from the TextGrid Digital Library[1]. Useful plot summaries from Kindlers Literatur Lexikon Online[2] are available for 215 of these novels. As the following experiments are partly based on direct speech, we have analysed the novels with regard to the direct speech they contain. We selected 58 novels with the highest possible amount of direct speech with as little errors as possible for which there was also a summary on hand.

Those 58 novels have been splitted into tokens and sentences with OpenNLP[3], POS-tagged and lemmatized by the TreeTagger [Schmid 1995] and further processed by the RFTagger [Schmid, Laws 2008] and the morphological tagger from MATE-Tools[4]. Additionally, we use the dependency parser by Bohnet [Bohnet, Kuhn 2012] to analyse the sentence structure. Named Entity Recognition is done with the tool by Jannidis et al. [Jannidis et al. 2015] and the rule-based component by Krug et al. [Krug et al. 2015] is used for Coreference Resolution. The detection of the speaker and the addressee for each direct speech passage is also part of the Coreference Resolution [Krug et al. 2015]. In the summaries from Kindler, Named Entities and Coreferences have been manually labeled by two annotators.

4. Methods

We use four different methods to identify the most central characters in the novels and evaluate their quality by comparison with the characters occurring in the summaries from Kindler.

The first method relies only on the frequencies of the characters in the text. In this approach, the most central characters are those appearing most often in the novel (coreferences resolved). The second methods counts only those entities that have at least once been detected as speaker or addressee of direct speech. The other methods each construct a different type of social network and make use of SNA to find the most central characters.

The first network is based on co-occurrences of characters in the same window of text: an edge between two characters exists if they are mentioned in the same paragraph and the weight of the edge is the number of paragraphs in which this is the case. The second network is created using the dialogue structure of the text. For each direct speech for which both speaker and addressee could be detected, an edge is drawn between those two. Longer dialogues consequently lead to higher edge weights between the participants. Thus, both network types are undirected and weighted. Examples for networks that were created with those methods are shown in Figure 1

To identify the most central characters we use the weighted degree of each node (i.e. the sum of the weights of all edges incident to a node) in decreasing order.

In the following paragraph, we compare the rankings with the summaries and discuss possible sources of error and their influence on the results.

[1] https://textgrid.de/digitale-bibliothek

[2] http://kll-online.de

[3] https://opennlp.apache.org/

[4] https://code.google.com/p/mate-tools/

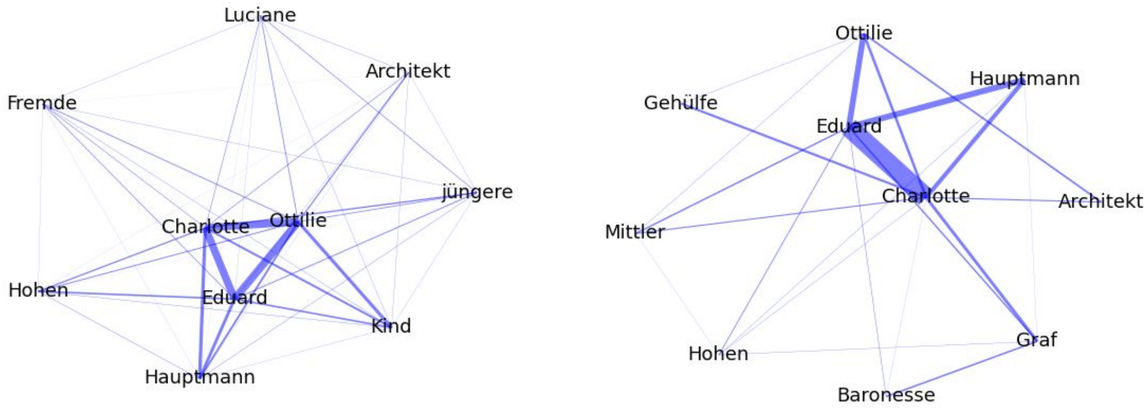

Figure 1: Automatically extracted social networks for Goethes: “Die Wahlverwandtschaften”. The left picture shows the ten most connected characters when an interaction is created for a common appearance in a paragraph. The right picture shows the corresponding network, that is created when only direct speech is used as interactions.

5. Evaluation

Evaluating automatically extracted social networks is not a trivial task and there are no established practices. Elson et al. [Elson et al. 2010b] validate literary hypotheses, [Park et al. 2013] and [Waumans et al. 2015] analyse typical distributions that they also expect of literary character networks. Agarwal et al. [Agarwal et al. 2013a] evaluate a machine-generated network of *Alice in Wonderland* against a manually conceived version by comparing typical SNA metrics like different centrality measures.

In this work, we want to compare the methods for identifying the most central characters as described in section 4. As a gold standard, we use the manually annotated Kindler summaries. The generated rankings for each novel, as well as the rankings from the summaries are first cleaned up: all mentions of a character known to the morphological tool SMOR [Schmid 2004], as well as titles and other references, are deleted so that (almost) only real names are left.

Our evaluation is based on the assumption that a summary contains all important characters. For each summary, we create a ranking of the mentioned characters by [a] the number of occurrences (gold_count from here) and [b] the order of occurrence (gold_order from here). As it cannot be guaranteed that such rankings represent the underlying structure of a novel, we select the top 5 (top 10) figures from the summary rankings and compare them against the top 5 (top 10) characters in the automatically obtained rankings for the novels. If the name of a character from the gold standard is exactly found in an automatic ranking, there is a match. Table 1 shows the resulting correspondences with the two gold rankings, averaged over all 58 novels.

Algorithm	DSN_5	DSN_10	PN_5	PN_10	DSC_5	DSC_10	Count_5	Count_10
gold_count	**40.5%**	50.2%	39.3%	51.6%	38.9%	49.0%	40.1%	**52.0%**
gold_order	38.6%	45.1%	**41.3%**	**48.6%**	37.5%	45.3%	41.2%	48.5%

Table 1: Overview of the successfully matched entities between the two rankings from the summaries (gold_count, gold_order) and the generated rankings for the top 5 and the top 10 entities (DSN= Direct Speech Network; PN = Paragraph Network;DSC = Direct Speech Count; Count = simple frequency)

6. Results and Discussion

Table 1 displays first results for the automatic extraction of meaningful social networks. Nevertheless, none of the methods yields very high scores for this kind of evaluation. Interestingly, the simpler approaches, namely mere counting and the networks based on co-occurrences in paragraph seem to be suited well for the task.
The low values can be explained by a variety of errors which can be grouped in three categories. Firstly, a character might not be among the top 10 of the ranking. If automatic matches to lower positions in the ranking are allowed, the score in Table 2 can be reached.

Algorithm	DSN_Max	PN_Max	DSC_Max	Count_Max
gold_count	55.1%	56.6%	53.8%	**57.3%**
gold_order	58.0%	**64.7%**	55.1%	**64.7%**

Table 2: Accuracy of the matching, independent of the position in the automatic ranking

We can see that approximately 60% of the characters can now be matched unambiguously. The highest percentage of errors is due to incorrectly resolved coreferences. Clusters of the same characters that have not been merged during the Coreference Resolution do not only create redundant elements in the rankings, wrongly merged clusters also mean, that one character can never be matched correctly. If coreference errors are ignored, the results are as shown in table 3.

Algorithm	DSN_Maxcr	PN_Maxcr	DSC_Maxcr	Count_Maxcr
gold_count	79.7%	**81.2%**	78.8%	**81.2%**
gold_order	58.6%	**65.3%**	55.6%	**65.3%**

Table 3: Accuracy of the matching, independent of the position in the automatic ranking, coreference resolution errors ignored

The third type of errors originates from different spellings of the same name which make an unambiguous matching very difficult (e.g. “Amanzéi” vs. “Amanzei”, “Lenore” vs. “Leonore”). Further reasons which render the matching more difficult or impossible respectively are missing or incorrectly detected Named Entities. The error analysis shows that future improvements are especially needed for the Coreference Resolution or procedures which avoid CR have a better chance to succeed.

7. Conclusion

In this paper we showed work in progress to extract social networks from German novels. We compared four different approaches to the distinction of central figures against manually annotated summaries. At least for this task, the more challenging approaches of determining speaker and addressee of direct speech and creating networks from the resulting interactions did score slightly lower than the more simpler approaches. To improve the results, future work especially needs to be invested into the creation of a less error-prone coreference resolution system.

8. References

Agarwal, Apoorv, Anup Kotalwar, and Owen Rambow. "Automatic extraction of social networks from literary text: A case study on alice in wonderland." *the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*. 2013a.

Agarwal, Apoorv, et al. "Sinnet: Social interaction network extractor from text." *Sixth International Joint Conference on Natural Language Processing*. 2013b.

Ardanuy, Mariona Coll, and Caroline Sporleder. "Structure-based clustering of novels." *EACL 2014* (2014): 31-39.

Bohnet, Bernd, and Jonas Kuhn. "The best of both worlds: a graph-based completion model for transition-based parsers." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012.

Celikyilmaz, Asli, et al. "The Actor-Topic model for extracting social networks in literary narrative." *NIPS Workshop: Machine Learning for Social Computing*. 2010.

Elson, David K., and Kathleen McKeown. "Automatic Attribution of Quoted Speech in Literary Narrative." *AAAI*. 2010a.

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. "Extracting social networks from literary fiction." *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010b.

Gruzd, Anatoliy A., and Caroline Haythornthwaite. "Automated discovery and analysis of social networks from threaded discussions." (2008).

Hassan, Ahmed, Amjad Abu-Jbara, and Dragomir Radev. "Extracting signed social networks from text." *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 2012.

He, Hua, Denilson Barbosa, and Grzegorz Kondrak. "Identification of Speakers in Novels." *ACL (1)*. 2013.

Hettinger, L.; Becker, M.; Reger, I.; Jannidis, F. & Hotho, A. Genre classification on German novels, *in* '*Proceedings of the 12th International Workshop on Text-based Information Retrieval*' . 2015

Jannidis, Fotis, et al. "Automatische Erkennung von Figuren in deutschsprachigen Romanen." *Conference Presentation at" Digital Humanities im deutschsprachigen Raum*. 2015.

Jing, Hongyan, Nanda Kambhatla, and Salim Roukos. "Extracting social networks and biographical facts from conversational speech transcripts." *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*. Vol. 45. No. 1. 2007.

Krug, Markus, et al. 2015 "Attribuierung direkter Reden in deutschen Romanen" *Digital Humanities im deutschsprachigen Raum 2016,* Manuscript submitted for publication.

Krug, Markus, et al. "Rule-based Coreference Resolution in German Historic Novels." *on Computational Linguistics for Literature* (2015): 98.

Park, Gyeong-Mi, et al. "Complex system analysis of social networks extracted from literary fictions." *International Journal of Machine Learning and Computing* 3.1 (2013): 107-111.

Schmid, Helmut, and Florian Laws. "Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging." *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008.

Schmid, Helmut, Arne Fitschen, and Ulrich Heid. "SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection." *LREC*. 2004.

Schmid, Helmut. "Improvements in part-of-speech tagging with an application to German." *In Proceedings of the ACL SIGDAT-Workshop*. 1995.

Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields for relational learning." *Introduction to statistical relational learning* (2006): 93-128.

Trilcke, Peer. "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft." *Philip Ajouri/Katja Mellmann/Christoph Rauen (Hg.), Empirie in der Literaturwissenschaft, Münster* (2013): 201-247.

Waumans, Michaël C., Thibaut Nicodème, and Hugues Bersini. "Topology Analysis of Social Networks Extracted from Literature." *PloS one* 10.6 (2015): e0126470.