A Reporting Tool for Relational Visualization and Analysis of Character Mentions in Literature

Barth, Florian
florian.barth@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Kim, Evgeny
evgeny.kim@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Murr, Sandra
sandra.murr@ilw.uni-stuttgart.de
Institut für Literaturwissenschaft, Universität Stuttgart, Deutschland

Klinger, Roman
roman.klinger@ims.uni-stuttgart.de
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland

Introduction and Motivation

The emergence of computational methods of text processing has created new paradigms of research in literary studies in recent years (Jockers & Underwood, 2016), for instance *distant reading* to find patterns and regularities (Moretti, 2005). Network analysis and extraction of information about relations between characters from literary texts is an example for distant reading methods. Such information can not only be helpful for better understanding of character interactions but can also facilitate the comparison of thereof in different texts.

Existing tools of text analysis and network visualization such as Voyant [1] or Gephi [2] are either missing modules for character network analysis or require preliminary steps on data preprocessing from the user and therefore are not easy-to-use for some humanities scholars who lack programming skills. Interactive tools in addition often lack features to ensure reproducibility of results.

We present our ongoing effort on closing this gap by developing a literary analysis reporting tool *rCAT* [3], whose primary purpose is to provide an easy-to-use, stable, and reusable solution for automatic extraction of relational information from text and to characterize these relationships automatically to provide the user with deeper qualitative insight. We opt for implementation as a web-based reporting tool instead of an interactive tool for two reasons: (1) automatically generated reports in PDF format can serve as a stable foundation for discussion and can be reused in publications and visualizations easily, and (2) the results are clearly connected to the chosen input parameters such that reproducibility of results is ensured.

As a use-case study, we apply *rCAT* to Johann Wolfgang von Goethe's epistolary novel *Die Leiden des jungen Werthers*. On the basis of this epistolary novel, we show that not only the network can be generated, but also the characteristic triangular relationship of the protagonists is easily identified. The goal is to automatically determine this triad in the original text and in the adaptations that have been published since the publication of *Werther* in 1774.

Previous Work

Previous research on social networks in literary fiction generally fall into one of the two categories: (1) works that explore methods for extracting and formalizing character networks (*cf.*, Elson et al. (2010), Agarwal et al. (2012, 2013), Park et al. (2012)), and (2) works that primarily focus on qualitative implications of network analysis (*cf.*, Rydberg-Cox (2011), Moretti (2011), Nalisnick & Baird (2013), Jayannavar et al. (2015)). It is common to address both tasks at the same time, as in Beveridge & Shan (2016), who introduce a number of formal measures for analyzing the centrality of the characters in *Game of Thrones* books, which results in both expected and surprising findings.

Building on graph theory extensively elaborated in the past fifty years (e.g., Bondy and Murty, 1976 or West, 2001), our work is similar to Beveridge & Shan (2016), in particular, in terms of the weighted degree measure, and to Park et al. (2012), in terms of distance measure for detecting closely related characters in a text.

Methods

In the following, we explain the different components in *rCAT*, which are available for text ana-

lysis. After that, we discuss the results based on a use-case study.

Character lists and character identification

To detect character mentions in the text we use a fundamental named-entity recognition approach based on dictionaries. This approach is suitable for scholars who analyze texts they already know. Consequently, we opt for a transparent and simple character recognition procedure: The user provides a list of character names to be included in the analysis specifying a canonical name form and all variations thereof she would like to take into account (*e.g.*, "Lotte" is the canonical name and "Lotten", "Lottens", "Lottgen", "Lottchen", "Charlotten S.". are its variants).

Relation detection and context words

We define the closeness of relationship between two characters using a *distance measure dist X (p,q)*, where *p* and *q* are the strings corresponding to these characters and *X* is the number of tokens between them (Park et al., 2012). In addition, we introduce the *context measure cont Y (p,q)*, where *p* and *q* are the strings corresponding to these characters and *Y* is the number of tokens before the character *p* and after the character *q*. While the former measure allows for detecting those characters that are closely related to each other, the latter one enables a contextual analysis of their relationship.

Network analysis

We visualize the network of characters with an undirected graph *G=(V,E)*, where *V* are the vertices, each vertex corresponding to one character, and each edge *E=(V i, ,V j)* corresponding to relations between pairs of characters. We output the following measures for each character node: *degree*, *edge weight*, *weighted degree* and *density*. The degree is the number of edges occurring with a given vertex. The edge weight, $w_{i,j} \geq 0$, is defined as the number of interactions between the vertices V *i* and V *j*. The weighted degree is the sum of weights of the edges occurring with a vertex *i*. Density is the ratio of occurring edges between two vertices and all possible vertex pairs.

Word clouds

Word clouds are an approach to visualize the vocabulary of a text. The size of one word corresponds to its frequency. We use two different kinds of word clouds: For each character in the character list, we show word clouds based on the context of a window size *n.* For each pair of characters occurring in the network, we present a word cloud based on the words between them as well as on the words found in the context. Both types of word clouds can be filtered to the specific word fields (words from specific domains) which is helpful in gaining a focused insight into the characters relations.

Word Field developments

We plot the timeline of multiple predefined world fields (specified by word lists) in the text. This feature is helpful in representing how certain fields (*e.g.*, concepts, emotions) develop throughout the narrative (Kim et al., 2017).

Implementation

The tool was developed using Python v.3.6 and the Flask [4] web development framework. The tool outputs a single PDF report. The resulting document contains information from the analysis modules described in the previous section. Network graphs included in the report are generated with *graphviz*. Additionally, the tool can generate a CSV file that can be used as input to Gephi.

Use-case Demonstration

For a use-case analysis, we apply *rCAT* to *Die Leiden des jungen Werther* by Johann Wolfgang Goethe with the following parameters: X=8, Y=5, stop words removed (previous work focused on this analysis without rCAT, *cf.* Murr, 2017).

In Goethe's epistolary novel, the protagonist Werther describes his unhappy love for Lotte, who is engaged to Albert. The characteristic triangular relationship in the novel arises from this constellation (protagonist - beloved woman - antagonist). With *rCAT* we expect to identify and characterize this relationship. Figures 1 and 2 show a sample network analysis output (tables are shown only partly).

The protagonist Werther shows a degree of 21, which is the number of characters with whom he interacts. The closest relationship measured by edge weight (Figure 2) is observed between Werther and Lotte (81 interactions). The antagonist Albert has a low degree of 3. However, his weigh-

ted degree is 36 (third highest after Werther and Lotte), which confirms his important role in the triangular relationship.

Degrees

Character (Node)	degree	weighted degree
Werther	21	184
Lotte	12	101
Albert	3	36
Wilhelm	3	34
Vetter	2	3
Magd	1	1
Schreiber	2	2
Hans	1	1
Marianne	1	1
Grafen von M . .	1	1
Graf v. C.	4	17

Illustration 1: Degrees and weighted degrees for most important characters of Goethe's Werther

Weights for Edges

Character Pair (Edge)	Weight
Werther -- Lotte	81
Werther -- Albert	26
Werther -- Wilhelm	32
Werther -- Vetter	2
Werther -- Magd	1
Werther -- Schreiber	1
Werther -- Hans	1
Werther -- Marianne	1
Werther -- Graf v. C.	12

Illustration 2: Edge weights

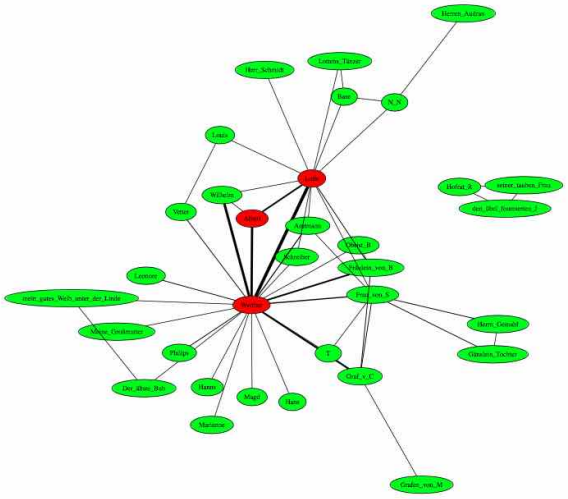

Illustration 3: Complete network of Goethe's Werther

Highlighted in red is the typical triangular relationship in Goethe's novel, which corresponds to the three highest weighted degrees. In further steps, we will use *rCAT* to analyze the adaptations of Goethe's novel with a focus on this triad.

To better characterize the edges, the tool outputs top- *n* word clouds sorted by edge weight (*n* is specified by the user) for character pairs and by degree for single characters. Figure 4 and 5 show examples of the word clouds for character pairs filtered to the words from the emotion domain.

Illustration 4: Word clouds for Werther-Lotte

Illustration 5: Werther-Albert

The word clouds enable first conclusions about the relationships of the characters. Werther and Lotte's word cloud characterizes their ambivalent relationship. The key words "Leidenschaft" and "Freude" reflect Werther's love, whereas the mentions of "sterben" and "Verblendung" are characteristic of the unrequited love, which leads Werther into his "disease unto death". As Werther and Albert's word cloud reveals, their relationship is dominated by the "Unruhe" that Werther feels through his adversary.

Additionally, the tool plots the development of the narrative (not bound to specific characters) based on the word fields, an example of which is shown on Figure 6. In this case we used words from the emotion domain (with emotion dictionaries by Klinger et al. (2016)).

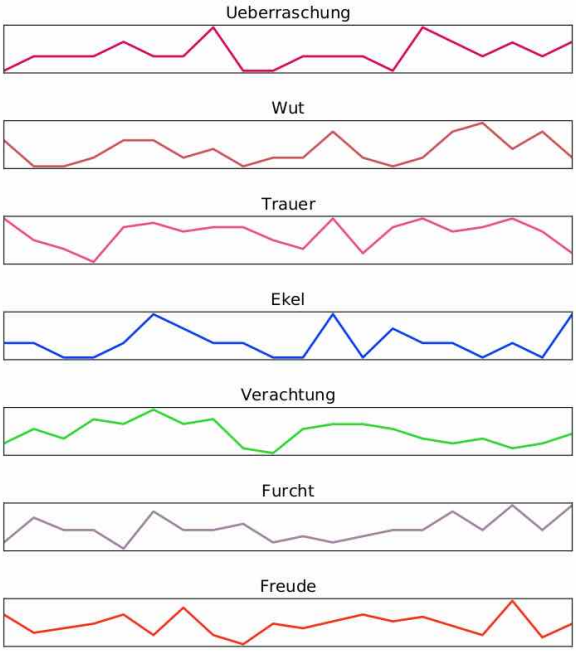

Illustration 6: Word field development for Goethe's Werther

The word field development can highlight the prevalence of individual emotion domains across the text. The accumulation of the negative emotion words (Wut,Trauer, Furcht) towards the end suggests, for example, that Goethe's novel has no "happy ending". The striking rash on "Freude", however, captures the last happy hours Werther spends with Lotte in the second part of the narration before he kills himself.

Future Work

The next version of the tool will include a character-oriented word field development calculated and plotted for the main characters of the stories. In addition, future releases will include more analysis features and bulk file processing.

Fußnoten

1. https://voyant-tools.org/
2. https://gephi.org/
3. *www.ims.uni-stuttgart.de/data/rcat*
4. http://flask.pocoo.org/

Bibliographie

Agarwal, A. / Corvalan, A. / Jensen, J. / Rambow, O. (2012): "Social Network Analysis of Alice in Wonderland", in: CLfL@ NAACL-HLT 88-96.

Agarwal, A. / Kotalwar, A. / Rambow, O. (2013): "Automatic Extraction of Social Networks from Literary Text. A Case Study on Alice in Wonderland", in: IJCNLP 1202-1208.

Beveridge, A. / Shan, J., (2016): "Network of thrones", in: Math Horizons, 23(4): 18-22.

Bondy, J.A. / Murty, U.S.R. (1976): Graph theory with applications (Vol. 290). London: Macmillan.

Burrows, J.F. (1987): "Word-patterns and story-shapes: The statistical analysis of narrative style", in: Literary & Linguistic Computing, 2(2): 61-70.

Elson, D.K. / Dames, N. / McKeown, K.R. (2010): "Extracting social networks from literary fiction", in: Proceedings of the 48th annual meeting of the association for computational linguistics 138-147. Association for Computational Linguistics.

Heuser, R., F. Moretti / E. Steiner (2016): The Emotions of London. Technical report. Stanford University. Pamphlets of the Stanford Literary Lab.

Jayannavar, P. / Agarwal, A. / Ju, M. and Rambow, O. (2015): "Validating Literary Theories Using Automatic Social Network Extraction", in CLfL@ NAACL-HLT 32-41.

Jockers, M.L. / Underwood, T. (2016): "Text-Mining the Humanities", in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): A New Companion to Digital Humanities 291-306.

Kim, E. / Padó, S. / Klinger, R. (2017): "Investigating the Relationship between Literary Genres and Emotional Plot Development", in: Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 17-26.

Klinger, R. / Sulliya S.S. / Reiter N. (2016): "Automatic Emotion Detection for Quantitative Literary Studies -- A Case Study on Kafka's 'Das Schloss' and 'Amerika'", in: Digital Humanities (DH), Conference Abstracts, Kraków, Poland, 2016.

Michel, J.B. / Shen, Y.K. / Aiden, A.P. / Veres, A. / Gray, M.K. / Pickett, J.P. / Hoiberg, D. / Clancy, D. / Norvig, P. / Orwant, J. / Pinker, S. (2011): "Quantitative analysis of culture using millions of digitized books", in: science, 331(6014) 176-182.

Moretti, F. (2005): Graphs, maps, trees: abstract models for a literary history. Verso.

Moretti, F. (2011). Network theory, plot analysis. Stanford Literary Lab Pamphlet Series 2. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf

Murr, S. / Barth, F. (2017): Digital Analysis of the Literary Reception of J.W. v. Goethe's 'Die Leiden des jungen Werthers', in: Digital Humanities (DH), Conference Abstracts, Montreal, Canada 2017.

Nalisnick, E.T. / Baird, H.S. (2013): "Extracting sentiment networks from Shakespeare's plays", in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on IEEE 758-762.

Park, G.M. / Kim, S.H. / Cho, H.G. (2013): "Structural analysis on social network constructed from characters in literature texts", in: Journal of Computers, 8(9): 2442-2447.

Rydberg-Cox, J., (2011): "Social networks and the language of greek tragedy", in: Journal of the Chicago Colloquium on Digital Humanities and Computer Science (Vol. 1, No. 3).

West, D.B. (2001): Introduction to graph theory (Vol. 2). Upper Saddle River: Prentice Hall.