A Shared Task for a Shared Goal: Systematic Annotation of Literary Texts

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
Stuttgart University, Germany

Evelyn Gius
evelyn.gius@uni-hamburg.de
Hamburg University, Germany;

Jannik Strötgen
jannik.stroetgen@mpi-inf.mpg.de
Max Planck Institute for Informatics, Germany

Marcus Willand
marcus.willand@ilw.uni-stuttgart.de
Stuttgart University, Germany

Introduction

In this talk, we would like to outline a proposal for a shared task (ST) in and for the digital humanities. In general, shared tasks are highly productive frameworks for bringing together different researchers/research groups and, if done in a sensible way, foster interdisciplinary collaboration. They have a tradition in natural language processing (NLP) where organizers define research tasks and settings. In order to cope for the specialties of DH research, we propose a ST that works in two phases, with two distinct target audiences and possible participants.

Generally, this setup allows both "sides" of the DH community to bring in what they do best: Humanities scholars focus on conceptual issues, their description and definition. Computer science researchers focus on technical issues and work towards automatisation (cf. Kuhn & Reiter, 2015). The ideal scenario– that both "sides" of DH contribute to the work in both areas– is challenging to achieve in practice. The shared-task scenario takes this into account and encourages Humanities scholars without access to programming "resources" to contribute to the conceptual phase (Phase 1), while software engineers without interest in literature per se can contribute to the automatisation phase (Phase 2). We believe that this setup can actually lower the entry bar for DH research. Decoupling, however, does not imply strict, un-crossable boundaries: There needs to be interaction between the two phases, which is supported by our mixed organisation team. In particular, this setup allows mixed teams to participate in both phases (and it will be interesting to see how they fare).

In Phase 1 of a shared task, participants with a strong understanding of a specific literary phenomenon (literary studies scholars) work on the creation of annotation guidelines. This allows them to bring in their expertise without worrying about the feasibility of automatisation endeavours or struggling with technical issues. We will compare the different annotation guidelines both qualitatively: by having an in-depth discussion during a workshop, and quantitatively: by measuring inter-annotator agreement. This will result in a community guided selection of annotation guidelines for a set of phenomena. The involvement of the research community in this process guarantees that heterogeneous points of view are taken into account.

The guidelines will then enter Phase 2 to actually make annotations on a semi-large scale. These annotations then enter a "classical" shared task as it is established in the NLP community: Various teams competitively contribute systems whose performances will be evaluated in a quantitative manner.

Given the complexity of many phenomena in literature, we expect the automatisation of such annotations to be an interesting challenge from an engineering perspective. On the other hand, it is an excellent opportunity to initiate the development of tools tailored to the detection of specific phenomena that are relevant for computational literary studies.

This talk has two purposes:

- To discuss these ideas and collect feedback and propositions. This is also an explicit invitation to contribute in the setup of this initiative. We are also welcoming a discussion about the phenomena that should be included.
- To advertise the idea of a shared task and to invite possible participants. The success of STs relies on a certain number of participants. Given that this has never been organized in the DH community before, we want to spread this idea throughout the community to

gather estimates of potential participants.

The Importance of Annotations

In computational literary studies, many phenomena cannot directly be detected from the text surface. To find and categorize such phenomena as, for example, the "narrated time" in a novel, it is first necessary to have an in-depth understanding of the text, knowledge about its author or literary conventions, or knowledge of the text's historical context. Therefore, instances of such phenomena need to be annotated either by human experts or software that is tailored to this task.

Unfortunately, many theories describing interesting phenomena are very difficult to apply to real texts. It has been shown numerous times (e.g., Reiter, 2015, Musi et al., 2016) that annotating theories or concepts directly can lead to very poor inter-annotator agreement (IAA): Different annotators have different interpretions of not only the text, but also descriptions of the theoretical concepts. Although subjective annotations have their merit, studying annotations on large scale depends on their consistency, i.e., a high IAA. In addition, many theories are underspecified and provide examples for illustrations only. Creators of annotation guidelines often have to interpret what is meant by a certain statement and extend definitions to cover examples found in real texts.

Annotation guidelines serve as a mediator between the annotators and a theory (that may use specialised vocabulary). Additionally, such guidelines often contain re-appearing instance patterns and their modes of annotation and/or exceptions, as well as many examples from real texts (see below).

We see the creation of annotation guidelines as one of the cornerstones of large scale text analysis in computational literary studies. Additionally, the creation of annotiation guidelines supports systematic disciplinary discussions about concepts and thus may lead to additional findings relevant for the theoretical discourse (e.g., Meister, 1995; Gius and Jacke, forthcoming). Experts from the field literary studies are well-suited to work with annotation guidelines, as annotation of literary phenomena in literary texts can be seen as a special form of close reading.

Phase One: Annotation Guidelines

The Shared Task

In theory, any phenomenon can be addressed in this fashion, as long as it can be defined inter-subjectively, is reasonably frequent, and is of interest in computational literary studies. As a starting point, we propose to address the issue of narrative levels (Pier, 2014). Narrative levels are a core concept in narrative theory (Genette, 1980; Bal, 1997) which in turn has shown to be a promising foundation for automatisation in literary theory (Bögel et al. 2015). The first reason for choosing to examine narrative levels is their ubiquity: every narrative text necessarily consists at least one, most texts contain many narrative levels; each element of a text can be assigned to a specific level. The second reason is our intuition that a definition of "level" in guidelines is as achievable as the automated detection of levels by computers.

Concretely, participating teams are asked to create guidelines for the detection and annotation of a) narrative levels and b) the relation of the narrator(s) to the narrated world (i.e., is the narrator part of the narrated world or not?). Participants are not bound to adhere to a specific narratological theory. The result of this phase, however, will be a fixation on a set of guidelines (that instantiate a theory).

We will select a number of literary narrative texts and provide copyright-free digitized corpora. All "official" texts (development and test sets) will be English literary texts. Naturally, the second step will be to extend this framework to other languages and/or phenomena.

Evaluation

In NLP shared tasks, the predictions of the systems are compared against a fixed test set– the "gold standard". Since there is no gold standard in Phase 1, we will evaluate the guidelines using an unseen data set. Each participating team annotates this set using their own guidelines before the guidelines are submitted. Submitted guidelines will be anonymized and re-distributed among the participants. Each participant is asked to annotate the evaluation data set using two other annotation guidelines. In addition, we will be collecting annotations from students.

The evaluation data set will thus be annotated according to each participant's guidelines four times (1x self, 1x student, 2x other participants).

This setup allows direct calculation of inter-annotator agreement. However, IAA should be only one aspect in evaluating the guidelines, but not the only one. Therefore, we will submit a workshop at DH 2018 to discuss the submissions and select final ones. This setup also allows the merging of different

annotation guidelines as well as adaptation according to the discussion during the workshop. Soon after the workshop, Phase 2 will commence.

Phase Two: Automatic Prediction

In Phase 2 of this endeavour, the selected guidelines of Phase 1 will be annotated on a large scale by student assistants. Since we do not yet know how long, complex and involved the guidelines will be, there should be close communication between the organizers and the team responsible for the selected guidelines.

As soon as the texts have been annotated, they will be made accessible to participants. For the deadline in Phase 2, participants will be asked to process a new data set – one that has not been released before – and return the predicted annotations to the organizers, who will then make evaluations using standard measures (e.g., accuracy). Finally, there will be a workshop in which each participating team presents its system. This workshop is projected to be coordinated with the LaTeCH workshop series, which has taken place at ACL conferences in the past years (one of the authors of this paper has been a workshop organiser in past years).

There will be no fixation on computational approaches, statistical models, programming languages or environments to tackle this problem. The main benefit of shared tasks towards automatisation is that different approaches can be compared directly. Restricting this possibility space would directly harm the goal.

Technical Details and Timeline

This proposal is innovative in a number of ways: Shared tasks are a new kind of framework in the DH/H community, as such, focalisation have not been investigated in this way before. Last but not least, a shared task with the goal of creating annotation guidelines has not been organized before (to our knowledge). We believe it is more important to do this right than to do this fast, hence we are looking at a rather lengthy timeline.

Date	Event
August 2017	Announcement talks at DH and LaTeCH 2017
November 2017	Finalisation of Phase 1 details, submission for DH 2018, Call for participation (Phase 1)
April 2018	Submission deadline for guidelines
June 2018	Phase 1 workshop (DH)
August 2018	Announcement talk for Phase 2 (LaTeCH 2018)
October 2018	Finalisation of Phase 2 details, submission for ACL 2019, Call for participation (Phase 2)
May 2019	Submission deadline for systems for Phase 2
August 2019	Phase 2 workshop (ACL/LaTeCH)

Attracting enough participants is the main challenge from the organiser's perspective. The main incentives we envision for contributors are excellent publication opportunities: All submitted and generated materials will be published online (open access) under the umbrella of the shared task. Each individual work will be citable. This includes the submitted annotation guidelines, the produced consensus guidelines, and explanation and commentary documents.

In addition to the publication incentive, we believe that our approach is an important contribution towards systematic text analysis in the DH realm. We count on the playfulness and curiosity (in the best sense!) of the DH community to take part in this experiment.

Bibliography

Bal, M. (1997). Narratology: Introduction to the Theory of Narrative. University of Toronto Press, 2nd edition.

Bögel, T, Gertz M., Gius, E., Jacke, J., Meister, J. C., Petris, M., and Strötgen, J. (2015). Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative. In DHCommons Journal, 2015. URL http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic.

Carlson, L., and Marcu, D (2001). Discourse tagging reference manual. Annotation Manual, University of Southern California, 2001. URL https://www.isi.edu/marcu/discourse/tagging-ref-manual.pdf.

Ferro, L., Gerber, L, Mani, I., , Sundheim, B, and Wilson, G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Genette, G. (1980). Narrative Discourse. An Essay in Method. Ithaca, N.Y: Cornell University Press.

Gius, E., and Janina Jacke, J. (2016). Zur Annotation narratologischer Kategorien der zeit. Annotation Manual 2.0, Hamburg University, November 2016. URL http://heureclea.de/guidelines

Gius, E., and Jacke, J. (n.d.)The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis. International Journal of Humanities and Arts Computing, forthcoming.

Hovy, E., and Lavid. J. (2010) Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. International Journal of Translation Studies, 22(1).

Kuhn, J., and Reiter, N. (2015) A Plea for a Method-Driven Agenda in the Digital Humanities. In Proceedings of Digital Humanities 2015, Sydney, Australia, June 2015.

Meister, J. C. (1995). Consensus ex Machina? Consensus qua Machina! Literary and Linguistic Computing, 10(4), pages 263–270.

Musi, E., Ghosh, D., and Muresan, S. (2016) Towards feasible guidelines for the annotation of argument schemes. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 82–93, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2810.

Pier, J. (2014). Narrative Levels (revised version; uploaded 23 April 2014). In Peter Hühn et al., editors, the living handbook of narratology. Hamburg: Hamburg University. URL http://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014

Reiter, N. (2015). Towards annotating narrative segments. In Kalliopi Zervanou, Marieke van Erp, and Beatrice Alex, editors, Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 34–38, Beijing,China, July 2015. Association for Computational Linguistics, Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-3705.

Santorini, B. (1992) Penn treebank: Part-of-speech tagging. Annotation Manual 3, University of Pennsylvania, 1992. URL ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (n.d.) TimeML Annotation Guidelines, Version 1.2.1. http://timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf

Styler, W., Savova, G., Palmer, M., Pustejovsky, J., O'Gorman, T., and de Groen, P. C. (2014) THYME Annotation Guidelines. Technical report. http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf