Biographical data exploration as a test-bed for a multi-view, multi-method approach in the Digital Humanities

André Blessing, Andrea Glaser and Jonas Kuhn

Institute for Natural Language Processing (IMS)
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

The present paper has two purposes: the main point is to report on the transfer and extension of an NLP-based biographical data exploration system that was developed for Wikipedia data and is now applied to a broader collection of traditional textual biographies from different sources and an additional set of structured biographical resources, also adding membership in political parties as a new property for exploration. Along with this, we argue that this expansion step has many characteristic properties of a typical methodological challenge in the Digital Humanities: resources and tools of different origin and with different accuracy are combined for use in a multidisciplinary context. Hence, we view the project context as an interesting test-bed for some methodological considerations.

Keywords: information extraction, visualization, digital humanities, exploration system

1. Introduction

CLARIN[1] is a large infrastructure project and has the mission to advance research in the humanities and social sciences. Scholars should be able to understand and exploit the facilities offered by CLARIN (Hinrichs et al., 2010) without technical obstacles. We developed a showcase (Blessing and Kuhn, 2014), which is called TEA[2] (Textual Emigration Analysis), to demonstrate how CLARIN can be used in a web-based application. The previously published version of the showcase was based on two data sets: a data set from the Global Migrant Origin Database, and a data set which was extracted from the German Wikipedia edition. The idea for the chosen scenario was to enable researchers of the humanities access to large textual data. This approach is not limited to the extraction of information, it also integrates interaction and visualization of the results. In particular, transparency is an important aspect to satisfy the needs of the researcher of the humanities. Each result must be inspectable. In this work we integrate two new data sets into our application:

- NDB - Neue Deutsche Biographie (New German Biography)
- ÖBL - Österreichisches Biographisches Lexikon 1815-1950 (Austrian Biographical Dictionary)

Furthermore we investigate new relations which are of high interest to researchers of the humanities, for example, if a person is or was a member of a party, company or a corporate body.

Next, we view the project context as an interesting test-bed for some methodological considerations.

1.1. The Exemplary Character of Biographical Data Exploration

The use of computational methods in the Humanities bears an enormous potential. Obviously, moving representations of artifacts and knowledge sources to the digital medium and interlinking them provides new ways of integrated exploration. But while this change of medium could be argued to "merely" speed up the steps a scholar could in principle take with traditional means, there are opportunities that clearly expand the traditional methodological spectrum, (a) through interaction and sharing among scholars, potentially from quite different fields (e.g., shared annotations (Bradley, 2012)), and (b) through scaling to a substantially larger collection of objects of study, which can undergo exploration and qualitative analysis, and of course quantitative analysis (Moretti, 2013; Wilkens, 2011).

However, these novel avenues turn out to be very hard to integrate into established disciplinary frameworks, e.g., in literary or cultural studies, and from the point of view of scholarly less erudite computational scientists, it often appears that the scaling potential of computational analysis and modeling is heavily under-explored (Ramsay, 2003; Ramsay, 2007). It is important to understand what is behind this rather reluctant adoption. Our hypothesis is that humanities scholars perceive a lack of control over the scalable analytical machinery and should be placed in a position to apply fully transparent computational models (including imperfect automatic analysis steps) that invite for critical reflection and subsequent adaptation.[3] An orthogonal issue lies in the fact that advanced scholarly research tends to target resources and artifacts that have not previously been made accessible and studied in detail. So the digitization process takes up a considerable part of a typical project and a bootstrapping cycle of computational tools and models (as it is common in methodologically oriented projects in the computational sciences) cannot be applied

[1] http://clarin.eu

[2] http://clarin01.ims.uni-stuttgart.de/geovis/showcase.html

[3] The bottom-up approach laid out in (Blanke and Hedges, 2013) seems an effective strategy to counteract this situation.

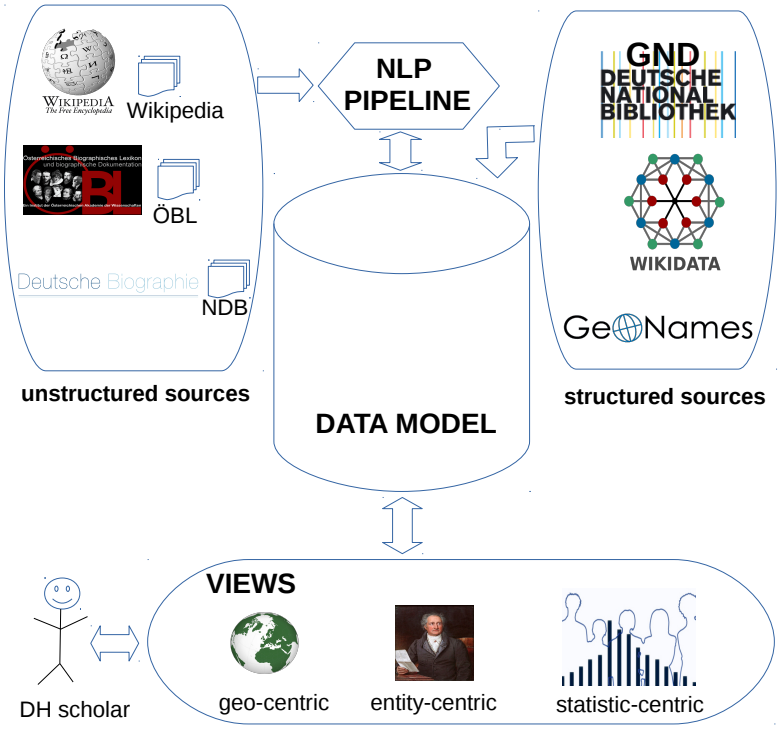

Figure 1: Overview of the NLP-based biographical data exploration system.

on datasets that are sufficiently relevant to the actual scholarly research question. We believe that biographical data exploration is an excellent test-bed for pushing forward a scalability-oriented program in the Digital Humanities: the compilation of biographical information collections from heterogeneous sources has a long tradition, and every user of traditional, printed resources of this kind is aware of the trade-off between the benefit of large coverage and the cost of high reliability and depth of individual entries. In other words, the intricacies that come from scalable computational models (concerning reliability of data extraction procedure, granularity and compatibility of data models, etc.) have pre-digital predecessors, and an exploration environment may invite to a competent negotiation of these factors. Here, a very natural multiple view presentation in a digital exploration platform can bring in a great deal of transparency: with a brushing-and-linking approach, users can go back and forth between an entity-centered view on biographical data (starting out from individuals or a visualization of tangible aggregates, e.g., by geographical or temporal affinity) and the sources from which information was extracted (e.g., natural language text passages or (semi-) structured information sources). This readily invites to a critically reflected use of the information. Methodological artifacts tend to stand out in aggregate presentations along an independent dimension, and it does not take specialist knowledge to identify systematic errors (e.g., in an underlying NLP component), which can then be fixed in an interactive working environment. Lastly, an important aspect besides this model character in terms of the interplay of resources and computational components and the natural options for multi-view visualization is the relevance of biographical collections to multiple different disciplines in the humanities and social sciences. Hence, sizable resources are already available and are being used, and it is likely that improved ways of providing access to such collections and encouraging interactive improvements of reliability, coverage and connectivity will actually benefit research in various fields (and will hence generate feedback on the methodological questions we are raising).

We are not the first who work on the exploration of different biographical data sets. The BiographyNet project (Fokkens et al., 2014; Ockeloen et al., 2013) tackles similar questions on reliability of resources, significance of derived output, and how results can be adjusted to improve performance and acceptance.

2. System Overview

Figure 1 shows the architecture of our approach. The system integrates different biographic data sources (top left). Additional biographic data sources can be integrated if they are based on textual data. Textual sources are processed by the NLP pipeline (top middle) which will be explained in the next section. In addition to textual data, structured

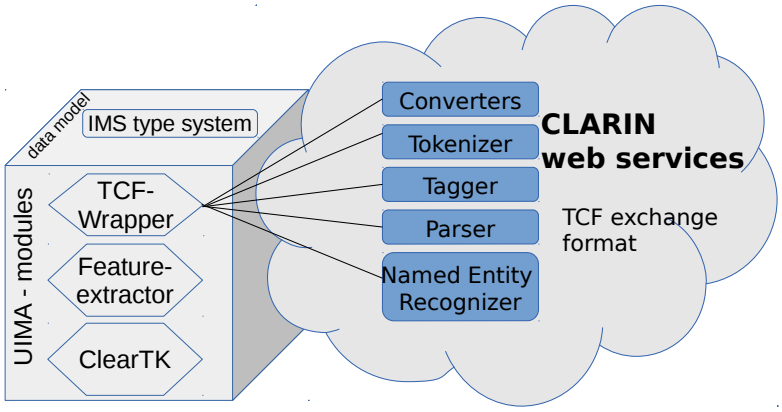

Figure 2: The used data model is based on the UIMA framework that interacts with CLARIN webservices.

data sets (top right) are used to enable real world inference (e.g. mapping extracted knowledge to a world map). We discuss the used structured data set in more detail later on. The data model (middle) central to our system includes the derived and extracted data and additionally all links to the sources. This enables transparency by providing access to the whole processing pipeline. Finally, several views of the data model (bottom) are provided. These allow the user to visualize the obtained data in different ways. A specific view can be used depending on the actual research question.

2.1. NLP Pipeline

Natural Language Processing (NLP) is typically done by chaining several tools as a pipeline. The right hand part of Figure 2 shows some basic tools (Mahlow et al., 2014) which are necessary. This pipeline includes normalization, sentence segmentation, tokenizing, part-of-speech tagging, coreference resolution, and named entity recognition. An important property is that these components are not rigidly combined. This allows the user to adjust or substitute single components if the performance of the whole system is not sufficient. The system is also language independent insofar as all NLP tools in one language can be replaced by tools in other languages. Table 1 gives more details about the used versions. These services are designed to process big data and do not require local installation of linguistic tools. This is often time consuming since most tools are using different input and output formats which have to be adapted.

2.2. Data Model

The data model of our system has to fit several requirements: i) store textual data and linguistic annotations; ii) enable interlinking and exploration of data; iii) aggregate results for visualization and data export; iv) store process meta data.

CLARIN-D provides its own data format called TCF (Heid et al., 2010) which is designed for efficient processing with minimal overhead. But, such a format is not adequate as core data model for an application. We decided to use the Unstructured Information Management Architecture (UIMA) framework (Ferrucci and Lally, 2004) as data model. The core of UIMA provides a data-driven framework for the development and application of NLP processing systems. It provides a customized annotation scheme which is called type system. This type system is flexible and makes it possible to integrate one's own annotation on different layers (e.g. part of speech tags, named entities) in the UIMA framework. It is also possible to keep track of existing structured information (e.g. hyperlinks in Wikipedia articles or highlighted phrases in a biographical lexicon) as the original text's own annotation in UIMA. Automatic annotation components are called analysis engines in the UIMA systems. Each of these engines has to be defined by a description language which includes the enumeration of all input and output types. This allows us to chain different engines including validation checks. UIMA is a well accepted data model framework, especially since the most popular UIMA-based application, which is called Watson (Ferrucci et al., 2010), won in the US show "Jeopardy" against human competitors. The flexible type system also enables the split of content-based annotation and process meta data annotations (Eckart and Heid, 2014) which allows keeping track of the processing history including versioning. Such tracking of process meta data can also be seen as provenance modeling (Ockeloen et al., 2013).

The combination of UIMA and TCF is simple since only a single bridge annotation engine is needed to map both annotation schemata. ClearTK is used as machine learning (ML) interface (Ogren et al., 2008). It integrates several ML algorithms (e.g. Maximum Entropy Classification). The extraction of relevant features is a customized component of the ClearTK framework. The used features are described in Blessing and Schütze (2010). At the current stage a standard feature set is used (e.g. part-of-speech tags, dependency paths, lemma information).

2.3. Textual Emigration Analysis

After the abstract definition of the requirements and architecture we give a more detailed view of the the extended TEA-tool. As mentioned before, we are using the already

Name	Description	PID which refers to the CMDI description of the service
Tokenizer (Schmid, 2000)	Tokenizer and sentence boundary detector for English, French and German	http://hdl.handle.net/11858/00-247C-0000-0007-3736-B
TreeTagger (Schmid, 1995)	Part-Of-Speech tagging for English, French and German	http://hdl.handle.net/11858/00-247C-0000-0022-D906-1
RFTagger (Schmid and Laws, 2008)	Part-Of-Speech tagging for English, French and German using a fine-grained POS tagset	http://hdl.handle.net/11858/00-247C-0000-0007-3735-D
German NER (Faruqui and Padó, 2010)	German Named Entity Recognizer based on Stanford NLP	http://hdl.handle.net/11858/00-247C-0000-0022-DDA1-3
Stuttgart Dependency Parser (Bohnet and Kuhn, 2012)	Bohnet Dependency Parser	http://hdl.handle.net/11858/00-247C-0000-0007-3734-F

Table 1: Overview of the used CLARIN webservices.

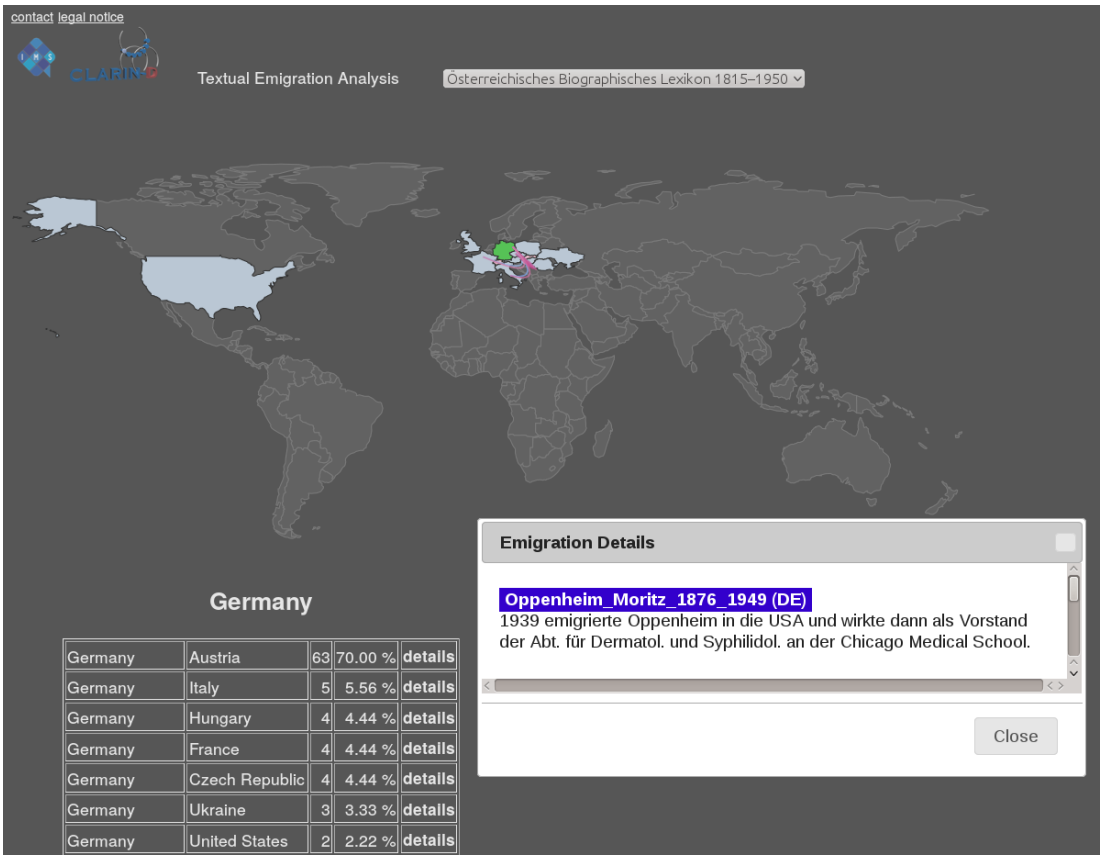

Figure 3: Using the TEA-tool to querying emigrations from Germany based on the ÖBL data set. The emigration details windows refers to ÖBL source which states that Moritz Oppenheimer emigrated 1939 from Germany to the US.

deployed web-based application that allows researchers to make quantitative and qualitative statements on persons who emigrated to other countries. The visualization of the results on a map helps to understand spatial aspects of the emigration paths, for example, if people mostly emigrate to nearby regions on the same continent or if they are spread over the whole world. The visualization contains a second view which aggregates and sums the emigration between two countries. The aggregated numbers can be inspected in a third view. Thereby, each number is decomposed by all persons who are part of the given emigration path. Not only the person names are shown, but the whole sentence stating this emigration can be visualized. In the expert mode such sentences can also be marked as correct or wrong by the user to increase the performance of the system through re-training or active learning. For more technical details on the base system please consider Blessing and Kuhn (2014).

The extended application, which contains the two new data sets, is shown in Figure 3. In this example the Austrian Biographical Data is used as data origin. The user selected the country Germany, and the extended system returned all persons who emigrated from Germany to other countries. This information is represented by arcs on the map and as a table at the bottom of the screen. A key feature of the application is that each number can be grounded to the underlying text snippets. This allows users interested in e.g., the two persons that emigrated from Germany to the US to click on the details to open an additional view that lists all persons

including the sentence which describes the emigration.
The three view types, geo-driven, text-driven and quantitative-driven of the TEA-application helps to explore the data set from different perspectives which allows researchers to identify inconsistencies. For example, the geo-driven view can be used to compare emigrations in a region by selecting adjacent countries. Such an analysis helps to find systematic geo-mapping errors (e.g. former USSR and the Baltic states). In contrast the text-driven view enables the identification of errors caused by NLP.

2.4. Challenges for extension of the TEA-system

To allow a smooth integration of the new biographic data sets, a few modifications in the NLP pipeline were needed. First, the import methods had to be adapted to allow the extraction of the textual elements from the new XML or HTML files. Second, the text normalization component had to be adjusted on biographic texts, because ÖBL or NDB use a lot more abbreviations which had to be resolved. This could easily be done using a list of abbreviations provided by the NDB website.
The integration of a new relation was more challenging: a new relation extraction component had to be defined and trained. For the emigration relation the whole process was done manually which is very time consuming. For the member-of-party relation we switched to a new system currently under development called 'extractor creator'. Since the system is in an early stage of engineering, the member-of-party relation was used as a development scenario. Figure 4 shows a screenshot of the extractor creator. Some of the basic methods of the interactive relation extraction component were published in Blessing et al. (2012) and Blessing and Schütze (2010). The novelty in the new system is that more background knowledge is integrated by using person identifiers (based on the German Integrated Authority File - GND) and Wikidata (Erxleben et al., 2014). This leads to a more effective filtering in the search which increases the performance of the whole system. The given example in Figure 4 shows the lookup of specific persons and the listing of all mentioned Körperschaften (corporate bodies) which are mentioned in the same Wikipedia article. A click on one of the corporate bodies opens the table on the right which lists all person who also mention this corporate body. A mouse-over function allows the user to see the textual context of the mention. The human instructor can then add relevant sentences as positive or negative training examples.

The first results of the novel relation extractor showed that unlike the emigration relation a more fine-grained syntactic feature set is needed in the scenario of corporate bodies. Figure 5 shows a simplified example that includes negations which occurred only rarely in the emigration scenario.

2.5. Entity disambiguation

Along with the extension of the core TEA system, we perform experiments with special disambiguation techniques that address named entities with multiple candidate referents. Often, people playing some role in a biography are mentioned very briefly, so unless the name is very rare, machine learning methods for picking the correct person have a hard time due to the very limited context. Many approaches rely on extracted features to learn something specific about people with ambiguous names, which requires enough training data. In our approach we use topic models for characteristic properties of the candidate referents. These properties can be for example nationalities, professions, or activities a person is involved in. We also apply topic models to the context of an ambiguous person in the biography and use the extracted properties to compute the similarity to the candidate referents. We then create a target-oriented candidate ranking.

3. Experiments

The largest data set consists of articles about persons which were extracted from the German Wikipedia edition. It covers 250,360 persons after filtering by the German Integrated Authority File (GND). The NDB data set contains 22,149 persons and the ÖBL data set 18,428 persons. Figure 6 depicts the overlap of the used data sets. Only 1,147 persons are part of all three data sets. We extracted 12,402 instances of the emigration relation from the Wikipedia person data set. For the NBD data set we found 1,932 instances of this relation and for the ÖBL data set we extracted 1,188 instances. Most of the persons found in Wikipedia are neither part of NDB or ÖBL which lead to the higher number of Wikipedia emigrations. Moreover, the overlap of all three data sets is small, meaning that we only have a few cases in which a person who emigrated is represented in all three data sets. An automatic comparison of the found instance for emigration is only possible to a limited extent since the different textual representations are not parallel for all facts. The member-of-party extraction is at an early development stage. Its performance has a high accuracy but the coverage is low. We started to use Wikidata for evaluation purposes since it also contains the same relation. However, the first results showed that Wikidata is not complete enough to be a sustainable gold standard. This observation was made by manually evaluating the membership relation in the Social Democratic Party of Germany. In this evaluation scenario our extractor found 18 persons which were not represented in Wikidata. This constitutes 20 percent of the extracted data. As a consequence, we need a larger manually annotated data set to enable a valid evaluation on precision.
Both experiments give evidence that we reached our first goal, which can be seen as a proof-of-concept. The chosen scenarios are not sufficient to enable an exhaustive evaluation since we have no well-defined gold standard data sets. However, components like the relation extraction provide enough parameters for optimization in the future.

4. Related Work

Since the Message Understanding Conferences (Grishman and Sundheim, 1996) in the 1990s, Information Extraction (IE) is an established field in NLP research. Chiticariu

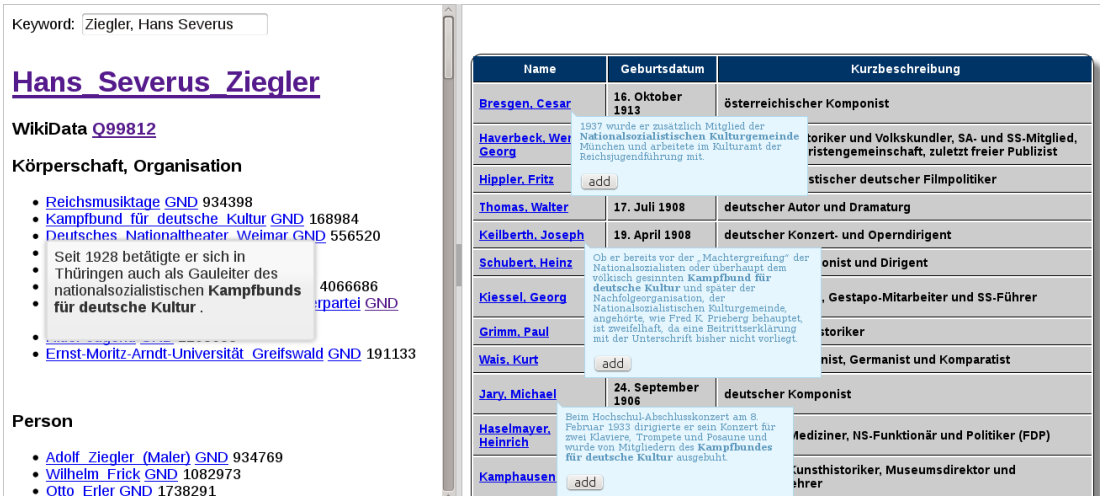

Figure 4: Prototype of the interactive relation extraction creator.

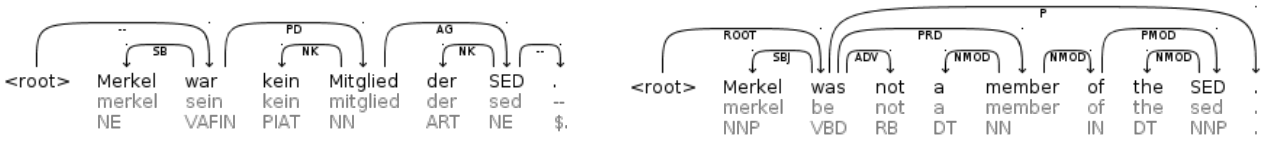

Figure 5: Dependency parse of the German sentence: Angela Merkel war kein Mitglied der SED.

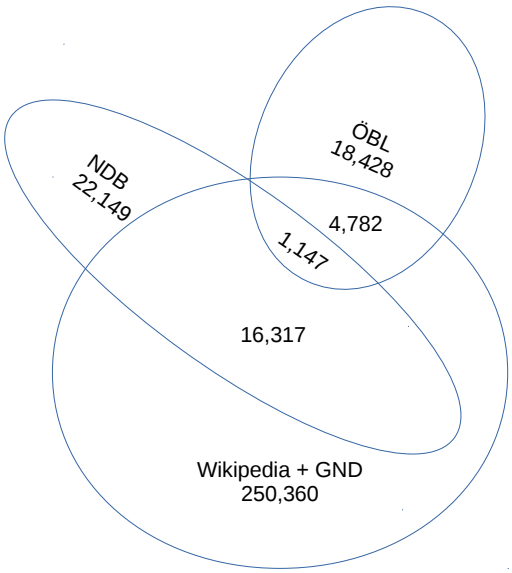

Figure 6: Size of used data sets.

et al. (2013) presented a study that shows that IE is addressed in completely different way in research than in industry. They showed that 75 percent of NLP papers (2003-2012) are using machine learning techniques and only 3.5 percent are using rule-based systems. In contrast, 67 percent of the commercial IE systems are using rule-based approaches (Li et al., 2012). One reason is the economic efficiency of rule-based systems which are expensive in development since the rules are hand crafted but later on the are very efficient without needing huge computational power and resources. For researchers such systems are not as attractive since their goals are different by working on clean gold standard data sets which allow exhaustive evaluation by comparing precision and recall numbers. In our system, we experimented with both, ML-based and rule-based approaches. Rule-based systems have the big advantage to provide transparency to the end users. On the other hand, small changes on the requested relations need a complete rewriting of the rules. We believe that a hybrid approach which allows the definition of some rule-based constraints to correct the output of supervised systems are the systems which provide the highest acceptance.

The drawback of ML-based IE systems (Agichtein and Gravano, 2000; Suchanek et al., 2009) is the need of expensive manually annotated training data. There are unsupervised approaches (Mausam et al., 2012; Carlson et al., 2010) to avoid training data but then the semantics of the extracted information is often not clear. Especially, for DH researchers, which have a clear definition of the information to extract, this is not feasible.

Another requirement of DH scholars is that they want to use complete systems which are often called end-to-end systems. PROPMINER (Akbik et al., 2013) is such a system which uses deep-syntactic information. For our use case such a system is not sufficient since they do not provide

several views on the data which also a big factor for the usability of system in the DH community.

5. Conclusion

We presented extensions of an experimental system for NLP-based exploration of biographical data. Merging data sources that have non-empty intersections provides an important access for quality control.

Offering multiple views for data exploration turns out useful, not only from a data gathering perspective, but quite importantly also as a way of inviting users to keep a critical distance from the presented results. Methodological artifacts that originate from NLP errors or other problems tend to stand out in one of the aggregate visualizations.

5.1. Outlook

We are collaborating with scholars of different fields of the humanities that are interested to use our system. Common questions are, which persons had certain positions at what time? Which persons are members of organizations or smaller groups at the same time? Which persons did their education at the same institutions? We will incrementally integrate such relation extractors in our system and observe the user experience. The mixture of data aggregation and being transparent is one of the crucial task to gain a high acceptance from DH scholars. We will also evaluate which additional factors are relevant for the acceptance of such a system.

Acknowledgements

We thank the anonymous reviewers for their valuable questions and comments. This work is supported by CLARIN-D (Common Language Resources and Technology Infrastructure, http://de.clarin.eu/), funded by the German Federal Ministry for Education and Research (BMBF) and by a Nuance Foundation Grant.

6. References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM Conference on Digital Libraries*, pages 85–94.

Alan Akbik, Oresti Konomi, and Michail Melnikov. 2013. Propminer: A workflow for interactive information extraction and exploration using dependency trees. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 157–162, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tobias Blanke and Mark Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Andre Blessing and Jonas Kuhn. 2014. Textual Emigration Analysis (TEA). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Andre Blessing and Hinrich Schütze. 2010. Self-annotation for fine-grained geospatial relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 80–88.

Andre Blessing, Jens Stegmann, and Jonas Kuhn. 2012. SOA meets relation extraction: Less may be more in interaction. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities*, pages 6–11.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both-worlds – a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87.

John Bradley. 2012. Towards a richer sense of digital annotation: Moving beyond a media orientation of the annotation of digital objects. *Digital Humanities Quarterly*, 6(2).

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th Conference on Artificial Intelligence*, pages 1306–1313.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 827–832. ACL.

Kerstin Eckart and Ulrich Heid. 2014. Resource interoperability revisited. In Ruppenhofer and Faaß (Ruppenhofer and Faaß, 2014), pages 116–126.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing wikidata to the linked data web. In *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)*, volume 8796 of *LNCS*, pages 50–65. Springer, October.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 129–133.

Daniel Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Christopher Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. Biographynet: Methodological issues when nlp supports historical research. In *Proceedings of the 9th International Conference on Language Resources and Evalua-*

tion (LREC 2014), Reykjavik, Iceland, May 26 - 31.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: the D-SPIN Text Corpus Format and its relationship with ISO standards. In *Proceedings of LREC-2010, Linguistic Resources and Evaluation Conference*, Malta. [CD-ROM].

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. Weblicht: Web-based lrt services in a distributed escience infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. electronic proceedings.

Yunyao Li, Laura Chiticariu, Huahai Yang, Frederick R. Reiss, and Arnaldo Carreno-fuentes. 2012. Wizie: A best practices guided development environment for information extraction. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 109–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cerstin Mahlow, Kerstin Eckart, Jens Stegmann, André Blessing, Gregor Thiele, Markus Gärtner, and Jonas Kuhn. 2014. Resources, tools, and applications at the CLARIN center stuttgart. In Ruppenhofer and Faaß (Ruppenhofer and Faaß, 2014), pages 127–137.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Niels Ockeloen, Antske Fokkens, Serge Ter Braake, Piek T. J. M. Vossen, Victor de Boer, Guus Schreiber, and Susan Legêne. 2013. Biographynet: Managing provenance at multiple levels and from different perspectives. In Paul T. Groth, Marieke van Erp, Tomi Kauppinen, Jun Zhao, Carsten Keßler, Line C. Pouchard, Carole A. Goble, Yolanda Gil, and Jacco van Ossenbruggen, editors, *Proceedings of the 3rd International Workshop on Linked Science 2013 - Supporting Reproducibility, Scientific Investigations and Experiments (LISC2013) In conjunction with the 12th International Semantic Web Conference 2013 (ISWC 2013), Sydney, Australia, October 21, 2013.*, volume 1116 of *CEUR Workshop Proceedings*, pages 59–71. CEUR-WS.org.

Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *UIMA for NLP workshop at Language Resources and Evaluation Conference*, pages 32–38.

Stephen Ramsay. 2003. Toward an algorithmic criticism. *Literary and Linguistic Computing*, 18:167–174.

Stephen Ramsay, 2007. *Algorithmic Criticism*, pages 477–491. Blackwell Publishing, Oxford.

Josef Ruppenhofer and Gertrud Faaß, editors. 2014. *Proceedings of the 12th Edition of the Konvens Conference, Hildesheim, Germany, October 8-10, 2014*. Universitätsbibliothek Hildesheim.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Helmut Schmid. 2000. Unsupervised learning of period disambiguation for tokenisation. Technical report, IMS, University of Stuttgart.

Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. 2009. SOFIE: A Self-Organizing Framework for Information Extraction. In *Proceedings of the 18th International Conference on World Wide Web*, pages 631–640.

Matthew Wilkens. 2011. Canons, close reading, and the evolution of method. In Matthew K. Gold, editor, *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis.