Social Network Extraction and Exploration of Historic Correspondences

Hui Li
Institute of Computer Science
Heidelberg University, Germany
hui.li@informatik.uni-heidelberg.de

ABSTRACT

Driven by the continuously increasing number of digitized and transcribed historic documents, natural language processing (NLP) and text analysis tasks are now frequently applied to historic texts to extract useful information and thus to enrich this cultural heritage. These tasks face several challenges, such as dealing with spelling variations, lack of orthography, and, oftentimes, missing reference language corpora.

In this paper, we present our approach to information extraction from a large corpus of correspondences written in the first half of the 16th century. For around 9,700 letters written in Early New High German by the German reformer Philipp Melanchthon, we develop and extend approaches to named entity recognition to extract information about geographic places, persons and organizations, and time. The extracted information serves as the basis for creating a social network structure that includes information about senders and addressees of letters as well as people mentioned in letters. This correspondence network structure is to be exploited in terms of analyzing the evolution of the correspondences, key persons and communities, and eventually topics covered by the letters over time, thus resembling key analysis tasks applied to today's typical social networks.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval—*Miscellaneous*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine Translation, Text Analysis*

General Terms

Theory

1. INTRODUCTION

Driven by major advancements in digitalization projects, an increasing number of historic documents such as letters, books, and reports are being transcribed, translated, and made accessible in digital form [22]. This is resulting in rich metadata repositories providing information about, e.g., the origin of the documents, short summaries, and even the documents' text itself, to name but a few. Digitized documents enable us to extract information from the documents' text, to analyze and correlate such information, to link it to other documents and to visualize structures and information, such as geographic information embedded in the texts.

In our project, the focus is on historic correspondences. Such correspondences, typically in the form of letters, provide a key sources for studying history as they describe or comment on events, provide information about actors and their roles in the development of history, and give valuable insights into opinions and trends over time. In the center of our study is a repository of letters written by the German reformer Philipp Melanchthon (1497-1560), a close friend of Martin Luther and humanist and professor in Wittenberg. These letters have been (and are still) collected and carefully analyzed over decades by the Melanchthon Research Center in Heidelberg, Germany [19]. Melanchthon is one of the most important reformers in the first half of the 16th century. In addition to his teaching, Melanchthon has published numerous books and commentaries in the areas of science, history, and theology. His contribution to the research and education system gave him the name of "Praeceptor Germaniae"(Teacher of Germany). He is also well known as the first systematic theologian of the Protestant Reformation, and he plays an important role in theology and church history. Melanchthon's correspondences not only provide an in-depth view of his life but also into the life of many other people during that period of time. These letters are considered a major source for studying German history during the early modern period.

Melanchthon's correspondences are written in Early New High German, mixed with Latin. Early New High German (ENHG) is the German Language used during the period from 1350 to 1560, which is also the beginning of the development of a German standard language [14]. According to the data we gathered, Melanchthon kept contacts to more than 7500 people. The information about the letters includes, among others, data about senders, addressees, time of the writing (sometimes associated with some uncertainty, as no day or even month is given or known) as well as summaries of the letters. For several letters the transcribed text and summaries are also accessible.

One major objective of this interdisciplinary project, spanning Computer Science, Computational Linguistics, and His-

Projects	Major Language	Time Period	Features
Correspondence of Thomas Bodley[4]	English	1585-1597	life events, interactive timeline
Early Modern Letters (EMLO)[12]	English	1550-1750	keyword analysis, Geonames, Wikipedia
17th-century Dutch Republic (CKCC)[7]	Dutch	17th century	visualization, topic modeling, keyword analysis
Bess of Hardwick's Letters[3]	English	1550-1608	life events, letter features
Darwin Correspondence Project[10]	English	1837-1883	theme classification
Thomas Gray's Archive[17]	English	1716-1771	index containing age of people
Mapping the Republic of Letters[23]	English	1400-1800	visualization

Table 1: Related projects on historical correspondences

tory, is to learn more about Melanchthon and his time. The key approach to this is to extract information about persons, places, and dates from these letters and associated metadata and to create a *correspondence network* that closely resembles today's *social networks*. It is our hypothesis that such a correspondence network provides a more holistic view of that period of time, its key players, personalities and circles of acquaintances than what would be perceivable through individual letters only. Who are the key persons in the correspondences of Melanchthon's, what communities do form over time and how do these communities evolve? Where are the geographic centers of communities? Who are the people Melanchthon corresponds with? Are these scientists, theologists, or more or less unknown people? For the latter aspect, obviously not only person (names) have to be extracted from the correspondences but such names also have to be correlated and linked to other data sources, such as DBpedia or repositories of historic person names. Also, can one identify major themes (topics) and events from the content or summaries of the letters and how can such information be correlated?

To achieve the above objectives, we incrementally build a comprehensive and extensible framework that addresses the following aspects. First, a pipeline for annotating correspondences is needed. The focus of this task is not on novel techniques in natural language processing (NLP) but on the adoption of methods for the extraction and exploration information latently embedded in historic textual documents, with a particular focus on temporal and geographic information as well as person information. This pipeline itself provides a valuable tool for scientists in respective areas in support of supervised annotation of historic documents. Second, based on the information extracted using the pipeline (and iteratively improving the extraction methods through thorough evaluations by experts), social network structures are built that initially contain information about senders, addressees, locations, and in particular temporal information about letters. Such a network structure alone already allows to derive, visualize, and explore measures of a network (such as key actors) over time. In a third step, the content of the letters, either in its raw form (non-translated) or as summaries (typically provided in Modern German) is explored in terms of detecting events (composed of time, location, and actors) and topics or themes. In the following, we briefly review existing work in these directions, and we then detail the questions to be answered in this project as well as the scientific approaches to be developed and employed.

2. RELATED WORK

The differences between historical languages and modern languages, such as spelling variations, orthography, and standardizations, have a significant influence on employing off-the-shelf natural language processing (NLP) approaches and tools [22]. Compared to modern languages, far less NLP resources and tools are available or even applied to historical language texts. One of the problems for historical documents is that the amount of digitized texts is relatively small compared to modern languages. Table 1 lists current well-known projects that deal with historical correspondences. As can be seen, most of the early modern correspondences are only available in English. It is quite difficult to find publicly available corpora of German correspondences that have been written during the early modern period. Furthermore, projects dealing with "old German" languages almost exclusively focus on corpus content analysis, such as manual annotation and mark up, instead of (automated) information extraction and linked data analysis.

Another problem is that the research on NLP for historical languages is just beginning, and the literature describing NLP tasks such as named entity recognition (NER) and part-of-speech (POS) tagging on these languages is not as comprehensive as for modern languages. A task most similar to traditional NER tasks on historical correspondences is done in the context of the Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic [7]. There, a semi-automatic, rule-based approach is used to build gazetteers of names. Names are obtained through manually annotated names, and then, names from specific books and rules are applied to generate more names. In this context the Aho-Corasick algorithm [2] is used for normalization of person names. However, this approach is very specific to a particular language and depends on a large amount of lexical evidence, which in most cases does not exist for historical languages [16].

As for the POS tagging task, the most straightforward approach is to create a POS tagger "from scratch". by annotating a corpus of that language and then training the tagger on it [22]. Dipper [11] tried this approach on Middle High German and found that it is better to apply normalization before tagging. Another frequently adopted approach is to use a modern-language POS tagger on a historical text, manually check the results, and then retrain the tagger on the corrected output. Scheible et al. [24] employed such an approach on Early Modern German texts (1650-1800). They were able to increase the accuracy of tagging by about 10% after normalization, compared to the non-normalized text tagging results.

Recently, topic modeling and event extraction has been adopted for the exploration of historical documents. Newman and Block [20] used a probabilistic topic decomposition approach to analyze topics in eighteenth century newspaper. Wittek and Ravenek [27] used random indexing to calculate

similarities between words and texts. Cybulska and Vossen [8] employed constructional clues and semantic type specifications to extract events from texts, but the precision and the recall are both no more than 60 %.

3. QUESTIONS TO BE ANSWERED

For our approach, transcriptions of letters are accessible in digital form (plain text and marked-up text). The letters are written mostly in dialects of low and high German or Latin. The letters content is available as plain text and/or semi-structured data in the form of regests (summaries of the original letter content). With this input data readily accessible, our research addresses problems in the following three areas:

Named-Entity-Recognition (NER). Existing resources and databases managing historic place and person names need to be exploited. For these, there is a plethora of sources such as Wikipedia, DBPedia, or GeoNames that vary in structure, detail, and quality. What are suitable data integration approaches in support of these NER tasks? Integrating and consolidating such sources in support of text tagging tasks is non-trivial. However, Linked Open Data infrastructures and techniques provide viable solutions in support of these tasks.

Deriving and Exploring Social Network Structures. The challenge in this task is to determine the nodes of a social network (typically persons, e.g., letter writers and addressees) and (labeled) links among nodes, typically referring to a correspondence. In particular, such a network can evolve over time as people appear or disappear (e.g., they died or are no longer involved in any correspondence). What are interesting structures that can be derived from such an evolving network? What are the relationships among correspondents and people mentioned in the letters? To answer questions like these, in particular the extracted temporal and geographic information will be used. While there has been substantial research on the analysis of social networks, there are only a few works that study the dynamics of networks, in particular with respect to location information or localized communities (see, e.g., [6]).

Event Analysis and Topic Models. The third theme of the project aims at investigating topics (themes) and events mentioned in the letters. What are significant events? What are the locations, actors and time periods that characterize events? What is a suitable model for describing events that are latently embedded in historical correspondences? What types of events are mentioned in the correspondences? If these questions can be answered effectively, then one would be able to analyze how topics and events evolve over time within the social network.

4. SCIENTIFIC APPROACHES

The scientific approaches to be employed in this project can be categorized according to the three problem areas stated above. Figure 1 illustrates the workflow with its tasks and data sources underlying our framework.

Information Extraction Part (NER tasks). One particular problem is normalization. Various spelling variations exist in historical texts and even the same author can spell words in different forms in one text [21]. Our approach is to first normalize the spelling in historical texts to modern language spelling, and then to use a tagger for modern languages to detect, extract, and normalize expressions (in a standard format) that refer to temporal, geographic, person, or organization entities occurring in the text. The languages to be considered in this context include Latin and different German dialects.

Statistical machine translation (SMT) techniques are chosen to be our approach and in that way the normalization process from historical language to modern language is regarded as a translation issue. Pettersson et al. [21] report that SMT-based approaches often give the best results, compared to two other main normalization approach, namely a simplistic filtering model and Levenshtein-based edit distance. In SMT, a translation model is trained on large parallel corpora (sentence-aligned corpora in two different languages) and the linguistic and structure information of the parallel data is extracted to infer translation correspondences based on different SMT systems [18]. However, SMT-based approaches typically require large corpora ("more is better"), which are usually not available for historic languages. A practical technique to deal with this issue is to duplicate the same data multiple times in the parallel corpora in order to increase the associated probabilities of the training data. However, regarding that some data might be more reliable than others, an alternative strategy such as confidence weighting will be used to ensure "good resources" [13].

Once the normalization process is completed, the Stuttgart TreeTagger [25] and the Stanford Named Entity Recognizer [15] is applied to tag the part-of-speech and named entities (e.g., person names and organization names) on the normalized text. Based on existing and newly developed resources, HeidelTime [26] will be extended to automatically extract temporal expressions from historic documents. Extracted and normalized information will be maintained in repositories that are used by subsequent text and network analysis tasks. The quality of the information extraction approaches to be developed will be evaluated based on the existing letter corpora and standard precision/recall as well as derived measures.

Social Network. Based on the extracted temporal, geographic, and person/organization information, social network structures are derived and analyzed. Of particular interest is the development of methods that focus on the evolving nature of such a network. Techniques and measures related to the exploration of social network structures will be applied for these tasks (see, e.g., [1] for an overview). For this, we employ measures such as connectivity, density, centrality, and cohesion. To investigate the evolution of a network, we aim at developing methods to detect stable clusters, i.e., groups of people who stay in contact via letters over time. Underlying these techniques are clustering approaches in social networks. Important input to determine the evolution of a network is information about the location(s) of sender/recipient(s) at a particular point in time and when the letters have been written (all information extracted from the letters). The approach we follow here first solely focuses on the geographic aspects of nodes (persons and organizations) and then on the temporal aspects (describing the movement and (dis)appearance of persons and organizations).

5. PRELIMINARY RESULTS

Normalization Trial. As the Moses system is the most dominant approach in the SMT field [18], it is applied in the ENHG normalization process. We notice that the Luther Bible was first written in early new high German (1545) and then transcribed in modern languages, i.e., the electronic version of Bilingual Bible is appropriate to be a sentence-align parallel corpus for machine translation. So a sample of the parallel Bible (the 1545 version and the modern version) is extracted and each version is divided into three parts for training (85,708 tokens), tuning (41,036 tokens), and test (434 tokens) separately. The translation result is very promising with around 0.60 BLEU score (translation from ENHG to modern German) and around 0.61 (translation from modern German to ENHG), compared to the normal score of limited resource of low density language (around 0.30).

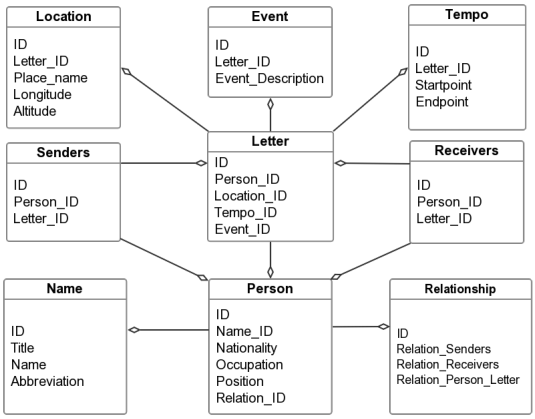

Figure 1: Data model to manage information and relationships for entities from letters, including senders, persons, receivers, names, events, and relationship.

Data Model. A conceptual data model has been created that builds the basis for managing information extracted from the correspondences and mapping them to data structure that can easily be managed in a repository or database. Figure 2 illustrates a part of the data model, which represents the information about the relations among persons and events. For instance, all the persons, including senders, receivers and people mentioned in the letters, are all stored in the table of "person"and the arrows between person and senders, or person and receivers show the inheritance relation between sender and person. The table of "Relationship"stores the information about the extent of closeness among all the persons in the correspondences. The objective of this data model is to minimize the redundancy and allows for flexibility and extensibility of the extracted data [5].

Visualization of Temporal and Geographical Information. Geo-spatial and temporal information has been extracted from the regests of the correspondences. A visualization tool "Dariah GeoBrowser"[9] has been used to visualize the locations of the senders of letters. Figure 3 shows the output of a visualization of around 200 letters sent from different locations between the years 1520 to 1561. The sparkling points on the figure represent different locations and the axis under the map is a timeline showing the number of letters sent over years. From this figure one can easily get a general view of the locations, the dates of correspondences, and the number of correspondences in time intervals. Besides,the amount of letters at each location suggests that people who are geographically close to each other tend to communicate more with each other than people who live far away from each other.

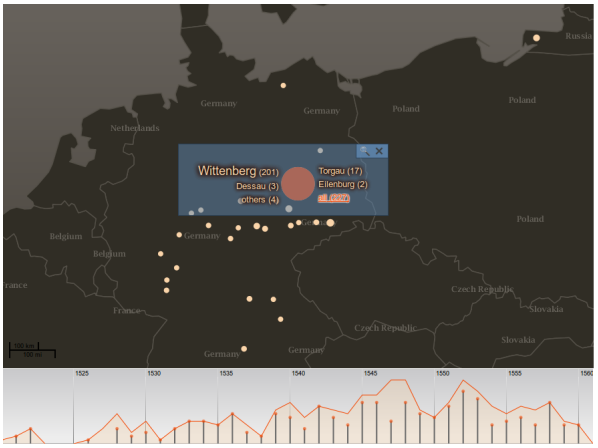

Figure 2: Visualization of geographic locations of senders, including the number of letters over time

6. SUMMARY AND ONGOING WORK

This interdisciplinary project focuses on the critical approach to information extraction of correspondences written in historical languages (here different dialects of German) and the exploration of social network structure of the German reformer Philipp Melanchthon and his time. The scientific approach and preliminary research demonstrate our effort on named entity recognition and deriving social network structures that are subject to a comprehensive analysis. Current research efforts focus in particular on the extraction of event information and integrating such information into the network structure.

7. ACKNOWLEDGEMENT

I would like to express my appreciation to my advisor, Prof. Dr. Michael Gertz for his guidance and support for this research project. I also would like to thank my colleagues for their time and effort for commenting on the proposed approaches and methods.

8. REFERENCES

[1] C. Aggarwal, editor. Social Network Data Analytics. *Springer*, 2011.

[2] A. V. Aho and M. J. Corasick. Efficient String Matching: An Aid to Bibliographic Search. *Communication of the ACM*, 18(6):333–340, 1975.

[3] Bess of Hardwick's Life. http://www.bessofhardwick.org/background.jsp?id=142.

[4] Thomas Bodley. http://www.livesandletters.ac.uk/bodley/bodley.html.

[5] M. Bollmann, F. Petran, S. Dipper, and J. Krasselt. Cora: A Web-based Annotation Tool for Historical and Other Non-standard Language Data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 86–90, 2014.

[6] T. V. Canh and M. Gertz. A Spatial LDA Model for Discovering Regional Communities. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 162–168, 2013.

[7] Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. http://ckcc.huygens.knaw.nl/.

[8] A. Cybulska and P. Vossen. Historical Event Extraction from Text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 39–43, 2011.

[9] Dariah Geobrowser. http://dev2.dariah.eu/e4d/.

[10] Darwin Correspondence Project. http://www.darwinproject.ac.uk/.

[11] S. Dipper. Pos-tagging of Historical Language Data: First Experiments. In *Semantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing*, pages 117–121, 2010.

[12] Early Modern Letters Online. http://emlo.bodleian.ox.ac.uk/.

[13] V. Eidelman, K. Hollingshead, and P. Resnik. Noisy SMS Machine Translation in Low-Density Languages. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 346–347, 2011.

[14] S. M. B. Fagan. German: A Linguistic Introduction. *Cambridge University Press*, 2009.

[15] J. R. Finkel and C. D. Manning. Hierarchical Joint Learning: Improving Joint Parsing and Named Entity Recognition with Non-jointly Labeled Data. In *Proceedings of ACL*, 2010.

[16] N. Freire, J. Borbinha, and P. Calado. An Analysis of the Named Entity Recognition Problem in Digital Library Metadata. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 171–174, 2012.

[17] Thomas Gray Archive. http://www.thomasgray.org/.

[18] P. Koehn. Statistical Machine Translation. *Cambridge University Press*, 2010.

[19] Melanchthon Research Center Heidelberg. http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/melanchthon/projekt.de.html.

[20] D. J. Newman and S. Block. Probabilistic Topic Decomposition of an Eighteenth-century American Newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767, 2006.

[21] E. Pettersson, B. Megyes, and J. Nivre. A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 32–41, 2014.

[22] M. Piotrowski. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers, 2012.

[23] Mapping the Republic of Letters. http://republicofletters.stanford.edu/.

[24] S. Scheible, R. J. Whitt, M. Durrell, and P. Bennett. Evaluating an "off-the-shelf" Pos-tagger on Early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, 2011.

[25] H. Schmid. Probabilistic Part-of-speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, 1994.

[26] J. Strötgen and M. Gertz. Event-centric Search and Exploration in Document Collections. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 223–232, 2012.

[27] P. Wittek and W. Ravenek. Supporting the Exploration of a Corpus of 17th-century Scholarly Correspondences by Topic Modeling. In *Proceedings of Supporting Digital Humanities: Answering the unaskable*, 2011.