Textual Emigration Analysis

Andre Blessing and Jonas Kuhn

IMS - Universität Stuttgart, Germany
clarin@ims.uni-stuttgart.de

Abstract

We present a web-based application which is called TEA (Textual Emigration Analysis) as a showcase that applies textual analysis for the humanities. The TEA tool is used to transform raw text input into a graphical display of emigration source and target countries (under a global or an individual perspective). It provides emigration-related frequency information, and gives access to individual textual sources, which can be downloaded by the user. Our application is built on top of the CLARIN infrastructure which targets researchers of the humanities. In our scenario, we focus on historians, literary scientists, and other social scientists that are interested in the semantic interpretation of text. Our application processes a large set of documents to extract information about people who emigrated. The current implementation integrates two data sets: A data set from the Global Migrant Origin Database, which does not need additional processing, and a data set which was extracted from the German Wikipedia edition. The TEA tool can be accessed by using the following URL: http://clarin01.ims.uni-stuttgart.de/geovis/showcase.html

Keywords: Digital Humanities, Visualization, Information Extraction

1. Introduction

One challenging aspect in the digital humanities is to enable researchers of the humanities access to large textual data. This not only includes 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 number of the results must be inspectable. Our showcase demonstrates how such an application can be designed and implemented. In our scenario, we focus on historians, literary scientists, and other social scientists that are interested in the semantic interpretation of text. Our application processes a large set of documents to extract information about people who emigrated. This application can be used in different scenarios: A researcher can use fiction literature from the 19th and 20th century to investigate which emigration destinations are popular in different time periods. Or to name another example a scientist can use biography data[1] to compare the emigration origins of different occupation groups.

Biographical texts are often rich in geographic information related to emigration: place of birth, place of death, emigration target country etc. This kind of information can be analyzed from a global perspective for understanding general trends in emigration (*which countries have how many people emigrated from Germany?*); it can also be evaluated on the level of individual biographies (*To which countries did Marlene Dietrich emigrate?*). Scientists often want to combine both perspectives, e.g., for understanding the general trend it is also relevant to analyze individual instances in detail, and vice versa. The brute-force method of merely counting co-occurring country names reveals general trends surprisingly well, but fails in many cases when looking at individual textual samples. Therefore, the challenge for an automatic analysis tool is to increase retrieval precision, e.g., by applying intelligent decoding for identifying emigration information expressed in natural language terms.

To enable easy access to the results which is a minimal requirement for the users of the humanities, we developed the system as a web-based application. We call this application Textual Emigration Analysis (TEA). The application can be used to display emigration source and target countries which includes emigration-related frequency information, and access to individual textual sources. The CLARIN infrastructure is an essential part of our application since already deployed tools can be used to process large amounts of textual data. Our current implementation integrates two data sets: A data set from the Global Migrant Origin Database[2], which does not need additional processing, and a data set which was extracted from the German Wikipedia edition.

2. Emigration extraction task

We aim to extract relation triples which can be aggregated and visualized. The goal is to get a triple like **emigrate**(ERIKA LUST, KAZAKHSTAN, GERMANY) for the sentence:

> Erika Lust grew up in Kazakhstan and emigrated to Germany in 1989.

The first argument refers to the person which has emigrated. The second argument links to the origin of the emigration and the third argument represents the target country. In contrast to the example just mentioned such a relation is often not entirely described in one sentence. The next example provides such a partial representation:

> In 1932, he emigrated to France.

That example represents only the third argument (*France*). For the other arguments more contextual information is need. For example, if we extract the text from Wikipedia we can resolve the subject *he* by exploiting the structured data of Wikipedia and the second argument can mostly be filled by the birth place of the person.

[1] E.g. from the German Biography: http://www.deutsche-biographie.de

[2] http://www.migrationdrc.org/research/typesofmigration/global_migrant_origin_database.html

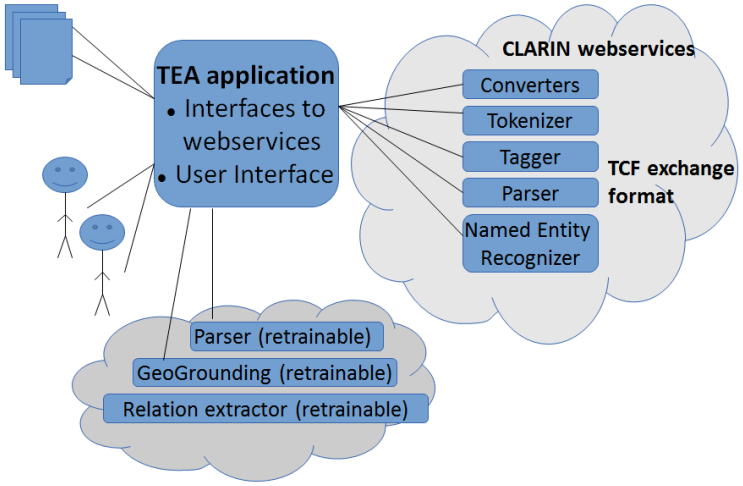

Figure 1: Overview of the architecture of our web application.

3. Design and Implementation

The goals of our showcase can be split into 5 key features:

1. exploit linguistic tools
2. accommodate concepts
3. aggregate information
4. link textual instances
5. adapt components

1) Linguistic tools can help to process text e.g. by providing named entities. The CLARIN infrastructure provides a large set of web services which can be freely used for academic research.
2) Researchers in the humanities searching for evidence in text to prove their hypothesis. We have to assist them in simplifying the task of finding such evidence in large corpora which cannot be achieved by manual reading of all content.
3) To enable large scale analysis it is necessary to process large amounts of textual data in short time and to aggregate the results across sentences and documents. This also includes to integrate external knowledge resources (e.g., gazetteers) which allow to map the results to real word entities (grounding).
4) The presentation of the final numbers for a quantitative analysis is not sufficient for the humanities scholars, since they want to reflect upon the results. Therefore it is required to link each number in the results back to the actual textual instance.
5) Researchers in the humanities are interested in a diverse range of texts including language variety and different domains. That can be solved by adaptable tools.

3.1. Advantages of using an infrastructure

The CLARIN infrastructure provides several tools as web services (Hinrichs et al., 2010) to process natural language text, e.g. convert different formats, sentence splitting, tokenizing, part-of speech tagging, morphological analysis, dependency parsing and named entity recognition. These services are registered in the Virtual Language observatory[3] (VLO). Table 1 lists all service that are used in our showcase. The last column contains the persistent and unique identifiers[4] (PID) for each service. These services are also integrated in WebLicht (CLARIN-D/SfS-Uni. Tübingen, 2012) which is an easy to use interface to the CLARIN infrastructure. However, for large applications like TEA a direct interaction with the services is more convenient. The WebLichtWiki[5] provides a developer manual which describes how to use the WLFXB library which allows to interact which the TCF format. There is also a tutorial[6] which describes how to interact with the RESTStyle web services of CLARIN in Java.

3.2. Architecture

Figure 1 shows the architecture of our showcase. The TEA application plays the central role with two functions: i) it provides a user interface to the researchers of the humanities which is implemented as web application; ii) it has interfaces to different web services to process textual content. These services can be further separated into two groups: i) services already integrated in the CLARIN infrastructure and ii) services that also include re-trainable methods not yet publicly released as CLARIN services. The latter allow domain adaptation for parsing, geo-grounding, and relation extraction.

3.3. User Interface

Figure 2 depicts the basic user interface of our application. The emigration data is visualized on a world map. The interface allows an interactive exploration of the data by clicking on countries to get more information. In this example the user selected *Iceland* and the top 3 countries for

[3] http://catalog.clarin.eu/vlo
[4] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-30
[5] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/
[6] http://hdl.handle.net/11858/00-247C-0000-0023-517B-8

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 CoreNLP	http://hdl.handle.net/11858/00-247C-0000-0022-DDA1-3
Stuttgart Dependency Parser (Bohnet, 2010)	Bohnet Dependency Parser for German	http://hdl.handle.net/11858/00-247C-0000-0007-3734-F

Table 1: Overview of the used CLARIN web services.

emigration from *Iceland* and the top 3 countries for immigration into *Iceland* are shown in a graphical (highlighted countries connected by arcs) and a textual manner (table below the map). The user can switch between different data sets.

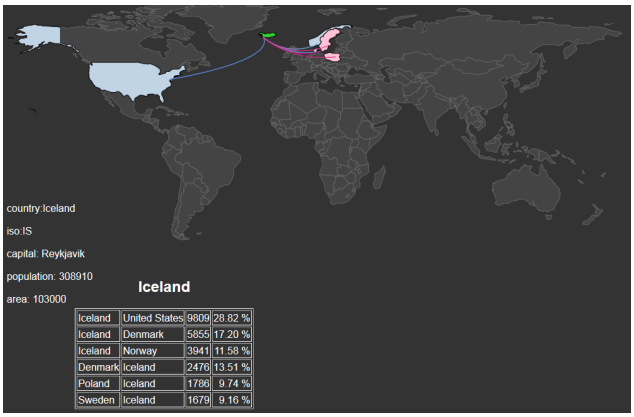

Figure 2: Screenshot of the user interface of our web application showing emigration from and to *Iceland*.

3.4. Relation extraction

We use the JWPL-API (Zesch et al., 2008) which allows a direct access to structured and unstructured data of Wikipedia. Structured data can be used instead of more expensive coreference resolution methods. Furthermore, Wikipedias structured data helps to infer missing information. In our previous example:

> In 1932, he emigrated to France.

we know that the article is about *Agostino Novella* who was born in *Genoa, Italy*. So we can infer the triple: **emigrate**(NOVELLA, ITALY, FRANCE) from the structured data of Wikipedia.

Besides using structured data, which can easily be extended to use any kind of Linked Open Data (e.g. from LOD initiative (Heath, 2009)), we included the output of NLP analyses to our extraction method. Only the combination of both enables a large coverage. The named entities of the NLP analyses help to find persons and geo-political entities in the text. The dependency and part-of-speech analyses are provide important features that allow us to classify the arguments of our relation. Therefore we process the text using CLARIN web services and extract the relations using a retrainable relation extraction system (Blessing et al., 2012).

3.5. Qualitative analysis

The application shown in Figure 3 is based on the data set extracted from the German Wikipedia edition. Again, the user selected *Iceland* and the table below the map shows the top country for emigrations and immigrations for *Iceland*. Since the data set is based on textual data the table contains of an additional column: details. This allows the user to link the aggregated results back to the original text passage which describes the emigration. In our screenshot we selected details about the emigration between *Iceland* and *Denmark*. In that case we have one person (Jon Törklánsson) that emigrated in 1981 from *Iceland* to *Denmark*. Additionally, the dialog about the emigration details provides hyperlinks to the complete Wikipedia article of each person and a hyperlink (show residence) that shows all countries which are mentioned in the article about the person.

3.6. Visualization and Mapping Details

Our application requires a JavaScript enabled web browser. We use the Raphaël JavaScript library[7] for the visualization. This library enables an integration of SVG[8] graphics which can be different geographical maps. E.g. if the researcher of the humanities is interested in movements of the 19th century the used world map can be replaced by a historical world map. Wikipedia is one good source to get such maps[9]. The grounding of the geo-political entities to a map is challenging, since we have to deal with ambiguous names. Our approach uses fuzzy search on a gazetteer and a ranking component. For instance the geo-political entity *Freiburg* is identified in a text but the gazetteer contains only the extended version of the name *Freiburg im Breisgau*. Our implementation uses two metrics to map a name to the gazetteer: string similarity and population size. This approach can be easily extended to use more features (e.g. previously mentioned geo-political entities). However, in

[7] http://raphaeljs.com/
[8] http://en.wikipedia.org/wiki/Scalable_Vector_Graphics
[9] http://en.wikipedia.org/wiki/Wikipedia:Blank_maps

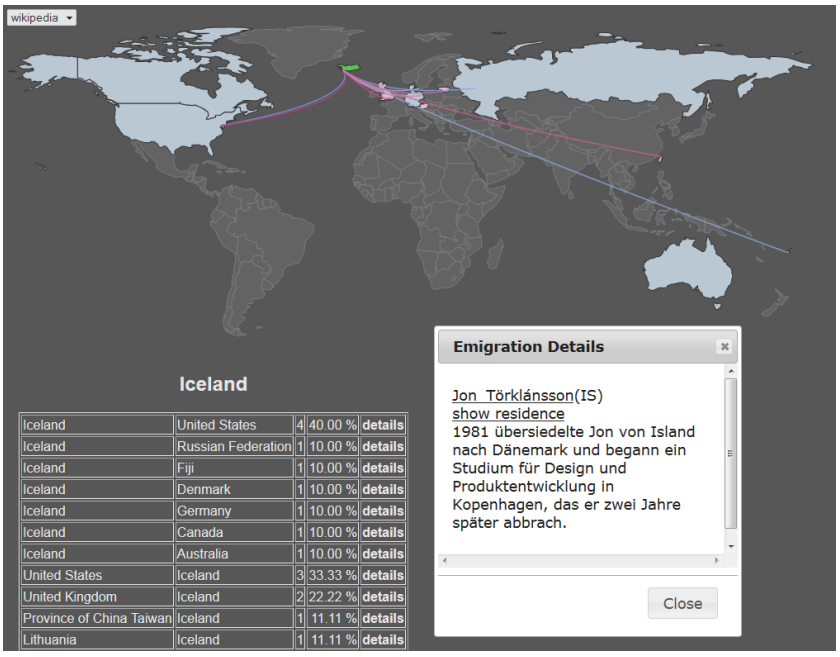

Figure 3: Visualized information about the extracted emigration to and from *Iceland* that are extracted from German Wikipedia edition.

some cases geo-political entities cannot be mapped since there are no corresponding toponyms in the gazetteer. In this work we use Geonames[10] as gazetteer which contains over 10 million geographical names, but if we have a sentence like:

> Sosonko emigrated in 1972 from the USSR to the Netherlands.

then Geonames has no representation for *USSR*. In that case we implemented a feedback method to ask for expert guidance: a user can suggest to use *Russia* as approximation for *USSR*.

4. Conclusion and Outlook

Our showcase consists of different components which, combined to a textual analysis system, aggregates and visualizes extracted data about emigrations. Yet, the system is restricted to German text but we are planning to extend this to other languages (Blessing and Schütze, 2012). This extension can be done because most of the web services in the CLARIN infrastructure support several languages. For instance, the dependency parser web service is based on the Bohnet mate-tools (Bohnet and Kuhn, 2012) which provides models for German, English, French, Chinese, Czech. We choose emigration as a first scenario and we are planning to integrate other relational concepts.

5. References

Blessing, Andre and Schütze, Hinrich. (2012). Crosslingual distant supervision for extracting relations of different complexity. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1123–1132.

Blessing, Andre, Stegmann, Jens, and Kuhn, Jonas. (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.

Bohnet, Bernd and Kuhn, Jonas. (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.

Bohnet, Bernd. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational*, pages 89–97.

CLARIN-D/SfS-Uni. Tübingen. (2012). WebLicht: Web-Based Linguistic Chaining Tool. Online. Date Accessed: 22 Mar 2014. URL https://weblicht.sfs.uni-tuebingen.de/ .

Faruqui, Manaal and Padó, Sebastian. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*.

Heath, Tom. (2009). Linked data - connect distributed data across the web. http://linkeddata.org/.

Hinrichs, Marie, Zastrow, Thomas, and Hinrichs, Erhard. (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.

Schmid, Helmut and Laws, Florian. (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.

Schmid, Helmut. (1995). Improvements in part-of-speech

[10] http://www.geonames.org

tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

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

Zesch, Torsten, Müller, Christof, and Gurevych, Iryna. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. electronic proceedings.