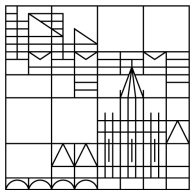
BACHELOR THESIS

VISUAL INTERACTIVE APPROACH FOR EXPLORING NAMED-ENTITY-RELATIONS IN MULTI-PARTY CONVERSATION DATA

RITA SEVASTJANOVA

Universität Konstanz



1.SUPERVISOR: Prof. Dr. Daniel Keim
2.SUPERVISOR: Junior-Prof. Dr. Bela Gipp

Department of Computer and Information Science
Data Analysis and Visualization Group
University of Konstanz

December 2015

Rita Sevastjanova: *Visual Interactive Approach for Exploring Named-Entity-Relations in Multi-Party Conversation Data,* © December 2015

SUPERVISOR:
Prof. Dr. Daniel Keim
Junior-Prof. Dr. Bela Gipp

ADVISOR:
Mennatallah El-Assady

LOCATION:
Konstanz, Germany

ABSTRACT

Language is an important medium for politics. Candidates debate policy positions during a campaign, politicians and experts debate controversial project approvals. The aim of the political science is to make inferences from such discussions. It can be done, by exploring what the participants of the discussions are saying and how they are responding to each other. Due to the large volume of data, it is a problematic task though. Therefore an automatic text analysis is needed to derive important and hidden information out of the data.

This thesis presents an approach to explore entities and their relations within multi-party conversation data. We use distance-restricted entity-relationship level, where two entities are seen as related if they are present in the same sentence within small distance to each other. The thesis shows, that such relations may incorporate more information than single entities. The exploration of the relations may present the topics of the discussion, and characterize, how similarly the participants of the particular discussion are expressing themselves. In our approach the user may observe the data from different perspectives, including temporal and geo-spatial information.

One of the contributions of the thesis is the description of a new distance-restricted entity relationship level. Such relations may be used as context descriptors of conversation data. We extract altogether 10 entity categories, therefore diverse entity-extraction methods are combined. Another contribution is the development of the visual, interactive user interface. We combine different frequently used visual representations with new approaches, to improve the data investigation process. One of the new approaches is the creation of data aggregation view, where user may define for him relevant data concepts, and explore relations between them.

ZUSAMMENFASSUNG

Sprache ist ein wichtiges Medium für die Politik. Die Kandidaten diskutieren ihre Standpunkte während politischer Kampagnen, Politiker und Experten diskutieren über Genehmigungen von umstrittenen Projekten. Das Ziel der Politikwissenschaftler ist das Ziehen der Schlussfolgerungen aus solchen Diskussionen. Dabei schauen sie, wie die Teilnehmer der Diskussion sich äußern und wie sie aufeinander eingehen. Allerdings, die große Menge der Daten ist schwer zu bearbeiten. Deshalb eine automatische Datenanalyse ist nötwendig um die versteckte, wichtige Information herleiten zu können.

Diese Arbeit präsentiert einen Ansatz für die Erforschung von Entitäten und deren Beziehungen in Mehrpartei-Konversitionsdaten. Dabei wird eine distanzbeschränkte Relationsart beschrieben, welche basagt, dass zwei Entitäten verbunden werden, wenn die in einem Satz nah an einander vorkommen. Diese Relationen beinhalten mehr Information als einzelne Entitäten. Die Erforschung von diesen Relationen zeigt, welche Themen diskutiert werden, und wie ähnlich die Diskussionteilnehmer sich äußern. Das System präsentiert Daten von verschiedenen Perspektiven, wie zum Beispel von der Perspektive der Zeit- und des Ortes.

Einer von in dieser Arbeit präsentierten Forschungsbeiträgen ist die Beschreibung von der distanzbeschränkten Relationsart der Entitäten. Diese Relationen werden als Inhaltsmerkmale von Konversationsdaten gesehen. 10 Entitäten-Kategorien werden extrahiert, dafür werden verschiedene Extraktionsverfahren miteinander kombiniert. Ein anderer Vorschungsbeitrag ist das Entwickeln von einer visuellen, interaktiven Benutzeroberfläche. Schon existierende visuelle Repräsentationen werden mit neuen Ansätzen kombiniert. Einer der neuen Ansätze ist das Entwickeln von Datenaggregations-Ansicht, wo der Nutzer kleinere Datenkonzepte definieren und visualizieren kann.

ACKNOWLEDGMENTS

I am grateful to the Professor Dr. Daniel Keim, for giving me the opportunity to write this thesis, and for his excellent classes which have inspired me to explore the issues of natural language processing. I would also like to thank the Junior-Professor Dr. Bela Gipp for his encouragement, I am looking forward for potential cooperation in the future.

I would like to gracefully acknowledge my advisor Mennatallah El-Assady, whose motivation, guidance and valuable ideas helped me to complete this work. She always found time to discuss the existing issues and new ideas. Without her support this work would not have been possible. Thanks, Menna!

I would also like to express my gratitude to Valentin Gold, who was always open for discussions. Thank you for your advises and for motivating your colleagues to take part in the user studies!

Dace, Sintija, Robert and Matthias, thank you so much for your help! I am so happy to have you.

And finally, I would like to thank my family and my partner Matthias, who are always there for me, who support me and believe in me. Love you!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CRF Conditional Random Fields

HMM Hidden Markov Model

IE Information Extraction

ME Maximum Entropy

MEMM Maximum Entropy Markov Model

MLN Markov Logic Networks

NER Named Entity Recognition

NLP Natural Language Processing

PMI Pointwise Mutual Information

POS Part Of Speech

SRL Semantic Role Labeling

SVM Support vector Machines

1 INTRODUCTION

Language is an instrument of communication. It is an important medium for politics, as political conflict often occurs in written and spoken language. Candidates debate policy positions during a campaign, politicians and experts debate controversial project approvals. To understand what politics is about you have to know what the participants of political discussions are saying. The scholars of politics have recognized the role of the language long time ago, however due to the data volume, it was difficult to make inferences. [40]

Therefore an automatic text analysis is needed to derive important and hidden information out of the data. Keim, D.A., Kolhammer J., et.al. [36] state, that "visual analytics provides technology that combines the strengths of human and electronic data processing." In this semi-automated analytics process, where the capabilities of the machine are combined with those of human, visualizations are used as medium. These visualizations present the data in an interactive manner, letting the user explore the data from different perspectives and at different levels. [36]

Named-entities and entity-relations are frequently used to derive hidden patterns out of large text data. Named-entities are concepts used in the Natural Language Processing (NLP) to refer to words for which one or many rigid designators stand for the referent. [51] They can be seen as important features also when exploring multi-party conversation data.

Multi-party conversation data.

Multi-party conversation data differs from simple text articles. Firstly, the data is unstructured and may be noisy. Such data embraces word-protocols of discussions; everything that has been said is transcribed, even if the utterances of the speakers are grammatically wrong. Secondly, the style of the transcripts depends on the person writing them down, as he can specify, how the statements are recorded (e.g. the abbreviations of words). The word "conversation" states, that the data consists of utterances of multiple persons, meaning, that both, what has been said in the discussion and who has said it, is important. "Multi-party" indicates, that the participants of the discussion may be divided in multiple groups, representing different parties or positions.

Entity relationship levels.

Multiple entity-relationship levels exist. Two relationship levels are frequently used in the literature: relations, representing semantic concern between entities, and relations, where entities are seen as related if they are present in the same document. However each of the relations have its limitations. The drawback of the semantic entity-relations is the time and effort, which need to be invested to extract them, especially, if many entity-categories are used. The relations, where entities are seen as related if they are present in the same document, are efficient to extract, however they are too general to be applicable on conversation data.

Due to the limitations of the above mentioned entity-relation levels, we focus on a distance-restricted relation (entity-pairs), which states, that entities are related only

if they are used close to each other within a sentence. This entity-relationship level can be efficiently extracted and it may reveal additional knowledge which relations representing large distance between entities do not contain. An example of extracted entity-pairs with maximum distance 3 is shown in the Figure 1.

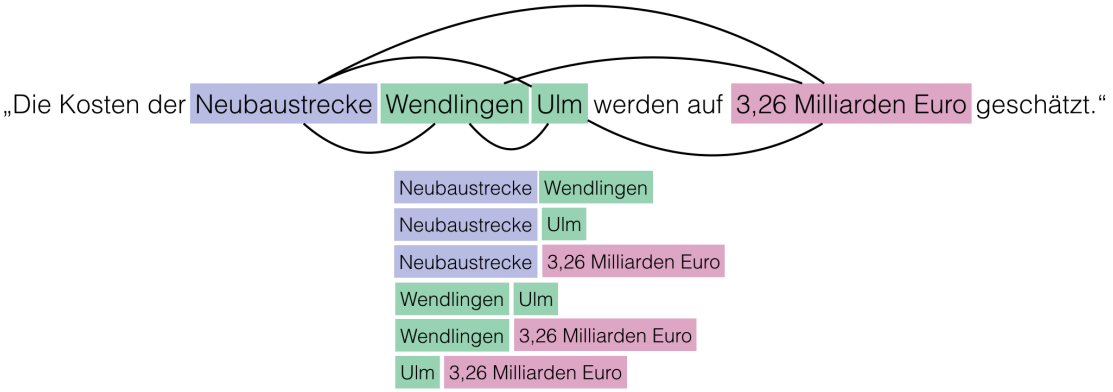


Figure 1: An example of entity-pairs with maximum distance 3.

In order to let the user analyze the discussed topics and compare single speakers and speaker-parties, we have created an approach, where multiple web-visualizations are interlinked, allowing to explore the data from different perspectives.

Our approach is part of the research project VisArgue[1]. The approach is done in four steps, shown in the Figure 2. First, the data is preprocessed and the relevant entities are extracted. Then the entity-pairs are searched and relevant information is mapped. When the data is transformed, the user interface is created, consisting of five interlinked visualizations. As the related work will show, a lot of solutions exist for named-entity extraction tasks. Therefore, the accuracy of the extracted entities is not our main goal. The focus of our approach is set on the visual user interface, letting the user explore new entity relationships. We combine different frequently used visual representations with new approaches, to improve the data investigation process. The evaluation of the approach is done, by carrying out user studies.

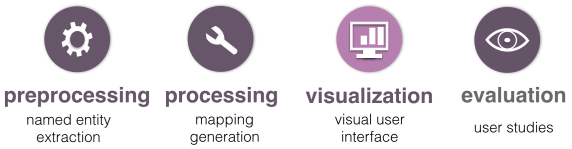


Figure 2: The workflow of the approach.

1 http://www.visargue.uni-konstanz.de/, accessed on 01.12.2015

1.1 STRUCTURE OF THE THESIS

The thesis is built up the following way: Chapter 2 presents the research challenges and the contributions of our approach. Chapter 3 presents the related work to existing named-entity and their relation extraction methods and entity exploration tools. The preprocessing step, containing data cleaning and named-entity extraction process, and the processing step, containing the generation of data mappings and extraction of entity-pairs, is explained in Chapter 4. After the data is processed, multiple interactive web-visualizations are created, which support the user in the data-exploration process. The visual decisions are described in the Chapter 5. The created visualizations are explained in Chapter 6. Chapter 7 presents the evaluation of the tool; Chapter 8 presents the conclusion of our approach. Chapter 9 lists possible improvements and future work.

2 RESEARCH CHALLENGES AND CONTRIBUTIONS

To make inferences about conversation data, multiple factors need to be taken into account. The word "conversation" states, that the data consists of utterances of multiple persons, meaning, that both, what has been said in the discussion and who has said it, is important. Therefore, the user should be able to observe the existing topics of the data and make inferences about single speakers and speaker-parties.

2.1 REQUIREMENTS OF MULTI-PARTY CONVERSATION DATA

- **R1:** To let the user explore the content of the discussion, and compare the utterances of single speakers, appropriate data descriptors are needed. They should not only present the summary of the text, but also reveal context information about smaller sections of the data.
- **R2:** The quality of the extracted knowledge from the data depends on the number of extracted entity categories. The more relevant categories are used, the more information can be received.
- **R3:** A new visual interface is needed, which allows the users observe the data in detail, make inferences about the discussed topic and explore similarities between participants of the discussion.
- **R4:** Not always a single discussion consists of a single topic. The changes of the discussed content may vary, therefore possibility to explore multiple subtopics is needed.

2.2 RESEARCH CHALLENGES

- **RC1:** The descriptors of the discussion should be general enough to present the summary of the text, and in the same time specific enough to reveal information about smaller subtopics of the data. The extraction of the descriptors should be efficient.
- **RC2:** Although multiple named-entity extraction methods exist, it is a highly challenging task to extract entities out of unstructured, noisy data. The methods, based on the syntax rules, are hardly applicably.
- **RC3:** The representation of the data in visual user interface needs to be pleasant and easy understandable. In the same time, the interface should represent the data from different perspectives, including the information about the topics of the discussion and about the participants, discussing them.

- **RC4:** Due to variable number of subtopics, which might be present in the discussion, it is problematic to specify a single parameter to limit them. Due to different preferences of the users regarding the relevant subtopics, it is difficult to specify and classify them.

2.3 CONTRIBUTIONS

- **C1:** To satisfy the **R1** and deal with the **RC1**, we describe and classify frequently used entity-relation types in the literature, and summarize their advantages and disadvantages. Due to their limitations, we describe a new distance-restricted entity relation, where two entities are seen as related, if they are used close to each other within a sentence.

- **C2:** To obey the **RC2**, we combine supervised and unsupervised computational methods of named-entity extraction task. We use 4 additional categories to the 6 named-entity categories, to receive even more context information from the data.

- **C3:** To deal with the requirement **R3**, we create and combine multiple interlinked interactive visualizations, allowing the user explore topics and subtopics of the data, and observe similarities and differences between single speakers and speaker parties with respect to their used entity-pairs.

- **C4:** To satisfy the **R4**, we create additional view for data-aggregation. The view lets the user define specific concepts (e.g. possible costs of discussed project). These concepts may be reused in the following sessions. That guarantees, that each user may explore exactly for him relevant subtopics of the discussion in more detail.

- **C5:** Additionally to the **C4**, we create a visualization to represent the user's defined concepts. The concept visualization includes the information about the persons discussing them. The Geo-Locations of the aggregated data in the visualization are placed according to their geo-coordinates, incorporating the geo-information in the force-graph layout. The timeline enables the presentation of concept changes over time.

3

RELATED WORK

This chapter presents the related work to named-entity extraction tasks and in the literature frequently used entity-relationship levels.

3.1 NAMED-ENTITY EXTRACTION

NLP is a field of computer science and linguistics concerning the interactions between computer and human (natural) languages. Since 1950s there are attempts to use symbolic methods to solve the problem of automatically processing the human language. In these methods the knowledge about the language is explicitly encoded in rules or other forms of representation. [21]

Many different subfields exist within NLP and one of them is Information Extraction (IE). IE deals with automatic extraction of structured information from unstructured sources. The topic of structure extraction engages many different communities spanning machine learning, information retrieval, database, web and document analysis. [55] However earlier - around 1996 - the main research topic dealt with identification and extraction of named-entities. [12]

Named-entity categories.

The word "named" in the expression "named-entity" aims to restrict the task to only those entities for which one or many rigid designators stand for the referent. Early works formulate the Named Entity Recognition (NER) problem as recognizing proper names [51], and most studied entity categories are Persons, Locations, Organizations. These types are collectively known as "enamex". The "timex" types like Date and Time and the "numex" types like Money and Percent are also quite predominant in literature. [53] However some recent works do not limit the number of named-entity categories. S. Sekine and Nobata [15] have defined a named-entity hierarchy which includes about 200 categories, trying to cover most frequent names mentioned in news articles. Different named-entity extraction techniques exist, which can be categorized in supervised, semi-supervised and unsupervised ones.

3.1.1 *Supervised Learning*

Entity extraction techniques.

Supervised learning is currently the dominant technique, when dealing with named-entity extraction tasks. In the last decades several machine learning models have been proposed like Hidden Markov Models (HMM) [45], Decision Trees [26], Maximum Entropy Models (ME) [3], Support Vector Machines (SVM) [61], and Conditional Random Fields (CRF) [59]. Systems, using supervised methods, read large annotated corpus and create disambiguation rules, which allow to extract entities, based on discriminative features associated with positive and negative examples. [53]

The main shortcoming of these methods is the need for a large annotated training corpus. [53]

3.1.2 *Semi-Supervised Learning*

The main technique for semi-supervised learning is called "bootstrapping" and involves a small degree of seed words, for starting the learning process. The system searches for sentences containing these seed words and tries to extract common contextual clues. Afterwards, the system inquiries for other instances that appear in similar contexts. The process is repeated and more related words can be found. [53]

The drawback of such methods is that any false seed "will mislead the next iteration of learning and might lead to even more false seeds, which degrades the performance of the whole learning process". [5]

3.1.3 *Unsupervised Learning*

The unsupervised learning is based on lexical resources (e.g.,WordNet[2]), on lexical patterns and on statistics computed on a large unannotated corpus. [53] For creation of patterns, multiple features may be used. These features can be divided in local knowledge, external knowledge and minor features.

Local knowledge features may be extracted from a token and its surrounding context, such as suffixes and prefixes of the token. [9] Also shallow parsing, or chunking may be used - it is the task of identifying non-recursive phrases, such as noun phrases, verb phrases, and prepositional phrases in text. [43]

One of the external knowledge features is Part-Of-Speech (POS) tagging. Generating POS tags for tokens may help to create more accurate rules, when extracting named-entities. Another feature which may be used is word clusters. One can gather named-entities from clustered groups based on their context-similarity and cover more possible entities. A simple way how to know whether a token or phrase is a named-entity is to look in a predefined entity list (gazetter). Such systems with large entity lists work pretty well if entities are not ambiguous. [9]

Number masks can be seen as a minor feature. They can be used when extracting dates (e.g. 10-12-1996: DD-DD-DDDD, 1999: DDDD). [9] A key orthographic feature for recognizing named-entities is token-capitalization, however it is not applicable for every language. [43]

One of the problem when using wordlists or gazetters is the entity-word ambiguity. (e.g. finding word *abend* as the surname *Abend*). Entity-entity ambiguity may occur if multiple categories share the same entity (e.g. last name *Eversberg* and part of the city *Eversberg*). Another common problem is the recognition of the beginning and end of a named-entity (e.g. finding *Frankfurt* instead of *Frankfurt am Main*). [54] POS tagging by itself, if used, can be a challenging task and even high quality POS taggers can lead to decrease of the classifier's precision and recall. [9]

2 https://wordnet.princeton.edu/, accessed on 01.12.2015

3.2 EXAMINATION OF NAMED-ENTITY RELATION LEVELS

Entity-relation levels in literature.

Two subtasks of IE are common and closely related - named-entity extraction and extraction of entity relations. [24] Two entity relationship levels dominate in the literature. They can be used to deal with different kind of tasks.

One of the entity relationships represents the semantic knowledge. Different relation types exist, like a role, which one person plays in organization (e.g. "member", "owner"), social relation (e.g. "parent", "sibling"), location relationships (e.g. "located" , "based-in") etc. [46] This relationship level is frequently used for question answering (e.g. Who is the chancellor of Germany?).

Another relationship level can be used to classify documents or summarize a single document. In this relationship all entities in a document are seen as related. Documents with the most common entities are seen as similar, and the most frequent entities of a single document presents a summary of the text.

3.2.1 *Entity-Relation Representing Semantic Knowledge*

Many applications in IE, NLP and information retrieval processes require an understanding of the semantic relations between entities. [50] Variety of approaches exist to fulfill this task, which can be divided in three groups - supervised, semi-supervised and unsupervised ones.

SUPERVISED TECHNIQUES Most of the relation extraction methods developed so far are based on the supervised learning, which requires a large collections of annotated text. The sentences in this data are hand-labeled for the presence of entities and relations between them. [20].

Supervised approaches, in general, can be divided in two categories - feature based and kernel based approaches. If an entity-pair and the sentence containing the pair is given, both of previously mentioned approaches analyze the sentence first. When the tokenization, partial or full syntactic parsing, and dependency parsing is done, then the feature based technique extracts diverse lexical, syntactic and semantic features. These features are used to train the system to identify entity-pairs and to classify them in predefined relations. [48] [42] [4] HMM are less suitable to be applied on a long range dependencies, which relation extraction often involves. Therefore Maximum Entropy Markov Models (MEMM) [22] are used to model more complex probability distributions, taking into consideration multiple text features. In contrast to feature based approach, the kernel based method does not extract syntactic or semantic features, but uses kernel functions to measure the similarities between two instances.[11]

The major drawback of the supervised learning methods is the need for a large annotated learning collections. Another common problem is the sparsity of lexical features. That means, if a relation is not mentioned in the training data, then it is difficult for both - the feature based and the kernel based systems - to detect it. [56]

One system which uses supervised learning techniques to extract semantic named-entity relations is presented by Wang, Li, et.al. [33] They use SVM in combination with diverse set of NLP tools to derive features. New features are introduced, like POS tags, entity subtype, entity class, entity role, semantic representation of sentences and WordNet synonym set. [33]

SEMI-SUPERVISED TECHNIQUES Semi supervised techniques try to avoid the problem of supervised techniques. For example, Sun, Grishman and Sekine [56] have presented a simple semi-supervised relation extraction system, which uses word clusters as additional features for relation extraction. Even if the training data has not the searched entity, it might be present in one of the clusters. That means - "the absence of lexical features can be compensated by the cluster features". [56]

The bootstrapping learning technique can be used also for named-entity relation extraction. Small number of seed patterns are needed and, used with a large corpus, in iterative fashion more instances and more patterns can be generated. [20]

Even if bootstrapping technique requires no labeled corpus, the performance of the technique relies on the input seeds. Similarly as for named entity extraction - false seeds may degrade the performance of the whole learning process. [41]

An example of bootstrapping technique is Snowball [37] system. Snowball introduced strategies for generating patterns and extracting tuples from plain-text documents. At each iteration of the extraction process, Snowball evaluates the quality of these patterns and tuples, calculating a confidence measures, and keeps only the most reliable ones as a new knowledge for the next iteration. The extraction patterns are mainly based on a strict keyword matching, which can be seen as one of the system's limitations. [49]

UNSUPERVISED TECHNIQUES Unsupervised extraction systems do not require human intervention. These systems, working in recursive manner, can discover new attributes, relations and instances in a fully automated manner. [6] Currently, the leading methods collect redundancy information from a local corpus, or use the Web as a corpus. The common workflow includes search for co-occurring token pairs with strings between them. Surface patterns are generated from these co-occurring tokens. [41] The frequent pattern mining is non-trivial, even if patterns are generated from well written text, as the "number of unique patterns is loose, but many patterns are non-discriminative and correlated". [41]

Gonza'lez and Turmo [30] have presented an unsupervised learning system, based on clustering. They see the relation detection between entities as a binary classification problem, where each pair of entities co-occurring in a sentence is classified as related or unrelated. The clusters are used to determine the scorer of an instance and filterer assigns the instance to the related or unrelated class. [30]

EXISTING VISUALIZATION TOOLS Siahbani M., Vadlapudi R. et.al. [8] have presented a system, which goal is to extract detailed facts about events from natural

language. To extract predicate-argument structures, they use Semantic Role Labeling (SRL) [18] approach based on a semantic data called Proposition Bank corpus, being annotated by linguistic experts. The extracted information is represented in an interactive visualization interface. They leverage three connected visualization components - the map represents the geographical information of the event, the timeline represents the temporal information and a faceted view presents the entities and their roles in the historical events. The system lets users explore information by applying multiple filters. These filters include entities classified as Person, Location, Country, and the roles between entities. System is useful for a specific task - to find geo- and temporal information on past events. The user may filter specific event (e.g. World War 2), and he is able to observe where and when this event took place. As this approach is created for the observation of events, it is too specific to be applicable on a conversation data.

Another system, which uses faced-based interface, is created by Hellrich J., Faessler E. et.al. [58]. It consists of interactive visualizations representing semantic relations among named-entities (protein-protein interactions), being automatically extracted from biomedical publications. The system combines two different modes of interaction. The first mode is faced-based interface, user can input a query and system uses taxonomic information to filter search results. The second mode is graph, it displays the proteins as nodes and relations as edges. By selecting a node, a link to this protein's entry in database called UniProt9 is displayed. When an edge is selected, a link to the abstract that describes this interaction is shown. This system is specialized to let the user search for articles, containing searched protein or function, and observe relations between proteins, therefore it is too specific to be applicable on a conversation data.

Oramas S., Sordo M. and Serra X. [60] have proposed a method to automatically extract meaningful knowledge from documents present in Digital Musical Document Libraries. The system classifies named-entities as "composers", "organizations", "places", and analyzes the relationships between them. The DBPedia Spotlight [14] is used for the identification of named-entities. For their specific dataset, each occurrence of the subject is treated as an entity, and linked to the correspondent DBpedia resource. For entity-relation extraction they use ad-hoc rule-based method based on dependency parsing trees, which "looks for paths among entities in the syntactic structure of the sentence, and filters them out according to the linguistic category of the words in between." [60] Afterwards the data is visualized as Web interface using the D3 Javascript library. Visualization is a graph, nodes are representing entities and edges - their relations. The information about selected entity can be observed in the DBpedia resource or in the website of Grove Music Online. This system is applicable only on a specific dataset, containing information about musicians and their work. Although new categories might be added, the visualization represents only the semantic relations between entities. No additional information about the original document is given, which might help to make conclusions about the data though.

Another system, which visualizes extracted named-entities, is called "Network of the Day" [38]. This system combines information, which has been extracted from on-

line newspaper articles and social media platform Twitter. [38] The goal of the "Network of the Day" is to show the "contrast of the presentation of events by the German online media and the reaction to the situation of a part of German online Twitter community" [38]. To find the relations between entities, the normalized Pointwise Mutual Information (PMI) scores [23] of their co-occurrence are calculated. Multiple interlinked views are used, giving an overview of the most popular entities (from categories Person and Organization), and their relationships. The user can get an overview of important topics over time, tag relations between entities and search for single entities. As already mentioned, the user may observe only two named-entity categories, therefore the information to explore is limited. The interface differs from our approach, as two separate resources are being compared. Therefore the goal of this program is too specific and can't be directly applicable on a conversation data.

CONCLUSION Semantic entity relations are used to present the semantic-role between two entities. By exploring these relations (if enough entity-categories are used), the user may find interesting subjects of the discussion, which is also the aim of our approach. But the main problem is the effort which needs to be invested to extract semantic relations. The more different relation-roles are needed, the more challenging is the extraction task. Therefore a simpler solution is required, which could be efficiently applicable on large number of entity categories and in the same time bear the entity-relatedness.

3.2.2 *Entity-Relation Representing Presence in the Same Document*

Multiple named-entity exploration tools exist, where entities are seen as related if they are present in the same document. This relationship level may be used to represent a summary of text document, where the most frequent entities are seen as topic representatives. This approach may be used also for document classification, where two documents are seen as similar if they have many common entities.

EXISTING VISUALIZATION TOOLS One of the tools which visualizes entity relations is called Jigsaw [35]. Entities which are found within a single report are seen as related. It lets the user observe, which entities are present in multiple documents, and explore relations between different entity-categories. The representation of entity-relations in the list view and in the scatter plot has one problem. The connections between entities may be presented only for distinct categories. In situations, where information about concrete subtopic is needed, which is represented by multiple categories (e.g. *10 Millionen Euro* categorized as Measure, *Neubaustrecke* categorized as Context-Keyword and *Wendlingen* categorized as Geo-Location), all present relations between these entities might not be displayed. Another drawback is the need to select the entities and visualize particular concepts in each session repeatedly. The storage of observed concepts might be useful though.

The goal of the system called Contexter [19] is to help experts to get an "efficient and quick understanding of large corpus of general news stories providing different levels of abstraction [19]". The user may explore entity relations, where two entities are seen as related if they are present in at least one common document. Authors use three

abstraction levels - plain text, bag of words and set of named-entities. In the visual interface the user can select an entity and explore the network around it (the most frequent entities in documents where selected entity is present). If desired, the user can view the actual context where the selected entity has appeared. The entity needs to be selected from a list, which has the problematic of the bag of words representations - it is difficult to get a good overview of data, if many data items are present. The possibility to query a concrete entity would improve the usability of system. Another drawback is the lack of used visual variables. The use of color to represent different entity-categories or use of size to represent the entity-frequency in the graph view would improve the readability and users perception.

Another visualization system to represent similar documents is created by Baumes J., Shepherd, J., and Chaudhary, A. [2] This system consists of multiple views, displaying entities classified as Person, Location, Organization and Date. Present entities of a single document are displayed in a list. The graph view shows entities present in the selected document and documents that reference these entities. Entities having geospatial tags are shown also in a map, created using Google Maps API. The system is useful to see a short summary of the news articles and search for documents containing queried entity. However some drawbacks still exist. Due to the lack of labels in the graph visualization, no quick overview of the used documents may be obtained. The user needs to hover over each node to get the complete information about the network. The entities present in a document are represented as a list. For a large document consisting of many entities the readability could decrease. A possibility to query entity, or sort the list to entity-frequency could improve the usability of the system. Exploration of subtopics of the discussion, what is important for our approach, is problematic.

Another tool to analyze and aggregate news articles is represented by Sebastian Arnold et.al [1]. They use "semantic techniques to extract named-entities, relations and locations from news sources" [1] with the aim to let the user analyze the news articles efficiently and visualize the extracted knowledge. This tool visualizes three types of categories - Persons, Locations and Organizations, where Locations are resolved to geo-coordinates using Google Maps API. Entities, which are frequently used in the most relevant news articles for the selected entity, are seen as related to the selected one. The visual representations - word clouds and histograms - give an overview of most mentioned entities in the respective date. They allow the user to recognize the trends of relationships and trace the entity back to the original document. Similarly as the previous system, it gives a quick overview which entities have been mentioned in the particular news articles. Although the visual interface is clear and pleasant, and different perspectives like geo-, and temporal information are displayed, the documents representing multi-party conversation data require additional representations. The probability, that in a conversation data multiple topics are being discussed is higher than in a news text. Therefore restrictions for entity-relations and additional representations for subtopics need to be added. Information about speakers of the conversation need to be integrated too.

Kintz M. and Finzen, J. [31] have presented a method for mining and visualizing company relations based on web sources. Two companies are seen as related, if they

are present in the same document. The user may explore the network around the chosen entity, being visualized as radial or force directed graph. The authors have mentioned themselves that, an obvious limitation is that relations simply represent co-occurrences of organizations in a document, all relations are presented in the same way, regardless of their meaning and importance. [31] Only one entity category is observable.

VizLinc [34] is a system, which integrates information extraction, search, graph analysis for the visual exploration of large data sets. It helps to find patterns and connections between entities, and one of the main goal of the tool is to narrow down the corpus to only relevant documents, containing filtered information. It lets the user see the text with highlighted entities, search for terms and entities, present locations on a map, display co-occurring persons in a graph visualization and show a summary of text as a word cloud with most frequent entities. Although the system incorporates different level information, it has some bottlenecks. First, only three entity-categories are used: Persons, Locations and Organizations. Second, the graph represents only the relations between persons, no network of other categories may be observed. Third, the geo-information is displayed separately, in a separate view, no connections between Locations and other categories may be observed. No temporal information is displayed, meaning, that the system has no overview on how the entities are used over time. This is the second tool found, where authors themselves emphasize, that seeing entities related, only because they are present in the same document, is too general. The authors say, that "multiple-term searches could also be improved by restricting the distance at which both terms can appear in a document". [34]

CONCLUSION As shown with the existing exploration tools, this relation may be used to summarize a single document or to search for similar documents (where similar documents have many common entities). Although this relation is efficient to extract, it is too general to be applicable on the conversation data. Especially in a long text documents, where multiple subtopics are present, entities do not always bear a real relatedness to each other. In such a data smaller subtopics, which are specific for a part of the data, might not be found. Therefore the comparison of the discussion's speakers is impossible, as no descriptors for utterances of single speakers may be found. Some improvements and restrictions of this entity relationship level is needed, to make it applicable on multi-party conversation data.

3.2.3 *Solution: Distance-Restricted Entity Relation*

Two previously explained entity-relationship levels have their advantages and disadvantages. The semantic relations between entities have a good quality of entity relatedness, however their extraction (especially if many entity-categories are used) may be very expensive. On the other hand, the relationship level, where entities are seen as related if they are present in the same document, is efficient to extract, however the relations are very general, especially where in a large document corpus diverse topics are present.

Therefore some generalization of semantic relationship is needed to reduce the extraction time and effort, and some restrictions of the second entity relation is desired to let the user observe not only the main topic of the text, but also provide understanding of smaller subtopics of the discussion.

To increase the probability, that two entities are genuinely related to each other, we shrink the possible distance between them. Due to the possibility, that in a single utterance multiple subtopics are mentioned, is still high, we say, that two entities are seen as related, if they are present close to each other in a sentence. It is difficult to set the most suitable distance, which would guarantee the real relatedness between entities. Too small distance could exclude relevant pairs, however too large distance could let the users make wrong conclusions about the data. Therefore the parameter of the distance between two entities to be seen as pair is set by the user before the data processing is started. The default value is maximum five tokens.

In comparison with the semantic entity relations, our method is more efficient to extract. Although the relations are more general as the semantic ones, due to the small distance they still may include semantic concern.

The only tool found, which has similarities with our approach, is called PosVis [10]. In this system named-entities are seen as related if they both appear at least once within a fixed text window, typically set by the user (e.g. a 20 word window, or a paragraph). Although the relation appears to be similar to our used one, it lacks on the explicit restrictions to guarantee the real relatedness between entities.

The aim of the system is to provide information about characters of a literally book or book collection. Word clouds and self-organizing graphs are used to let the user review the vocabulary of one or more entities, filter POS tags, compare different text segments, and explore the network of the characters of the literary work.

Although the system uses "word windows" to obtain the entity-relatedness, which is similar to our idea, the main functionalities of the system significantly differ from our approach. The authors of the PosVis state, that "the term name entity is used loosely to refer to names that can be extracted automatically (typically proper names)" [10]. Even though they are talking about named-entities, no specific categories like Persons or Locations are extracted. All named-entities are classified as proper-names. Hence, when searching for a concrete person, organization or location, all entities classified as proper-name need to be examined. Much time and effort need to be invested to receive information about concrete entity or entity-relation. Therefore it is not appropriate for exploration of conversation data.

4

WORKFLOW OF DATA PROCESSING

This chapter presents the preprocessing and processing steps in data-mining process, used in our approach. First the data is cleaned and named-entities and additional category entities are extracted. As the extraction of entities may be time consuming, the preprocessing step needs to be executed only once per file. The preprocessed data is stored in a separate XML file, which is reused in the following sessions. When the entities are extracted, the data is processed and entity-pairs are found. The preprocessing and processing of the data is done, using Java programming language.

4.1 DATA CLEANING AND NAMED-ENTITY EXTRACTION

10 entity categories.

The number of used named-entity categories depends on the system's task. To explore multi-party conversation data, the following named-entity categories are used: Person, Geo-Location, Organization, Measure, Measuring-Unit, Date-Time.

The Person names allow to explore, if speakers are mentioning other speakers in their talks, letting to foresee how active the discussion is. The category Geo-Location provides information on territories which are being discussed. It lets the users explore, if persons with different opinions are indicating distinct regions. The category Organization presents knowledge on organizations mentioned in the discussion. Knowing the political position of the speaker may explain the use of particular organization names in his talks. The category Measure lets the user investigate, if speakers are prepared for the discussion, knowing some statistical data (e.g. how mush particular project will cost, how many persons will be involved etc.). The category Date-Time presents the time periods, which are being considered in the discussion. The category Measuring-Unit presents some additional context words (e.g. *Tonne*, *Kilometer*).

To reveal even more context information out of data, 4 additional entity categories are used, which can't be classified as named-entities, but which are important to understand speaker's attitude to the discussed topic and to other participants.

While exploring the active discussion between the participants, one could observe, if these persons are polite to each other. Therefore additional category Politeness-Indicator is used, which represents the politeness words such as *Danke*, *Bitte* etc. By exploring them, one can find out, if the particular speaker is using politeness word, when addressing his statement to another participant of the discussion (e.g. *Danke, Herr Kefer*). The category Context-Keyword represents the context (topic) keywords of the discussion. They show, if speaker is talking about relevant issues of the discussion topic. The category Positive-Emotion-Indicator contains tokens incorporating positive emotion or sentiment (e.g. *schön*). The category Negative-Emotion-Indicator, in contrast to the previous category, represents the tokens incorporating some negative emotion or sentiment (e.g. *furchtbar*).

4.1.1 *Data Cleaning*

The aim of the preprocessing step is to extract entities from the input XML file. For this purpose, the data first needs to be cleaned. The existing dashes are removed, as they are seen as single words and therefore could negatively influence the search for entity-pairs. The superfluous signs are found, using regular expressions.

4.1.2 *Named-Entity Extraction*

The next step includes named-entity extraction. As mentioned in Chapter 2, multiple learning techniques exist to fulfill the entity extraction task. Methods, features and subcategories, which are used to extract the final entity-categories are shown in the Figure 3.

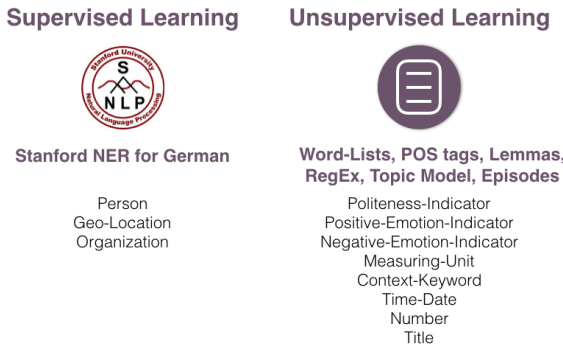


Figure 3: Methods and sub-categories, which are used to receive the final entity-categories. Source of the Stanford NLP logo: https://digitalprojectstudio.wordpress.com

SUPERVISED LEARNING Most of the work for named-entity extraction tasks has been done for English. A major reason for this situation is "the (un-)availability of labelled development data in the respective language" [52]. The Faruqui M. and Pado S. [52] have developed the first NER system for German, which is freely available for academic purposes. They use Stanford's Named Entity Recognition system[3] , which uses a linear-chain CRF [17] to predict the most likely sequence of named-entity labels. The system uses multiple features, like the lemma and POS tag of word, n-gram features, the capitalization of word, numbers, etc.

This NER system is used as the first step in our entity extraction pipeline. After testing both classifiers provided (Huge German Corpus-generalized classifier and deWac-generalized classifier), although the authors suggest to use the second classifier for data which is not news-wire based, the first classifier gives better results on the tested discussion documents, creating less false categories.

Because the Stanford NER for German may be applicable for extracting only such categories like Person, Geo-Location and Organization, additional unsupervised methods need to be used to extract all remaining entity types.

UNSUPERVISED LEARNING One of the unsupervised learning techniques used for our system is the predefined word-lists (gazzeters), which are created manually. They

3 http://www.nlpado.de/, accessed on 01.12.2015

are directly applicable to extract tokens or phrases which may be classified as Measuring-Unit (e.g. *Kilometer*, *Tonne*, *Prozent* etc.), Date-Time (e.g. *Mai*, *Monntag* etc.), Politeness-Indicator(e.g. *Danke*, *Bitte* etc.), Positive-Emotion-Indicator (e.g. *geehrter*) and Negative-Emotion-Indicator (e.g. *frustriert*). The category Context-Keyword is also created using gazzeter, but this word-list is generated, using topic descriptors of the LDA Topic Modelling [27] and Episode lemmas [39]. In parallel to Stanford NER, the external knowledge feature POS tag is used to tag corresponding tokens as nouns, lowering the number of tokens which requires the search in predefined gazzeters.

The category Measure does not contain single tokens, but word phrases. To classify them, additional category Number is used. First, using gazzeter containing numbers as text, the system checks the presence of tokens like *vier*, *hundert* etc. Using regular expressions, the numerals are extracted. Sometimes numbers consist of multiple parts like *100 Millionen*, where the first part(s) is a numeral and the following is written in words. Therefore the system proves, if multiple tokens categorized as Number are following each other, in such a situation they are concatenated to one single entity. Afterwards the system proves if n-grams are present, consisting of category Number followed by the category Measuring-Unit or Context-Keyword. These n-grams are classified as Measure. Some n-grams, containing statistical data, have a preposition like "pro" in the phrase *250 Euro pro Kubikmeter*. The system proves, if such statistical data followed by preposition, which is afterwards followed by Measure or Context-Keyword entity exists and classifies it as Measure.

Also the category Date-Time may consist of word phrases (e.g. *Jahr 2010*, *14 Uhr*). therefore additional rules are used to match the patterns *Jahr** followed by Number or Number followed by *Uhr*.

To improve the quality of the data, the title (e.g. *Herr*) of the person mentioned in the discussion needs to be added to the entity categorized as Person (e.g. *Herr Volker Kefer*). To accomplish this task, another gazzeter is used, containing tokens representing category Title, such as *Dr.*, *Frau*, *Kollege*. It is important to have these words added to the person names, as otherwise it would negatively impact the search for entity-pairs.

The categories Positive-Emotion-Indicator and Negative-Emotion-Indicator are created by the linguists. The original input XML file already contains the annotated emotion-tokens, they are used to classify the particular entities.

4.1.3 *Possible Improvements*

As mentioned in the Chapter 1, the precision of the entity extraction task is not the focus of our approach. However the precision could be improved, using additional features, like word-clusters or databases to generate larger gazetters. If the transcript data is too unstructured and noisy, especially for the categories, extracted with the supervised method, additional features are highly relevant. A recommender system, where user corrects the wrongly categorized entities, could be used to improve the extracted data quality. These additional methods are explained more detailed in the Chapter 9.

4.2 PROCESSING AND ENTITY-PAIR EXTRACTION

After the named-entities are extracted and new XML file is stored, the data is processed and all relevant information, including entity-pairs, are obtained.

4.2.1 *Data-Mapping Generation*

When the data is preprocessed, the necessary basic information is stored to be later used as an input for visualizations. The stored information includes words/entities of the discussion, speakers of the discussion, frequency of a single entity (with respect to its category), length of each sentence and each utterance.

The lemma of single entities and the Levenshtein [28] edit distance between two entities is used for the recommendation system, when containers with single concepts are created. The system recommends the most similar entities which could be added to the newly created container automatically. The topic of each entity is used to suggest the possible candidates which could be semantically related to the already selected entities. The n-grams of named-entities and additional category entities are stored separately in the concatenated form (used for visualizations as class names or ids).

Each speaker, which is mentioned in the utterance, is checked for if this person has been spoken 1-2 utterances before the current utterance or 1-2 utterances after the current one. These utterances are used to show, how active the discussion is.

Complete table of generated mappings is shown in the Table 1.

4.2.2 *Entity-Pairs*

Observation of entity-pairs with small distance is not a common practice, when exploring related entities in the text. However, such pairs may result with some additional knowledge, which single entities or other entity-relationships might not reveal. For example, knowing, that the speaker uses entity pair *Wendlingen Ulm* in his talk instead of separate entities *Wendlingen* and *Ulm* might mean, that not only the single locations are important, but also the route. Presence of entity pairs *Magistrale Paris*, *Magistrale Bratislava* and *Paris Bratislava* lets you conclude, that the *Magistrale Paris Bratislava* is being discussed. If some of the speakers close to the Context-Keyword entity (e.g. *Neubaustrecke*) use a positive expression like *wunderbar* and some of the speakers use negative emotion entities like *furchtbar* this could lead to observation, that some speakers are positively minded against the particular topic in contrast to others.

The most frequent entity-pairs may represent the topic of the discussion. The frequent pairs used by a single speaker may present a short summary of his talk. However not only the frequent pairs are important, but also pairs, which slightly differ from each other (e.g. *Kosten 1000 Euro*, *Kosten 1050 Euro*) might present different speaker opinions about specific subtopics of the discussion.

sentence	entities entity pairs length
utterance	entities entity pairs length speaker political politeness positions
speaker	entity pairs number of utterances party
speaker pair	common entity pairs
entity	entity pairs frequency edit distance lemma topic
entity pairs	frequency weight (sum of distances/frequency)
entity's position in discussion	entity category position in sentence

Table 1: In the processing step generated data mappings.

The order of the entities is important. The pair *Wendlingen Ulm* and *Ulm Wendlingen* are seen as two separate entity-pairs. Especially for entity-pairs containing two locations, the arrangement, in which entities are mentioned, may be of a high relevance. Different speakers using opposite order of two identical entities could mean, that they want to emphasize exactly the concrete direction.

When the pairs are found, a weight (sum of distances/frequency) is calculated for each of them, representing the average distance between entities of the pair.

5 VISUAL DECISIONS

This chapter presents the visual decisions for representation of entities and speakers of conversation data. To visually represent the data transformed in the processing step, it needs to be mapped to the visual variables. How Maciejewski, R. [47] states that visual representation helps the users perceive salient aspects of their data, and "these visual representations augment the cognitive reasoning process with perceptual reasoning, which enhances the underlying analysis." [47]

5.1 DATA TYPES AND VISUAL VARIABLES

For each specific task the appropriate visual methods need to be chosen. To make the best choice, the used data types need to be taken in account. Four different data types exist: nominal (each data element is defined by a label), ordinal (each data element has a specified rank ordering), interval (data with specified distances between levels of an attribute) and ratio (data having a zero point that indicates the absence of the item being measured). [47]

Diverse data types may be represented in diverse ways, using different visual variables. Bertin [29] proposed seven variables, which might be mapped to data elements: position (the spatial variables), size, value (tone), texture, orientation, shape, and hue. He distinguishes between selective, ordered, associative and quantitative visual variables. If symbols of the data can be isolated, forming a groups of similar symbols, then selective (e.g. color, hue) variable can be used. If it is possible to perceptually group symbols, based on some characteristic, then associative (e.g. shape) variable may be used. If it is possible to rank symbols based on a characteristic, then ordered (e.g. darker or lighter shading) variable is appropriate. If it is possible to perceptually quantify the degree of variation of symbol, then quantitative (e.g. size) variable is the best choice. [57]

When using color as visual variable, it is crucial to choose the appropriate color scale. Three types of univariate color schemes exist: qualitative, sequential, and divergent color scale. Harrower and Brewer [7] utilize quantitative color scheme, when working with nominal data, where each data category can be separated in its own color. In sequential color scale, single scalar variable is mapped to its brightness. It may be used to represent the ordinal, interval, and ratio data. [47] Rheingangs [44] defines the divergent color scale as two sequential color schemes which are pasted together, sharing an end point. It is appropriate for ratio data type. [47]

5.2 VISUAL VARIABLES FOR ENTITY CATEGORIES AND SPEAKER PARTIES

Visual design decisions to represent single entity.

Entity categories represent nominal data. Single elements can be isolated from each other and similar elements may be grouped. Therefore color or shape are visual variables, which might be used to represent them. As our approach in total consists of 10 entity categories, to choose appropriate shapes or colors is not an easy task. Due to the need to visually represent the frequency of an entity (representing it with the visual variable "size"), we decided to use color instead of a shape, as it might be difficult to compare the size of different shape elements. Representation of entities in the whole approach is equivalent. Due to the repeated use of graph layout, the single entities are visualized as circles.

Figure 4: The final representation of entity categories.

The decision of the most appropriate 10 qualitative colors to represent the different categories is not an easy task. After multiple unsuccessful tries, the final choice (shown in the Figure 4) is based on the semantically meaningful concept-color associations idea [16]. This idea is based on the linguistic information about the terms, to generate semantically meaningful colors. In this approach the co-occurrences of color name frequencies from Google n-grams are combined with the representative color from Google Images, and additionally symbolic relationships defined by WordNet can be used to select identity colors for categories such as countries or brands. We are using a simplified method of the previously described one. The most representative color from Google Images for the respective category is chosen. The category Person is represented with the color "orange", characterizing the color of skin. Category Geo-Location is represented with the color "green", describing the color of land. The category Organization is represented with the color of skyscaper - "light blue". The Date-Time is represented with the most present color for clock in Google Images - "gray". The Measuring-Unit is represented with the color of measure tape - "yellow". The category Politeness-Indicator could be represented as hand-shake, therefore similar color to the Person is used, we chose the color "brown". The Positive-Emotion-Indicator can be seen as permission or "yes", therefore the "green" color is used, representing the green traffic light. The Negative-Emotion-Indicator, on the contrary, represents "no" and is presented with the "red" color of traffic light. For the category Measure, presenting statistical data, and category Context-Keyword, presenting topic descriptors, no unique color in Google Images could be found, therefore two qualitative colors which differ from all previously chosen ones are selected - "pink" for Measure and "violet" for Context-Keyword. As the recognition in the usability is much more desirable

as recall [32], to discriminate different categories, svg icons in addition to the colors are used.

Similarly as entities, speaker positions represent nominal data, therefore they are also mapped to the variable "color". To distinguish entities from speaker profiles, the rectangle shape is used to present the last ones. The top line of the rectangle indicates the particular speaker's party. The speaker photos are used to let the user better distinguish the speakers from each other. The used colors for the speaker parties are assigned in the input file's metadata. An example of speaker-profile node is shown in the Figure 5.

Figure 5: Example of speaker-profile node.

6 STRUCTURE OF THE TOOL

This chapter presents the visual user interface. The interface is created, using JavaScript, HTML and D3 library. It consists of 6 tabs, representing 5 different visualizations, and an aggregation tab (shown in the Figure 6). The functionality, interactivity, local settings and the limitations of each of the 6 views are explained separately. Two sidebars are used to display the detail and settings information. They are described in the end of the chapter.

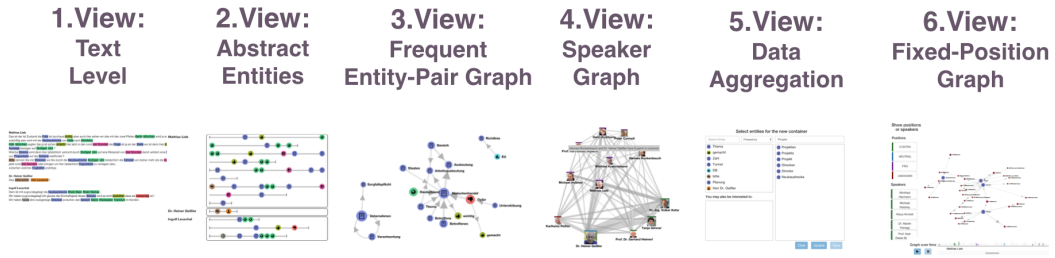


Figure 6: Views of the visual user interface.

6.1 GOAL OF THE APPROACH

The goal of the visual-interactive approach is to let the user get an overview of entity-pairs used in the discussion, filter relevant information, using different interaction methods, and get better understanding of the data. To achieve this goal, multiple interlinked views are created. As Keim, D.A., Kolhammer J., et.al. [36] state, the guide to visually exploring data by Shneidermann "Overview first, zoom/filter, details on demand" [13], in the context of visual analytics "can usefully be extended to "Analyse first, show important, zoom/filter, analyse further, details on demand" [25] indicating, that it is not sufficient to just retrieve and display the data using a visual metaphor; rather, it is necessary to analyse". [36]

The style of an interface, the shapes, fonts, colors, and graphical elements that are used and the way they are combined, can influence the emotional impact to the user. [32] Therefore the tool is created in plain manner, by offering more place for visualizations themselves and placing additional settings and details in the sidebars.

To support the user and ease the exploration process, it is important to provide a constant feedback. [32] Therefore, if additional user input is needed, the system shows an alert message, reminding the user about the missing step. If after the execution of some function no results are found, an alert message is shown. A spinner is used to indicate, that the operation is still in progress.

6.2 1.VIEW: TEXT LEVEL

The first visualization represents the whole text, the entities being colored in appropriate colors, with respect to their categories (shown in the Figure 7). Each utterance is separated, the speaker's name is placed on the left side of it. This visualization lets the user read the original document. Visualization is supposed to serve as the source of original text, therefore only one interaction is possible - by selecting a color-element in the color legend the particular category is highlighted or faded out.

LIMITATIONS The visualization serves as the source of original text. As the text may be large, it is important to create the visualization efficiently. The view consists of large number of elements, as each of the entities is created as a single span tag. Therefore the creation of these elements for large files may be time consuming. Although the DocumentFragment is used to reduce the number of nodes added to the DOM, the improvement of the performance is still desired.

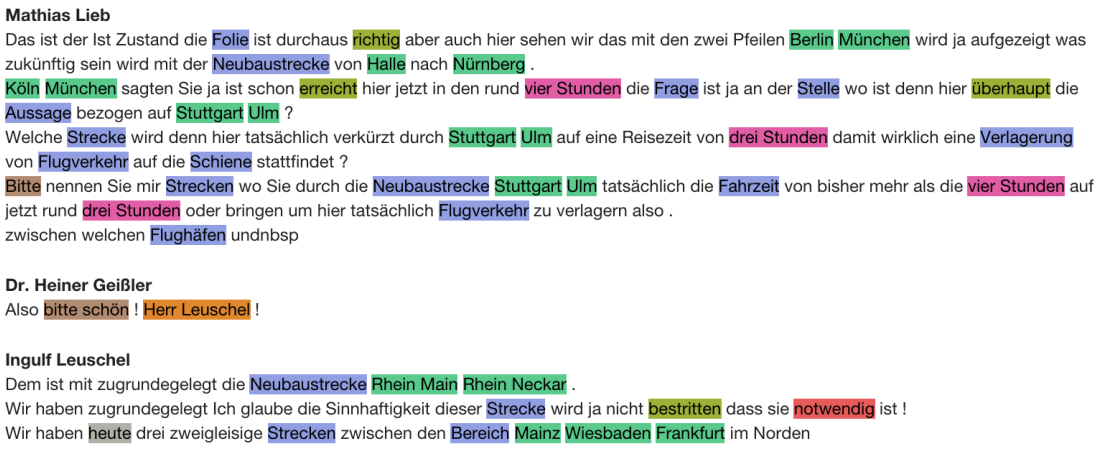


Figure 7: View with text and colored entities.

6.3 2.VIEW: ABSTRACT ENTITIES

The second visualization (shown in the Figure 8) represents the whole data too, however this one is created in an interactive manner. The main components of this visualization are rectangles, representing utterances, filled with horizontal lines and circles. The horizontal lines serve as sentences, length being scaled to the number of words in the particular sentence. The vertical lines present the beginning and end of the sentence. The circles represent entities, tooltip shows the token or phrase of it. The first visualization is partly included in this one - by hovering over a sentence line, the text of this sentence is shown on the left side of the view, letting to know what exactly is said. An overview of all present utterances is placed on the right side of this visualization. They are represented as horizontal lines with the length scaled to the number of words in the utterance. A tooltip shows the speaker name of the particular utterance.

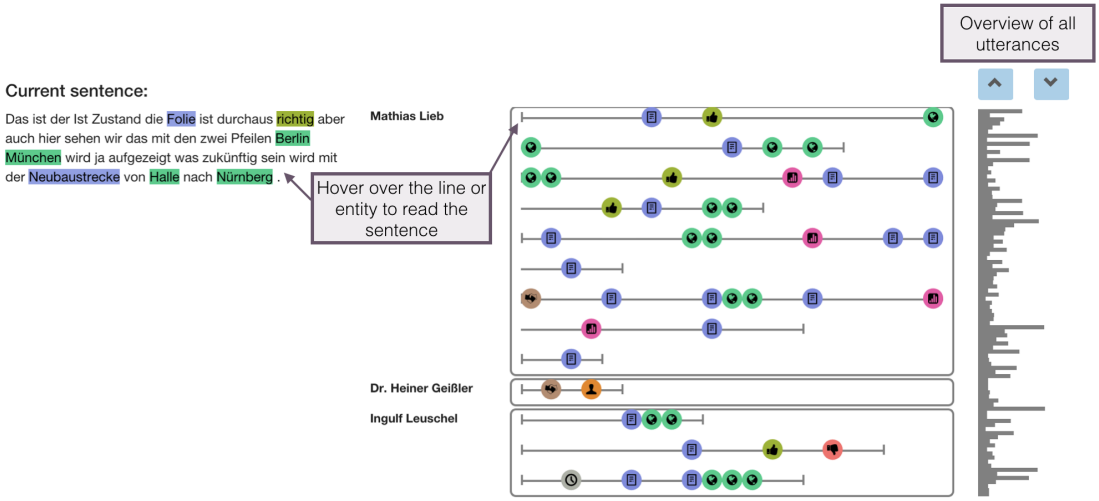


Figure 8: Entity abstraction view.

INTERACTIVITY The circles, representing single entities, serve as filter. The left-click on a single entity highlights the selected entity in all utterances, others being faded out. An example with highlighted entity pair is shown in the Figure 9.

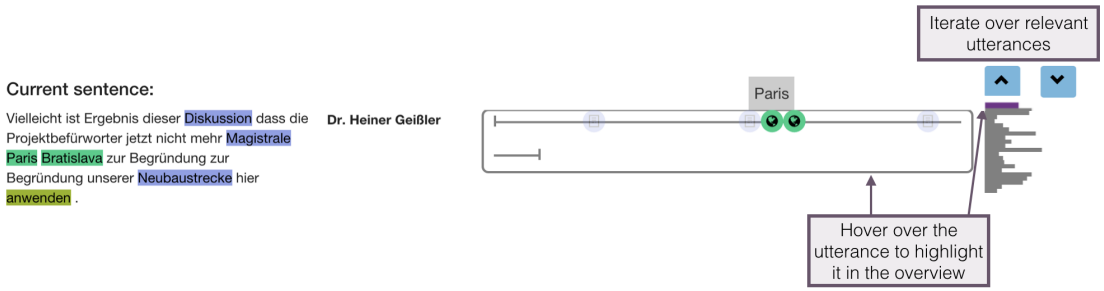


Figure 9: Highlighted entity-pair *Paris-Bratislava* in abstract entities view.

The utterances on the left and the overview of utterances on the right side are connected. By hovering over an utterance on the left, the appropriate utterance in the overview is scrolled to the top of the view-port and is highlighted with respect to the particular speaker party's color. That lets the user know, where the currently explored data-section is located in the discussion. The connectivity of these two visualization parts works the other way around too. By clicking on an utterance in the utterances overview, the view on the left side is automatically scrolled to the appropriate position. That helps the user retrace where the concrete utterance is positioned in the discussion, and easily find the utterances which contain the filtered elements. The 1.view is linked to this one - the text with colored entities is scrolled to the position of currently clicked utterance in the utterances overview, letting to explore the content in more details.

When a single element is filtered, or a new concept is visualized, the utterances containing this data are highlighted. The view is scrolled to the position of the first

relevant utterance. The user may iterate over the relevant utterances, clicking on the buttons, placed above the utterances overview. The iteration simplifies the exploration of the relevant data.

LOCAL SETTINGS User may filter utterances, where some information about politeness is present. This is done in following way: all utterances are highlighted, where the speaker has mentioned another speaker, who has spoken 1-2 utterances before or 1-2 utterances after the current one. Entities, categorized as Politeness-Indicator and Person, are highlighted, letting the user explore, which speakers are directly addressing statements to other participants and if they are doing it politely.

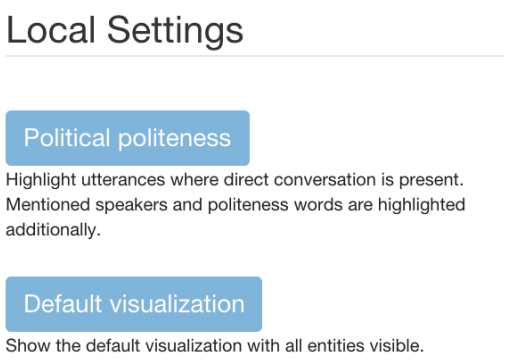


Figure 10: Local settings for abstract entities view.

By the left-click on the "Default visualization", all entities of the visualization are highlighted. The view of the local settings is shown in the Figure 10.

LIMITATIONS The number of circle elements being displayed in the visualization depends on the number of extracted entities. Due to large amount of elements, the performance of the visualization's display may decrease. One of the solutions could be to display only the data visible in the view-port and update it, when the view is scrolled, so reducing the number of elements needed to render.

6.4 3.VIEW: FREQUENT ENTITY-PAIR GRAPH

The third visualization represents all entity-pairs in the discussion, being visualized as a force directed graph (shown in the Figure 11). Graph layout supports relation representation, where a single entity is represented as node and relation - as link. This visualization gives an overview of all frequent entity pairs, giving a clue, what is the topic of the discussion.

Not only the main topic of the discussion, but also smaller subtopics might be relevant to present the content in more details. Not always the smaller subtopics are characterized with frequent entities though. Such subjects like the costs of the project could not be found within frequent entity-pairs, as the entities representing the costs could be quite diverse (e.g. *100 Millionen Euro*, *150 Millionen Euro*). To get an idea about discussed topics, the user may reduce the number of categories, display the graph with entity-pair frequency "1" and get an impression of used entities.

The frequency of a data element is a quantitative value, therefore the size is appropriate mapping variable to use. Radius of the node represents, how often the particular

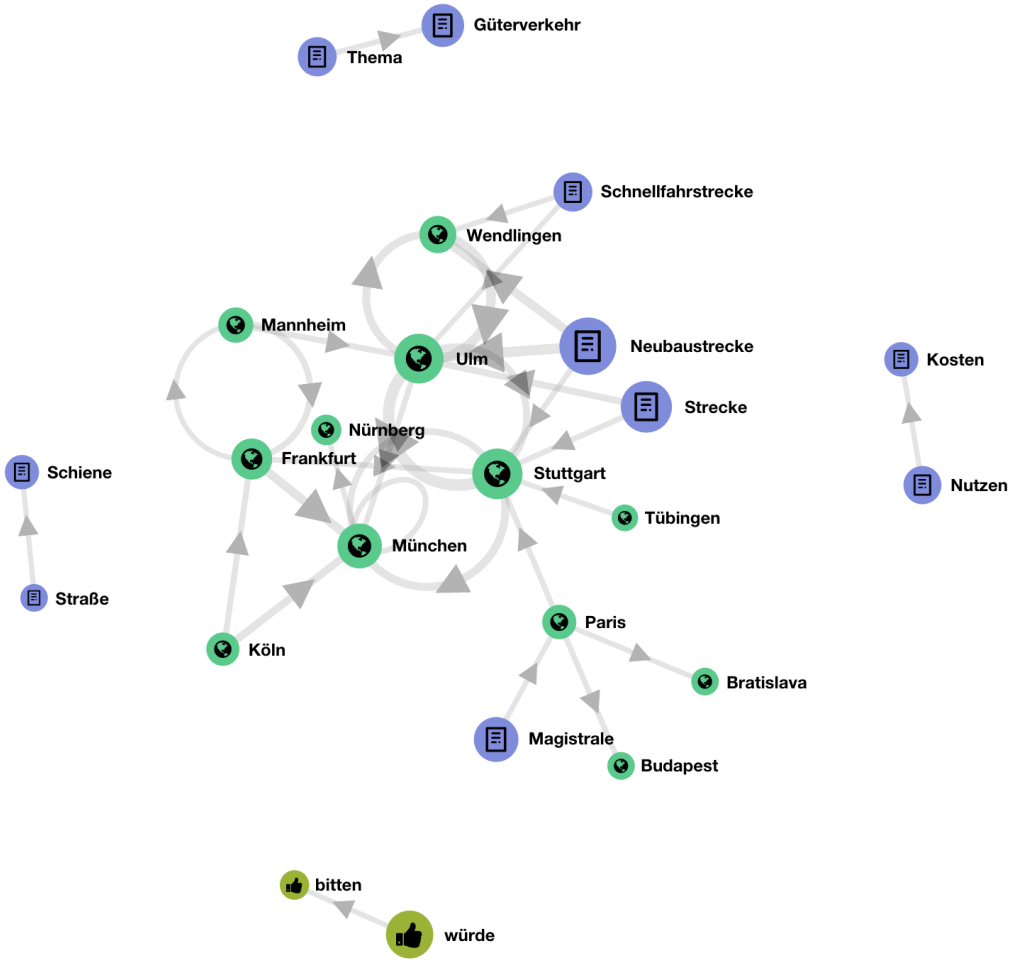


Figure 11: Frequent entity-pair graph.

entity is mentioned in the discussion, the width of a link represents how frequently the entity-pair is used. In processing step calculated entity-pair weight is mapped to the length of the link, representing the average distance between particular entities in a pair. Tooltips show the frequency values.

INTERACTIVITY The user may select a single entity, then the entity is highlighted in the abstract entities view (2.view) and in fixed-position graph (6.view), if present. The user may select an entity-pair by clicking on the edge between two entities. Then the selected pair is highlighted in abstract entities view (2.view), in fixed-position graph (6.view), if present, and also in speaker graph (4.view) - all speaker pairs are highlighted which have been mentioned the selected entity-pair. An example of selected elements is shown in the Figure 12.

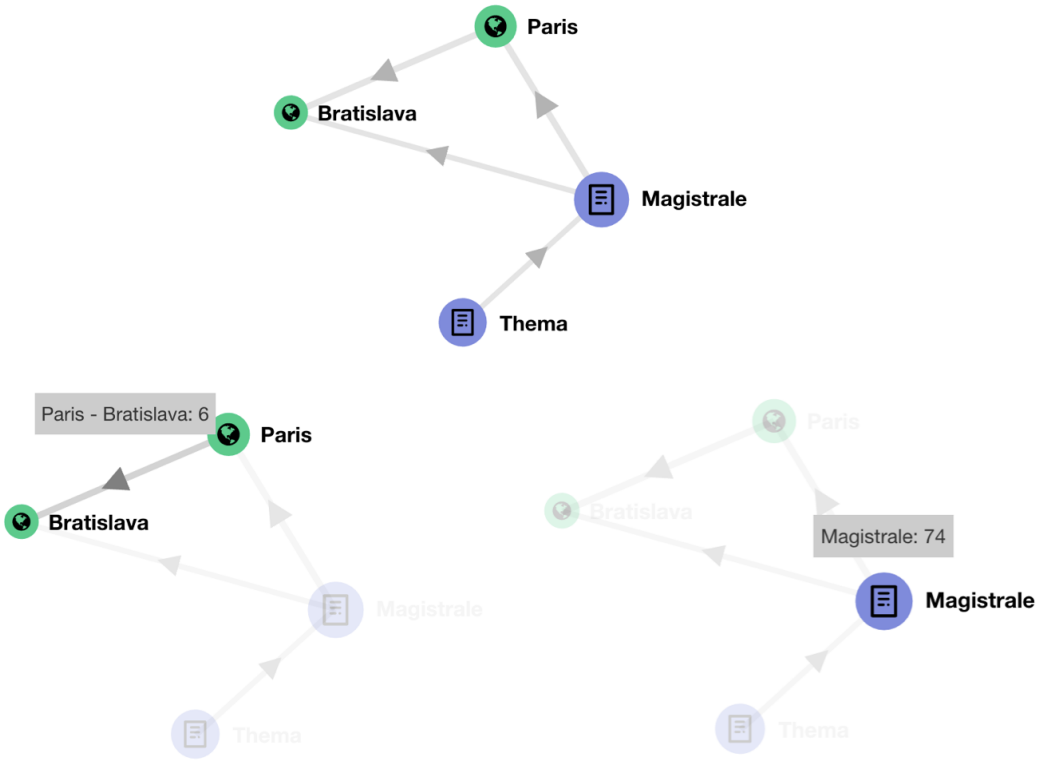


Figure 12: Selection and highlighting of single elements in the frequent entity-pair graph.

To improve the readability of the view, the user may double click on a node, then all entities connected to the selected one are highlighted, others being faded out. Additionally, the edges between the neighbor entities are highlighted, if present. The user may also hover over a node or edge, when the whole graph is displayed - then the neighbor entities are highlighted. By scrolling on the graph, it is zooming in and out, and by dragging the background of the graph, it is moving.

LOCAL SETTINGS In default all pairs which are mentioned at least three times in the discussion are displayed. However the user may change this parameter and show all entity-pairs or only those pairs, which are more persistently used.

The user may control the label charge value parameter. By adopting the charge value, the overlap of nodes may be avoided.

The default visualization with all entities visible may also be shown, if needed. The view of the local settings is shown in the Figure 13.

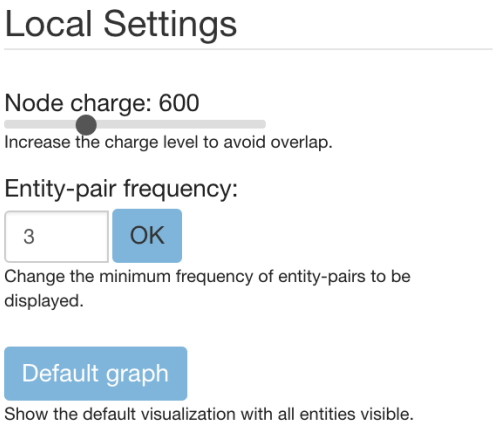


Figure 13: Local settings for frequent entity-pair graph.

LIMITATIONS The graph is useful to display the most frequent entity-pairs, letting the user observe the main topic of the discussion. By decreasing the number of minimum frequency, and showing only single category, smaller subtopics may be found. However to explore these topics in more details another visualization is needed.

One of the visualization's drawbacks is the decrease of the readability, if all entity-categories with low frequency are displayed. Therefore some data reduction, representing only the most important entities is needed. One of the solutions could be to highlight those entities, which are more "trendy" for the particular topic. It could be done, by decreasing the visibility of entities, which are frequently used in the German language.

6.5 4.VIEW: SPEAKER GRAPH

The fourth visualization represents speakers, who are using common entity-pairs, being visualized as force directed graph (shown in the Figure 14). Speakers are represented as nodes, which are connected by an edge, if they have at least one entity pair in common.

The size of the speaker node represents the number of utterances, being said by him. The thickness of the edge shows the number of common entity-pairs. By hovering over an edge the user may explore, which entity-pairs are said by both speakers. The distance between two nodes represents the speaker-positions. Speakers representing the same party are located closer to each other as speakers, representing different parties. It lets the user get an overview, if speakers from the same party are talking about the same topics, or exactly opposite.

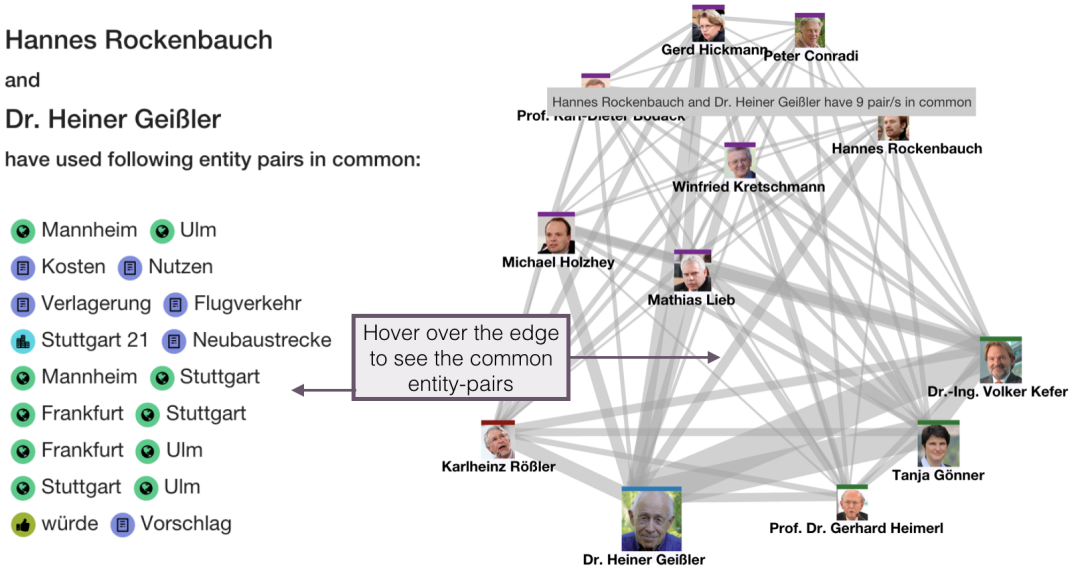


Figure 14: Speaker graph.

INTERACTIVITY The common entity pairs, which are displayed on the left side of the view, are interactive. The user may select a single pair and see which speakers have mentioned it, observing the changes of the displayed graph (shown in the Figure 15). This pair is highlighted also in the 2.view, in 3.view and in 6.view, if present. The user may also select a single speaker and observe in the abstract entities view (2.view), how active he is taking part in the discussion and which entities he is using. The graph can be zoomed in and out, and dragged similarly as the frequent entity-pair graph.

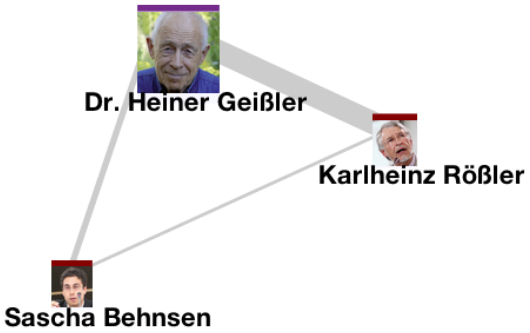


Figure 15: Highlighted entity-pair *Paris-Bratislava* in speaker graph.

To improve the readability of the view, the user may double click on a node, then all entities connected to the selected one are highlighted, others being faded out. The user may also hover over a node or edge, when the whole graph is displayed - then the neighbor entities are highlighted.

The user may explore the profile of the speaker in more details, by selecting a speaker node and opening the detail sidebar.

LOCAL SETTINGS The user may control the label charge parameter. By adopting the charge value, the overlap of nodes may be avoided. The default visualization with all entities visible may be shown, if needed. The view of the local settings is shown in the Figure 16.

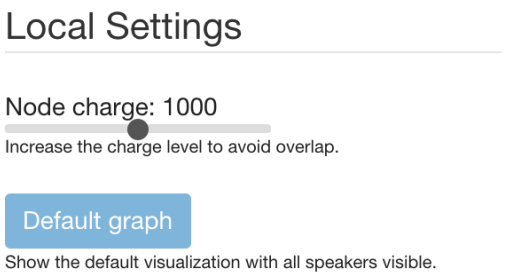


Figure 16: Local settings for speaker graph.

LIMITATIONS The visualization gives an opportunity to compare different speakers with respect to their used entity-pairs. The graph shows, if speakers from the same or different parties are talking about similar topics.

The limitation is in the fact, that a single subtopic may be represented by different entity-pairs. Therefore, even if speakers are discussing the same subtopic, but using different entity-pairs, this information is not visible. We have tried to solve this drawback by letting the user visualize single concepts in a new visualization, where speaker profiles are included. Some additional features, to represent similar (but not equal) entity-pairs, should still be added to this visualization though.

One of the solutions could be to introduce the system with smaller subcategories. The category "Measure" could be divided in such subcategories like "Money", "Distance", "Weight" etc. In such a manner all entities classified as one of the subcategories would be seen as similar. Therefore, if multiple speakers were talking about "Money" (e.g. *Strecke 100 Millionen Euro*, *Strecke 200 Millionen Euro*), these entity-pairs would be seen as similar and displayed in the graph, even though they were not equal.

6.6 5.VIEW: AGGREGATION OF ENTITY-PAIRS

Data aggregation, creation of containers.

Even for relatively small data set, the number of mentioned entity-pairs can be high. Although the user may control the pair-frequency parameter, reducing the amount of visualized data and improving the readability of the frequent entity-pair graph, sometimes exactly infrequent entity-pairs are important. Infrequent use of expected entity-pair might mean, that the speakers are avoiding to talk about the particular topic (e.g. infrequent use of entity-pair *Paris-Magistrale*, *Bratislava-Magistrale* in Stuttgart21 discussions). Therefore other possibility of data reduction is needed, to increase the readability of the entity-pair graph visualization, without losing relevant data.

To reduce the amount of visualized data, we offer the user an option to aggregate the entity-pairs. Aggregation-tab lets the user create containers with only relevant data for some specific concept. These containers are saved on the server side, for each user separately, allowing to reuse them in the following sessions.

The system supports the user, when entities are selected. First of all, the user is able to order the entity list alphabetically, by category or entity-frequency. The user may also query a single entity. By default, all entities containing the same lemma as the selected one, or edit distance less than the set threshold, are dragged automatically too. The system provides additional information which could help to choose entities for the container. When a single entity is dragged, then topic descriptors of the particular entity's topic (if it is present in at least one topic) are suggested. An example of the creation of new container is shown in the Figure 17.

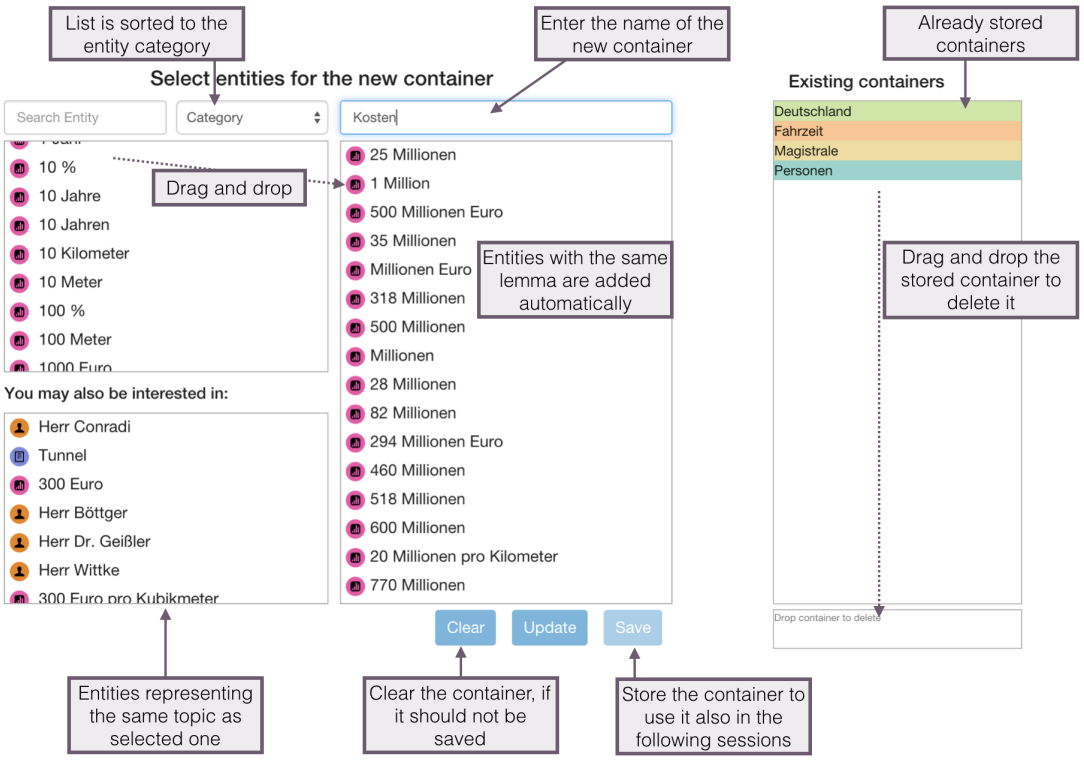


Figure 17: Workflow to create a new container.

When all relevant entities are chosen, the user needs to name the container and save it. Afterwards the color of the container may be set (shown in the Figure 18). The selection of color is done by the user (selecting a color from color-picker), as the number of created containers is not controlled by the system. Therefore multiple containers may be colored in the same color. The color may be changed limitless.

The user may also observe, which entities are stored in already saved containers, and update them (shown in the Figure 19). By dragging a single container to the middle of the view (on the part, where new container may be created), all entities present in the container are displayed. The user may add or remove entities from the container and update it, or rename it and save as a new container.

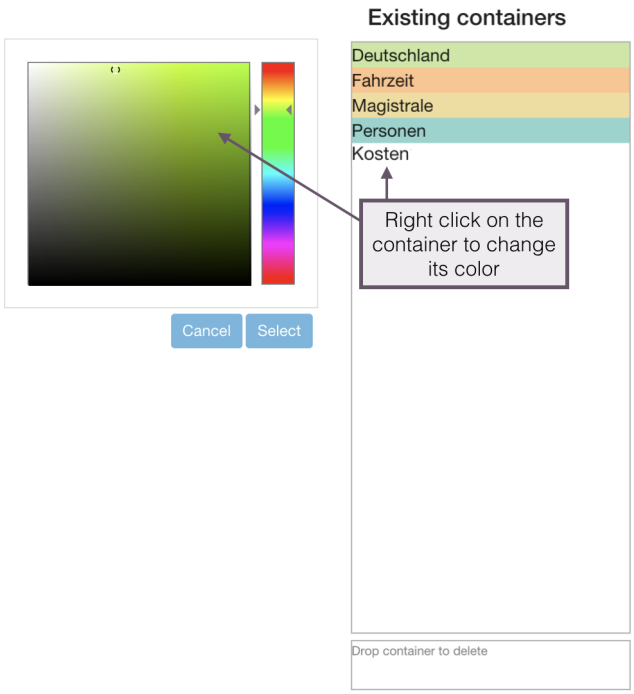


Figure 18: Workflow to set a color for the new container.

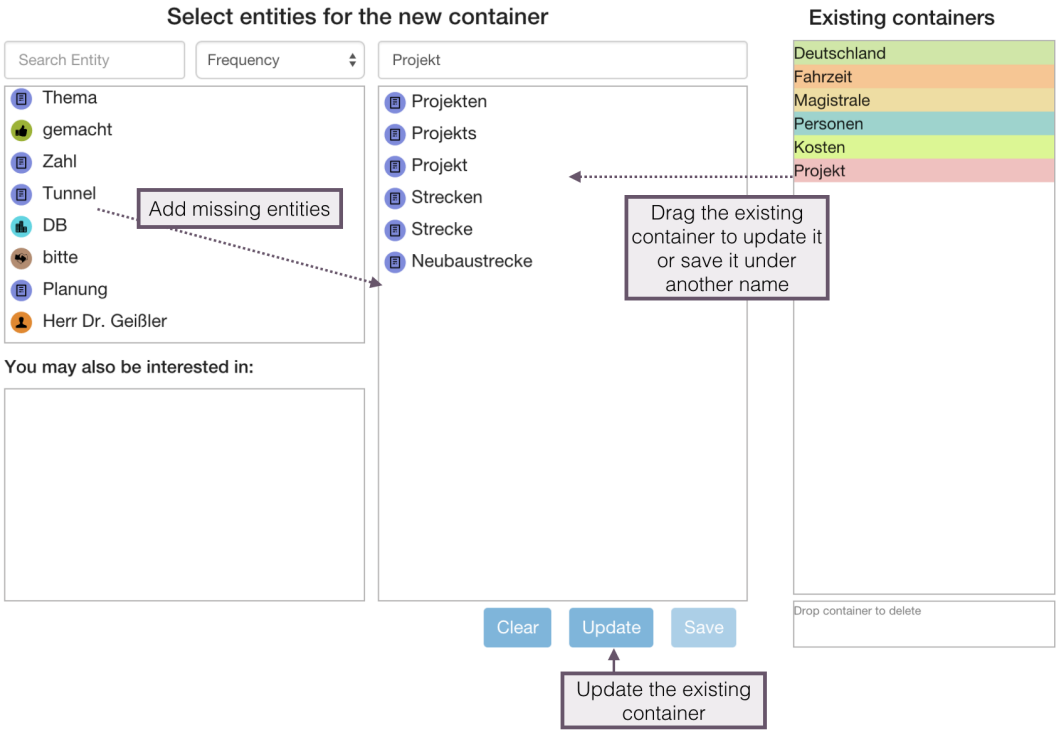


Figure 19: Workflow to update existing container.

After at least two containers are created, the user may select them to search for entity-pairs (shown in the Figure 20). Selection of multiple containers is also possible. Then the entity-pairs found are those, where each entity represents different container.

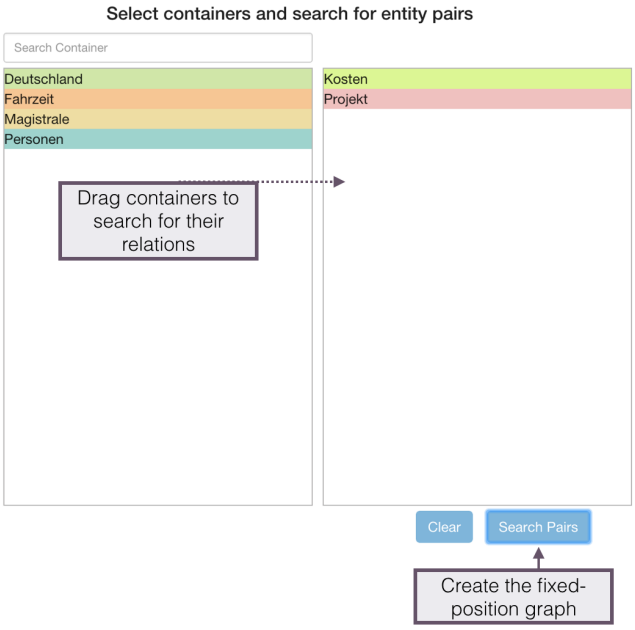


Figure 20: Workflow to create fixed-position graph.

The entity-pairs are visualized in a new tab as forced-directed graph. These pairs are highlighted in the abstract entities view (2.view) and frequent entity-pair graph (3.view). Speakers who have mentioned these entity-pairs in their talks are also highlighted in the speaker graph (4.view).

LOCAL SETTINGS The user may control the parameter, which sets the similarity distance between entities. The user may also choose to add only the selected entity to the new container. The view of the local settings is shown in the Figure 21.

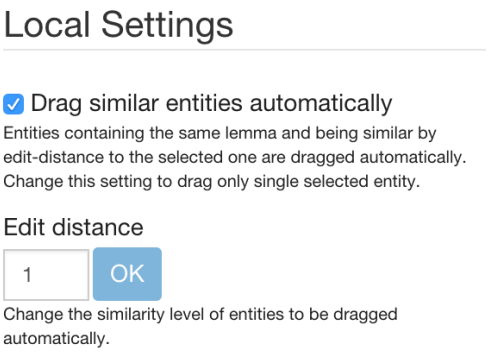


Figure 21: Local settings for data aggregation.

LIMITATIONS The system supports the user, when a new container is created or existing one is updated. However, additional features are desired. The user should be able to query not only a single entity, but also an entity-category. When an entity or category is queried, then the ability to add the whole search result to the new container would improve the usability.

The decision, which entities are good representatives for the particular subtopic, is not simple. Therefore additional features to the already used topic descriptors should be proposed. One of the solutions is to display those entities in a separate list, which are related to the same entities as the currently selected one (e.g. if the user selects *Neubaustrecke*, then the system suggests *Strecke*, as both of them are related to *Wendlingen*, *Ulm*, *Stuttgart* etc.).

6.7 6.VIEW: FIXED-POSITION GRAPH

Visualization of container-relations.

Although the reduced data is visualized similarly as all entity-pairs - by using a force directed graph (shown in the Figure 22), this graph has some important additional features. Firstly, the containers of single entities are visualized, using border color/s of nodes, bearing the container information.

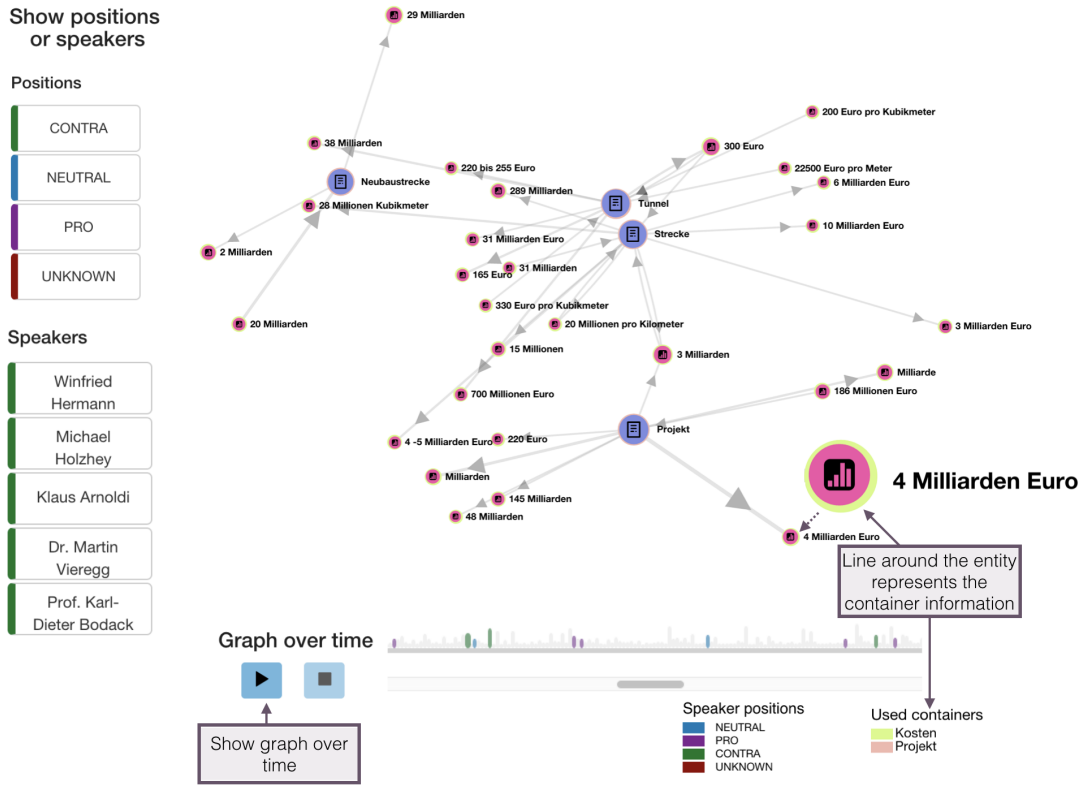


Figure 22: Fixed-position graph.

Secondly, entities, which are classified as Geo-Location are positioned with a fixed position, with respect to the actual location's geo-coordinates. Multiple tools for named-entity exploration exist, which use the information on their Geo-Locations. Most of them use the Google-Maps API, representing the locations in a separate Google Maps visualization. Our approach differs from the existing tools though. Our aim is to incorporate the locations in a graph visualization, not losing the context, in which locations have been mentioned. They are displayed just like other nodes of the forced

directed layout graph, but additionally being fixed to the relative geo-coordinate position to each other. That lets the user get an overview of the discussed territories, and see, which entities from other categories are related to the particular locations (e.g. *Neubaustrecke Wendlingen Ulm*).

To receive the location geo-coordinates we use the GeoNames[4] database, that contains the geo-coordinates of the countries of the world and of the cities with population greater than 1000. However, the position of the location is not always relevant. In that case the user may ignore the Geo-Location coordinates, displaying the basic force directed graph.

This visualization lets the user compare the single positions of speakers, as the speaker profile nodes are made fixed, when dragged to a new position. With the help of the force layout, the user may see, which entity-pairs have been mentioned by the particular speaker or speaker's party.

The user may explore the graph over time, observing the development of the single subtopic. The timeline presents information, which speakers in which point of the time are talking about the particular topic. An example of the visualization is shown in the Figure 23.

INTERACTIVITY The user may select a single entity in the graph, then the entity is highlighted in the abstract entities view (2.view), and in the frequent entity-pair graph (3.view), if present. The user may select an entity-pair by clicking on the edge between two entities. Then the selected pair is highlighted in 2.view, in 3.view, and in speaker graph (4.view) - all speaker pairs are highlighted which have mentioned the selected entity-pair. The user may explore in the abstract entities view, when the particular speaker has mentioned entity pairs containing the particular entity by selecting an edge between speaker profile and entity. That improves the observation process - the user does not need to explore single entity one after another, but complete concept may be explored at once.

The user may explore the profile of the speaker in more details, by selecting a speaker node and opening the detail sidebar.

The user may explore the use of entity-pairs over time. The entity pair, particular speaker and position in the discussion are highlighted. When highlighting over time is started, the user may select other position in the discussion, from which the highlighting should be moved on. He may pause the highlighting and explore already visualized part of the concept. Then, by selecting an utterance in the timeline, the view with abstract entities is opened and the respective position is shown.

To improve the readability of the view, the user may double click on a node, then all entities connected to the selected one are highlighted, others being faded out. Additionally, the edges between the neighbor entities are highlighted, if present.

4 http://www.geonames.org/, accessed on 01.12.2015

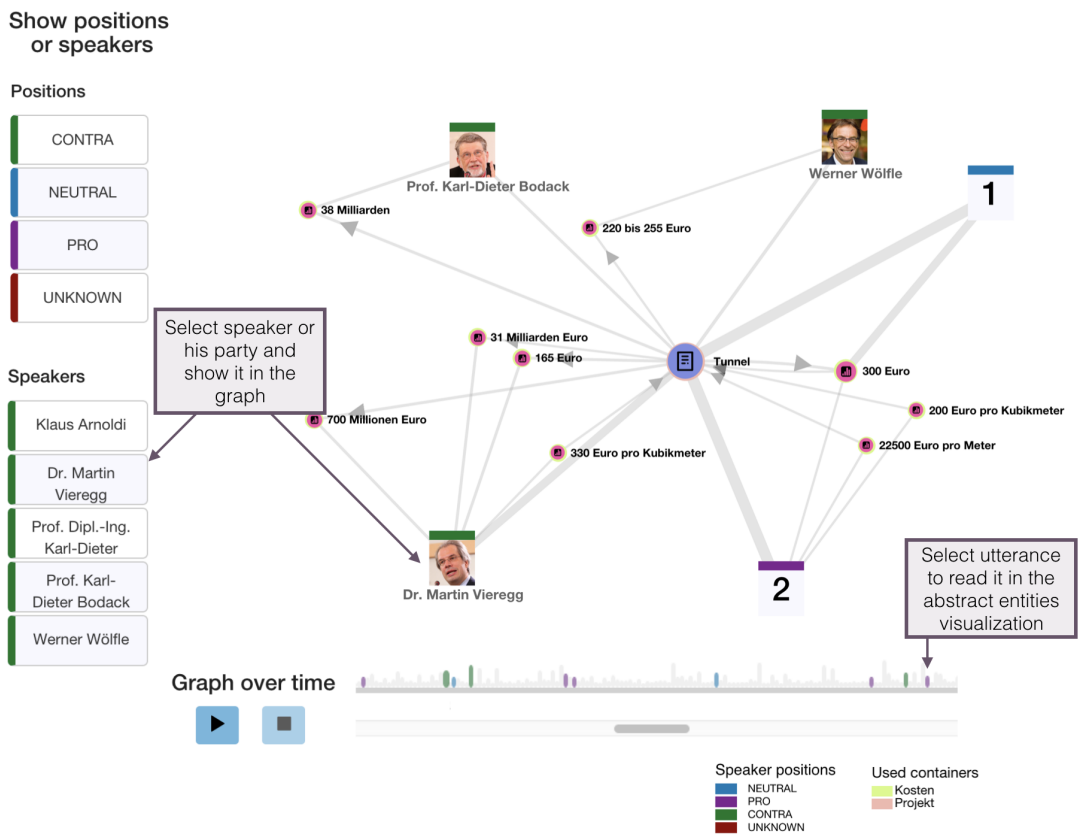


Figure 23: Fixed-position graph with integrated speaker profiles.

The user may hover over a node or edge, when the whole graph is displayed - then the neighbor entities are highlighted. The graph can be zoomed in and out, or moved to the sides.

LOCAL SETTINGS The user may control the label charge parameter. By adopting the charge value, the overlap of nodes may be avoided. The user may adapt the speed of the transitions, which represent the changes of the graph over time.

The default visualization with all entities visible may be shown, if needed. The view of the local settings is shown in the Figure 24.

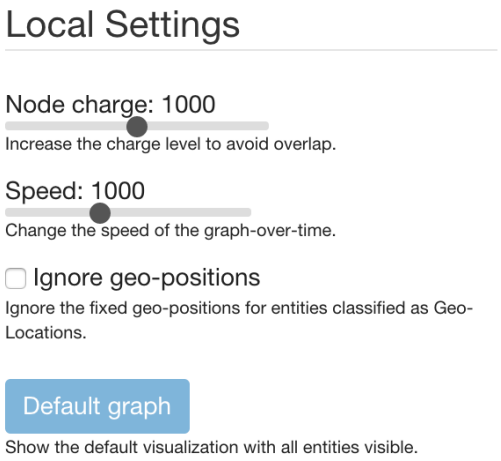


Figure 24: Local settings for fixed-position graph.

LIMITATIONS This visualization lets the user observe smaller subjects or concepts, which the user might not explore in other visualizations. It combines the force-directed layout with speaker profile nodes, allowing to compare the single speakers and parties. The timeline shows, how the particular topic develops over time.

The visualization still has drawbacks. The readability of the view depends on the size of displayed containers. The more nodes and edges are visualized, the more difficult it becomes to get a quick overview of the displayed relationships. Therefore some improvements for the entity-pair representation is needed. One of the solutions could be to visualize an entity-pair as a single node. It would not decrease the number of nodes displayed in the visualization, but the observation of pairs might be improved though. However, due to the use of fixed positions to represent the Geo-Locations, it would be problematic to display multiple pairs on the same position. Therefore an alternative solution is needed.

6.8 METADATA

6.8.1 *Color Legend*

In the footer of the window color legend is placed. In the development process, the position of it was changed from the left side to the bottom of the window, leaving more horizontal place for the visualizations. The legend is divided in two parts: all entity categories are displayed on the left side,- and all speaker positions in the discussion are displayed on the right side. The color items for categories serve as filter - the specific category may be blended out in the first three views, if needed. The purpose of this is to reduce the number of represented colors in the visualizations, hence, increasing the level of perception on the kept categories.

6.8.2 *Setting Sidebar*

As the explored data may be large, it is important, that the maximum of the existing space of the window can be used for the display of visualizations. Therefore all settings are placed in the right sidebar.

On the top of the sidebar currently filtered element is displayed, helping the user retrace, which element is visualized.

The sidebar is divided in global and local settings. The global settings (shown in the Figure 25) include filtering of entities, entity-pairs and speakers. These elements are then highlighted in the abstract entities visualization, entity-pairs graph, speaker graph and fixed-position graph.

Before filtering a single entity, the user may sort the entity-list alphabetically, by frequency or category. The list containing entity-pairs may be sorted alphabetically or by frequency. The entity, entity-pair or speaker may also be queried, simplifying the search for a specific string.

The order of entities in the entity-pair is relevant. Therefore, if the order should be ignored (e.g. if by selecting pair *Wendlingen Ulm* both pairs *Wendlingen Ulm* AND *Ulm Wendlingen* should be highlighted), the user may select the checkbox 'Ignore Direction'.

The local settings include the specific settings for each visualization separately. They were described after the representation of the particular visualization.

6.8.3 *Detail Sidebar*

The details sidebar offers additional information on single entities and speakers. When an entity categorized as Geo-Location or Organization is selected, the abstract from DBpedia is shown, if present.

If a single speaker is selected, then speaker profile is shown, consisting of short information (added by the user) on his profession, his party and the most frequently used entity-pairs (an example shown in the Figure 25). Entity-pairs work as filter.

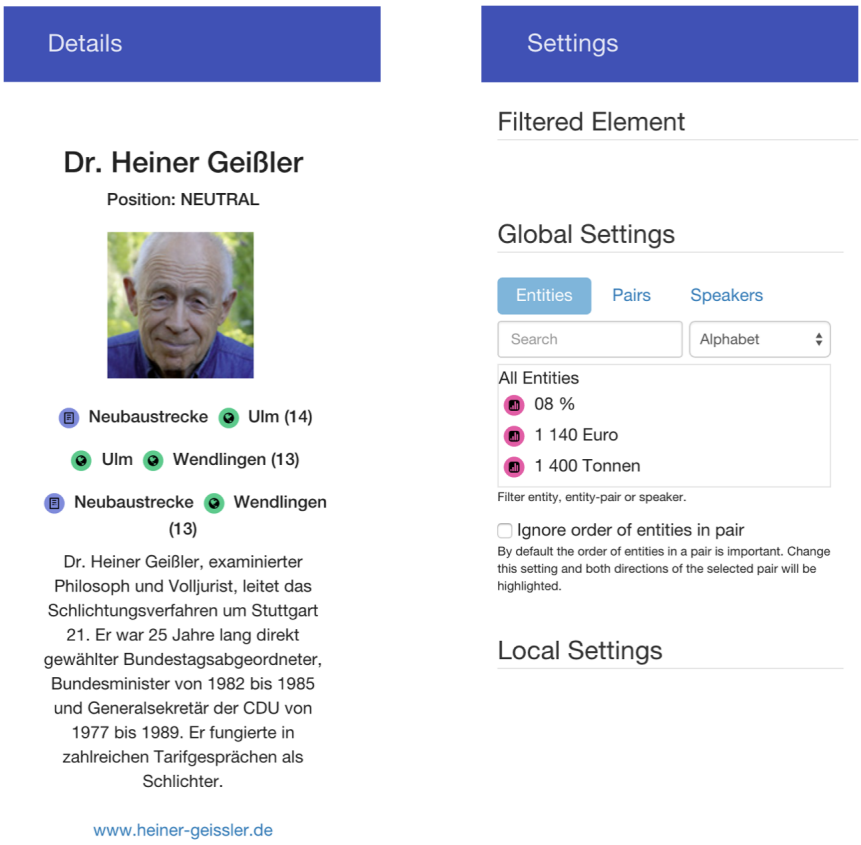


Figure 25: Detail and setting sidebar.

7

EVALUATION, USER STUDIES

To evaluate our approach, we carried out user studies with seven participants (four PhD students of political sciences, one master student of political sciences and two bachelor degree students of computer sciences).

Each person received an example of political discussion (50 pages) and multiple questions regarding the discussion. They were allowed to use existing tools or just read the data, and they were suggested not to spend more than 1 hour for the answers. Afterwards each person was met separately to carry out the study. They were introduced with the tool and had time to try out its functionalities. Afterwards they answered the same questions as previously, and additionally executed some example use cases. Examples are shown in the Appendix A. Five of the seven participants used the tool on the same discussion dataset, which they read before. Their findings using the tool were compared with their previously gathered knowledge of the text. Four of the participants explored additional discussion text, showing, if the tool helps to understand discussed topics of unknown dataset.

Their executed steps and methods were registered. The participants were suggested to think aloud and explain, why they are using the particular method to answer the question. Additionally, the participants filled a questionnaire about the usability and visual design of the tool. The user studies of each participant took 1-1,5 hours.

7.1 THE OBJECTIVES OF THE EVALUATION

The size of the text may be a challenge, when a quick overview of the discussed topics is needed. Knowing the topics of the text could ease the decision, if the particular document is important, and if it should be read and observed in details. However, in situation when user is already familiar with the topic of the text, he might want to review, if expected subtopics or subjects are really mentioned and how they are used by different speakers over time.

Therefore, the aim of the evaluation is to find out, if the user can conclude what is the main topic of the discussion, by exploring most present entities and their relations in the text, and additionally to observe the most active and similar speakers. It is important to see, if entity-pairs may reveal additional information and help to understand the text better than single entities. We are also interested, if our approach supports the user to understand the topic in more details and if it presents some smaller subtopics to be explored. It is also important to find out, if the user may get an understanding of the development of single subtopics, and if the similarities and differences between speakers and speaker parties with respect to the particular subtopics may be observed.

7.2 UNDERSTANDING THE DISCUSSION'S TOPIC

All participants stated, that the tool gives an overview of the discussed topic. For the determination of the topic, each of them observed the frequent entity-pair graph, saying, that not only the single entities, but also the entity-relations may help to understand the discussed matter. Depending on the dataset and the explored topic, the entity-pairs may reveal additional information that single entities might not contain. The tool is suited to be used, when the data is too large, to be read, and when a quick overview of discussed topic is needed.

Two of the participants admitted, that when reading the text, without using any tool, they can't bear the whole information and some details on the topic are forgotten. They stated, that they can summarize the explored text better, when using the tool, as the most frequent entities and their relations are present and observable during the whole exploration process.

One of the participants found it problematic, to understand the relations in the graph, as the exploration of graph-visualizations as such is difficult and problematic for him. However another participant emphasized, that with the help of the network it is easier to quickly find the main descriptors of the topic.

The frequent entity-pair graph does not reveal complete information about the discussion though. If more information is needed, the users would explore the entity, or entity-pair list, sorted to the element frequency.

7.3 UNDERSTANDING AND EXPLORING THE SUBTOPICS

Almost all of the participants stated, that it may be possible to find out subtopics of the discussion, by exploring the frequent entity-pair graph. However it depends on the discussion text, and on the selected minimum frequency of the pairs.

In situations, when some subtopics are found, or when the user is already familiar with the discussion topic and is interested in specific subtopics, he should be able to explore them in more details. Participants confirmed, that the creation of containers and visualization of this specific data help to explore the subtopics. Yet the decision, which entities should be selected to find relations, is not trivial. It is easy to check, if speakers are talking about some subtopic, which is expected to be discussed. However it may be very time consuming to find out other subtopics, as the right entities need to be selected. Although the system suggests the topic descriptors, which might be added to the container, the decision, with which entity to begin, is quite problematic.

The timeline, representing the flow of the discussion, and the geo-coordinates (if Geo-Locations are present in the selected containers), support the user to understand the topic and observe the changes in the discussion's flow. Speaker's profiles help to compare different parties and speakers for the particular subtopic.

One of participants stated, that if too many containers are present and too many relations are visualized, then the readability of the fixed-position graph decreases

and the user must spend more time to understand the multilevel relations. Therefore improvements for this visualization are needed, to improve the perception of the entity-relationships.

7.4 COMPARING SINGLE PARTIES AND SPEAKERS

When a single topic or subtopic of the discussion is explored, the users confirmed, that it is possible to get an outline of the active speakers. To find out, which persons are talking about the subtopics, the participants observed the speaker pair graph, abstract entities view and fixed-positions graph, if containers were created.

One of the participants stated, that "speaker photos used to represent single speakers of discussion, help to remember the most relevant speakers, and to make better conclusions about the discussion flow". The participants found helpful, that the utterances where some filtered element is present, are highlighted in the speaker-party's color. It helps to see more easily, which parties are responding to others on a specific topic.

One of the participants stated, that she found more persons talking about the particular topic, when she was using the tool, than when she was simply reading the text. However another participant mentioned, that by reading the text, he noticed more speakers talking about the particular topic as the system would show. That means, that too few entity-pairs were found. Therefore, it leads to conclusion, that it is important, that the maximum distance between two entities in an entity-pair is set by the user, before the data is preprocessed. It lets him/her to try out the parameters to find the most appropriate distance for his needs, receiving more or less entity-pairs respectively. The maximum boundary should be the length of the sentence though, as the entities in the pair should still maintain the relatedness to each other.

7.5 USABILITY AND VISUAL DESIGN

We created a questionnaire about the usability and visual design of the tool, to find out, if it is intuitive to use and if some improvements could be done. As the results show (Table 2), most of the participants were satisfied with the usability and design of the system, and they confirmed that they would definitely use the system to explore conversation data. The participants pointed to additional functions and improvements, which could be desirable to make the usability even better.

First of all, three of the seven participants stated, that it is not always clear what is happening, when some function is executed. Partly it could be explained with the need to get used to the system and learn all functions and their execution steps. However, constant feedback for users is still very important. Especially then, when multiple views are updated simultaneously, the user should be provided with feedback, which views have changed their conditions. Additional spinners are created to support the user and show if the system is still executing some functions. Moreover, the filtered element is included in settings side-bar, so helping the user retrace, which element

Question	Strongly Agree	Agree	Disagree	N/A
I would like to use this system, if I needed to explore discussion text	7	0	0	0
Using the tool is intuitive	0	6	1	0
The various functions of the tool are well integrated	3	4	0	0
The system is inconsistent	0	1	6	0
I needed to learn lot of things before I could get going with the system	0	2	5	0
It is always clear what is happening	1	3	3	0
Included graphics are meaningful	4	3	0	0
The information on the screen is presented in a clear and pleasant manner	5	2	0	0
The meaning of buttons and clickable regions are easily perceived	2	5	0	0
Menu option titles match item to which they refer	7	0	0	0
Searching for single items is simple	4	3	0	0
There is a clear way to return to the starting point of the visualization	7	0	0	0

Table 2: The results of the usability and visual design.

is currently visualized. However the highlighting of all relevant interlinked views should still be done, when an element is filtered.

One of the participants expected, that, by highlighting speakers who have used selected entity-pair, the edges of the speaker graph should represent the frequency of the use of searched entity-pair instead of the information on all common entity pairs. Due to the lack of time to implement additional features, we added this one to the future work.

When a single word is queried in entity or entity-pair list, the participants would prefer, that all hits, containing the query, would be highlighted simultaneously. This feature is added to the future work.

7.6 POSSIBLE ADDITIONAL FEATURES

Although all of the participants admitted, that the system is good to find out the topics of the discussion, explore the development of subtopics in more detail and compare the different speakers and parties, most of them would like to have some additional features, which would reveal some statistical data about the discussion to work with. An extra view with some summaries and statistics would help to make conclusions about the discussion. It is added to the future work.

Additionally to the statistics view, it would be preferred if the data could be exportable in a CSV file or in other data format, to be used in another data exploration tool afterwards.

One of the users suggested, that some data reduction, like letting the user explore only the most important entity-pairs, could improve the perception of visualizations. Therefore a method to classify the entity pairs as important or less important is needed. One of the solutions could be to use some interestingness measure, where important entities are determined. This measure could be used to exclude such words, which are very frequently used in the German language. And the words, which are more specific for the particular topic could be higher rated. Because of the lack of time to implement additional features, this one is included in the future work.

Another user confirmed our idea, which has already been included in the future work list. It is important, that the user is able to correct the wrong entity-categories, if needed. This feature would improve the data quality and ensure more correct conclusions about the discussion text.

7.7 CONCLUSION ON THE EVALUATION

Although many suggestions and improvements were given, the goal of the evaluation is reached. The participants confirmed, that the system is useful in finding out the topic of the discussion, exploring the development of smaller subtopics and comparing different speakers and parties with respect to the particular subtopic. They admitted that entity-pairs help to understand the topics better than single entities,

"presenting more context information". All the participants stated, that they would definitely use this tool to explore the discussion data. Four of the seven participants confirmed, that they would like to use this system to explore other data than political discussions. They would like to explore the topics and their changes in a role-games data and transcripts of talk-shows "to observe the use of specific context keywords and look, how topic develops over time". Two of the participants admitted, that they would use the tool to compare different documents or papers, to find out the most relevant ones, which they would read afterwards.

The evaluation shows, that the tool can be used to answer questions about topics and speakers of the discussion. However, additional features and improvements of existing visualizations are needed to make the exploration process better and satisfy even more user needs.

8 CONCLUSIONS

This thesis presents an approach for exploring entity usage in multi-party conversation data. We use distance-restricted entity relation and show, that entity-pairs may be seen as new efficiently extractable entity relation, containing additional context information, what single entities may not reveal. This entity-relationship level may be seen as generalization of semantic-relation and restriction of the relation, where two entities are seen as related, if they are present in the same document. The restriction is done in the following way: entities are seen as related only if they are used close to each other in a single sentence. This restriction lets the user explore not only the main topic of the discussion, but also smaller subtopics (e.g. travel time between different routes, positive or negative attitude to the discussed topic, costs of the discussed project, specific concepts like *Magistrale Paris Bratislava* etc.), and compare utterances of single speakers.

The approach, consisting of five interlinked interactive visualizations, lets the user explore the data from different perspectives, and from different levels of details. The system is useful in finding out the topic of the discussion, exploring the development of smaller subtopics and comparing different speakers and parties with respect to the particular subtopic. The user studies have shown though, that additional features and extensions are desirable. First of all, data reduction, where only the most important entities are displayed, could improve the readability of the visualizations. Secondly, the visualization with fixed-position graph should be improved, as the readability decreases, when multiple containers are displayed. Additionally, some new features like statistical data about single speakers or utterances with respect to the mentioned entities and their-pairs could improve the approach and satisfy more possible users.

9

FUTURE WORK

The main focus of our approach lies on the interlinked visualizations, representing the discussed topics and similarities between discussion's participants. Although high precision and recall of extracted entities is not our goal, the extraction task might still be improved. First of all, some additional data resources (e.g. WordNet, DBpedia etc.) could be used for the creation of entity-gazzeters. Secondly, the user should be able to add untagged entities, or remove wrongly classified entities, so improving the data quality. In the graph visualizations, the category would be changed for all entities containing the particular string. In the text view and abstract entities view the category would be changed for a single entity, with respect to its position in the discussion.

If an entity or pair is queried in the sidebar, then all hits, containing the query, should be highlighted simultaneously. It would help to observe the complete subject simultaneously.

The representation of the speaker-graph could be improved. The speaker-graph currently displays those speaker pairs, which have at least one entity-pair in common. In situations, when speakers are talking about the same topic, using slightly different entities, the visualization comes to its limit. For this purpose smaller subcategories could be introduced. The category "Measure" could be divided in such subcategories like "Money", "Distance", "Weight" etc. In such a manner all entities classified as one of the subcategories would be seen as "similar" (e.g. *100 Millionen Euro*, *200 Millionen Euro*). Pairs containing an entity classified as one of the subcategories would be represented as "entity + subcategory" (e.g. *Strecke* Money). That allows the user to see, which speakers are discussing on similar topics.

The speaker graph could be expanded, showing the information about all pairs and entities used by a single speaker, when the node of his profile is hovered. Additionally, if a single entity-pair is filtered, the edges of the speaker graph should represent the frequency of the use of searched entity-pair instead of the information on all common entity-pairs.

The decision, which entities are good representatives for the particular subtopic, when a new entity-container is created, is not simple. Therefore additional features to the already used topic descriptors should be proposed. One of the solutions is to display those entities in a separate list, which are related to the same entities as the currently selected one (e.g. if the user selects *Neubaustrecke*, then the system suggests *Strecke*, as both of them are related to *Wendlingen*, *Ulm*, *Stuttgart* etc.).

The readability of the fixed-position graph should be enhanced. Solution for better representation of multilevel relationships is needed. Additionally, the user should be able to observe how the size of the nodes and links of the graph increase over time. It

would show, how frequently a particular entity-pair has been used in a specific point of time of the discussion.

To visualize the Geo-Locations in the fixed-position graph, the respective geo coordinates are used. Currently we are using the GeoNames database, however a web-service for geo-coding could be used to improve the performance.

As the evaluation of the tool shows, the participants of the user studies have pointed to many new features, which could be included to improve and expand the tool. First of all, additional view with some statistical data or simple listing of relevant entities or pairs for specific filtering could be added, to provide some concrete inferences about the data, without the need to interact with the tool greatly. Secondly, the statistical data should be exportable in a CSV file or in other data format, to be used in another data exploration tools afterwards.

The reduction of data could improve the perception of visualizations. Therefore a method to classify the entity pairs as important or less important is needed. Some trendiness measure could be used to exclude such words, which are very frequently used in the German language. Words, which are more specific for the particular topic could be higher rated. More important entities could be highlighted to emphasize their role in the text.

To create the tool even more intuitive, the specific representation of all updated views are needed, when a single element is highlighted. It could be done with the transition of the tab element, highlighting the updated tabs for a short time period.

A

APPENDIX

A.1 QUESTIONS FOR USER STUDIES

- Which words are crucial for the discussion?
- Which word pairs are crucial for the discussion?
- What might be the topic of the discussion?
- Which speakers are talking about the main topic?
- Which other subjects („subtopics“) are discussed?
- Which speakers are talking about similar subjects?
- Do they respond to each other, when discussing them?
- Do they respond more to the same party's speakers or to the different party's speakers?
- Do they respond politely?
- Which stories can you find?

A.2 USE CASE 1

- Which abroad locations are mentioned, being related to each other?

  „Die Region gewinnt international an Bedeutung: Anbindung an das europäische Hochgeschwindigkeitsnetz. Mit Stuttgart 21 und der Neubaustrecke Wendlingen–Ulm wird die Region Stuttgart an das europäische Hochgeschwindigkeitsnetz angeschlossen. Beide Projekte sind zentrale Teile der Magistrale Paris–Bratislava. Nur mit ihnen ist der Südwesten Deutschlands auch künftig an den nationalen und internationalen Fernverkehr angeschlossen.“ [Neubauprojekt Stuttgart–Ulm, Neue Strecken, neues Verkehrskonzept für die Region, Deutschland und Europa, DB]

- Is the *Magistrale Paris Bratislava* discussed in the talk?

  (Dr. Volker Kefer is a member of the management board of Deutsche Bahn AG, it is expected, that he talks about Paris-Bratislava.)

- Who is actually talking about it? Are these speakers representing different parties?
- Who begins this subtopic? Do other speakers immediately react to it or ignore it?

- What is actually said?
- What is the conclusion? How important is this *Magistrale* for the Stuttgart21 project?

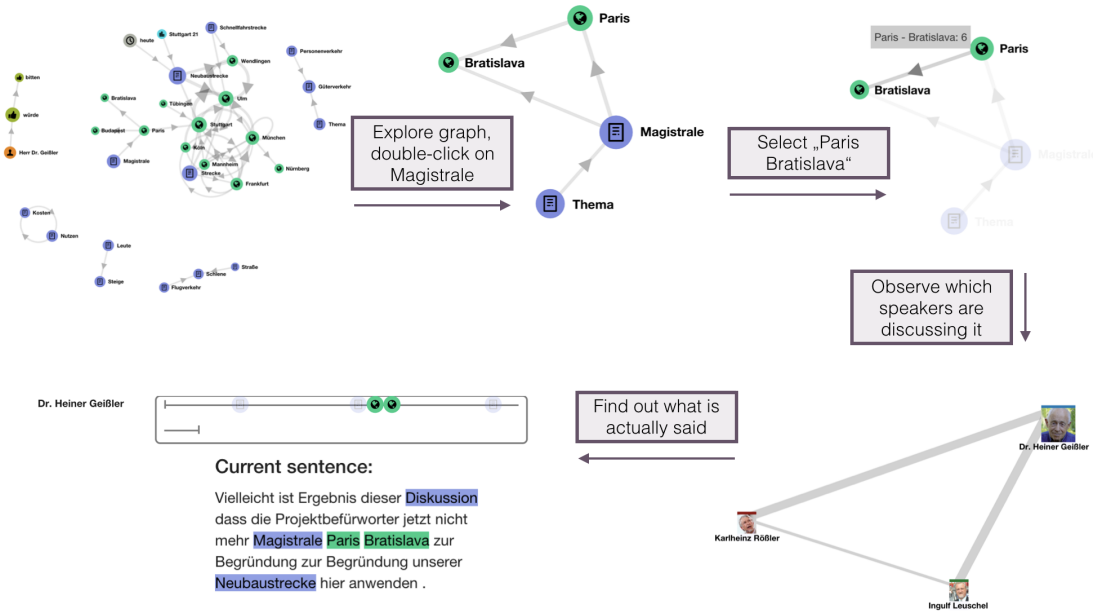


Figure 26: Use case 1 "Role of the *Magistrale Paris-Bratislava* in Stuttgart 21 Project".

A.3 USE CASE 2

- Which regions are mentioned in the discussion?
- Which problems are discussed? How can they be categorized? Which locations are related to them?
- Which organization is mentioned, being related to poisoning?
- Does someone suggest how to fight the existing problem in Germany?

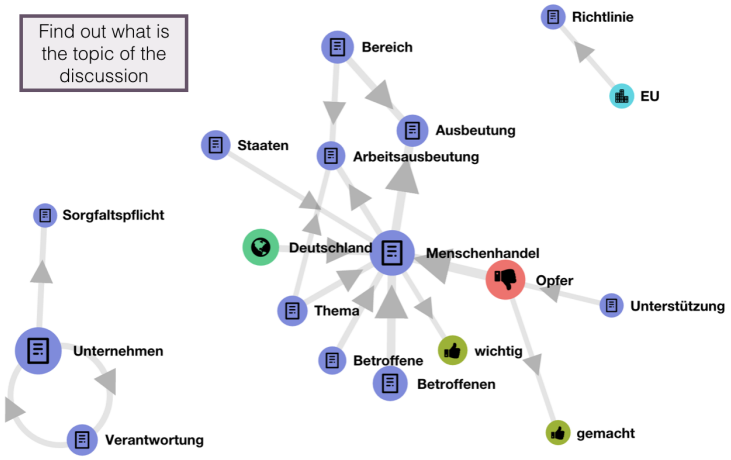


Figure 27: Use case 2 "Exploring the topic Human Rights".

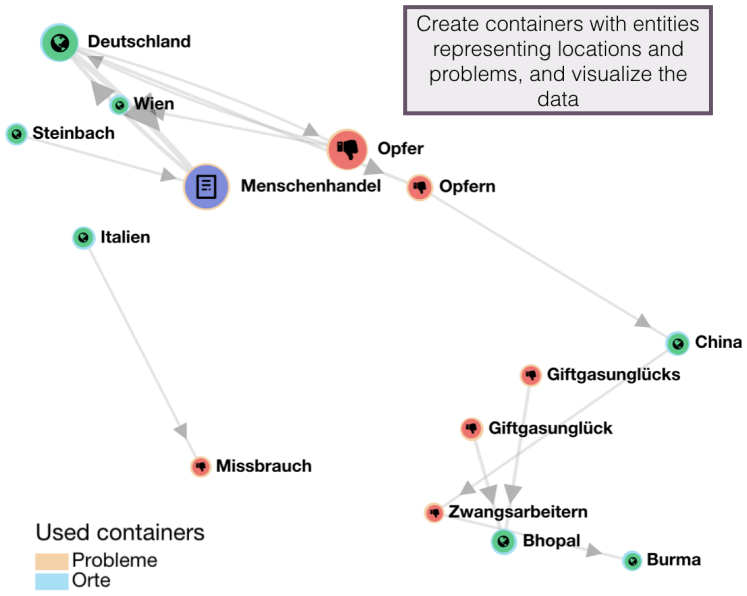


Figure 28: Use case 2 "Exploring the problems of Human Rights".

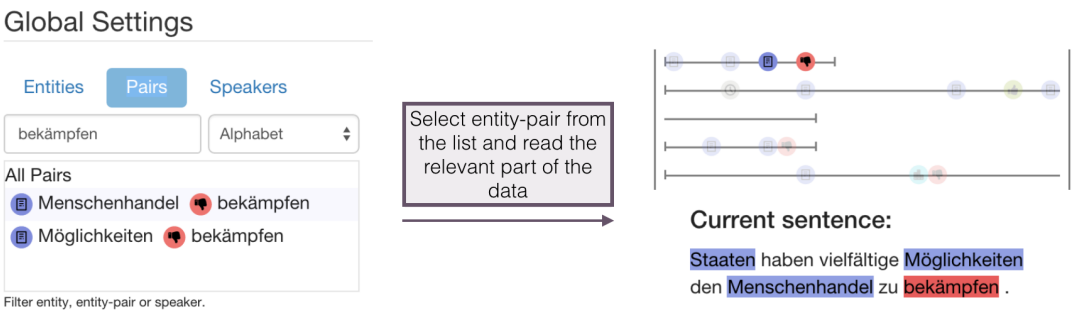


Figure 29: Use case 2 "Exploring the solutions to deal with the problems of Human Rights".

BIBLIOGRAPHY

[1] Arnold S., Burke D., Doersch T., Loeber B., Lommatzsch A. *News Visualization Based on Semantic Knowledge.* International Semantic Web Conference, (2014).

[2] Baumes J., Shepherd J., Chaudhary A. *Geospatial and Entity Driven Document Visualization for Non-proliferation Analysis.* VisWeek 2011, Workshop on Interactive Visual Text Analytics for Decision Making, (2011).

[3] Borthwick A. "Decision Tree Method for Finding and Classifying Names in Japanese Texts." In: *Processings of the 6th ACL Workshop on Very Large Corpora* (1998).

[4] Boschee E., Weischedel R., Zamanian A. "Automatic information extraction." In: *Proceedings of the International Conference on Intelligence Analysis* (2005).

[5] Dang V. B., Aizawa A. "Multi-Class named entity recognition via bootstrapping with dependency tree based patterns." In: *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining* (2008), pp. 76–87.

[6] Etzioni O., Cafarella M., Downey D., Popescu A. M., Shaked T., Soderland S., Weld D. S., Yates A. "Unsupervised Named-Entity Extraction from the Web: An Experimental Study." In: *Artificial Intelligence* 165.1 (2005), pp. 91–134.

[7] Harrower M., Brewer C. A. "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps." In: *The Cartographic Journal* 40 (2003), 27—37.

[8] Siahbani M., Vadlapudi R., Whitney M., Sarkar A. "Knowledge Base Population and Visualization Using an Ontology based on Semantic Roles." In: *Proceedings of the 2013 workshop on Automated knowledge base construction* (2013), pp. 85–90.

[9] Tkachenko M., Simanovsky A. "Named Entity Recognition: Exploring Features." In: *Proceedings of KONVENS* (2012), pp. 118–127.

[10] Vuillemot R., Clement T., Plaisant C., Kumar A. "What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections." In: *IEEE VAST* (2009), pp. 107–114.

[11] Zelenko D., Aone C., Richardella A. "Kernel Methods for Relation Extraction." In: *Journal of Machine Learning Research* 3 (2003), pp. 1083–1106.

[12] Grishman R., Sundheim B. "Message understanding conference-6: a brief history." In: *Proceedings of the 16th conference on Computational linguistics* (1996), pp. 466–471.

[13] Schneidermann B. "The eyes have it: A task by data type taxonomy for information visualization." In: *IEEE Symposium on Visual Languages* (1996), pp. 336–343.

[14] Mendes P. N., Jakob M., García-Silva A., Bizer C. "DBpedia Spotlight: Shedding Light on the Web of Documents." In: *I-Semantics '11 Proceedings of the 7th International Conference on Semantic Systems* (2011), pp. 1–8.

[15] Sekine S., Nobata C. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." In: *Proc. Conference on Language Resources and Evaluation* (2004), pp. 1977–1980.

[16] Setlur V., Stone M. C. *A Linguistic Approach to Categorical Color Assignment for Data Visualization.* The IEEE Information Visualization Conference, (2015).

[17] Finkel J.R., Manning C.D. "Nested named entity recognition." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2009), pp. 141–150.

[18] Gildea D., Jurafsky D. "Automatic labeling of semantic roles." In: *Computational Linguistics* 28.3 (2002), pp. 245–288.

[19] Grobelnik M., Mladenic D. "Visualization of News Articles." In: *Informatica Journal* 28.4 (2004), pp. 375–380.

[20] Mintz M., Bills S., Snow R., Jurafsky D. "Distant supervision for relation extraction without labeled data." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (2009), pp. 1003–1011.

[21] Kumar E. *Natural Language Processing.* New Delhi: I.K. International Publishing House, (2011).

[22] McCallum A., Freitag D., Pereira F. "Maximum en- tropy Markov models for information extraction and segmentation." In: *ICML-2000* (2000), pp. 591–598.

[23] Bouma G. "Normalized (Pointwise) Mutual Information in Collocation Extraction." In: *Proceedings of the International Conference of German Socienty for Computational Linguistics and Language Technology* (2009), pp. 31–40.

[24] Hong G. "Relation Extraction Using Support Vector Machine." In: *Natural Language Processing – IJCNLP 2005* (2005), pp. 366–377.

[25] Keim D. A., Mansmann F., Schneidewind J., Ziegler H. "Challenges in visual data analysis." In: *Proceedings of the conference on Information Visualization* (2006), pp. 9–16.

[26] Sekine S., Grishman R., Shinnou H. *Decision Tree Method for Finding and Classifying Names in Japanese Texts.* Fifth Conference on Applied Natural Language Processing, (1998), pp. 171–178.

[27] Blei D. M., Ng A. Y., Jordan M. I. "Latent Dirichlet Allocation." In: *Journal of machine Learning research* 3 (2003), pp. 993–1022.

[28] Levenshtein V. I. "Binary codes capable of correcting deletions, insertions and reversals." In: *Sov. Phys. Dokl. 10* (1966), pp. 707–710.

[29] Bertin J. *Semiology of graphics: diagrams, networks, maps.* Madison WI: University of Wisconsin Press, (1983).

[30] Gonzalez E., Turmo J. "Unsupervised Relation Extraction by Massive Clustering." In: *Ninth IEEE International Conference on Data Mining* (2009), pp. 782–787.

[31] Kintz M., Finzen J. *A simple method for mining and visualizing company relations based on web sources.* 7th International Conference on Web Information Systems and Technologies (WEBIST), (2011).

[32] Rogers Y., Sharp H., Preece J. *Interaction design: beyond human-computer interaction. 3rd edition.* United Kingdom: John Wiley and Sons Ltd, (2011).

[33] Wang T., Li Y., Bontcheva K., Cunningham H., Wang J. "Automatic Extraction of Hierarchical Relations from Text." In: *The Semantic Web: Research and Applications* (2006), pp. 215–229.

[34] Acevedo-Aviles J.C., Campbell W.M., Halbert D.C., Greenfield K. "VizLinc: Integrating information extraction, search, graph analysis, and geo-location for the visual exploration of large data sets." In: *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics* (2014), pp. 10–18.

[35] Stasko J., Görg C., Liu Z., Singhal K. "Jigsaw: Supporting Investigative Analysis through Interactive Visualization." In: *Information Visualization archive* 7.2 (2008), pp. 118–132.

[36] F. Keim D., Kohlhammer J., Mansmann F., May T., Wanner. *Mastering the Information Age Solving Problems with Visual Analytics.* Goslar: Eurographics Association, (2010).

[37] Agichtein E., Gravano L. "Extracting Relations from Large Plain-Text Collections." In: *Proceedings of the fifth ACM conference on Digital libraries* (2000), pp. 85–94.

[38] Benikova D., Fahrer U., Gabriel A., Kaufmann M., Yimam S. M., von Landesberger T., Biemann C. "Network of the Day: Aggregating and Visualizing Entity Networks from Online Sources." In: *Proceedings of the 12th KONVENS 2014* (2014).

[39] Gold V., Rohrdantz C., El-Assady M. *Exploratory Text Analysis using Lexical Episode Plots.* Eurographics Conference on Visualization (EuroVis), (2015), pp. 85–89.

[40] Grimmer J., Stewart B. M. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." In: *Political Analysis* (1974), pp. 667–673.

[41] Yan Y., Okazaki N., Matsuo Y., Yang Z., Ishizuka M. "Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web." In: *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* (2009), pp. 1021–1029.

[42] Kambhatla N. "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations." In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (2004), pp. 178–181.

[43] Ritter A., Clark S., Etzioni M., Etzioni O. "Named Entity Recognition in Tweets: An Experimental Study." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), pp. 1524–1534.

[44] Rheingans P. "Task-based Color Scale Design." In: *Proceedings Applied Image and Pattern Recognition* (1999).

[45] Bikel D. M., Miller S., Schwartz R., Weischedel R. *Nymble - a High-Performance Learning Name-finder.* Fifth Conference on Applied Natural Language Processing, (1998), pp. 194–201.

[46] Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel R. "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation." In: *Proceedings of LREC 2004* (2004), pp. 837–840.

[47] Maciejewski R. *Data Repesentations, Transformations, and Statistics for Visual Reasoning*. Morgan and Claypool Publishers, (2011).

[48] Miller D., Boisen S., Schwartz R., Stone R., Weischedel R. "Named entity extraction from noisy input: speech and OCR." In: *Proceedings of the sixth conference on Applied natural language processing* (2000), pp. 316–324.

[49] Zhu J., Nie Z., Nie X., Zhang B., Wen J. R. "StatSnowball: a Statistical Approach to Extracting Entity Relationships." In: *World Wide Web Conference Series - WWW* (2009), pp. 101–110.

[50] Bach N., Badaskar S. *A Review of Relation Extraction.* Language Technologies Institute, Carnegie Mellon University, (2007).

[51] Coates Stephens S. "The Analysis and Acquisition of Proper Names for the Understanding of Free Text." In: *Computers and the Humanities* 26 (1992), pp. 441–456.

[52] Faruqui M., Padó S. "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization." In: *Proceedings of KONVENS 2010* (2010), p. 129.

[53] Nadeau D., Sekine S. "A survey of named entity recognition and classification. Named Entities: Recognition, classification and use." In: *Special issue of Lingvisticæ Investigationes* 30.1 (2007), pp. 3–26.

[54] Nadeau D., Turney P. D., Matwin S. "Unsupervised Named-Entity Recognition: Generating Gazetters and resolving Ambiguity." In: *Proceedings of the 19th international conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence* (2006), pp. 266–277.

[55] Sarawagi S. *Information Extraction.* 10th. Boston: now Publishers, (2007).

[56] Sun A., Grishman R., Sekine S. "Semi-supervised Relation Extraction with Large-scale Word Clustering." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (2011), pp. 521–529.

[57] Garlandini S., Fabrikant S.I. "Evaluating the Effectiveness and Efficiency of Visual Variables for Geographic Information Visualization. Spatial Information Theory." In: *9th International Conference, COSIT* (2009), pp. 195–211.

[58] Hellrich J., Faessler E., Buyko E., Hahn U. "Visualizing Semantic Metadata from Biological Publications." In: *Proceedings of the International Workshop on Intelligent Exploration of Semantic Data (IESD 2012) at EKAW 2012* (2012).

[59] McCallum A., Li W. "Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons." In: *Processings Conference on Computational Natural Language Learning* (2003), pp. 188–191.

[60] Oramas S., Sordo M., Serra X. *Automatic Creation of Knowledge Graphs from Digital Musical Document Libraries.* Conference in Interdisciplinary Musicology, (2014).

[61] Asahara M., Matsumoto Y. "Japanese Named Entity Extraction with Redundant Morphological Analysis." In: *Proc. Human Language Technology Conference - North American Chapter of the Association for the Computational Linguistics* (2003), pp. 8–15.