VisLR:
Visualization as Added Value in the Development, Use and Evaluation of Language Resources

31 May 2014

ABSTRACTS

Editors:

Annette Hautli-Janisz, Verena Lyding, Christian Rohrdantz

Workshop Programme

09:00 – 10:30 – Morning Session, Part I

09:00 – 09:10 – Introduction

09:10 – 09:40
Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, *An Interactive Visualization of Crosslinguistic Colexification Patterns*

09:40 – 10:00
Roberto Theron and Eveline Wandl-Vogt, *The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually*

10:00 – 10:30
Pierrick Bruneau, Olivier Parisot, Amir Mohammadi, Cenk Demiroğlu, Mohammad Ghoniem and Thomas Tamisier, *Finding Relevant Features for Statistical Speech Synthesis Adaptation*

10:30 – 11:00 Coffee break

11:00 – 13:00 – Morning Session, Part II

11:00 – 11:30
Florian Stoffel, Dominik Jäckle and Daniel A. Keim, *Enhanced News-reading: Interactive and Visual Integration of Social Media Information*

11:30 – 11:50
Markus John, Florian Heimerl, Andreas Müller and Steffen Koch, *A Visual Focus+Context Approach for Text Comparison Tasks*

11:50 – 12:10
Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe and Daniel A. Keim, *V1 in Icelandic: A Multifactorical Visualization of Historical Data*

12:10 – 12:40
Stefan Jänicke, Marco Büchler, Gerik Scheuermann, *Improving the Layout for Text Variant Graphs*

12:40 – 13:00 – Closing Discussion

Workshop Organizers

Annette Hautli-Janisz	University of Konstanz
Verena Lyding	European Academy of Bolzano/Bozen
Christian Rohrdantz	University of Konstanz

Workshop Programme Committee

Noah Bubenhofer	Dresden University of Technology
Miriam Butt	University of Konstanz
Chris Culy	University of Tübingen
Christopher Collins	University of Ontario Institute of Technology
Annette Hautli-Janisz	University of Konstanz
Gerhard Heyer	Leipzig University
Kris Heylen	University of Leuven
Daniel Keim	University of Konstanz
Steffen Koch	University of Stuttgart
Verena Lyding	European Academy of Bolzano/Bozen
Thomas Mayer	Philipps-Universität Marburg
Daniela Oelke	DIPF Frankfurt
Christian Rohrdantz	University of Konstanz

Preface to the VisLR Workshop

The VisLR workshop aims at bringing together people from visual analytics and computational linguistics to discuss the potentials and the challenges related to visualizing language data and in particular language resources. Linguistics has a long tradition of visually representing language patterns, from tree representations in syntax to spectograms in phonetics. However, the large amounts and ever-increasing complexity of today's resources call for new ways of visually encoding a multitude of abstract information on language in order to assure and enhance the quality and usability of these language resources.

We invited submissions on research demonstrating the development, use and evaluation of visualization techniques for language resources. This includes work applying existing visualization techniques to language resources as well as research on new visualization techniques that are specifically targeted to the needs of language resources.

The workshop contributions comprise visualization approaches for lexicographic data, text resources as well as speech data. Mayer et al. and Theron & Wandl-Vogt present two visualizations for facilitating the interactive exploration of lexicographic data. Bruneau et al. show how to make use of visual tools for analyzing high-dimensional models for speech synthesis adaptation. The visualizations for text and corpus data propose visual approaches that can be applied to support enhanced news-reading (Stoffel et al.), distant reading (John et al.) and the study of language change (Butt et al.) as well as the comparison of different editions of a text (Jänicke et al.).

Morning Session, Part I

Saturday 31 May, 9:00 – 10:30
Chairperson: Annette Hautli-Janisz

An Interactive Visualization of Crosslinguistic Colexification Patterns

Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban

In this paper, we present an interactive web-based visualization for the CLICS database, an online resource for synchronic lexical associations (colexification patterns) in over 200 language varieties. The associations cover 1,288 concepts and represent the tendency for concepts to be expressed by the same words in the same languages and language varieties of the world. The complexity of the network structure in the CLICS database calls for a visualization component that makes it easier for researchers to explore the patterns of crosslinguistic colexifications. The network is represented as a force-directed graph and features a number of interactive components that allow the user to get an overview of the overall structure while at the same time providing an opportunity to look into the data in more detail. An integral part of the visualization is an interactive listing of all languages that contribute to the strength of a given pattern of colexification. Each language in the list is thereby attributed a different color depending on its genealogical or areal affiliation. In this way, given associations can be inspected for genealogical or areal bias.

The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually

Roberto Theron and Eveline Wandl-Vogt

Historical dictionaries and non-standard language corpora can greatly benefit from a visual access in order to grasp the inherent tangled and complex nature of the knowledge encapsulated in them. Although visual analytics has been used to tackle a number of language and document related problems, most dictionaries are still reproducing the book metaphor in which Web pages substitute the paper and the user experience is only enhanced by means of hyperlinks. Although fields such as dialectology and dialectal lexicography have incorporated Geographic Information Systems and advanced computational linguistics features, spatio-temporal dynamics can be discovered or understood if appropriate visual analytics techniques are used to surpass the idea of these linguistic resources as alphabetically ordered lists. In this paper we present the work carried out in this direction for the Dictionary of Bavarian Dialects in Austria. By means of multiple-linked views an access that fosters the exploratory analysis of the data is enabled.

Finding Relevant Features for Statistical Speech Synthesis Adaptation

Pierrick Bruneau, Olivier Parisot, Amir Mohammadi, Cenk Demiroğlu, Mohammad Ghoniem and Thomas Tamisier

Statistical speech synthesis (SSS) models typically lie in a very high-dimensional space. They can be used to allow speech synthesis on digital devices, using only few sentences of input by the user. However, the adaptation algorithms of such weakly trained models suffer from the high dimensionality of the feature space. Because creating new voices is easy with the SSS approach, thousands of voices can be trained and a Nearest-Neighbor (NN) algorithm can be used to obtain better speaker similarity in those limited-data cases. NN methods require good distance measures that correlate well with human perception. This paper investigates the problem of finding good low-cost metrics, i.e. simple functions of feature values that map with objective signal quality metrics.

We show this is a ill-posed problem, and study its conversion to a tractable form. Tentative solutions are found using statistical analyzes. With a performance index improved by 36% w.r.t. a naive solution, while using only 0.77% of the respective amount of features, our results are promising. Deeper insights in our results are then unveiled using visual methods, namely high-dimensional data visualization and dimensionality reduction techniques. Perspectives on new adaptation algorithms, and tighter integration of data mining and visualization principles are eventually given.

Morning Session, Part II

Saturday 31 May, 11:00 – 13:00
Chairperson: Verena Lyding

Enhanced News-reading: Interactive and Visual Integration of Social Media Information

Florian Stoffel, Dominik Jäckle and Daniel A. Keim

Today, everyone has the possibility to acquire additional information sources as supplement to articles from newspapers or online news. The limitations of classical newspaper articles and restrictions of additional materials on online newsportals often lead to the situation where the reader demands additional news sources and more detailed information. When using the Internet, exploiting new information sources is a trivial task. Besides professionally administered information sources, like for example large newsportals such as cnn.com, there is a growing amount of user generated content available. Services like Twitter, Facebook or Reddit allow free discussion of any subject, giving everyone the possibility to participate. In this paper, we demonstrate an approach that combines professionally generated news content with user-generated data. This approach effectively enriches the information landscape and broadens the context of a given subject. For the presented system, we focus on Reddit, one of the biggest web portals for user-generated contents. Taking the general nature of user generated content into account, we exploit metadata and apply Natural Language Processing (NLP) methods to allow users to filter additional information, which is also supported visually.

A Visual Focus+Context Approach for Text Comparison Tasks

Markus John, Florian Heimerl, Andreas Müller and Steffen Koch

The concept of distant reading has become an important subject of digital humanities research. It describes a mode of textual work in which scholars are aided by automatic text analysis and visualization to directly find and access information relevant to their research questions in a large volume of text. While such techniques have proven to be effective in saving time and effort compared to extracting the information by linearly reading through the text, they introduce a new abstract level of analysis that masks the original source text. In this work, we present a flexible focus+context approach that facilitates scholarly textual work while at the same time supports efficient distant reading techniques. Users have full access to digital text sources for the perusal of a single text passage or the comparison of multiple ones. For each selected passage, an interactive visual summarization of its respective context allows users to effortlessly switch back and forth from close to distant reading. We demonstrate the capabilities of our approach with a usage scenario from the comparative study of poetics. The applicability and usefulness based on expert feedback is discussed afterwards.

V1 in Icelandic: A Multifactorical Visualization of Historical Data

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe and Daniel A. Keim

We present an innovative visualization technique for the analysis of historical data. We illustrate our method with respect to a diachronic case study involving V1 word order in Icelandic. A number of interacting factors have been proposed by linguists as being determinative of matrix declarative V1. The significance of these factors in contributing to declarative V1 can be explored interactively via our multifactorial visualization within a given text, but also comparatively over time. We apply the visualization to a corpus study based on the IcePaHC historical corpus of Icelandic and show that new results emerge very clearly out of the visualization component and that the appearance of declarative V1 is not confined to the situations identified so far by linguists. We demonstrate that the multifactorial visualization opens up new avenues for the exploration of alternative explanations. The visualization can be applied to any linguistic problem studying an interaction between several factors across time.

Improving the Layout for Text Variant Graphs

Stefan Jänicke, Marco Büchler, Gerik Scheuermann

Sentence Alignment Flows are visualizations for Text Variant Graphs that show the variations between different editions of texts. Although the resultant graph layouts are a substantial improvement to standard tools that are used in the corresponding Digital Humanities research field, the visualization is often cluttered due to large amounts of edge crossings and the occlusion of edges and vertices. In this paper, we present methods for the layering of vertices, the bundling of edges and the removal of overlaps between edges and vertices to reduce clutter, and therefore, to improve the readability for such graphs. Finally, we present the results of our survey with participants from the humanities and computer science, who had the task to compare the readability of Sentence Alignment Flows to the layouts generated by our improved method.