Abstracts of Oral and Poster Presentations of the 3rd Swiss Text Analytics Conference (SwissText 2018)

Contents

CEUR-WS.org/Vol-2226/abstract1.pdf

1 Text Zoning for Job Advertisements with Bidirectional LSTMs

Ann-Sophie Gnehm
University of Zurich

We present an approach to text zoning for job advertisements with bidirectional LSTMs (BiLSTMs). Text zoning refers to segmenting job ads into eight classes that differ from each other regarding content. It aims at capturing text parts dedicated to particular subjects (e.g. the publishing company, qualifications of the person wanted, or the application procedure) and hence facilitates subsequent information extraction. As we have 38,000 job ads in German available, published in Switzerland from 1950 to 2014 (Swiss Job Market Monitor corpus), each labeled with text zones, we benefit from a large amount of training data for supervised machine learning. We use BiLSTMs, a class of recurrent neural networks particularly suited for sequence labeling. Our best model reaches a token-level accuracy of 89.8%, which is 2 percentage points above results from previous approaches with CRFs and implies an error rate reduction by 16%. Models with task-specific embeddings perform better than models with pretrained word embeddings, due to the large amount of labeled training data. When optimizing the model for future application on recently published job ads, the inclusion of older training data lowers performance, as some sort of out-of-domain effect counteracts the effect of more training data. Ensembling, i.e. to aggregate classification decisions of five models, brings the largest improvement of all optimization steps, raising accuracy by 0.5 percentage points. In conclusion, we succeeded in building a high performing solution to automatic text zoning for job ads based on neural networks.

2 Rupert the Nutritionist, the Efficient Conversational Agent

Jacky Casas[1], *Nathan Quinteiro Quintas*[2], *Elena Mugellini*[1], *Omar Abou Khaled*[1]
University of Applied Sciences of Western Switzerland; Fribourg[1], Ecole Polytechnique Fédérale de Lausanne[2]

Rupert is a conversational agent, or chatbot, whose goal is to coach people who wish to improve their food lifestyle. Two choices are available to the user: reduce his consumption of meat or consume more fruits and vegetables. The data gathering method is easy and especially fast in order to simplify the life of the user. The follow-up is done daily with a more comprehensive summary at the end of each week.

The purpose of this chatbot is to demonstrate that an interaction via a chat interface can be fast and efficient, and has nothing to envy to existing coaching mobile applications.

The daily interaction duration is up to the user, from few seconds to several minutes. Users who want to know more about different topics related to food can start a discussion with the agent. A guided conversation is then launched and users will learn interesting facts like meat consumption in Switzerland, meat ”budget” and other tips

and statistics about food in general. Since the target audience is Switzerland, the information relates to this country. The chatbot currently speaks French, but will also speak German and Italian in the future. A coach must be friendly and speak the language of the user. Rupert is developed to run on Facebook Messenger.

User tests will be organized during the beginning of 2018 to compare this way of doing with standard coaching apps.

3 The surprising Utility of NLP Algorithms on non Text Data

Ruben Wolff
D-ONE

Natural Language Processing methods are designed to work on datasets with vocabulary sizes in the hundred thousands and trained on milion row corpora. The mathematical representation of a sentence is often just a one-hot boolean vector, a representation which is natural for a wider array of sparse category of datasets. In this talk we will detail one real world customer story which appeared to require training of 60 000 different binary classification decision planes in a input space of 200 000 dimensions. By using sent2vec on this categorical data we were able to learn a joint representation and then take a nonparametric nearest neighbour approach to perform the binary classifications in one unified model. We hope our story can inspire others to identify the potential of NLP methods in their high dimensional sparse datasets.

4 A Morphological Toolset for Rumantsch Grischun

Simon Clematide
University of Zurich

Rumantsch Grischun (RG) is a written language created in 1982 in order to provide a standard that unifies the Romansh idioms spoken in the canton of Graubünden. The idioms as well as RG are officially recognized as the 4th national language of Switzerland, however, only RG is used by the cantonal and Swiss authorities. Although the electronic dictionary "pledari grond" is openly available, there is a lack of readily available Natural Language Processing (NLP) tools or components for RG. We developed efficient morphological tools based on finite-state methods for analyzing inflected words as well as for generating specific word forms from base forms plus morphological feature specifications. Underspecified feature sets allow flexible generation of inflection paradigms. Additionally, we annotated more than 8,000 tokens on the level of lemmas, part of speech tags, and morphological features. Using supervised sequence classification methods as Conditional Random Fields, we achieve an mean (10-fold cross validation) accuracy of 87.4% in predicting detailed morphological analyses (case, number, tense, etc.) in running text. For the simpler task of predicting the correct out of 21 parts of speech, we achieve a mean accuracy of 88.5% . The ongoing annotation of more sentences will improve this performance in the near

future. The carefully built lexicon (mostly derived from pledari grond) contains 5,200 adjectives, 3,500 verbs, 26,700 (proper) nouns. Our tools and resources are licensed under CC 4.0 BY.

5 SlowSoft Speech Technology Components for Swiss German and Romansh

Christof Traber
SlowSoft GmbH

Speech technology components for Swiss German and Romansh (speech recognition/ASR, text-to-speech/TTS, and spoken dialog systems) are not currently available from big international companies. SlowSoft GmbH, a Zurich-based company, develops speech components for these "neglected" Swiss languages. SlowSoft has a focus on TTS in Swiss German and Romansh and on overall speech applications. An initial Romansh ASR was developed in-house, and ASR in Swiss German is done in collaboration with partner companies.

The SlowSoft TTS system has been developed from scratch since 2015. The system is suitable for fully embedded as well as for server-based solutions. The system currently supports one Romansh idiom (Vallader, Lower Engadine idiom) and one Swiss German dialect (Grisons dialect). Further idioms and dialects will be developed over time.

The initial TTS system follows a classical, rule- and statistics-based approach. Deep learning components (for prosody and signal generation) are planned to be added to the system in the course of 2018.

The contribution will give an overview of the company and the TTS system, and it will describe the major difficulties and the adopted solutions in the treatment of Swiss German dialects. The contribution will provide demonstrations of the current status of the Swiss German and Romansh TTS in the framework of small speech dialog applications, which were also developed by SlowSoft.

6 Building a Question-Answering Chatbot using Forum Data in the Semantic Space

Khalil Mrini, *Marc Laperrouza*, *Pierre Dillenbourg*
Ecole Polytechnique Fédérale de Lausanne

We attempt to combine both conversational agents and conceptual semantic representations by creating a Chatbot based on an online autism forum, and aimed at answering questions of parents of autistic children.

First, we collected the titles and posts of all threads in two forums. We filter the threads based on their titles, such that only threads which titles are questions are kept. We remove threads without replies and obtain 36,000 threads, totalling 600,000 replies.

Then, to provide the best answers, we use Amazon Mechanical Turk to obtain usefulness labels on five levels for a part of the data set: about 10'000 replies. We train a variety of models to learn from these labels and apply them on the unlabelled replies. We use seven standardized continuous features, with three features on sent2vec cosine similarity. The Random Forest Classifier came on top with an F1-score of 0.66.

Afterwards, we compute the sentence vectors of questions and replies by averaging word2vec embeddings. When the Chatbot is asked a question, its sentence representation is computed and compared to all forum questions. The replies of the most cosine-similar question are first filtered to keep the ones with the highest usefulness label, and then the most cosine-similar reply is returned as answer.

An example of how the Chatbot works is its answer to "What is Autism?": "Autism has always been difficult for some people to to explain, but I do know what it is not: Pretty colors and sparkly gems".

7 Auto-Generated Email Responses: Boost Your Email Efficiency

Lorenz Bernauer, *Patrick Bernauer*
Pinsoris GmbH

Emails are a time-consuming and poorly automated task within today's business world. Repetitive emails can be under-challenging and frustrating for highly skilled workforce. This pain is addressed by Mailytica, a new service for Email Smart Responses based on Natural Language Processing and Artificial Intelligence.

Mailytica analyzes emails by applying Text Analysis and Machine Learning pipelines in order to derive topics without human input. All new incoming emails are constantly examined and referred to existing topics. Subsequently, Mailytica does generate emails responses automatically on basis of answers which were given in the past to the same topic. This especially automates emails of repetitive business transactions.

Mailytica significantly differs from comparable systems: 1. The training is done by itself. No manual training needed. This is achieved by classifying a dataset of old emails and by analysing how the user responded to these emails. 2. Deployment of most up-to-date NLP and AI algorithms. Mailytica does understand not only single words, but also the real meaning and context of all sentences and messages. 3. Self-learning algorithms. Ensuring that the solution is advancing constantly. 4. Incorporation of different functionalities like classification, routing, analytics of emails and auto-generated Smart Responses.

Based on a real world case study, we demonstrate how Mailytica works, how it integrates and how users benefit from it.

8 Optimizing Information Acquisition and Decision Making Processes with Natural Language Processing, Machine Learning and Visual Analytics

Albert Weichselbraun, *Philipp Kuntschik*, *Norman Süsstrunk*, *Fabian Odoni*, *Sandro Hörler*, *Adrian M.P. Brasoveanu*
Swiss Institute for Information Research, University of Applied Sciences Chur

The Web, social media and textual resources within organizations contain a growing wealth of data that is potentially useful to improve an organization's products, services and business processes. Key technologies such as natural language processing, machine learning and visual analytics provide means for acquiring data from external and internal sources, extracting information, and performing analytics that help in obtaining insights in regard to specific industries, companies, products and services.

Applying these methods to real-world use cases, therefore, has the potential to play a pivotal role in improving an organization's competitiveness. Nevertheless, combining data sources, technology and domain knowledge to use cases that unfold the economic potential of these methods is a challenging task.

The following presentation, therefore, introduces three industry research projects from the domains of (i) recruiting, (ii) company valuation, and (iii) digital asset management to demonstrate how natural language processing, machine learning and visual analytics can be used to improve the efficiency of information gathering processes, enable data-driven decision making and improve customer value. We will discuss data sources (Web, Deep Web and internal data) relevant to each use cases, methods for processing these data, their adaptation to the domain and integration in the company's business processes.

9 Text-based Indexation and Monitoring of Corporate Reputation using the R Package "sentometrics"

Samuel Borms[1], *David Ardia*[1], *Keven Bluteau*[1], *Kris Boudt*[2]
Université de Neuchâtel[1], Vrije Universiteit Brussel[2]

This presentation provides a methodological and practical introduction to optimized textual sentiment indexation applied to reputation monitoring. The reputation of a firm, a political party or an individual is largely determined through the tone of news spread out across different media channels. Increasing data availability gives the opportunity to use sentiment analysis to capture the information value regarding reputation within textual data of all sorts. First, we set forward a generic methodology that accounts for the metadata and other features of texts as well as the various degrees of freedom in computing and aggregating sentiment, leading to multiple textual sentiment time series. Second, these time series are related to several reputation proxies, given appropriate econometric techniques. The end result is a reputation index that can be monitored at any frequency, re-assessed easily and decomposed into its underlying drivers. Above workflow is implemented in the freely available R package 'sentometrics'. We point

to the problem and usefulness of reputation monitoring using texts, and demonstrate the methodology using 'sentometrics'. For the demonstration, we analyse a corpus of texts covering several well-known Swiss companies and work step-by-step towards the creation of their respective reputation indices before briefly studying their relative evolution.

10 Enhancing Search with Facets, Categories and Clusters extracted from Free Text

Alexandros Paramythis, *Doris Paramythis*
Contexity AG

When searching unstructured text, automatically overlaying various types of structure over the search corpus (e.g., in support of interactive disambiguation and refinement) is only possible through sophisticated forms of analysis that go beyond traditional term-matching approaches.

We use a variety of techniques to augment free-text search with interactive facilities that help users to quickly and easily reach the results they are really interested in. One such technique is the extraction of attributes and categories with the assistance of generic or domain-specific dictionaries, thesauri and ontologies. This makes it possible to offer faceting/filtering and hierarchical categorization along dimensions of interest to the users.

A different technique is based on automatically clustering "on top" of text features extracted through NLP analysis. These features may be concrete entities (e.g., persons, locations), or entirely "unconstrained" like clusters based on noun phrases with high discriminatory value. These clusters are used both for interactive disambiguation, and for exploratory navigation through the search space.

We will demonstrate how the above can be used in various application domains, starting from enterprise search, where the techniques are used to "connect the dots" in scattered corporate content, all the way to search solutions for business clusters, where the goal is to identify the members' unique characteristics, and the interconnections between them.

11 Parsing Approaches for Swiss German

Noëmi Aepli, *Simon Clematide*
University of Zurich

This paper presents different approaches towards universal dependency parsing for Swiss German. Dealing with dialects is a challenging task in Natural Language Processing because of the huge linguistic variability, which is partly due to the lack of standard spelling rules. Building a statistical parser requires expensive resources which are only available for a few dozen high-resourced languages. In order to overcome the low-resource problem for dialects, approaches to cross-lingual learning are exploited. We apply different cross-lingual parsing strategies to Swiss German, making use of the

Standard German resources. The methods applied are annotation projection and the model transfer. The results show around 60% Labelled Attachment Score for all approaches and provide a first substantial step towards Swiss German dependency parsing. The resources are available for further research on NLP applications for Swiss German dialects.

12 Automated Transformation of Documents into Data Sources

Aaron Richiger, *Martin Keller*, *Patrick Emmisberger*
turicode AG

One of the most frequently encountered challenges for any text analysis task is the procurement of usable input data. The tedious nature of this task stems from the fact that human-generated text is usually not available in a clean, machine-readable format. Quite the opposite is the case: Documents created with a human reader in mind are made to please our eyes and adhere to design-principles that primarily serve the physiological aspects of reading and content consumption in general. Extracting the textual information from such documents is a highly difficult undertaking, which is not only aggravated by seemingly illogical formatting rules, but also the use of illustrations, tables and figures to convey content.

In order to efficiently extract data captured in documents meant for human reading, turicode AG developed an intelligent automated method that can cope with large corpora and complex document layouts and information representations. Our solution – named MINT.extract – combines a purpose-built PDF parser and query language with the power of machine learning, facilitating the transformation of documents into data sources.

We would like to demonstrate the novel ways in which we can interact with documents using MINT.extract, as well as illustrate the possibilities that arise for the text analysis community.

13 History of Patents in Switzerland

Arnaud Miribel, *Yu Yamashita*, *Soufiane El Badraoui*
Ecole Polytechnique Fédérale de Lausanne

Patents capture a lot of information. They enable to identify the state-of-the-art in a specific technological field, and they also reflect the countries competitiveness. But with over 1,000,000 new patent applications published every year, it is impossible to do a manual analysis on this particular data, that's why we leveraged machine learning tools to do it. On the 56,611 english-speaking patents that were filed in Switzerland from 1956 to 2012, we tried to find how addressed topics changed over time. That is, we plotted the semantic path of patents abstracts and our results clearly underline that innovation in Switzerland shifted from Chemistry to Information Sciences.

14 Encoder-Decoder Methods for Text Normalization

Massimo Lusetti, *Tatyana Ruzsics*, *Anne Göhring*, *Tanja Samardzic*, *Elisabeth Stark*
University of Zurich

Text normalization is the task of mapping non-canonical language, typical of speech transcription and computer mediated communication (CMC), to a standardized writing. For example, Swiss German words viel, viil, vill and viu all map to the normalized form viel. It is an important up-stream task, necessary to enable the subsequent direct employment of standard natural language processing (NLP) tools for text mining applications and indispensable for languages such as Swiss German that have no written standard. Text normalization has been addressed with a variety of methods, most successfully with character-level statistical machine translation (CSMT). In the meantime, machine translation has changed and the new methods, known as neural encoder-decoder (ED) models, resulted in remarkable improvements. Text normalization, however, has not yet followed. A number of neural methods have been tried, but CSMT remains the state-of-the-art. In this work, we normalize Swiss German WhatsApp messages using the ED framework. We exploit the flexibility of this framework, which allows us to learn from the same training data in different ways. In particular, we modify the decoding stage of a plain ED model to include target-side language models operating at different levels of granularity: characters and words. Our systematic comparison shows that our approach results in an improvement over the CSMT state-of-the-art.

15 Spitch - Technology Stack & Applications

Igor Nozhov, *Juerg Schleier*
Spitch AG

This talk covers the technology stack developed at Spitch (www.spitch.ch) as well as the real-life applications for Swiss German. The first part of the presentation introduces the state-of-the-art implementation of speech-to-text, voice biometrics including speaker verification and identification, semantic modeling for IVR system, and sentiment analysis of a live conversation.

The second part of the presentation is dedicated to the voice search in railway timetable scheduling, automatic subtitling for TV weather forecasts, and sentiment analysis of the live dialogues recorded in the call centers. The applications for the corresponding use cases will be shown during the demonstration.

16 Differences between Swiss High German and German German via data-driven methods

Gerold Schneider
University of Zurich

This study uses data-driven methods to detect and interpret differences between the

High German used as standard language of written communication in Switzerland, and German German. The comparison is based on a comparable web corpus of two million sentences, one million from Switzerland and one million from Germany. We describe differences at the levels of lexis, morphosyntax, and syntax, and compare to previously described differences. We show that data-driven methods manage to detect a wide range of differences.

17 Cutter - a Universal Multilingual Tokenizer

Johannes Graën, *Mara Bertamini*, *Martin Volk*
University of Zurich

We present Cutter, a pattern-based tokenizer available for many languages, including German, French, Italian and Romansh. The patterns that identify tokens are derived from empiric examples and developed in a test-driven manner. They form sets of language-specific and language-independent rules. Both this property, the modular architecture of our rule system and our test-driven development approach render it possible to integrate Cutter into any NLP pipeline and easily adapt the tokenizer to other languages (e.g. dialects) and domains, genres or historical text variants, which often do not have reasonable tokenization tools. In fact, tokenization is commonly regarded as a solved problem. Yet, we are often forced to adapt either a tokenizer or its output to our particular needs in corpus processing – especially in cases where the token boundaries do not simply consist of a whitespace but (additionally) contain hyphens, apostrophes, slashes, abbreviation points and the like.

Our tool works as follows: Given at least one (language-/domain-specific) list of abbreviations, Cutter protects every occurrence of those from further tokenization. The actual tokenization consists of applying the rules in the order provided and thereby identifying tokens by patterns: Once a pattern matches a token, the whole input is split into token and non-token parts. Subsequently, the same patterns are applied to the non-token parts; when no further pattern is applicable, the leaves of the derived token tree correspond to the final token sequence. We organize rules in different sets, from the most specific to the most general ones, thereby interweaving common with language-specific rules.

A collection of unit tests for each tokenization rule assures that new rules in the form of token-defining patterns do not conflict with existing ones when added, and thus guarantees consistency.

18 On the Effectiveness of Feature Set Augmentation using Clusters of Word Embeddings

Georgios Balikas[1], *Ioannis Partalas*[2]
Kelkoo[1], Expedia[2]

Word clusters have been empirically shown to offer important performance improvements on various Natural Language Processing (NLP) tasks. Despite their importance,

their incorporation in the standard pipeline of feature engineering relies more on a trial-and-error procedure where one evaluates several hyper-parameters, like the number of clusters to be used. In order to better understand the role of such features in NLP tasks we perform a systematic empirical evaluation on three tasks, that of named entity recognition, fine grained sentiment classification and fine grained sentiment quantification.

19 "What can I ask you?" - The User Experience of Conversational Interfaces

Sibylle Peuker
Zeix AG

Conversational Interfaces are considered the new heroes of human-machine interaction. Smart assistants like Alexa from Amazon and Google Home are catching up in more and more households, and there's hardly any industry that does not experiment with chatbots.

But can humans and machines really understand each other? We wanted to know if such conversational interfaces are as easy to use as is claimed. For this purpose, we have investigated many chatbots and other conversational interfaces within a year, we observed many people (individuals and groups) interacting with such interfaces, and we also built chatbots.

To substantiate the hypotheses built during this year, we tested the user experience and acceptance of language assistants and chatbots in an exploratory study with 16 users. In in-depth interviews, we saw that think big is okay, but start small is the way to success.

In the talk, we discuss what we have learned from user research: • What are the hurdles in the communication between man and machine? • For which tasks is a conversational interface suitable? • When is it useful? • When should the chatbot rather hand over to a human employee? • How do I design a conversation so that it feels natural and meaningful to the user?

We back up our user research with 17 years of experience with user behavior in usability tests. We will discuss several examples to show the potential and the pros and cons of conversational interfaces.

20 Swiss German Language Detection in Online Resources

Pius von Däniken, *Mark Cieliebak*
SpinningBytes AG

A problem when trying to gather Swiss German phrases from online forums and social media is that there are no common language detection tools for Swiss German. This leads to a situation where we have to rely on other meta-data to identify possible Swiss

German phrases, which can lead to a large number of false positives. We explore different one-class and multiclass classification approaches to identify Swiss German phrases and evaluate them on the SB-CH corpus, a large corpus of phrases gathered from Swiss German speaking communities on social media.

21 Mining Goodreads: A Text Similarity Based Method to Measure Reader Absorption

Simone Rebora[1], *Moniek Kuijpers*[2], *Piroska Lendvai*[3]
University of Verona[1], University of Basel[2], University of Göttingen[3]

Our study addresses the automatized evaluation of reader absorption in texts from an online social reading platform. Our approach is to support empirical aesthetic research with computational linguistics tools. The method for reaching this objective is to compare reader reviews published on the Goodreads website with statements on the Story World Absorption Scale (SWAS, Kuijpers et al., 2014). This approach enables to investigate multiple dimensions : (1) the potential of converting Goodreads texts into an extensive corpus for computational analysis of reader responses; (2) validation of the SWAS scheme; and (3) conducting comparative analyses of readers' absorption across different genres. Our manual analysis on a corpus of 180 reader reviews confirmed that sentences in such user-generated content substantially overlap with SWAS statements (241 semantically-correlated pairs out of a total of 54,990, identified and cross-checked by two annotators). In order to automate the process, we tested two technologies: the detection of textual entailment (EOP tool) and of text reuse (TRACER tool) on a development corpus of 284,396 reviews of three genres: Fantasy, Romance, and Thriller. By keeping their default configurations, machine learning experimental results indicate that both tools need adaptation to the social reading genre (e.g. by defining new synonym lists and by training on a corpus consistent with the task), but they are comparable with performance on other types of user-generated content. Post-processing TRACER output shows that precision can be increased through syntactic simplification of the SWAS statements.

22 Is Reading Mirrored in the Face? A Comparison of Linguistic Parameters and Emotional Facial Expressions

Egon Werlen, *Ivan Moser*, *Christof Imhof*, *Per Bergamin*
Institute for Research in Open-, Distance- and eLearning (IFeL), Swiss Distance University of Applied Sciences (FFHS)

Digital reading and the digitisation of knowledge management and learning are increasing. The humans' nature evaluating every kind of event in relation to positive and negative also concerns reading. This evaluation, often referred to as primary (unconscious) and secondary (conscious) appraisal, is a component of emotions. It's just

the digitisation that facilitates measuring emotional characteristics of texts (e.g. lexical emotional valence), and emotional reactions and face expressions (e.g. facial emotional valence). We hypothesize that lexical emotional valence predicts readers' facial emotional valence. A three-part text was lexically analysed with the revised Berlin Affective Word List (BAWL-R). While reading, 91 subjects were filmed, and the videos were analysed with a facial emotion recognition software (FaceReader®) calculating the facial emotional valence. The result of a generalized linear mixed model predicting facial emotional valence by lexical emotional valence is highly significant but explains nearly no variance (0.3%; fixed effects). The subjective emotional valence of participants after reading confirms emotional reactions to the text. Generally, measuring emotional face expressions works well enough. So, the mostly neutral face expressions of our participants may be a result of the non-social reading situation. Therefore, we need other measurement or analysis methods.

23 Automatic Hyperlinking on a Juridicial Corpus via an RDFized Terminology

Markus S.T. Pilzecker
Die Wissens-Ingenieure

In collaboration with the terminology section of the Bundeskanzlei, the Bundesarchiv Bern, in its role as national coordination body for Open Data, commissioned a project, whose intention it was to demonstrate the usefulness of Linked Data technologies on open terminologic data.

The two principal parts of this project were: - "TERMDAT-LD", the RDFification of the juridical subdomain of TERMDAT, an official Swiss multilingual terminology - Automatic Hyperlinking on the "Systematische Sammlung des Bundesrechts (SR)" and TERMDAT itself

Automatic Hyperlinking is a technology, where a machine enriches an otherwise unaltered web document with hyperlinks to other web addresses. In contrast to a human, a machine places such hyperlinks, following a strategy, incarnated in its implementation.

Juridicial documents are rich and comparably precise in refering other juridicial documents. But even in the carefully, manually edited SR, only few of such references are made explicit as hyperlinks, placed by a human. And this is due to a simple fact: cost.

We decided to combine Automatic Hyperlinking with Linked Data technologies and, instead of chosing as linking strategy a direct link to the authoritative web page, we went over the web surface page of the TERMDAT-LD entry of the title of the legal document as a pivot point.

Since the pivot page exposes the whole RDF graph structure of the terminology, for the price of "some clicks more", it allows to study many more juridicial workflow scenaria than just direct links.

30000 terms and phrases (DE, EN, FR, IT, RM) triggered construction of a bit less than a million new links in about 120k documents. Precision was better than 99%. Due

to limitations in budget, we could not determine recall. A first inspection showed, that it is considerably below 50%, since our primitive algorithm is not very robust against syntactical noise, e.g. HTML markup in longer phrases.

24 BioMedical Text Mining Activities of the BioMeXT group

Fabio Rinaldi
University of Zurich

The OntoGene/BioMeXT group at the Institute of Computational Linguistics, University of Zurich, has been active in biomedical text mining for more than a decade. They have developed several applications for mining domain entities and relationships from the scientific literature, and more recently other biomedical textual sources.

In the presentation I will describe in some details the current projects, dealing with applications from veterinary medicine to electronic health records from several Swiss hospitals. Additionally, I will present our award-winning tools, such as OGER and the Bio Term Hub.

For more details, see http://www.ontogene.org/

25 Automated Detection of Adverse Drug Reactions in the Biomedical Literature Using Convolutional Neural Networks and Biomedical Word Embeddings

Diego Saldana
Novartis Pharma A.G.

Monitoring the biomedical literature for cases of Adverse Drug Reactions (ADRs) is a critically important and time consuming task in pharmacovigilance. The development of computer assisted approaches to aid this process in different forms has been the subject of many recent works.

One particular area that has shown promise is the use of Deep Neural Networks, in particular, Convolutional Neural Networks (CNNs), for the detection of ADR relevant sentences. Using token-level convolutions and general purpose word embeddings, this architecture has shown good performance relative to more traditional models as well as Long Short Term Memory (LSTM) models.

In this work, we evaluate and compare two different CNN architectures using the ADE corpus. In addition, we show that by de-duplicating the ADR relevant sentences, we can greatly reduce overoptimism in the classification results. Finally, we evaluate the use of word embeddings specifically developed for biomedical text and show that they lead to a better performance in this task.

26 NOAH 3.0: Recent Improvements in a Part-of-Speech Tagged Corpus for Swiss German Dialects

Noëmi Aepli[1], *Nora Hollenstein*[2], *Simon Clematide*[1]
University of Zurich[1], IBM[2]

NOAH is a part-of-speech (POS) tagged text corpus of different Swiss German dialects with 115,000 tokens overall. It includes articles from the Alemannic Wikipedia, a special edition of the newspaper "Blick am Abend", an annual report of Swatch, extracts of novels by Viktor Schobinger as well as some blogs. Our corpus uses STTS – the standard POS tagset for German as known from the TIGER treebank – and extends it for phenomena frequently found in written Swiss German, e.g. fused words. For version 3.0, we applied machine learning to efficiently spot annotation inconsistencies. We manually verified and corrected the differences between a POS tagger's output and the hitherto gold standard. The verification also led to some guideline changes for problematic cases. Anticipating future dependency annotations, we systematically chose the syntactically more consistent alternative if in doubt. For modal verbs' past participles, we deviate from the TIGER scheme and categorize them as such (VMPP) even though they morphologically look like infinitives (ich/PPER ha/VAFIN das/PDS müesse/VMPP läse/VVINF; I had to read that). A 10-fold cross-validation of NOAH 3.0 gave the following evaluation results (sentences were split randomly): TNT tagger (92.4% mean accuracy), Conditional Random Fields tagger (92.4%), BTagger (93.5%). The best-performing BTagger model improved by 2.86 percentage points compared to the last published results on NOAH.

27 Evaluating Neural Sequence Models for Splitting (Swiss) German Compounds

Don Tuggener
Zurich University of Applied Sciences (ZHAW)

This paper evaluates unsupervised and supervised neural sequence models for the task of splitting (Swiss) German compound words. The models are compared to a state-of-the-art approach based on character ngrams and a simple heuristic that accesses a dictionary. We find that the neural models do not outperform the baselines on the German data, but excel when applied to out-of-domain data, i.e. splitting Swiss German compounds. We release our code and data, namely the first annotated data set of Swiss German compounds.

28 Quantifying Collaboration in Synchronous Document Editing

Adrian Pace[1], *Louis Baligand*[1], *Stian Håklev*[2], *Jennifer Olsen*[1], *Nore de Grez*[3], *De Wever Bram*[3]

Ecole Polytechnique Fédérale de Lausanne[1], University of Toronto[2], Ghent University[3]

Collaborative synchronous writing tools like Google Docs and Etherpad let multiple users edit the same document, and see each others edits in near real-time to simplify collaboration and avoid merge-conflicts. These tools are used extensively across many domains, including education and collaboration in both research and industry. The very nature of needing to constantly synchronize state between multiple users means that very granular editing data is automatically captured and stored. In theory, this data could provide important insights into the editing process, the contributions of the different users, how the text developed over time, and other questions relevant to researchers studying writing from different theoretical and methodological angles. However, this extreme granularity of the data (down to individual key presses), makes analysis very complex. Most of the research focused on automatic analysis of collaborative writing to date has focused on asynchronous writing, and looked at the "diffs" between one editing session and the next. In this paper, we present a method and a tool to construct informative operations from text data, as well as preliminary metrics for measuring the collaborative writing process. Additionally, our method adds to previous work in that it can be used to access the writing throughout the writing process rather than just being applied to an end product.

29 A Genetic Algorithm for Combining Visual and Textual Embeddings Evaluated on Attribute Recognition

Ruiqi Li[1], *Guillem Collell*[2], *Marie-Francine Moens*[2]
University of Applied Sciences and Arts Western Switzerland (HES-SO Valais-Wallis)[1], Katholieke Universiteit Leuven[2]

We propose a genetic-based algorithm for combining visual and textual embeddings in a compact representation that captures fine-grain semantic knowledge—or attributes—of concepts. The genetic algorithm is able to select the most relevant representation components from the individual visual and textual embeddings when learning the representations, simulating how complementary visual and linguistic knowledge is combined. We evaluate the proposed model in an attribute recognition task and compare the results with a model that concatenates the two embeddings and models that only use monomodal embeddings.

30 A Supervised Dataset for Stylometric Analysis of Swiss Text

Alireza Ghasemi
ELCA Informatique SA

Stylometry is the problem of recognising, verifying, and profiling an individual from

their writing habits, i.e. their style of writing. An old and well-known problem in linguistics, it has recently seen a revive of interest with the emergence of novel machine learning techniques, hoping to see ground-breaking performance improvements. The applications of stylometry range from detecting plagiarism in academic and scholar work, to market analysis and customer studies, and further to investigative journalism and massive text mining.

We provide for the first time, to the best of our knowledge, a supervised dataset, comprising articles from the Swiss news website SwissInfo, tagged with both an Author ID as well as multiple topic labels, allowing topic-dependent, and topic-independent stylometry. The articles are written in Swiss German, French, and Italian, which we believe will help to determine distinctive factors of the stylometry of Swiss variants of these languages for further studies and applications.

Moreover, in certain cases, translated articles are distinguished from original ones, which makes the dataset useful for developing a system for determining translated text, an interesting problem by itself. We hope that providing this dataset to the public will help advance the stylometry research for Swiss languages and enable better, tailor-made text-analysis solutions for the Swiss market.

31 Bilingual Term Extraction with Big Data.

Rémy Blättler[1], *Mark Cieliebak*[2], *Jan Deriu*[3]
Supertext[1], SpinningBytes AG[2], ZHAW[3]

We are building a system that automatically extracts a translation memory (TM) and a terminology list from a multilingual website. This works by first scanning the whole website with a JavaScript capable spider and then links the related websites to each other. From those linked websites, a TM is created by running an alignment algorithm that uses Machine Translation (MT) to find related sentences (Example: Bleualign https://github.com/rsennrich/Bleualign). From the TM, special words and phrases are extracted into a terminology list. With GENSIM we find connected words. The EU corpus is used to build a base level on how frequent specific words are used in common texts.

32 Swiss Twitter Corpus

Christoforos Nalmpantis[1], *Fernando Benites*[2], *Michaela Hnizda*[1], *Daniel Kriech*[2], *Pius von Däniken*[1], *Ralf Grubenmann*[1], *Mark Cieliebak*[1]
SpinningBytes AG[1], Zurich University of Applied Sciences[2]

The Swiss Twitter Corpus is a collection of tweets related to Switzerland. Since early 2018, approximately 3 million tweets with a certain "swissness" have been added. Many applications such as word of the year, sentiment per canton, media and reputation monitoring, etc. can profit from dynamically growing corpus. The key challenge in building the Swiss Twitter Corpus is to determine the "Swissness" of a tweet. Since there is no generally accepted definition of a "national" tweet, we decided to use a

two-tier approach: In the first tier, we download tweets that are potentially related to Switzerland, e.g. by typically Swiss keywords ("Matterhorn", "Coop", "Roger Federer"), geolocation, or authors. This download runs for the three main languages of Switzerland (German, French, Italian) plus English and Swiss German. Due to the broad selection criteria, this phase grabs many tweets that are not really Swiss-related (for instance since "Coop" refers not only to the Swiss retailer). For this reason, we introduced a second tier, where we filter more precisely and tag each tweet with all matching relevance criteria. This phase includes, among other things, a language classifier for Swiss German, entity extraction for geonames, topic classification, company name disambiguation, etc. The result is a multilingual collection of tweets which can be easily selected and processed for further analysis.

33 Merging Haystacks to Find Matching Needles: A Transliteration Approach

Fernando Benites[1], *Gilber Duivesteijn*[2], *Pius von Däniken*[3], *Mark Cieliebak*[3]
Zurich University of Applied Sciences[1], Deep Impact[2], SpinningBytes AG[3]

Finding entities in multi-language and multi-script corpora is a critical task with many applications, from social network graphs to fighting terrorism online. On the other side, it is particularly challenging, since names and places are specific qualifiers as well as very unique. They are usually embedded into the language and its script. When converting a proper noun from a source script into a different target script (which has also another sound inventory), standard machine translation approaches might not achieve satisfying results. Consequently, generative transliteration models are more likely to succeed if they have their own specific training data. However, the few existing datasets for name transliteration are very specific and usually limited to one language pair.

We present a new dataset for name transliteration. The extensive JRC-Names dataset is merged with crowdsourced transliterations that can be gathered using the Wikipedia "lang" template. Based on that template, we searched for transliterations of the name title in each wiki-page. Further, we used other smaller and more specific datasets, including sanction lists. As a result, the size of the dataset is considerable: about 850 thousand entries and over 180 languages. In this talk, we will visualize the data and present our main results regarding classification of entities-matching with string-based and deep-learning methods.

34 Text and Data Mining Workflow to Make Scientific Publications Accessible

Donat Agosti[1], *Guido Sautter*[1], *Reto Gmür*[2]
Plazi[1], Factsmission[2]

Scientific publications ought to contribute to the dissemination of research results. Whilst the quintessence of current scientific publications lays in creating largely un-

structured natural language publications from often highly structured primary data, the Swiss company Plazi does the opposite, i.e., creates structured, findable, citable, and reusable data from natural language scientific publications. This workflow is based on standalone open source applications written in plain Java, with modular Natural Language Processing tools. These tools can be chained into highly customizable pipelines, allowing TDM to be highly automated, including discovery of anything ranging from named entities to entire specific text blocks, or figures and their associated captions, and deposit them both in TreatmentBank (TB) and in the Biodiversity Literature Repository (BLR). Because data is not work in a legal sense, and thus is not copyrighted, and Swiss law allows to make temporary copies of publications for mining, the Plazi workflow can operate legally under Swiss law – a competitive advantage for Switzerland. TreatmentBank includes text based entities such as 220,000 taxonomic treatments extracted from 31,000 articles, including millions of extracted facts, close to 1M bibliographic references. BLR, a community within CERN's Zenodo repository, includes 172,000 extracted scientific illustrations and 32,000 article deposits, with a daily input. Each of the deposits includes meta data, links to related items and a digital object identifier (DataCite DOI). Upon upload, this data is immediately submitted to some of the world's largest science infrastructure, such as the NCBI and the Global Biodiversity Information Facility, and names fed into the Swiss Linked Open Government Portal LINDAS.

35 Recommendations from Unstructured Data for Investment Banking

Ethan Brown, *Saurabh Jain*
Squirro

Senior bankers at investment banking firms spend much of their time manually scrolling news feeds for potential deals. Here we suggest a way to augment and automate this process leveraging natural language understanding tools. More specifically, we use a combination of entity extraction and classification to surface events that are relevant to a particular investment banker. For the entity extraction step, we have a domain-trained word embedding feed into a bi-directional recurrent neural network and end in a single-layer classifier which projects into 10 phrase classes in the investment banking domain. Next, an additional extraction is performed on the classified phrase to pull out relevant details, e.g. company name, deal size, region, etc. Following this, we have two routes to provide recommendations. In the first, each extracted phrase is scored by a simple regressor, with some of the extracted details used as input features. These scores are then aggregated by one of the extracted details (not used in the input), e.g. company name. The second approach is to simply predict details missing from the phrase extraction, and use these predictions as a recommendation. To tune both underlying models, the banker is able to provide both explicit and implicit feedback. This approach has been successfully deployed at several banks around the world.