5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS): Abstracts of the Applied Track

Contents

1 FLIE: Form Labelling for Information Extraction

Ela Pustulka-Hunt, Thomas Hanne, Phillip Gachnang and Pasquale Biafora
Information extraction from forms is a challenging topic with high practical relevance, in particular for the insurance industry in Switzerland. We have gathered over 20'000 anonymized insurance policies and related documents in German, French, English and Italian and have prototyped an automated method for information extraction. We tested this method with three policy types in German.

Given a user schema, expressed as a list of attributes to be found in an insurance policy, we extract the relevant information and map it to the attributes. To do that, we first extract the text from pdf and generate the bounding boxes as a csv. We then reconstruct a page, group the text boxes into horizontal groups and columns within groups and annotate the geometry. 24 policies coming from various insurers and representing three policy types were annotated manually by the user with the desired attribute names. Machine learning was used to propagate this annotation in two steps: first, text was tagged as being metadata or data, and in the second step, attribute names were mapped to the extracted text. The accuracy of the first step is now at 88%, and in the second step we can map the attributes which appear over 8 times in the documents with similar accuracy, while other attributes are often singletons and cannot be mapped yet. Data extraction uses those annotations to produce the required output for the user. With more annotated data, we will be able to reach the required accuracy of over 90%.

2 Ranking of Social Reading Reviews Based on Richness in Narrative Absorption

Piroska Lendvai, Uwe Reichel, Simone Rebora and Moniek Kuijpers
Book reviews on social platforms are generated in large quantities by non-specialist avid readers, and contain subjective evaluations pertaining to one's own reading experience. Social reading reviews often feature an under-researched phenomenon: Narrative Absorption, i.e. the extent to which immersion inro the book's narrative took place during reading. Absorption can be reflected by statements such as 'I was completely hooked' and pertain to a complexity of dimensions such as attention, emotional engagement, mental imagery, and transportation. Based on a set of user-generated reviews that we manually annotated (cf. Rebora et al. 2020), the detection of reading absorption with NLP approaches has been investigated in e.g. Lendvai, Rebora and Kuijpers (2019), Lendvai et al. (2020).

We work on a pipeline to retrieve and rank absorption-rich user reviews from a large, unlabeled document dump (6+ million reviews in English), in order to allow for the preselection of subsets of the dump that undergo manual annotation. We fine-tuned BERT (Devlin et al., 2018) for a supervised absorption detection task on 16k review sentences absorption-annotated by us (Absorption vs. Nonabsorption), and evaluated it on a held-out dataset of 149 reviews, achieving .75 macro F1 mean (support: 1,011 vs. 3,510 sentences).

Our current focus was to create a model that aggregates sentence level prediction scores on the document level. To this end, BERT's sentence level absorption probabilities were averaged per review and were used to train a linear regression model on the full corpus to predict Absorption Richness, defined as the proportion of sentences annotated as expressing absorption in a review. Review-level Absorption Richness regression lowers classification error relative to the baseline, defined as the review-level proportion of absorption classifications by taking the argmax of BERT's logits (Mean Average Errors of .08 vs. .11 and Spearman correlation of .73 vs. .65, respectively). The increase of the Spearman's rank correlation coefficient directly expresses that a review ranking by linear regression predictions corresponds more closely to the

ground truth ranking than a ranking solely based on BERT. We utilize the regression model in Absorption-Richness-based document filtering, to facilitate the benchmarking and analysis of social reading reviews in our large document dump.

References:

J. Devlin, M.W. Chang, K. Lee, K. Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

P. Lendvai, S. Rebora, M. Kuijpers (2019). Identification of Reading Absorption in User-Generated Book 2019 Reviews. In: Proc. of the 15th Conference on Natural Language Processing (KONVENS 2019): Kaleidoscope Abstracts. Erlangen, Germany: German Society for Computational Linguistics Language Technology, pp. 271-272.

P. Lendvai, S. Daranyi, C. Geng, M. Kuijpers, O. Lopez de Lacalle, J.C. Mensonides, S. Rebora, U. Reichel (2020). Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation. In: Proc. of The 12th Language Resources and Evaluation Conference (LREC 2020), pp. 4835-4841.

S. Rebora, P. Lendvai M. Kuijpers (2020). Annotating Reader Absorption. In: Proc. of Digital Humanities Conference (DH2020).

3 The "Multilingual Anonymisation Toolkit for Public Administrations" (MAPA) Project

Paula Reichenberg, Artūrs Vasiļevskis and Manuel Herranz

The European Union's new 'Open Data Directive' aims to stimulate the publishing and sharing of dynamic data by public administrations, thus furthering the development of language technologies, NLP research and translation. However, such data sharing is only possible subject to compliance with the General Data Protection Regulation (GDPR). For this reason, the European Commission has commissioned the development of a multilingual anonymization toolkit for public administrations.

Pangeanic and Tilde, together with CNRS (www.cnrs.fr), ELDA (www.elra.info/en), the University of Malta (www.um.edu.mt), Vicomtech (www.vicomtech.org) and SEAD (Spanish Agency for Digital Advancement) have been awarded EU funds to develop such open-source toolkit for all EU languages, able to detect and de-identify personal data (name, addresses, emails, credit card and bank accounts, etc.). The anonymisation toolkit is based on Named-Entity Recognition (NER) techniques, using neural networks approaches. Pre-trained models such as BERT (Delvin et al., 2018) and preprocessing of text using regular expressions are included.The toolkit will provide support to EU public administrations in complying with GDPR requirements, in particular in the health and legal fields.

In this short presentation, Manuel Herranz, CEO of Pangeanic, and Artūrs Vasiļevskis, Head of Machine Translation Solutions at Tilde, will discuss the challenges of the MAPA project, their strategy, the results reached so far and the perspective it opens for public administrations and the industry.

4 Showcase: Language analytics and semantic search for unknown document varieties

Holger Keibel, Elisabeth Maier and Tobias Christen

HIBU is a proprietary solution platform based on which Karakun (Basel) builds customer solutions around Enterprise Search and Text Analytics. In this talk, we present a solution by DSwiss (Zürich): high-security digital safes which allow users to store, exchange, but also search any

type of documents and other security-relevant data. The focus will be on the text analytics aspects of the solution developed with HIBU. Since the uploaded data can contain any sort of content, the solution supports users to organize their data in two ways: by a hierarchical folder structure and by means of facets (search filters). Some of the default facets are derived from structured metadata as file format or date, while others are populated dynamically by semantic taggers and classifiers as e.g. semantic document type, persons, locations mentioned in the document. Especially these filters have proven very useful to support document and data retrieval. We touch on the challenges of analyzing and indexing documents in a highly secure, multiple-encrypted environment and will then discuss joint ongoing work to support the individual needs of users even better: (1) use state-of-the-art neural network architectures to classify and extract more types of information from documents to provide a broader range of filters; (2) personalize the trained models that create the search filters; and (3) add a workflow engine with text-based triggers (e.g. proposing a specific folder when uploading a document).

5 Towards a regionally representative and socio-demographically diverse resource of Swiss German

Péter Jeszenszky, Burcu Demiray, Carina Steiner and Adrian Leemann

When it comes to representing its vast regional diversity, Swiss German is under-resourced for text-to-speech and speech-to-text tasks. The database we aim to build enriches existing resources by representing low resource regional varieties and by matching dialect variation to diverse socio-demographic backgrounds.

We plan to compile a database based on two projects. The SDATS [1] (Swiss German Dialects Across Time and Space) project, focusing on language variation and change, collects about 2000 hours of recordings from 125 survey sites (8 speakers/locality). Local dialects of the respondents, women and men from two age groups, with different professional backgrounds, are recorded. The ongoing structured interviews (to be finished by Summer 2021) involve prompting certain words and phrases, reading a text previously translated from Standard German to the local dialect by the speaker, semi-structured speech and spontaneous general interaction with the interviewer. The audio recordings come with rich background information (mobility, social networks, personality, attitude etc.), which enables the characterisation of sociolinguistic variation beside regional variation.

EAR [2] data contains non-intrusive records of spontaneous speech from healthy older individuals, mainly including everyday interactions in Swiss German. We invite EAR participants for SDATS interviews, making it possible to match linguistic variables across the spontaneous EAR records and the structured and spontaneous parts of the SDATS interview with the same person.

We plan the automated phonetic transcription of the data and aligning the results to Standard German. With the combination of the two sources, a more realistic picture about spontaneous language use will be available, which, especially annotated with the rich metadata, can become a useful resource in Swiss German text-to-speech and speech-to-text tasks.

At the conference, we plan to present the roadmap of data collection, cleaning, matching and analysis. Besides, we plan to show some sound samples along with potential future uses of the database.

Péter Jeszenszky Resume: Péter is a geographer and data scientist interested in linguistic variation and its geographic and socio-demographic causes. He finished his PhD in Geographic Information Science at the University of Zurich in 2018, mainly working with Swiss German morphosyntactic data. He was a postdoctoral researcher at the Ritsumeikan University in Kyoto, Japan, with the SNF Early PostDoc.Mobility grant, where he studied spatial and historical

variation of Japanese dialects. He is now at the University of Bern working in the SDATS project as a postdoc.

Intended Audience: Stakeholders interested in the following topics: enriching their Swiss German databases with spatially and socio-demographically diverse data; machine translation and transcription of Swiss German; validating present Swiss German databases using matched spontaneous and clearly uttered speech; generating Swiss German speech or text.

6 Deep learning and visual tools for analyzing and monitoring integrity risks

Albert Weichselbraun, Christian Hauser, Sandro Hörler and Anina Havelka

Risks jeopardizing the integrity of an organization are widespread. According to a 2018 study by PricewaterhouseCoopers, almost 40% of Swiss companies haven been affected by illegal and unethical behavior, such as embezzlement, cybercrime, corruption, fraud, money laundering and anti-competitive agreements. Although the number of cases within Switzerland is relatively low, the financial impact of these incidents is still above the global average. The University of Applied Science of the Grisons conducts research that applies web intelligence and deep learning to the task of supporting Swiss companies in identifying and mitigating integrity risks. Historical data is used for training an LSTM classifier to recognize national and international media coverage on corruption. Afterwards, we apply transfer learning techniques to the task of adapting the classifier to a wide range of integrity topics such as human rights, labor conditions and sustainability. The adapted classifier assigns scores to News articles that indicate their relevance to the topic of integrity. Sophisticated visual tools use the annotated documents for (i) tracking and visualizing past integrity management gaps and their respective impacts, (ii) identifying whether organizations have been mentioned positively or negatively in these events, and (iii) leveraging media coverage on upcoming integrity stories for predicting and discovering existing blind spots within a company's governance.

7 Exploring German BERT model pre-training from scratch

Branden Chan, Stefan Schweter and Timo Möller

In this work we provide interesting insights into BERT model pre-training from scratch for German. We experiment with different corpora and subword masking techniques.

The two current available BERT models for German (from Deepset and DBMDZ) were trained on similar amounts of data (16GB). With the availability of larger corpora such as the OSCAR corpus, that has an uncompressed size of 145GB for German and the recently introduced whole word masking technique that is used in the preprocessing step, we train BERT base and large models with different subword masking techniques and training data sizes, ranging from 16GB up to 160GB of text.

In order to show which subword masking technique improve or harm performance and if larger training corpora really improve performance significantly, we perform an extensive evaluation of our models over the course of pre-training on various German downstream tasks. Our BERT large model achieves new state-of-the-art results on GermEval 2018.

All trained models will be made publicly available to the research community.

Branden Chan is a Stanford graduate in computational linguistics. He now works for deepset.ai as a machine learning engineer bringing the latest NLP techniques to the industry. He is part of the team that open sourced German BERT and a regular contributor to the transfer learning framework FARM. Currently he is experimenting with German language model pre-training with a range of different architectures.

The intended audience ranges from researchers to developers. Researchers might be interested in our detailed evaluation. Developers might be interested in the integration of our models into the Hugging Face Transformers library and Deepset's FARM.

8 Speech-to-Text Insights

Manuela Hürlimann, Malgorzata Anna Ulasik, Philippe Schläpfer, Fernando Benites de Azevedo E Souza, Katsiaryna Mlynchyk, Pius von Däniken, Flurin Gishamer, Lina Scarborough, Olesya Ogorodnikova, Tracey Etheridge, Nitin Kumar, Badrudin Stanicki and Mark Cieliebak
Generating high quality transcripts from spoken dialogues (e.g. meetings or interviews) is not a trivial task. Many different Automatic Speech Recognition (ASR) engines exist, both commercial and open source. Two key tasks need to be solved: partitioning the speech according to the different speakers (Diarization), and recognizing what is being said (Speech-to-Text). The quality of the resulting transcript and its usability are influenced by many different factors. In this talk we are going to present multiple insights and techniques which can improve the output quality of ASR. We will address topics such as:

- the recording setting, e.g. which microphone setup is going to give the best results?
- error analysis, e.g. what are typical errors? How can we measure only semantically meaningful errors?
- confidence scoring, e.g. how can we create more reliable confidence scores for the STT and diarization output?

The main goal of our contribution is to present best-practice approaches which can improve both the diarization as well as the transcription quality. Our insights are based on extensive research and experiments, including an evaluation of 10 STT engines and error analysis of more than 70 hours of transcribed speech in German and English.

9 Enabling conversational-based leadership training through advanced natural language understanding

Daniele Puccinelli, Sandra Mitrovic, Denis Broggini, Giancarlo Corti, Luca Chiarabini, Riccardo Mazza, Fabio Rinaldi and Andrea Laus
SkillGym (www.skillgym.com) is a computer-based training system that enables in-role and prospective leaders to develop their communication skills by presenting them with realistic simulations of workplace situations. SkillGym walks the end user through a sequence of videos related to a specific management situation by showing a rich set of alternatives as text boxes. SkillGym also provides extensive feedback, which enables users to review a conversation step by step, and learn the implications of their behavior at each step.

Feedback from SkillGym users praises its engaging training environment. To make simulations even more realistic, our goal is to move from the existing point-and-click interface to a voice-based interface. Achieving this goal requires cutting-edge natural language understanding to interpret the user input in the context of the ongoing flow of the simulated interaction. Our proposed solution is to carry out feature extraction based on the output of a commodity speech-to-text engine so that a dialog state tracker can select the next video based on the user input. Notably, the user must be guided through textual hints to ensure that she provides input that is coherent with the training goals of SkillGym. Moreover, the dialog state tracker must handle all situations where the user input is not aligned with the training goals (e.g. off-topic comments, disambiguation).

Short CV: Daniele Puccinelli is a senior lecturer and researcher at SUPSI and holds a Ph.D. from the University of Notre Dame (USA). His current research interests lie in human-computer interaction.

Intended Audience: practitioners

10 Interactive Poem Generation: when Language Models support Human Creativity

Andrei Popescu-Belis, Aris Xanthos, Valentin Minder, Àlex R. Atrio, Gabriel Luthier and Antonio Rodriguez

Neural language models, which are probability distributions over sequences of words or characters, have recently enabled the generation of fluent sentences and even short texts. However, controlling such models in order to convey specific meanings remains difficult. To study how language modeling can be constrained with text-level features, we have designed a system for interactive poem generation, which enables the joint writing of a poem by a human and a machine. The human first selects the intended form of the poem, e.g. a sonnet or a haiku, although internally any numbers of stanzas and lines are allowed. Using a general-domain neural language model at the character-level, trained on French poems, the system generates a first draft respecting the form. The draft can be modulated according to a desired combination of specific topics (e.g. art, love, or nature) by modifying a number of words using topic-specific language models. Similarly, the draft can be modulated in terms of emotions (happiness, sadness, or aversion). To express their creativity and improve the readability of the poem, humans are allowed to edit it at any stage of the creative process. A strategy to improve rhyming patterns is currently explored. The system has been active since mid-February in the Digital Lyric exhibition. All poems are logged in a database, from which descriptive statistics can be extracted. The system can be demonstrated live at the conference using a large touchscreen.

Bio of the presenter: Andrei Popescu-Belis is professor of computer science at HEIG-VD / HES-SO and a lecturer at EPFL. He is a graduate of the École Polytechnique, with a PhD from the University of Paris-Sud. He has been a researcher in human language technology at the University of Geneva and at the Idiap Research Institute. His interests are in machine translation, information retrieval and human-computer interaction. He has published over 150 refereed papers and edited 12 books/proceedings.

Intended audience: This talk will be of interest to researchers and developers of language technologies, especially those using deep neural language models to generate texts. The talk will also be relevant to those interested in digital humanities and creativity support tools.

11 A conversational recommender system based on neural NLP models

Sandra Mitrović, Vani Kanjirangat, Denis Broggini, Lorenzo Cimasoni, Marco Alberti, Alessandro Antonucci and Fabio Rinaldi

Abstract: In this project, we focus on conversational recommender systems that allow users to specify their preferences through a sequence of dynamically customized interactions, as contrasted to traditional ones. In particular, we seek to improve an online recommendation platform of Stagend (stagend.com) that aims at finding the most suitable performer (”an item”) for a particular event specified by an event organizer (”a user”). In the first phase, an adaptive, Bayesian methods-based approach was used to sequentially update the model given a new piece of information, e.g. performer’s answer to organizer’s question. However, in a real-time setting, delayed/incomplete interactions (e.g. missing reply), can hamper the system efficiency.

To overcome this issue, and also to avoid unnecessary burden on performer (in cases when the answer is already available in performer's biography or previous events' conversations), we investigate the ways of enhancing the Bayesian approach with NLP methods. Specifically, we adopt a question-answering BERT-based approach to either provide a confident automated answer based on the existing information, or to indicate uncertainty and thus, the necessity of contacting the performer. Additionally, given that Stagend operates in multilingual markets, we benchmark different multilingual models such as multilingual BERT and XLM-RoBERTa, as well as compare these with separate language models per each of the target languages (DE + Swiss DE challenge, FR, IT, EN).

Short CV: Sandra Mitrović is a postdoctoral researcher at IDSIA (Dalle Molle Institute for Artificial Intelligence) since November 2019. She has a background in Applied Mathematics and Computer Science (University of Montenegro). She did Masters in Data Mining and Knowledge Management at Université Pierre et Marie Curie, Paris 6 and her PhD at KU Leuven. Her research interests encompass natural language processing, representation learning, (social) network analysis and machine learning in general.

Intended Audience: project managers, developers

12 Swiss German Speech-to-Text with Kaldi

Iuliia Nigmatulina, Tannon Kew and Tanja Samardžić

Recent improvements in speech technology enable its increasing use in a range of applications, including chatbots, online speech translation and smart home devices, among others. While speech technology already achieves strong results for standardised languages, for languages without orthography, with high regional variation and limited training resources, such as Swiss German, it remains a considerable challenge. A high degree of dialectal variability combined with a lack of standardisation leads to extremely sparse data that decreases the quality of alignments between the acoustic signal and its labels and, therefore, the final accuracy.

To tackle the challenge of speech-to-text for Swiss German, we built a speech recognition system using an adapted Kaldi toolkit recipe on multi-dialectal speech data from the ArchiMob corpus. The system was separately trained on two types of writing in the target texts: a) an approximate acoustic transcription that provides a close correspondence between labels and the acoustic signal and b) a normalised writing that potentially reduces the lexical variability. We find that the system trained on the normalised transcriptions currently achieves better results in word error rate (40.81% vs. 54.39%) but underperforms the system trained on the acoustic transcriptions on the character level (character error rate) (23.19% vs. 22.19%). We investigate possible improvements of both approaches and present the outcomes.

CV: Iuliia Nigmatulina received her MA degree in Psycholinguistics and Phonetics from St.Petersbutg State University. She is now a master student in Computational Linguistics and Speech Processing, at the University of Zürich. She is writing currently her master thesis about Acoustic modelling for Swiss German ASR. Her research interests are in the area of automatic speech recognition, sound analysis, phonetics and human-computer interaction.

Tannon Kew: I am a master's student in Computational Linguistics at the University of Zurich in Switzerland. I have a background in Linguistics and language teaching. Throughout my studies, I have worked on multiple projects relating to the development and applicability of large parallel language corpora. In my current research project, I have focused on language representation and modelling for Swiss German speech-to-text systems, under the supervision of Dr. Tanja Samardžić.

The intended audience: developers, project managers, data specialists.

13 Biomedical relation extraction with state-of-the-art neural models

Vani Kanjirangat and Fabio Rinaldi

Typically text mining systems are based upon the identification of mentions of domain entities of relevance (entity recognition and linking), and the identification of their relationships, such as the role of genes in certain diseases, or protein-protein interactions.

We experimented the efficacy of state-of-the-art neural models for extracting high-quality relations from biomedical abstracts. The transformer models, BERT and its biomedical counterpart, BIOBERT were tested as classification models as well as embeddings features.

Experiments were conducted on reference datasets such as the CHEMPROT dataset (Chemical-Protein relations) and the CDR dataset (Chemical-Disease relations). Depending on the dataset used, the tasks varied from binary to multi-class classification and intra-sentential to inter-sentential relation spans. By modelling the problem as a sentence pair classification task, we found that our approach had comparable results with the SOTA models and specifically improved inter-sentential results.

Our research centers on improving the relation extraction models, by analyzing the features captured by the current models. Experiments are done on visualizing the attention flow to exploit the features that were involved in deciding the relations by existing models. These analysis are quite important, especially when the black-box nature of the neural models is considered to be a main pitfall specifically restricting their practical applications.

Short CV: I am currently working as a Researcher in the Natural Language Processing (NLP) lab of IDSIA, Switzerland. I completed my PhD in NLP, which was primarily centered on integrating machine learning and NLP techniques for text plagiarism detection. Ongoing research work includes Biomedical Text Mining, Semantic Shift Detection and Visual Summary Generations using NLP techniques, Temporal Embeddings, Transformers and other Deep Learning models. Alongside, I am working on projects aligned with application of deep learning models in financial and question answering domains.

Intended Audience: project managers, developers

14 MedMon: multilingual social media mining for disease monitoring

Joseph Cornelius, Tilia Ellendorff, Nico Colic, Lenz Furrer, Albert Weichselbraun, Raul Rodriguez-Esteban, Philipp Kuntschik, Mathias Leddin, Juergen Gottowik and Fabio Rinaldi

The MedMon project ("Monitoring of internet resources for pharmaceutical research and development") is a collaborative InnoSuisse project between the University of Zurich, University of Applied Science of the Grisons, and Roche. The project aims to monitor different social platforms on the internet (e.g., Twitter, Reddit, and Medical Forums) to assess patients' perception of their specific disease burden and to discover unmet medical needs. By automating the gathering of patient insights, we enable a more patient-centered drug development and surveillance, particularly for rare diseases.

Bringing together various sources of multilingual micro-posts for disease monitoring has the advantage of ensuring a complete picture by integrating information from all source-types. However, all monitored source-types are inherently different, each posing their own challenges for computational processing.

We discuss specific characteristics, advantages and disadvantages of each source-type and condition (e.g. Parkinson, Multiple Sclerosis, Angelman Syndrome) in the context of automatic medical monitoring. Using the sub-task of personal health mention recognition as an example, we showcase how we addressed these challenges in practice.

Our results give further insights on how to optimally benefit from these multilingual resources and how to integrate them into an efficient model which can be applied in the context

of different disease patterns.

Additionally, in the context of this project the academic partners participated in an international challenge about social media mining for health, achieving top results in two tasks, using deep-learning BERT-based models. Specific methods and results will be presented.

Short CV: Joseph Cornelius works as a research assistant at the Institute of Computational Linguistics at the UZH. He holds a MSc in Neural System and Computation from UZH and ETH. During his master's studies, he has been working on automatic text summarization. His research focuses on state-of-the-art deep learning methods (BERT, BioBERT, etc) for NLP in the biomedical domain. He participated in scientific challenges focusing on social media mining for health, obtaining top scores in two of them.

Intended Audience: project managers, developers

15 A Methodology for Creating Question Answering Corpora Using Inverse Data Annotation

Jan Deriu, Katsiaryna Mlynchyk, Philippe Schläpfer, Alvaro Rodrigo, Dirk von Grünigen, Kurt Stockinger, Eneko Agirre and Mark Cieliebak

In this paper, we introduce a novel methodology to efficiently construct a corpus for question answering over structured data. For this, we introduce an intermediate representation that is based on the logical query plan in a database called Operation Trees (OT). This representation allows us to invert the annotation process without losing flexibility in the types of queries that we generate. Furthermore, it allows for fine-grained alignment of query tokens to OT operations.

In our method, we randomly generate OTs from a context-free grammar. Afterwards, annotators have to write the appropriate natural language question that is represented by the OT. Finally, the annotators assign the tokens to the OT operations. We apply the method to create a new corpus OTTA (Operation Trees and Token Assignment), a large semantic parsing corpus for evaluating natural language interfaces to databases. We compare OTTA to Spider and LC-QuaD 2.0 and show that our methodology more than triples the annotation speed while maintaining the complexity of the queries. Finally, we train a state-of-the-art semantic parsing model on our data and show that our corpus is a challenging dataset and that the token alignment can be leveraged to increase the performance significantly.

This work has been partially funded by the LIH-LITH project supported by the EU ERA-Net CHIST-ERA; the Swiss National Science Foundation (20CH21174237); the Agencia Estatal de Investigación (AEI, Spain) projects PCIN-2017-118and PCIN-2017-085; the INODE project supported by the European Unions Horizon 2020 research and innovation program under grant agreement No863410.

16 Following, understanding, and supporting service-oriented person-to-person communications

Alexandros Paramythis, Doris Paramythis and Andreas Putzinger

Automation in enterprise service provision has proliferated in recent years. In service-based communications, such automation typically has the form of Chatbots or Interactive Voice Response systems, of varying sophistication. Despite very significant improvements achieved in the corresponding technologies, recent studies show that in the domain of service-oriented communications, person-to-person interaction is highly more effective and efficient. This has given rise to a new generation of products that seek to empower humans engaging in such interaction, rather than replace them.

The main prerequisites for providing support during person-to-person communication are: on the one hand, being able to observe the ongoing interaction as it happens, bringing it to a computable form in (near) real time (e.g., through automatic speech recognition); and, on the other hand, being able to semantically interpret utterances in context. The second part specifically entails natural language understanding coupled with a semantic representation of the domain of intercourse that can be used for reasoning.

In this presentation we outline our experiences with applying approaches from the fields of natural language processing and ontological domain modeling for the interpretation of dialogue acts, and also for the analysis of domain-specific data (e.g., product documentation), targeted to identifying the pieces of information most relevant to an ongoing person-to-person dialogue in real time.

17 Text Mining Technologies for Animal Health Surveillance

Fabio Rinaldi, Anne Goehring, Corinne Gurtner, John Berezowski, Michele Bodmer, Irene Zuehlke and Celine Faverjon

We describe the outcomes of a collaborative project between the Vetsuisse faculty of the University of Bern and the Institute of Computational Linguistics of the University of Zurich, aimed at exploiting text mining technologies in the analysis of pathology reports from multiple Swiss veterinary laboratories. An online tool has been developed which allows the dynamic processing of batches of reports for the extraction of relevant signals, which in turn can be used for statistical analysis in epidemiological studies. The process is based on the identification in the reports of terminological items referring to relevant domain concepts. The terminologies used in the project are sourced from several ontological resources. We have also developed a semi-automated process to cross-map our ontological resources through a reference ontology, such as the UMLS.

In a first step we evaluated the completeness and validity of the necropsy data. In a second step, we combined information extracted from the three necropsy data sources, and investigated factors associated with necropsy submissions at three different levels – "national" , "farm" and "individual" – and according to age, region and time of the year.

An interactive dashboard application enables data exploration. The combined pathology data from several veterinary pathology laboratories can be spatially and temporally displayed for different types of analysis. All aspects of the projects have been assessed for their potential benefits for animal health surveillance.

Short CV of the presenter: Fabio Rinaldi leads the NLP research group at the Dalle Molle Institute for Artificial Intelligence Research (IDSIA). Previously he was a lecturer and senior researcher at the University of Zurich, as well as a PI in numerous research projects, which he acquired and managed.

He has an academic background in computer science and more than 25 years experience in NLP research, with a specific focus on applications in the biomedical domain, such as automatic analysis of the scientific literature, of clinical reports, and health-related social media discussions. He also authored more than 100 scientific papers (including more than 30 journal papers).

Intended Audience: decision makers, project managers

18 Assigning Grant Applications to Reviewers via Text Analysis

Anne Jorstad

The Swiss National Science Foundation normally finds the most appropriate expert reviewer for

each grant application by hand. However, when the pool of reviewers is known in advance, this process can be performed more efficiently using text mining.

An application can be represented by the text of its title, keywords, and abstract. Potential reviewers can be defined by similar texts from their publications. We have tested a variety of techniques to define the similarity between pairs of texts, followed by an optimization procedure to determine the final matching, given constraints about the number of applications allowed per reviewer.

The biggest challenge is due to the fact that the amount of discriminatory information provided in these texts varies widely between disciplines. Humanities and social sciences texts tend to use standard language vocabulary such as "law" or "urban", while the hard sciences include very specific terminology like "SARS-CoV-2" or "latent semantic analysis". And some expressions overlap, but carry different meanings in different fields, such as "family" or "support", which are generally not meant in the context of "family of algorithms" or "support vector machines".

We aim to develop a system that can appropriately assign applications to reviewers for funding schemes as multi-disciplinary as Spark ("rapid funding of unconventional ideas") and as mono-disciplinary as our new Coronavirus call. We note that this algorithm will not be applied for all funding schemes at the SNSF.

Intended Audience: Developers and decision makers. Specifically those who need to pair texts from a variety of topics simultaneously. We would also like to get feedback from researchers in related fields to improve all aspects of our algorithm.

Author CV:

Professional Experience:

- Swiss National Science Foundation, Data Scientist, 2014-Present
- Ecole Polytechnique Fédérale de Lausanne (EPFL), Postdoc, 2012-Present
- Johns Hopkins Applied Physics Lab, Research Intern, 2008-2010 (summers)

Education:

- PhD, Applied Mathematics, University of Maryland, USA, 2012
- Visiting Doctoral Student, ENS Cachan, Paris, France, 2010
- Master, Mathematics, University of Wisconsin, USA, 2007
- Bachelor, Mathematics (Computer Science Concentration), Cornell University, USA, 2005

19 Named entity recognition for job description mining

Dina Wieman, Khan Ozol, Natalia Korchagina, Claudio Bonesana, Anastassia Shaitarova and Fabio Rinaldi

In a collaborative project with a major pharma company we explored name entity recognition (NER) strategies applied to job/resume mining tasks. In the project we leveraged advanced NER approaches in order to identify job titles, organization names, and geographical locations which are the essential parts of a job mining task, such as recruiting, tracking job candidates and job recommendation. This process is currently based on the manual analysis of hundreds of CVs, often with no relevance for a specific position or a profile.

Despite the existence of many commercial providers of similar services, there are no publicly available datasets to evaluate the advertised algorithms. The existing pre-trained NER models such as spaCy models, and Stanford NER models were trained on blogs, news and

media. Their performance drops significantly when applied on the sentences taken from the resumes, since titles, locations and organization names in a resume are often written in the manner of a heading.

We asked domain experts to manually annotate a reference dataset of free-text job title description extracted from CVs, used it to train a deep-learning model, and compared the results against the reference models mentioned above. We were able to outperform both pre-trained models by a significant margin. Our NER models have been integrated in a prototype system which demonstrates a more dynamic and flexible data analysis compared to baseline commercial solutions.

Short CV: Fabio Rinaldi leads the NLP research group at the Dalle Molle Institute for Artificial Intelligence Research (IDSIA). Previously he was a lecturer and senior researcher at the University of Zurich, as well as a PI in numerous research projects, which he acquired and managed.

He has an academic background in computer science and more than 25 years experience in NLP research, with a specific focus on applications in the biomedical domain, such as automatic analysis of the scientific literature, of clinical reports, and health-related social media discussions. He also authored more than 100 scientific papers (including more than 30 journal papers).

Intended Audience: decision makers, project managers

20 Company Name Disambiguation

Ahmad Aghaebrahimian and Mark Cieliebak

Company Name Disambiguation (CND) is a form of Named Entity Disambiguation where different textual representations of a company name are linked to its formal name. For instance, the company 'ArcelorMittal SA' is often referred to as 'Arcelor Mittal Group', 'Mittal Steel', or simply 'Mittal Co.'. The task of mapping these surface forms to the same company formal name is known as CND or in a more general term, Named Entity Disambiguation (NED). NED is a crucial task in many Natural Language Processing applications such as entity linking, record linkage, knowledge base construction, or relation extraction, to name a few. It has been shown that parameter-less models for NED do not generalize to other domains very well. On the other hand, parametric learning models do not scale well with a large number of candidate names which is often the case for CND since the number of company formal names usually exceeds hundreds of thousands of instances. Yet another challenge is multilingual NED; while company formal names are often in English, texts and company mentions are in another language which makes string matching impractical.

In this talk, I elaborate on a wide range of techniques we use to tackle these challenges for a proprietary CND system. I will talk about our parameterized and non-parameterized models, string normalization, encoding and disambiguation on the scale. Eventually, I present the audience with the state-of-the-art results we obtained on three publicly available datasets using our CND system.