6th Swiss Text Analytics Conference (SwissText): Abstracts of the Applied Track

Contents

1 Automatic Classification of Service Desk Tickets

Michael Zemp, Don Tuggener, Stephen Cryan and Mark Cieliebak

Manual routing of support tickets (or service desk tickets) within a large organization is difficult, prone to error, and requires a lot of domain-specific knowledge. Even service desk agents struggle with finding the affected services and support groups. To help these central support units assign tickets, statistical models and neural networks can be trained based on actual historical support cases.

We present a solution implemented for the Julius Baer Group, a medium-sized internationally active private bank. The enterprise has multiple service desks for different sort of inquiries (business or general) and regions. They accept re-quests from the organization through different channels: phone, mail or intranet forms. Fortunately, they all share the same data model as well as a common ticketing system.

Results: Using the historic tickets, text classification with deep learning has shown a 92% validation accuracy in predicting the affected service and therefore enabling an initial routing of the tickets. Despite poor text quality and mixed languages, a macro F1-score of 0.75 is achieved over more than 300 classes. Especially preprocessing has shown a significant impact on how individual classes are being predicted. There is even more automation potential, as the tickets can also be classified into requests and incidents with a 93% accuracy. Finding the responsible support group itself that can resolve an incident has shown promising results, yet the available data itself lacks the needed information. It still succeeds to present a good selection of responsible groups, out of which the top 4 can resolve the issue with a probability of 96Methods: Different neural networks such as CNNs, LSTMs and Transformers have been tried with different configurations as well as different tokenization techniques. Overall, trained word embeddings in combination with convolutional layers achieve the best results in accuracy and macro F1-score. Further pretraining the language model of RoBERTa, a state-of-the-art transformer-based ma-chine learning technique, outperformed the best CNN by 0.018 in macro F1-score. However, when it comes to real-world application and usability, their predictions suffer from the low text quality and lack the domain-specific terms.

Real-World Application: During development, the practicability was evaluated frequently by experienced service desk agents, who were provided the prediction features for their daily work. The developed models are used now on a daily basis on the production environment. One measurable benefit is that quality managers have less work in terms of adjusting tickets and can instead focus on their other tasks. The support group predictions, which are implemented in the live system as well, help service desk agents to route the tickets to the corresponding services and support groups. Based on these successes, there is already the requirement to start fully automated labelling of inquiries which the bank receives through channels like email and intranet form.

2 What's Hot in the News Today?

Claudia Schulz

To stay tuned about the daily news, you probably browse your favourite news outlet every morning. However, not everything you read is equally "hot". Some of the articles you come across will be covered by other news providers too. These are the "hot topics" everyone will be discussing throughout the day. So which headlines are "hot", i.e. most covered, among the thousands of articles published in print media as well as online every day?

To answer this question, we evaluated various off-the-shelf document clustering and topic extraction algorithms and libraries, all resulting in poor performance. We thus created our own pipeline to determine daily hot topics, using a mixture of state-of-the-art neural models and

old-school NLP techniques to tackle the document clustering and text summarisation tasks. In addition to performance results, this presentation covers challenges encountered, such as dealing with articles in multiple different languages, and lessons learned.

3 Blackbox Testing for Conversational Systems

David Klaper

Conversational systems in production vary widely in quality. Testing such systems is the most important way to ensure quality. This talk focuses on raising awareness for testing conversational systems with insights from practical experience.

Testing NLP systems can be tricky. Interpreting the underlying models is difficult and in some cases they are not in our control. This hinders building a conversational solution that improves over time without breaking the existing understanding. We would like to share our approach to this problem and outline our key learnings from voice projects over various use cases. It focuses on testing without looking inside the model, so-called blackbox testing.

A central lesson is that blackbox testing can reveal meaningful insights into the model and training data. Suddenly failing tests are not just a nuisance but also a chance to find structural problems. Thus, there are some strategies for analyzing test failures and fixing them the right way.

Then, conversational systems can be tested with different purposes and levels. We show what tests we have right now and what additional test types we consider useful for the future.

Finally, conversational tests can serve as a crucial communication tool between developers, project managers and business experts. In contrast to testing other kinds of software, it is much more natural to formulate tests in a language easily understandable for a wider range of stakeholders than just developers.

4 Swiss Voice Project: Swiss German Speech Synthesis

Julian Mäder

At the ETH Media Technology Center, we started the Swiss Voice Project to research the technical possibilities required to build Swiss German voice assistants, with a focus on text-to-speech models. Swiss German is a low-resource language with a dialect continuum, for which only little data and no standardized written form is available. This makes building text-to-speech models challenging, since state-of-the-art models for languages such as English and German are data-driven and rely on large data sets. The Swiss Voice project tries to address those challenges with the following contributions.

In this talk we will highlight the key findings and outcomes from this project. We introduce the first annotated parallel corpus of spoken Swiss German across 8 different dialects, with Swiss German transcripts and parallel standard German reference translations. This data set will enable the NLP community to research and evaluate Swiss German speech synthesis and the use of High German resources for transfer learning. Based on our data set we trained powerful deep learning models for machine translation and speech synthesis and deployed the Swiss Voice REST API, which takes High German text as input and produces audio in all 8 supported dialects. Building on our API and the Speech Recognition Technology from recapp AG, we created the first voice assistant prototype that understands and speaks Swiss German. The jointly developed prototype answers everyday requests, such as reading news headlines and weather forecasts in any of the 8 dialects currently available.

5 How We Crowdsourced a Large Swiss German Speech-to-Text Dataset: "Die Schweizer Dialektsammlung"

Manuela Hürlimann, Michel Plüss, Mark Cuny, Alla Stöckli, Malgorzata Anna Ulasik, Manfred Vogel and Mark Cieliebak

Speech-to-Text is essential for many innovative applications such as automated customer support or automatic transcription of meetings. While this is a solved problem for languages such as English, Spanish or Chinese, there are currently no solutions for Swiss German because of the lack of suitable training data.

Our project "Die Schweizer Dialektsammlung" ("The Swiss Dialect Collection") creates such a large training dataset using crowdsourcing by prompting users to speak a provided Standard German text in their dialect. Our goal is to collect at least 2'000 hours of Swiss German audio - equivalent to approximately 650'000 samples. We will discuss how we organised such a large-scale endeavour, and how we motivated the Swiss population to take part.

We will discuss the following aspects of the project:

- Demonstrate the web application for data collection, based on the Mozilla CommonVoice open source code base.
- Present our strategy for reaching and addressing various target groups so as to motivate them to participate in the data collection. We will show the different strategies (including gamification/rewards) and messages and evaluate their success.
- Provide insights into the collected data (e.g. geographical distribution, quality) and how high-quality Speech-to-Text models for Swiss German can be trained on the collected data.

6 Automatic Document Parsing with Weak Supervision from Spreadsheets

Johannes Rausch, Susie Xi Rao, Ce Zhang and Peter Egger

Automated processing of electronic documents is a common downstream ML task in industry and research. Here, the lack of structures in formats such as PDF files or scanned documents remains a major obstacle. In practice, extensive engineering and ad-hoc code are required to recover document structures, e.g., for headings, tables, or nested figures.

ML-based systems require large amounts of labeled training data. In DocParser, Rausch et al. [2021] have shown a robust way to parse complete document structures from rendered inputs. This problem is solved through a novel weak supervision approach that automatically generates training data from structured LaTeX source files of readily available scientific articles.

However, the application in new target domains, e.g., the business context, is hindered by the mismatch between source and target document layouts. Furthermore, weak supervision often cannot be applied, if business documents are generated from different source files, e.g., office software.

As a result, producing sufficient training data sets demands high human labeling efforts.

We contribute a new system to automatically generate weakly labeled data for ML-based document parsers in real-world settings. We achieve this by developing a weak annotation scheme that can utilize source file types such as spreadsheets. In a preliminary case study, we are able to generate 12,000 annotated samples from Excel files. We demonstrate that training our system on this weakly-annotated data allows accurate parsing of structured data in real-world office documents. Previous findings in DocParser suggests that pretraining systems

using such weak supervision scheme can lead to a significant reduction of labeling complexity. (Acknowledgement: We thank Mr. Livio Kaiser for supporting this project during his master thesis.)

7 Boosting the Quality of Experience of a Leadership Training System with Natural Language Processing

Oscar Lithgow-Serrano, Denis Broggini, Giancarlo Corti, Luca Chiarabini, Daniele Puccinelli, Fabio Rinaldi and Andrea Laus

SkillGym is a computer-based training system designed to improve communication skills by walking the end-user through a sequence of videos related to specific management situations. As part of the Innosuisse-funded project "Boosting the SkillGym Quality of Experience with Artificial Intelligence" (BOOST), our team is developing a novel technological solution that leverages voice-based interaction and natural language processing to boost SkillGym's Quality of Experience. This has presented various demanding challenges. Here we present the practical path that we have followed to achieve our goal. At each simulation step, the user is offered a selection of hints on what to say to the system. The hints are carefully crafted to move the simulation along based on the available pre-recorded video segments. The user is free to express herself in her own words. In general terms, the objective is to select the pre-recorded video segment whose content provides the best fit as a response to the user utterance. The freedom afforded to the user requires the system to leverage a multi-stage approach to sort out the practical difficulties of dealing with free natural language interaction. The tasks addressed include the detection of unsuitable utterances, the detection and completion of incomplete utterances, the detection and correction of spurious words introduced by the voice-to-text transcription, and fine-grained utterance classification. The implemented solutions range from simple heuristics based on grammatical features to fine-tuned deep learning models with some hybrid approaches in between.

8 Speaker Recognition

Jan Deriu, Amin Moghaddam, Malgorzata Anna Ulasik, Katsiaryna Mlynchyk and Mark Cieliebak

When we transfer insights from the research setting to a real-world application, problems often arise, which never occurred in the clean lab setting. In this talk, we showcase this problem in the context of speaker diarization or speaker recognition, i.e., the task of segmenting a recorded conversation into "Who spoke when?". Potential use-cases include the creation of transcripts of interviews, or for authentication of speakers.

One major issue is that it is not essential in the lab setting if the change of one speaker to the other is precise to the tenth of a second. However, in the real-world setting, this is crucial. For instance, when we want to get an automatic transcript of an interview, one-tenth of a second might shift a couple of words to the wrong speaker, which leads to the software being perceived as low quality.

This talk aims to showcase how we tackled this issue using speech processing and linguistic features. Furthermore, we show how the setting in which diarization is trained in a lab setting differs from the actual use-case and how this can be leveraged (e.g., in our use-case, we have information from an ASR system, and we know the number of speakers).

9 FLIE with Rules

Ela Pustulka-Hunt, Thomas Hanne and Lucia de Espona

FLIE (Form Labelling for Information Extraction) allows us to extract information from Swiss insurance policies. Insurance policies are forms which are weakly aligned and do not lend themselves to automated data extraction without preprocessing. Our preprocessing annotates data with geometry and combined with manual training data generation gives the extraction accuracy of over 80% for a subset of attributes which have been seen 8 times or more.

In this paper we extend FLIE with rules. The aim is to compare machine learning used in FLIE to the standard industry approach of using rules to extract data. We hand crafted rules (regular expressions in Python) for the KTG insurance (27 rules), UVG insurance (29 rules), and UVG-Z (23 rules), for each insurance type covering around 20 attributes. We also generated rules for building insurance policies which we were new to (16 rules encoded in SpaCy). In all cases we saw that using rules alone gives us a similar accuracy in data extraction to machine learning (around 80%). In the case of building insurance the accuracy is higher, above 96%, with precision and recall around 89-92%. To support annotation and experimental evaluation, we created an annotation GUI and a GUI which automates the ML experiment. Planned work includes a comparison of rule based and ML approaches and extension to further policy types.

10 CareerCoach: Automatic Knowledge Extraction and Recommender Systems for Personalized Re- and Upskilling suggestions

Albert Weichselbraun, Roger Waldvogel, Philipp Kuntschik, Alexander van Schie and Andreas Fraefel

Competition, advances in science and technology and crises, such as the COVID-19 pandemic, trigger serious disturbances in the job market that devalue certain skills and job profiles while generating demand for new ones. Re- and up-skilling are efficient mitigation strategies but given the sheer amount of available educational programs, up-skilling decisions are far from trivial. The CareerCoach project develops a recommender system that combines the user's personal context such as age, qualification and interests with real-time data on the demand within the job market to suggest re- and up-skilling options that optimize the user's chances within the market. In order to do so, the CareerCoach recommender applies sophisticated knowledge extraction methods such as named entity linking and slot filling to web pages from educational programs and institutions. The extracted knowledge is then integrated in the industry partner's domain ontology, enabling content-based recommendations that consider graph-based methods, embeddings, hybrid approaches and state-of-the-art neural models. Once completed, CareerCoach will provide a comprehensive knowledge graph of educational programs and a recommender system that aids individuals in their personal re- and up-skilling decisions. On a societal level, the project will support the drafting of up-skilling strategies for quickly reintegrating job-seekers into new positions and assist policies that mitigate disruptions of the job market.

11 Use-Case Driven Tokenization for Non-Standard Corpora

Niclas Bodenmann

Emerging, non-standard domains such as computer-mediated communication are becoming more and more prevalent in digital humanities research. But starting with preprocessing, researchers often use tools that are not made for the idiosyncrasies of their data. In the context of text-based research, this is also prevalent in tokenization.

My BA-thesis and the corresponding poster discuss the requirements for tokenization in the context of digital humanities. I argue for a white-box tokenization approach, that is still able to adapt to idiosyncrasies that are emergent in the sense that they haven't necessarily been seen outside the data at hand. The first premise calls for a rule-based tokenizer, while the second one calls for a tokenizer that is informed by the corpus itself.

This tension is resolved by using Cutter (Graën et al., 2018) and a way to inductively identify difficult sentences in the corpus. Cutter is a tokenization framework that facilitates writing tokenization rules once a few gold-standard sentences have been tokenized by hand. Instead of thinking up artificially difficult sentences for this, I use a novel approach based on "Inter-Tokenizer-Disagreement" to identify such sentences: Several off-the-shelve-tokenizers are tasked to tokenize the same utterance, and the more their results deviate, the more difficult the utterance is deemed.

This approach has been tested on a corpus of user comments concerned with COVID from swiss newspaper websites.

12 Understanding structural information in scanned documents

Holger Keibel

HIBU is a proprietary solution platform based on which Karakun (Basel) builds customer solutions around enterprise search and text analytics. In our customer projects we often have to deal with scanned documents which are digitized through OCR software. While for unstructured documents (running text), text analytics algorithms can directly operate on the OCR output as a subsequent and independent processing step, doing the same on structured sections of a document (such as forms and tables) tends to produce poor results because the OCR result does not capture the non-sequential relations in such structured sections (e.g. interpret a table cell relative to its column title).

In this talk we illustrate how this can be improved if the text analytics algorithms are integrated into the OCR step for these structured document sections. Example use cases from customer projects will be presented.

13 From Normalized Swiss German to Standard German with Synthetic Parallel Data

Benjamin Suter, Larissa Schmidt and Josef Novak

When working with Swiss German, the raw data (text or audio in various dialects) is usually mapped to a normalized representation. The exact format of this representation can vary, but a common practice is to map each Swiss German word to a unique Standard German word. However, this approach generally does not result in entirely correct Standard German, as both syntax and word choice may differ from correct Standard German. For instance, relative clauses in normalized Swiss German may be introduced by the particle ¡wo¿, whereas they need to be introduced by a correctly declined form of ¡der/die/das¿ in Standard German.

We present an approach to transform normalized Swiss German into Standard German with fully synthetic parallel data. Starting from a corpus of Standard German which will be used as the target-side corpus, we automatically generate a source-side corpus by probabilistically applying text transformations and corruptions, including insertions, deletions, swaps, and substitutions of certain words. The process requires some lingustic expertise, but no manual creation of a parallel corpus. The same approach can be used to additionally predict correct punctuation and capitalization of words in order to fully restore Standard German orthography. We show

that a sequence-to-sequence model trained on the described synthetic parallel corpus achieves very good quality in terms of both word error rate and human judgment.

14 Question classification in German dialogues

Alexandros Paramythis and Eleni Adamantidou

The classification of Dialogue Acts (DAs) is a crucial component of conversational systems. Existing approaches have focused predominantly on dialogues in English. As far as German is concerned, there are approaches "dedicated" to the language, as well as applicable multi-lingual ones; in either case, however, these fall behind the current state-of-the-art (SotA) for English.

In Contexity we are developing smartCC, an intelligent assistant for customer care services, which "observes" dialogues between agents and customers, and provides live support. In such a context, the classification of any type of DA with high precision is critical; even more so for questions asked, or requests made, by either party.

To improve smartCC's capabilities in this respect, we have started an open-source project, in which we are experimenting with SotA approaches to DA classification, applied to German. Specifically, we are testing deep learning-based techniques, successfully applied in the recent past in English, as well as ensembles of the said techniques with ones based on the analysis of utterance syntax and semantics.

We anticipate the result of this work to be an open-source standalone system –with accompanying models– that can classify German dialogue utterances as specific question types (e.g., yes/no-, wh-, declarative-questions, etc.) Our participation in SwissText would allow us to share our results with the community, and draw attention and external contributions to the project.

15 A complete End-to-End Human Resources system backed by NLP and Deep Learning

Alejandro Jesús Castañeira Rodríguez

Within the field of human resources there are still many challenges ahead, by example, keyword based searches or manual resume screening could contribute to the omission of valuable opportunities for all the parties involved or to lengthy hiring procedures. Therefore, we would like to present a Recommender System in the field of HR, which comprises a series of Natural Language Processing techniques and Deep Learning models, that allow achieving a fully automated process which will propose semantically related and explainable suggestions to job seekers and companies in real time. The described system is running in production at Janzz.technology where it's used to match candidates and job seekers on a daily-basis so we would like to show a live demo. The architecture of the system combines the JANZZon! Ontology with several NLP techniques like Named Entity Recognition, Text Classification, and Entity Relationships based on HR data collected by Domain Specialized Curators and it also includes Language Detection, specific Text Preprocessing and Language Models pretrained over domain specific data, which allows to extracts and processes more than 50 different characteristics from job postings and resumes such as occupations, skills, education, etc., with various levels of granularity across multiple languages.

16 Neural Machine Translation in Academic Contexts

Alice Delorme Benites and Fernando Benites

English is the main lingua franca for the international scientific dialogue. However, several

studies show that a lesser command of English is a factor of exclusion for many non-native researchers. In medicine, English competence robustly predicts chances of publication in leading journals – with a predictive weight that surpasses total financial investment in research. Non-native researchers often turn to free online neural machine translation tools (NMT) to spare the additional temporal and financial costs imposed by journals' expectation that submissions are "proofread by a native speaker". However, general NMT cannot account for the specificities of German academic texts: terminology (i.e. German ad hoc compounding, frequent neologisms, popular "German" terms existing parallel to scientific Greco-latin terms), syntax (i.e. presenting constructions, grammatical subjects and rhematized long subjects) and hedging (i.e. German modal verbs) are the three main issues. By fine-tuning a state-of-the-art neural machine translation system on specific academic corpora, we hope to tackle these issues and to compare which errors were corrected and which were not. Such a domain specific approach could be adopted by institutions to train in-house systems and become a standard tool provided by the institution to their researchers like any other intern application.

17 SST-BERT: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces

Vani Kanjirangat, Sandra Mitrović, Alessandro Antonucci and Fabio Rinaldi

Lexical semantic change detection (also known as semantic shift tracing) is the task of identifying words that have changed their meaning over time. The task of unsupervised semantic shift tracing in SemEval2020 is particularly challenging. This task comprised two subtasks; The first one refers to a binary clustering task, where we have to identify whether a word gains or loses a sense over the time period. The second subtask was a ranking task, to measure the degree of lexical semantic change. Given the unsupervised setup, in this work, we propose to identify clusters among different occurrences of each target word, considering these as representatives of different word meanings or senses. As such, disagreements in the obtained clusters naturally allow quantifying the level of semantic shift per each target word in the four given target languages (German, English, Swedish, and Latin). To leverage this idea, clustering is performed on contextualized (BERT-based) embeddings of word occurrences. The ranking task is done using Shannon-Jensen distance. The obtained results show that our approach performs well both measured separately (per language) and overall, where we surpass the provided SemEval baselines. The approach would be interesting with respect to any language, especially to understand the linguistic properties and variations of the words within the language and how they have evolved over time.

Original Paper: Kanjirangat, V., Mitrovic, S., Antonucci, A., Rinaldi, F. (2020, December). SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 214-221). (https://github.com/vanikanjirangat/SST-BERT-SEMEVAL-TASK1)