7th Swiss Text Analytics Conference (SwissText 2022): Abstracts of the Applied and Demo Tracks

Contents

1 Classification of Large Patent Descriptions

Fernando Benites, Dominik Frefel, Joshua Meier and Daniel Perruchoud

Classification of long documents is still a domain for classical machine learning techniques such as TF-IDF or BM25 with Support Vector Machines. Transformers and LSTMs do not scale well with the document length at training and inference time. For patents, this is a critical handicap since the key innovation is often described towards the end of the patent description, which varies in structure and length and can be relatively long.

Furthermore, because the class ontology for patents is very deep, specific classification can only be performed by looking at the differences that might be named in any part of the document. Therefore, it is advantageous to process the whole patent and not only specific parts.

We investigate hierarchical approaches that break down documents into smaller parts and other heuristics, such as summarization and hotspot detection, for Bert and PatentBERT and compare them to classical methods. The dataset was downloaded from the European patent office (EPO).

2 Swissdox@LiRI: Swiss Media Data for Big Data Research at your Fingertips

Noah Bubenhofer, Stefan Bircher, Johannes Graën, Daniel McDonald, Igor Mustač, Nikolina Rajović and Jonathan Schaber

Swissdox@LiRI offers unique access to Swiss media for Big Data analyses. In cooperation with the Swiss media database SMD, the technology platform "Linguistic Research Infrastructure" (LiRI) provides access to over 260 media titles in Switzerland, including newspapers such as Neue Zürcher Zeitung, Tages-Anzeiger, CH-Media and Blick, but also to public service broadcasting services SRG.

The service is intended for academic Big Data research: computational and corpus linguistics, but also social sciences, history and economics and many other disciplines and methods can benefit from the data. Funding is provided through subscriptions: the universities of Zurich, Bern and Basel, the ZHAW and the ETH are already supporters. This means that all members of these institutions, whether students or researchers, have free access to the data: The data can be processed using NLP, used for data mining or as a basis for language models.

In addition to academic use, commercial use of the data is also possible. In this case, the data can also be used via Swissdox@LiRI, whereby the SMD charges adapted rates for this.

The project is unique for several reasons: Normally, it is difficult to give research access to media data for copyright reasons. Moreover, existing offers, e.g. in the context of libraries, usually do not allow Big Data use, but only enable the download of individual articles. Swissdox@LiRI, however, enables extensive use of the data for research on a legally sound basis.

3 Herlock.ai - Finding the missing Link in 500 federal folders comprising of over 100'000 pages of evidence

Jacqueline Stählin and Jürg Burkhalter

For the compliance and legal profession, the exponential growth of data is both a threat and a promise. A threat, because finding crucial facts buried in hundreds of thousands of documents is hard. A promise, because proper management, analysis, and interpretation of data provides a competitive advantage and full transparency during all stages of investigations.

In this talk, we present our novel text analytics platform Herlock.ai to leverage these possibilities.

Herlock.ai finds mentions of persons, dates, and locations in the corpus and makes these findings available to the user. In order to achieve this, several hurdles have to be overcome. Paper documents need to become machine readable. Even when the digital version exactly replicates the paper, the data is not available for analysis because of human inconsistencies and errors.
Herlock.ai fixes these problems and provides excellent content. In order to support the user in their work, Herlock.ai needs to be easy to use and understand, e.g. by splitting documents into meaningful parts, by comparing different variants, and by marking textual anomalies.

We will do a demo of Herlock.ai. The platform has been used in a recent Swiss legal case that has received high media coverage. 500 federal folders that physically fill entire walls of shelves were an unprecedented challenge for the involved parties. The quick search and navigation is a key tool and the analytics we provided were used for official submissions to the court.

4 Closed-Domain Task-Oriented Conversational Agents with Embedded Intelligence

Reda Bousbah, Simone Sguazza, Luca Chiarabini, Massimo Coluzzi, Hichem Ferchichi, Omran Ayoub, Aldo Polledro, Michele Giacobazzi, Daniele Puccinelli and Roberto Mastropietro

Swiss software company WellD and its SaaS hotel tech spinoff TellTheHotel wish to develop a closed-domain task-oriented conversational agent for the hospitality industry.

As part of the Innosuisse-funded project TACO “Closed-domain task-oriented conversational agents with embedded intelligence”, SUPSI is helping WellD and TellTheHotel leverage the state of the art of NLP to build a custom conversational agent. The main objective is the development of a multi-language, multichannel, digital concierge that enables end users to complete a hotel reservation as well as ancillary activities through a natural conversational flow.

The conversational agent is based on customized cutting-edge NLP techniques and the RASA framework. Basic hotel reservation requests are handled with multi-language intent detection, which is carried out through the application of BERT-based cross-lingual sentence embeddings, with substantial benefits compared to translation-based systems. User queries that go beyond room reservations are handled with a BERT model fine-tuned for question answering.

From an architectural point of view, the project is developed based on micro-services and deployed as a Kubernetes cluster to ensure scalability.

5 Document Classification and Information Extraction for Workflow Selection

Luca Chiarabini, Fabio Landoni, Olmo Barberis, Giancarlo Corti, Daniele Puccinelli, Davide Vosti and Sandro Pedrazzini

Karakun AG and DSwiss AG have developed a leading-edge product that enables intelligent full-text search across user documents in encrypted repositories. As part of the Innosuisse-funded project "Document classification and information extraction for workflow selection", SUPSI, in collaboration with Karakun and DSwiss, has developed the EXTRA module, a novel technological solution that leverages NLP for document classification and information extraction to streamline the user experience by automatically processing the content of user documents.

The EXTRA module can classify various common kinds of user data and extract relevant pieces of information for specific document types, such as invoices. The implemented solutions range from transfer learning to rule-based approaches and computer vision.

6 Multitask Prompted Engineering for Patents

Dimitrios Christofidellis, Antonio Berrios Torres, Ashish Dave, Manuel Roveri, Kristin Schmidt, Sarath Swaminathan, Hans Vandierendonck, Dmitry Zubarev and Matteo Manica

Patents are one of the principal tools for establishing ownership rights to the creations and inventions in industry. They can be used to generate tangible economic benefits to their owner. Submitting a patent is an expensive task both in terms of time and money. An extensive prior art search should take place to determine the novelty and value of the patent followed by a careful drafting of the patent to precisely capture the idea. Participating in the drafting process typically requires a great expertise both in the technical field and in the right writing. Undoubtedly, natural language tools can be utilized for the patent domain and facilitate both their generation and analysis. Yet, the prominent transformer based language models, such as BERT and GPT-2, have been trained in corpora which do not resemble the vocabulary and syntax of the patent domain. As such, application of these models does not guarantee good performance. To this end, we propose a language model able to understand the patent domain and capable of performing several related tasks. Following the recent trend of multitask prompt-based language modeling, we cast three main patent related tasks (generation, editing, quality check) as a text generation problem using the proper prompts and we fine-tune the standard GPT-2 model for them. The achieved performance indicates that our model can be reliably incorporated in patent generation and analysis pipelines and speed up the patent related processes.

7 Interscriber: Automatic Meeting Transcription and Analysis (Demo)

Mark Cieliebak, Alexandros Paramythis, Daniel Neururer, Nicola Good, Yuliya Parfenova, Nauman Arif, Panagiotis Gkikopoulos, Don Tuggener, Karim Zerika and Zhivar Sourati

Interscriber is an innovative AI-based system that can be used to transcribe audio recordings of interviews, meetings and discussions automatically. This also includes speaker recognition and auto-generated time stamps. For this purpose, we have implemented automatic systems that combine several systems for speech-to-text and speaker diarization via ensembling methods. At the moment, Interscriber is available for English and German, and a prototype implementation in Swiss German.

Since the automatic transcription usually contains errors, Interscriber also offers a customized editor that can be used to post-process and improve the texts. Generated and edited texts can be subsequently exported in several formats.

Finally, Interscriber allows to explore the generated transcripts interactively and to generate summary texts. In the demo, we will showcase the quality of the resulting transcripts and discuss with the audience general challenges when implementing speech processing in a real-world product.

8 “Schweizer Dialektsammlung” - Results, Learnings and Next Steps

Manuela Hürlimann, Michel Plüss, Jan Milan Deriu, Christian Scheller, Manfred Vogel and Mark Cieliebak

The project “Schweizer Dialektsammlung” (“Swiss Dialect Collection”) has been running since spring 2021. Its goal is to collect a large dataset with Swiss German audio samples and their transcriptions to Standard German text. So far, we have crowdsourced 200 hours of audio from nearly 4000 volunteers via a web recording platform, equivalent to over 150’000 text prompts. The dataset is called SDS-200 and will be released for research purposes.

In a related project funded by the Schweizer Nationalfonds (SNF), we are using SDS-200 together with parallel dialect data to find out how Swiss German Speech-to-text (STT) systems can better recognise dialects for which little annotated data is available. Initial experimental results show that including SDS-200 as part of the training data significantly enhances STT performance: the BLEU score on the All Swiss German Dialects Test Set improves from 48 to 65 when we add SDS-200.

We are also planning the next phase of “Schweizer Dialektsammlung”, where users can form teams and compete for prizes.

We will

- present the project and the data collected so far
- discuss our Speech-to-Text experiments and results
- talk about lessons learnt
- provide an outlook of planned future activities in data collection and systems development

9 PGMasker: Human-in-the-Loop Models for Text Anonymization (Demo)

Carlos Durá Santonja, Mercedes García-Martínez, Carmen Grau Lacal, Amando Estela and Manuel Herranz

The development and expansion of new technologies poses an additional challenge, ensuring the privacy of their users. People need to protect their personal data due to new GDPR law. Unstructured documents are difficult and costly to process. Thus, finding sensitive information in a document requires a lot of effort due to the complexity of the language. Also, users need to anonymize specific information that general systems cannot identify. PGMasker is an automatic anonymization solution that combines machine learning, regular expressions and dictionaries guarantying privacy.

Users can select the tags (e.g. person, organization, location, bank account or email) to be anonymized. In addition, users can customize models and create their own regular expressions to detect specific entities such as websites and use dictionaries that can force words to be anonymized or not. Public administrations or private companies would require this system. A possible use case is a hospital, as they may want all their patients to be anonymized and on the other hand they may not want the doctors to be anonymized. Among all, PGMasker offers the option to review anonymized text by annotators what is called 'human-in-the-loop' in another platform called PECAT that allows to add tags. Using the reviewed text to fine tune models, PGmasker obtains more accurate anonymization.

10 Data Augmentation in Russian for Neural Machine Translation in the Language Industry

Nikita Teslenko Grygoryev, Mercedes García-Martínez, Amando Estela, Manuel Herranz and Francisco Casacuberta Nolla

The objective of this work is to create text examples in Russian from existing ones using Machine Learning techniques with the propose of generating new data which preserve the quality of the input text. Data from public domain have been used.
The system generates new texts using information from embeddings trained with a huge amount of data in neural language models (LMs). RuRoBERTa-large predicts a new word that is forced to be same gender and number as the original word. An acceptance threshold was established empirically to 0.07 during the evaluation of generated words.
The data augmentation methods (adverb, noun, adjective and a mix of them substitution techniques) search for an alternative word until the maximum number of words is reached for the generation of new sentences, which increases the size and quality of the in-domain corpus used to train Machine Translation (MT) models. We produce English side by MT provided by Pangeanic.
On the one hand, those models have been compared by performing quantitative analysis, comparing the different methods by means of automatic metrics used in MT (BLEU, TER, chrF, COMET, BEER and NIST). For English->Russian translation, mixture data augmentation method gives the best scores. On the other hand, they have been compared through a qualitative analysis, where a sample of translations generated by the trained MT models have been compared with the baseline by people who know both Russian and English.

11 Using Transformers to Build Scalable Automation Solutions

Kevin Kipfer, Samuel Schaffhauser, Alvin Pyngottu and Gero Gunkel

Large scale transformer models have become the de-facto standard in academia. However, until now few (non tech) companies have actually developed and globally scaled transformer-based data products, leading to a dearth of industry case studies. That said, we developed a new general purpose transformer based document extraction solution and then scaled it to over 10 markets globally, enabling the automated processing of millions of long, highly complex and diverse input documents (emails, pdfs, scans, tables, infografics, images etc.).

In this presentation the team will present the solution, outline the opportunities and challenges of scaling such models in the financial services industry, outlining key technology and business considerations to successfully deploy and scale them in an industry setting.

12 Leadership Training with Human-Computer Interaction and Natural Language Processing

Oscar William Lithgow Serrano, Giancarlo Corti, Luca Chiarabini, Fabio Rinaldi, Daniele Puccinelli and Andrea Laus

SkillGym is a leadership training software that presents in-role or prospective leaders with simulations of delicate instances of inter-personal interaction in corporate settings.

In the Innosuisse-funded project “Boosting the SkillGym Quality of Experience with Artificial Intelligence” (BOOST), our team has developed a system that leverages natural language processing to boost SkillGym’s Quality of Experience.

The system relies on a voice-based interface that enables the end user to speak to a character played by a professional actor. The character can only communicate based on a sequence of pre-recorded videos, but the user feels like she can speak freely to the character within the well-defined context of the simulation.

The key NLP challenge is to ensure a seamless simulation by choosing the most appropriate next video segment based on the user's utterance. We address this challenge with a hybrid combination of transfer learning and rule-based solutions.

The key human-computer interaction challenge is to ensure as lifelike an experience as possible, for instance by minimizing the latency of each exchange. We address this challenge through the application of software engineering best practices.

13 SwissText Scripting from UBS Business Solutions AG

Daniel Mazzolini and Vladimir Matic Lugo

Our Firm UBS is the preeminent universal bank in Switzerland. Drawing on our network of around 200 branches and 4,600 client advisors, complemented by modern digital banking services and customer service centers, we are able to reach approximately 80% of Swiss wealth.

In Conversational Banking, our vision is to bring natural language as a new way for interacting with digital clients and employees, enhancing user experience and increasing efficiency.
For our clients, we want to offer digital services via conversational interfaces, and for our employees we aim to provide a virtual assistant for knowledge workers, call agents and client advisors along most important business domains
Why Conversational Baking? In 2020 more than 3 million requests were raised to our support units - about 40% of these are trivial in nature and have an associated self-service option or info materials .
Common requests for queries are information, navigation, update, order. Common questions are: What is it? Where can I find it? How to do it? We will present how to leverage cloud cognitive services for Conversational Banking use cases.

14 Automated Bookkeeping to Save Valuable Time by Classifying Payments with ML in Production

Camilo Gordillo, Manuel Guth, Christian Reser and Andreas Meinel

Lexoffice is a cloud software for small-to-medium business (SMB) companies that simplifies and automates bookkeeping for more than 150k customers. In a nutshell, a core task in bookkeeping is assigning a bookkeeping category and tax rate to each payment of an SMB. This multi-task problem is very demanding due to the large number of classes and often requires domain expert knowledge.

Methods: Based on historic bookkeeping data, we trained a logistic regression model on various transaction-related features including textual ones such as the payment description. Overall, we predict the correct bookkeeping category with a weighted precision of 77 % on 15 different categories on a validation set. In addition, we use a confidence score describing the certainty of a single prediction. We ignore low confidence predictions using a category-specific threshold, which is calibrated to minimize the rejection rate.

Results: We successfully deployed a machine learning (ML) model to production and users now receive real-time bookkeeping category proposals for their payments. Within our cloud-based web app, users can easily interact with these proposals. Based on this direct closed-loop, we achieved an across customer averaged rejection rate of 19 %. To conclude, our ML model spares valuable time while reducing the SMBs need for tax domain knowledge.

Based on this success story, we aim to foster the usage of more complex models to mimic customer-specific needs for bookings.

15 Witty is a Writing Assistant for Inclusive Language

Elena Nazarenko

Every day we write some text. We try to write grammatically correct and politely where we have spelling and grammar checkers to support us. But many people struggle to ensure that their writing does not contain deterring words. Research shows that using non-inclusive language (especially gendered, ageism, and racist words) has a particularly strong impact on missing a large share of potential talents in the labor market. We developed a smart tool Witty that can assist in automatically detecting and suggesting alternatives to deterring language, enabling inclusive writing.

The core Witty algorithm is based on Natural Language Processing (NLP) advanced technologies. We combine a rules-based approach and modern transformer architecture. We created our own glossaries (German, English) with inclusive and non-inclusive words with the help of our highly trained language specialist and based on studies and research in that field. We use NLP pre-trained models for German and English correspondingly (SpaCy). The algorithm transforms the words in the text into lemmas, performs linguistic analysis, extracts the linguistic features from the text, and also does named entity recognition. Our approach can handle acronyms, and idiomatic multi-words expressions as well.

Currently, we are implementing the transformers (BERT, Hugging Face) to identify the right meaning of the words properly and classify the job-related text and perform the sentiment analysis of the text.

16 BeSTT - A Framework for Evaluation of STT Benchmarks

Elena Adamantidou, Daniel Aschauer, Mark Cieliebak, Katsiaryna Mlynchyk, Daniel Neururer, Alexandros Paramythis, Doris Paramythis and Yuliya Parfenova

Automatic Speech Recognition (ASR) has numerous applications, including chatbots, transcription of meetings, subtitling of TV shows, or automatic translation of conference presentations. For this reason, Speech-to-Text (STT) is a very active field of research, and tremendous progress has been made in the past years, in particular by using pretrained language models such as wav2vec and its derivatives. On the other hand, several ready-to-use solutions exist, from international corporations such as Google or IBM to specialized providers such as Trint or Speechmatics to open-source frameworks such as Fairseq or DeepSpeech.

But how do you find the "best" ASR engine? Grounded decisions in this respect typically require an in-depth comparison of the performance of ASR engines on various annotated corpora. In order to simplify this process, we have developed a framework that allows to easily run and evaluate benchmarks on arbitrary ASR engines.

In this presentation, we introduce the framework itself as well as insights from our research on extensive benchmark experiments on various ASR engines. Among other things, we answer the following questions: How well do ASR engines perform on different types of speech, e.g. spontaneous vs. read-aloud? Can you combine several engines to achieve better results? How can you distinguish automatically between minor errors (e.g. singular vs. plural) and semantically significant errors (e.g. "cat" instead of "car")

17 Integrating ML-based Classifiers into an Enterprise Search System (Demo)

Sandro Pedrazzini and Holger Keibel

HIBU is a proprietary software platform that we use to build customer solutions around enterprise search and multilingual text analysis. Its architecture provides two analysis pipelines: a first one embeds basic NLP steps, based on the detected document language and used to pre-elaborate the document’s content; a second one contains a sequence of high-level annotators, able to discover information in the document. Some examples are extracting entities from the text, such as persons, places and organizations, identifying paragraphs containing confidential information etc.

Both pipelines use the framework Apache UIMA to combine the annotators that are relevant for the target application. Each single one can be adapted and switched on and off by configuration. Moreover, the framework allows us to add new annotators based on the individual customer's needs.

In this context, we recently integrated some new ML-based annotators as part of an Innosuisse project carried out in collaboration with SUPSI and DSwiss ("EXTRA", presented separately, leveraging a fine-tuned version of the pre-trained BERT model and other ML technologies). These annotators allow us to provide scalable document classification, as well as customized information extraction, to be used by applications for further workflow-based functionalities.

In this demo we will show how we wrap the new BERT-based functionalities into the base platform to further enrich the final results.

18 Adapted Convolutional Architecture for Table Detection and Structure Recognition

Pratyush Singh, Moritz Platscher, Manuel Günther and Timothy Man

While tables remain a wide-spread format in different domains, lots of information remains inaccessible to automatic data extraction methods. This is due to the vast variety of layouts, occurrences of nested cells, and implied or missing separators, among others. Rule-based approaches have long dominated table data extraction, but their applicability is brittle and limited to specific domains. With the advent of various deep learning techniques, new methods for cross-domain table detection and extraction have emerged. While the nature of the issue is inherently tied to NLP, its solution often relies on multi-modal features, i.e., adding visual features, layout and reading order. Our contributions are: A new, asymmetric evaluation metric for table detection. A uniform data format for extensive cross-testing. Carefully designed model parameters adapting the neural network to the peculiarities of tables. We find that our modified metric aligns our qualitative findings and the quantitative outcomes and reduces the dependency of the F1-scores on the chosen IoU threshold. Similarly, we find that given the intrinsically different difficulties in detecting rows, columns, and cells compared to generic object detection, it is crucial to modify the model parameters accordingly. Based on this intuition, we have devised a fully convolutional model that achieves SOTA results on public datasets while requiring significantly less training data and performing more robustly across domains.

19 A New Platform for Automatic Skills Extraction and Validation of Profiles to Identify Talents

Natasa Sarafijanovic-Djukic, Lilian Roos, Priska Burkard and Beatrice Paoli

We present a joint project between the Laboratory for Web Science at the FFHS (Fernfachhochschule Schweiz) and a start-up company, Skills Finder AG. It is funded by InnoSuisse. The goal of the project is to build a platform for processing job application documents that automates the extraction of relevant information from a candidate's CV and the validation of this information with the candidate's references and certificates. Our processing pipeline uses the most recent advances in the field of both image document processing and Natural Language Processing (NLP).

The first step in the processing of any document is to extract the text from it by taking a proper reading order into consideration. As CVs have very diverse layouts, none of the existing tools could correctly extract the text from them. To detect these complex layouts, we train a CV Layout Model by using a Deep Layout Parser (layout-parser.github.io), a unified toolkit for deep learning-based document image analysis.

The next step is an NLP component for the information extraction. We use Named Entity Recognition (NER) with a pretrained BERT multi-language model, where we fine tune the model on our dataset with a custom labeling.

The last step in our processing pipeline is the information validation. We calculate a semantic similarity between word phrases by using output feature vectors from the BERT model and the mean pooling. Our model achieves more than 80% accuracy on the skills extraction.

20 Evaluation of Commercial Swiss German Speech-to-Text Systems

Yanick Schraner, Christian Scheller, Manfred Vogel and Michel Plüss

Transcribing Swiss German speech automatically into Standard German text is a challenging Speech-to-Text (STT) translation task. For the last two years, we at FHNW worked on the development of an end-to-end system to solve this task. In cooperation with ZHAW we also created a 35 hours test corpus which contains 7 x 5 hours of audios with transcripts of Swiss newspaper articles spoken in 7 Swiss German dialects (Basel, Bern, Graubünden, Innerschweiz, Ostschweiz, Wallis, and Zürich). Thus, for each region, we collected a total of 3600 spoken sentences from at least 10 different speakers.

We use this test set to objectively quantify the quality of our STT system and compare it to two commercial STT services for Swiss German to Standard German.

We evaluated all three STT systems using our test set and we present a fair comparison using our carefully designed test corpus. We discuss weaknesses and strengths of the three models in terms of the different dialects and other aspects.

21 Ukraine-Russia - First Insights into Recent Twitter Posts about this Conflict

Zhivar Sourati Hassan Zadeh, Pius von Däniken and Mark Cieliebak

After the Russia and Ukraine conflict escalated, people started to chat about different aspects of this conflict on Twitter. As one of the platforms that are hosting lots of opinions from different people, Twitter can, to some degree, reflect peoples' points of view on this matter. On this ground, we gather a rich set of tweets containing over 12M posts which we analyze from various angles using state-of-the-art NLP and network analysis methods. On top of that, we investigate the presence of bots in people's interaction network. This gives us structured insights into what topics are relevant in people's minds, how these topics are connected and evolve over time, and how they correlate with events during the conflict. We use Dirichlet Multinomial Mixture with Gibbs Sampling (GSDMM) for topic modeling and various emotion lexicons for our study. Our results showed that bots express significantly different levels of emotions towards topics compared to normal users and can be influential in the network. Also, we observed that although there are topics prevalent in all communities, there are topics more specific to certain communities as well that can relate to their chief characteristics.

22 Using Dependency Parsing to extract Structured Data from Expert Texts

Philipp Thomann

Libraries of technical expert information have been written in free text. Such texts are usually authored by experts in a semi-formal style. How can this valuable information be extracted into structured and useful representations?

The human touch of these texts renders rule based approaches useless. Annotating enough samples for an ML model might be too expensive. We show an approach essentially combining both worlds by using an off-the-shelf dependency parser together with tree-based rules of extraction.

Any syntax tree - even if it happens to be incorrect as in the below illustration - with its phrases and their relations do have the right format for rule-based extraction in this context (illustration only visible in PDF version). Rules detect syntactic relations and adpositions and build the structured output. Enumerations can easily be determined and extracted as lists.

We demonstrate the effectiveness of this methodology on a catalog from CRB, which has been standardizing processes in the Swiss construction industry for over 60 years. Expert authors write books of detailed definitions on walls, tunnels, canalisation etc. and specify which types of concrete can be used, for which purpose the wall is meant, what dimensions they should have. This is then used by contractors to formulate clear offers. For the next generation of standards a structured representation of the legacy catalogs is extracted by the above methodology.

23 Legal Prompt Engineering: A Case-Study on Multilingual Legal Judgement Prediction

Dietrich Trautmann, Alina Petrova and Frank Schilder
In supervised classification tasks, a machine learning model is provided with an input, and after the training phase, it outputs one or more labels from a fixed set of classes. Recent developments of large pre-trained language models (LLMs), such as BERT, T5 and GPT-3, gave rise to a novel approach to such tasks, namely prompting.
In prompting, there is usually no further training required (although fine-tuning is still an option), and instead, the input to the model is extended with an additional text specific to the task – a prompt. Prompts can contain questions about the current sample, examples of input-output pairs or task descriptions. Using prompts as clues, a LLM can infer from its implicit knowledge the intended outputs in a zero-shot fashion.

Legal prompt engineering is the process of creating, evaluating, and recommending prompts for legal NLP tasks. It would enable legal experts to perform legal NLP tasks, such as annotation or search, by simply querying large LLMs in natural language.
In this presentation, we investigate prompt engineering for the task of legal judgement prediction (LJP). We use data from the Swiss Federal Supreme Court and the European Court of Human Rights, and we compare various prompts for LJP using multilingual LLMs (mGPT, GPT-J-6B, etc.) in a zero-shot manner. We find that our approaches achieve promising results, but the long documents in the legal domain are still a challenge compared to single sentence inputs.

24 Hateful Social Media Users - Can we Predict their Behavior?

Pius von Däniken, Zhivar Sourati and Don Tuggener
Hate speech and other types of harmful online content have received a lot of attention in recent years. Many researchers have contributed to the field and literature in different ways, which can fall under the following categories. Some have tried to make a connection between hateful content creation and the events that are happening in society, often focused on a specific domain. Others kept their center of attention on each post regardless of the user activity timeline and context.Lastly, most similar to us, some have attempted to classify or identify posts in terms of their hatefulness taking into account the users' history of activity. However, to our knowledge, there is no study of the transition that underlies the potential for creating such content nor forecasting this transition. In the context of the DOSSMA research project at ZHAW CAI we are currently working on methods to forecast hateful behavior of users on Twitter. We are investigating whether this is possible at all, and if so, how far into the future we can make useful predictions. We also look into which features are most predictive, including content based and network based features.
We will present our preliminary results and discuss the many challenges that arise when tackling this task.

25 Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling

Ada Wan

We perform systematically and fairly controlled experiments with the 6-layer Transformer to investigate the hardness in conditional-language-modeling languages which have been traditionally considered morphologically rich and poor. We evaluate through statistical comparisons across 30 possible language directions from the 6 languages of the United Nations Parallel Corpus across 5 data sizes on 3 representation levels --- character, byte, and word. Results show that performance is relative to the representation granularity of each of the languages, not to the language as a whole. On the character and byte levels, we are able to eliminate statistically significant performance disparity, hence demonstrating that a language cannot be intrinsically hard. The disparity that mirrors the morphological complexity hierarchy is shown to be a byproduct of word segmentation. Evidence from data statistics, along with the fact that word segmentation is qualitatively indeterminate, renders a decades-long debate on morphological complexity irrelevant in the context of computing. The intent of our work is to help effect more objectivity and adequacy in evaluation as well as fairness and inclusivity in experimental setup in the area of language and computing. Multilinguality is real and relevant in computing not due to concepts such as morphology or "words", but rather standards such as character encoding --- something which has thus far been sorely overlooked in our discourse and curricula.

26 QnA Bot "Katie" -- Answering Questions in Communication Channels such as Slack, MS Teams or Mailing Lists (Demo)

Michael Wechner, Yannis Schmutz, Simon Sterchi and Erik Graf

Background:

Access to knowledge in professional contexts is a long-standing challenge whereas the following limitations prevail in today’s solutions:
- People often find it difficult to document their knowledge in Wiki form and such pages are in danger of getting outdated very quickly
- People also do not like to answer the same questions repeatedly, and messaging channels do not scale well beyond a certain number of users and messages

To overcome these limitations, we have devised an intelligent bot that participates in the collaborative communication platforms.

Solution:

The potential for knowledge sharing on collaborative communication platforms is demonstrated by their wide use. MS Teams topped 270 million monthly active users in Q4 2021, and Slack had more than 10 million users in 2021.
Our mission is to make these interactions and the vast amount of knowledge they represent more accessible. Our intelligent bot system named Katie, can be deployed into these platforms in order to analyze their communication patterns.

Katie acts as a smart participant and is capable of distinguishing between chit-chat, and different types of direct and indirect question. It can detect duplicated questions based on semantic approaches and is open to integrate external knowledge bases into its answers.

Demo:

We demonstrate the capability of Katie based on an integration into a public Slack channel.

27 The Future of Work Project: Early Projections on Results and Impact

Philipp Kuntschik and Albert Weichselbraun

The Swiss labor market is changing dramatically. Projections regarding technological advances and automation show that "60 percent of occupations have at least 30 percent of constituent work activities that could be automated" (Bick et al. 2020). Furthermore, due to effects of the Covid-Pandemic, the pace of change is even accelerated with millions of workers in advanced economies facing significant change in their occupations (Lund et al. 2021). Both employers and employees have to adapt to new, more flexible working structures to be ready for the labor market of the future, whose requirements are increasingly hard to predict.

Future of Work develops methods to decomposes current job profiles in their elementary tasks and to provide predictions on their future relevance in terms of automation, offshoreability and required qualifications. Future of Work therefore, i) enables employers to proactively analyze and adapt the tasks and skill sets of their employees based on predictions made and ii) allows workers to gain a better understanding of the value of their particular skill and highlights opportunities to efficiently increase their personal future readiness factor.

Future of Work describes an ongoing research project funded by Innosuisse. We will discuss the current project state and its future road map as well as technologies applied. We will present early results and give a glance on what we expect to achieve in the next year.

28 Automated Grading of Open-Ended Answers in Adaptive Tasks Using Semantic Analyses

Egon Werlen, Behnam Parsaeifard, Sukanya Nath and Ioan Sorin Comsa

Adaptive online tasks require immediate feedback even on open questions. To get this feedback, we relied on students' self-assessment, comparing their answers with sample answers. Our aim is to replace the self-assessment with an automated grading system using NLP. For this purpose, we proposed a deep learning model based on transformers and conducted analyses on a self-curated German language dataset of nearly 7000 answers. Having introduced the attention mechanism, transformers take into account the relation between the words of a sentence. Our model comprises of two transformer encoders for student response and sample answer. The encoders take as input the pre-trained word embeddings and output new representations. By applying a simple feed forward neural network followed by a softmax to the difference between encoders' outputs we predict the grade. Our approach results in an accuracy of 60% and outperforms the simple random classifier (25%). Different answer styles of students compared to sample answers, spelling mistakes, synonyms, abbreviations, and other irregularities are limiting the accuracy of our model. Therefore, we plan to correct the texts for spelling errors, to create a dictionary for synonyms and abbreviation lexicon of technical words and lemmatise all words. In addition, we might change the analysis strategy by taking keywords from the sample answers and searching for them in the student answers in order to grade the answers.