efl insights 02|2020

An efl – the Data Science Institute Publication

efl

the Data Science Institute

Digitalizing Asset Management – The Way Forward

Economic Value of Data

DBPal: A Novel Lightweight NL2SQL Training Pipeline

Why Open Innovation in B2B Needs a Push

DEUTSCHE BÖRSE GROUP

DZ BANK Gruppe

finanz informatik

AMERICAN EXPRESS

Deutsche Leasing

FACTSET

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

TECHNISCHE UNIVERSITÄT DARMSTADT

Impressum

Redaktion
Prof. Dr. Peter Gomber
Dr. Jascha-Alexander Koch

Herausgeber
Prof. Dr. Wolfgang König
Vorstandsvorsitzender
efl – the Data Science Institute e. V.

Prof. Dr. Peter Gomber
Stellvertretender Vorstandsvorsitzender
efl – the Data Science Institute e. V.

Kontakt
info@eflab.de
www.eflab.de

Gestaltung
Novensis Communication GmbH
Bad Homburg

2. Ausgabe, 2020
Auflage: 200 Stück (Print) / 2.000 Stück (E-Paper)

Printed in Germany
ISSN 1866-1238

Editorial

Digitalizing Asset Management – The Way Forward

Alexander Lichtenberg

Alexander Lichtenberg
Member of the Executive Board
Union Investment

My first months at Union Investment have been very challenging. Not only with regard to the company or the people I am working with, but mainly because of the Corona pandemic. It was not what I was expecting to get into, but it helped me to get to know the asset management industry very fast including the challenges we and every other asset manager have to face now.

I was and still am very impressed by the rapidness we switched to a telework mode, and merely 90% of our staff worked from home during shutdown time. For me, this showed how digitalized we are already in our company. However, there are many other challenges coming up:

(1) Find a way to offer clients our products without visiting the bank

Before the corona pandemic, we could already see a trend to more digitalization in the asset management industry. For example, Union Investment established its own FinTech VisualVest in 2017. It enables us to give new digital trends a chance without integrating it right from the beginning in the infrastructure of a big asset manager. VisualVest and our tailored solution MeinInvest for our retail banks, Volksbanken Raiffeisenbanken – the cooperative financial network, allow banks to offer our investment fund products via a robo advisor to their clients. Investors do not need to visit their bank in person to invest in investment funds. This can be easily done by the Internet and a robo advisor.

(2) Make sustainability in portfolio management measurable and investable

Digitalization is integrated in our sustainable investment processes. The Sustainable Investment Research Information System (SIRIS) is our proprietary digital research tool for sustainability in portfolio management. In-depth fundamental analysis, active portfolio management, and consideration of environmental, social, and governance (ESG) criteria throughout the investment process make it possible to systematically exploit market inefficiencies in order to achieve risk-adjusted outperformance for our clients. In the context of portfolio management, the analysis of ESG risks and opportunities is firmly embedded within our research work. Closer fundamental integration enables us to make better investment decisions and generate a positive impact on investment performance. The numbers of SIRIS are quite impressive: We cover 32,000 issuers, 5 asset classes, over 94,000 securities, 114 countries, and manage EUR 50.4 billion sustainable assets under management so far.

(3) Humanistic artificial intelligence will affect our business

Artificial intelligence (AI) can be used in various areas. It covers the area of documents and texts (e.g., chatbots, CV scanning, sentiment analysis, document processing), image and video (e.g., gesture/face/handwriting recognition, object detection), time series (e.g., forecasts, predictive maintenance, fraud detection, supply chain analytics), data analysis (e.g., pattern recognition, insight minings) and audio (e.g., voice detection, noise reduction, audio fingerprinting). Union Investment wants to become a leading player in the area of artificial intelligence for asset management. From our point of view artificial intelligence can develop our business in various ways and areas. But our overall goal at Union Investment stays the same with AI. We always have the best interests of our investors in mind and will use AI to improve our data and service level towards our banking partners and customers. But, with all the fancy and interesting things, you can do with AI, we shall never lose the focus and control over the machines and robotics. Therefore, Union Investment supports a humanistic AI approach. There always needs to be a person who teaches and controls the machines and stays responsible for the findings. Hence, our AI approach combines economic benefit and ethical principles.

Research Report

Economic Value of Data

FIRMS COLLECT A LARGE AMOUNT OF DATA BY ENGAGING HEAVILY IN THE COLLECTION AND STORAGE OF ONLINE USER ACTIVITY VIA VARIOUS USER TRACKING TECHNOLOGIES. RECENT POLICY INITIATIVES AIM AT RESTRICTING THIS PRACTICE TO PROTECT CONSUMER PRIVACY. WE STUDY EMPIRICALLY THE CONSEQUENCES OF SUCH RESTRICTIONS FOR ONLINE PUBLISHERS, SUCH AS NEWS WEBSITES, BECAUSE THEY STRONGLY RELY ON REVENUES THAT ARE GENERATED BASED ON USER DATA. WE FIND A PRICE DECREASE OF CA. 30% FOR ONLINE ADS WHEN NO DATA FROM USER TRACKING IS AVAILABLE. THE POTENTIAL REVENUE LOSS COULD BE MORE THAN EUR 14 BILLION IN THE EU AND MORE THAN USD 27 BILLION IN THE US.

Rene Laub

Klaus Miller

Bernd Skiera

Motivation

According to the Internet Advertising Bureau (IAB), in 2018, the online advertising industry generated revenues of more than USD 107 billion in the US and more than EUR 55 billion in the EU, with steady year-over-year growth of more than 10%.

Transactions between publishers and advertisers create the vast majority of these revenues. Publishers, like the Financial Times or Spiegel Online, make most of their money by selling ad space on their websites to advertisers, who use the ad space to display a specific ad to an individual user of the publisher website. With increasing data science capabilities, advertisers draw on a large amount of data to personalize the content of the ad to the interests of each individual user in real-time; thereby, they increase the relevance of the ad for each user. For that purpose, the online advertising industry collects and stores a record of a user's activity on the Internet via various tracking technologies – with third-party cookies as the most predominant form. Third-party cookies are text files that contain unstructured data about a user's browsing history and are shared among advertisers and publishers across websites. Assume a user is browsing the Internet visiting real estate websites, like immobilienscout24.de in Germany. A small piece of text would be saved in the third-party cookie. Advertisers can access this data and deduct that this user might be interested in a mortgage and target this user with a matching ad.

With the growing discussion on the protection of consumer privacy, the tracking of a user's browsing history is under fire. Policy makers have put forward regulation to restrict the collection, storage, and processing of user data as introduced with the General Data Protection Regulation (GDPR) in 2018 in the EU or the California Consumer Privacy Act (CCPA) in 2020 in the US. Web browsers like Mozilla's Firefox, Apple's Safari, and Google's Chrome have already disabled or plan to disable tracking technologies, like third-party cookies, by default. This development might have important economic consequences for publishers, which rely heavily on data-based online advertising income to finance their free editorial content.

If online user tracking technologies are no longer available, online advertisers (1) lose the ability to profile users and personalize ad content, and (2) are no longer able to measure the success of their online ads, e.g., by observing click-through rates. As a result, advertisers' willingness to pay (WTP) for displaying a specific ad to a specific user might decrease and, consequently, also prices for ads and, thus, advertising revenues of publishers might decrease.

While disabling user tracking technologies potentially declines advertisers' WTP and, thereby, leads to reduced ad prices, there are also arguments that point in a different direction. Disabling user tracking technologies might increase ad prices by increasing competition between advertisers. If user tracking is possible, data about a user's interests and preferences deduced from a user's browsing history allows advertisers to segment the large amount of online users and target only users that fit the profile of an advertiser's target audience. Given that not all advertisers are interested in the same users, user data narrows down user segments and thereby decreases competition for a specific user resulting in lower prices. Without user data, the effect might reverse. More advertisers could compete for the same users, leading to thicker markets, higher ad prices, and, thereby, higher publisher revenues (Levin & Milgrom, 2010).

So far, the potential effect of disabling user tracking technologies is unclear, theoretical predictions are mixed and only very little and conflicting empirical evidence exists (Johnson et al., 2020). We address this gap by empirically investigating a unique data set of millions of ad transactions to understand changes in ad prices. We, thereby, inform policy makers and industry participants about the monetary consequences of the actions restricting user tracking.

Description of Empirical Study

To assess the potential changes in publisher revenues, we estimate the difference in prices of an ad that are paid with and without user data. We, therefore, examine the prices of more than 42 million ad impressions sold via a large European ad exchange within a period of two weeks. These ad impressions account for approximately 5% of all ad impressions of the data provider during the observation period. Around 85 % of the ad impressions have associated user data available via a third-party cookie and around 15% of the ad impressions are without associated user data. The ads are shown to more than 1.3 million different online users and comprise 100 different online and mobile publishers, covering a broad variety of topics like cars, computer and technology, finance, games, health and lifestyle, or sports.

The average price, measured in cost per thousand (CPM), for an ad impression is EUR 0.63. With an average CPM of EUR 0.69, prices for ads with associated user data are higher than prices for ads without user data that have an average CPM of EUR 0.28. Thus, prices for ads with user information are EUR 0.41 (ca. 146%) higher than prices for ads without such information. Yet, this price difference cannot directly be interpreted as the increase that user data causes. The reason is that this difference does not account for other factors that impact differences between the prices, such as contextual data of the publisher, location of the user, content of the publisher website, or characteristics of the ad space (e.g., size or position).

We, therefore, consider these factors in a regression and also account for selectivity concerns. Selectivity concerns arise because the (un-)availability of user data is not random but a result of a user's deliberate choice of web browser or the installment of privacy management apps blocking user tracking. Therefore, users who do not allow for tracking might be systematically different to users who allow for tracking, e.g., in their preference for online advertising, and those (unobserved) differences might impact ad prices. If these unobserved differences influence the probability of having user information associated, then we will have an imbalance of these (also price influencing) factors between ads with and without user information. We, therefore, use augmented inverse probability weighting (AIPW). AIPW is a two-step procedure: We, first, estimate for each ad the propensity of having user information associated using a logistic regression. We, then, use in the second step the inverse of this propensity in a linear regression of ad prices on all observable ad price determinants. This weighting creates a balance in the unobserved difference between ads with and without user information.

Empirical Findings

Controlling for all observable ad impression price determinants, we estimate an average CPM price of EUR 0.64 for ad impressions with user data and EUR 0.44 for ad impressions without user data. As a result, user information yields prices that are EUR 0.20 (ca. 45%) higher. Stated differently, disabling user tracking could, therefore, lead to a reduction in ad impression prices of ca. 30%, as Figure 1 depicts. Given that the total share of ad impressions with user data accounts for around 85% of all ad impressions, the vast majority of ad transactions in the market could suffer from a severe price reduction. Assuming that the total advertising revenue numbers stated in the outset came from up to 85% of ad transactions with user data, the potential loss in the EU would be more than EUR 14 billion and in the US more than USD 27 billion.

We, further, investigate whether the potential price reductions differ between publishers. We observe the highest price reduction for publishers that provide content related to (1) cars and (2) computer and technology products. User data in these industries seems to be especially valuable. Publishers with content related to shopping and lifestyle products indicate the lowest price reduction. User data seems, therefore, to be especially valuable related to high-priced products (e.g., cars or computers). Advertisers could have a higher WTP for data of users interested in these products due to higher expected profits compared to lower-priced products in the shopping and lifestyle area. User preferences for high-priced products are usually also more specific, compared to broader user preferences in lower-priced products, which could make user data more valuable in high-priced product environments.

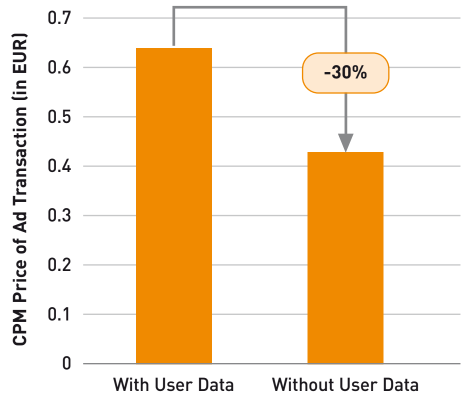


Figure 1: Price Differences of Ad Impressions

Conclusion

Our study demonstrates the high economic value of user data. We find that prices for ad impression drop by ca. 30% when no user data is available. Since many companies base their business models on advertising revenues, our prediction could guide managers of these companies when designing new pricing schemes or switching to other business models. For example, companies could offer users the possibility of not collecting user information and in exchange charge a subscription fee that compensates for losses in advertising revenue. Our research results inform policy makers about the potential economic losses when balancing user privacy interest and interests of companies in the future.

References

Levin, J; Milgrom, P.:
Online Advertising: Heterogeneity and Conflation in Market Design.
In: American Economic Review, 100 (2010) 2, pp. 603-607.

Johnson, G. A.; Shriver, S. K.; Du, S.:
Consumer Privacy Choice in Online Advertising: Who Opts out and at What Cost to Industry?
In: Marketing Science, 39 (2020) 1, pp. 33-51.

Research Report

DBPal: A Novel Lightweight NL2SQL Training Pipeline

NATURAL LANGUAGE (NL) IS A PROMISING ALTERNATIVE INTERFACE TO DATABASE MANAGEMENT SYSTEMS (DBMSs) BECAUSE IT ENABLES NON-TECHNICAL USERS TO FORMULATE COMPLEX QUESTIONS. RECENTLY, DEEP LEARNING HAS GAINED TRACTION FOR TRANSLATING NATURAL LANGUAGE TO SQL. HOWEVER, THE CORE PROBLEM WITH EXISTING DEEP LEARNING APPROACHES IS THAT THEY REQUIRE AN ENORMOUS AMOUNT OF MANUALLY CURATED TRAINING DATA IN ORDER TO PROVIDE ACCURATE TRANSLATIONS. WE PRESENT DBPAL THAT USES A NOVEL TRAINING PIPELINE TO LEARN NL2SQL INTERFACES WHICH SYNTHESIZES TRAINING DATA AND, THUS, DOES NOT RELY ON MANUALLY CURATED TRAINING DATA.

Benjamin Hättasch

Nadja Geisler

Carsten Binnig

Introduction

In order to effectively leverage their data, DBMS users are required to not only have prior knowledge about the database schema (e.g., table and column names, entity relationships) but also a working understanding of the syntax and semantics of SQL. Unfortunately, despite its expressiveness, SQL can often hinder non-technical users from exploring and making use of data stored in a database. These requirements set "a high barrier to entry" for data exploration and have, therefore, triggered new efforts to develop alternative interfaces that allow non-technical users to explore and interact with their data conveniently. For example, imagine a doctor wants to look at the age distribution of patients with the longest stays in a hospital. To answer this question, the doctor would either need to write a complex nested SQL query or work with an analyst to craft the query. Even with a visual exploration tool (e.g., Tableau, Vizdom), posing such a query is nontrivial, since it requires the user to perform multiple interactions with an understanding of the nested query semantics. Alternatively, with an NL interface, the query is as simple as stating: "What is the age distribution of patients who stayed longest in the hospital?"

Based on this observation, a number of "natural language interfaces to databases" tools (NLIDBs) have been proposed that aim to translate natural language to SQL (NL2SQL). The first category of solutions are rule-based systems, e.g., NaLIR (Li and Jagadish, 2014), which use fixed rules for performing translations. Although effective in specific instances, these approaches are brittle and do not generalize well without substantial additional effort to support new use cases. More recently, deep learning techniques have gained traction for NL2SQL, since similar ideas have achieved success in the related domain of machine translation. For example, generic sequence-to-sequence (seq2seq) (Zhong et al., 2017) models have been successfully used in practice for NL2SQL translation, and more advanced approaches like Syntax SQLNet (Yu et al., 2018), which augments deep learning models with a structured model that considers the syntax and semantics of SQL, have also been proposed.

However, a crucial problem with deep learning approaches is that they require an enormous amount of training data in order to build accurate models. The aforementioned approaches have largely ignored this problem and assumed the availability of large, manually-curated training datasets (e.g., using crowd-sourcing). In almost all cases, however, gathering and cleaning such data is a substantial undertaking that requires a significant amount of time, effort, and money.

Moreover, existing approaches for NL2SQL translation attempt to build models that generalize to new and unseen databases, yielding performance that is generally decent but does not perform as well as running new queries on the databases used for training. That is, the training data used to translate queries for one specific database, such as queries containing words and phrases pertaining to patients in a hospital, does not always allow the model to generalize to queries in other domains, such as databases of geographical locations or flights.

In order to address these fundamental limitations, in a recent SIGMOD 2020 paper (Weir et al., 2020), we proposed DBPal, a fully pluggable NL2SQL training pipeline that can be used with any existing NL2SQL deep learning model to improve translation accuracy. DBPal implements a novel training pipeline for NLIDBs that synthesizes its training data using the principle of weak supervision.

The basic idea is to leverage various heuristics and existing datasets to automatically generate large (and potentially noisy) training data instead of manually handcrafting training examples. In its basic form, only the database schema is required as input to generate a large collection of pairs of NL queries and their corresponding SQL statements that can be used to train any NL2SQL deep learning model.

In order to maximize our coverage across natural-linguistic variations, DBPal also uses additional input sources to automatically augment the training data through a variety of techniques. One such augmentation step, as an example, is an automatic paraphrasing process. The goal of these augmentation steps is to make the model robust to different linguistic variations of the same question (e.g., "What is the age distribution of patients who stayed longest in the hospital?" and "For patients with the longest hospital stay, what is the distribution of age?").

In the evaluation of our SIGMOD publication, we show that DBPal, which requires no manually crafted training data, can effectively improve the performance of a state-of-the-art deep learning model for NL2SQL translation. Our results demonstrate that an NLIDB can be effectively bootstrapped without requiring manual training data for each new database schema or target domain. Furthermore, if manually curated training data is available, such data can still be used to complement our proposed data generation pipeline.

Overview of DBPal

In the following, we first discuss the overall architecture of a NLIDB and, then, discuss DBPal, our proposed training pipeline based on weak supervision that synthesizes the training data from a given database schema.

Figure 1 shows an overview of the architecture of our fully functional prototype NLIDB, which consists of multiple components, including a user-interface that allows users to pose NL questions that are automatically translated into SQL. The results from the user's NL query are then returned to the user in an easy-to-read tabular visualization.

At the core of our prototype is a neural translator, which is trained by DBPal's pipeline, that translates incoming NL queries coming from a user into SQL queries. Importantly, our fully pluggable training pipeline is agnostic to the actual translation model; that is, DBPal is designed to improve the accuracy of existing NL2SQL deep learning models (e.g., SyntaxSQLNet) by generating training data for a given database schema.

Training Phase. During the training phase, DBPal's training pipeline provides existing NL2 SQL deep learning models with large corpora of synthesized training data (Figure 2). This training pipeline consists of three steps to synthesize the training data: (1) generator, (2) augmentation, and (3) lemmatizer. Once training data is synthesized by DBPal's pipeline, it can then be used (potentially together with existing manually curated training data) to train existing neural translation models that can be plugged into the training pipeline.

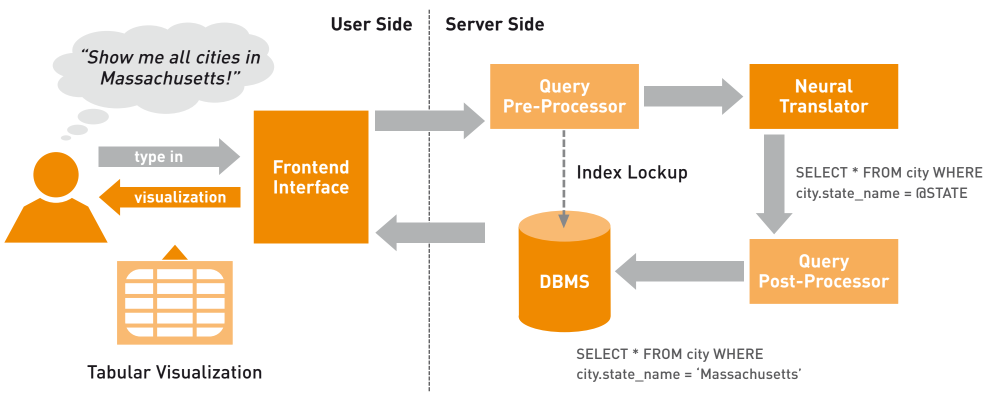
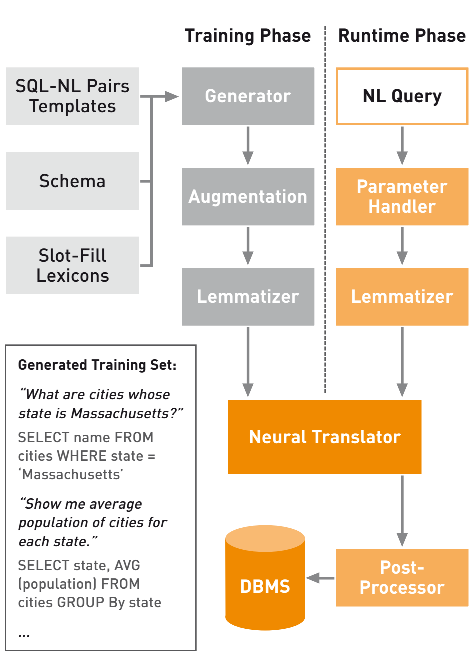


Figure 1: Overview of DBPal's Architecture

Runtime Phase. The runtime phase can leverage a model (neural translator) that was trained by DBPal, as shown on the right-hand side of Figure 2. The parameter handler is responsible for replacing the constants in the input NL query with placeholders to make the translation model independent from the actual database and help to avoid retraining the model if the underlying database is updated. For example, for the input query shown in Figure 2 (i.e., "What are cities whose state is Massachusetts?"), the parameter handler replaces "Massachusetts" with the appropriate schema element using the placeholder @STATE. The lemmatizer then combines different variants of the same word to a single root. For example, the words "is", "are", and "am" are all mapped to the root word "be". Then, the neural translator works on these anonymized NL input queries and creates output SQL queries, which also contain placeholders. In the example shown in Figure 2, the output of the neural translator is: SELECT name FROM cities WHERE state = @STATE. Finally, the post-processor replaces the placeholders with the actual constants such that the SQL query can be executed.

DBPal's Training Pipeline

The basic flow of the training pipeline is shown on the left-hand side of Figure 2. In the following, we describe the training pipeline and focus in particular on the data generation framework.



Generated Training Set:

"What are cities whose state is Massachusetts?"

SELECT name FROM cities WHERE state = 'Massachusetts'

"Show me average population of cities for each state."

SELECT state, AVG (population) FROM cities GROUP By state

...

Figure 2: Training and Runtime Phase

Generator. In the first step, the generator uses the database schema along with a set of seed templates that describe typical NL-SQL pairs to generate an initial training set. In the second step, augmentation, the training data generation pipeline then automatically adds to the initial training set of NL-SQL pairs by leveraging existing general-purpose data sources and models to linguistically modify the NL part of each pair.

The main idea is that each seed template covers a typical class of SQL queries (e.g., a SELECT-

FROM-WHERE query with a simple predicate). Composing the seed templates is only a minimal, one-time overhead, and all templates are independent of the target database, i.e., they can be reused for other schemas. Furthermore, in DBPal, we assume that the database schema provides human-understandable table and attribute names, but the user can optionally annotate the schema to provide more readable names if required. Deriving readable schema names automatically is an orthogonal issue.

The schema information is, then, used to instantiate these templates using table and attribute names. Additionally, manually predefined dictionaries (e.g., to cover synonyms) can be used to instantiate simple variations of NL words and phrases (e.g., "Show me" and "What is" for the SELECT clause). Currently, DBPal contains approximately 100 seed templates. A typical training set that can be generated from these templates contains around one million NL-SQL pairs for a simple, single-table database schema and around two to three million for more complicated schemas.

Augmentation. A core aspect of our pipeline is the augmentation step that automatically expands the training data produced by our generator in order to offer more accurate and linguistically robust translations. During augmentation, the training data generation pipeline automatically adds new NL-SQL pairs by leveraging existing general-purpose data sources and models to linguistically vary the NL part of each pair. The goal of the augmentation phase is to cover a wide spectrum of linguistic variations for the same SQL query, which represent different versions of how users might phrase the query in NL. This augmentation is the key to make the translation model robust and allows DBPal to provide better query understanding capabilities than existing standalone approaches.

Lemmatization. Finally, in the last step of the data generation procedure, the resulting NL-SQL pairs are lemmatized to normalize the representation of individual words. During this process, different forms of the same word are mapped to the word's root in order to simplify the analysis (e.g., "cars" and "car's" are replaced with "car"). The same lemmatization is applied at runtime during the aforementioned pre-processing step.

Conclusions and Future Work

We presented DBPal, a fully pluggable natural language to SQL (NL2SQL) training pipeline that generates synthetic training data to improve both the accuracy and robustness to linguistic variation of existing deep learning models. In combination with our presented data augmentation techniques, which help improve the translational robustness of the underlying models, DBPal is able to improve the accuracy of state-of-the-art deep learning models by up to almost 40%.

Longer term, we believe that an exciting opportunity exists to expand DBPal's techniques to tackle broader data science use cases, ultimately allowing domain experts to interactively explore large datasets using only natural language (Rogers et al., 2017). In contrast to the typical notion of one-shot SQL queries currently taken by DBPal, data science is an iterative, session-driven process, where a user repeatedly modifies a query or machine learning model after examining intermediate results until finally arriving at some desired insight, which will, therefore, necessitate a more conversational interface. These extensions would require the development of new techniques for providing progressive results (Turkay et al., 2018) by extending past work on traditional SQL-style queries and machine learning models.

Finally, we believe there are also interesting opportunities related to different data models, e.g., time series (Eichmann et al., 2017) and new user interfaces, e.g., query-by-voice (Lyons et al., 2016).

References

Eichmann, P.; Crotty, A.; Galakatos, A.; Zgraggen, E.:
Discrete Time Specifications in Temporal Queries.
In: CHI Extended Abstracts; Denver (CO), US, 2017.

Li, F.; Jagadish, H.:
NaLIR: An Interactive Natural Language Interface for Querying Relational Databases.
In: Proceedings of the ACM SIGMOD Conference; Snowbird (UT), US, 2014.

Lyons, G.; Tran, V.; Binnig, C.; Çetintemel, U.; Kraska, T.:
Making the Case for Query-by-Voice with EchoQuery.
In: Proceedings of the ACM SIGMOD Conference; San Francisco (CA), US, 2016.

Rogers, J. L. J.; Potti, N.; Patel, J.:
Ava: From Data to Insights Through Conversations.
In: Proceedings of the 8th Biennial Conference on Innovative Data Systems Research (CIDR); Chaminade (CA), US, 2017.

Turkay, C.; Pezzotti, N.; Binnig, C.; Strobelt, H.; Hammer, B.; Keim, D.; Fekete, D.; Palpanas, T.; Wang, Y.; Rusu, F.:
Progressive Data Science: Potential and Challenges.
In: Working Paper, 2018.

Weir, N.; Utama, P.; Galakatos, A.; Crotty, A.; Ilkhechi, A.; Ramaswamy, S.; Bhushan, R.; Geisler, N.; Hättasch, B.; Eger, S.; Cetintemel, U.; Binnig, C.:
DBPal: A Fully Pluggable NL2SQL Training Pipeline.
In: Proceedings of the ACM SIGMOD Conference; Portland (OR), US, 2020.

Yu, T.; Yasunaga, M.; Yang, K.; Zhang, R.; Wang, D.; Li, Z.; Radev, D.:
SyntaxSQLNet: Syntax Tree Networks for Complex and Cross-Domain Text-to-SQL Task.
In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); Brussels, Belgium, 2018.

Zhong, V.; Xiong, C.; Socher, R.:
Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning.
In: Working Paper, 2017.

Insideview

Why Open Innovation in B2B Needs a Push

INTERVIEW WITH SVEN SIERING

Sven Siering
Head of Digital Innovation Unit
Deutsche Leasing AG

Open innovation means opening up the innovation process and joining forces with the outside world, involving experts from other companies, start-ups, and universities. Especially in the B2B sector, there is still a lot of catching up to do taking this step. Often enough, prevailing competition bars the way to progress. Which advantages do you see in taking a step towards open innovation?

Opening up enables companies to challenge problems faster, validate new ideas, search for, and find suitable partners as well as test solutions. Including various perspectives delivers results that are more valid, adds new insights, and taps additional potentials. Plus, you create a trustful network. This builds the crucial base for implementing jointly explored solutions at a later stage.

Why do you see great potential here, especially in the B2B sector?

Today's value chains and networks evolved and manifested themselves through the years. Thus, it is often difficult to break new ground. Due to the ever faster changing framework conditions caused by new technologies and trends and the resulting changes in demand, established companies in particular are under considerable strain to innovate. Newcomers often find it easier because they are consistently committed to innovation and take unconventional paths. Open innovation helps established companies, especially in the B2B sector, to understand their customers' requirements and to react to changes at an earlier stage. At the same time, it also makes it easy to compensate for missing competences, as larger, more diverse teams draw on more resources, of course. To support this step, we are currently creating a cross-industry open innovation platform, B2Innovate.

What kind of solution do you provide and who can participate?

With B2Innovate (launch planned in August 2020), we address any companies regardless of their size in the context of Industry 4.0. No matter if it is, for example, a mechanical engineering company, a telecommunications company, a construction company, or a supplier. On the platform, we offer the chance to collaborate on problem areas, to develop suitable problem solutions, or to find suitable startups to tackle their challenges. In addition, we will organize accompanying events and personal meetings to strengthen commonalities within this project.

Why are universities so interesting?

Research and development are universities' inherent competences. There, researchers develop new technologies, research human behavior, and make new connections from all this. Additionally, the university environment started to encourage directly founding startups. Thus, there is a valuable potential for companies here and my suggestion is to tap that. Our platform intends to support in this respect. With the Goethe University Frankfurt and the Technical University of Darmstadt, two partners are involved right from the start. We are already in touch with other universities.

Where do you see challenges ahead for B2Innovate?

Open innovation in the B2B sector is not a matter of course. There have been attempts before trying to establish such a platform. At Deutsche Leasing, we deliberately chose a neutral, unbranded approach. The goal is to find a common path that meets everyone's requirement. Therefore, the platform is also free of charge for users. The currency for participation is the feedback the users provide and their active involvement.

What do you wish for the future?

I would like to see a big community on B2-Innovate, especially in the midst and aftermath of the current challenges so we can shape our future together by open innovation.

Thank you for this interesting conversation.

Infopool

News

FIRM Dissertation Price

Dr. Benjamin Clapham was awarded the Research Prize 2020 of the Frankfurt Institute for Risk Management and Regulation for his dissertation "Integrity and Efficiency of Electronic Securities Markets: Fraud Detection, Safeguards, and the Role of High-Frequency Trading". The dissertation was supervised by Prof. Gomber. Congratulations!

efl Jour Fixes via Online Video Conference Due to COVID-19

We have changed the format of our monthly efl Jour Fixes. If you are interested in taking part in our digital Jour Fixes, please e-mail to: *info@eflab.de*. All dates are on our website.

efl Spring Conference 2020

Thank you for your participation in the efl Spring Conference 2020 "The value of data in business, research and society", on February 18th, 2020, organized by Prof. Hinz and his team. Over 350 persons from 68 organizations gave their time and resources to attend and to contribute. Participants and speakers made the efl Spring Conference 2020 a success. We would like to give special thanks to the speakers. Hopefully, all participants enjoyed both the scientific part and the social "get together" after the conference and used the opportunity to extend existing networks. We hope to see you at the efl Spring Conference 2021!

New Colleague

In April 2020, Tino Cestonaro joined the Chair of Prof. Gomber as a doctoral student. He holds a Master's Degree in Business Administration from Goethe University with a concentration in Finance. During his doctoral studies, Tino will focus on big data applications in Finance.

Haushaltskrisenbarometer – A Bi-Weekly Updated View on the Economic Situation of German Households During the Corona Crisis

Prof. Hackethal has been operating the "Haushaltskrisenbarometer" together with Prof. Inderst since March 2020. Core element is the evaluation of survey responses of a large sample of households in the Nielsen Consumer Panel. It provides a reliable and timely picture of the economic situation, (consumer) behavior, and expectations of the entire population. Additionally, it provides insights into the actual purchasing behavior of all households. FAZ, WELT, Spiegel, and Wirtschaftswoche, among others, have reported on the results.

Successful Funding

The team of Prof. Hinz succeeded in raising a seeding fund with the proposal "From Machine Learning to Machine Teaching (ML2MT) – Making Machines AND Humans Smarter" in the call "Artificial Intelligence and the Society of the Future" by the Volkswagen Stiftung. The seeding of about EUR 150,000 should be used to write a full proposal in 2020.

Outstanding Reviewer Award 2019

Prof. Dr. Gomber received the Outstanding Reviewer Award 2019 from "Electronic Markets – The International Journal on Networked Business". The renowned journal recognizes commitment and reliability of reviewers as well as outstanding quality of their reviews.

Selected efl Publications

Dimitrios, K.; Meyer, S.; Uhr, C.:
Google Search Volume and Individual Investor Trading.
Forthcoming in: Journal of Financial Markets, 49 (2020).

Frank, M.:
Sharing Information Security Failure: The Role of Social Context and Social Environment.
In: Proceedings of the 23rd Pacific Asia Conference on Information Systems (PACIS); Dubai, UAE, 2020.

Hättasch, B.; Meyer, C. M.; Binnig, C.:
Interactive Summarization of Large Document Collections.
In: Workshop on Human-In-the-Loop Data Analytics (HILDA); Amsterdam, Netherlands, 2019.

Hilprecht, B.; Schmidt, A.; Kulessa, M.; Molina, A.; Kersting, K.; Binnig, C.:
DeepDB: Learn from Data, not from Queries!
In: Proceedings of the 48th International Conference on Very Large Data Bases; Tokyo, Japan, 2020.

Kimmerl, J.:
When Moms Go Online – Exploring Motives Determining Mothers' Participation in Maternal Online Communities.
In: Proceedings of the 28th European Conference on Information Systems (ECIS); Marrakesh, Morocco, 2020.

Kimmerl, J.:
Understanding Users' Perception on the Adoption of Stablecoins – The Libra Case.
In: Proceedings of the 23rd Pacific Asia Conference on Information Systems (PACIS); Dubai, UAE, 2020.

Lausen, J.; Clapham, B.; Siering, M.; Gomber, P.:
Who Is the Next "Wolf of Wall Street"? Detection of Financial Intermediary Misconduct.
Forthcoming in: Journal of the Association for Information Systems.

Shang, Z.; Zgraggen, E.; Buratti, B.; Kossmann, F.; Eichmann, P.; Chung, Y.; Binnig, C.; Upfal, E.; Kraska, T.:
Democratizing Data Science through Interactive Curation of ML Pipelines.
In: Proceedings of the SIGMOD Conference; Amsterdam, Netherlands, 2019.

Teso, S.; Hinz, O.:
Challenges in Interactive Machine Learning – Toward Combining Learning, Teaching, and Understanding.
Forthcoming in: German Journal of Artificial Intelligence.

For a comprehensive list of all efl publications see *http://www.eflab.de/publications*

Infopool

RESEARCH PAPER: ALGORITHMIC BIAS? AN EMPIRICAL STUDY OF APPARENT GENDER-BASED DISCRIMINATION IN THE DISPLAY OF STEM CAREER ADS

Using a field test on Facebook across 191 countries, the authors provide empirical evidence that an ad serving algorithm specifically instructed to be gender-neutral can lead to a discriminatory outcome where STEM (science, technology, engineering, and math) career ads are 20% more likely to be shown to men. Notably, the algorithmic discrimination does not reflect learned patterns from consumer behavior or country-specific preferences. Instead, discrimination is the consequence of a market price externality caused by the bidding for the ad placement. By trying to allocate the ad cost-effectively, the algorithm starts to discriminate because advertisers from other industries prize young women's attention higher than the attention of men. Extending to other platforms, these findings seem to represent an idiosyncrasy of the online advertising ecosystem.

Lambrecht, A.; Tucker, C.
In: Management Science, 65 (2019) 7, pp. 2966-2981.

RESEARCH PAPER: NORTHSTAR: AN INTERACTIVE DATA SCIENCE SYSTEM

In order to democratize data science, we need to fundamentally rethink the current analytics stack, from the user interface to the "guts". Most importantly, enabling a broader range of users to unfold the potential of (their) data requires a change in the interface and the "protection" we offer them. On the one hand, visual interfaces for data science have to be intuitive, easy, and interactive to reach users without a strong background in computer science or statistics. On the other hand, we need to protect users from making false discoveries. Furthermore, it requires that technically involved (and often boring) tasks have to be automatically done by the system so that the user can focus on contributing their domain expertise to the problem. In this paper, the author presents Northstar, the Interactive Data Science System, which was developed to explore designs that make advanced analytics and model building more accessible.

Kraska, T.
In: Proceedings of the VLDB Endowment,11 (2018) 12, pp. 2150-2164.

efl insights

The efl publishes the insights in the form of a periodic newsletter which appears two times a year. Besides a number of printed copies, the efl insights is distributed digitally via E-mail for reasons of saving natural resources. The main purpose of the efl insights is to provide latest efl research results to our audience. Therefore, the main part is the description of two research results on a managerial level – complemented by an editorial, an interview, and some short news.

For receiving our efl insights regularly via E-mail, please subscribe on our homepage www.eflab.de (> NEWS > SIGN UP EFL INSIGHTS) as we need your E-mail address for sending the efl insights to you. Alternatively, you can mail your business card with the note "efl insights" to the subsequent postal address or send us an E-mail.

Prof. Dr. Peter Gomber
Vice Chairman of the efl – the Data Science Institute
Goethe University Frankfurt
Theodor-W.-Adorno-Platz 4
D-60629 Frankfurt am Main

insights@eflab.de

Further information about the efl is available at www.eflab.de.

The efl – the Data Science Institute is a proud member of the House of Finance of Goethe University, Frankfurt.
For more information about the House of Finance, please visit www.hof.uni-frankfurt.de.

THE EFL – THE DATA SCIENCE INSTITUTE IS AN INDUSTRY-ACADEMIC RESEARCH PARTNERSHIP BETWEEN FRANKFURT AND DARMSTADT UNIVERSITIES AND PARTNERS DEUTSCHE BÖRSE GROUP, DZ BANK GROUP, FINANZ INFORMATIK, AMERICAN EXPRESS, DEUTSCHE LEASING, AND FACTSET LOCATED AT THE HOUSE OF FINANCE, GOETHE UNIVERSITY, FRANKFURT.

For further information please contact:

Prof. Dr. Peter Gomber
Vice Chairman of the
efl – the Data Science Institute
Goethe University Frankfurt
Theodor-W.-Adorno-Platz 4
D-60629 Frankfurt am Main

Phone +49 (0)69 / 798 - 346 82
E-mail gomber@wiwi.uni-frankfurt.de

Press contact
Phone +49 (0)69 / 798 - 346 82
E-mail presse@eflab.de

or visit our website
http://www.eflab.de