Report on the first funding phase of the research grant

entitled:

Complex Community-Based Question Answering on Heterogeneous Data for Educational Information (QA-EduInf)

Automatisches Beantworten komplexer benutzergenerierter Fragen auf heterogenen Daten für die Bildungsinformation (QA-EduInf)

5th June 2019

Contents

1 General Information

DFG-Geschäftszeichen: GU 798/18-1 and RI 803/12-1

1.1 Applicant

Prof. Dr. Iryna Gurevych, Department of Computer Science, Technische Universität Darmstadt

1.2 Topic

Complex Community-Based Question Answering on Heterogeneous Data for Educational Information
Automatisches Beantworten komplexer benutzergenerierter Fragen auf heterogenen Daten für die Bildungsinformation

1.3 Reporting and Funding Period

December 1, 2014 – May 31, 2017 (extended use of funds: September 30, 2019)

1.4 Project Publications

- Y. Gao, C. M. Meyer, M. Mesgar, and I. Gurevych. Good rewards are all you need: Reward learning for efficient reinforcement learning in document summarisation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, page (to appear), Macao, China, Aug 2019. Association for Computational Linguistics
- N. S. Moosavi, L. Born, M. Poesio, and M. Strube. Handling domain shift in coreference evaluation by using automatically extracted minimum spans. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, page (to appear), Florence, Italy, Jul 2019. Association for Computational Linguistics
- S. Eger, G. G. Şahin, A. Rücklé, J.-U. Lee, C. Schulz, M. Mesgar, K. Swarnkar, E. Simpson, and I. Gurevych. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 1634–1647, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics
- A. Rücklé, N. S. Moosavi, and I. Gurevych. Coala: A neural coverage-based approach for long answer selection with small data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, page (to appear), Honolulu, USA, Jan. 2019. Association for the Advancement of Artificial Intelligence
- A. Rücklé, K. Swarnkar, and I. Gurevych. Improved cross-lingual question retrieval for community question answering. In *Proceedings of the 2019 World Wide Web Conference (WWW-19)*, pages 3179–3186, San Francisco, USA, May 2019. Association for Computing Machinery
- Y. Gao, C. M. Meyer, and I. Gurevych. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4120–4130, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics

- D. Sorokin and I. Gurevych. Interactive instance-based evaluation of knowledge base question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 114–119, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics
- S. Eger, A. Rücklé, and I. Gurevych. Pd3: Better low-resource cross-lingual transfer by combining direct transfer and annotation projection. In *5th Workshop on Argument Mining, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 131–143, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics
- I. Kuznetsov and I. Gurevych. Corpus-driven thematic hierarchy induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, pages 54–64, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics
- I. Kuznetsov and I. Gurevych. From text to lexicon: Bridging the gap between word embeddings and lexical resources. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 233–244, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics
- D. Sorokin and I. Gurevych. Modeling semantics with gated graph neural networks for knowledge base question answering. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 3306–3317, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics
- D. Sorokin and I. Gurevych. Mixing context granularities for improved entity linking on question answering data across entity categories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, pages 65–75, New Orleans, Louisiana, June 2018. Association for Computational Linguistics
- S. Momtazi and I. Gurevych. Corrigendum to unsupervised latent dirichlet allocation for supervised question classification. *Information Processing & Management*, 54(3):380–393, 2018
- D. Sorokin and I. Gurevych. End-to-end representation learning for question answering with weak supervision. In *ESWC 2017 Semantic Web Challenges*, Portoroz, Slovenia, Oct. 2017
- A. Rücklé and I. Gurevych. Representation learning for answer selection with lstm-based importance weighting. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, Montpellier, France, Sept. 2017. Association for Computational Linguistics
- D. Sorokin and I. Gurevych. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics
- A. Rücklé and I. Gurevych. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations (ACL 2017)*, pages 19–24, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics

2 Work Report and Findings

2.1 Motivation and Research Goals of the Project

Over the past years, the amount of data in community-based question-and-answer (cQA) platforms on the web has been steadily growing. Such platforms contain explanations, descriptions and advice for a wide range of different topics (among other types of information). This is especially useful in the age of life-long learning where it becomes more and more common for people from different domains to search for information to educate themselves—e.g., to gather knowledge on cultural topics, to learn new programming languages, to get insights on academic procedures, and many more. However, accessing the relevant information across heterogeneous cQA platforms can be costly because users need to manually assess a large number of retrieved search results. Automatically analyzing relevant information from cQA platforms will help the users to access the required pieces of information with minimal effort.
In particular, the first funding phase of QA-EduInf had the following goals reflected in the defined work packages:

1. Research on, evaluation, and analysis of robust frame-semantic parsing methods in the context of cQA, including the German data.
2. Classification, retrieval, and ranking of complex questions and answer passages in cQA, with a focus on low-resource scenarios.
3. Integration of the novel components in an end-to-end cQA system and and demonstrating its capabilities in a working prototype.

2.2 Status Report and Findings

With respect to all three goals, the project's team has provided multiple contributions that were published at leading conferences in the field. During the reporting period, the research field of natural language processing was transformed substantially through novel technological advances—in particular deep neural networks—which achieved unprecedented improvements over the previous state of the art. Hence, over the course of this project, we utilized, adapted, and advanced these novel methods, favoring state-of-the-art neural networks over the planned feature-based approaches whenever appropriate.
In the following, we describe our contributions for each QA-EduInf work package in detail.

WP1: Corpus acquisition and analysis and classification of user questions. We collected and analyzed data from different German and English cQA platforms. From our project partner GuteFrage.net[1] we obtained more than 10 million German user-generated questions and their answers. From publicly available sources—seven different StackExchange sites[2]—we further collected over 10 million English questions and their answers. For our datasets, we performed basic preprocessing such as tokenization, sentence splitting, POS tagging, and syntactic parsing. The collected and processed data built the foundation for our research in the further work packages, and in particular, we extensively made use of the datasets to creation of our working prototype.

[1] https://www.gutefrage.net

[2] https://stackexchange.com/sites, the sites we used were stackoverflow, askubuntu, apple, aviation, cooking, travel, and academia.

Based on our collected data, we developed a classification approach for German user-generated questions. We randomly sampled 1,000 questions from our German GuteFrage.net dataset and annotated them with their question types and question categories. Prior to the annotation, we developed an annotation scheme for question classification that consists of 12 different question types (e.g., recommendation, opinion, description, etc.) and 50 different question categories (e.g., "Bildung", "Gesundheit", "Sport", etc.). We employed three German native speakers that, in accordance with our annotation scheme, assigned one type and up to five categories to each question. Based on this dataset, we developed a multi-label classification approach which uses a combination of unsupervised Latent Dirichlet Allocation (LDA) to extract topics from unlabeled questions and supervised training to estimate topic-category distributions based on labeled questions [35]. Our proposed approach benefits from a large number of unlabeled instances and therefore only requires a small amount of training data. It achieves consistent improvements over other linear feature-based models and increases the accuracy for question category classification by 12.9% over other state-of-the-art methods on a German question classification dataset.
To alleviate the problem of non-availability of training languages other than English, in [38] we performed experiments for question type classification (in addition to other sentence classification tasks) in a cross-lingual scenario. Here we trained models on English data and applied them to texts in different target-languages, e.g., German. Importantly, the classifier uses novel cross-lingual sentence embeddings as an input, which are obtained by concatenating different types of word embeddings and then summarizing the concatenated word embeddings of the words in the sentence with different pooling mechanisms. We performed an extensive comparison with different state-of-the-art sentence embedding techniques in monolingual and cross-lingual scenarios which show that our sentence embeddings outperform state-of-the-art sentence embeddings on question-type classification monolingually and cross-lingually by 2.2% and 1.6% accuracy, respectively.

WP2: Frame-semantic parsing for cQA. We conducted a major study on domain dependence of FrameNet semantic role labeling applied to cQA data. In [26] we have created and made available a novel, substantially sized test set based on question and answer texts (YAGS), annotated with FrameNet frames and semantic roles. We have used this test set to assess the performance FrameNet role labeling on out-of-domain data. We have found that frame disambiguation is a major bottleneck in out-of-domain FrameNet role labeling, since (1) frame disambiguation is most prone to domain shifts, and (2) role labels in FrameNet are conditioned on frame. As a solution we have proposed a frame disambiguation system based on distributed word representations, which performs on par with state of the art in-domain and outperforms the reference by 2 (YAGS) up to 11 (MASC) accuracy points out-of-domain. The system (SimpleFrameId) was made available to the research community.
Elaborating on these results, [33] explores the fundamental relationship between lexicons and distributed word representations. We show that (1) there exists a conceptual gap between word representations and lexicon entries, and (2) this gap can be partially addressed by simple preprocessing techniques: lemmatization and POS-disambiguation. The study introduces a new benchmarking method for distributed word representations – word class suggestion – and evaluates the performance of enhanced word representations on VerbNet and WordNet class prediction. The results show that lemmatization and POS disambiguation are crucial for verb similarity, consistently improving the SimVerb benchmark performance by up to 0.1 Spearman ρ and the word class suggestion performance by up to .08 F1. These results are applicable to FrameNet and any other lexical resource that groups lexemes into word classes.

Our joint work with the Univ. of Heidelberg [27] reports experiments on the German data. The goal of this comparative study is to assess the performance of three major SRL frameworks: FrameNet, PropBank and VerbNet. To enable this assessment, the first parallel corpus with FrameNet, PropBank and VerbNet-style annotations for German (SR3DE) was created and made available to the community. Using this data we evaluate the complexity of the corresponding labeling schemes, as well as the labeling performance and generalization capability of the frameworks. We find that FrameNet has the lowest generalization capability due to data sparsity and frame-conditioning of the roles, while PropBank and VerbNet roles generalize well.

One of the core features of semantic roles is their contextual dependency, which is usually modeled via global optimization of the role labeling outputs. In [32], we propose an alternative approach based on *thematic hierarchies*, a concept widely used in theoretical semantic role research but so far not applicable to SRL due to the restricted scope of the existing hierarchies. We show that thematic hierarchies for VerbNet (1) can be induced from small amounts of training data, (2) bear similarities to the proposals in the theoretical literature, and (3) to a certain extent apply cross-lingually between English and German, the latter evaluation made possible by our previously released SR3DE corpus.

Applying SRL to new domains is a challenging task due to the lack of new-domain training data and in our work [25] we show that the performance of models that were trained on smaller labeled gold data drop substantially in these scenarios. We therefor propose a method for automatic training data generation where we utilize data from linked lexical knowledge bases to label large-scale corpora (34–823k instances) with frames and semantic roles in a distant supervision setup. Due to the increased amount of training data our models improve the in-domain verb sense disambiguation performance (VSD) by 0.097 F1 in our English evaluation and by 0.007 F1 in our German evaluation. In our English out-of-domain evaluation the models that were trained on our generated training data substantially improve the VSD performances on three datasets by 0.163, 0.129, and 0.292 F1. Our evaluation thus shows that our knowledge-based approach results in suitable training data for SRL.

WP3: cQA on heterogeneous information sources. For *normalizing and preprocessing user generated data*—i.e., user queries and retrieved texts—we applied DKPro-UGD [7], a toolset developed by UKP Lab for the cleansing of noisy texts. We furthermore preprocessed the input data by applying lemmatization, POS tagging, named entity recognition and dependency parsing using the DKPro framework [10].

Passage retrieval and paraphrase recognition. One of the limiting factors in state-of-the-art neural passage retrieval methods is that they are data-hungry, i.e., they need large amounts of annotated data such as duplicate questions for training. Because such data is not available in many cQA platforms, in our ongoing work we propose to use question generation in order to automatically generate large quantities of duplicate questions (i.e., training instances). Our method does not require any labeled data and only relies on questions and their associated descriptions, which are commonly available in cQA. Our results show that models trained with generated questions outperform adversarial domain transfer and unsupervised methods with the same number of training samples and perform on-par with models that were trained on labeled in-domain duplicate questions when we use all available generated duplicates. The work is currently under review for EMNLP 2019.

Additionally, in a master thesis [55], we trained question retrieval models using different weakly supervised techniques based on the retrieval of question descriptions instead of relying on annotated question duplicates. In an ongoing bachelor thesis [54], we currently extend this work and perform an

in-depth comparison of this method across different datasets and tasks. Preliminary results show that our training method outperforms unsupervised training, adversarial domain transfer, and unsupervised retrieval baselines such as BM25.

When dealing with user queries in other languages, additional challenges arise. For example, when users formulate queries in German, it would substantially limit the usefulness of our end-to-end cQA system to use only search through German cQA data for question answering because English platforms are predominant in many cases and thus offer considerably more data, e.g., in technical domains. In [42] we therefore experiment with cross-lingual question retrieval, i.e., we retrieve similar questions from an English cQA platform for a given user query in German. We found that first applying machine translation to translate the user query to English and then continuing with a monolingual question retrieval model leads to performance decreases of 3.1–6.6pp mean average precision (MAP) compared to monolingual cases in technical domains. We close this gap by up to 36% by improving the NMT model with back-translated texts from the cQA domain and by training a more robust question retrieval model with additional noisy training data that we obtain with back-translation.

Graph-based passage selection. We divided our efforts into answer selection and graph-based question answering. This separation has been caused by the methodological shift in NLP towards neural network-based approaches after the start of the project and it allowed us to build on the recent state of the art (rather than using the planned feature-based techniques). Further, it allowed us to carefully extend neural networks with graph-based structures and to evaluate the effectiveness of such methods in QA setups.

For answer selection, in [40] we proposed a representation learning approach with a self-attentive component. The approach learns the importance of text segments in questions and answers with a separate LSTM and independently for the question and answer texts. Our approach outperforms different classical attention mechanisms, e.g., it improves the answer selection accuracy compared to the bidirectional attention mechanism proposed in [44] by 0.9pp on InsuranceQA v1 [16] and by 5pp on InsuranceQA v2. To cope with small-data scenarios in cQA, which are common when we deal with more specialized domains or with languages other than English, in [41] we presented a neural coverage-based method that test whether all aspects of the question are covered by an answer. We define aspects as the n-grams of a text and also experiment with extensions that incorporate the syntactic structure of the texts. Our approach has only few trainable network layers and it can be trained with a small number of instances—e.g., it outperforms standard information retrieval baselines already when trained with 15 questions. When using all available data, our best approach outperforms a state-of-the-art text matching approach by 6.1pp accuracy over six datasets from different domains (on average). Our analysis revealed that our approach is especially suitable to deal with long answers, i.e., ones that contain more than 150 words.

For graph-based question answering, we focused on the graph-based modelling of the semantics of the input question. The input question is represented as a graph of real-world entities, the answer node and connections between them. The explicit graph-based representation allows us to compare the question with a structured source of information, such as a knowledge base, and extract the answer. In [46] and in [49], we developed methods for an improved joint extraction of relations and identification of real-world entities in questions. We combined relations and entities into graph-based question representations and developed a novel graph neural network model to process them in [50]. The results on popular question answering datasets have shown that the graph-based modeling of the question in combination with the graph neural networks outperforms the previous approaches by 5pp

F1. We demonstrated that the proposed architecture is especially beneficial for complex questions, i.e., for questions that require two or more relations to the correct answer.

Answer summarization. We tested several general text summarization approaches in the cQA domain. In addition to an existing answer summarization dataset in English [51], we obtained German reference summaries for our GuteFrage.net dataset by employing three annotators (German native speakers) that extracted the most important sentences from all answers that were given to a question. For our dataset, we created 170 reference summaries, i.e., we manually summarized the answers to 170 questions. We subsequently used both the existing English dataset and our own German dataset and proposed an evaluation framework to test different automatic summarization approaches. We evaluated MMR [5] with the word mover's distance [31], REAPER [37], Sum Basic [53], and LexRank [14], and a number of simple baselines (e.g., choosing all sentences from the best answer as a summary). All evaluated approaches outperformed simple baselines but otherwise achieved similar results. Only a feature-based approach to answer summarization [51] achieved a substantially better result (according to ROUGE metrics). In our error analysis, we found that this is mostly due to the domain-specific adaptations of the feature-based approach with cQA specific features, e.g., answer quality, user authority, answer relevance. Because our general approaches to text summarization were not trained on cQA data, they could not make use of such information. Modern methods such as neural summarization approaches can learn such properties. However, they require a large amount of training data, which is not available in realistic answer summarization scenarios. This has been a departing point for our new preference-based summarization approach with reinforcement learning [17], which can utilize user feedback instead of learning from gold summaries. We find that our method, by using a novel objective function that is separated in active preference learning and reinforcement learning phases, substantially outperforms the existing preference-based approach SPPI [45] on DUC'04, with the same round of interaction (e.g. 10 rounds) by 18% (ROUGE-2), and outperforms non-interactive baseline by 25% (ROUGE-2) because of its ability to incorporate user feedback more efficiently. At the same time, our approach performs on-par with state-of-the-art multi-document summarization approaches on the DUC'01 and DUC'02 corpora already when trained with as few as 10 user feedback instances. Our approach is thus applicable to a wide range of scenarios where no gold summaries exist, including and beyond the cQA domain of QA-EduInf.

Evaluation in CQA. We evaluated our passage retrieval methods on several realistic datasets from a number of different educational domains. For question retrieval in [42], we used the AskUbuntu dataset [34], and data from StackOverflow. For the answer selection in [40], we evaluated on InsuranceQA v1/v2 [16] and in [41], we additionally performed experiments on WikiPassageQA [6] and on six datasets from different domains in StackExchange that we collected in WP1. For all experiments, we measured the accuracy of the ranking methods and computed ranking metrics, such as mean reciprocal rank (MRR) and mean average precision (MAP).
We did not perform experiments on DIPF datasets because of their small size (only hundreds of question-answer pairs), and because most answers follow similar patters—e.g., they often only contain weblinks. We instead used data from the German cQA platform GuteFrage.net for answer summarization and we investigated cross-lingual question retrieval for the German-English language pair [42].

WP4: Information quality assessment to enhance cQA. Our approaches to text quality assessment are based on modern neural techniques rather than on feature-based approaches. In particular,

we utilized mono- and cross-lingual sentence embeddings of WP1 for classifying texts as subjective or objective, measure sentiment, opinion polarity, and argumentation [38]. For making them applicable to German, we translated all English datasets to German using Google Translate and trained a classifier on English training data using our cross-lingual sentence embeddings as input features. We applied the resulting model to the German test data. Results show that our proposed sentence embeddings are considerably better compared to other cross-lingual adaptations of more complex state-of-the-art monolingual techniques because of a smaller cross-language performance drop. Most importantly, our approach can utilize existing English training data and effectively train models that can also be applied on the German data.

A challenge when predicting the quality of user-generated comments (e.g., in cQA platforms) is that offensive or toxic comments are often obfuscated by users to fool classifiers and filtering strategies. A common method is to apply visual modifications of the input text, e.g., by using "leet speak", which is a writing-style where users replace characters with similar looking digits (e.g., "BAD" $\rightarrow$ "B4D"). More advanced methods could further replace characters with other similar-looking alternatives from Unicode. In a collaborative project [13], we consider these types of obfuscations as visual adversarial attacks and we observe that current approaches cannot properly deal with them. For example, we observe that a toxic comment classifier, when under visual adversarial attack, predicts different (i.e., non-toxic) labels for toxic comments in more than 24% of all cases when we visually perturb 10% randomly selected characters in the input text. We propose different methods to shied approaches from such attacks, which are based on adversarial training and on visual character embeddings. Our shielding methods substantially improve the robustness of models and decrease the success rate of visual adversarial attacks when identifying toxic comments from 24% to 12%. We show that similar trends can be observed for a number of sentence-level, word-level, and character-level tasks. Our approach can thus also be used in cQA scenarios to more robustly predict the quality of user-generated answers.

WP5: User-based evaluation. In [39], we presented a web-based prototype of an end-to-end cQA system. Our web interface allows the user to enter a question and to retrieve relevant answers from a cQA platform in regard to this question. Important advantages of our prototype are the seamless switching between different answer selection models for retrieval and an interactive visualization of neural attention components. Our demonstrator integrates our results from WP3, specifically passage retrieval and passage selection, and has been made available as open-source software to the research community.

In our ongoing work, we extended our prototype with our cross-lingual question retrieval approach [42] to allow users to ask questions in German and in English. In the remainder of this project, we plan to integrate data from multiple cQA domains by first training a classifier using BERT [8] to determine the most relevant domain of the user question and then by retrieving relevant answers from this domain. Finally, we will conduct a user study to evaluate the usefulness of our prototype in an educational scenario. For this user study, we will compare three different variants of our prototype: (1) a variant that uses IR baselines, (2) the neural state of the art, (3) the neural state of the art with cross-lingual search capabilities. We will formulate 10 educational questions (scenarios) that should be answered by the subjects using our prototypes by formulating own queries (queries will not be pre-defined). We will employ 10 subjects per prototype (30 subjects in total), who all perform the same tasks, and ask them to complete questionaries after using our system. We will investigate whether neural CQA approaches improve the user satisfaction, better 'understand' the queries, and whether users benefit

from using cross-lingual search features.
We will design the experimental setup and the evaluation in coordination with Judit Bar-Ilan, Department of Information Sciences, Bar-Ilan University, who has vast experience in designing user studies. This also includes the overall setup of our study and the collection of questions.

2.3 Team and Cooperation Partners

The research staff that contributed to this project consisted of PhD students and PostDocs. Andreas Rücklé (PhD student, expected to finish his PhD in 2020) contributed to passage retrieval, passage selection, information quality assessment, and user-based evaluation. Daniil Sorokin (PhD student, expected to finish his PhD in late 2019) contributed to graph-based passage selection and question answering. Nafise Moosavi (PostDoc) contributed to passage retrieval and passage selection. Yang Gao (PostDoc) contributed to summarization. Additionally, Silvana Hartmann (now in industry) contributed to this project with her expertise on semantic role labelling. Saeedeh Momtazi (now assistant professor at Amirkabir University of Technology) contributed to question classification. Ilia Kuznetsov (PhD student, supported by FAZIT Stiftung) brought in his experience on distributional semantics.
We also established a number of important external collaborations, which are listed in a separate subsection below.

2.3.1 Outreach

We disseminated our work by publishing in a number of highly-recognized venues, including journal publications and top venues in NLP and AI including TACL, NAACL, ACL, EMNLP, CONLL, AAAI, and *SEM. In addition we gave multiple invited talks in both academia and industry, listed below.

- I. Gurevych. Interactive Machine Learning (iML) for Digital Humanities, August 2018. NTU: Nanyang Technological University, Singapore (Iryna Gurevych covered the APRIL framework for preference-based summarization)
- I. Gurevych. Out-of-Domain Semantic Role Labeling, March 2018. Center for Information and Language Processing, Ludwig-Maximilians-Universität Munich, Germany
- I. Gurevych. Adaptive Information Preparation from Heterogeneous Sources; Challenges and Tools, October 2015. Netzwerk Recherche. Dortmund, Germany (Iryna Gurevych presented the research framework and the scenario of QA-EduInf)
- I. Gurevych. Making Sense of Ubiquitous Knowledge with Natural Language Processing, June 2015. Computer Science Department at the University of Jaén, Spain (Iryna Gurevych presented the research framework and the scenario of QA-EduInf)
- I. Gurevych. Catching Up With the Trends: Language Resources and Tools for Less-Resourced Languages, September 2015. Keynote at the Conference “Language and Modern Technologies”, Tbilisi, Georgia (Iryna Gurevych covered the QA-EduInf research for less-resourced languages)

2.3.2 Cooperation partners

The UKP group set up a collaboration with the group of Prof. Dr. Anette Frank from the Department of Computational Linguistics, University of Heidelberg as part of the CLARIN project. Further, we

closely cooperated with the GRK AIPHES in the area of Question Answering (Prasetya Ajie Utama, PhD student) and Multimodal Frame Identification (Teresa Botschen, successfully defended her PhD in early 2019). Additionally, we initiated a collaboration with FactMata Ltd. (London, United Kingdom), where Daniil Sorokin spent three months as an intern in the area of credibility assessment of web-based information. Finally, we collaborated with the SoftwareCampus project "Intelligente Suche nach Informationen im Sozialen Web" (funded by BMBF) through the evaluation of our novel passage retrieval components in the context of the use-case scenarios defined by the project partner DATEV eG.

Selected publications in cooperation projects

- A. Rücklé, K. Swarnkar, and I. Gurevych. Improved cross-lingual question retrieval for community question answering. In *Proceedings of the 2019 World Wide Web Conference (WWW-19)*, pages 3179–3186, San Francisco, USA, May 2019. Association for Computing Machinery
- T. Botschen, D. Sorokin, and I. Gurevych. Frame- and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 90–96. Association for Computational Linguistics, 2018
- T. Botschen, I. Gurevych, J.-C. Klie, H. Mousselly Sergieh, and S. Roth. Multimodal frame identification with multilingual evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1481–1491. Association for Computational Linguistics, 2018
- T. Botschen, H. Mousselly Sergieh, and I. Gurevych. Prediction of frame-to-frame relations in the framenet hierarchy with frame embeddings. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 146–156. Association for Computational Linguistics, 2017

2.3.3 Participation in software challenges

We participated in the Question Answering over Linked Data (QALD) challenge, which aims at providing an up-to-date benchmark for comparing approaches that mediate between a user—who expresses his or her information need in natural language—and RDF data [52]. We won the Task 4 of QALD-7 and we outperformed the second best system by 4pp macro F1. Further, our system description paper has won the Best Challenge Paper award among all accepted challenge papers and SharedTask papers at ESWC 2017.

- D. Sorokin and I. Gurevych. End-to-end representation learning for question answering with weak supervision. In *ESWC 2017 Semantic Web Challenges*, Portoroz, Slovenia, Oct. 2017

2.4 Qualification of Young Scholars

The qualification of young scholars in this project included the qualification of graduate and undergraduate students, PhD students, as well as PostDocs.
Dr. Saeedeh Momtazi was a PostDoc at UKP (2013–2016) while contributing to this project. Since October 2016, she is an Assistant Professor in the Department of Computer Engineering and

Information Technology at Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. Dr. Yang Gao was a PostDoc at UKP (2017–2019). Since June 2019 he is a Lecturer at the Department of Computer Science, Royal Holloway, University of London, Egham, UK. Nafise Moosavi started as a PostDoc in 2018 and contributed to passage retrieval in this project.
The following theses, internships, and courses have been completed and taught (or are expected to finish in due time) in the context of this project:

PhD theses:

- S. Hartmann. *Knowledge-based supervision for Domain-adaptive Semantic Role Labeling.* PhD Thesis, Computer Science Dpt., Technische Universität Darmstadt, September 2016
- D. Sorokin. *Lexical-Semantic and World Knowledge for Semantic Parsing* PhD Thesis, Computer Science Dpt., Technische Universität Darmstadt, Upcoming (2019)
- I. Kuznetsov. *Semi-Supervised Semantic Role Labelling* PhD Thesis, Computer Science Dpt., Technische Universität Darmstadt, Upcoming (2019)
- A. Rücklé. *Large-Scale Semantic Search in Community Question Answering* PhD Thesis, Computer Science Dpt., Technische Universität Darmstadt, Upcoming (2020)

Diploma-/Master theses:

- R. Jain. Using coreference resolution in question answering and how to make the best of it. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2019
- J. Wiedmeier. Enhanced representation learning for question retrieval with transfer learning. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017
- M. Hassan. Efficient knowledge base access for relation extraction and semantic parsing. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017
- P. Dubs. Large-scale semantic question retrieval in community question answering. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017
- R. Bora. Extracting references for knowledge base facts using a semantic parser. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2016
- A. Rücklé. Real-time summarization of big data streams. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2015
- S. Henß. Retrieving and summarizing educational document collections using reinforcement learning. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2013
- D. Wu. Text classification for determining good answers on stack exchange user forums. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2014

Bachelor theses:

- J. Vatter. Exploring training strategies for cqa retrieval tasks. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, ongoing
- D. Faber. Neural sequence to sequence approaches for knowledge base question answering. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2018
- I. Zelch. Frame semantic based approach for semantic textual similarity. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2018

- N. Geisler. Enhancing complex question answering with rule-based question understanding. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017
- L. Wolf. Neural sequence to sequence approaches for mapping natural questions to sparql queries. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017

Internships:

- K. Swarnkar. 2018 (3 months). Retrieval and ranking of questions in a cross-lingual German-English setting utilizing neural machine translation and neural network-based learning to rank methods.
- X.-S. Vu. 2016 (4 months). Creation of a software framework to answer argumentative questions, which included data acquisition, literature research, creation of an argument retrieval system, and argument paraphrase detection.
- O. Zayed. 2016 (4 months). Evaluation of different answer summarization approaches in comparable setups and the development of an evaluation framework for answer summarization.
- Y. Shao. 2015 (6 months). Summarization of relevant answers for questions posed in community question answering websites.
- I. Kuznezov. 2015 (12 months). Implementation and extension of a monolingual annotation projection method that transfers semantic role labelling annotations from a source corpus to target corpus via dependency graph alignment.

Teaching:

- Seminar. *Text Analytics*. 2017, summer semester. Research, critical assessment, comparison, and presentation of community question answering literature and state-of-the-art approaches. 13 students participated in this course.
- Software project. *Data Analysis Software Project for Natural Language*. 2016, summer semester. Development of an end-to-end question answering system that uses community question answering data from the web to retrieve answers. 9 students participated in this course.

Student research projects ("Studienarbeiten"):

- J. Wiedmeier. *Deep-Learning-Based Answer-Ranking in CQA*. 2016/2017, winter semester. 9CP. Adaptation of neural answer selection methods to answer ranking in community Question Answering and creation of an extensible software architecture for their evaluation.
- R. Rieb. *Automatic Alignment of Role Schemata*. 2015, summer semester. 15CP. Development of methods for the automatic alignment of resources on different levels of semantic representation, namely semantic role schemata (FrameNet and VerbNet, semantic predicates and semantic roles).

3 Summary and Conclusions (Zusammenfassung)

In the first funding phase of QA-EduInf, we have used large quantities of data from community question answering (cQA) forums to automatically answer new user questions. We have explored several foundational components of the cQA pipeline, with a particular focus on frame semantic parsing—including semantic role labeling with German texts from educational CQA domains—and neural methods to retrieve and select relevant question and answer passages. Based on our published work, we have developed prototypical systems for automatic question answering using cQA data. Whenever applicable, the data and source code produced by UKP Lab has been made publicly and openly available with the corresponding scientific publications.

Based on our findings, we see several important areas that can benefit from the work we have completed in the first funding phase of QA-EduInf. For instance, dialog systems could benefit from our research to frame semantic parsing in related conversational scenarios. Further, our methods could be used to improve the retrieval of information from cQA forums in dialogues. Finally, our novel datasets as well as our free and open source code can benefit a wide range of researchers in related areas, from building end-to-end QA systems to performing preference-based summarization with a small number of instances.

References

[1] R. Bora. Extracting references for knowledge base facts using a semantic parser. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2016.

[2] T. Botschen, I. Gurevych, J.-C. Klie, H. Mousselly Sergieh, and S. Roth. Multimodal frame identification with multilingual evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1481–1491. Association for Computational Linguistics, 2018.

[3] T. Botschen, H. Mousselly Sergieh, and I. Gurevych. Prediction of frame-to-frame relations in the framenet hierarchy with frame embeddings. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 146–156. Association for Computational Linguistics, 2017.

[4] T. Botschen, D. Sorokin, and I. Gurevych. Frame- and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 90–96. Association for Computational Linguistics, 2018.

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[6] D. Cohen, L. Yang, and B. W. Croft. WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval. In *SIGIR*, pages 1165–1168, 2018.

[7] R. E. De Castilho and I. Gurevych. Dkpro-ugd: a flexible data-cleansing approach to processing user-generated discourse. In *Onlineproceedings of the First French-speaking meeting around the framework Apache UIMA, LINA CNRS UMR*, page 33, 2009.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *ArXiv preprint*, 2018.

[9] P. Dubs. Large-scale semantic question retrieval in community question answering. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017.

[10] R. Eckart de Castilho and I. Gurevych. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[11] S. Eger, A. Rücklé, and I. Gurevych. Pd3: Better low-resource cross-lingual transfer by combining direct transfer and annotation projection. In *5th Workshop on Argument Mining, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 131–143, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[12] S. Eger, G. G. Şahin, A. Rücklé, J.-U. Lee, C. Schulz, M. Mesgar, K. Swarnkar, E. Simpson, and I. Gurevych. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 1634–1647, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] S. Eger, G. G. Şahin, A. Rücklé, J.-U. Lee, C. Schulz, M. Mesgar, K. Swarnkar, E. Simpson, and I. Gurevych. Text processing like humans do: Visually attacking and shielding nlp systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, page to appear, Februar 2019.

[14] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[15] D. Faber. Neural sequence to sequence approaches for knowledge base question answering. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2018.

[16] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou. Applying Deep Learning to Answer Selection: A Study and an Open Task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820, 2015.

[17] Y. Gao, C. M. Meyer, and I. Gurevych. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4120–4130, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[18] Y. Gao, C. M. Meyer, M. Mesgar, and I. Gurevych. Good rewards are all you need: Reward learning for efficient reinforcement learning in document summarisation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, page (to appear), Macao, China, Aug 2019. Association for Computational Linguistics.

[19] N. Geisler. Enhancing complex question answering with rule-based question understanding. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017.

[20] I. Gurevych. Adaptive Information Preparation from Heterogeneous Sources; Challenges and Tools, October 2015. Netzwerk Recherche. Dortmund, Germany (Iryna Gurevych presented the research framework and the scenario of QA-EduInf).

[21] I. Gurevych. Catching Up With the Trends: Language Resources and Tools for Less-Resourced Languages, September 2015. Keynote at the Conference "Language and Modern Technologies", Tbilisi, Georgia (Iryna Gurevych covered the QA-EduInf research for less-resourced languages).

[22] I. Gurevych. Making Sense of Ubiquitous Knowledge with Natural Language Processing, June 2015. Computer Science Department at the University of Jaén, Spain (Iryna Gurevych presented the research framework and the scenario of QA-EduInf).

[23] I. Gurevych. Interactive Machine Learning (iML) for Digital Humanities, August 2018. NTU: Nanyang Technological University, Singapore (Iryna Gurevych covered the APRIL framework for preference-based summarization).

[24] I. Gurevych. Out-of-Domain Semantic Role Labeling, March 2018. Center for Information and Language Processing, Ludwig-Maximilians-Universität Munich, Germany.

[25] S. Hartmann, J. Eckle-Kohler, and I. Gurevych. Generating training data for semantic role labeling based on label transfer from linked lexical resources. *Transactions of the Association for Computational Linguistics (TACL)*, 4:197–213, Mai 2016.

[26] S. Hartmann, I. Kuznetsov, T. Martin, and I. Gurevych. Out-of-domain framenet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 471–482. Association for Computational Linguistics, April 2017.

[27] S. Hartmann, É. Mújdricza-Maydt, I. Kuznetsov, I. Gurevych, and A. Frank. Assessing srl frameworks with automatic training data expansion. In *Proceedings of the 11th Linguistics Annotation Workshop (LAW XI) at EACL 2017*, pages 115–121. Association for Computational Linguistics, April 2017.

[28] M. Hassan. Efficient knowledge base access for relation extraction and semantic parsing. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017.

[29] S. Henß. Retrieving and summarizing educational document collections using reinforcement learning. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2013.

[30] R. Jain. Using coreference resolution in question answering and how to make the best of it. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2019.

[31] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[32] I. Kuznetsov and I. Gurevych. Corpus-driven thematic hierarchy induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, pages 54–64, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[33] I. Kuznetsov and I. Gurevych. From text to lexicon: Bridging the gap between word embeddings and lexical resources. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 233–244, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[34] T. Lei, H. Joshi, R. Barzilay, T. Jaakkola, K. Tymoshenko, A. Moschitti, and L. Marquez. Semi-supervised Question Retrieval with Gated Convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 1279–1289. Association for Computational Linguistics, 2016.

[35] S. Momtazi and I. Gurevych. Corrigendum to unsupervised latent dirichlet allocation for supervised question classification. *Information Processing & Management*, 54(3):380–393, 2018.

[36] N. S. Moosavi, L. Born, M. Poesio, and M. Strube. Handling domain shift in coreference evaluation by using automatically extracted minimum spans. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, page (to appear), Florence, Italy, Jul 2019. Association for Computational Linguistics.

[37] C. Rioux, S. A. Hasan, and Y. Chali. Fear the reaper: A system for automatic multi-document summarization with reinforcement learning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 681–690, 2014.

[38] A. Rücklé, S. Eger, M. Peyrard, and I. Gurevych. Concatenated p-mean embeddings as universal cross-lingual sentence representations. *arXiv*, 2018.

[39] A. Rücklé and I. Gurevych. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations (ACL 2017)*, pages 19–24, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

[40] A. Rücklé and I. Gurevych. Representation learning for answer selection with lstm-based importance weighting. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, Montpellier, France, Sept. 2017. Association for Computational Linguistics.

[41] A. Rücklé, N. S. Moosavi, and I. Gurevych. Coala: A neural coverage-based approach for long answer selection with small data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, page (to appear), Honolulu, USA, Jan. 2019. Association for the Advancement of Artificial Intelligence.

[42] A. Rücklé, K. Swarnkar, and I. Gurevych. Improved cross-lingual question retrieval for community question answering. In *Proceedings of the 2019 World Wide Web Conference (WWW-19)*, pages 3179–3186, San Francisco, USA, May 2019. Association for Computing Machinery.

[43] A. Rücklé. Real-time summarization of big data streams. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2015.

[44] C. d. Santos, L. Barbosa, D. Bogdanova, and B. Zadrozny. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 694–699. Association for Computational Linguistics, 2015.

[45] A. Sokolov, J. Kreutzer, S. Riezler, and C. Lo. Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1489–1497, 2016.

[46] D. Sorokin and I. Gurevych. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[47] D. Sorokin and I. Gurevych. End-to-end representation learning for question answering with weak supervision. In *ESWC 2017 Semantic Web Challenges*, Portoroz, Slovenia, Oct. 2017.

[48] D. Sorokin and I. Gurevych. Interactive instance-based evaluation of knowledge base question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 114–119, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[49] D. Sorokin and I. Gurevych. Mixing context granularities for improved entity linking on question answering data across entity categories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, pages 65–75, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[50] D. Sorokin and I. Gurevych. Modeling semantics with gated graph neural networks for knowledge base question answering. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 3306–3317, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[51] M. Tomasoni and M. Huang. Metadata-aware measures for answer summarization in community question answering. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 760–769. Association for Computational Linguistics, 2010.

[52] R. Usbeck, A.-C. N. Ngomo, B. Haarmann, A. Krithara, M. Röder, and G. Napolitano. 7th open challenge on question answering over linked data (qald-7). In *Semantic Web Evaluation Challenge*, pages 59–69. Springer, 2017.

[53] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

[54] J. Vatter. Exploring training strategies for cqa retrieval tasks. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, ongoing.

[55] J. Wiedmeier. Enhanced representation learning for question retrieval with transfer learning. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017.

[56] L. Wolf. Neural sequence to sequence approaches for mapping natural questions to sparql queries. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2017.

[57] D. Wu. Text classification for determining good answers on stack exchange user forums. Master's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2014.

[58] I. Zelch. Frame semantic based approach for semantic textual similarity. Bachelor's Thesis, Computer Science Dpt., Technische Universität Darmstadt, 2018.