Towards Synthesizing E-Mail Conversations as Part of Knowledge Work Datasets with Large Language Models

Desiree Heim[1,2,*], Christian Jilek[1], Adrian Ulges[3] and Andreas Dengel[1,2]

[1]*Smart Data and Knowledge Services Department, German Research Center for Artificial Intelligence (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany*

[2]*Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany*

[3]*Department DCSM, RheinMain University of Applied Sciences, Kurt-Schumacher-Ring 18, 65197 Wiesbaden, Germany*

Abstract

Data-driven evaluations or optimizations of knowledge work support tools are challenging due to the absence of a generally usable, comprehensive dataset that provides sufficient information about the backgrounds of users and their documents. Since data collections suffer from issues like data incompleteness due to data protection measures and lack of thorough annotations, we develop a configurable dataset generator, called KnoWoGen, that simulates collaborative, task-based knowledge work. While in the past a major problem of synthesizing such a dataset was the generation of authentic and diverse documents, the emergence of Large Language Models (LLM) enables it. Hence, in the KnoWoGen, an LLM is prompted to generate task-related documents. Hereby, task configurations include a domain or general topic which is used to randomly generate a more specific subtopic at simulation time to condition the generation of the related document. Additionally, the KnoWoGen stores all available contextual information about the documents and the simulation environment in a knowledge graph.

As a proof of concept, we study the generation of e-mail conversations as relevant representatives of knowledge work documents reflecting collaboration. Such threads are particularly difficult to collect in real environments since the involvement of third parties typically hinders their publication and, in laboratory settings, require a substantially higher amount of resources to plan and simulate. In a study conducted to assess the quality of generated e-mail threads, participants rated them regarding their naturalness, coherence, answer quality, and content advances. Overall, two-thirds got the highest or second-highest score on a 5-point scale.

Keywords

Conversation generation, Knowledge work datasets, Large Language Models, Knowledge graphs, Evaluation of knowledge work support tools

1. Introduction

Compared to user studies, data-driven evaluations of knowledge work support tools, like task predictors or document recommenders, enable comparisons between tools and offer more objective, reproducible insights into the tools' performance and its backgrounds, such as reasons for issues or correct results.

However, as also Gonçalves [1] stated, collecting a comprehensive dataset is challenging not least because of the required, extensive data annotations that would be necessary to get sufficient information about the users' and their documents' background and ground truth data for evaluations. Even if data collections are annotated with contextual information the annotations might not be sufficient for different evaluation use cases. Moreover, issues like data incompleteness due to privacy-, confidentiality- and copyright-preserving, such as censoring and deletion, remain for real-life data collections. From all publicly available knowledge work datasets published over the years, the more recent RLKWiC dataset [2] provides the most background information but is still subject to the aforementioned issues.

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands

*Corresponding author.

desiree.heim@dfki.de (D. Heim); christian.jilek@dfki.de (C. Jilek); adrian.ulges@hs-rm.de (A. Ulges); andreas.dengel@dfki.de (A. Dengel)

https://www.dfki.uni-kl.de/~heim/ (D. Heim); https://www.dfki.uni-kl.de/~jilek/ (C. Jilek); https://www.cs.hs-rm.de/~ulges/ (A. Ulges); https://www.dfki.uni-kl.de/~dengel/ (A. Dengel)

0000-0003-4486-3046 (D. Heim); 0000-0002-5926-1673 (C. Jilek); 0000-0002-6100-8255 (A. Dengel)

Because of these disadvantages, we are currently working on a paper elaborating in detail the state of the art regarding knowledge work datasets and why generating datasets can be advantageous over collecting data. This motivation also led to our recently proposed knowledge work dataset generator KnoWoGen [3]. It simulates configurable scenarios in which multiple knowledge workers complete tasks, create and utilize documents, and collaborate with others. All documents are generated during the simulation by prompting a Large Language Model [4] with task-specific instructions and all relevant contextual information. For instance, in the configuration, domains are defined for the tasks. At the simulation time, more fine-granular topics of this domain are generated and one is randomly selected and given in the prompt to generate the document. The main advantage of the simulation is that all modeled or controlled background information about the knowledge workers and their documents is known and can be stored alongside the simulation process. To make the contextual data easily accessible for later simulation steps or at evaluation time, KnoWoGen stores it in the form of a Knowledge Graph [5].

In a previous paper [3], we showed that the KnoWoGen can generate authentic documents that humans cannot reliably distinguish from real documents. While, in that paper, the focus was on single documents without any interdependencies, in this paper, we concentrate on e-mail conversations.

In knowledge work, e-mails are an important type of document. On average, roughly 347 billion e-mails are sent per day [6] in total, and approximately half of them are business e-mails [7]. However, when conducting data collections, e-mails are particularly sensitive since typically third parties are involved which impedes their publication. Alternatively, in laboratory settings, it would be possible to avoid such issues by requesting the participants to collaborate. Nevertheless, it would require a substantially higher amount of resources to plan the collection setting and ensure proper collaboration. Besides, it could be difficult to imitate realistic collaboration processes. Although there exist two popular, business-focused e-mail datasets, Enron [8] and Avocado [9], they are unsuitable as knowledge work benchmarks due to their lack of contextual information about the e-mails and involved persons including other related documents like text files. Thus, relevant input information or ground truth data is lacking. To the best of our knowledge, there is also no synthetic dataset containing e-mails and comprehensive contextual information about them and their environment.

This paper focuses on how our current KnoWoGen prototype generates e-mail conversations. Hereby, we investigate, in a user study, whether the KnoWoGen can generate threads of high quality regarding the aspects of naturalness, coherence, answer quality, and content advances. In the paper, we first introduce the general functionality of the KnoWoGen (Sect. 2), explain how e-mail threads are generated (Sect. 3), present the aforementioned user study (Sect. 4), and conclude with an outlook on future work (Sect. 5).

2. KnoWoGen - The Knowledge Work Dataset Generator

The general functionality of the KnoWoGen is shown in Figure 1: First, an engineer of a knowledge work support tool, who wants to evaluate their tool, specifies the configuration. Subsequently, the simulation environment with the knowledge workers, their tasks, and other relevant entities like projects, companies, or products is set up according to the configuration. All information about this environment is stored in a knowledge graph. In the succeeding simulation steps, tasks are assigned to agents, and task-correspondent documents are generated by prompting a Large Language Model (LLM) with task-specific instructions and all other relevant contextual information about involved entities or related artifacts using a suitable parameterized prompt template. These synthesized documents are stored in a document base. Again, all contextual information utilized during the simulation is stored in the knowledge graph. This knowledge graph is built upon an extended version of the PIMO ontology [10], a well-known personal information management ontology. Finally, the generated knowledge work dataset composed of the document base and the knowledge graph can be used to evaluate tools that, for instance, predict directly related documents, detect tasks, or classify documents concerning parameters used in the simulation. More details about the general design of the KnoWoGen can be found in Heim et al. [3].

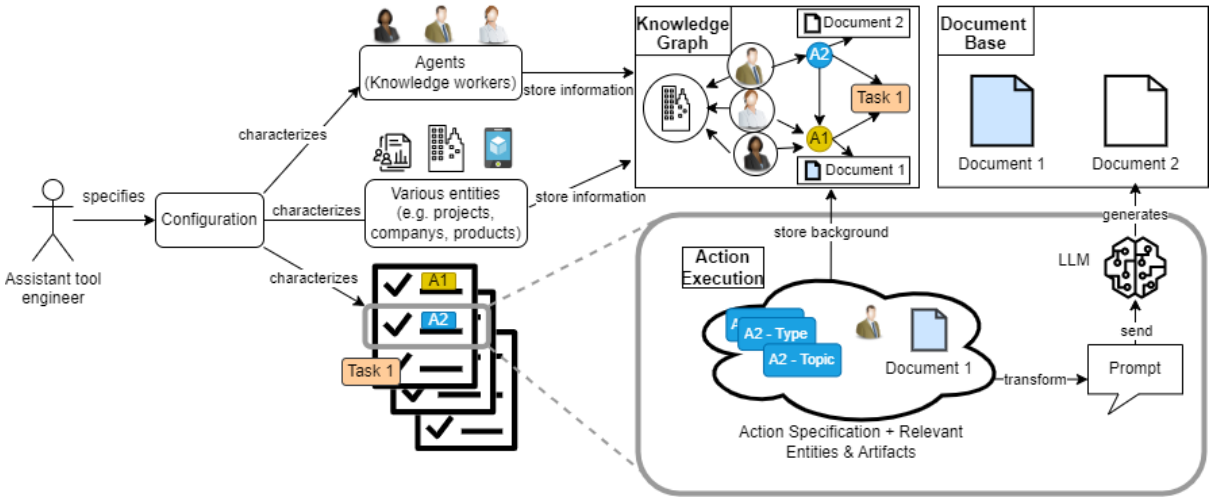

Figure 1: Overview of the KnoWoGen. First, the simulation is configured and set up, and then in the simulation steps, actions, e.g., Action A2 in this example, are executed and mainly result in documents generated by an LLM.

3. Synthesizing E-mail Conversations

E-mails are specific knowledge work documents. They are also generated by a Large Language Model (LLM) in the context of specifically defined actions, i.e., substeps of larger tasks, and can depend on other documents, as explained in Section 2. However, compared to other document types, e-mails are designed more openly since the rough contents, document dependencies, and other action-specific characteristics, like how formal the tone should be, are given for the whole e-mail thread and not for every single e-mail. In the current prototype, these specifications are represented in the prompt to generate the first e-mail. For the consecutive e-mails, we noticed in pretests that including the entire previous e-mail, when generating replies, results in unsubstantial answers addressing too many details of this e-mail. To address this issue, we implemented two mechanisms to get a higher focus on essential aspects.[1]

The first condensation mechanism is *summarization*. Here, the previous e-mail or thread is summarized in a few sentences. The following reply generation prompt includes this summarization with the instruction to answer the previous e-mail. If only this mechanism is used, the group of recipients who should reply to the e-mail has to be defined externally. Per selected recipient, one reply is generated.

The other mechanism is *question generation*. For this mechanism, there are two variants. First, in the *implicit variant*, questions are encouraged by including a dedicated instruction in the prompt generating the initial e-mail that should be answered with respect to these questions. Alternatively, questions can be generated based on an existing e-mail (*explicit variant*). In this case, the generated questions are included in a second e-mail from an initial receiver to the sender who should answer them.

For the first question generation variant, implicitly included questions are extracted in a structured list of question-addressee tuples using the Langchain framework's OutputParsers[2]. Finally, for each recipient addressed with questions, a reply is generated with the instruction to answer the respective questions. In the other variant, the addressee is always the sender of the initial e-mail and, since questions were produced in a separate step, they are known and do not have to be extracted. Since questions only address the initial sender, the number of chosen questioners determines the final number of replies.

Both condensation mechanisms, the summarization, and the question generation, can be combined. This is especially meaningful for longer e-mail threads or long e-mails to avoid reaching prompt length

[1]Examples including prompts, generated e-mails and accompanying knowledge graph excerpts can be found online: https://purl.archive.org/knowogen/examples/email_threads

[2]Langchain is a framework for working with LLMs. See also: https://python.langchain.com/

limits. While only utilizing the summarization mechanism can potentially generate more diverse answers since it is less focused on specific questions asked, the question-generation process offers more background information about the e-mail conversation. Thus, extracted questions with their inquirer, addressee, the question text, and the replies in which they are answered can be stored in the knowledge graph. This enriches the dataset with more background information about the e-mails that can also serve as ground truth in later evaluations targeting, for instance, in which e-mail certain questions were answered.

4. Evaluation

Setup. To examine the quality of the generated e-mail conversations, we conducted a user study, in which participants rated the synthetic single-turn, i.e., an e-mail and its reply, and multi-turn conversations regarding their naturalness, thread coherence, answer quality, and content advances[3] on a 5-point Likert scale [11]. Before the experiment, the participants were told that the e-mail threads had been generated.

The conversations have been generated by version 0.2 of the Mistral-7B-Instruct model [12]. We have chosen this LLM since, at the time of the experiment, conduction, it showed a good ratio between model size and achieved scores on several benchmarks [4]. Moreover, it supported a comparably high context size of 32k tokens and thus was able to also generate of longer documents.

The generated conversations had various topics. Hence, agents discussed about planning a language course, organizational questions about a course, strategic planning of job interviews, and planning a company party. The two single-turn conversations were generated according to the implicit question generation variant as explained in the previous section. We selected the implicit variant as a representative of question generation mechanisms because we perceived the questions slightly more natural. The single-turn conversations were initial e-mails and respectively one reply of one recipient. Similarly, the KnoWoGen generated the first two e-mails of the two longer conversation chains composed of four e-mails. For comparison, the next and last two replies were generated with the summarization mechanism. Again, only one part of the thread involving the sender and one recipient of the initial e-mail as senders was considered.

The first aspect examined in the experiment was the naturalness of single-turn e-mail conversations. In an earlier experiment conducted on single documents [3], we noticed that participants judged naturalness often based on social aspects, such as how an author refers to colleagues. Hence, we asked two separate questions to evaluate how naturalness is judged - on a social and linguistic level. Additionally, participants were also asked about the coherence of the reply with the earlier e-mail and how well it addressed the posed questions. For the longer conversation chains, we asked whether the conversation led to content advances and whether e-mails respect the entire preceding e-mail thread.

In total, 49 participants aged between 18 and 54 with 34 males and 15 females completed the study. The majority were students, researchers, or software engineers. 43% had a background in Computer Science. Almost all participants stated that their English language proficiency was B1 or higher. 65% of the participants used LLMs regularly, 27% occasionally, and the rest had not used LLMs before.

Results. Overall, the participants gave high ratings to the threads' quality. Figure 2 depicts the score distribution per question. Especially for the questions about single-turn conversations, the participants had a high agreement, and respectively 75% gave a score of 4 or 5. The coherence and answer quality of single-turn conversations achieved the highest score, while the content advances and the answer quality of the multi-turn interactions received a lower rating and higher variance.

Most comments given for the single-turn conversations addressed the naturalness of the e-mails. Participants stated, in particular, that some e-mails were too enthusiastic and that the language and

[3]The generated e-mails and study questions are available here: https://purl.archive.org/knowogen/experiments/email_threads
[4]We consulted the Huggingface Leaderboard for Open LLMs [13] which summarizes the performance of a range of LLMs on several common benchmarks

tone did not fit the social roles of the involved persons. Besides, there were only a few other remarks stating that one reply did not contain the names of all persons included in the previous e-mail and two comments about questions that were not properly addressed in the reply. Overall, no comments were mentioning major issues regarding the coherence of the e-mails or the reply quality. In contrast, there were several comments for communication chains indicating that from the second reply on, which was the first one without an explicitly introduced or extracted question, there were barely any advances in content but rather the content from the first two e-mails was repeated. Moreover, some participants stated that involved communication partners confused their role in the discourse and answered their questions or addressed their ideas as if they were proposed by another person.

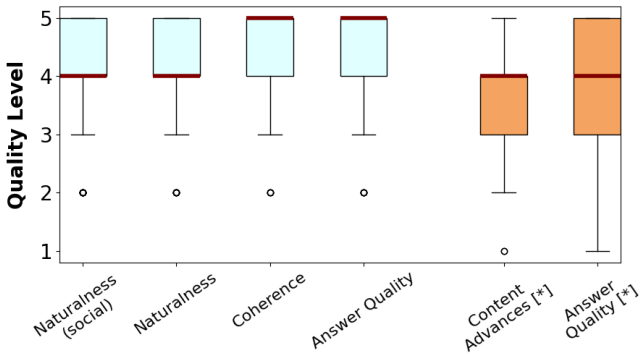

Figure 2: Boxplots showing the dispersion of given ratings. The color-coded grouping shows which questions were asked in a bundled way for the same generated artifacts. The first group of questions (blue) was about a single-turn conversation, and the other group (orange, also marked with [*]) was about a multi-turn conversation.

5. Conclusion

This paper gave a brief introduction of our knowledge work dataset generator KnoWoGen, and focused on how e-mail threads are currently implemented. The user study showed that especially single-turn e-mail threads, in which replies focused on specific questions from the previous e-mail, were perceived as natural, coherent, and the response as accurate. However, there were some issues with the identity of the sender and the summarization-focusing method led to little content advances. In future versions, prompts of consecutive e-mails could, for example, use an instruction stating that the Large Language Model should take the role of the sender and make it clearer what the sender and others contributed to previous e-mails. Moreover, since participants especially perceived replies without a preceding question as unsubstantial, a dynamic decision of whether e-mails without explicit questions require a reply could be incorporated. Apart from the mentioned options to improve the e-mail generation even more, the study indicated that the current KnoWoGen version is already a solid fundament for further works. In future experiments, since the user study showed that the e-mail are overall of a reasonable quality, generated e-mail conversations can be additionally evaluated automatically, i.e., by employing, for instance, an LLM to verify that all questions are answered in a reply or checking the consistency among e-mails. This would also allow to test various settings or LLMs and compare the quality of more generated documents without having a high manual effort.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) in the project SensAI (grant no. 01IW20007).

References

[1] D. Gonçalves, Pseudo-desktop collections and PIM: The missing link, in: ECIR 2011 workshop on evaluating personal search, 2011, pp. 3–4.

[2] M. Bakhshizadeh, C. Jilek, M. Schröder, H. Maus, A. Dengel, Data collection of real-life knowledge work in context: The RLKWiC dataset, in: Information Management, Springer, 2024, pp. 277–290.

[3] D. Heim, C. Jilek, A. Ulges, A. Dengel, Using large language models to generate authentic multi-agent knowledge work datasets, in: INFORMATIK 2024, Gesellschaft für Informatik e.V., Bonn, 2024, pp. 1347–1357.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. Wen, A survey of large language models, CoRR abs/2303.18223 (2023).

[5] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2022) 71:1–71:37.

[6] The Radicati Group, Email statistics report, 2023-2027, 2023. URL: https://www.radicati.com/wp/wp-content/uploads/2023/04/Email-Statistics-Report-2023-2027-Executive-Summary.pdf.

[7] The Radicati Group, Email statistics report, 2015-2019, 2015. URL: https://www.radicati.com/wp/wp-content/uploads/2015/03/Email-Statistics-Report-2015-2019-Executive-Summary.pdf.

[8] B. Klimt, Y. Yang, The enron corpus: A new dataset for email classification research, in: Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings, volume 3201 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 217–226.

[9] D. Oard, W. Webber, D. A. Kirsch, S. Golitsynskiy, Avocado research email collection, 2015. URL: https://catalog.ldc.upenn.edu/LDC2015T03.

[10] L. Sauermann, L. van Elst, K. Möller, Personal information model (PIMO) ontology v1.3, Online, 2013. URL: https://www.semanticdesktop.org/ontologies/2007/11/01/pimo/.

[11] R. Likert, A technique for the measurement of attitudes., Archives of psychology (1932).

[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. S. et al., Mistral 7B, CoRR abs/2310.06825 (2023).

[13] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, T. Wolf, Open LLM Leaderboard v2, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.