AFRL-AFOSR-UK-TR-2022-0018

Knowledge Graph Construction from Situated Multimodal Dialogue

Paul Groth
UNIVERSITEIT VAN AMSTERDAM
SPUI 21
AMSTERDAM, NOORD-HOLLAND, 1012 WX
NLD

01/21/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20220121	Final	START DATE 20191101	END DATE 20211031

4. TITLE AND SUBTITLE
Knowledge Graph Construction from Situated Multimodal Dialogue

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA8655-20-1-7005	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Paul Groth

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITEIT VAN AMSTERDAM
SPUI 21
AMSTERDAM, NOORD-HOLLAND 1012 WX
NLD

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2022-0018

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Computationally mediated dialogue (e.g. chat) has emerged as a fundamental mechanism for team performance in terms of coordination, planning and action. Being able to extract information from both dialogue and its surrounding context (e.g. shared documents, organizational embedding) can enable new forms of analytics and help power intelligent interventions. Current research has focused on the development of conversational agents and not on the analytics over dialogue itself. In particular, there is a paucity of research that considers the multimodal nature of a dialogue's context with respect to dialogue itself.. This project focuses on the extraction of information about dialogue and its context in the form of a knowledge graph – a knowledge base describing entities (e.g. people, organizations) and their relationships to one another. Knowledge graph construction combines the areas of information extraction, knowledge fusion, and graph refinement, thus making it a challenging problem. Specifically, the project aims to define a set of benchmark tasks and associated datasets for extracting knowledge graphs from dialogue along with its context. Additionally, it will develop a state-of-the-art baseline system for the defined tasks.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	12

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANDINI IYER	314-235-6161

FA8655-20-1-7005
Knowledge Graph Construction from Situated Multimodal Dialogue

Prof. Paul Groth (PI) and Valentin Vogelmann
University of Amsterdam

November 1 2019 – October 31, 2021
Final Research Report

1 Summary

This report documents the results of the project Knowledge Graph Construction for Situated Multimodal Dialogue. The project has achieved a number of results, including:

- benchmark dataset and task definitions for the problem of extracting knowledge graphs from complex conversational data;
- open-source software for extracting such knowledge graphs and obtaining benchmark data;
- baselines for the aforementioned tasks;
- usage of this the software by AFRL collaborators; and
- definition of follow-on research agenda.

Software can be found at `https://github.com/INDElab/conversationkg`.

2 Introduction

This report describes the results of the project Knowledge Graph Construction for Situated Multimodal Dialogue. The project aimed to provide a foundation for understanding computationally mediated dialogue in its overall context. Such dialogue is a fundamental part of team performance in terms of coordination, planning and action [12, 1]. In such a setting, (e.g. shared chat or email) it is important not only to understand the back and forth of the dialogue itself but also how that dialogue is situated in a larger context whether that be common documents or shared background information. This project hypothesized that knowledge graphs provide a powerful way to capture not only the structure of such dialogue but the information that provides the situation in which that dialogue resides. Knowledge graphs – graph structured knowledge bases that capture information about the entities, their attributes and relations between entities [8, 15] are a frequently used approach to integrate information, fuel downstream analysis and support other AI tasks [11, 9] Specifically, then, the project investigated the establishment of a research foundation for addressing the problem of knowledge graph construction for such situated dialogue. This problem can be summarized in the research question: Can we learn to detect individuals, their types, and how they have interacted and what they have interacted about from dialogue-based situated multilingual conversation. To build this foundation the project aimed at three objectives:

1. the definition of benchmark tasks for this problem;
2. the identification and construction of benchmark corpus or corpora;

3. the development of a baseline system.

In addition to these objectives, the project aimed to establish a collaborative relationship with AFRL's System Analytics Research Area. This report builds on our midterm report as a majority of the results were completed in that period.

3 Accomplishments

Here we outline the key accoplishments of the project divided into the methods and then the specific outcomes.

3.1 Methods, Assumptions and Procedures

We discuss the benchmark datasets and tasks and system we developed.

3.1.1 Benchmark dataset

We surveyed a number of existing datasets to use a basis for benchmark development. This included:

- Ubuntu Dialogue corpus of IRC chat data [7, 5]
- Reddit forum data [3]
- Spotify Podcasts Dataset[1]
- The Enron Email Corpus [4]
- World Wide Web Consortium (W3C) emails about standards development[2] [16]

We were guided by a number of principles in our survey of datasets. First, we wanted to have data that was representative of more complex dialogues or conversations that occurred over time. Second, it was imperative that these conversations reference external multimodal information that situated the conversations. Third, the data should be potentially easily distributable to the community. Finally, the data should provide a reasonable analogue to data of interest to AFRL.

Based on these factors, we decided that the W3C emails were a good starting point. The World Wide Web Consortium (W3C) is a core standardization body of the internet, developing and deciding standards such as HTML and providing support for such standards. In their mailing lists, members of the W3C as well as other organisations, private

[1] https://engineering.atspotify.com/2020/04/16/introducing-the-spotify-podcast-dataset-and-trec-challenge-2020/

[2] `https://tides.umiacs.umd.edu/webtrec/trecent/parsed_w3c_corpus.html`

persons and automated mail agents interact, inquire and announce around concepts, standards and events pertaining to the internet. These mailinglists exist since 1997 and are publicly available on the W3C's mailinglist archives.[3]

W3C emails are public and describe complex back-and-forths about deeper technical subjects over quite a lot of time scales. Many actors and organizations are covered. Additionally, because they reference standards specifications - which contain complex multimodal information (e.g. diagrams, cited works, code) - these dialogues are situated in rich background material. The Enron Email corpus was also used as a comparison point.

For our first developments, we used the corpus scrapped by Wu in 2004. This contained 170,000 emails. However, during our usage, we found a number of issues in terms of textual processing (e.g. tokenization, whitespace, unicode) that made it difficult for downstream use. Additionally, we wanted to create a much larger corpus. Thus, we built a new scraper and re-scraped the entirety of the W3C corpus up-to 2020 (See Results).

3.1.2 Benchmark Tasks

The overall problem is to construct a complete knowledge graph that represents the totality of information useful to end-users. This is obviously a complex task but importantly we miss the the *ground truth* for such a task. Annotating such data by hand, in particular for complex domains where domain experts are not available, is beyond the scope of this project and indeed makes it difficult to do this in practice. Given this we looked for "natural" tasks where we could source ground truth from the dataset itself. Here, two tasks were investigated.

1. Entity Resolution as Author Attribution. Entity resolution - determining which entity descriptions refer to the same real world entity [2] is a key task in knowledge graph construction. In the context of emails, we are able to extract ground truth data. Specifically, we formulate the task as follows: We consider entities (persons, organisations, topics, etc.) to be represented by the texts they produce. In this context, then, an entity is represented by the set of e-mails sent in its name, therefore, entity resolution becomes equivalent to authorship attribution.

Our basic assumption is that if entity labels X and Y for emails e_1 and e_2 are the same, then e_1 and e_2 really are from the same author (i.e. entity). Such data are positive samples. Importantly, if we randomly sample two emails e_1 and e_2, then with overwhelming probability, they are not from the same author (regardless of the entity labels of e_1 and e_2).

Thus, the specific task is: given emails e_1 and e_2 with labels X and Y, decide whether e_1 and e_2 are from the same author (i.e. the labels X and Y identify the same entity). This can be phrased as $P(author(e_1) = author(e_2))$.

In order perform entity resolution, sample email pairs are formed from all / some labels X and Y and merge the labels if the emails pair has high probability of being from the

[3] `https://lists.w3.org/Archives/Public`

same author (labels becomes aliases) . The advantage here is if X and Y each have multiple emails associated with them, then we can rely voting strategies to increase confidence and reliability.

2. Reconstruction of Email Metadata Knowledge Graph. A crucial ingredient in using knowledge graphs, and the large majority of machine learning, is of course ground truth, a set of verified data to measure performance against. To overcome the difficulty of sourcing ground truth for conversational knowledge graphs, we realized that the email corpus includes metadata of email servers' transactions and that this information is arguably verified. This argument allows us to extract knowledge graphs on different levels of the conversations and hence different levels of verification of the information in them. In brief:

- EmailKG represents the ground truth of email conversations; it contains the emails' meta-information, such as sender and receiver information, and additionally any information that can be extracted with high certainty, such as mentioned links or addresses.
- TextKG contains only information that can be obtained from the conversations' content itself, i.e. without reference to the metadata; any information that can be extracted from the emails' bodies, e.g. by ML, may be added to the TextKG

Both the EmailKG and TextKG contain the scaffolding of the the email corpus: there are temporally ordered conversations which consist of exchanges of temporally ordered emails.

We can then use this information to create a *node classification* task. Specifically, we use heuristics on the EmailKG to label nodes with particular classes of interest. An example is a "major organization" is based on the number of people that belong to them (this is a parameter to the heuristic); for anyone, who isn't part of a major organization we introduce a label. We then cross connect the EmailKG to the TextKG so that we can transfer these labels to the TextKG. Standard approaches to the generation of test, training and validation data splits can then be employed.

3.1.3 Collaboration

We actively collaborated with the Machine Learning Technologies Section of the 711th Human Performance Wing. They have provided testing and integrating our software with datasets of interest. This included monthly meetings and conversations through a shared Slack channel. The feedback has helped us improve the usability and efficacy of the approach. Unfortunately, due to COVID we were not able to physically meet.

3.2 Results and Discussion

Beyond the specification of the methods and procedures above, the project has resulted in the following outcomes.

3.3 System Implementation

The most important outcome is an implementation of the ideas above in a reusable system. The system *conversationkg* that constructs the two types of knowledge graphs above. In addition it, performs named entity recognition, topic modeling and baseline entity resolution on input email corpus. It provides convenient python data structures that can be easily extended with different algorithms. In addition, we support the export of the knowledge graphs to various formats (e.g. JSON, CSV, Neo4j) for further analytics. Included in the system repository are baseline examples of the tasks above as well as sample data. As part of the package, a completely redone web scrapper for the W3C email lists is also provided.

The code and documentation are available at:
`https://github.com/INDElab/conversationkg`
We have archived it using Zenodo [13]

3.3.1 W3C Email Dataset

Using the aforementioned mailing list scrapper, we have scraped the entire W3C email corpus dating back to 1997 and have the data in an easily accessible json form. This includes:

- 1192 mailing lists
- 37,536 time periods
- 705,201 conversations
- 1,876,156 emails.

This can be easily used by *conversationkg* to generate the knowledge graphs above.

3.3.2 Entity Resolution as Author Attribution

We spent significant time investigating the idea of entity resolution through author attribution. We investigated baseline scenarios as well as neural language model (e.g. BERT) based methods for tackling the task. Concretely, we developed the following baselines.

- Jaccard Similarity: arguably a most naive baseline; no learning involved, i.e. no parameters. Given emails e_1 and e_2, simply compute the multiset Jaccard similarity between them.
- Averaged BERT: given a tokenised text, the pre-trained BERT model returns embeddings for each token in the text. This baseline computes the average embedding of the tokens in e_1 and e_2, respectively, and then the cosine similarity of the averages

model	threshold	F1	accuracy	balanced accuracy	precision	recall
Jaccard	0.15	0.69	0.82	0.77	0.72	0.66
Jaccard	0.18	0.67	0.83	0.76	0.81	0.57
avg BERT	0.90	0.63	0.71	0.75	0.51	0.83
avg BERT	0.93	0.63	0.76	0.74	0.59	0.69
BERT+LSTM	0.76	0.43	0.67	0.60	0.44	0.42
BERT+LSTM	0.90	0.31	0.70	0.56	0.49	0.23

Table 1: Performance scores for baseline methods on the task of Entity Resolution for Author Attribution on Emails

.

however: [10] demonstrate that averaged BERT embeddings do not perform better than averaged GloVe embeddings and [14] explain why BERT does not directly allow for language model interpretations. Moreover, initial visualisations of the embedding space suggest that true positives are not likely detected by this model (embeddings of emails from the same author labels are uniformly spread across the embedding space).

- Classifier BERT: instead of averaging, train an LSTM to aggregate the embeddings from BERT into a single vector. The LSTM is explicitly trained s.t. emails from the same label are embedded into vectors with high cosine similarity.

Table 1 shows the results using 100,000 email pairs from the W3C email list. 30% of that data are positive examples.

In general, the results show that Jaccard similarity is a very strong baseline and could actually provide a useful model. The results show that the task is feasible in principle and additional heuristics could further improve performance. The more advanced neural language models (e.g. BERT) did not show improved performance. Indeed, BERT embeddings might perhaps obfuscate more than help in detecting the relevant signal in the data. The notion of author attribution is interesting but given that each author has relatively little source information it might be hard to extract the relevant signal about the source. Indeed, this challenging task could be a useful contribution to extending the state of the art in NLP.

3.3.3 Knowledge Graph Extraction

We are able to extract knowledge graphs useful for analytics. Figure 1 shows a visualization of such an extraction within Neo4J. We are also able to extract the TextKG and EmailKG

variants discussed above. We have developed initial node classification models based on Relational Graph Convolutional Networks but have yet to comprehensively test them.

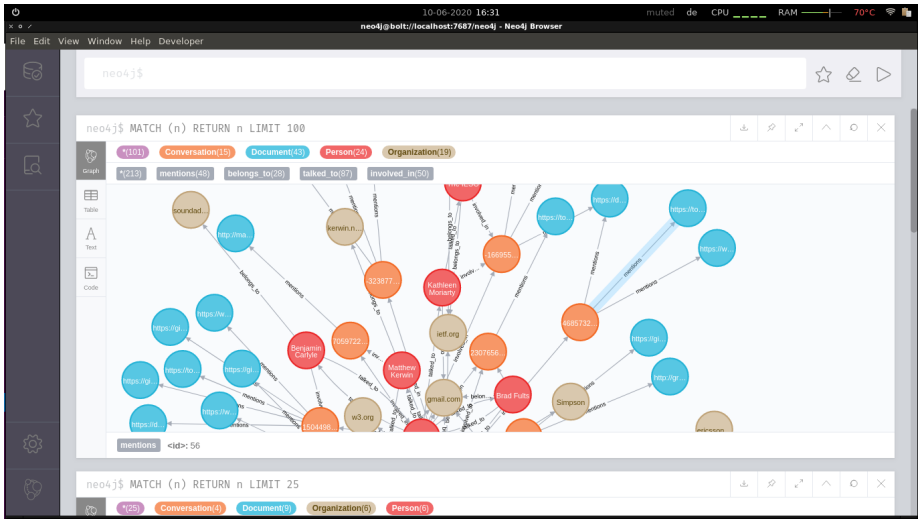

Figure 1: Portion of a knowledge graph extracted from W3C emails in Neo4j

3.3.4 Goals not met

While we were able to transition our technology to AFRL teams including contributing to an invention disclosure (see below), the project did not produce published work. The project instead focused on documenting its results in terms of software and working with our collaborators. The researcher working on the project with the PI transitioned to another job after the completion of this project and was not able to work on finalizing a publication. Additionally, the PI spent time working on the research roadmap going forward.

4 Impacts

Our work has primarily made impact in the principle disciplines in the project specifically in colloboration and forming a basis for subsequent research work in terms of new results and defining a research roadmap.

4.1 Collaboration

Our collaboration with the Machine Learning Technologies Section has produced several results. First, they have been able to use *conversationkg* to extract a knowledge graph

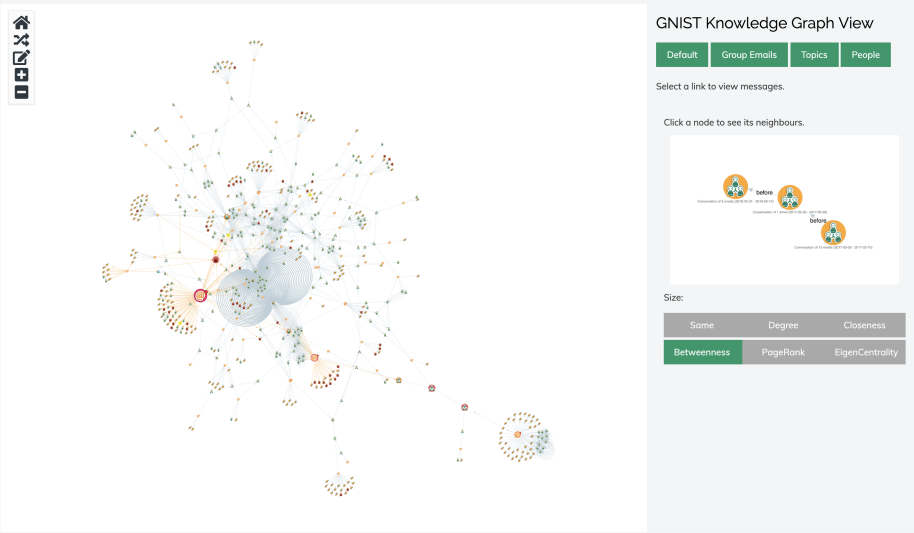

Figure 2: Knowledge Graph in Analytic

from their own email data. The subsequent knowledge graph was then imported into their specialized analytic environment used for analysts. The knowledge graph view provides a different perspective than co-ocurrence based methods. This is shown in Figure 2. Additionally, we contributed to an IP disclosure on forensic pattern discovery made by the section.

4.2 Future Work

The work performed in this project led to subsequent research work within INDE lab, in particular, we have been investigating the challenges of natural language processing tasks. Here, we are also looking at email corpora including the W3C corpora developed in this project. One of the first works based on this investigation was looking at the difficulties of co-reference resolution in the domain of complex conversations, which has been recently published at the 2021 K-Cap conference [6].

As part of the project, we have defined a future research roadmap based on using causality to help preform systematic multimodal systematic data fusion. The idea is to investigate the use of causal theory to provide a strong theoretically grounded approach to integrate different data sources (audio, yet, video); different qualities of data under different assumptions. The project outcomes along with the future roadmap were presented in a meeting with the AFRL stakeholders on January 19, 2021. This research roadmap was articulated in-depth in a follow-on proposal entitled "CausalFusion: Causal Knowledge Extraction and Fusion from Multiple Modalities" (FA8655-22-1-7155). The learnings in this project provided a basis for this roadmap. The datasets and methods explored in this

project will provide a critical foundation for this further investigation.

5 Conclusions

Overall, we made strong steps towards development of a foundation for the extraction of situated dialogues in the form of knowledge graphs. We have explored a number of important tasks and obtained initial results in this challenging domain. We released an open source software package based on these tools. Additionally, we outlined a research roadmap for developing new theory to systematically address key challenges for data fusion of situated multimodal dialogue.

References

[1] April M. Courtice. *Chat Communication in a Command and Control Environment: How Does It Help?* Dissertation, Wright State University, 2015.

[2] Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. An overview of end-to-end entity resolution for big data. *ACM Comput. Surv. 53*, 6 (Dec. 2020).

[3] Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.-H., Vulic, I., and Wen, T.-H. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI* (jul 2019). Data available at github.com/PolyAI-LDN/conversational-datasets.

[4] Klimt, B., and Yang, Y. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning* (2004), Springer, pp. 217–226.

[5] Kummerfeld, J. K., Gouravajhala, S. R., Peper, J. J., Athreya, V., Gunasekara, C., Ganhotra, J., Patel, S. S., Polymenakos, L. C., and Lasecki, W. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 3846–3856.

[6] Li, X., Magliacane, S., and Groth, P. The challenges of cross-document coreference resolution for email. In *Proceedings of the 11th on Knowledge Capture Conference* (New York, NY, USA, 2021), K-CAP '21, Association for Computing Machinery, p. 273–276.

[7] Lowe, R., Pow, N., Serban, I., and Pineau, J. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Prague, Czech Republic, Sept. 2015), Association for Computational Linguistics, pp. 285–294.

[8] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE 104*, 1 (2016), 11–33.

[9] Pujara, J., and Singh, S. Mining Knowledge Graphs From Text. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2018), WSDM '18, ACM, pp. 789–790. event-place: Marina Del Rey, CA, USA.

[10] Reimers, N., and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3982–3992.

[11] Stoyanchev, S., and Johnston, M. Knowledge-Graph Driven Information State Approach to Dialog. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence* (June 2018).

[12] Uthus, D. C., and Aha, D. W. Multiparticipant chat analysis: A survey. *Artificial Intelligence 199-200* (June 2013), 106–121.

[13] Vogelmann, V., Groth, P., and Grier, B. conversationkg, 1 2022.

[14] Wang, C., Li, M., and Smola, A. J. Language models with transformers. *CoRR abs/1904.09408* (2019).

[15] Wilcke, X., Bloem, P., and de Boer, V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Science 1*, 1-2 (Jan. 2017), 39–57.

[16] Wu, Y., and Oard, D. W. Indexing emails and email threads for retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (2005), pp. 665–666.

6 List of Symbols, Abbreviations and Acronyms

- KG - Knowledge Graph
- W3C - World Wide Web Consortium.
- LSTM - Long Short Term Memory
- BERT - Bidirectional Encoder Representations from Transformers
- JSON - JavaScript Object Notation