AFRL-RI-RS-TR-2023-139

NARRATIVE BASED HYPOTHESIS GENERATION USING EVENT COHERENCE AND PROBABILISTIC INFERENCE

UNIVERSITY OF TEXAS AT AUSTIN

JULY 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-139 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
NANCY A. SEPULVEDA
Work Unit Manager

/ S /
MATTHEW J. KOCHAN
Technical Advisor,
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
JULY 2023	FINAL TECHNICAL REPORT	JANUARY 2018	JANUARY 2023

4. TITLE AND SUBTITLE

NARRATIVE BASED HYPOTHESIS GENERATION USING EVENT COHERENCE AND PROBABILISTIC INFERENCE

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-18-2-0017	N/A	62303E

5d. PROJECT NUMBER	5e. TASK NUMBER	5f. WORK UNIT NUMBER
		R2FV

6. AUTHOR(S)

Katrin Erk

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas at Austin
305 E 23rd ST B5100
Austin TX 78712

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)
Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2023-139

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The aim of the DARPA AIDA program was to detect disparate hypotheses about a common claim or question within text and image data, using three domains/scenarios: 1) the war in Ukraine 2013-14, 2) unrest in Venezuela, and 3) claims regarding the COVID virus. The program was separated into three main tasks, where Technical Area 1 (TA1) extracted knowledge graphs from text and images, TA2 fused knowledge graphs using coreference reasoning, and TA3 extracted hypotheses. The UTexas AIDA project was a TA3 project: It detected distinct hypotheses in knowledge graphs extracted and assembled by TA1 and TA2 teams at other sites. Research and development in the UTexas AIDA project resulted in two distinct approaches, one for the Ukraine and Venezuela scenarios, and a separate approach for the COVID scenario.

For the first approach, the central idea was to use narrative coherence as a criterion for assembling hypotheses. We trained neural networks on the task, using for training synthetic data at scale by automatically generating “Story Salads”, mixtures of multiple narratives. In the evaluations, especially the 2020 hackathon, this architecture proved to be simple but effective, and able to quickly adapt to changes in the textual data associated with knowledge graph nodes.

For the COVID scenario the task was to determine compatibility relations between pairs of claims. Here our approach used two types of classifiers working on English raw text data: a relatedness classifier to roughly group claims, and a natural language inference system to determine supporting and refuting relations.

15. SUBJECT TERMS

neural networks, narrative coherence, hypothesis generation, natural language inference

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	27

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANCY A. SEPULVEDA	N/A

PREVIOUS EDITION IS OBSOLETE.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

The aim of the DARPA AIDA program was to detect disparate hypotheses about a common claim or question within text and image data, using three domains/scenarios: 1) the war in Ukraine 2013-14, 2) unrest in Venezuela, and 3) claims regarding the COVID virus. The program was separated into three main tasks, where Technical Area 1 (TA1) extracted knowledge graphs from text and images, TA2 fused knowledge graphs using coreference reasoning, and TA3 extracted hypotheses. The UTexas AIDA project was a TA3 project: It detected distinct hypotheses in knowledge graphs extracted and assembled by TA1 and TA2 teams at other sites. Research and development in the UTexas AIDA project resulted in two distinct approaches, one for the Ukraine and Venezuela scenarios, and a separate approach for the COVID scenario.

For the first approach, the central idea was to use narrative coherence as a criterion for assembling hypotheses. We used a neural network for the task of hypothesis construction; to obtain large amounts of train-ing data, we automatically constructed synthetic "Story Salads", mixtures of multiple narratives. In the evaluations, especially the 2020 hackathon, this architecture proved to be simple but effective, and able to quickly adapt to changes in the textual data associated with knowledge graph nodes.

For the COVID scenario the task was to determine compatibility relations between pairs of claims. Here our approach used two types of classifiers working on English raw text data: a relatedness classifier to roughly group claims, and a natural language inference system to determine supporting and refuting relations.

2.0 INTRODUCTION

The aim of the DARPA AIDA program was to detect disparate hypotheses about a common claim or question. For example, when Malaysia Airline Flight 17 was destroyed in July 2014, there were initially multiple hypotheses about what happened: 1) it was an accident, 2) there was a bomb on board, or 3) the plane was mistaken for a military plane and shot down. The latter hypothesis turned out to be the correct one, but the aim of AIDA was not to check facts but to separate out the different hypotheses and their associated evidence in large amounts of textual and image data.

The AIDA program separated the task into three phases. TA1 extracted information from text and images into knowledge graphs. TA2 fused knowledge graphs based on coreference. TA3, finally, extracted disparate hypotheses from the resulting knowledge graph. The UTexas AIDA project was a TA3 project, which extracted hypotheses from knowledge graphs extracted and assembled at other sites.

The main idea of the UTexas AIDA project was to treat the task as relying on script knowledge [1]. Scripts, also called narrative schemas or situation knowledge, constitute knowledge about typical events, both at the level of individual events with their roles and participants, and at the level of longer event sequences [2]. Script knowledge can be used for inference, for example when listeners infer steps in a longer event sequence that are not stated explicitly [1]. Information about scripts, or narrative schemas, can be learned from large text collections [3,4] using many different machine learning frameworks [5,6,7].

In the UTexas AIDA project, we focused on hypothesis expansion. We first determined "hypothesis seeds." These are minimal answers to the question of what happened. For the Malaysia Airline example from above, one such hypothesis seed could be "militants attacked Flight 17." We then add more pieces to the hypothesis seed to flesh out the story. In our example, we ideally would want to add: the militants were affiliated with Russia; they were located in Ukraine, fighting Ukrainian forces; they mistook the plane for a military plane. We trained a system to expand a hypothesis seed by statements that form a coherent narrative with the hypothesis seed. The idea was to use narrative reasoning to identify elements that are mutually compatible and that “make sense together.” Hypothesis expansion draws on event coherence prediction, a task for which neural models are eminently suited, but also needs to obey logical and ontological constraints, which can be best enforced by symbolic inference. Accordingly, the project used a mixture of machine learning and rule-based inference. For narrative reasoning, we used a neural network operating over a knowledge graph [8]. Logical constraints on the co-occurrence of knowledge elements were formulated as symbolic rules, to be applied during and after neural inference. In this approach, we treated the TA3 task not as a general-purpose reasoning task but as first and foremost a clustering task enhanced by inference.

The DARPA AIDA program used three scenarios for evaluation: 1) the war in Ukraine 2013-14, 2) unrest in Venezuela, and 3) claims regarding the COVID virus. For the first two scenarios, hypotheses consisted of a central event with one or more queried roles, for example the event of

the destruction of Malaysia Airline Flight 17, where the queried role would be the Attacker of an attack event, or the Cause in case it was an accident event. For these two scenarios, the hypothesis expansion approach was appropriate, as there were usually additional knowledge elements relevant to each hypothesis that were present in the graph.

For the third scenario, COVID, the hypotheses were claims that consisted of a single sentence each, without additional knowledge elements to be added. For this scenario, the evaluation task was also different, with a focus on predicting relations of "supporting" and "refuting" between pairs of claims. For this scenario, we therefore designed a separate system that relied on an interplay of two types of classifiers: a relatedness classifier, and a system for natural language inference.

In the remainder of this report, we describe the methods used and approaches developed within the UTexas AIDA project (Sec. 3), and present and discuss results (Sec. 4).

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

In the UTexas AIDA project, we pursued two approaches. The first combined knowledge elements into hypotheses based on narrative coherence. The second approach used relatedness classifiers and natural language inference (NLI) to judge relations between claims.

The central idea of the first approach was to use narrative coherence as a criterion for assembling hypotheses: We aimed to select, as parts of a common hypothesis, all elements that are mutually compatible and that "make sense together", and that fit a common narrative. We used neural networks for predicting narrative coherence, paired with symbolic methods for checking ontological and logical consistency. For narrative coherence, we first explored text-based methods (Sec. 3.1) as a precursor to the graph-based methods we used in later evaluations (Sec. 3.3). For logical and ontological consistency reasoning, we experimented with probabilistic inference for logical consistency (Sec. 3.2) before settling on to non-probabilistic symbolic inference (Sec. 3.3).

We report on the second approach in Sec. 3.4.

Within the scope of the AIDA project, we also explored how to represent entities for better inferences, in both short and very long texts (Sec. 3.5). When testing natural language inference systems for our second approach, we confirmed previous findings that natural language inference systems were bad at handling negation, and found that the AIDA data had a high number of negations; for that reason, we explored how to improve handling negation in NLI (Sec. 3.6).

3.1 Story salads and narrative coherence reasoning with plain text data

As mentioned above, we wanted to use neural networks to predict narrative coherence between elements that might or might not belong with the same hypothesis. However, neural networks need large amounts of training data. There is not a large dataset of naturally occurring data on which to learn narrative coherence. So our project centrally relied on synthetic data to train our models. Here, the main idea was to create "story salads", artificial mixtures of texts, where of course we know which original text each passage is from. The task, then, is to sort apart a "story salad" into the original texts.

To explore the feasibility of the task, we first used "story salads" made of raw text in English. In [9], we created synthetic texts that were mixtures of different articles and used neural supervised clustering to separate the "story salads". We tested mixtures of arbitrary Wikipedia or New York Times articles, as shown in Figure 1, and also mixtures made of closely related articles. For Wikipedia, we made use of the Wikipedia hierarchy to determine relatedness. Figure 2 shows an example. As can be seen, this mixture is much more challenging, and provides some interesting inference problems, for example it would be helpful to infer that a person born in 1855 would not be in high school in 1913. In [10] we again address the task of raw-text "story salads", this time formulating the task as one-class clustering to obtain a stronger model. Here, the idea is that we have a sentence that constitutes a starting point, a "story seed", and the model iteratively extends this story seed with additional sentences that fit with the story so far.

(A) Some of the prisoners were survivors of the Battle of Qala-i-Jangi in Mazar-i-Sharif. (A) Chechnya came under the influence of warlords. (B) The U.S. invaded Afghanistan the same year when several Taliban prisoners were shot. (A) Russian federal troops entered Chechnya and ended its independence. (A) The Russian casualties included at least two commandos killed and 11 wounded. (B) The dead were buried in the same grave under the authority of Commander Kamal.

Figure 1: Example of a "story salad". Sentences labeled (A) are from the first original text, sentences labeled (B) from the second.

(A) John K. Randle was born on 1 February 1855, son of Thomas Randle, a liberated slave from an Oyo village in the west of what is now Nigeria. (B) In 1913 he was enrolled in Eko Boys High School but dropped out. (B) Martins joined the theatre and from there took on various theatre jobs to survive. (A) Born in Sierra Leone , he was one of the first West Africans to qualify as a doctor in the United Kingdom. (B) He also worked as a wrestler (known as "Black Butcher Johnson").

Figure 2: Example of a "story salad" from related documents belonging to the Wikipedia category "Nigerian people of World War I". Sentences labeled (A) are from the first original text, sentences labeled (B) from the second.

3.2 Probabilistic inference for logical and ontological consistency

Our original plan had been to interleave two models for an incremental construction of hypotheses: 1) a neural model that would add knowledge elements based on narrative consistency, and 2) a probabilistic reasoning system that would filter incremental hypotheses based on logical and ontological constraints. The idea was to use a Sequential Monte Carlo (particle filter) model for the probabilistic reasoning, implemented in the probabilistic programming language WebPPL [11]. However, we found that WebPPL was not able to handle the very large knowledge graphs of AIDA, and its successor Pyro [12], while able to handle much larger amounts of data, only had very limited support for categorical data. Accordingly, we switched to a non-probabilistic implementation of logical rules.

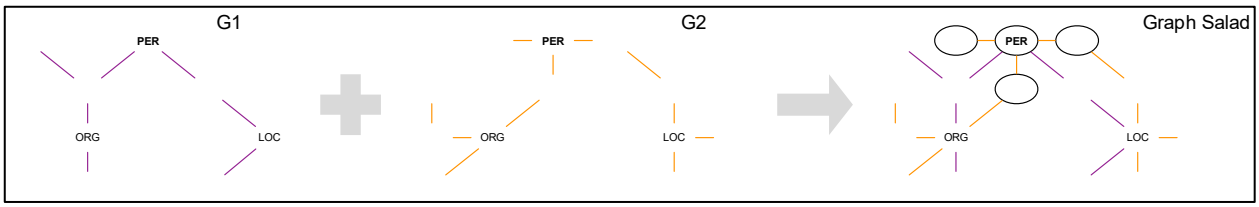

Figure 3: Illustration of how two knowledge graphs are fused into one by artificially merging nodes. Here: a Person node of the purple graph has been merged with a Person node in the orange graph, and likewise an Organization node with an Organization node, and a Location with a Location.

3.3 Story salads and narrative coherence reasoning on knowledge graphs

While our first experiments with sorting apart "story salads" were on plain text, the challenge in AIDA is to work with a knowledge graph that is extracted from multiple texts and images, where the texts may use different source languages. Accordingly, we reformulate the story salad task to work on knowledge graphs instead of raw text. This work is explored in a masters thesis [13] and published in [14].

For training, we again create synthetic data using Wikipedia, but we transform Wikipedia articles to knowledge graphs using the TA1 tool of the GAIA team. We then create story salads by artificially fusing entity or event nodes from different knowledge graphs, as illustrated in Figure 3. The artificial merging preserves the knowledge graph type, so that for example a person is merged with a person, a location with a location. This creates artificial disparate hypotheses, for example it might create a person that is a merge of Napoleon and Putin, who was involved in wars in both the 19th and the 21st century. To best match the domain of test documents, we use the Wikipedia hierarchy to identify only Wikipedia articles about armed conflicts, obtaining about 70k articles. Figure 4 shows a snapshot from an actual graph story salad, as visualized by our in-house visualization tool.

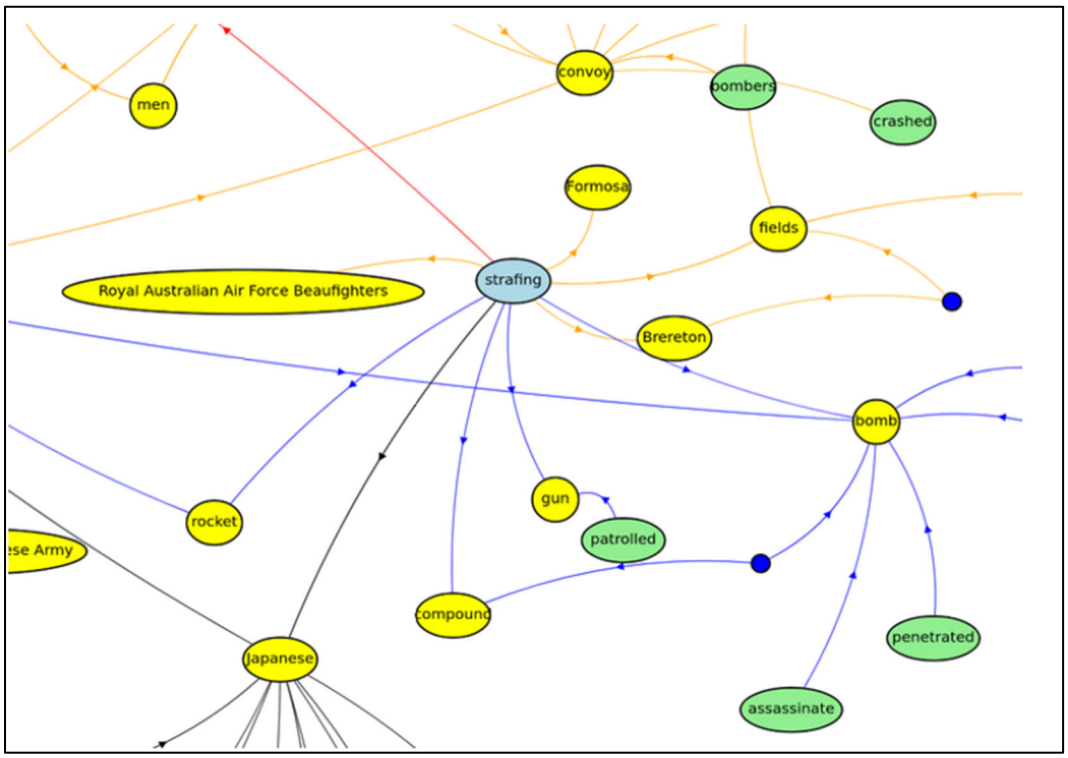

Figure 4: Part of a graph "story salad" as visualized by our visualization tool. Black edges are from the target knowledge graph, blue and orange edges are from distractor knowledge graphs, and red edges are part of target knowledge graph and the "hypothesis seed".

For narrative inference over the resulting knowledge graphs, we use graph convolutional networks [8]. Each node in the graph is associated with an embedding. Initially, this embedding is based on the knowledge graph label of the node, as well as any textual information associated with the node. In subsequent graph convolution steps, the embedding of each node is updated to include information from its neighbors, where this update is learned to best support the eventual classification task. In our case, the classification task is: Given a subgraph that has so far identified as belonging to a single hypothesis, and given an edge connecting the subgraph to the rest of the graph, should this edge be added to the hypothesis-so-far or not? Figure 5 illustrates how the model worked, where the purple nodes are the nodes selected so far as being part of the hypothesis, and the orange nodes are the nodes directly adjacent to the current hypothesis. The next step of the algorithm will decide which of the orange nodes to add to the hypothesis. This approach builds on the one-class clustering setting of [10].

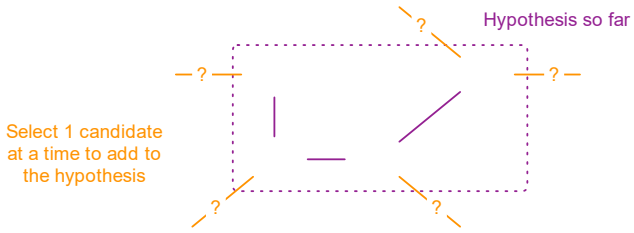

Figure 5: Snapshot of the iterative hypothesis expansion within the graph convolutional network

The graph convolutional network forms part of the overall UTexas AIDA pipeline, which is illustrated in Figure 6. The input to the system consists of a knowledge graph in AIDA Interchange Format (AIF) along with statements of information need (SINs). The first component in the pipeline, Entry Point Matching, identifies subgraphs in the knowledge graph that match each SIN. These form the hypothesis seeds. After the graph is converted to an internal JSON format, the graph convolutional network expands on each hypothesis seed. The hypotheses created by the network are then filtered using symbolic rules, and converted to AIF for output.

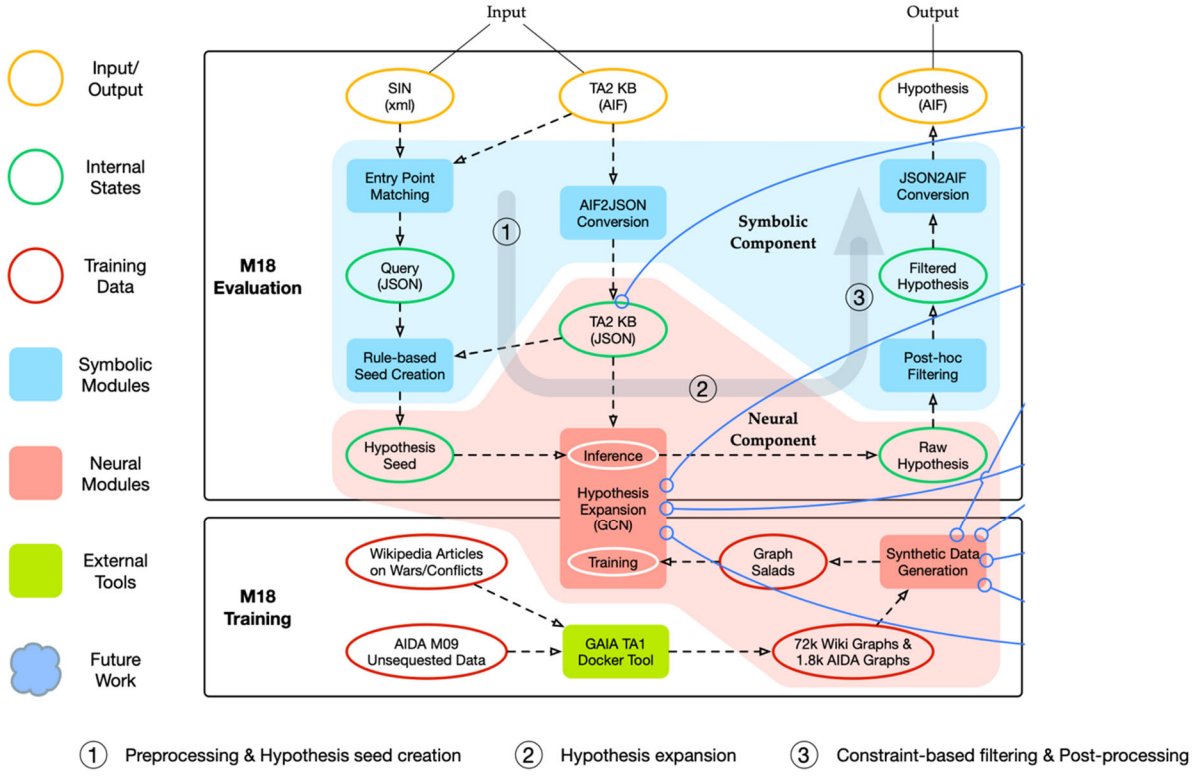

Figure 6: System overview. The parts with a blue-shaded background are symbolic, the parts with a red-shaded background are neural.

The pipeline shown in Figure 6 is the pipeline from the Month 18 evaluation. For subsequent evaluations, we improved the neural model, and we expanded the post-neural processing by a redundancy module to detect duplicate hypotheses. We also interleaved neural hypothesis expansion with logical rules, such that inconsistent hypotheses could be removed early.

Originally, TA1, TA2, and TA3 had been separated in such a way that TA3 was not able to see any plain text, only TA1 could see it, and the only text information except from knowledge graph ontology labels that were passed on to TA3 were named entity labels. During the Winter 2020 hackathon for the Venezuela domain, this "membrane" was removed, and TA1 was able to pass on plain text information associated with knowledge graph nodes to TA3. At this point, we changed our system to make use of mention strings of arbitrary length in neural hypothesis expansion. We also improved the weighting of entry points (hypothesis seeds), implemented a rule-based query expansion module that reformulated Statements of Information Need, and worked on reducing redundancy of the generated hypotheses. We also implemented a system for an in-depth qualitative analysis of extracted hypotheses to better identify strengths and weaknesses of our system.

3.4 Predicting claim relations with relatedness classifiers and natural language inference

While the two first AIDA scenarios comprised narratives involving multiple events and participants, this was not the case for the COVID scenario, where the focus was on individual claims. In addition to the change in data, the task also changed. In the last AIDA evaluation, the focus was on matching claims to overarching topics/questions, and determining relations of either "supporting", "refuting" or "neutral" between pairs of claims. For this changed task, we designed a separate system that relied on an interplay of two types of classifiers: a relatedness classifier, and a natural language inference system. Claims in the COVID domain were given both as raw text and in graph format; in order to quickly ramp up to a new system, we used claims in raw format, and tested whether deep learning systems would be able to detect Supporting and Refuting relations among claims given as raw text.

Our system had two main components. The first was a relatedness detection component, which we used both to test whether a claim was related to a question, and to test whether two claims addressed a common question. This system was trained on a paraphrase detection task, where we found that lowering the threshold for a positive answer allowed the system to also detect claims that were related rather than actual paraphrases. The second component was a Natural Language Inference (NLI) system for determining, for a given claim pair, whether they were supporting, refuting, or neutral.

Here is the pipeline for Condition 5, where each claim was labeled with a query (question) to which it pertains. The pipeline for the other two conditions was similar:

1. Determine Query-Claim relatedness: **Relatedness classifier**
   Restrict to claims with same topic and subtopic as query
   - Relatedness scorer: threshold set manually on dry-run data
2. For each related query-claim pair: judge supporting/refuting/neither using an **NLI system**:
   - entails = supporting
   - contradicts = refuting

    - neutral = related
3. For any two claims related to the same query: use **Relatedness classifier** to determine similarity
4. Rank claims for each query:
    - Starting claim: most related to query
    - Subsequent claims sorted by maximum dis-similarity to previous claims

3.5 Representing entities for narrative reasoning

Within the AIDA project, we also explored how to represent information about entities in narrative schemas for improved inference. In [15,16], we address the task of retrieving implicit arguments, arguments of semantic roles that are not specified explicitly in the text but that can be inferred. An example is: "More than 2,600 people have been infected by Ebola in Liberia, Guinea, Sierra Leone and Nigeria since the outbreak began in December, according to the World Health Organization. Nearly 1,500 have died." Here, the Disease argument of "outbreak" is implicit, as is the Reason argument of "died". In [17], we explored different methods for injecting information about coreference into neural models to improve their reasoning. In [18], we explored entity representations over very long stretches of text, specifically literary works.

3.6 Improving negation handling in natural language inference

We observed that claims in the COVID domain showed a high incidence of negation, and that NLI systems did badly on data points involving negation. To address this problem, we did a study to improve the handling of negation in NLI. We followed the hypothesis that standard pretraining did not give large language models enough information on the inference potential of negation, that is, on the impact of negation on the appearance of other words in the context. We used ELECTRA, a large language model that, in addition to the masked prediction task, includes a discrimination task of distinguishing words from the original text from words supplied by masked prediction. We used this second, original/replaced prediction task for additional pretraining with negation examples where the words to be predicted as original/replaced were in the negation focus. We found that (a) existing NLI datasets are very different in the amount of negation they contain, and consequently in the usefulness of additional negation-related pretraining, that (b) our approach is successful in improving the distinction of entailment from contradiction. This work is currently under review.

4.0 RESULTS AND DISCUSSION

In this section, we report on our results and discuss them. Sections 4.1 and 4.2 discuss results of our narrative coherence approaches, where Sec. 4.1 is about text-based and 4.2 about graph-based inference. Section 4.2 also reports on results of AIDA evaluations that use graph convolution for hypothesis expansion. Section 4.3 reports on results of our second approach, which combines relatedness classifiers with natural language inference classifiers for claim relations. Section 4.4 discusses results of our approaches towards improved entity representations, and Sec. 4.5 has results on improved negation handling in natural language inference. Sections 4.6 and 4.7 report on software created in the project and on students supported in the project.

4.1 Results and discussion: Story salads and narrative coherence reasoning with plain text data

In [9], as discussed above, we first explored the "story salad" and hypothesis expansion task for plain-text data. We created synthetic texts that were mixtures of pairs of different articles, either random articles or closely related articles, and used neural supervised clustering to separate the "story salads". Performance was measured in terms of clustering accuracy, the ratio of correctly clustered sentences in a document mixture, averaged over test mixtures. For random documents, we obtained **clustering accuracie**s of **84.9** on New York Times and **81.8** on Wikipedia articles, and for mixtures of closely related documents, we obtained clustering accuracies of **68.0** on New York Times articles and **66.6** on Wikipedia data.

In [10], as discussed above, we again addressed the task of raw-text "story salads", using stronger neural models in a one-class clustering setting. In this paper, we tested stories with either one or multiple random distractor narratives mixed in. On mixtures of New York Times articles, we obtained an **F-score** of **0.92** for mixtures with a single distractor narrative, and **0.84** with four distractor narratives in a task setting when the number of statements to extract was fixed. When the model was also tasked in deciding how many statements to extract, F-scores were **0.73** (single distractor) and **0.65** (four distractors). We also collected, from crowd workers, 100 topics with two disparate hypotheses each (Human100). On the Human100 data, **F-scores** ranged from **0.56** to **0.62** depending on the synthetic training data used. (Human performance at the task was at an F-score of 0.82.)

Discussion. The results of both [9] and [10] show that the task is feasible with neural architectures that learned narrative coherence in a task-based manner. The results of [9] also show that mixtures of closely related articles indeed pose more challenging inference problems. In [10] we see that the stronger model does improve performance. We also see a clear difference in performance between New York Times article mixtures and the Human100. One likely contributor to this difference in performance is data sparseness: The Human100 hypotheses, being shorter texts, are sparser than the New York Times articles in terms of the information they provide about entities and events.

4.2 Results and discussion: Story salads and narrative coherence reasoning with knowledge graphs

For the knowledge graph version of the hypothesis expansion task, [13] evaluates on synthetic data created by mixing 3 knowledge graphs derived from 3 Wikipedia articles. To connect the knowledge graphs, 3 merge points (artificially merged nodes) were created, always merging nodes of the same ontological type. On this data, he reports an **average precision of 0.86** for knowledge graphs mixing one narrative with 2 distractors, for a setting where the 3 merge points were mutually reachable. On the subset of data where the merge points were not mutually reachable, average precision was **0.79**. (All reported numbers are from the best of several attention variants tested).

The thesis [13] also explored the use of Reinforcement Learning to better explore longer paths in the graph, but was not able to obtain better results. It also explored the use of contextualized embeddings (ELMO) to initialize node embeddings, again without improvement. The main problem for this setting was to create training data in a way that prevented "bleed over" between nodes: If the contextualized embedding of one node was able to "see" the text that forms the contextualized embedding of another node, that would be a giveaway that the two nodes stem from the same original knowledge graph.

Table 1: Results, 2019 TAC SM-KBP task. These are results on the Ukraine domain.

	GAIA_1-OPERA_3. Colorado_1. UTexas_2	OPERA_4. Colorado_1. UTexas_2	OPERA_TA1a_aditi_V5. OPERA_TA2_aditi_V5. UTexas_2	LDC_2. LDC_2. UTexas_3
edge correctness	0.253	0.3139	0.3655	0.8017
edge coherence	0.3607	0.3691	0.4475	0.8612
KE coherence	0.6108	0.6012	0.5851	0.8979
KE relevance (strict)	0.1566	0.2763	0.3295	0.8312
KE relevance (lenient)	0.6552	0.4543	0.4836	0.9034
argument coverage	0	0.0031	0.0032	0.01

Table 2: Evaluation results of the UTexas TA3 system, with different TA1 and TA2 knowledge bases, Venezuela scenario, Hackathon Fall 2020. Lines highlighted in yellow show post-hackathon runs. Lines not highlighted are pre-hackathon runs.

SIN	Prec	Recall	F1	Run
E201	0.6000	1.0000	0.7500	OPERAGAIA_TA1_CO_TA2_UTexas_TA3_052
E201	0.4286	0.5000	0.4615	GAIA_TA1_CO_TA2_UTexas_TA3_052
E201	0.0833	0.6667	0.1481	GAIA_TA1tv_OPERA_TA2_UTexas_TA3_046
E201	0.0000	0.0000	0.0000	OPERA_TA1_COLORADO_TA2_UTexas_TA3_046
E202C	0.2500	0.5000	0.3333	GAIA_TA1_CO_TA2_UTexas_TA3_052
E202C	0.1429	1.0000	0.2500	OPERAGAIA_TA1_CO_TA2_UTexas_TA3_052
E202C	0.0000	0.0000	0.0000	OPERA_TA1_COLORADO_TA2_UTexas_TA3_046
E202C	0.0000	0.0000	0.0000	GAIA_TA1tv_OPERA_TA2_UTexas_TA3_046
E203	0.4286	1.0000	0.6000	OPERAGAIA_TA1_CO_TA2_UTexas_TA3_052
E203	0.3333	0.6667	0.4444	GAIA_TA1_CO_TA2_UTexas_TA3_052
E203	0.1304	1.0000	0.2308	OPERA_TA1_COLORADO_TA2_UTexas_TA3_046
E203	0.0400	0.6667	0.0755	GAIA_TA1tv_OPERA_TA2_UTexas_TA3_046

We now turn to evaluations on actual DARPA AIDA data. Table 1 shows results from the SM-KBP task at the 2019 TAC conference. These were results on the first scenario (Ukraine). Results on the second scenario, Venezuela, are shown in Table 2. These are results from the 2020 Hackathon, where the lines with yellow highlights are post-hackathon, and the non-highlighted lines are pre-hackathon. As mentioned before, the task and data were changed at that time in that the "membrane" between TA1, TA2 and TA3 was removed, and TA3 was able to see plain text information associated with nodes in the knowledge graph. The results in Table 2 reflect this change in task and data, and subsequent changes that were made to the UTexas system.

Discussion. The results of our system on knowledge graphs have improved drastically over the course of the AIDA program. However, it is possible that the biggest improvement came not from changes to the model but changes to the data, in particular the removal of the "membrane", which made the knowledge graphs much more informative. Interestingly, while we adapted the preprocessing during the 2020 hackathon, the model was not retrained. It was able to take advantage of the additional textual information attached to the nodes simply because all embeddings were situated in the same semantic space.

Comparing results across knowledge graphs in Table 2, we see that richer knowledge graphs, combined across two TA1 sources, generally led to better performance. Another thing to note in Table 2 is that model performance differed drastically across the three different hypothesis complexes, with F-scores on joint GAIA/Opera knowledge graphs of 0.75, 0.25, and 0.6, respectively; the reason for this drastic variation in performance is not quite clear, but based on our qualitative analysis it looks like it might stem from differences in sparseness in information about the core entities in the different hypothesis complexes.

Continuing on the theme that it is important for the information not to be sparse, we again see better performance on the synthetic knowledge graph "salads," which tend to be less sparse than the AIDA evaluation knowledge graphs.

4.3 Results and discussion: Predicting claim relations with relatedness classifiers and natural language inference

We now turn to results on the third domain, COVID-related claims. As discussed in Section 3, we built a separate system for this domain, combining relatedness classifiers and natural language inference (NLI).

For NLI, we tested state-of-the-art systems based on both RoBERTa and DeBERTa using a small dataset of hand-annotated claims. We found that both systems over-assigned the Neutral category, but RoBERTa showed overall better performance. We also tested training data in either a general news domain or the biomedical domain, hypothesizing that the latter would be more appropriate for COVID-related claims. However, systems trained on an available biomedical dataset exhibited a highly problematic bias towards prediction of Entailment, so that we stayed with the general news domain.

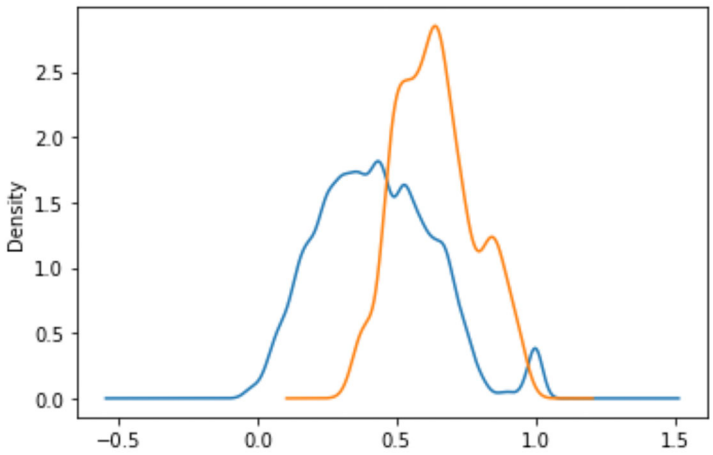

Figure 7: Using a relatedness classifier to determine whether two claims are the same (redundant): Density plot for relatedness values, blue = different claims, orange = same claim.

For detecting redundancy among extracted claims, we again used the relatedness classifier. Figure 7 illustrates that relatedness values are higher for pairs of claims that are mutually redundant (orange) than for pairs of claims that are not (blue), but as can be seen, overlap of the scores for same-claim pairs and not-same-claims is high.

Overall, we found the COVID domain to be challenging. Our system achieved high precision but very low recall, for an overall very low **F-score** of around **0.01 in condition 5, 0.06-0.12 in condition 6.**

Discussion. Due to the very short development time for this second system, there is not much that can be learned from the results. However, I think we can say that current NLI systems are not well suited for detecting compatibility among claims. A dedicated system considering similarity and compatibility of particular roles/participant slots in the claim would likely do better.

4.4 Results: Representing entities for narrative reasoning

In [15], we studied implicit arguments, finding that it is possible to train models for implicit argument prediction on a simple cloze task, for which data can be generated automatically at scale. A neural model trained on this data, and constructed to use narrative coherence and entity salience for predictions, obtained an accuracy of 47.8 on (length-restricted) OntoNotes, and an F-score of 49.6 on the Gerber&Chai data, an existing set of 960 human-annotated implicit argument sentences. In [16] we use a different neural model, a pointer network, on the task, improving performance to an accuracy of 58.12 on OntoNotes, while performance on Gerber&Chai stayed the same.

In [17], we consider natural language understanding, specifically the LAMBADA dataset, which has the masked word prediction task but focuses on cases that are easy for humans but hard for language models; often, the datapoints involve complex coreference problems with many possible antecedents. We add supervised attention to the BiDAF natural language understanding model to nudge the model to attend to entities across all their mentions. This leads to a performance on the LAMBADA dataset on par with the performance of GPT-2, but with an order of magnitude fewer parameters.

In [18], we learn entity representations over the course of very long texts, namely novels. We use a "masked entity prediction" task in which the mention of a character is masked, and the model must guess which it is. We find that the model is able to do the task reasonably well, with an accuracy of 74%. The resulting entity embeddings contain some information about the characters in question: They can be used to predict the character's gender with an accuracy of 72%, and they can in some cases be used to recognize the characters from character descriptions in Wikipedia. This is a first step in learning in-depth character representations for very long texts. In follow-up work, we have been studying entity embeddings based on what characters say rather than what they do, finding this to lead to more informative embeddings.

4.5 Results: Improving negation handling in natural language inference

As mentioned above, we have explored continued pre-training in ELECTRA to teach a language model about the inference potentials of negation, with the aim of improving NLI examples involving negation. We find that continuing to pretrain ELECTRA-Small's discriminator leads to substantial gains on a variant of RTE (Recognizing Textual Entailment) with additional negation. On SNLI (Stanford NLI), there are no gains due to the extreme underrepresentation of negation in the data. Finally, on MNLI (Multi-NLI), we find that performance on negation cues is primarily stymied by neutral-labeled examples.

4.6 Software

Software for the system developed for the Ukraine and Venezuela scenarios was submitted to the DARPA AIDA Integration Contractor, CACI.

A Docker of the COVID Scenario system has been made available at https://hub.docker.com/repository/docker/katrinerk/utexas_aida/general

4.7 Students supported

This project has supported 14 students, 11 of which have since graduated. Of the 14 students, 11 were graduate students and 3 were undergraduates. Of the graduate students previously supported by this project, one has joined Google Research, two have joined Bloomberg (one of them Bloomberg Research), one went to INDEED, and one joined Microsoft as a software engineer.

5.0 CONCLUSIONS

Within the DARPA AIDA program, the UTexas TA3 team has pursued two approaches: 1) an approach to hypothesis construction that focused on narrative coherence, and 2) an approach to judging relations between claims based on natural language inference and relatedness classifiers. Due to the large difference in development times, conclusions can be drawn only about the first approach. Of this approach, we learn that it is feasible to treat narrative coherence-based inference as a learning problem, and that it is possible to learn narrative coherence from data. In our case, we focused on synthetic training data in the form of "story salads". We found this to be a useful form of data that allowed the model to learn narrative coherence; it is also good that this synthetic data can be calibrated in its difference based on how many texts are mixed, and how similar the underlying texts are that form part of the "salad". In the case of knowledge graph "story salads", data can also be calibrated by how many artificially merged nodes are constructed.

In AIDA evaluations, we found this machine learning based setting to perform relatively well, and in particular we found that it adapted well to changes in how much textual data was associated with graph nodes, without re-training the model. However, the approach seems to be sensitive to how sparse the data is, with dramatically better performance if more information is given about the entities and events in the narrative.

The problem of distinguishing disparate hypotheses or claims remains a relevant one; it shows up again for example in fact checking and argumentation analysis. However, even though the synthetic "story salad" data is good, it would be good for development on the task to have datasets with naturally occurring data.

6.0 REFERENCES

[1] R. Schank and R. Abelson. *Scripts, plans, goals, and understanding*. Psychology Press, 1977.

[2] A. Sandford and S. Garrod. *The role of scenario mapping in text comprehension*. Discourse Processes 6:2-3, 159-190. 1998

[3] N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In Proceedings of ACL, 2008.

[4] N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In Proceedings of ACL, 2009.

[5] K. Pichotta and R. Mooney. Learning statistical scripts with LSTM recurrent neural networks. In Proceedings of AAAI, 2016.

[6] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. Voss. Connecting the Dots: Event Graph Schema Induction with Path Language Modeling. In Proceedings of EMNLP, 2020.

[7] N. Weber, L. Shekhar, H. Kwon, N. Balasubramanian, and N. Chambers. Generating Narrative Text in a Switching Dynamical System. In Proceedings of CoNLL, 2020.

[8] T. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In Proceedings of ICLR, 2017.

[9] S. Wang, E. Holgate, G. Durrett, and K. Erk. Picking Apart Story Salads. In Proceedings of EMNLP 2018.

[10] S. Wang, G. Durrett, and K. Erk. Query-Focused Scenario Construction. In Proceedings of EMNLP 2019.

[11] N. Goodman and A. Stuhlmüller. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org, 2014.

[12] E. Bingham, J. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall and N. Goodman. Pyro: Deep Universal Probabilistic Programming. arXiv 1810.09538, 2018

[13] A. Tomkovich. Graph-based narrative coherence. Masters thesis, University of Texas at Austin.

[14] P. Cheng, A. Tomkovich, E. Holgate, S. Wang, and K. Erk. The UTexas system for TAC 2019 SM-KBP Task 3: Hypothesis detection with graph convolutional networks. In Proceedings of the 2019 Text Analysis Conference (TAC).

[15] P. Cheng and K. Erk. Implicit argument prediction with event knowledge. Proceedings of NAACL 2018.

[16] P. Cheng and K. Erk. Implicit Argument Prediction as Reading Comprehension. Proceedings of AAAI 2019.

[17] P. Cheng and K. Erk. Attending to entities for better text understanding. Proceedings of AAAI 2020.

[18] E. Holgate and K. Erk. "Politeness, you simpleton!" retorted [MASK]: Masked prediction of literary characters. Proceedings of IWCS 2021.

APPENDIX A – PUBLICATIONS AND PRESENTATIONS

Publications

Su Wang, Eric Holgate, Greg Durrett, and Katrin Erk. Picking Apart Story Salads. Proceedings of EMNLP 2018.

Pengxiang Cheng and Katrin Erk. Implicit argument prediction with event knowledge. Proceedings of NAACL 2018.

Pengxiang Cheng, Alexander Tomkovich, Eric Holgate, Su Wang, and Katrin Erk. The UTexas system for TAC 2019 SM-KBP Task 3: Hypothesis detection with graph convolutional networks. Proceedings of the 2019 Text Analysis Conference (TAC).

Pengxiang Cheng and Katrin Erk. Implicit Argument Prediction as Reading Comprehension. Proceedings of AAAI 2019.

Su Wang, Greg Durrett, Katrin Erk. Query-Focused Scenario Construction. Proceedings of EMNLP 2019.

Pengxiang Cheng and Katrin Erk. Attending to entities for better text understanding. Proceedings of AAAI 2020.

Eric Holgate and Katrin Erk. "Politeness, you simpleton!" retorted [MASK]: Masked prediction of literary characters. Proceedings of IWCS 2021.

Gunjan Bhattarai and Katrin Erk. To Learn or Not to Learn: Exploring the effectiveness of replaced token detection on learning the meaning of negation cues. Conference paper, currently under review.

Presentations

conference paper presentations at:

- EMNLP 2018
- NAACL 2018
- AAAI 2019
- EMNLLP 2019
- AAAI 2020
- IWCS 2021

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

DOD	Department of Defense
NLI	Natural Language Inference
AIDA	Active Interpretation of Disparate Alternatives
SIN	Statement of Information Need
DARPA	Defense Advanced Research Projects Agency
COVID	Corona Virus Disease
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT
DeBERTa	Decoding-enhanced BERT with disentangled attention
ELMO	Embeddings from Language Model
SM-KBP	Streaming Multimedia Knowledge Base Population
TAC	Text Analysis Conference
LAMBADA	Language Modeling Broadened to Account for Discourse Aspects
BiDAF	Bi-Directional Attention Flow
GPT	Generative Pre-trained Transformer
RTE	Recognizing Textual Entailment
SNLI	Stanford NLI
MNLI	Multi-NLI