



AFRL-RI-RS-TR-2023-215

GAIA: GENERATING ALTERNATIVES FOR INTERPRETATION AND ANALYSIS

UNIVERSITY OF SOUTHERN CALIFORNIA INFORMATION
SCIENCES INSTITUTE

NOVEMBER 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

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-2023-215 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 /

SCOTT D. PATRICK
Deputy Chief
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	
NOVEMBER 2023		FINAL TECHNICAL REPORT		START DATE	END DATE
				JANUARY 2018	JUNE 2023
4. TITLE AND SUBTITLE					
GAIA: GENERATING ALTERNATIVES FOR INTERPRETATION AND ANALYSIS					
5a. CONTRACT NUMBER		5b. GRANT NUMBER		5c. PROGRAM ELEMENT NUMBER	
FA8750-18-2-0014		N/A		62303E	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER	
				R2FM	
6. AUTHOR(S)					
Marjorie Freedman, Shi-Fu Chang, Filip Ilievski, Heng Ji, Kathleen McKeown, Daisy Wang					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)				8. PERFORMING ORGANIZATION REPORT NUMBER	
University of Southern California Information Sciences Institute 4676 Admiralty Way, Suite 1001 Marina del Rey CA 90292					
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-2023-215
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					
This final report describes the work performed by the GAIA team under DARPA's AIDA program. The GAIA team was a collaboration between: (1) The University of Southern California Information Sciences Institute (USC-ISI, PI: Marjorie Freedman); (2) The University of Florida (UofF, PI: Daisy Wang); (3) Rensselaer Polytechnic Institute (RPI, PI: Heng Ji) and The University of Illinois Urbana Champagne (UIUC, PI Heng Ji); and (4) • Columbia University (Columbia, PI: Shih-Fu Chang). The AIDA program's research focused on extracting and fusing multilingual, multimodal- content and synthesizing the information to identify conflicting claims. We explored diverse extraction targets in the form of knowledge elements (KEs) about events, relations, entities, sentiment, and claims. We explored knowledge fusion of both named and non-named concepts. We explored a variety of ontology targets from relatively small (derived from previous program's definitions) to enormous (with all of Wikidata as target). Software developed under this effort has been delivered to AFRL as dockized components integrated to run the full claim detection system. In addition to this report, the GAIA project resulted in numerous publications.					
15. SUBJECT TERMS					
Information Extraction, Multimedia Processing, Natural Language Processing, Computer Vision					
16. SECURITY CLASSIFICATION OF:				17. LIMITATION OF ABSTRACT	
a. REPORT	b. ABSTRACT	c. THIS PAGE			18. NUMBER OF PAGES
U	U	U	SAR		102
19a. NAME OF RESPONSIBLE PERSON					19b. PHONE NUMBER (Include area code)
NANCY SEPULVEDA					N/A

Table of Contents

LIST OF FIGURES	III
LIST OF TABLES	V
1 SUMMARY	1
2 INTRODUCTION	1
3 METHODS, ASSUMPTIONS, AND PROCEDURES	1
3.1 EXTRACTION FROM TEXT.....	1
3.1.1 <i>Sentiment, Stance, and Cognitive Relations, and Document Internal Claims</i>	1
3.1.2 <i>Question Answering Based Document Level Claim Detection</i>	9
3.1.3 <i>Text Joint Entity Relation Event Extraction</i>	9
3.1.4 <i>Text Document-level Event Argument Extraction</i>	10
3.1.5 <i>Text Entity and Event Coreference Resolution</i>	10
3.2 EXTRACTION FROM IMAGES AND VIDEO	11
3.2.1 <i>Visually Anchored Knowledge Elements</i>	11
3.2.2 <i>Multimodal Event Extraction</i>	17
3.2.3 <i>Zero-Shot Image Event Extraction</i>	17
3.2.4 <i>Video Event Extraction based on Argument State Tracking</i>	18
3.2.5 <i>Few-Shot Video Event Prediction</i>	19
3.2.6 <i>Integrating Pretrained Models for Visual Input to Claims</i>	19
3.2.7 <i>Audio within Video</i>	20
3.2.8 <i>Text within Video and Images</i>	20
3.3 AGGREGATING A KNOWLEDGE BASE	21
3.3.1 <i>Linking to Wikidata</i>	21
3.3.2 <i>Constructing a Knowledge Base of Extracted Knowledge Elements</i>	21
3.3.3 <i>Knowledge Graph Tool Kit</i>	22
3.4 DISCOVERING CLAIMS	27
3.4.1 <i>Cross-Claim Relation Detection:</i>	28
3.4.2 <i>Claim Selection through Combination of Structural and NLP Results</i>	29
3.4.3 <i>Claim Frame (CF) Clustering</i>	29
3.4.4 <i>Claim Frame (CF) Ranking</i>	29
3.4.5 <i>Claim Frame (CF) Answer Selection:</i>	29
3.5 INTEGRATION INTO PROGRAM TESTBED	30
3.6 VISUALIZING AND INTERACTING WITH EXTRACTED KNOWLEDGE	33
3.6.1 <i>Claim Radar</i>	33
3.6.2 <i>SmartBook: AI-Assisted Situation Report Generation</i>	35
4 RESULTS AND DISCUSSION	36
4.1 EXTRACTION FROM TEXT.....	36
4.1.1 <i>Sentiment, Stance, and Document Internal Claims</i>	36
4.1.2 <i>Question Answering Based Document Level Claim Detection</i>	44
4.1.3 <i>Text Joint Entity Relation Event Extraction</i>	45
4.1.4 <i>Text Document-level Event Argument Extraction</i>	45
4.1.5 <i>Text Entity and Event Coreference Resolution</i>	45

4.2	EXTRACTION FROM IMAGES AND VIDEO	46
4.2.1	<i>Visual Content</i>	46
4.2.2	<i>Multimodal Event Extraction</i>	50
4.2.3	<i>Zero-Shot Image Event Extraction</i>	51
4.2.4	<i>Video Event Extraction based on Argument State Tracking</i>	52
4.2.5	<i>Few-Shot Video Event Prediction</i>	52
4.2.6	<i>Integrating Pre-Trained Vision Models</i>	52
4.2.7	<i>Audio and Text within Video and Images</i>	55
4.3	AGGREGATING A KNOWLEDGE BASE	55
4.3.1	<i>Linking to Wikipedia</i>	55
4.3.2	<i>Constructing a Knowledge Base of Extracted Knowledge Elements</i>	56
4.3.3	<i>Knowledge Graph Tool Kit</i>	56
4.4	DISCOVERING CLAIMS	62
4.5	INTEGRATION INTO PROGRAM TESTBED	66
4.6	VISUALIZING AND INTERACTING WITH EXTRACTED KNOWLEDGE	67
4.6.1	<i>Claim Radar</i>	67
4.6.2	<i>SmartBook: AI-Assisted Situation Report Generation</i>	72
5	CONCLUSIONS	72
6	REFERENCES	74
7	APPENDIX A - PUBLICATIONS AND PRESENTATIONS	82
8	SYMBOLS, ABBREVIATIONS, AND ACRONYMS	87

List of Figures

Figure 1: Typical sentences in our dataset, which includes entity (boxed) and relation (as arrows) annotations with human rationales (underlined).	2
Figure 2: Architecture of TGA Net. Enc indicates contextual conditional encoding, GTR indicates Generalized Topic Representations and TGA indicates Topic-grouped Attention.	4
Figure 3: event-guided denoising: using date-marked events from Wikipedia to select a diverse group of sentences that describe the same relation (indicated with a check) from news articles, filtering out noisy examples (indicated with an X).	5
Figure 4: sentiment, epistemics and claimer detection, modeled as NLI.	7
Figure 5: the multilingual discourse connective modeling task.	8
Figure 6: An overview of our claim detection framework based on zero-shot QA	9
Figure 7: An illustration of our end-to-end joint information extraction framework ONEIE at the test stage	9
Figure 8 Our argument extraction model using conditional generation.	10
Figure 9 Our mention-pair encoder uses CDGMs (Context Dependent Gated Module) to incorporate symbolic features	11
Figure 10: Multimodal event extraction and argument role labeling.	12
Figure 11: Multimodal transformer for joint event extraction and argument role labeling. In the target sequence, blue-gray and light orange are for textual and visual decoding heads.	13
Figure 12: Multimodal event coreference resolution from self-supervision.	15
Figure 13: Spatio-temporal grounding of visual event from self-supervision.	15
Figure 14: Spatio-temporal localization model.	16
Figure 15: Spatio-temporal inference.	16
Figure 16: Overview of Multimodal Event Extraction.	17
Figure 17: Architecture of CLIP-Event.	18
Figure 18: Video event detection via tracking argument status.	18
Figure 19: Candidate retrieval based on entity profiling	21
Figure 20: KGTK's Web Browser.	23
Figure 21: KGTK's web browser browser functionalities. a) incoming links from related items; b) profiling data.	23
Figure 22: DPR query to retrieve spouse fact support documents.	26
Figure 23: Temporal information extraction architecture.	27
Figure 24: Pipeline for HypoGator Claim Extraction	28
Figure 25: Overall workflow for constructing SmartBook	36
Figure 26: Performance as F-score Varying Number of Rationales	38
Figure 27: Visualization of event extraction results on one video segment. We mask faces (orange boxes) for privacy.	47
Figure 28: Event coreference resolution visualization. The bold sentence is correctly selected as coreferential within the article by the model.	48
Figure 29: Visualization on GroundingYoutube dataset. The red box is the annotation and heatmap is the prediction from the model.	50
Figure 30: A caption example using LDC image L0C04AEUW.jpg. For reporting purposes to avoid propagated personally identifiable information we have masked the face.	53

Figure 31: Evaluation of the time needed for participants to retrieve the correct answer with our browser.	57
Figure 32: Brute-force top-k similarity query for philosophers similar to Socrates (Q913) with BERT text embeddings.	57
Figure 33: Generalized vector join query with multiple inputs.	58
Figure 34: DPR query to retrieve spouse fact support documents.	59
Figure 35: Number of wrong models per question.	62
Figure 36: Results Comparing HypoGator Claim Approach on LDC Annotated Data	63
Figure 37: An initial view of the Extended Claim Radar Visualizer with some corrections made to Claimers.	67
Figure 38: In Extended Claim Radar, the edit value view and the dropdown options upon clicking an editable cell in the table	68
Figure 39: In Extended Claim Radar, the Claim Details view of a claim when logged in.	68
Figure 40: In Claim Radar, an example of a Knowledge Element table in Claim Details	69
Figure 41: In Extended Claim Radar, the display of associated knowledge elements after clicking “awarded.”	69
Figure 42: In Extended Claim Radar, the display from Figure 41 when hovering over “scientists.”	70
Figure 43: In Extended Claim Radar, a set of filter queries along with the first result	70
Figure 44: In Extended Claim Radar, A claim detail modal dialog displays a claim including a source image.	71
Figure 45: An example showing the summaries produced by SmartBook and other baseline approaches.	72

List of Tables

Table 1: F-score (for Include- $\emptyset$ columns) and accuracy (for Exclude- $\emptyset$) columns for various baselines and our sentiment method.	37
Table 2: Performance breakdown on GFBB including $\emptyset$	37
Table 3: Precision, recall and F-scores for each label on MPQA (including $\emptyset$), averaged across 5 folds. Each entry shows the performance with trained attention, with the improvement over AttnLSTM baseline in parentheses. Results are averaged over 5 folds.....	37
Table 4: Macro-averaged F1 for stance on the test set. * indicates significance of TGA Net over BERT-joint, $p < 0.05$	39
Table 5: For stance, accuracy on varying phenomena in the test set. I indicates examples with the phenomenon, O indicates examples without.....	40
Table 6: For stance, F1 on the test set for examples with a majority sentiment polarity (M) and conversion between sentiment polarities (e.g., $M^+ \rightarrow M^-$). The direction the score a sentiment-susceptible model is expected to change is indicated with the up and down arrows.	40
Table 7: Zero-shot (no fine-tuning) accuracies on the FewRel development set.....	41
Table 8: F1-scores on two standard supervised cognitive relation extraction benchmarks for (left) English (SemEval 2010 Task 8) and (right) Spanish (TAC KBP 2016). EvtGD indicates event-guided denoising.	41
Table 9: Single task, multitask and NLI-based performance on PolNEAR (claimer), MPQA (sentiment), FactBank (epistemics) and GAIA’s manual annotations (claimer and epistemics).....	43
Table 10: Zero-shot transfer (from English to other languages) performance on XNLI.	44
Table 11: Topic-wise F1 and overall accuracy (both in %) for topic	44
Table 12: IE Results on ACE2005 (F-score, %).....	45
Table 13: Argument Extraction Results (%) on WIKIEVENTS	45
Table 14: End-to-end results on ACE 2005 for Coreference.....	46
Table 15: Video event extraction and argument role labeling.	46
Table 16: Multimodal event coreference resolution.	47
Table 17: Video event extraction and argument role labeling with self-supervision.....	49
Table 18: Spatio-temporal grounding with self-supervision.....	50
Table 19: Event and argument extraction results (%). We compare three categories of baselines in three evaluation settings.....	51
Table 20: Evaluation results and ablation studies on image event extraction.....	51
Table 21: Results on semantic role prediction. Results are the average (Avg) over 10 runs with standard deviation (Std) reported.	52
Table 22: Video event prediction (%).	52
Table 23: The GAIA Unified-IO-Inference VizWiz-Captions scores on test run.	53
Table 24: Performance of alternative prompts used for image classification. Accuracies are calculated on a set of 200 images belonging to 10 different classes sampled from ObjectNet.	54
Table 25: Overall entity linking results of micro F1 scores (%).	55
Table 26: Zero-shot evaluation scores of pre-trained models over the fact validation tasks.	59
Table 27: F1 of Claim Radar in Condition 5 on Various Metrics.....	64
Table 28: F1 of the Ontopic Measure for Evaluation Condition 6	65

1 SUMMARY

This final report describes the work performed by the GAIA team under DARPA's AIDA program. The GAIA team was a collaboration between

- The University of Southern California Information Sciences Institute (USC-ISI, *PI: Marjorie Freedman*),
- The University of Florida (*UofF, PI: Daisy Wang*),
- Rensselaer Polytechnic Institute (*RPI, PI: Heng Ji*)
- The University of Illinois Urbana Champagne (*UIUC, PI Heng Ji*)
- Columbia University (*Columbia, PI: Shih-Fu Chang*)

The AIDA program's research focused on extracting and fusing multilingual, multimodal-content and synthesizing the information to identify conflicting claims. We explored diverse extraction targets in the form of knowledge elements (KEs) about events, relations, entities, sentiment, and claims. We explored knowledge fusion of both named and non-named concepts. We explored a variety of ontology targets from relatively small (derived from previous program's definitions) to enormous (with all of Wikidata as target). Software developed under this effort has been delivered to AFRL as dockized components integrated to run the full claim detection system. In addition to this report, the GAIA project resulted in numerous publications which are listed in Section 7.

2 INTRODUCTION

In the following report, we describe the approaches and results used to address AIDA's diverse modalities and languages. We organize this report with respect to both input type (e.g. *text or multimodal*); output target (e.g. *document-local extraction, synthesized knowledge base, and claims*). We further describe how the overall system was integrated into a shared program environment for evaluation and team internal work to provide interactive visualizations of the system's results.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Extraction from Text

3.1.1 Sentiment, Stance, and Cognitive Relations, and Document Internal Claims

Sentiment Relations: While the general task of textual sentiment classification has been widely studied, much less research looks specifically at sentiment between a specified source and target.

To tackle this problem, we experimented with a state-of-the-art relation extraction model. Surprisingly, we found that despite reasonable performance, the model’s attention was often systematically misaligned with the words that contribute to sentiment. Thus, we directly trained the model’s attention with human rationales and improved our model performance by a robust 4-8 points on all tasks we defined on our data sets.

Sentences

“He endorsed a proposal for targeted sanctions against Mugabe.”

“The sale infuriated Beijing, which regards Taiwan an integral part of its territory.”

Relations

He $\xrightarrow{+}$ proposal, proposal $\xrightarrow{-}$ Mugabe
endorsed against

The sale $\xrightarrow{-}$ Beijing, Beijing $\xrightarrow{+}$ Taiwan
infuriated regards

We developed a method for extracting opinion relations that features a technique for training attention to focus on human rationales, along with a series of quantitative metrics to evaluate the quality of the model’s attention. Our approach builds on a state-of-the-art relation extraction method [ZH17] and augments it with a loss function based on KL-divergence between the model’s attention and the human rationales in our training data. We evaluated our system on two separate annotated corpora which carry different kinds of sentiment: the Multi-Perspective Question Answering (MPQA) corpus [WI08] which has been extensively used in sentiment analysis, and the Good-For Bad-For (GFBF) corpus [DE13] which contains sentences expressing whether a person or event is good or bad for another person or event. Both datasets have been annotated with rationales, enabling the research we do here. We also developed our system for two separate scenarios, one in which it must distinguish between pairs of entities that have sentiment relations and pairs that do not, and another in which it receives only entity pairs that do participate in sentiment relations where the goal is to determine polarity.

We compared our approach against a wide range of relation extraction systems showing a significant absolute gain in performance of 8 points. We experimented with different quantities of human-annotated rationales and show that only a small number of human rationales was enough to yield half of the gain in performance[ZH19].

Stance: To complement our work in Phase 1 on sentiment identification, in Phase 2 we developed a novel technique for zero-shot stance detection. This expanded focus was informed by an observation that the kinds of targeted sentiment relations relevant to the AIDA program were regularly expressed implicitly. For example, consider the relation expressed in, “Russian state media cheered the dispersal of ‘the violent thugs from the Maidan.’” While Russian state media’s sentiment toward the “violent thugs” (likely identified as protestors by an upstream coreference system) is clear to any reader, standard sentiment classification techniques struggle with its implicit expression. By incorporating additional scores from stance detection, our system was more robust to the diverse ways sentiment is expressed in AIDA.

Given the large set of entities and events in the ontology (and their paucity in the available training data), we opted for a zero-shot approach that allows for stance prediction for any user-specified topic. We collected and released a new corpus, VAST, that includes 5,634 unique topics debated in 23,525 news article comments, and we use this dataset to evaluate a novel architecture for zero-shot stance detection that implicitly captures relationships between topics using generalized topic representations. Our experimental results show that this improves over the baseline on several challenging linguistic phenomena. We discuss here the work on VAST[AL20], but have a larger body of work on stance published over the course of the co-operative agreement, including work using an adversarial approach to few-shot stance detection [AL21], an approach using rationales[JA21] and a recently accepted paper on few-shot approaches in multilingual stance to appear in *Frontiers in Artificial Intelligence*.

Our dataset, **V**ARied **S**tance **T**opics (VAST), spans a large range of topics covering broad themes, such as politics (e.g., ‘a Palestinian state’), education (e.g., ‘charter schools’), and public health (e.g., ‘childhood vaccination’). It consists of comments collected from The New York Times ‘Room for Debate’ section, drawn from the Argument Reasoning Comprehension (ARC) Corpus. We first heuristically extract specific topics from the stance positions provided by the ARC corpus. We define a candidate topic as a noun-phrase in the constituency parse, generated using Spacy, of the ARC stance position. To reduce noise, we filter candidates to include only noun-phrases in the subject and object position of the main verb in the sentence. If no candidates remain for a comment after filtering, we select topics from the categories assigned by The New York Times to the original article the comment is on. Then, we use Amazon Mechanical Turk to collect crowd-sourced annotations. We present each worker with a comment and first ask them to list topics related to the comment, to avoid biasing workers toward finding a stance on a topic not relevant to the comment. We then provide the worker with the automatically generated topic for the comment and ask for the stance, or, if the topic does not make sense, to correct it. Workers are asked to provide stance on a 5-point scale which we map to 3-point pro/con/neutral.

Each topic-comment pair is annotated by three workers. For all examples, we select the majority vote as the final label. Finally, as ARC is focused on positive and negative expressions, we expand our set of neutral examples by randomly permuting topics and comments, ensuring semantic distance by disallowing lexical overlap.

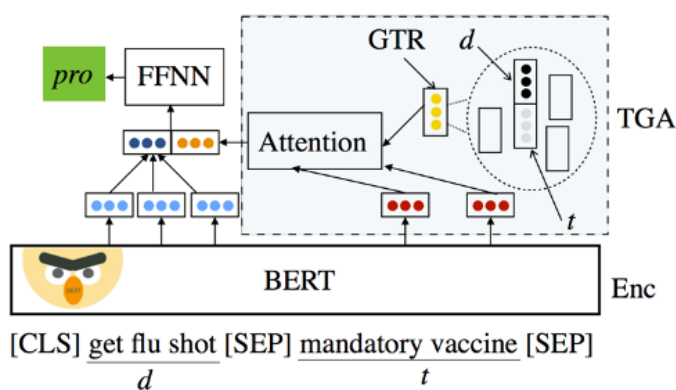


Figure 2: Architecture of TGA Net. *Enc indicates contextual conditional encoding, GTR indicates Generalized Topic Representations and TGA indicates Topic-grouped Attention.*

Using VAST, we developed Topic-Grouped Attention (TGA) Net: a model to implicitly construct and use relationships between the training and evaluation topics without supervision. The model consists of a contextual conditional encoding layer, followed by topic-grouped attention using generalized topic representations and a feed-forward neural network (see Figure 2).

We first begin by independently embedding our training documents d and topics t with BERT[DE19], calculating a TFIDF-weighted average of each, and concatenating these averages to produce a joint representation of each training example’s document and topic. Then, we perform hierarchical clustering to discover “generalized topic representations” r_{dt} .

During training (and inference), we embed a document and topic jointly using BERT, treating the document and topic as a sentence pair, and obtain two sequences of text embeddings, one for the document d' and another for the topic t' . These document embeddings are conditioned on the topic and vice versa. We mean pool the document embeddings and calculate a “topic-grouped attentional” weighting over the topic embeddings by calculating the similarity of each to the training example’s nearest “generalized topic representations” r_{dt} . The resulting pooled document and topic representations are passed to a feed-forward network for label prediction.

We evaluated TGA Net against several state-of-the-art baselines including BiCond[Au16], CrossNet, and BERT in both zero and few shot settings and found that only BERT is competitive

with TGA Net. Nevertheless, our extensive error analysis demonstrates TGA Net performs better on several particularly difficult linguistic phenomena.

Cognitive relations (sponsorship, blame, hoaxfraud, source): we augmented our pattern-based system for the identification of *sponsorship* and *blame* relations with a neural system for multilingual (English, Spanish and Russian) relation extraction. As with stance detection, we opted for a few-shot approach which allowed us to flexibly adapt to changes in the AIDA ontology. In the months before the evaluation, in addition to *sponsorship* and *blame*, we were able to use this system to identify instances of *hoaxfraud* relations as well as identify the source of knowledge elements to help power TA3 perspectives.

To build this system, we developed a method for automatically extracting high quality relation training examples from date-marked news articles. This technique leverages the predictable distributional structure of such articles to build a corpus that is denoised (i.e., where sentences with the same entities only express the same relation, see Figure 3). We use this corpus to learn general purpose relation representations and evaluate their quality on few-shot and standard relation benchmarks in English and Spanish with little to no task-specific fine-tuning, achieving comparable results to a significantly more data-intensive approach that was the state-of-the-art.

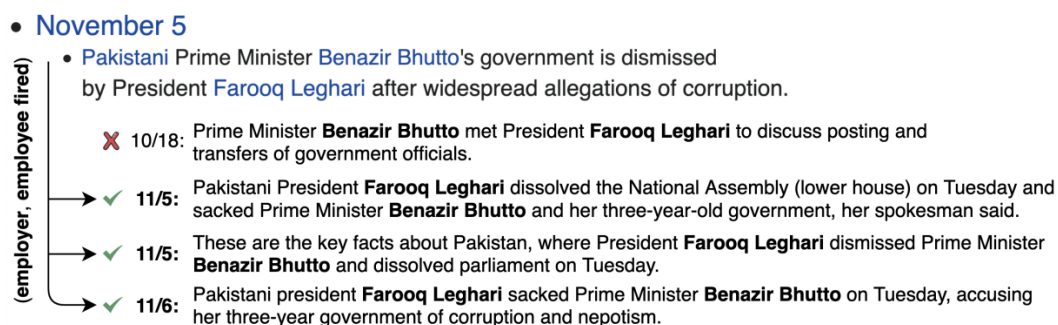


Figure 3: event-guided denoising: using date-marked events from Wikipedia to select a diverse group of sentences that describe the same relation (indicated with a check) from news articles, filtering out noisy examples (indicated with an X).

The state-of-the-art model, “Matching the Blanks” or MTB , is a distant supervision [LI09] technique that provides large gains on many relation extraction benchmarks and builds on Harris’ distributional hypothesis [HA54] and its extensions [LI01]. Soares et al. (2019) assume that the informational redundancy of very large text corpora (e.g., Wikipedia) results in sentences that contain the same pair of entities generally expressing the same relation. Thus, an encoder trained to collocate such sentences can be used to identify the relation between entities in any sentence s by finding the labeled relation example whose embedding is closest to s . While Soares et al. (2019) achieve state-of-the-art on FewRel [HA18] and SemEval 2010 Task 8 [HE19] their approach relies on a huge amount of data, making it difficult to retrain in English or any other

language with standard computational resources: they fine-tune BERT large [DE19], which has 340mil parameters, on 300mil+ relation pair statements with a batch size of 2048 for 1mil steps. In contrast our method, with only 50k relations statements and a language model one-third the size, achieves comparable performance when fine-tuned on little to no task-specific data.

Our main contribution is a distant supervision approach in which we assume that sections of news corpora exhibit even more informational redundancy than Wikipedia. Specifically, news in the days following an event (e.g., the 2006 World Cup) frequently re-summarizes the event before adding new details. As a result, news exhibits a strong form of local consistency over short rolling time windows where otherwise fluid relations between entities remain fixed. For example, the relation between Italy and France as expressed in a random piece of text is dynamic and context-dependent, spanning a wide range of possibilities that include “enemies”, “neighbors” and “allies”. But, in the news coverage following the 2006 World Cup, it is static – they are sporting competitors. Therefore, by considering only sentences around specific events, we extract groups of statements that express the same relation and are relatively free of noise (i.e. statements that express a different relation).

We train multilingual BERT (mBERT)[DE19] on our denoised corpus, adopting the training methodology for learning relation representations employed by Soares et al. (2019) (MTB) in its entirety. That is, we train an encoder $f_{\theta}(r) = h$ to produce relation representations $h \in \mathcal{R}^d$ that determine whether two relation statements r_i, r_j describe the same relation by computing their inner product. During training, relation statements are collected into groups of positive examples (statements that align with the same relation type according to our denoising techniques) and negative examples (statements that align with a different relation type) for each selected entity pair. Following Soares et al. (2019), negative examples include both “easy” and “hard” mentions. “Easy” mentions include no part of the entity pair and “hard” mentions include exactly one of the entities in the entity pair, suggesting that these “hard” examples describe similar but different relations that require disambiguation. With probability α , each entity in a relation statement is replaced with a [BLANK] token and with probability β each token in a relation statement is replaced with a [MASK] token. We train a parameterization of f_{θ} that minimizes both a masked language modeling loss [TA53]; [DE19] and a simple binary cross-entropy loss that encourages the representations of positive examples for a given entity pair to be closer to one another than to the negative examples.

We evaluate the quality of these representations on FewRel and SemEval 2010 Task 8, producing near state-of-the-art results when finetuned on little to no task-specific data. In addition to the strong performance of our approach in English, it is easily generalizable to other languages, requiring only news corpora and event descriptions from Wikipedia to build a high-quality training corpus. We evaluate this in Spanish and find our method outperforms mBERT on the TAC KBP 2016 relation corpus[EL15].

Epistemics & Source: In Phase 3, as AIDA pivoted toward COVID claim detection, our group worked on the identification of sentiment, epistemic status and claimer for claims in English, Spanish and Russian. As in the prior phases, due to a lack of available training data, we leveraged the semantic entanglement among the three tasks in a multi-task framework. When evaluating against benchmarks like PolNeAR[NE18], MPQA[DE15], FactBank[SA09] and GAIA’s manually collected COVID-19 claims, our multi-task RoBERTa model[LI19] outperformed its single task variants. Due to relative abundance of NLI training data [BO15], [CO18] and its good transfer performance [CH22], we “intermediate”-trained our model on NLI (see Figure 4), resulting in additional performance improvements on both FactBank and GAIA.

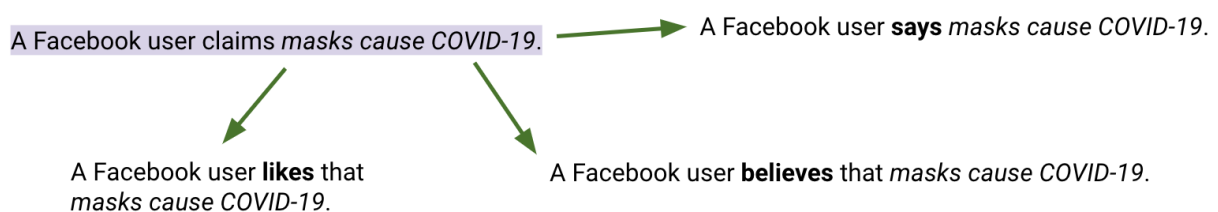


Figure 4: sentiment, epistemics and claimer detection, modeled as NLI.

To enhance the improvement due to NLI, we focused on ways to distantly supervise the collection of NLI pretraining corpora to improve its performance and enable better zero-shot transfer from English to other languages. After an initial investigation into automatically collecting claims of opposing epistemic status from news articles discussing contentious topics (in politics and about COVID), we determined that the resulting extracted corpora were too narrow. Instead, building on work from the literature on discourse [SI19], we re-oriented toward the task of multilingual discourse connective prediction for sentence pair representation learning,

informed by the insight that often markers like “moreover” and “nevertheless” indicate the semantic alignment between adjacent sentences (see Figure 5).

“We formed a government which many believed was an impossible one — we formed it in order to stop the terrible tailspin Israel was in the midst of,” Mr. Bennett said in his speech. “Together we were able to pull Israel out from the hole,” he added.

[MASK] the government was ultimately unable to overcome its contradictions. Its members clashed regularly over the rights of Israel’s Arab minority, the relationship between religion and state, and settlement policy in the occupied West Bank — clashes that ultimately led two key members to defect, and others to vote against government bills.

Aunque la opinión pública estuvo dividida, las élites políticas y militares señalaron que creían que el líder de izquierda en funciones en aquel momento, Evo Morales, debía dejar el cargo y prácticamente lo obligaron a hacerlo.

[MASK] cuando el remplazo de derecha de Morales no pudo evitar meses de inestabilidad y turbulencia y luego se dispuso a posponer las elecciones, muchas de esas mismas élites presionaron para que estas se realizaran rápidamente, lo que benefició al sucesor elegido por Morales.

PDTB discourse relation classes: contrast || elaboration || specification || temporal || ...

Figure 5: the multilingual discourse connective modeling task

Using automatically translated variants of the Penn Discourse TreeBank[PR08], we trained an XLM-R[CO19] based model (across 16 languages) to predict whether instances of discourse connectives connect the pairs of sentences which they adjoin. We used this model to extract a large corpus of multilingual discourse connective delimited sentence pairs from newswire and Wikipedia. Additionally, we augmented our NLI model with a novel masked cross-attention architecture that iteratively builds a sentence pair representation using token representations weighted by their similarity to tokens in the other sentence.

We’ve pre-trained this model with our extracted corpus on the task of discourse connective prediction (exploring predicting the connective and its underlying PDTB discourse relation class). Shallow classification heads trained on English-only data using these sentence pair representations outperform baselines like XLM-R on Spanish, French, Russian and Chinese XNLI by 10-15% points in accuracy. We are currently exploring different approaches that enable fine-tuning this model on English only data while preserving its strong zero shot multilingual transfer performance before deploying it for improved epistemic and claimer detection in our system.

3.1.2 Question Answering Based Document Level Claim Detection

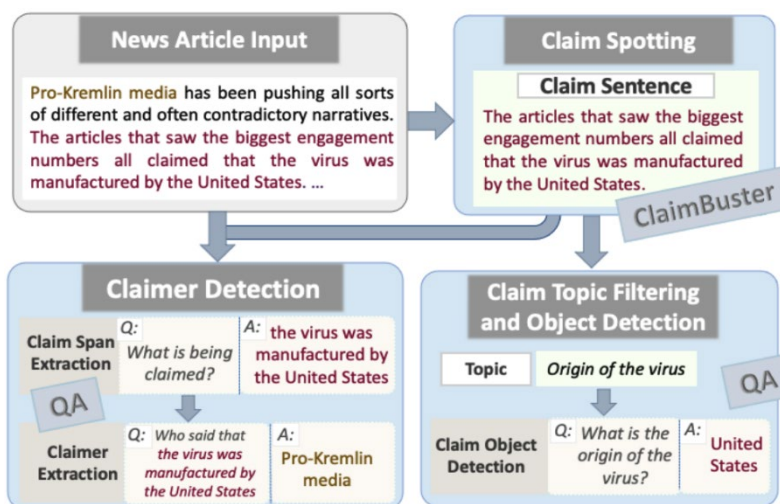


Figure 6: An overview of our claim detection framework based on zero-shot QA

A notable bottleneck of existing claim detection approaches is their portability to emerging events and low-resource training data settings. In this regard, we propose a fine-grained claim detection framework that leverages zeroshot Question Answering (QA) using directed questions to solve a diverse set of sub-tasks such as topic filtering, claim object detection, and claimer detection. We leverage a QA model as the pre-trained model to feed directed questions as input. Specifically, we outline our zero-shot claim detection framework which leverages the above model for sub-tasks such as topic filtering, claim object detection and claimer detection.

3.1.3 Text Joint Entity Relation Event Extraction

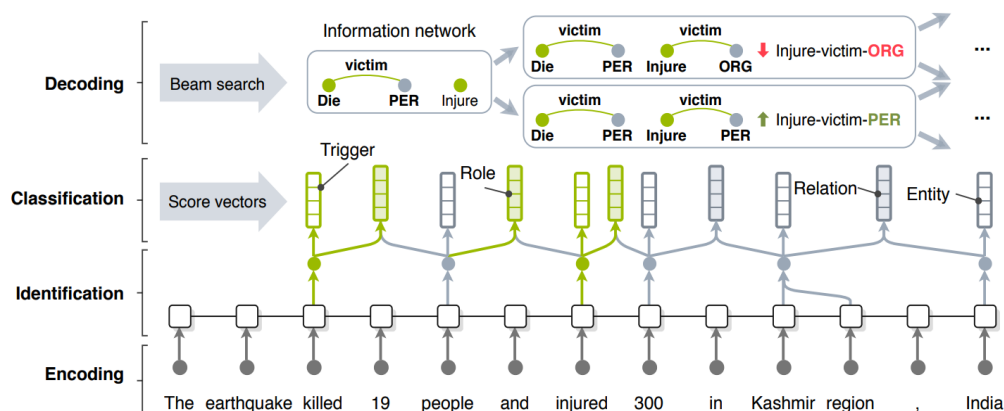


Figure 7: An illustration of our end-to-end joint information extraction framework ONEIE at the test stage

Existing methods use separate task-specific classifiers in the final layer and do not explicitly model interdependencies among tasks and instances. Instead of predicting separate knowledge elements using local classifiers, OneIE aims to extract a globally optimal information network for the input sentence. OneIE performs end-to-end IE in four stages: (1) Encoding a given sentence as contextualized word representations; (2) Identifying entity mentions and event triggers as nodes; (3) Computing label scores for all nodes and their pairwise links using local classifiers; (4) Searching for the globally optimal graph with a beam decoder.

3.1.4 Text Document-level Event Argument Extraction

We propose a new end-to-end document-level event argument extraction model by framing the problem as conditional generation given a template. Conditioned on the unfilled template and a given context, the model is asked to generate a filled in template with arguments. In detail, we use a conditional generation model for argument extraction, where the condition is an unfilled template and a context. The template is a sentence that describes the event with `<arg>` placeholders. The generated output is a filled template where placeholders are replaced by concrete arguments.

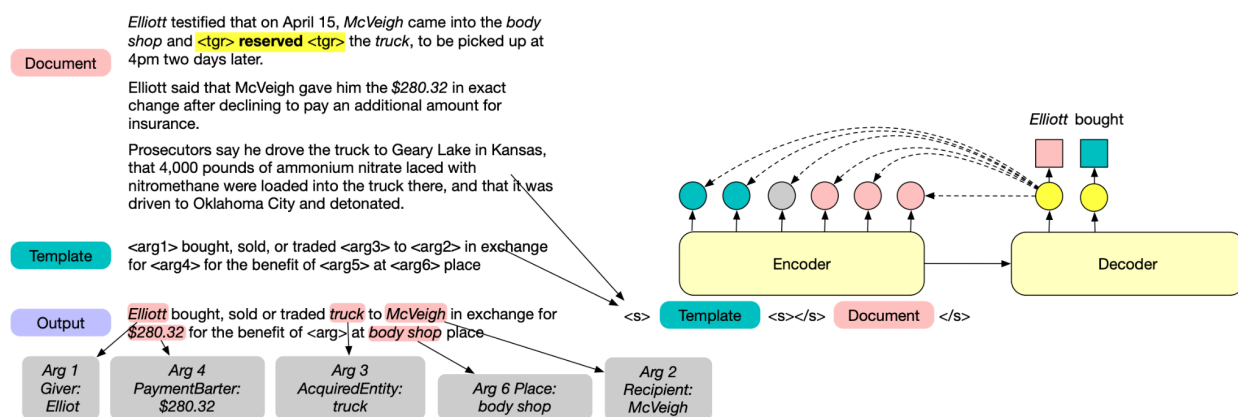


Figure 8 Our argument extraction model using conditional generation.

3.1.5 Text Entity and Event Coreference Resolution

We propose general and effective methods for incorporating a wide range of symbolic features into event coreference resolution. Simply concatenating symbolic features with contextual embeddings is not optimal, since the features can be noisy and contain errors. Also, depending on the context, some features can be more informative than others. Therefore, we propose

Context Dependent Gated Module (CDGM), which uses a gating mechanism to extract information from the input symbolic features selectively.

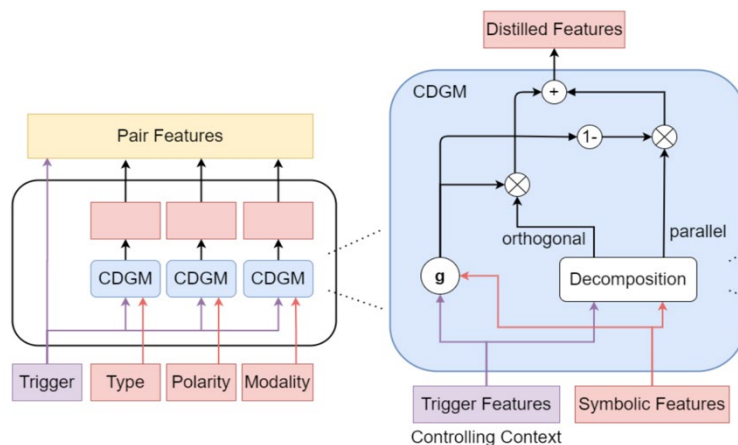


Figure 9 Our mention-pair encoder uses CDGMs (Context Dependent Gated Module) to incorporate symbolic features

3.2 Extraction from Images and Video

3.2.1 Visually Anchored Knowledge Elements

Learning to recognize real-time situations has drawn sufficient attention to the community. Aside from utilizing the text alone, multimodal information, especially vision, can serve as a complementary source to (1) add additional information in the context of the text and (2) verify the information from the text to be correct when both modalities had consensus. In the GAIA vision team, we specifically investigate the following directions for richer multimodal understanding:

- Multimodal event extraction and argument role labeling
- Learning visual event from multimodal self-supervision
- Grounded visual events from multimodal self-supervision

Multimodal event extraction and argument role labeling. Traditional event extraction methods target a single modality, such as text, images, or videos. However, in the AIDA scenario, it distributes news via multimedia. By randomly sampling 100 multimedia news articles from the Voice of America (VOA), we found that visual and text serve as complementary information. Our previous work [LI20b], has tried to solve the problem by jointly extracting the visual and textual event. we extend the model to extract events from videos for several reasons. First,

images contain a visual summary only, while video contains more action events that can be extracted. Second, we find some argument roles are hard to be recognized by a single image, while video provides a better understanding of the different arguments. Third, in video events, some argument roles exist across different video frames, which is more realistic than the image event.

The task of video m2e2 will be given a document with its corresponding video; we want to extract the events and argument roles from both domains. After we obtain the events/argument roles from each domain, we will link them if they represent the same events/argument roles. Multimedia event extraction/argument role labeling is given the video segment-sentence pair created by multimodal event linking; we perform the event classification and argument role labeling jointly from both domains.

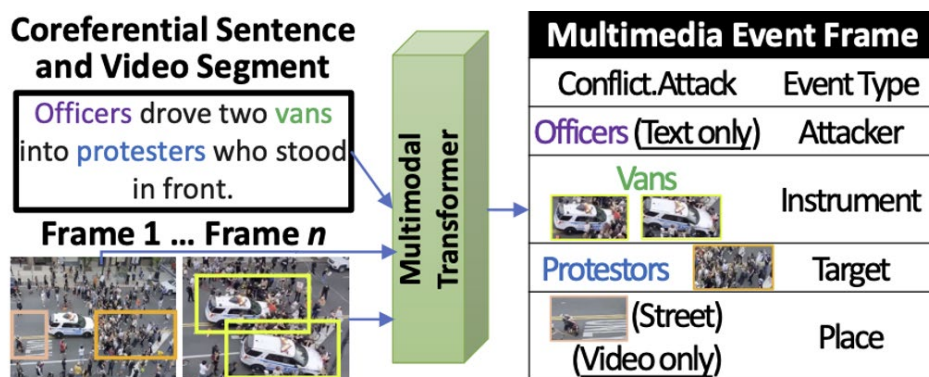


Figure 10: Multimodal event extraction and argument role labeling

We employ a joint multimodal transformer (JMMT) to extract events and arguments from a paired text and video clip for the joint event extraction and argument role labeling. The proposed JMMT has an encoder-decoder structure: the encoder extracts and fuses information from both modalities (text and video), while the decoder is more complex. The decoder consists of two heads: one for detecting trigger words, event types, and arguments from text and the other for classifying video event types and predicting bounding boxes for visual arguments. This joint encoder enables JMMT to effectively leverage contextual information to extract events and label argument roles, as shown in Figure 11. JMMT takes the text and visual tokens as input. For text tokens, we follow the standard T5 model [RA20] to embed text tokens. For visual tokens, we use four feature types to represent both global and local information comprehensively:

- (1) Video-level features extracted from the whole video segment capture the global event context.
- (2) Frame-level object labels produced by an object detector.

- (3) The frame-level region features extracted from bounding boxes detected by the object detector provide fine-grained argument information.
- (4) Frame-level object coordinates are also provided by the object detector for the localization of arguments.

Note that we sample 3 frames and 15 objects with the highest confidence scores in the video. The text and visual tokens are then stacked as a sequence and input to the encoder for joint processing.

Our encoder and decoder are initialized from transformers pretrained on text corpora. Our decoding head for text event extraction is borrowed from the state-of-the-art text event extraction model OneIE [RE20]. For text decoding, we take encoder outputs as input and first merge these multimodal contextualized embeddings of word pieces to obtain a representation for each word in the input sequence. Then we process these word representations for identification, classification, and decoding. We leverage the decoder from the T5 pretrained text transformer and cast the task as a sequence-to-sequence prediction problem for the video decoding head. The text decoding head is supervised based on the objective of OneIE, and the video decoding head is trained using a standard teacher-forcing strategy with cross-entropy loss. Its loss term supervises each decoding head, and the gradients are back-propagated to the encoder. The overall objective is the sum of the two losses. This way, the proposed JMMT can effectively fuse multimodal information and jointly extract events and arguments.

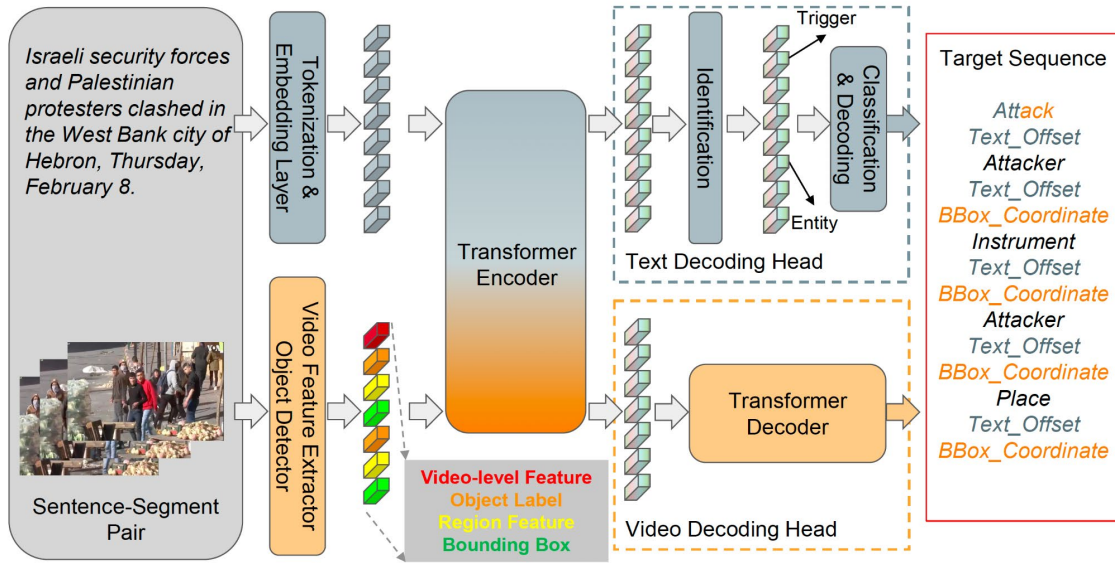


Figure 11: Multimodal transformer for joint event extraction and argument role labeling. In the target sequence, blue-gray and light orange are for textual and visual decoding heads.

Learning visual event from multimodal self-supervision. In the context of real-time information extraction, constructing a new vocabulary requires much human annotation. Hence, we investigate the self-supervised direction where we can learn common space representation from freely available video and ASR pairs in the YouTube videos. Later, our learned common space representation can perform zero-shot event extraction and argument labeling via our grounding component.

Apart from using human annotation, we have developed a self-supervised training method for the multimodal event coreference task that extends the NCE contrastive loss [RA16] to multi-instance learning by considering region features in the video frame. The NCE contrastive loss approach learns the common space across video and text by sampling positive and negative pairs within a batch of videos along with its ASR transcript. The positive pairs are video and ASR instances that co-occur in time, and the negative pairs are all the mismatch pairs within the batch. The intuition is that video instances and text instances should be closer in semantics while mismatched pairs do not. The video instance's representation was a max-pooled 2D and 3D feature over time, and the text instance was the max-pooled Glove features over the whole sentence. We trained on 3000 videos with their ASR transcript and tested on the annotated 1500 video article pairs. We further extend the positive pairs from the ASR transcript and the video segment to the ASR transcript paired with objects in the video segment. This formulation considers the objects (arguments) that participate in the event when we do event coreference across modalities.

For the multimedia event extraction/argument role labeling, we utilize the common space learned for multimodal event coreference for zero-shot event extraction. Given a video segment, we compare its feature to one of our 20 classes for event classification. Since video and text features are in a common space, the features are directly comparable.

Grounded visual event from multimodal self-supervision: To better the arguments in a certain multimodal event, we developed a self-supervised method for visual event extraction with the grounding ability to adapt to various domains without additional human annotation. To achieve this goal, we aim to learn multimodal representations with spatial/temporal localization ability from uncured video with its associated audio and ASR transcript. Given an event trigger extracted from the text IE as a query, we task the model to localize its start time/end time and spatial location in the video. We observe that we can learn spatially localized interactions with self-supervision on a large corpus of videos with accompanying transcribed narrations. Also, audio provides useful signals to determine whether a certain action starts, such as the yelling and marching sound in the protest event. Beyond the cross-modality supervision, we further design within-video modality supervision where arguments in the same event should share certain consistency. Also, we serve the temporal and spatial information and reciprocal information where we can select better grounding samples by our temporal model and better arguments to track via the spatial model. To this end, we develop a training pipeline to capture semantic similarities across modalities at a more granular level than in previous works. The model can be

applied to multimodal event co-referencing by computing sentence-to-video feature similarity, as shown in Figure 12, and event spatio-temporal grounding, shown in Figure 13.

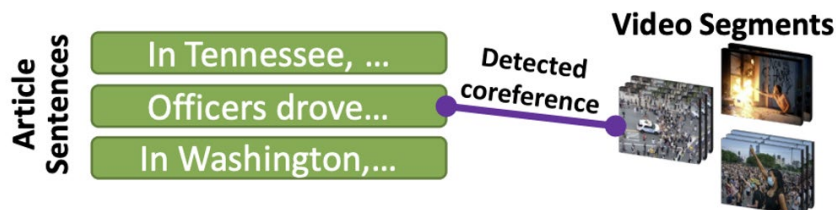


Figure 12: Multimodal event coreference resolution from self-supervision.

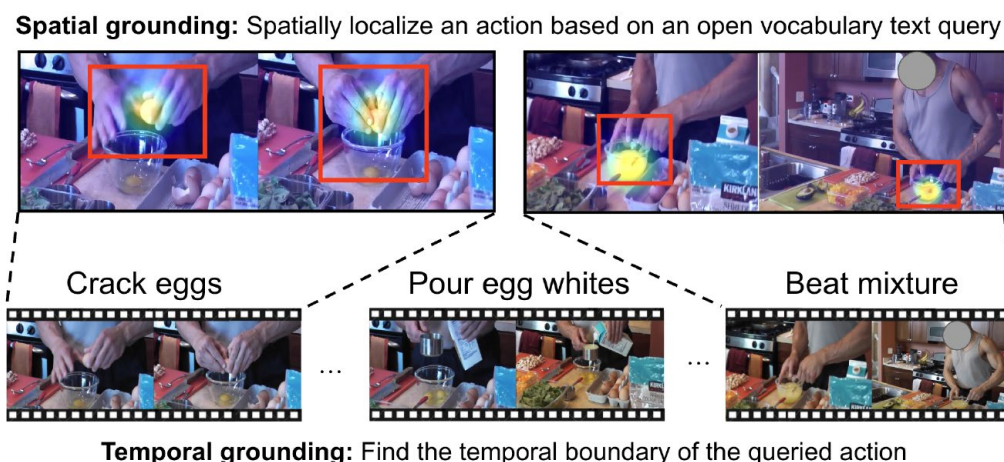


Figure 13: Spatio-temporal grounding of visual event from self-supervision.

Instead of learning a single vector representation for a video segment or sentence, we observe that we can learn spatially localized interactions with self-supervision on a large corpus of videos with accompanying transcribed narrations (ASR). Also, audio provides useful signals to determine whether a certain action starts, such as the yelling and marching sound in the protest event. Beyond the cross-modality supervision between video and text, we further design within-video modality supervision where arguments in the same event should share certain consistency. Also, we serve the temporal and spatial information and reciprocal information where we can select better grounding samples by our temporal model and better arguments to track via the spatial model. In addition, we use word features from CLIP transformer architecture [RA21] to select possible start time/end time given the long video. This is motivated by the imperfect supervision between video and its ASR pairs, where the text and video are likely to misalign. We can explore longer video supervision and better text pairs with our proposed frame selection strategy by Sinkhorn optimal transport, as shown in Figure 14. To this end, we develop a training pipeline to capture semantic

similarities across modalities at a more granular level than in previous works. In Figure 14, (a) we first select the most relevant words in the sentence using the [CLS] sentence embedding. (b) we then select frames with possible groundable objects by matching the selected, projected word features with respective frames features utilizing the Sinkhorn optimal transport. (c) based on the selected frames, we learn a local representation to ground the action description to the spatial region and (d) a global representation to allow for temporal localization.

As shown in Figure 15, during inferencing, our model can localize the event query and localize the corresponding region related to the event. We utilize both temporal and spatial representation during inferencing for spatio-temporal grounding. We start from predicting the action boundary on untrimmed videos. Spatial grounding is then performed using the predicted label as a query to find corresponding regions.

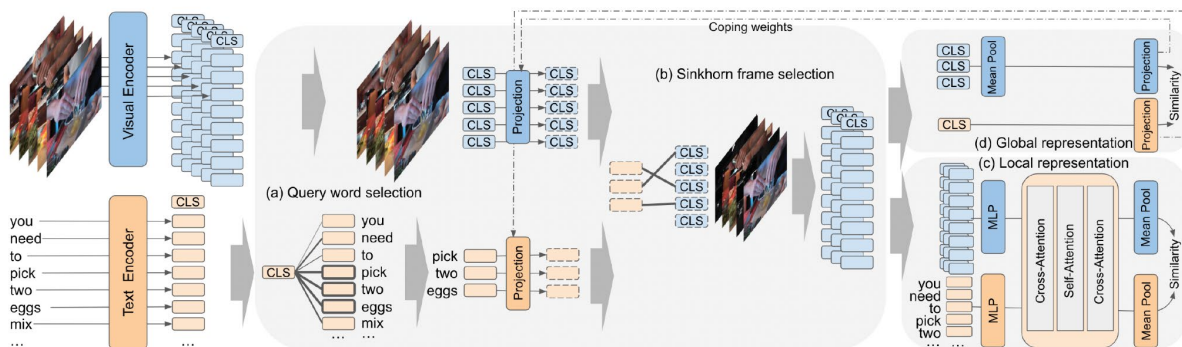


Figure 14: Spatio-temporal localization model.

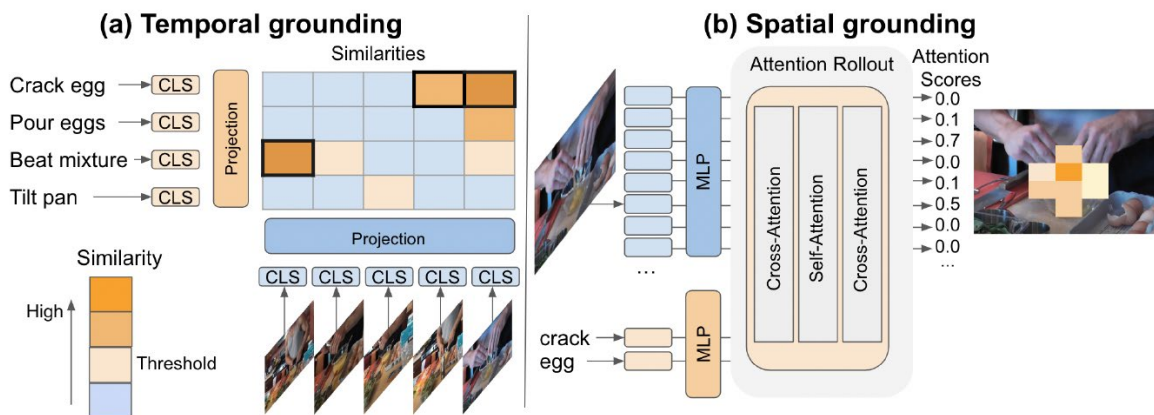


Figure 15: Spatio-temporal inference.

3.2.2 Multimodal Event Extraction

We propose a method to extract events and arguments from multiple modalities, illustrated in Figure 16. For a multimedia document, we represent each image/sentence as a graph, where each node represents an event or entity, and each edge represents an argument role. The node and edge embeddings are represented in a multimedia common semantic space, as they are trained to resolve event co-reference across modalities and to match images with relevant sentences.

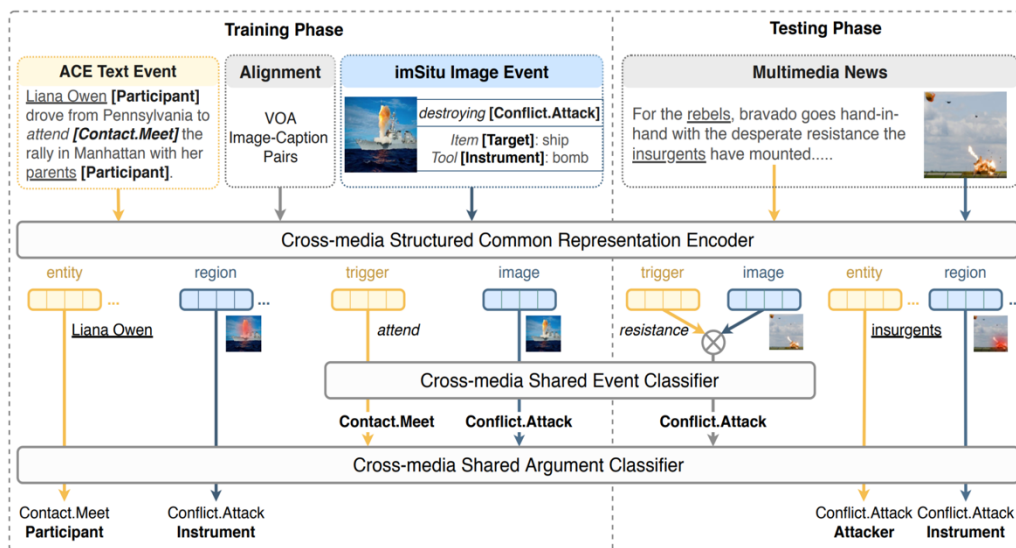


Figure 16: Overview of Multimodal Event Extraction.

3.2.3 Zero-Shot Image Event Extraction

There has been little research on extracting event structures from news images, yielding limited support for event knowledge acquisition needed in downstream applications. Thus, we propose to leverage text information extraction technologies, which have been well researched in natural language processing, to automatically extract event structures from captions. Furthermore, to transfer knowledge of argument structures, we explicitly construct event graphs consisting of event types and argument roles in vision and text. We introduce a fine-grained alignment between two event graphs, aligning the objects in images with the corresponding text entities and their argument roles. We employ optimal transport to encourage a global alignment based on the structures of two graphs, which enables the model to capture the interactions between arguments.

For example, objects with similar visual features tend to be aligned to the same argument role. This is illustrated in Figure 17.

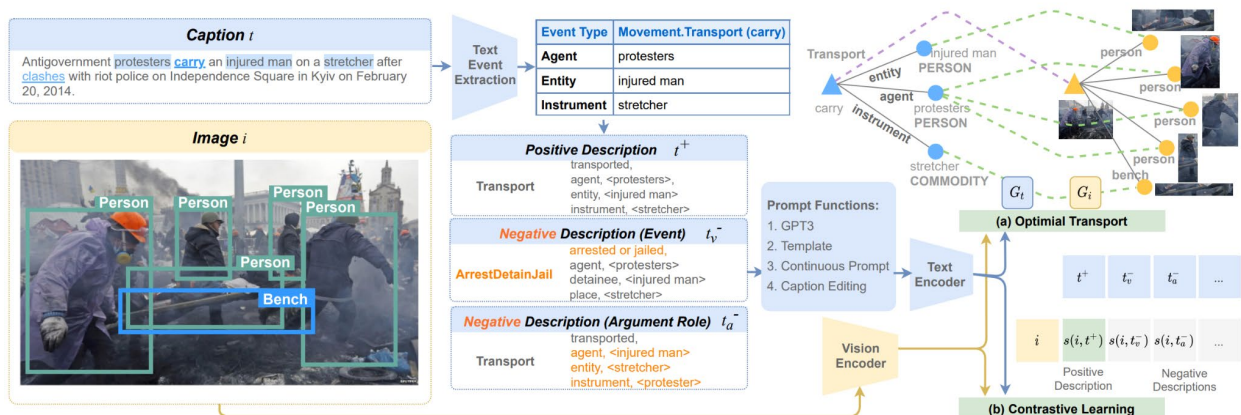


Figure 17: Architecture of CLIP-Event.

3.2.4 Video Event Extraction based on Argument State Tracking

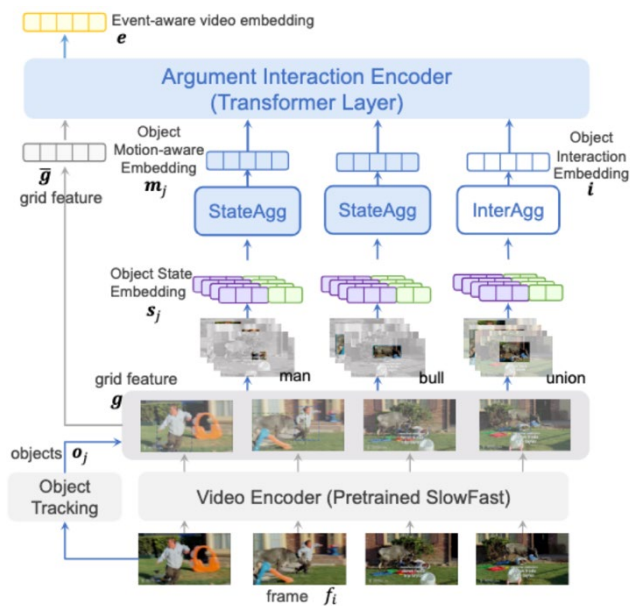


Figure 18: Video event detection via tracking argument status.

Due to the high variation of temporal dynamics and visual features presented in videos, it remains to be a challenging and understudied task. Existing methods focus on capturing the

overall visual scene of each frame, ignoring fine-grained argument-level information. We propose to encode the argument visual state changes from three perspectives: (1) Single-frame state is characterized by in-bounding-box pixel changes and displacements of bounding boxes. Therefore, we propose Object State Embedding (OSE) by encoding both pixel changes and displacements. (2) The motion of objects across multiple frames, which we introduce Object Motion-aware Embedding (OME) to aggregate the changes of object states from different frames. (3) The interaction between objects, which we propose to learn a Multi-Object Interaction Embedding by encoding the interaction area. This is illustrated in Figure 18.

3.2.5 Few-Shot Video Event Prediction

To address the unique challenges of videos, we first leverage a pretrained image-language contrastive model CLIP[RA21] to perform visual tokenization. Next, in order to capture the overall semantics at the frame level, we employ the pretrained image captioner in the **image-language** model BLIP[LI2022c] to obtain frame captions. Finally, we instruct a pretrained large language model using in-context learning to interpret visual tokens and frame captions into the target textual output.

3.2.6 Integrating Pretrained Models for Visual Input to Claims

Over the time period covered by the GAIA effort, there was extensive external to the AIDA program progress in computer vision, including the release of pre-trained general-purpose vision-to-text models (e.g. Unified-IO). In the months after the final evaluation, ISI explored the use of such pre-trained models as a means of extracting claim-relevant content from video and image data. To support this, a docker image was developed and tested to include all necessary packages for using a containerized model. The docker image definition included custom scripts for handling image archives in multiple formats and for running scheduled jobs across ISI's high performance computing cluster (HPCC).

We explored two approaches to integrating the pre-trained vision model. First, applying automatic captioning followed by applying text-based knowledge extraction and second, using the model's question answering capabilities as a form of labeling. Captioning like ASR and OCR serves to leverage a well-trained text-to-knowledge element model, not to produce knowledge elements.

The captioning task is well defined with available datasets and leaderboards, which we used as means of confirming that our integration of the pre-trained model achieved expected performance. The Unified-IO-Inference model was trained to perform computer vision tasks, including captioning, by responding to text prompts. The baseline prompt for captioning is "*what does the image describe?*". We compared the captioning performance of the baseline model to alternative prompts. The alternative prompts were selected with an intention of generating fuller responses from the model.

For image classification abilities of Unified-IO through two experiments. Routines to swap input prompts and output labels were created and used to create a docker image. In the first experiment, the model's image classification performance was measured using a set of handcrafted alternative prompts. To use the knowledge contained within the model, we replaced the category labels with descriptions of the categories in the second experiment. These descriptions were created manually for each of the ten categories. The model was expected to classify a given image into classes based on these descriptions instead of the class labels. For example, the class label 'wheel' was replaced by the class *description* 'A circular component that is intended to rotate on an axle bearing.'

We explored expanding the GAIA TA1 pipeline's document-level claim generation capabilities by using the Unified-IO-Inference model. Here, we used question-answering to identify scene location and other claim details. The AIDA program task of generating claims is not one of the standard computer vision tasks that the Unified-IO-Inference model was trained to perform. A claim has many properties in its full definition, including the topic, claimer, location, etc. We used custom prompts to construct detailed claims. The model was used to respond to a batch of prompts. The resulting answers would then be aggregated to populate a claim. As an example, 'What does the image claim?' was used to produce the foundation of a claim. Additionally, 'Where does the image take place?' was used to add a location to the claim. To form a single claim, multiple model responses were required to be aggregated.

3.2.7 Audio within Video

While automatic speech recognition (ASR) was not a primary focus of the work under GAIA, in Phases 1 and 2, we incorporated Kaldi (<https://kaldi-asr.org/>) models for languages for which there was sufficient training. In Phase 3, audio segments were excluded from evaluation and thus the ASR component was not included in our final system.

3.2.8 Text within Video and Images

Similarly, the extraction of text from images was not a primary focus of GAIA's work, but we explored the inclusion recognized text from text annotations within both images and video using Optical Character Recognition (OCR) using a combination of YOLO based approaches to detect the presence of text within an image and ISI's VISTA OCR to recognize the text. We explored improving domain-specific accuracy by augmenting the models with unknown vocabulary (typically names) found within the text data for a domain. This approach improved the ability to e.g. recognize non-English names. Within the overall GAIA pipeline, recognized text was passed as sentences to the text-based knowledge extraction elements. We found two primary challenges for the inclusion of OCR in the overall GAIA pipeline. First, distinguishing between content bearing text (e.g. a video caption running below a news report); and second, we found that there were sufficient differences between the linguistic characteristics of the text within video (e.g. frequently a name appearing in isolation to identify the speaker), that the knowledge element

models were less accurate. Further, in review of the evaluation results we often found that the content extracted via the OCR path was overlapping with information found through the core text path, thus in our final Phase 3 evaluation system we did not include the OCR-based path.

3.3 Aggregating a Knowledge Base

3.3.1 Linking to Wikidata

We propose a novel candidate retrieval paradigm for Wikidata based on entity profile generation. Wikidata entities and their textual fields are first indexed into a text search engine (e.g., Elasticsearch). Given an entity mention and its context, we use a seq2seq model to generate the profile of the target entity, which consists of its title and description. The profile is then used to query the indexed search engine to retrieve candidate entities. Our technique is applicable to virtually any KB, not just Wikipedia or Wikidata. It also complements the dictionary-based approaches, enabling us to further design an effective hybrid method for candidate retrieval.

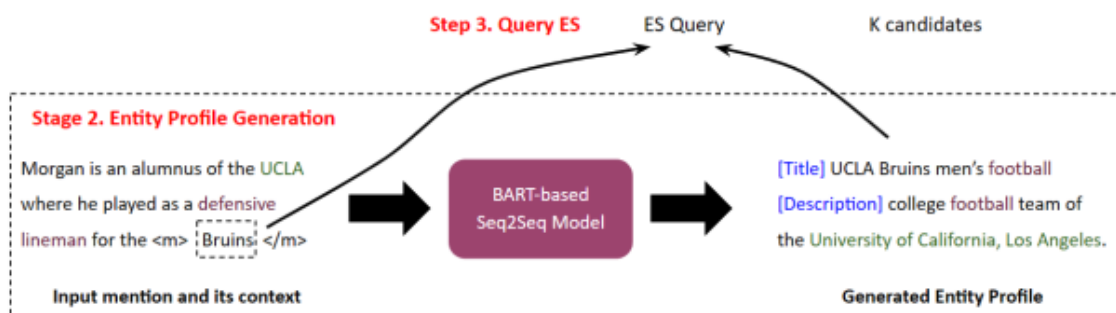


Figure 19: Candidate retrieval based on entity profiling

3.3.2 Constructing a Knowledge Base of Extracted Knowledge Elements

To support program evaluation, we fused the document-specific knowledge base comprised of the knowledge elements extracted from a single published (e.g. the images and text associated with a single news article) across a stream of documents. For purposes of program evaluation, concrete, named entities served as the primary entry point into the knowledge graph. For these entities, we relied on either the global IDs that were the result of linking to Wikidata (see Section 3.3.1) or name string similarity, heuristically selecting a best string from the set of co-referent mentions within a document. While we explored other features (e.g., incorporating events and relations from the text extractions as an additional means of disambiguation), we found that such features were sufficiently sparse and noisy that they did not yield improved performance.

Similarly, while we explored approaches to aggregating unnamed knowledge elements (e.g. the events, instances of e.g. masks), we found that there was sufficient possibility for overlinking to cause catastrophically large clusters (e.g. treating all attacks as a single attack), that we did not incorporate non-named linking into our overall pipeline.

3.3.3 Knowledge Graph Tool Kit

In addition to developing approaches to aggregate knowledge elements extracted from disparate sources into a unified knowledge base, the GAIA effort developed various tools to enable easier development with large knowledge graphs, and to support AIDA's more general shift to a Wikidata based ontology. We refer to these capabilities as the Knowledge Graph Tool Kit. Below, we describe specific elements of KGTK developed under GAIA.

3.3.3.1 Browsers for intuitive knowledge exploration

We used Web technologies to develop an intuitive and scalable browser for knowledge graphs represented in the KGTK format. A snippet of our browser, with information about the concept dog (Q144) in Wikidata, is shown in Figure 20. The browser is developed by using state-of-the-art Web technologies for its front-end design, and by employing scalable backend tools to support quick rendering. Specifically, we use our in-house query language Kypher for retrieving of the browser data, and we use Elasticsearch to perform text search. The browser is designed to support the KGTK data model, which encapsulates other structured formats like the Wikidata model and variants of RDF.

The browser has several unique functionalities. Besides a text search, it also shows the incoming links for the rendered object, which is sorely missing in most knowledge browsers. As shown in Figure 21a, from the incoming links we obtain, e.g., specific instances that point to the concept of dog or diseases that afflict dogs. Another feature of the KGTK browser are its item profiles: aggregated graph statistics that provide statistics about each item in Wikidata. For instance (Figure 21b), we see that nearly half of the dogs in Wikidata have a country of origin and 70% have an image.

The entire browser code is available on GitHub: <https://github.com/usc-isi-i2/kgtk-browser>.



Figure 20: KGTK's Web Browser.

From Related Items			Profiling Data		
instance of	←	Buddy	pagerank (undirected)	4.889954573788569e-06	
		Jasper		series ordinal	4723
		Kratu		properties for this type including subcl...	
subclass of	←	Leuk	proportion	100.00 %	
		Mabrouk		quantity	300
		Max		instance of	
afflicts	←	Ned the Dog	proportion	73.00 %	
		Peter the Great		quantity	219
		Tarzan		image	
animal breed	←	Yoda: use the power	proportion	69.33 %	
		< 1 2 3 4 5 ... 45 >		quantity	208
		447 values		Freebase ID	
appears in the form of	←	Austrian Black and Tan Hound	proportion	68.33 %	
		Corpschund		quantity	205
		Frenchton		Commons category	
nature of statement	←	inuit sledge dog	proportion	68.00 %	
		Japanese dog		quantity	204
		male dog		country of origin	
rarely	←	Mantiqueira Shepherd	proportion	48.00 %	
		Molossus of Epirus		quantity	144
		Swinford Bandog		topic's main category	
	←	Q1111549	proportion	45.67 %	
		< 1 2 3 4 5 ... 18 >		quantity	137
		178 values		Commons gallery	
	←	diabetes in dogs	proportion	39.67 %	
		Dirofilaria repens		quantity	119
		limber tail syndrome		YSO ID	
	←	Bond Forger	proportion	31.67 %	
		NIIC the Singing Dog		quantity	95
		Office Assistant		catalog code	
	←	< 1 2 3 4 5 ... 14 >	proportion	31.33 %	
				quantity	94

Figure 21: KGTK's web browser browser functionalities. a) incoming links from related items; b) profiling data

3.3.3.2 Kypher-V for hybrid data science queries

Building hybrid systems that jointly search over embeddings and structured knowledge is needed for different applications, such as similarity-based search, recommendation systems, and question answering. Hybrid systems that combine structured and unstructured data have the promise to amplify the best features of both while compensating for their weaknesses. For example, structured data is richly described and provides useful constraints for high precision results, but it lacks a notion of similarity and is generally incomplete. Unstructured data is often very large-scale, high-coverage, captures most relevant semantic dimensions automatically and can be used to query by example, but it lacks constraints and becomes less precise with increasing number of objects. Effective combination of structured and unstructured sources is still an active research area and generally requires extensive experimentation.

To support this emerging new area of large KG + vector-based systems, we built Kypher-V: an extension to KGTK's Kypher query language and processor, that allows users to augment Wikidata and other KGs with large-scale vector data and efficiently query it through a hybrid query system using a structured query language based on Neo4J's Cypher language. Kypher-V allows users to augment Wikidata and other KGs with large-scale vector data and efficiently query it through a hybrid query system using a familiar structured query language based on Cypher. Both Kypher and Kypher-V use the KGTK representation and can therefore be used not just on Wikidata but also RDF KGs, for example, DBpedia, and others.

Kypher-V represents vectors as special kinds of literals that are linked to KG objects via edges in KGTK format. When vector data is queried, it is imported and cached if necessary just like any other KGTK data processed by Kypher. The main difference is that for faster query processing, vector literals get translated into binary format first before they are stored in the Kypher graph cache based on an indexing directive. Various up-front transformations such as data type conversions, normalization, storage type, etc. can also be controlled here. Finally, vector files are treated as large contiguous and homogeneous arrays, so that vectors can be referenced and accessed efficiently by vector table row ID instead of having to use an index.

This scheme stores vectors on disk and brings them into main memory on demand during a query just as any other KG data. Since vector literals can be quite large, it is important that these accesses are fast and minimized to maximize query speed. For more limited applications, it is possible to precompute similarities and only store the results instead of the full vector data. However, we chose to explicitly store vectors to allow more open-ended queries, comparison against dynamic data (e.g., vectors from user queries), as well as vector management in general which can facilitate downstream applications such as machine learning. Kypher-V can store and simultaneously query any number of different types of embedding vectors. The only requirement is that when two vectors are compared in a query, they must have the same dimensionality and should come from the same vector set, since it does not make sense, for example, to compute a similarity between a ComplEx and TransE vector. Just as the core Kypher system it is based on,

Kypher-V is optimized for read-only processing. Nevertheless, updates and dynamic inputs can be handled but should be limited to small-batch data for best performance.

The following types of vector queries are supported by Kypher-V:

1. Vector computation: given one or more vectors accessed in a query, compute a function based on those vectors. Currently supported vector functions are inner product, cosine similarity and Euclidean distance.
2. Vector search: traditional similarity search that finds the top- k similar vectors and associated nodes for a given input. Only cosine similarity is supported at the moment.
3. Vector join: a generalized form of vector search that filters input vectors and output vectors based on arbitrary Kypher query patterns.
4. Multi-vector computations: a planned extension is to be able to aggregate similarities from multiple vector types, as needed, for example, by the topsim computation of KGTK's similarity service.

3.3.3.3 Automatic completion methods for knowledge engineering

KGs are important sources of information assisting in various analytical and decisive applications, for which carefully engineered knowledge is essential. Despite being a rich source, KGs are often incomplete, which can be either due to the occurrence of new events or due to some facts that were missed or deliberately left out while creating the KG. Hence, KG completion (filling in information gaps in a KG using other sources) is a promising avenue to enhance the usability of KGs, e.g., for downstream tasks. We experiment with two orthogonal approaches: knowledge distillation from another KG (DBpedia) and temporal information extraction from Wikipedia.

We also extracted 2K facts for training via distant supervision. We generated an intermediate dataset containing fact-evidence pairs using DPR. We automatically labeled the fact-evidence pair and create 4 different types of training datasets using the following strategy:

1. We used the Bert-based ner-extraction model to extract person entities from the evidence.
2. We attached the extracted person entity to the spouse-pair if either the extracted name is part of a spouse's name or vice-versa.
3. If the evidence contains a Person entity for both the spouses, we add it to the direct set with the label "yes", having 2 fact-validation samples based on the questions in Table 1. Again, if the evidence contains partial names, we add it again to the indirect_pos set with 2 fact-validation questions formed using partial names.
4. If the evidence doesn't contain a Person entity for any of the spouses, we add it to the direct set with the label "no", having 2 fact-validation samples based on the questions in .
5. Also, we create negative samples using extracted person entities which don't match any person in the spouse fact. We add it to the indirect_neg set.
6. Based on the above-created sets, we have 4 types of completely balanced datasets with equal counts of each label having sizes of 500, 1000, 2000, 4000, 8000, and 20000. These 4 types of datasets are:
 - a. Direct dataset: Contains samples from only direct set
 - b. Indirect_pos dataset: Contains samples from direct and indirect_pos sets.

- c. Indirect_neg dataset: Contains samples from direct and indirect_neg sets.
- d. All dataset: Contains samples from all the sets.

The training datasets can be found at: <https://github.com/usc-isi-i2/wikidata-fact-verification/tree/master/data/unifiedQA>

To distill knowledge from DBpedia, we first extract all the spouse data facts (P26 in Wikidata and property:spouse in DBpedia), which are present in DBpedia but missing in Wikidata, creating the list of candidate facts for addition. For all the candidate facts, we perform document retrieval using Dense Passage Retrieval (DPR) from Wikipedia pages by forming queries as shown below. For DPR, we use the pyserini library.¹ After extracting multiple support/evidence for each spouse-data fact, we validate the candidate facts by applying fact-validation over the supports for each fact.² Hence, we expect to obtain higher quality validated new facts for enhancing Wikidata. As data, we use information that we extracted from DBpedia in another work [ZH22], which provides DBpedia statements corresponding to infoboxes in Wikipedia.

DPR Spouse data queries	1. Is {person_one} married to {person_two}? 2. Is {person_two} married to {person_one}?
-------------------------	--

Figure 22: DPR query to retrieve spouse fact support documents.

The extraction of temporal qualifier information is motivated by the sparsity of this information. For example, 68,159 spouse (P26) triples in Wikidata have a time qualifier (P580) and 69,207 triples do not. In this effort, we extract time qualifiers by combining two state-of-the-art tools: DPR for retrieving documents and UnifiedQA for text generation (Figure 23). Given a triple in Wikidata, e.g., A-spouse-B, we formulate the question by using our Kypher query language: When did A marry B? We obtain k=100 top document parts from Wikipedia by using DPR. These paragraphs are filtered based on the entity name of the first entity. For the remaining paragraphs, we use UnifiedQA and provide it the paragraph(s) as evidence. We also fine-tune UnifiedQA on artificially created data based on the links between Wikidata and Wikipedia.

¹ <https://github.com/castorini/pyserini>

² <https://huggingface.co/fractalego/fact-checking>

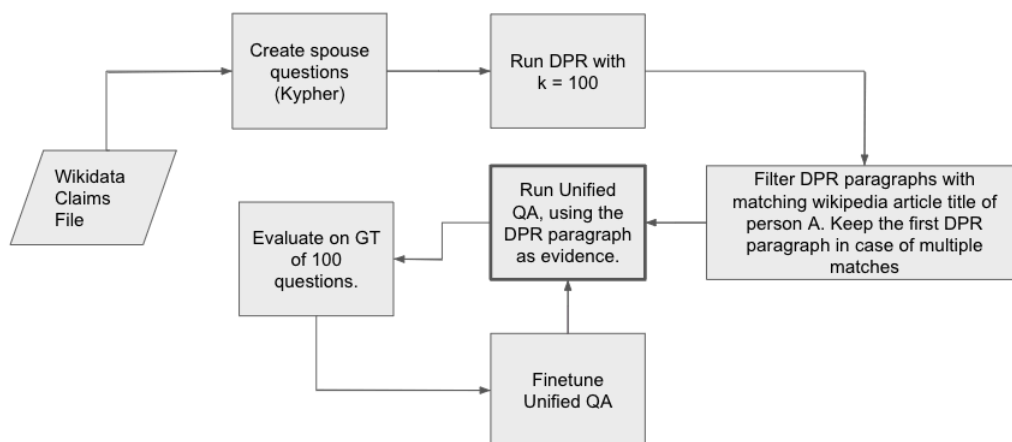


Figure 23: Temporal information extraction architecture.

3.4 Discovering Claims

HypoGator, was developed as a part of GAIA specifically to generate alternative claims given query claims both with and without query relations, and with just a just a query topic. It includes the structural output from the TA1(document level) and TA2 (corpus integrated) components. Input also includes natural language descriptions of all candidate claims.

To select the proper candidate claim, our system first learns and detects the cross-claim relations between different claim pairs, then selects the candidates with the ideal cross-clam relation. To better leverage different types of information provided for each claim, we implement two cross-claim detection modules: one using the structural information and another exploiting the natural language description for each claim. All candidates are then clustered based on structural attributes and ranked with their relevance to the query topic and the popularity of their clusters. Finally, HypoGator selects the answer claims with one or two steps of claim selection based on different queries to guarantee the diversity of claims in our answer claim list.

Figure 24 illustrates the HypoGator pipeline, which includes several modules, initially to discover claims using the extraction results and then to compose, rank and select claim frames (CFs). The result is a set of diverse, relevant claims.

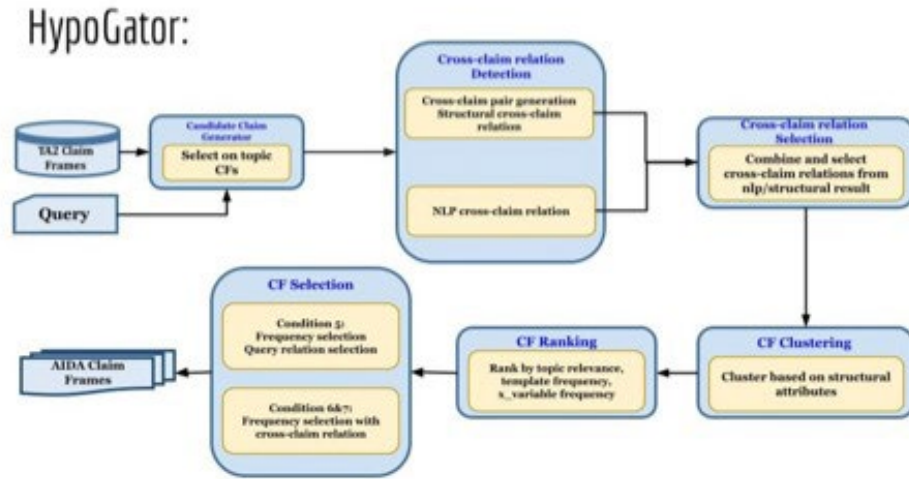


Figure 24: Pipeline for HypoGator Claim Extraction

3.4.1 Cross-Claim Relation Detection:

Structural Cross-claim relation detection using KGTK: Cross-claim relation detection is one of the most important steps in TA3 end-to-end system. Following the definition of different relations (refuting, supporting, and related), HypoGator can categorize given claim frame pairs into proper groups, leveraging different attributes of claim frames and KGTK's Qnode similarity service. Specifically, we implement a decision tree with attributes including topic, x-variable, claimer, and epistemic status. Intuitively, we use string similarity for topic, x-variable, and claimer matchings. Since there are dependencies between x-variables and claimers that can not be captured by just the string, we further leveraged KGTK's Qnode similarity service for x-variable and claimer matching. Since different embeddings may capture different types of relations, we explored various types of embedding provided in DWD's Qnode similarity system for better performance.

Also, we integrated KGTK's bulk mode into HypoGator which improves the efficiency of HypoGator. Considering multiple type Qnodes may help to more accurately identify cross-claim relations, we discuss with our TA1 group and advanced our model so that it can efficiently deal with more than 1 Qnodes for each entity. We also work on a link prediction method with certain relations to further improve the quality of our cross-claim relation detection result.

NLP-based Cross-Claim Detection: NLP-based cross-claim detection aims to classify the relationship between two natural language claims into one of three classes: supporting, refuting, and neutral. To study this task, we construct a new dataset that contains 10K natural languages claim pairs. We experiment with a state-of-the-art language model through a two-step pre-

training regime. In particular, we fine-tune the pre-trained mT5 on the MultiNLI dataset, then fine-tune the resulting model again on our newly annotated 10K claim pair dataset and on the LDC’s 20K dataset respectively. Two different claim expression styles are compared: manufactured (Template + epistemic status) and natural (provenance sentence). Our best-performed model is fine-tuned over the LDC’s annotation using natural claim pairs.

3.4.2 Claim Selection through Combination of Structural and NLP Results

After exploring different groups of TA1 & TA2 data, we observe that NLP models achieve better performance in identifying supporting & refuting relations with a much higher F1-score on supporting relations. Besides, the NLP-description model outperforms the NLP-template model in supporting and refuting relations identification. Based on the above, we conclude that textual information is helpful for cross-claim relation detection and description provides more information than template + x-variable. Thus, we apply different models for different relations detection as follows: the structural model is used for identical and related claims detection. In contrast, NLP model is used for supporting and refuting claims detection. With such a combination, HypoGator is now able to leverage both structural and natural language information of claims when identifying relations between claims.

3.4.3 Claim Frame (CF) Clustering

To select more diverse claims into the final answer list, we cluster all candidate claim frames based on their x-variables and templates. The same Qnode service for x-variable similarity as in cross-claim relation detection is used here while exact string matching is used for template matching.

3.4.4 Claim Frame (CF) Ranking

To guarantee diversity in final answer selection, we rank those claims in less frequent clusters and epistemic status are ranked higher by the system. Since we also prefer claims that are more related to the query topic or query claim, we leverage the confidence score provided by our TA1 and the similarity score between x-variables to identify more relevant to the query. For efficiency concerns, we keep the previous x-variable similarity result in the cross-claim detection step locally and reuse it for the cluster and ranking modules.

3.4.5 Claim Frame (CF) Answer Selection:

The general idea of this module is to first select claims from uncommon clusters, then select refuting claims from other uncommon clusters (if possible) to include diverse claims in the answer list. Answer selection is tailored to the conditions of the evaluation which are:

- Condition 5: In this task, a complete query claim is provided with a query relation. The system is asked to generate a list of claims that hold the query relation with the query claim.
- Condition 6: In this task, the input to the system is a query claim with a template and a topic. The system is expected to generate 3 lists of claims that share the same topic and have related, supporting, and refuting relations with the query claim.
- Condition 7: In this task, instead of a complete claim, only a topic is provided as the input. The system is still expected to generate 3 lists of claims that share the same topic and have related, supporting, and refuting relations with the query claim.

To address the specific needs of Condition 5, HypoGator divides answer selection into two steps:

- **Global selection:** In this step, HypoGator uses the cross-claim relations generated in the corresponding module from each cluster following the ranking. The idea is still to include more diverse claims in the answer list.
- **Query relation selection:** To include more claims that are more related to the query claim, we focus on selecting more refuting and supporting claims which usually have more similar x-variable to the query claim. x-variable similarity and claim confidence score are reused in this step.

For Condition 6 and 7, the tasks focus more on relevancy with the query topic and relations between candidate claims themselves. Thus, we only keep the global step for them, where we still use cross-claim relations among candidate claims and clusters to keep diversity and TA1 claim frame confidence score for the relevancy concern.

3.5 Integration into Program Testbed

The GAIA TA1 pipeline was composed of many independent, portable docker images. In addition to the docker images, the GAIA TA1 pipeline definition also included container orchestration. The component docker images in the GAIA TA1 pipeline ran in serial with component orchestration that detailed how files would be shared from one step to the next. The GAIA TA2 and TA3 implementations were each a single docker image. For all technical areas, each executable component was defined as a docker image with its own embedded resources. All resources including software code, packages, and large pre-trained models needed for computation had to be packaged in a docker image for evaluation.

CACI implemented a cloud-based Kubernetes environment where all TA1 pipelines were run. The program requirements directed that we create portable docker images to be submitted to the CACI environment. CACI's Kubernetes environment provided a virtual location to execute pipelines in a repeatable manner. CACI's environment also hosted the LDC corpus including all the media files to be processed by the pipelines. The Phase III evaluation data was not available to performers outside of CACI's environment until after the evaluation had completed.

Debugging of components was not allowed on the CACI infrastructure. Instead, the GAIA team developed file-based logging that was available at the conclusion of a pipeline run.

The evaluation LDC corpus included media in many formats and multiple languages. The large volume of data presented a challenge in testing a pipeline composed of many serial components. The pipeline design allowed components to run in sequence only. Errors in the earlier steps could cause pipeline failures. In these cases, the steps at the end of the pipeline were not executed and therefore any hidden errors in those components could not be easily identified until the earlier components were corrected. This placed a high priority for ISI on testing and correcting the earlier steps in the TA1 pipeline so that execution of the entire pipeline could be completed.

At ISI we worked with team members to develop and test the docker images. We followed the guidance of CACI by adopting common conventions on the use of the docker images. These conventions included using environmental variables passed into the docker containers that could be used by the containers at run-time to configure their performance. Additional conventions included naming container volume mounts for organizing inputs, outputs, and intermediate results.

While using the CACI environment we encountered some difficulties in testing. Execution times were very long, and diagnostic information was difficult to acquire. With CACI, we developed improved methods for logging system output and errors. We also constructed a small test corpus made of a subset of available LDC documents. By running on the small test set we were able to identify errors more quickly and to verify when updated docker images successfully corrected those errors.

One challenge in creating the docker images was the inclusion of large pre-trained models. Any component that used a large pre-trained model needed to include the model in the docker image delivered to CACI. We primarily built the images and published them to Dockerhub where they were available to the CACI environment. Each time a model was rebuilt or any software code within the component was changed the docker image would need to be rebuilt and given an updated label. The updated docker image would then be republished to Dockerhub, which would then require CACI to pull down the image which would often be many GB in size and would take time to download.

Another challenge in testing and delivering the docker images was in the use of graphical processing units (GPUs). The use of GPUs was required by components that used CUDA libraries to make efficient use of GPUs for computation. In order to operate the GPU hardware, the NVIDIA device driver was required. Consequently, docker image definitions needed to include the installation of the NVIDIA device driver inflating the size of the images.

The most difficult errors to identify and correct were those that were found only at run-time. These errors could have a variety of causes, and in some cases only appeared when processing a problematic input, such as a document containing unusually long sentences or unexpectedly

large images. Occasionally out-of-memory errors occurred in these cases. To address these issues, we worked to replicate the errors at ISI. Then we explored memory configuration settings that properly allowed the components to handle the challenging input. Once these memory settings for a component were determined these were included in the pipeline definition so that the same memory settings were applied at ISI and at CACI.

The GAIA TA1 pipeline included components from our team members at UIUC, Columbia University, and USC. Throughout the program, and especially leading up to evaluation, dockers images needed to be updated for many purposes. There were corrections, updated models, and required software package updates.

ISI implemented machine translation as a first step in the TA1 pipeline. This made all non-English text documents available as English translations. Sentences were parsed and translated by a pre-trained model capable of processing all languages used in the AIDA program.

ISI provided support by authoring tools to access LDC documents. GAIA tools were used to present uniform, expected inputs to each GAIA team member contributing a component docker image. The ISI team also contributed a claim detection and semantic extraction docker image that followed the program conventions and was included in the Phase 3 evaluation.

The program defined barriers between TA1, TA2, and TA3 meant that only specified data could be shared from one TA to the next. After the Phase 2 evaluation the separation between the technical areas was relaxed. Additionally, there were changes to the AIDA Interchange Format (AIF). During Phase 3, the AIF was amended to define claim frames, refine attributes on entities and events, and NIST validation rules were updated.

Testing at CACI only allowed the entire TA1 pipeline to be executed. At ISI, we also tested individual docker images. By using docker images to launch docker containers at ISI, we could inspect the running processes and debug issues in the containerized code.

Docker presents security challenges on high-performance computing cluster. Any running docker container essentially has root access to its host machine. This is a concern on a shared resource such as the ISI HPCC because a docker container process could access system files that are normally unavailable to users. This required ISI and team members to be aware of the behavior of their software and make efforts to ensure no system files were impacted.

At both CACI and ISI docker containers ran based on the same docker images. CACI and ISI environments used different hardware and different orchestration. At CACI, cloud provider offered a different class of machines than available at ISI (eg.g g4dn.4xlarge on AWS vs ISI's RTX machines. In both cases recent GPUs were required to be provisioned with updated drivers. Some errors could only be detected by running at scale. We encountered failures when not giving enough shared memory to docker containers when running over large documents. We were able to correct these failures on subsequent runs by determining and specifying required memory resources. Kubernetes at CACI used cloud resources that were inaccessible to GAIA team members.

3.6 Visualizing and Interacting with Extracted Knowledge

3.6.1 Claim Radar

An initial version of Claim Radar was developed to demonstrate the extraction of claims from text for a single GAIA team specifically on COVID-related data. We then extended this initial capability to achieve greater interoperability and explore additional user-focused interactions.

Interoperability. We achieved interoperability with the program's official format (AIF) via a conversion script that transformed AIF to Claim Radar internal JSON. This was designed as a stand-alone data-format conversion process to ensure that the core interoperability capability would be maintained with the originally developed Claim Radar, independently of modifications that ISI made of Claim Radar.

The converter takes the official AIF output from the programs three teams: GAIA, OPERA, and BBN, and converts them into JSON. This JSON is preprocessed before being fed into UIUC's application. Some of the challenges in developing the conversion method are as follows.

Precise span offsets. One of our biggest challenges was obtaining the span offsets of text justifications. While this data was usually provided for English data, not all justifications were given these offsets, and sometimes the only offsets provided were for the spans in the source language. We envisioned displaying the text justification values in English, but this was almost impossible in some cases. As a result, text from OPERA and BBN ended up being displayed in their source languages.

Variation in value assignment. Certain values were assigned to different fields, including text justifications and event types. Some were not assigned to a certain field, but rather embedded in the object ID. The methods used to extract these values depended on provenance.

Missing types from GAIA TA3. Some type labels from TA2 did not carry over to the TA3 output. We relied on reading those types from TA2 to provide the type labels.

Nonexistent values. Some teams did not provide the values that others did. A common reason was because they were not required for evaluation. One example was for the sentence values of claims; some teams did not provide these, so we resorted to using the claims themselves as the "sentences." For the most part, we had to accept the inability to display all fields for each team.

Interactivity and Additional Content. Additionally, we extended the initial published COVID Claim Radar, which served primarily as a browser of extracted information, to incorporate richer user interactions around editing information. In preparation for this, ISI performed a reimplement of Claim Radar's interfaces to support a more flexible software environment. This work included extensions as follows.

User authentication. We implemented a simple user authentication system to prevent unauthorized users from creating changes to the claim data.

Claim data editing. When logged in, a user can take these exclusive actions on the fields Claimer, Claimer Affiliation, Stance, and Claim Object:

- Reject the individual value
- Reject the value for that field across all claims (useful for values that don't make sense in any context, e.g. a Claimer value of "said")
- Undo the rejection

Users cannot modify a value that has already been rejected by another user.

When a value is "rejected," it does not alter the data in the original input file. To ensure preservation of the original data, change data is saved to a separate document. The data in that document gets applied to the information on display.

Allowing simple rejection was intended to be the first step, with the ability to replace the old value with a custom one to potentially come later.

Viewing Modes. There are three table viewing modes available in the visualizer.

- **View all corrections:** displays the claim data with corrections made by all registered users.
- **View your corrections:** displays the claim data with only the corrections made by the active user – if not logged in, no corrections will be shown.
- **Hide corrections:** displays the original claim data. To prevent potential confusion, corrections cannot be made in this mode.

New KE Information. In the main table view, we added two additional columns: *Associated KEs*, which lists the text representing each knowledge element (excluding relations) related to a given claim, and *Associated KE Types*, which lists each unique type label assigned to the claim's associated KEs.

We also added and modified the content under Associated Knowledge Elements in Claim Details. New in our version is a table listing each knowledge element (excluding relations) with their ID, text values, and type label. Each KE also has three values associated with it: the number of KEs throughout all claims that have both the same text and type label, the number of KEs that have the same text but another type label, and the number of claims that have the same associated KE.

In the document context of the associated KEs, we highlight any arguments a KE may have by clicking on the KE. When an argument KE is highlighted, hovering over it will display the argument role it fills.

Three new fields can be queried for: Language, Associated KEs, and Associated KE Types. Querying KE text and KE types will return claims that contain those respective text or type values. If the user selects "Bind types to KE text" before executing the filter, only claims with the queried KE text which have the queried KE type(s) will be returned. These additional filters would potentially reveal more interesting patterns in the claim data.

Images. Finally, we introduce support for image display. For claims identifying an image source, the image is displayed in the claim details dialog.

As results, we illustrate some of these extensions with screenshots in in Section 4.6.1.

3.6.2 SmartBook: AI-Assisted Situation Report Generation

As an additional visualization, we developed an automatic situation report which is organized in the form of timelines to provide a coherent, chronological representation of event developments, facilitating users tracking and understanding of situation context. Major events identified via clustering news articles within the 2-week timeline duration serve as the foundation for corresponding chapters. To further guide detailed chapter analysis, we incorporate a logical structure through automatically generating section headings in the form of strategic questions relating to each major event. SmartBook generates chapter content with strategic perspective by using such questions covering various strategic aspects of the event. The individual sections comprise the core content of our situation report and contain grounded, query-focused summaries which address the strategic questions providing readers with a comprehensive understanding of event context and implications. The generated summaries feature citation linking, which grounds each summary fragment in its factual input context and allows experts to cross-check and verify information as needed. Ultimately, the intuitive and structured incorporation of timeline chunking, major event chapter clustering, and query focused section summaries within chapters enables our proposed SmartBook formulation to generate reliable, insightful reports for time-sensitive, emerging situations.

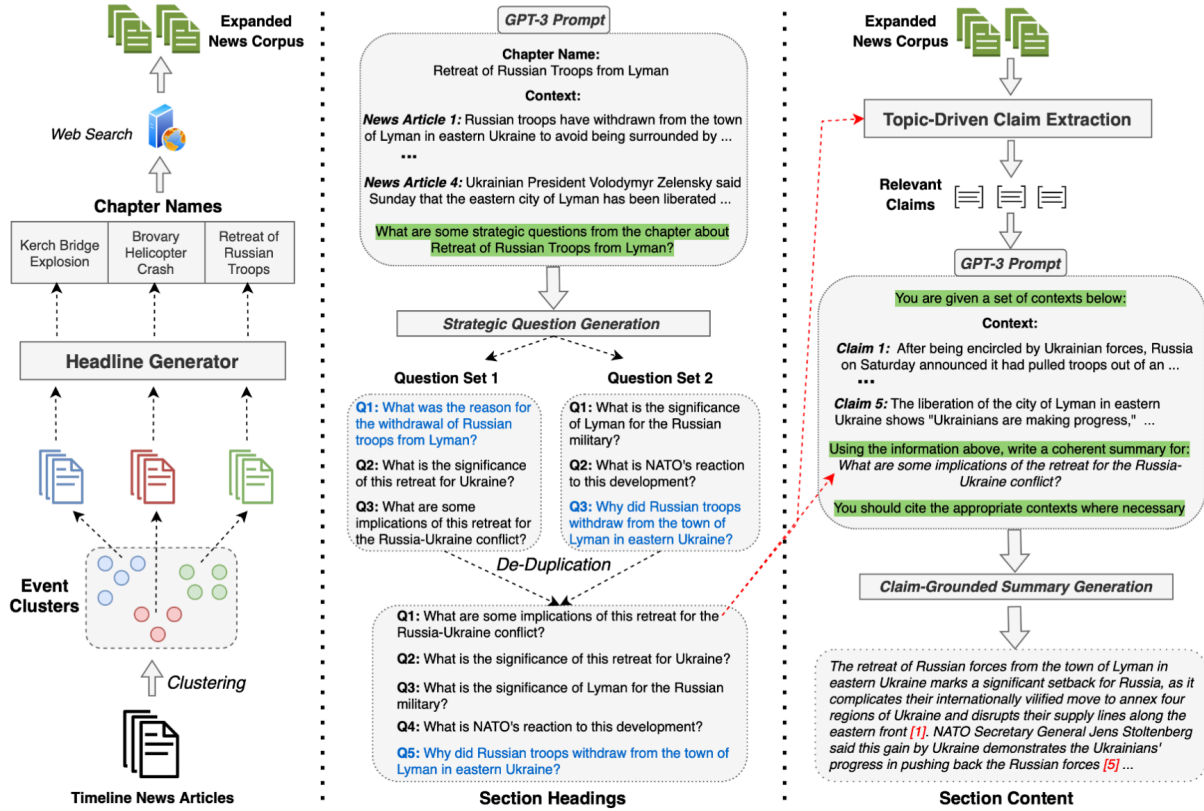


Figure 25: Overall workflow for constructing SmartBook

4 RESULTS AND DISCUSSION

4.1 Extraction from Text

4.1.1 Sentiment, Stance, and Document Internal Claims

Sentiment Relations: Table 1 shows the performance of our model (trained attention, last row) in comparison to several baselines. AttnCNN is the CNN model by Wang et al. (2016); TreeLSTM is a tree model with LSTM units by Tai et al. (2015); SDP is the Shortest Dependency Path model by Xu et al. (2015); AttnLSTM is our baseline model by Zhang et al. (2017). Pred-rationales is a popular multi-task learning method for leveraging rationales, and trained-attn is our novel method for training the attention. Both are applied to AttnLSTM. Improvements of Trained-attn over non-trained attention, shown in parentheses, are statistically significant with $p < 0.002$. Likewise, improvements over Pred-rationales have $p < 0.02$.

Trained attention brings a significant improvement on F-score over untrained attention (4-8 points); additionally, it improves all the label-wise precision, recall and F-scores on both MPQA and GFBF (see Table 2 and Table 3). It also outperforms the popular multi-task learning method (Pred-rationales) in all tasks.

Table 1: F-score (for Include- $\emptyset$ columns) and accuracy (for Exclude- $\emptyset$) columns for various baselines and our sentiment method.

Model	MPQA Include- $\emptyset$	MPQA Exclude- $\emptyset$	GFBF Include- $\emptyset$	GFBF Exclude- $\emptyset$
AttnCNN	20.3	38.7	34.5	53.6
TreeLSTM	29.9	56.1	46.4	70.6
SDP	34.2	60.7	60.1	81.0
AttnLSTM	32.7	62.4	60.4	80.4
Pred-rationales	34.9	63.0	61.4	85.1
Trained-attn	37.6(+4.9)	68.7(+6.3)	65.4(+5.0)	88.7(+8.3)

Table 2: Performance breakdown on GFBF including $\emptyset$.

label	Precision	Recall	F-score
goodfor	60.2(+5.9)	66.9(+6.3)	63.4(+6.7)
badfor	63.4(+0.3)	71.2(+7.4)	66.8(+3.5)
Average	61.8(+3.1)	69.1(+6.9)	65.1(+5.1)

Table 3: Precision, recall and F-scores for each label on MPQA (including $\emptyset$), averaged across 5 folds. Each entry shows the performance with trained attention, with the improvement over AttnLSTM baseline in parentheses. Results are averaged over 5 folds.

label	Precision	Recall	F-score
Negative	36.4(+2.1)	38.7(+2.5)	36.9(+1.5)
Neutral	30.8(+3.0)	36.5(+4.6)	32.9(+3.6)
Positive	41.2(+9.0)	44.7(+8.1)	42.3(+8.2)
Average	36.1(+4.7)	40.0(+5.1)	37.4(+4.5)

We also evaluated how well the model performs when given varying amounts of rationales. Since it may be hard to manually produce rationales for all instances in a dataset, it is helpful to know how many are required to improve performance. Figure 26 plots this performance against the fraction of data points with human rationales annotation. On the left/right hand side is F-score, Performance/attention loss improvement (y-axis) vs. using different percentages of human-annotated rationales (x-axis). We calculate the difference from attention loss when training with all rationales. To draw the plot, for MPQA, we sample 100, 200, 400 and 800 rationales (corresponding to 4%, 8%, 16% and 33% of all rationales); and for GFBF, 50, 100, 200 and 400 rationales (7%, 13%, 27% and 53%).

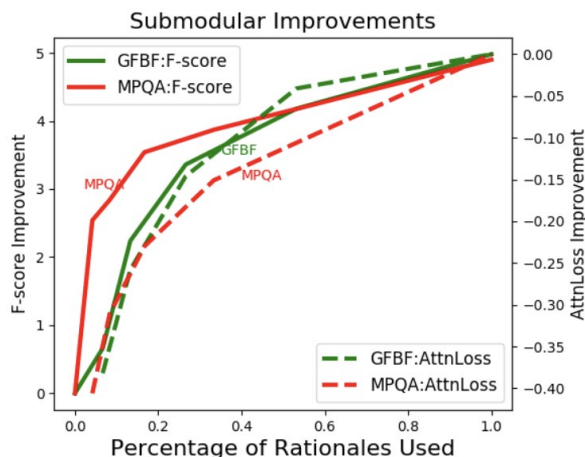


Figure 26: Performance as F-score Varying Number of Rationales

As expected, the more rationales given, the better the performance. Notably, however, the improvement is sub-modular: the marginal improvement per rationale decreases as more rationales are used, and so the first few rationales end up being the most helpful.

Just 100 rationales (4% of all rationales in MPQA, or 7% in GFBF) brings about 2.5 points of improvement. Attention loss follows a similar increasing and submodular trend, for both datasets. Even more surprisingly, we found that while a model could develop a reasonable attention mechanism by itself with just a good amount of data without rationales, training on a few rationales can 'nudge' the attention in the right direction, resulting in significant gains in attention performance. In a separate experiment, we trained a model on just 200 data points and all their rationales, and this performed worse than training on all the data points but a smaller fraction of rationales -- 50 rationales (4%) on GFBF, and 100 (7%) on MPQA. Therefore, fewer rationales can be more effective if there is enough other data.

Stance: We evaluate TGA Net and our baselines using macro-average F1 calculated on three subsets of VAST (see Table 4): all topics, topics only in the test data (zero-shot), and topics in

the train or development sets (few-shot). We do this because we want models that perform well on both zero-shot topics and training/development topics.

Table 4: Macro-averaged F1 for stance on the test set. * indicates significance of TGA Net over BERT-joint, $p < 0.05$.

	F1 All			F1 Zero-Shot			F1 Few-Shot		
	pro	con	all	pro	con	all	pro	con	all
CMa j	.382	.441	.274	.389	.469	.286	.375	.413	.263
BoWV	.457	.402	.372	.429	.409	.349	.486	.395	.393
C-FFNN	.410	.434	.300	.408	.463	.417	.413	.405	.282
BiCond	.469	.470	.415	.446	.474	.428	.489	.466	.400
Cross-Net	.486	.471	.455	.462	.434	.434	.508	.505	.474
BERT-sep	.4734	.522	.5014	.414	.506	.454	.524	.539	.544
BERT-joint	.545	.591	.653	.546	.584	.661	.544	.597	.646
TGA Net	.573*	.590	.665	.554	.585	.666	.589*	.595	.663

We observe that BiCond and Cross-Net, state-of-the-art systems for stance detection, both perform poorly on VAST. Although these were designed for cross-target stance, a more limited version of zero-shot stance, they suffer in their ability to generalize across many targets when few examples are available for each.

We see that while TGA Net and BERT-joint are statistically indistinguishable on all topics, the topic-grouped attention provides a statistically significant improvement for few-shot learning on ‘pro’ examples (with $p < 0.05$). Note that conditional encoding is a crucial part of the model, as this provides a large improvement over embedding the comment and topic separately (BERT-sep). Moreover, TGA Net exhibits a smaller drop in performance on topics that are not lexically similar to ones seen during training compared to BERT (only 0.04 F1 compared to 0.06).

Additionally, we examine the performance of TGA Net and BERT-joint on five challenging phenomena in the data: i) Imp – the topic phrase is not contained in the document and the label is not neutral (1231 cases), ii) mlT – a document is in examples with multiple topics (1802 cases), iii) mlS – a document is in examples with different, non-neutral, stance labels (952 cases), iv) Qte – a document with quotations, and v) Sarc – sarcasm, as annotated by Habernal et al. (2018). Overall, we find the TGA Net performs better on these difficult phenomena (see Table 5). These phenomena are challenging for both models, as indicated by the generally lower performance on examples with the phenomena compared to those without, with the *mlS* especially difficult. We observe that TGA Net has particularly large improvements on the rhetorical devices (*Qte* and *Sarc*), suggesting that topic-grouped attention allows the model to learn more complex semantic information in the documents.

Table 5: For stance, accuracy on varying phenomena in the test set. *I* indicates examples with the phenomenon, *O* indicates examples without.

		<i>Imp</i>	<i>mlT</i>	<i>mlS</i>	<i>Qte</i>	<i>Sarc</i>
BERT	I	.600	.610	.541	.625	.587
joint	O	.710	.748	.713	.657	.662
TGA	I	.623	.624	.547	.661	.637
Net	O	.713	.752	.725	.663	.667

Finally, we investigate the connection between stance and sentiment vocabulary. Specifically, we use the MPQA sentiment lexicon (Wilson et al., 2017) to identify positive and negative sentiment words in texts. We observe that in the test set, the majority (80%) of pro examples have more positive than negative sentiment words, while only 41% of con examples have more negative than positive sentiment words. That is, con stance is often expressed using positive sentiment words but pro stance is rarely expressed using negative sentiment words and therefore there is not a direct mapping between sentiment and stance. As sentiment can be a misleading heuristic for stance prediction, we investigate how susceptible both TGA Net and BERT are to changes in sentiment using corrupted examples. For examples with majority polarity p , we randomly replace sentiment words with a WordNet synonym of opposite polarity until the majority polarity for the example is $-p$. When examples are changed from the opposite polarity ($- \rightarrow +$ for pro, $+ \rightarrow -$ for con), a model that relies too heavily on sentiment should increase performance. Conversely, when converting to the opposite polarity ($+ \rightarrow -$ for pro, $- \rightarrow +$ for con) an overly reliant model’s performance should decrease. Our results show TGA Net is less susceptible to these replacements than BERT-joint (see Table 6).

Table 6: For stance, *F1* on the test set for examples with a majority sentiment polarity (M) and conversion between sentiment polarities (e.g., $M+ \rightarrow M-$). The direction the score a sentiment-susceptible model is expected to change is indicated with the up and down arrows.

		BERT joint	TGA Net
Pro	$M+$	.73	.77
	$M-$	.65	.68
	$M+ \rightarrow M- (\downarrow)$	.71 $\rightarrow$.69	.74 $\rightarrow$.67
	$M- \rightarrow M+ (\uparrow)$	.71 $\rightarrow$.74	.71 $\rightarrow$.70
Con	$M+$	.74	.70
	$M-$	.79	.80
	$M+ \rightarrow M- (\uparrow)$	.76 $\rightarrow$.80	.70 $\rightarrow$.74
	$M- \rightarrow M+ (\downarrow)$	.75 $\rightarrow$.71	.75 $\rightarrow$.74

Cognitive relations (sponsorship, blame, hoaxfraud, source): We evaluate the performance of our multilingual relation embedding space on FewRel, SemEval 2010 Task 8 and on the Spanish subset of TAC KBP 2016 with little to no fine-tuning on task-specific data.

FewRel is a few-shot relation benchmark where, given a set of n relation types with k relation statement examples for each, the task is to identify which of the n relation types is expressed by a query relation statement q . To do so, we embed q and all candidate relation statements $c_1^l \dots c_n^k$, find the candidate c_i^j whose embedding is most similar to q and predict relation type i . In addition to English, we translate 10% of FewRel’s development set with Google Translate (which performs well on Spanish).

SemEval 2010 Task 8 is a standard relation classification benchmark. TAC KBP 2016 is a Spanish knowledge base population corpus; though it is typically used to evaluate slot-filling tasks, it contains documents with labeled relations. For both corpora, following Soares et al. (2019), we append individual dense layers to our embedding space to map relation representations $h \in R^d$ to their respective classes.

Table 7: Zero-shot (no fine-tuning) accuracies on the FewRel development set.

Denoising	Model	Langs	# Ex.	$n = 5, k = 1$		$n = 10, k = 1$	
				en	es	en	es
	mBERT base	n/a	n/a	62.9	63.8	58.7	50.6
	BERT base	n/a	n/a	68.7	-	57.1	-
	BERT large	n/a	n/a	72.9†	-	62.3†	-
	BERT large + MTB	en	300mil+	80.4†	-	71.5†	-
date-window grouped	mBERT	en	100k	72.4	67.5	59.3	56.7
event-guided (EvtGD)	mBERT	en	50k	78.7	75.1	69.0	64.3
event-guided (EvtGD)	mBERT	en, es	70k	77.5	76.7	68.1	64.4

Table 8: F1-scores on two standard supervised cognitive relation extraction benchmarks for (left) English (SemEval 2010 Task 8) and (right) Spanish (TAC KBP 2016). EvtGD indicates event-guided denoising.

Model	Corpus Size	1%
BERT large	n/a	28.6†
BERT large + MTB	300mil+ ex	31.2†
mBERT	n/a	28.5
mBERT + EvtGD	50k ex	33.3

Model	1%	10%
mBERT	33.0	72.8
mBERT + EvtGD	41.5	76.8

Our event-guided denoising approach performs well with little to no fine-tuning on task-specific data. We achieve results comparable to or better than those of Soares et al. (2019) (MTB) under similar fine-tuning constraints despite training a much smaller model (110mil vs 340mil parameters) on much less training data (50k vs 300mil+ examples).

On FewRel, with no fine-tuning for either model, we achieve only a 2% reduction in accuracy (on a benchmark where fine-tuning with even a small amount of data produces gains of over 5%) (Table 7). For each test case, k is the number of relation types and n is the number of examples per relation type. The development set is used for comparison with the dagger-ed results from MTB.

On SemEval, when both models are fine-tuned on just 1% (80 examples) of its training data, our model outperforms Soares et al. (2019) (Table 8 left). These scores are F1-scores on a random held out slice of 1500 examples from the SemEval 2010 Task 8 training set (for comparison with the dagger-ed results from MTB). Models are fine-tuned on 80 examples. These results align with an observation from their subsequent fine-tuning on the FewRel training set – namely, what was most critical to improved performance was the number of different relation types seen during fine-tuning, not the number of examples for each relation type. Thus, we hypothesize that our event-guided approach is producing a small but diverse set of high-quality relations.

Our multilingual approach also outperforms mBERT on the FewRel (auto-translated) and TAC KBP Spanish subsets (Table 7, Table 8 right). Table 8 right shows F1-scores on the TAC KBP 2016 Spanish relation corpus (from our en/es model, fine-tuned on 1% and 10% of its data and evaluated on the rest. On TAC KBP, our approach achieves a higher F1 than mBERT when fine-tuned on the same small percentages of task-specific data. After 20% (> 500 examples), the two models' performances converge. These results, in English and Spanish, demonstrate the efficacy of our event-guided denoising approach on resource-constrained relation extraction tasks.

Epistemics & Source: We evaluate the performance of our models on PolNeAR (Newell et al., 2018), MPQA (Deng et al., 2015), FactBank (Sauri et al., 2009) and GAIA's manually annotated COVID-19 claims, using standard classification metrics to choose the best performing variants for the evaluation. As is clear from the results (Table 9), all three tasks see improved performance when included as part of a unified multi-task architecture, revealing the semantic entanglement among them. Moreover, moving to an NLI based model led to clear improvements across several of the tasks.

Table 9: Single task, multitask and NLI-based performance on PolNeAR (claimer), MPQA (sentiment), FactBank (epistemics) and GAIA’s manual annotations (claimer and epistemics).

	PolNeAR (accuracy)	MPQA (macro F1)	FactBank (macro F1)	GAIA Claimer (accuracy)	GAIA Epistemics (accuracy)
claimer only	0.750			0.501	
sentiment only		0.715			
epistemics only			0.715		
multitask	0.752	0.771	0.777	0.513	
multitask + NLI	0.751	0.628	0.782	0.555	0.615 (5-class) 0.717 (3-class)

To improve the performance of our NLI-based system, we explored intermediate training of our model on a corpus of discourse connective delimited sentences extracted from newswire and Wikipedia across a wide set of languages. We evaluated this system in a zero-shot setting on XNLI (Conneau et al., 2018) by training a classification head on its sentence representations using English-only training data. In addition to exploring the effects of our intermediate training, we measure the improvement due to our novel cross-attentional architecture. As is clear from Table 10, both of these methods result in improvements over the baseline. We are currently exploring different approaches to fine-tune the entire model on English only data while preserving this strong zero shot multilingual transfer performance.

Table 10: Zero-shot transfer (from English to other languages) performance on XNLI.

XNLI Lang	XLM-R Base <i>en -> target lang F1</i>	XLM-R Base + Discourse Intermediate Training (cls) <i>en -> target lang F1</i>	XLM-R Base + Discourse Intermediate Training (cross) <i>en target lang F1</i>
ar	0.483	-	0.597
de	0.568	-	0.640
el	0.560	0.575	0.632
es	0.490	0.638	0.668
fr	0.482	0.622	0.672
ru	0.542	0.585	0.643
sw	0.484	0.509	0.562
th	0.534	0.513	0.587
tr	0.524	0.546	0.618
ur	0.469	0.452	0.552
vi	0.585	0.570	0.633
zh	0.488	0.569	0.619

4.1.2 Question Answering Based Document Level Claim Detection

Table 11: Topic-wise F1 and overall accuracy (both in %) for topic

Model	Or.	Trans.	Prot.	Cure	All
NLI	56.9	45.1	54.5	3.3	46.6
QA	85.9	64.7	63.9	66.5	72.3

We measure the performance for both zero-shot topic classification and the subsequent filtering that yields claims. As shown in Table 11, our zero-shot QA approach considerably outperforms zero-shot NLI, demonstrating that QA can be better at measuring topic relevance.

4.1.3 Text Joint Entity Relation Event Extraction

Our joint model for relation and event extraction consistently outperforms DYGIE++[WA19] and BERT BASELINE on ACE05-R and ACE05-E as shown in Table 12.

Table 12: IE Results on ACE2005 (F-score, %)

Dataset	Task	DYGIE++	BASELINE	ONEIE
ACE05-R	Entity	88.6	-	88.8
	Relation	63.4	-	67.5
ACE05-E	Entity	89.7	90.2	90.2
	Trig-I	-	76.6	78.2
	Trig-C	69.7	73.5	74.7
	Arg-I	53.0	56.4	59.2
	Arg-C	48.8	53.9	56.8

4.1.4 Text Document-level Event Argument Extraction

Our event argument extraction model effectively captures all arguments, while the baseline model struggles with rare event types and the QA model is hindered by over-generation. Results are in Table 13.

Table 13: Argument Extraction Results (%) on WIKIEVENTS

Model	Arg Identification		Arg Classification	
	Head F1	Coref F1	Head F1	Coref F1
BERT-CRF	69.83	72.24	54.48	56.72
BERT-QA	61.05	64.59	56.16	59.36
BERT-QA-Doc	39.15	51.25	34.77	45.96
BART-Gen	71.75	72.29	64.57	65.11

4.1.5 Text Entity and Event Coreference Resolution

As shown in Table 14, the magnitude of the differences in scores between our the Context Dependent Gated Module (CDGM) with noise and the state-of-the-art methods indicates the effectiveness of our methods.

Table 14: End-to-end results on ACE 2005 for Coreference

ACE (Cross-Validation)	CoNLL	AVG
SSED + Supervised _{Extended} (2016)	55.23	52.53
SSED + MSEP (2016)	53.80	51.38
ACE (Test Data)	CoNLL	AVG
Baseline	58.93	55.78
Simple (All Features)	57.55	54.79
CDGM (All Features)	58.99	56.32
Noise (All Features)	60.43	57.85
CDGM + Noise (All Features)	62.07	59.76

4.2 Extraction from Images and Video

4.2.1 Visual Content

Multimodal event extraction and argument role labeling: The Joint Multimodal Transformer(JMMT) significantly improves the event extraction performance over baseline methods, as shown in Table 15. Compared to text-only OneIE or video-only JSL [PR20] baselines, the JMMT produces at most 74% relative gain in event extraction, demonstrating the importance of leveraging multimodal information for understanding complex events. Compared to previous methods on image-text multimedia event extraction, the superior performance of JMTT verifies 1) the effectiveness of the powerful transformer model for multimodal information fusion and 2) the importance of modeling dynamics in videos.

Table 15: Video event extraction and argument role labeling.

Input	Model	Text Evaluation						Video Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
Text	OneIE	38.5	52.1	44.3	16.6	21.8	18.8	-	-	-	-	-	-	38.5	52.1	44.3	16.6	21.8	18.8
Video	JSL	-	-	-	-	-	-	24.1	17.1	20.0	2.2	2.8	2.4	24.1	17.1	20.0	2.2	2.8	2.4
	JMMT _{Video}	-	-	-	-	-	-	26.6	29.2	27.8	8.9	10.1	9.5	26.6	29.2	27.8	8.9	10.1	9.5
Multimedia	WASE	33.6	53.8	41.4	15.2	22.1	18.0	20.4	14.0	16.6	2.8	1.3	1.7	34.0	54.0	41.8	15.3	22.1	18.1
	JMMT	39.7	56.3	46.6	17.9	24.3	20.6	32.4	37.5	34.8	9.2	10.6	9.9	41.2	56.3	47.6	18.8	24.7	21.3

We also visualize the results of event extraction. As shown in Figure 27, the center frame is very hard to extract events from due to the occlusion of arguments. With only this image as input, WASE fails to extract events and arguments. However, our JMMT successfully recognizes the

event and detects “Jailer” in the image with the help of video-level dynamics and the context of the previous and next frames. This example illustrates the importance and difficulty of multimedia event extraction from videos and articles. We also observe that our JMMT fails to recognize the “Detainee” because of occlusion. This indicates the possibility of leveraging entity tracking to further improve VM2E2, which we leave as future work.

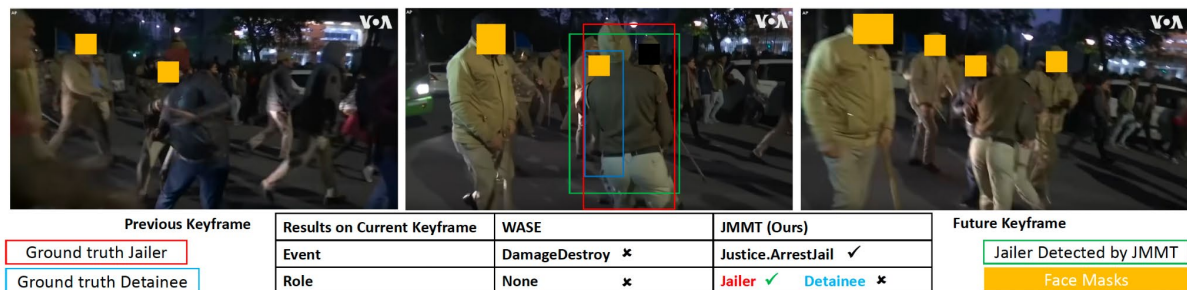


Figure 27: Visualization of event extraction results on one video segment. We mask faces (orange boxes) for privacy.

Learning visual event from multimodal self-supervision. We first examine the results of the multimodal event coreference resolution task in Table 16. All the methods we compared shared the same text feature extractor. For visual feature extraction, HowTo100M [MI20b], NCE, and our method apply ResNet-152 (R152) and ResNeXt-152 (RX101), followed by [CH21c]. MIL-NCE [MI20a] uses a more advanced video feature extraction backbone, S3D-G [SA18]. Our model significantly outperforms all previous methods using the same architecture and those with a trainable (TR) and more powerful visual backbone [MI20a].

Table 16: Multimodal event coreference resolution.

Method	Visual Model	TR	P	R	F_1	Acc
HowTo100M	R152+RX101	N	32.2	62.8	44.3	55.2
NCE	R152+RX101	N	35.5	68.3	45.5	47.5
Ours	R152+RX101	N	38.4	76.4	51.5	59.6
MIL-NCE	S3D-G	Y	37.8	75.0	50.6	59.2

We visualize results from our event coreference resolution model in Figure 28. We observe that the model correctly selects the most appropriate sentence for a given video segment. Also, we find that the model learns to associate object regions with the words in the sentence. For example, the first sentence had a high similarity score with the object ‘Crane’ since it mentioned ‘*building some new island.*’

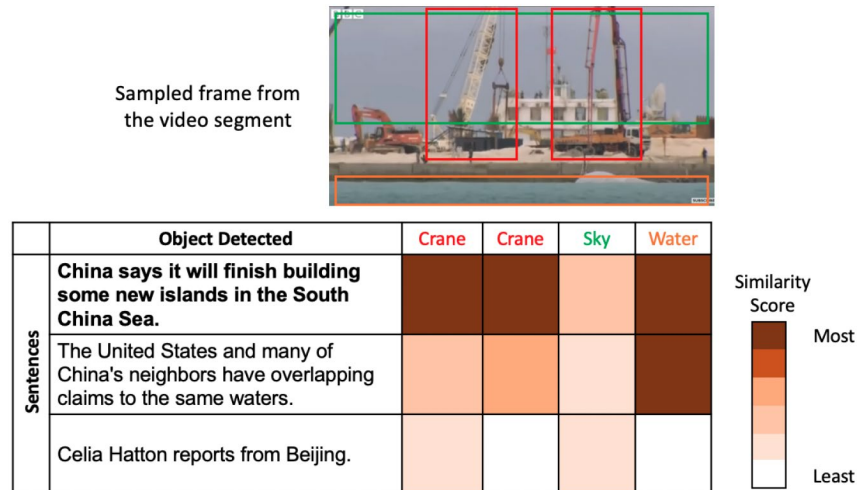


Figure 28: Event coreference resolution visualization. The bold sentence is correctly selected as coreferential within the article by the model.

Self-supervised event extraction and argument role labeling. In addition to utilizing supervised annotation as the JMMT model shown above, we explore the multimodal common space by video-to-ASR text self-supervision to train our Multimodal Clustering Network [CH21c]. As shown in Table 17, Result demonstrates that we achieved comparable results with the SOTA transformer model in event extraction tasks. We assume the performance drop in argument role labeling is due to the lack of bounding box supervision, which is not provided in the self-supervision model.

Table 17: Video event extraction and argument role labeling with self-supervision.

	Supervision	Video Evaluation					
		Event Mention			Argument Role		
		Precision	Recall	F1	Precision	Recall	F1
JSL	Supervised	24.1	17.1	20.0	2.2	2.8	2.4
JMMT	Supervised	26.6	29.2	27.8	8.9	10.1	9.5
Ours	Self-supervised	28.4	19.2	19.5	2.1	2.4	2.2

Spatio-temporal grounding: In the spatio-temporal grounding, we combined the spatial and temporal grounding from previous work to form the evaluation. The entire video and the respective action instructions were provided. The model needs to localize each action step in temporal (start-time/end-time) and spatial (location in the video) as described in Figure 13. We evaluate in two metrics: IoU + Pointing game combines the evaluation setting from the spatial grounding and temporal grounding metrics. The frame in the video is counted as correct when the predicted class is correct. Also, given the predicted class as a query, the maximum point of the heatmap lies within the desired bounding box. We then compute the IoU over all the predictions with the GT to acquire the final score. We also follow previous spatio-temporal evaluations, which compute video mAP, where we set IoU threshold between GT and predicted spatio-temporal tubes. A prediction is counted as correct when it surpasses the IoU threshold. We then compute the mean over the Average precision over all classes. We form a 3D prediction mask following Figure 15 and compute IoU between our 3D heatmap and 3D tube. As shown in Table 18, our method outperforms the other baselines by a significant margin. The result demonstrates our model’s ability to incorporate both global (temporal) and local (spatial) representations to perform untrimmed video spatio-temporal action localization. Models designed for trimmed videos (CoMMA [TA21]) or trained with aligned image-text (CLIP [RA21]) without specific loss designs for spatial grounding (GLIP [LI22]) were not able to ground the action in the correct region.

Table 18: Spatio-temporal grounding with self-supervision.

Method	Backbone	Data	Super.	Mod.	YouCook-Inter	Grounding Youtube	
					Acc	Acc	mAP
MIL-NCE [37]	S3D-word2vec	HT100M	Self	VT	23.67	27.45	8.21
CoMMA* [50]	S3D-word2vec	HT200K	Self	VT	48.63	47.68	23.38
Ours	S3D-word2vec	HT200K	Self	VT	53.98	60.62	44.93
CLIP [41]	CLIP	HT200K	Weak	IT	14.10	12.50	3.49
CoMMA† [50]	CLIP	HT200K	Self	VT	52.65	47.56	36.42
GLIP [31]	Swin-L	Cap24M	Weak	IT	52.84	53.62	24.73
Ours	CLIP	HT200K	Self	VT	57.10	55.49	43.12

We visualize our spatio-temporal result on the GroundingYoutube dataset, as shown in Figure 29. For the GLIP model, we output the bounding box with the highest confidence score and visualize its center point. We found GLIP model focuses on the salient object while our model focuses more on human-object interaction.

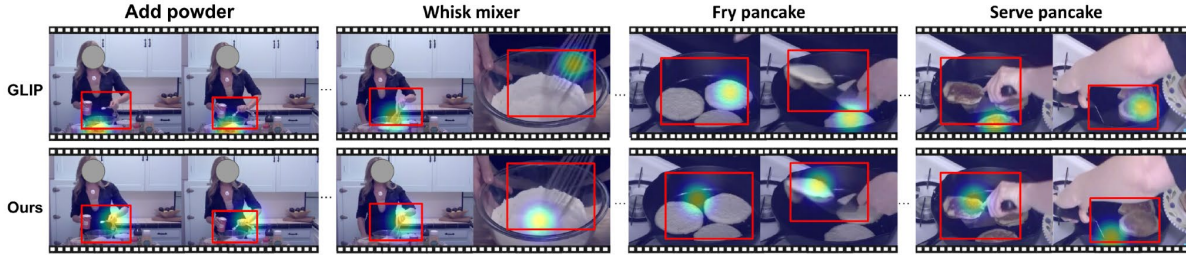


Figure 29: Visualization on GroundingYoutube dataset. The red box is the annotation and heatmap is the prediction from the model.

4.2.2 Multimodal Event Extraction

As shown in Table 19, our multimodal event extraction model achieves better performance on text-only and image-only events, which demonstrates the effectiveness of multimedia training framework in knowledge transfer between modalities. We contrast our model, WASE[Li20] with JMEE[LI18], and GAIL[ZH19].

Table 19: Event and argument extraction results (%). We compare three categories of baselines in three evaluation settings.

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	32.4	37.3	22.7	12.8	16.4
	WASE ^T	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	33.1	36.7	20.1	13.0	15.7
Image	WASE ^I _{att}	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	23.0	25.4	2.9	6.1	3.8
	WASE ^I _{obj}	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	22.4	24.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat _{att}	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat _{obj}	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE _{att}	37.6	66.8	48.1	27.5	33.2	30.1	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	19.9
	WASE _{obj}	42.8	61.9	50.6	23.5	30.3	26.4	43.1	59.2	49.9	14.5	10.1	11.9	43.0	62.1	50.8	19.5	18.9	19.2

4.2.3 Zero-Shot Image Event Extraction

The gains achieved by pretraining on news data are significantly amplified with the help of structural event knowledge, showing that the proposed optimal transport alignment effectively captures the argument structures, which previous vision-language pretraining models fail.

Table 20: Evaluation results and ablation studies on image event extraction.

Setting	Model	Multimedia Event Extraction (M ² E ²)						Grounded Situation Recognition (SWiG)				
		Event			Argument			Event	Argument			
		P	R	F ₁	P	R	F ₁		verb	value	value-all	ground
Zero-shot	CLIP	29.5	65.7	40.7	9.2	12.7	10.7	28.3	19.3	12.8	13.2	4.1
	CLIP pretrained on news	31.7	64.7	42.6	9.7	13.1	11.1	29.9	20.1	13.2	13.9	5.3
	CLIP-Event	36.4	70.8	48.1	13.9	16.0	14.8	31.4	23.1	14.1	14.8	6.3
	w/o OptimalTransport	35.0	59.3	44.1	11.0	12.6	11.9	30.2	21.6	13.6	14.1	5.4
	Single Template	32.3	71.4	44.4	11.9	15.6	13.2	30.4	22.0	13.5	14.0	5.7
	Composed Template	33.9	72.8	46.3	12.7	15.3	13.9	30.9	22.5	13.9	14.2	5.8
	Continuous Prompt	33.6	75.7	46.5	11.1	16.7	13.3	30.4	21.9	13.0	14.1	5.9
	Caption Editing	30.9	71.4	43.2	11.6	13.8	12.6	30.1	20.5	13.0	13.8	5.3
	GPT-3 Prompt	34.2	76.5	47.3	12.1	16.8	14.1	31.1	22.1	13.2	14.2	6.0
Supervised	State-of-the-Art [17, 25]	43.1	59.2	49.9	14.5	10.1	11.9	39.9	31.4	18.9	24.9	9.7
	CLIP finetuned on SWiG	38.1	71.6	49.8	20.9	12.8	15.9	42.6	32.6	19.2	25.2	10.2
	CLIP-Event⁺SWiG	41.3	72.8	52.7	21.1	13.1	17.1	45.6	33.1	20.1	26.1	10.6
	w/o OptimalTransport	40.3	71.3	51.5	20.8	13.0	16.0	44.7	32.9	19.4	24.4	10.1

4.2.4 Video Event Extraction based on Argument State Tracking

As shown in Table 21, on both verb classification and semantic role prediction tasks in VidSitu, our model achieved a new state of the art, outperforming state-of-the-art 3D convolutional neural networks.

Table 21: Results on semantic role prediction. Results are the average (Avg) over 10 runs with standard deviation (Std) reported.

Model	CIDEr		CIDEr-Verb		CIDEr-Arg		ROUGE-L	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std
GPT2 [†]	34.67		42.97		34.45		40.08	
I3D [†]	47.06		51.67		42.76		42.41	
SlowFast [†]	45.52		55.47		42.82		42.66	
SlowFast	44.49 ±2.30		51.73 ±2.70		40.93 ±2.42		40.83 ±1.27	
Ours (OSE-pixel + OME)	47.82 ±2.12		54.51 ±3.00		44.32 ±2.45		40.91 ±1.32	
Ours (OSE-pixel/disp + OME)	48.46 ±1.84		56.04 ±2.12		44.60 ±2.33		41.89 ±1.12	
Ours (OSE-pixel/disp + OME + OIE)	47.16 ±1.71		53.96 ±1.32		42.78 ±2.74		40.86 ±2.54	

4.2.5 Few-Shot Video Event Prediction

To our surprise, our 10-shot model outperforms state-of-the-art fully-supervised baseline, i.e., MERLOT (Zellers et al, 2021), by a large margin ($\sim 4\%$) for video event extraction. This shows that our model has strong few-shot ability not only on video-language understanding but also on prediction.

Table 22: Video event prediction (%).

Method	#video _{FT}	Acc
VLEP [23]	20142	67.5
MERLOT [68]	20142	68.4
VidIL(ours)	10-shot	72.0
Human	-	90.5

4.2.6 Integrating Pre-Trained Vision Models

As previously described, following the program’s final evaluation, we explored the use of newly-available pre-trained visual models as a means of extracting knowledge from images. We specifically applied Unified-IO (Lu, et al, 2022) on captioning and claim generation tasks. Using ten GPUs concurrently, 8000 images were captioned in under nine hours.

To establish that our instance of the pre-trained model was operating as expected, we measured our model’s performance on the VizWiz-Captions Challenge 2021 validation dataset (VizWiz, 2022). Table 23 includes scores for the multiple metrics reported as a part of the VizWiz Captions challenge. These results are comparable to those in 9th through 12th places on the VizWiz-Captions Challenge 2021 leaderboard as of the time of writing, without any specific targeting of the VizWiz task.

Table 23: The GAIA Unified-IO-Inference VizWiz-Captions scores on test run.

	B1	B2	B3	B4	ROUGE-L	METEOR	CIDEr	SPICE
GAIA Unified-IO-Inference docker	55.59	37.76	25.32	16.71	38.71	17.56	48.25	13.13

We explored a variety of prompts to improve performance. One alternative prompt outperformed the baseline in ISI testing: *“What does the image describe including every detail?”* This prompt, which explicitly asks for “every detail” but is otherwise the same as the baseline prompt, produced captions that better matched the VizWiz gold-standard captions. In ISI testing, the baseline captioning produced a 16.3 METEOR score compared to the alternative prompt score of 17.2 on the same 500-image sample.



Figure 30: A caption example using LDC image L0C04AEUW.jpg. For reporting purposes to avoid propagated personally identifiable information we have masked the face.

An example image is shown above in Figure 30. The baseline and alternate captions are as follows:

- *what does the image describe?*
 - a woman wearing a burgundy dress in front of a room of beds.
- *what does the image describe including every detail?*
 - a woman in a mask and a purple shirt stands in a hospital room.

The claim formation prompts and responses are as follows:

- *what does the image claim?*
 - the lady in the maroon shirt is working in the hospital.
- *where does the image take place?*
 - in a hospital room.

For the image captioning experiments, the approaches were evaluated using 200 images belonging to 10 different classes sampled from ObjectNet. In the alternative prompts experiment, the alternative prompts performed on par with or worse than the default classification prompt. Table 24 presents the results of classification accuracy using different prompt variants.

Table 24: Performance of alternative prompts used for image classification. Accuracies are calculated on a set of 200 images belonging to 10 different classes sampled from ObjectNet.

Prompt	Accuracy (%)
what is this in the image? (<i>default prompt</i>)	16.00
classify this image.	15.50
put this image in a class.	16.00
what type is this?	11.50
image, classify.	15.50
categorize this image.	15.50

Class descriptions were manually curated for these 10 categories. Using class descriptions instead of class labels in combination with the default prompt resulted in a decrease in accuracy from 16% to 12%.

For GAIA’s goals, claims (not captions) are the desired output. Thus, we further explored integrating this model into claim generation. Claims produced by the new component were successfully exported in JSON format. The detailed claims were searchable and viewable within the GAIA Claim Radar visualization as described in Sections 3.6.1 and 4.6.1.

4.2.7 Audio and Text within Video and Images

In practice, we found four challenges that limited the benefit of the incorporating ASR and OCR into the overall multimodal pipeline and at detection time. (1) Errors in the underlying ASR and OCR which while high performing are still not perfect; (2) a mismatch between the training data for the text-based models that were applied to extract content from ASR and OCR and produce knowledge elements; (3) the overlap between content in the ASR/OCR and content elsewhere in the corpus, which limited the gains in recall and made the precision errors more problematic; and (4) the evaluation setup itself which targeted visual information, not the content in the text/audio of the video and images. From error analysis, we would expect the degree to which ASR and OCR contribute to knowledge extraction from a data stream to be impacted by the nature of the corpus.

4.3 Aggregating a Knowledge Base

4.3.1 Linking to Wikipedia

Our approach to linking to Wikidata (EPGEL Table 25 below) outperforms a variety of SOTA techniques across all datasets, showing that having an effective candidate retrieval can still be helpful even during this era of large language models.

Table 25: Overall entity linking results of micro F1 scores (%).

Methods	RSS-500 (test)	ISTEX-1000 (test)	TweekiGold (test)	WikipediaEL (dev)
EPGEL	76.4	92.7	69.3	92.3
<i>Effects of Candidate Retrieval Strategy</i>				
◆ Simple Query	66.4	87.6	66.0	81.9
◆ Wikipedia Dictionary	71.2	91.6	68.8	89.8
◆ Profiling-Based Query [Title + Desc]	68.4	92.6	69.1	88.4
<i>Previous Methods</i>				
GENRE * (Cao et al., 2021)	68.2	88.4	62.4	86.3
BLINK * (Wu et al., 2020)	73.5	88.5	65.9	90.5

4.3.2 Constructing a Knowledge Base of Extracted Knowledge Elements

Our approach to constructing a knowledge base of knowledge elements from multimodal data relied on unifying the entries for knowledge elements linked to a particular Wikidata entry and the features of extracted name strings. This approach to linking limits, linking to named concepts and does not capture event-event coreference or the linking of concepts (e.g. *Sam's mask*). Thus, we consider approach a high-precision approach to cluster formation. We explored and manually reviewed approaches that expanded the class of linkable concepts to both events and non-named entities. However, we found that the granularity of the linking decision often was governed by the particulars of the claim of interest, thus for example in some contexts one would want to distinguish between *a blue and red mask, both owned by Sam*; where in other contexts, one wants a single cluster for *all of Sam's masks (e.g. Sam washes his masks weekly)* or *even all masks in general (Masks help prevent the spread of COVID-19)*. We found that allowing these linking decisions to be made on-demand by the claim generation system yielded better performance in the program evaluations. These observations suggest two possible paths for a GAIA-like system in the future: The first, as we have chosen delays hard coreference decisions and optimizes them for a specific query. The second would involve exploring a richer semantics of coreference involving, e.g. part-whole relationships and partial-overlap. From our own investigations, a challenge of the latter approach is uncertainty of many of these decisions, which when applied at scale requires far more edges within the knowledge graph.

4.3.3 Knowledge Graph Tool Kit

4.3.3.1 Browser evaluation

We evaluated the possibility of deploying different knowledge graph formats with our browser and ran a user study to compare the ability of users to consume knowledge with different representations [IG--]. Our results with 5 query types and 45 participants show that knowledge exploration with any of the models is relatively quick with our browser. Among the different representations, we noted better performance for the browser with the Wikidata model and the standard reification RDF approach, and worse performance for the n-ary relationships model (Figure 31). We also asked participants to write a programmatic query based on first exploring the graph with the browser. In this experiment, we noted worse results for the RDF reification representation. However, over all these experiments, we noticed that participants were able to solve any of their tasks relatively quickly (typically within a few minutes). In terms of accuracy, most participants were able to answer the queries correctly, however, here we also noticed that the qualifier (Wikidata) model yielded best results. As it can be expected, we noted higher times and lower accuracies on the more complex queries (e.g., queries with multiple clauses). These evaluations demonstrate that the browser is adequate to enable knowledge exploration, especially when coupled with the Wikidata representation model.

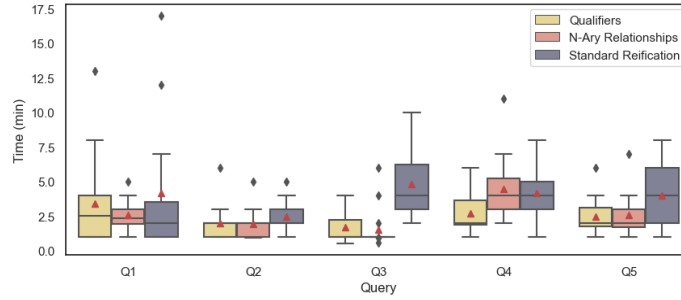


Figure 31: Evaluation of the time needed for participants to retrieve the correct answer with our browser.

Hybrid query evaluation We experimented with several different embedding models based on the popular BERT, ComplEx, and TransE methods [CH22]. The size of the embeddings ranges from 26GB for 100-dimensional TransE and ComplEx to 182GB for 1024-dimensional BERT. The import time for TransE and ComplEx is 25 mins and 1.4 hours, whereas the import time for BERT is 8 hours. The indexing time ranges between 1.1 and 11.5 hours. We investigated the effects of different queries, which we summarize here, described in detail in our publication [CH22].

First, we used Kypher-V. It can perform typical vector operations, such as similarity between two items and retrieval of top-K embeddings with structural restrictions (philosophers). Example of the latter for Socrates is shown in Figure 32. In fact, the implementation of this query in Kypher is optimized to scale, resulting in a 250x speedup in comparison to the brute-force version [CH22]. A more advanced computation based on three inputs is shown in Figure 33.

```
!$kypher -i short_abstracts_textemb -i labels -i claims
--match 'emb: (x:Q913)-[]->(xv), (y)-[]->(yv),
        claims: (y)-[:P106]->(:Q4964182),
        labels: (x)-[]->(xl), (y)-[]->(yl)'
--return 'xl as xlabel, yl as ylabel, kvec_cos_sim(xv, yv) as sim'
--order 'sim desc' --limit 5
```

xlabel	ylabel	sim
'Socrates'@en	'Socrates'@en	1.00000
'Socrates'@en	'Heraclitus'@en	0.93445
'Socrates'@en	'Hippocrates'@en	0.93040
'Socrates'@en	'Diogenes Laërtius'@en	0.91655
'Socrates'@en	'Ammonius Hermiae'@en	0.90550

Figure 32: Brute-force top-k similarity query for philosophers similar to Socrates (Q913) with BERT text embeddings.

```

!$kypher -i short_abstracts_textemb -i labels -i claims
--match 'emb:      (x)-[]->(xv),
                (xv)-[r:kvec_topk_cos_sim {k: 100, nprobe: 4}]->(y),
                claims: (y)-[:P31]->(:Q5),
                labels: (x)-[]->(x1), (y)-[]->(y1)'
--where 'x in ["Q859", "Q868", "Q913"] and x != y and r.similarity >= 0.9'
--return 'x1 as xlabel, max(r.similarity) as sim, y1 as ylabel'

xlabel      sim      ylabel
'Aristotle'@en 0.97378 'Bryson of Achaea'@en
'Plato'@en    0.92298 'Franciszek Kasparek'@en
'Socrates'@en 0.93465 'Anytus'@en

```

Figure 33: Generalized vector join query with multiple inputs.

Kypher-V, thus, extends the capabilities to store, index and query large sets of embedding vectors in a fully integrated way, allowing users to efficiently combine large amounts of structured and unstructured data while still preserving the ability to use the system on modest resources such as a laptop. The reported execution times are mostly anecdotal and a more systematic evaluation of system performance has to be reserved for future research.

We do not make any claims about the quality and usefulness of the different types of embeddings used in the examples. They are all flawed in different ways, and in fact combining multiple embedding types as supported by Kypher-V has the potential to improve overall system performance. Kypher-V is completely agnostic to the types of embeddings used, and any reasonable embeddings that can be linked to the nodes and edges of a KG can be queried by it. The contribution of the evaluation is to show that KGTK, Kypher and Kypher-V are effective tools for complex analytic use cases. We demonstrate here and in [CH22] that Kypher-V supports a variety of hybrid queries that are difficult to implement with existing tooling. Kypher-V allows researchers and developers to investigate use cases on their own laptop, exploring extensions of Wikidata with large amounts of vector data, using minimal setup, modest resources, and a simple query language. Kypher-V is part of the latest KGTK version and can be installed as customary via pip to be used in various use cases.

4.3.3.2 Wikidata completion evaluation

We finally evaluated the accuracy of our two completion methods.

The fact validation method based on DBpedia is evaluated on 200 samples based on 50 randomly selected fact-evidence pairs: for each fact-evidence pair, we created two samples using two different fact-validation questions like those in Figure 34 . The zero-shot results in Table 26 are encouraging, showing that the best model UnifiedQA-large achieves 0.905 F1-score off-the-shelf, without any tuning. The other UnifiedQA variants perform similarly, with the base model obtaining best recall and the largest (11b) model obtaining maximum precision.

DPR Spouse data queries	1. Is {person_one} married to {person_two}? 2. Is {person_two} married to {person_one}?
-------------------------	--

Figure 34: DPR query to retrieve spouse fact support documents.

Table 26: Zero-shot evaluation scores of pre-trained models over the fact validation tasks.

Models	Precision	Recall	F1
Fact-checking	0.639	0.919	0.754
UnifiedQA			
• base	0.803	0.963	0.876
• large	0.941	0.872	0.905
• 3b	0.957	0.81	0.879
• 11b	0.963	0.705	0.814

Encouraged by these results, we also attempted to fine-tune our T5-based UnifiedQA (base) with different types of training dataset size 2000. Table 3 shows the results of the experiments. The results show that fine-tuning the model significantly increased its overall performance from 0.876 to 0.945 (7 absolute points). We also noted that the choice of the dataset has a significant impact on the obtained scores, with the direct dataset performing best overall. We also experimented with different batch sizes ranging from 2 to 32 and observed optimal F1-score of 0.95 for batch size 16. This means that, in most of the cases, the model performs well in the initial part of the training. As the model starts to overfit the training data, the evaluation performance drastically decreases. We performed further engineering of our data, by allowing positive labels only when the DPR score is over 0.83, which ultimately resulted in best F1-score of 0.96.

Table 3: Fine-tuning of UnifiedQA base trained with batch size of 8.

Train Dataset	epoch	Precision	Recall	F1	Accuracy
w/o tuning	-1	0.803	0.963	0.876	0.85
direct	0	0.938	0.964	0.951	0.945
all	6	0.913*	0.954	0.933*	0.925*
indirect_pos	1	0.922	0.973	0.947	0.94
indirect_neg	0	0.944	0.927*	0.935	0.93

Our evaluation of the extraction of temporal information from Wikipedia also showed encouraging, yet lower, results. We created seven different experimental settings that differ in terms of their model and training set, reaching an optimal F1-score of 0.92 and accuracy of 0.85 (Table 4). We also looked closer at the different models, investigating how many models answer a question wrongly (Figure 35). We observed that in most cases all models get the question right, and in 92% of the cases a minority of the models are wrong. This highlights the promise of our approach and indicates that ensemble models are a promising path forward.

Table 3: Fine-tuning experiments for temporal information extraction.

Fine-tuning	Base Model	Training Dataset	Precision	Recall	F1	Accuracy
-	Bart	-	0.84	0.97	0.90	0.75
-	T5 large v2	-	0.81	1.0	0.90	0.68
v1	Bart	Training/Dev Set 1	0.843	1.0	0.91	0.81
v2	Bart	Training/Dev Set 2	0.833	1.0	0.90	0.75
v3	Bart	Training/Dev Set 3	0.91	0.91	0.91	0.85
v4	Bart	Training/Dev Set 4	0.91	0.92	0.92	0.83
v5	Bart	Training/Dev Set 5	0.91	0.92	0.92	0.83
v6	T5 large v2	Training/Dev Set 5	0.947	0.9	0.92	0.83

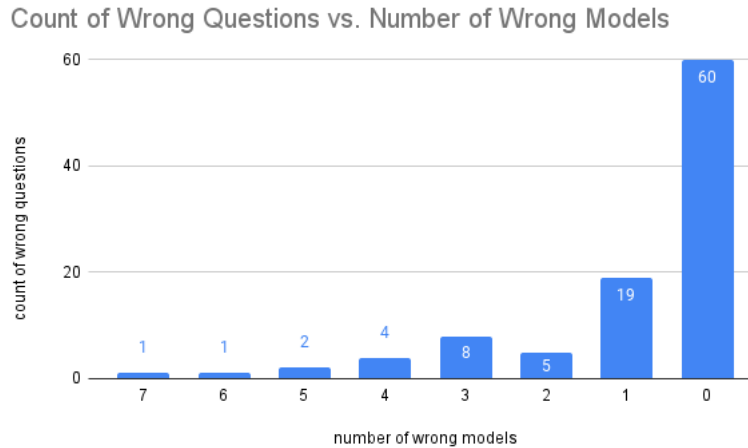


Figure 35: Number of wrong models per question.

Our error analysis showed several common challenges: 1) names in the question and in the Wikipedia article do not match; 2) the second entity is not in the paragraph or is truncated by DPR’s model; 3) mismatch in the precision of the dates between Wikidata and Wikipedia; 4) no mention of the event that corresponds to the relation (e.g., no mention of marriage for the spouse relation). These errors can likely be circumvented by further engineering efforts, which will ultimately be justified by the large amount of knowledge that can be distilled from Wikipedia and other unstructured sources into Wikidata.

4.4 Discovering Claims

Figure 36 shows a comparison of various approaches to finding claims as measured internally on LDC annotated claim frames. While the two NLP models outperform the structural model for supporting and refuting claims. The structural model outperforms the two NLP models for Related claims.

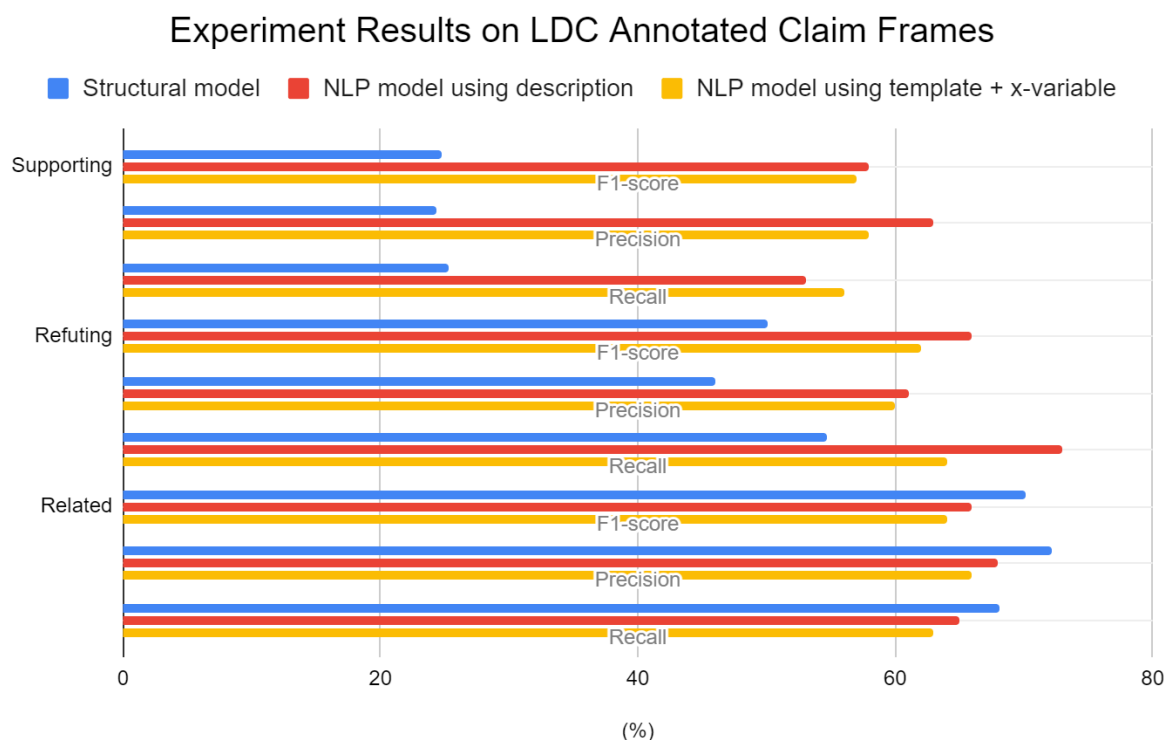


Figure 36: Results Comparing HypoGator Claim Approach on LDC Annotated Data

We further evaluate HypoGator on the program’s evaluation datasets. We compare HypoGator with another competing system and report one evaluation matrix. We define metrics based on the following features of the program’s complex evaluation.

F1 is computed using *NumUniqueSubmittedValues* as the numerator for Precision and Recall. *NumUniqueSubmittedValues* is defined as follows:

1. For *ClaimRelation* in {'ontopic'}: *NumUniqueSubmittedValues* is the number of unique combinations of correct values of required F1 fields for the correct claims in the set of claims induced by taking the top Cutoff number of assessed claims in the submitted ranking for the Condition and Query.
2. For *ClaimRelation* in {'nonrefuting', 'nonsupporting', 'refuting', 'related', 'supporting'}: *NumUniqueSubmittedValues* is the number of unique combinations of correct values of required F1 fields for the correct claims in the set of claims induced by taking the top Cutoff number of assessed claims in the submitted ranking for the Condition and Query, such that their cross-claim-relation to the Query claim (as labeled in the submitted rankings list for the Query claim) is compatible with the *ClaimRelation*.

Six variants of F1-based scores are generated by the scorer. These are broken down into two types (1 and 2) and three subtypes (A, B, and C).

- **Type 1:** When scoring, consider all assessed claims in the ranking that was submitted for the QueryID and ClaimRelation.
- **Type 2:** When scoring, consider the top C assessed claims in the ranking that was submitted for the QueryID and ClaimRelation, where C is the pooling depth for the QueryID and ClaimRelation.
- **Subtype A:** required F1 fields comprise: claimTemplate, claimEpistemic, xVariable, and claimer
- **Subtype B:** required F1 fields comprise: claimTemplate, claimEpistemic, and xVariable
- **Subtype C:** required F1 fields comprise: claimTemplate and xVariable

Based on these types and subtypes, 6 variants of F1 are reported by the scorer:

- Variant 1A: Type 1 - Subtype A
- Variant 1B: Type 1 - Subtype B
- Variant 1C: Type 1 - Subtype C
- Variant 2A: Type 2 - Subtype A
- Variant 2B: Type 2 - Subtype B
- Variant 2C: Type 2 - Subtype C

Table 27 and Table 28 present results on evaluation for various metrics on Conditions 5 and 6 comparing the GAIA system (HypoGator) with those from a different high-scoring team. While we participate in Condition 7, no results were provided.

Table 27: F1 of Claim Radar in Condition 5 on Various Metrics

Metric	Claim Relation	Team	Score
F1Score-V1A	Ontopic	GAIA TA3 (#1)	0.1015
	Ontopic	OPERA TA3 (#2)	0.0578
F1Score-V1A	Supporting	GAIA TA3 (#1)	0.0918
	Supporting	OPERA TA3 (#2)	0.0451
F1Score-V1A	Refuting	GAIA TA3 (#3)	0.0837
	Refuting	OPERA TA3 (#1)	0.0939

F1Score-V2A	Ontopic	GAIA TA3 (#2)	0.1582
	Ontopic	OPERA TA3 (#1)	0.2426
F1Score-V2A	Supporting	GAIA TA3 (#1)	0.0970
	Supporting	OPERA TA3 (#2)	0.0528
F1Score-V2A	Refuting	GAIA TA3 (#2)	0.0901
	Refuting	OPERA TA3 (#1)	0.0979

Table 28: F1 of the Ontopic Measure for Evaluation Condition 6

Metric	Claim Relation	Team	Score
F1Score-V1A	Ontopic	GAIA TA3 (#2)	0.0476
	Ontopic	OPERA TA3 (#1)	0.0521
F1Score-V2A	Ontopic	GAIA TA3 (#2)	0.154
	Ontopic	OPERA TA3 (#1)	0.2164

From the results in Table 27 and Table 28, we found that HypoGator generally performs better on Condition 5 than on Condition 6. The difference in HypoGator is that the NLP module is only used in condition 5 which, again, shows that the description information does improve the quality of relation detection.

By comparing the result on tasks V1 and V2 in condition 5 (Table 27), we find that HypoGator is relatively better on V1 tasks. One possible reason is that, as defined, the V2 task considers the top-C assessed claims in the ranking. Thus, the result may not include more supporting & refuting claims that the two-step answer selection module in HypoGator adds to the end of the answer list.

Besides, HypoGator achieves a better ranking on supporting/non-refuting relation detection. This result, again, indicates the effectiveness of the NLP cross-claim relation detection module, since it improves the most on supporting relation detection compared with the structural module.

4.5 Integration into Program Testbed

The GAIA pipeline successfully completed Phase 3 evaluation, including TA1, TA2, and TA3. There were many iterations of the pipeline through development. The Confluence wiki maintains a history:

<https://nextcentury.atlassian.net/wiki/spaces/AIDAC/pages/2659942401/Phase+3+-+Evaluation>

The full results on the TA1 pipeline were collected by CACI upon completion within their Kubernetes system. All steps in the processing, including components for natural language processing and computer vision, ran in serial and stored incremental results along with the final NIST AIF output.

- *Job YAML:* `ta1-all-caci-dev-20220614.yaml`
- *Run Time:* ~51 hours (g4dn.2xlarge)
- *Run Output:* `s3://aida-phase2-ta-performers/TA1a-GAIA/phase3/ta1-all-caci-dev-20220614.tar.gz`

GAIA TA2 was implemented as a single Docker image. The GAIA TA2 docker image was used to run containerized processing at both CACI and at ISI.

- *Job Yaml:* `gaia-ta2-job.yaml`
- *Runtime:* 4h8m (g4dn.4xlarge)

GAIA TA3 was also implemented as a single Docker image. Each revision to the Docker image was made available as a uniquely tagged Docker image. Each update required testing and validation.

- *Job Yaml:* `gaia-ta3-job.yaml`
- *Runtime:* 6h13m (g4dn.4xlarge)

In addition to running GAIA pipeline on the CACI environment, we also tested the components at ISI. This demonstrated portability of the Docker-based pipeline and also allowed additional run outputs on the ISI high-performance computer cluster.

TA1, TA2, and TA3 were separated throughout the evaluation. This allowed for creating full pipeline runs that used components from different teams. We were able to reproduce results using OPERA and BBN TA1 and TA3 components.

Following the Phase 3 evaluation, additional runs with improvements and corrections were made. These were submitted to CACI s3. The GAIA pipeline results were incorporated in the Claim Radar software for analysis.

4.6 Visualizing and Interacting with Extracted Knowledge

4.6.1 Claim Radar

As described in section 3.6.1, to provide general AIDA compatible, user-facing access to system extractions about claims we expand on Covid Claim Radar. As results, we present screenshot of some of the core extensions to this capability.

Covid-19 Claim Radar

Welcome, guest! Log in to edit.

☒ View all corrections
 ☐ View your corrections
 ☐ Hide corrections

Username: Password: Login

-Select Topic- -Select Claimer- -Select Claimer Affilia... -Select Claim Object- -Select Language- -Select Location- Search Reset

-Select KE text- -Select KE type- ☐ Bind types to KE text

Topic	Claimer	Claimer Affiliation	Stance	Language	Location	Claim Object	Sentence	Associated KES	Associated KE Types
Contracting the virus	Jack Dorsey		Affirm	EN		Toddlers and babies	Toddlers and babies are showing up severely ill already, in Italy, in China, in Iran.	Kids, China, Some, some, coronavirus, who, tweet, whom, CDC, die, Toddlers, Ryan, Iran, Italy, them, children, Elon Musk, babies, followers, asshole, his	GPE.NonSovereignCountry, PER.Politician, Physical.LocatedNear, PER.Politician.HeadOfGovernment, ORG.CommercialOrganization, Contact.Broadcast, MHI, Life.Die, PER
Contracting the virus	(REMOVED)		Affirm	EN	Egypt	Anyone	"Anyone can get coronavirus."	Anyone, people, confirmed, Philadelphia Inquirer, folks, she	PER.Professional.Scientist, PER.Serviceman, ORG.CommercialOrganization, Contact.Contact, PER
Contracting the virus	(REMOVED)		Affirm	EN		Black folks	To be sure, that means that Black folks can indeed contract the illness, just like everybody else.	whose, father, everybody, Gobert, its, CDC, illness, take, folks, victims, mother	PER.SituationalRole.Victim, socialbehavior.complyobey.unspecified, PER.Serviceman, ORG.CommercialOrganization, MHI, PER.Professional.Athlete, PER
							"Or it could be that gender roles linked to behaviors of men and women— access and control over	awarded, Amuguni,	

Figure 37: An initial view of the Extended Claim Radar Visualizer with some corrections made to Claimers.

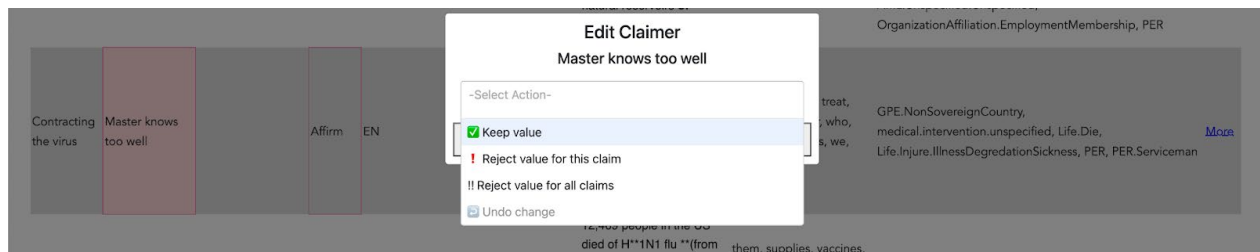


Figure 38: In Extended Claim Radar, the edit value view and the dropdown options upon clicking an editable cell in the table

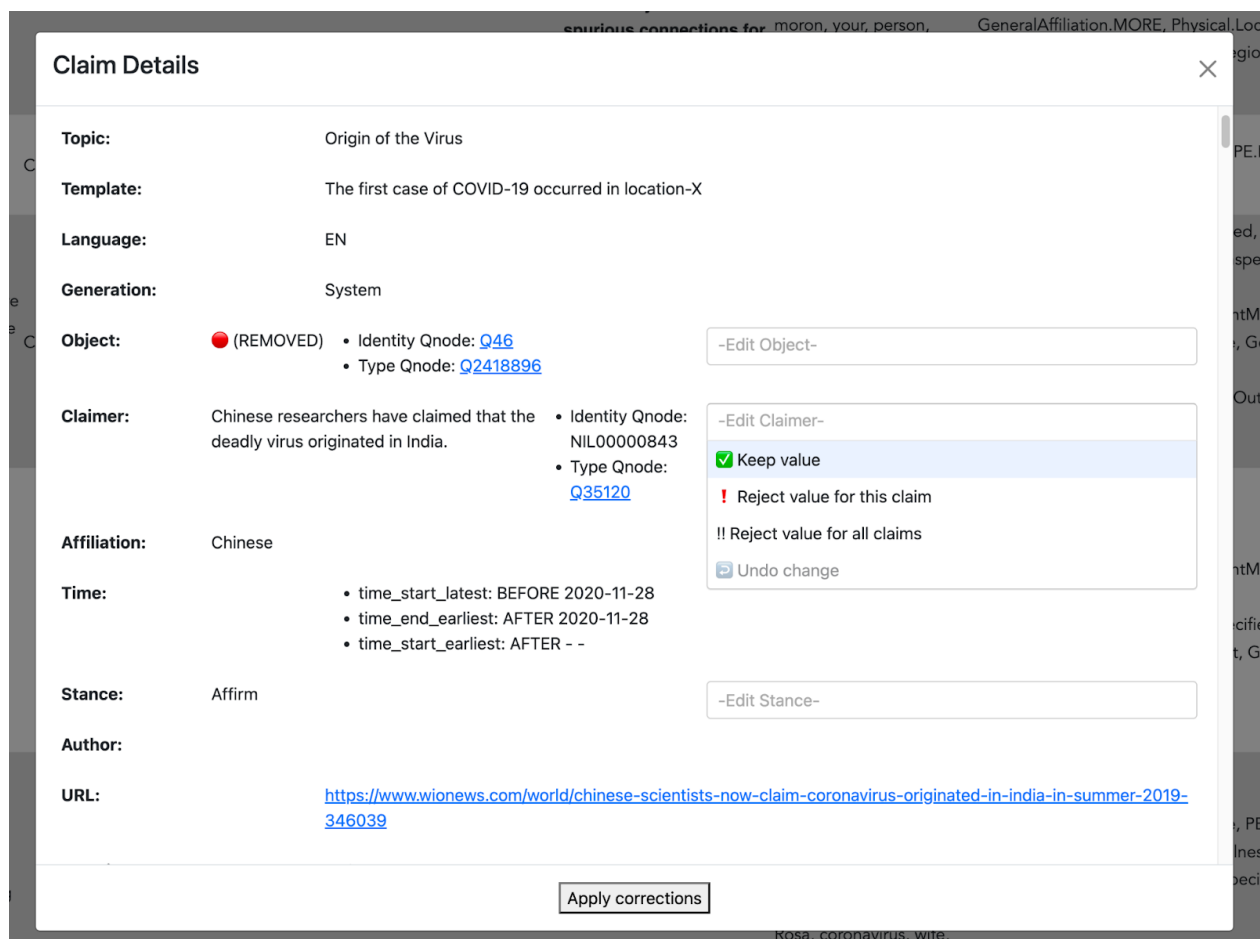


Figure 39: In Extended Claim Radar, the Claim Details view of a claim when logged in.

Claim Details					
ID	Text	Type	KEs with same text and type	KEs with same text, different type	Claims with same KE
EN_Entity_EDL_ENG_0000001	Chinese, Chinese	GPE.NonSovereignCountry	226	0	77
EN_Entity_EDL_ENG_0004507	scientists, researchers	PER.Professional.Scientist	177	90	1
EN_Entity_EDL_ENG_0000007	coronavirus, it	MHI	59	35	15
EN_Entity_EDL_ENG_0000024	India, India, India	GPE.NonSovereignCountry	21	3	4
EN_Entity_EDL_ENG_0000032	Beijing	GPE.NonSovereignCountry	2	9	2
EN_Event_001708	outbreak	Disaster.DiseaseOutbreak.DiseaseOutbreak	14	43	1

Figure 40: In Claim Radar, an example of a Knowledge Element table in Claim Details.

Claim Details

Topic score: 0.452388227

Associated Knowledge Elements:

.....er and sex perspective," said Amuguni [PER.Professional.Scientist], an expert [PER.Professional.Scientist] on the relationship between gender roles and infectious disease [MHI] in global [LOC] health.

"It might be that the men's immune systems are built in a different way from women's," she [PER.Professional.Scientist] said.

"Or it could be that gender roles linked to behaviors of men [PER] and women—access and control over resources and opportunities, power dynamics, and cultural norms that determine what men [PER] or women [PER] do daily—are playing an underlying role in determining who [PER] contracts this infection [life.infect.unspecified]."

To study the many factors potentially at play, Amuguni has assembled a team of researchers from Cummings School and the Tufts School of Medicine. [OrganizationAffiliation.EmploymentMembership] Amuguni [PER.Professional.Scientist] has assembled a team [PER] of researchers [PER.Professional.Scientist] from Cummings School [ORG] and the Tufts School of Medicine [ORG.EducationalInstitution.College].

The scientists [PER.Professional.Scientist] have been awarded [Transaction.TransferMoney] \$50,000 in COVID-19 Rapid Res.....

Equivalent Claims:

- The laboratory focused on the study of viruses that are transmitted from animals to humans. [open](#)
- Coronaviruses are transmitted from animals to humans. [open](#)

Figure 41: In Extended Claim Radar, the display of associated knowledge elements after clicking "awarded."

Claim Details

Topic score:

0.452388227

Associated Knowledge Elements:

.....er and sex perspective," said Amuguni [PER.Professional.Scientist], an expert [PER.Professional.Scientist] on the relationship between gender roles and infectious disease [MHI] in global [LOC] health.

"It might be that the men's immune systems are built in a different way from women's," she [PER.Professional.Scientist] said.

"Or it could be that gender roles linked to behaviors of men [PER] and women—access and control over resources and opportunities, power dynamics, and cultural norms that determine what men [PER] or women [PER] do daily—are playing an underlying role in determining who [PER] contracts this infection [life.infect.unspecified]."

To study the many factors potentially at play, Amuguni has assembled a team of researchers from Cummings School of Medicine, Tufts School of Medicine. [OrganizationAffiliation.EmploymentMembership] Amuguni [PER.Professional.Scientist] has assembled a team [PER] of researchers [PER.Professional.Scientist] from Cummings School [ORG] and the Tufts School of Medicine [ORG.EducationalInstitution.College]. The scientists [PER.Professional.Scientist] have been awarded [Transaction.TransferMoney] \$50,000 in COVID-19 Rapid Res.....

Equivalent Claims:

- The laboratory focused on the study of viruses that are transmitted from animals to humans. open
- Coronaviruses are transmitted from animals to humans. open

Figure 42: In Extended Claim Radar, the display from Figure 41 when hovering over “scientists.”

Covid-19 Claim Radar

View all corrections

View your corrections

Hide corrections

Welcome, guest! Log in to edit.

Username

Password

Login

Origin of the Virus

-Select Claimer-

-Select Claimer Affilia...

Wuhan

bats

-Select Language-

-Select Location-

Search

Reset

-Select KE text-

-Select KE type-

Bind types to KE text

Topic	Claimer	Claimer Affiliation	Stance	Language	Location	Claim Object	Sentence	Associated KEs	Associated KE Types
Origin of the Virus	<AUTHOR>: View My Complete Profile		Affirm	EN		bats	"The project includes studying viral diversity in animal (bats) reservoirs, surveying people that live in high-risk communities for evidence of bat-coronavirus infection, and conducting laboratory experiments to analyze and predict which newly discovered viruses pose the	bats, funding, globally, NIH, that, director, grant, human, EcoHealth Alliance, disease, communities, infection, SARS, people, US, nonprofit snakeperson	ORG.Government.Agency, disaster.diseaseoutbreak.unspecified, AML.Unspecified.Unspecified, life.infect.unspecified, Transaction.TransferMoney, ORG.EducationalInstitution.College, Physical.LocatedNear, PER.Artist.Musician, PER.CivilServant.Spokesperson, LOC,

Figure 43: In Extended Claim Radar, a set of filter queries along with the first result.

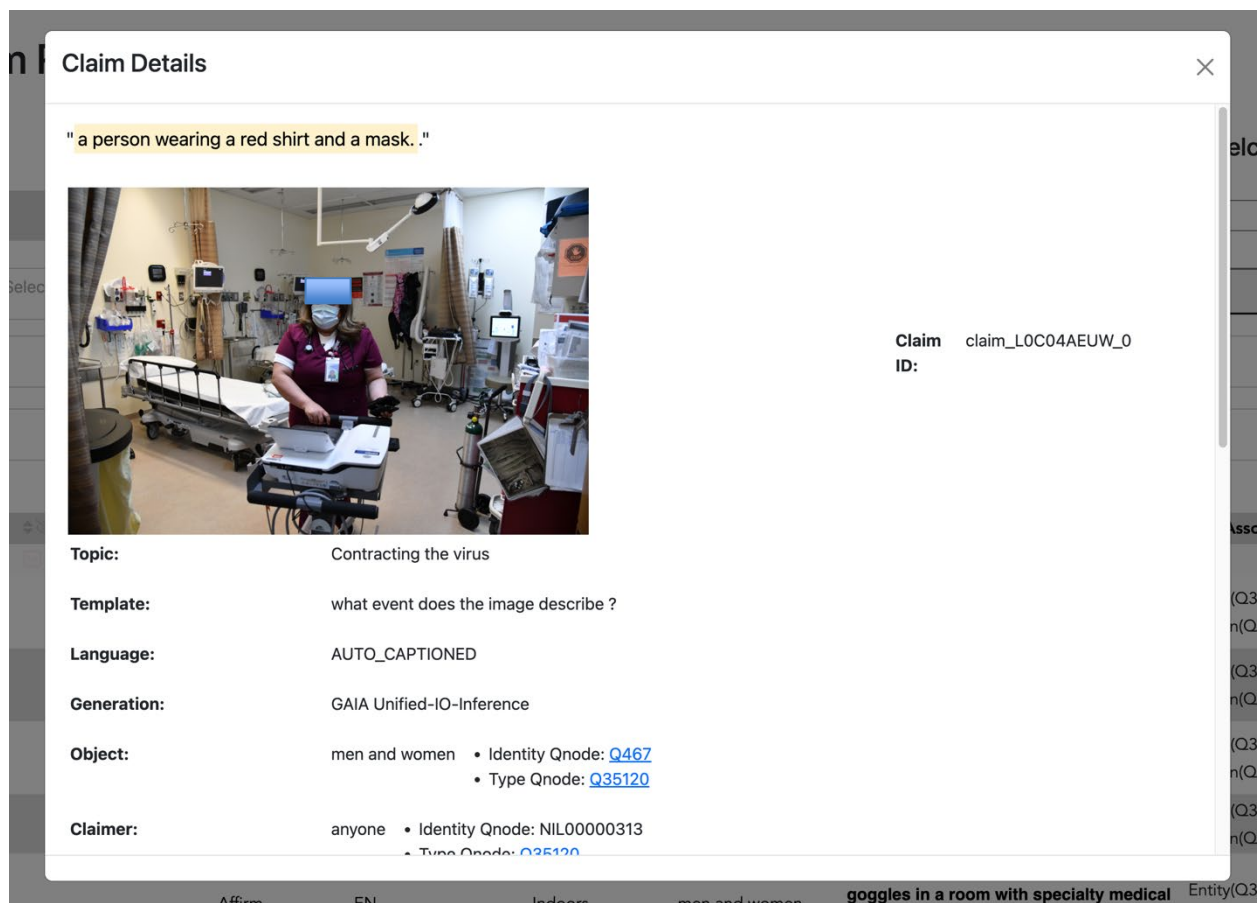


Figure 44: In Extended Claim Radar, A claim detail modal dialog displays a claim including a source image.

4.6.2 SmartBook: AI-Assisted Situation Report Generation

Figure 54 illustrates summaries produced by SmartBook and other baseline approaches. Qualitative evaluations indicate over 80% of questions probe for strategic information, and over 90% of summaries produce tactically useful content. Experts tend to add more information to the generated reports than they remove (under 2.5% of existing tokens), indicating SmartBooks provide analysts with a valuable foundation for situation reports.

Question: How have kamikaze drones changed the nature of the Russia-Ukraine conflict?

Query-Focused	Web Search + LLM	SmartBook
Kamikaze drones, or suicide drones, have been used by Russia in its attacks on Ukrainian cities such as Kyiv, Vinnytsia and Odesa. They are known as a “loitering munition” because they are capable of waiting in an area identified as a potential target and only strike once an enemy asset is identified. Russia is using Iranian-made Shahed-136 drones, which have a payload of approximately 50 kilograms (110 pounds) and are hard to detect due to their small size. Ukraine has responded by developing an app that civilians can use to report sightings of incoming drones and missiles, and increasing the proportion shot down before they hit the ground.	Russia and Ukraine have been using Iranian-made Shahed-136 drones in the conflict. These drones are known as ‘kamikaze’ drones due to their ability to self-destruct on impact. The use of these drones has been highlighted by Russia’s growing reliance on them and their use as both military weapons and psychological weapons. On October 13, 2022, kamikaze drones and missiles hit north and south Ukraine, though the Kremlin denied any involvement from President Putin in talks between Russia and the West to end the war in Ukraine.	The Russia-Ukraine conflict has seen a paradigm shift in the nature of Russian tactics, with the increased reliance of kamikaze drones over artillery fire. These drones, which are designed to crash into a target and explode, can spray potentially deadly debris and are being used to plug gaps in the deployment of Russian forces along the front line. Russia is reportedly using Iranian-made Shahed series drones, which are showcased at Kashan Airfield south of Tehran. This shift in tactics has seen an escalation in the aerial bombardment of densely populated urban areas and an attack on critical infrastructure and power sources.

Figure 45: An example showing the summaries produced by SmartBook and other baseline approaches.

5 CONCLUSIONS

The GAIA project led to numerous publications and advances in multimedia extraction of knowledge elements, integrating those knowledge elements into a knowledge base, and extracting claims from that knowledge base (see: Section 7). In addition to publications, the

project resulted in software that has been provided to both AFRL and to integrated into the program's testbed run by the test and integration performer. In addition to the overall system, the GAIA team developed several interactive visualizations that demonstrate how extracted knowledge elements could be used for analytic tasks. In this report we have described methods and results in each of these areas.

6 REFERENCES

- [AL20] Allaway, Emily, and Kathleen Mckeown. "Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
- [AL21] Allaway, Emily, Malavika Srikanth, and Kathleen Mckeown. "Adversarial Learning for Zero-Shot Stance Detection on Social Media." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.
- [AU16] Isabelle Augenstein, Tim Rocktaschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In EMNLP.
- [BO15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326 (2015).
- [CH21a] Yifan Chen*, Qi Zeng*, Heng Ji and Yun Yang. Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method. Proc. Thirty-fifth Annual Conference on Neural Information Processing Systems (NeurIPS2021).
- [CH21b] Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with selfattention networks. In ACL.
- [CH21c] Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A., ... & Chang, S. F. (2021). Multimodal clustering networks for self-supervised learning from unlabeled videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8012-8021).
- [CH22] Hans Chalupsky, Pedro Szekely (2022). Hybrid Structured and Similarity Queries over Wikidata plus Embeddings with Kypher-V. Wikidata workshop, co-located with the International Semantic Web Conference (ISWC) 2022.
- [CH22a] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji and Shih-Fu Chang. Joint Multimedia Event Extraction from Video and Article. Proc. The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021) Findings.
- [CH22b] Yifan Chen*, Qi Zeng*, Dilek Hakkani-Tur, Di Jin, Heng Ji and Yun Yang. Sketching as a Tool for Understanding and Accelerating Self-attention for Long Sequences. Proc. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022).
- [CO18] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. "XNLI: Evaluating cross-lingual sentence representations." arXiv preprint arXiv:1809.05053 (2018).
- [CO19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin

- Stoyanov. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
- [DE13] Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 120–125.
- [DE15] Lingjia Deng, and Janyce Wiebe. "Mpqa 3.0: An entity/event-level sentiment corpus." *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2015.
- [DE19] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *NAACL-HLT*.
- [DU21a] Xinya Du and Heng Ji. Retrieval-Augmented Generative Question Answering for Event Argument Extraction. *Proc. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022)*.
- [DU21b] Xinya Du, Sha Li and Heng Ji. Dynamic Global Memory for Document-level Argument Extraction. *Proc. The 60th Annual Meeting of the Association for Computational Linguistics (ACL2022)*.
- [EL15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- [HA18] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.
- [HA54] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- [HE19] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.
- [HU20] Lifu Huang and Heng Ji. Semi-supervised New Event Type Induction and Event Detection. *Proc. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP2020)*.
- [IG--] (under review) Ana Iglesias-Molina, Kian Ahrabian, Filip Ilievski, Jay Pujara, Oscar Corcho. Comparison of Knowledge Graph Representations for User Consumption Scenarios. *ESWC*.
- [IL--] (under review) Filip Ilievski, Kartik Shenoy, Hans Chalupsky, Nicholas Klein, Pedro Szekely. A Study of Concept Similarity in Wikidata. *Semantic Web Journal*.
- [II23] KGTK: User-friendly Toolkit for Manipulation of Large Knowledge Graphs. Filip Ilievski, Kian Ahrabian, Ke-Thia Yao, Gleb Satyukov, Jay Pujara. *AAAI 2023 Lab*.

- [JA21] Jayaram, Sahil, and Emily Allaway. "Human rationales as attribution priors for explainable stance detection." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.
- [KA--] Kaldi ASR. <https://kaldi-asr.org/>
- [KA15] Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In TAC.
- [KG--] <https://kgtk.isi.edu/browser>, accessed on 1/4/2023.
- [KL22] Nicholas Klein, Filip Ilievski, Hayden Freedman, Pedro Szekely (2022). Identifying Surprising Facts in Wikidata. Wikidata workshop, co-located with the International Semantic Web Conference (ISWC) 2022.
- [LA21] Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt and Walter Chang. A Context-Dependent Gated Module for Incorporating Symbolic Semantics into Event Coreference Resolution. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021).
- [LA22] Tuan Lai, Heng Ji and ChengXiang Zhai. Improving Candidate Retrieval with Entity Profile Generation for Wikidata Entity Linking. Proc. The 60th Annual Meeting of the Association for Computational Linguistics (ACL2022) Findings.
- [LE22] Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. "Where to start? Analyzing the potential value of intermediate models." arXiv preprint arXiv:2211.00107 (2022).
- [LI01] Dekang Lin and Patrick Pantel. 2001. Dirt@ sbt@ discovery of inference rules from text. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 323–328.
- [LI18] Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018a. Jointly multiple events extraction via attentionbased graph information aggregation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1247–1256.
- [LI19a] Manling Li, Ying Lin, Joe Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani and Heng Ji. Multilingual Entity, Relation, Event and Human Value Extraction. Proc. NAACL-HLT2019 Demo Track.
- [LI19b] Ying Lin and Heng Ji. An Attentive Fine-Grained Entity Typing Model with Latent Type Representation. Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)
- [LI19d] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. CoRR, abs/1906.03158.
- [LI19e] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

- [LI20a] Ying Lin, Heng Ji, Fei Huang, Lingfei Wu. A Joint Neural Model for Information Extraction with Global Features. Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020).
- [LI20b] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji and Shih-Fu Chang. Cross-media Structured Common Space for Multimedia Event Extraction. Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020).
- [LI20c] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski and Marjorie Freedman. GAIA: A Fine-grained Multimedia Knowledge Extraction System. Proc. The 58th Annual Meeting of the Association for Computational Linguistics (ACL2020) Demo Track. [ACL2020 Best Demo Paper Award]
- [LI21a] Sha Li, Heng Ji and Jiawei Han. Document-Level Event Argument Extraction by Conditional Generation. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021).
- [LI22] Li, Liunian Harold, et al. "Grounded language-image pre-training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
- [LI22a] Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan M. Lai, Pengfei Yu, Zixuan Zhang and Heng Ji. COVID-19 Claim Radar: A Structured Claim Extraction and Tracking System. Proc. The 60th Annual Meeting of the Association for Computational Linguistics (ACL2022) Demo Track.
- [LI22b] Ying Lin, Liyuan Liu, Heng Ji, Dong Yu and Jiawei Han. Reliability-aware Dynamic Feature Composition for Name Tagging. Proc. The 57th Annual Meeting of the Association for Computational Linguistics (ACL2019).
- [LI22c] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji and Shih-Fu Chang. CLIP-Event: Connecting Text and Images with Event Structures. Proc. Conference on Computer Vision and Pattern Recognition (CVPR2022).
- [LU18] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL2018), 2018.
- [LU20] Lu, J., Clark, C., Zellers, R., Mottaghi, R., & Kembhavi, A. (2022). Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916.
- [LU21] Di Lu, Clare R. Voss, Fangbo Tao, Xiang Ren, Rachel Guan, Rostyslav Korolov, Tongtao
- [MI09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011.

- [MI20a] Miech, Antoine, et al. "End-to-end learning of visual representations from uncurated instructional videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- [MI20b] Miech, Antoine, et al. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- [NE19] Edward Newell, Drew Margolin, and Derek Ruths. "An attribution relations corpus for political news." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
- [PR08] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. "The Penn Discourse TreeBank 2.0." In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 2008.
- [PR20] Pratt, Sarah, et al. "Grounded situation recognition." European Conference on Computer Vision. Springer, Cham, 2020.
- [RA16] Rao, Jinfeng, Hua He, and Jimmy Lin. "Noise-contrastive estimation for answer selection with deep neural networks." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.
- [RA20] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.
- [RA21] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
- [RA21] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR
- [RE20] Reddy, Revanth Gangi, et al. "UIUC TAC 2020 RUFES System Description." Relation 63 (2020): 67-5.
- [RE22] Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R. Fung, Kathryn Summerville Conger, Ahmed S. Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small and Heng Ji. NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge. Proc. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP2022).
- [RE22] Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi Fung, Kevin Small and Heng Ji. A Zero-Shot Claim Detection Framework using Question Answering. Proc. The 29th International Conference on Computational Linguistics (COLING2022).
- [RE22] Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Li, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avi Sil, Shih-Fu Chang, Alexander Schiwing and Heng Ji. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. Proc. Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI2022).

- [RE23] Revanth Gangi Reddy, Heba Elfardy, Hou Pong Chan, Kevin Small and Heng Ji. SumREN: Summarizing Reported Speech about Events in News. Proc. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023).
- [RE23] Revanth Gangi Reddy, Md Arafat Sultan, Heng Ji and Avi Sil. Entity-Conditioned Question Generation for Robust Attention Distribution in Neural Information Retrieval. Proc. The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2022).
- [SA09] Roser Saurí and James Pustejovsky. "FactBank: a corpus annotated with event factuality." *Language resources and evaluation* 43.3 (2009): 227-268.
- [SA18] Xie, Saining, et al. "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
- [SI19] Damien Sileo, Tim Van-De-Cruys, Camille Pradel, and Philippe Muller. "Mining discourse markers for unsupervised sentence representation learning." *arXiv preprint arXiv:1903.11850* (2019).
- [TA21] Tan, Reuben, et al. "Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos." *Advances in Neural Information Processing Systems* 34 (2021): 14476-14487.
- [TA53] Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.
- [To--] (under review) Riccardo Tommasini, Filip Ilievski, Thilini Wijesiriwardene. The Internet Meme Knowledge Graph. ESWC.
- [VI22] Viz Wiz Captioning Challenge. 2022. <https://eval.ai/web/challenges/challenge-page/739/leaderboard/2006>
- [Via--] VistaOCR. <https://github.com/isi-vista/VistaOCR>
- [WA21] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, David Liem, Ahmed Elsayed, Martha Palmer, Jasmine Rah, Clare Voss, Cynthia Schneider, Boyan Onyshkevych. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021) Demo Track. [NAACL2021 Best Demo Paper]
- [WA22] Jiang Wang, Filip Ilievski, Pedro Szekely, Ke-Thia Yao (2022). Augmenting Knowledge Graphs for Better Link Prediction. *International Joint Conference on Artificial Intelligence (IJCAI)*.
- [WA22a] Zhenhailong Wang, Hang Yu, Manling Li, Han Zhao and Heng Ji. Rethinking Task Sampling for Few-shot Vision-Language Transfer Learning. Proc. COLING2022 Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models.

- [WA22b] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal and Heng Ji. Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners. Proc. The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS2022).
- [WE21a] Haoyang Wen and Heng Ji. Utilizing Relative Event Time to Enhance Event-Event Temporal Relation Extraction. Proc. The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021).
- [WE21b] Haoyang Wen, Yanru Qu, Heng Ji, Qiang Ning, Jiawei Han, Avi Sil, Hanghang Tong and Dan Roth. Event Time Extraction and Propagation via Graph Attention Networks. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021).
- [WI08] Theresa Ann Wilson. 2008. Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states. Ph.D. thesis, University of Pittsburgh.
- [WI17] Theresa Wilson, Janyce Wiebe, and Claire Cardie. 2017. Mpqa opinion corpus.
- [YU21] Pengfei Yu, Heng Ji and Premkumar Natarajan. Lifelong Event Detection with Knowledge Transfer. Proc. The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021).
- [YU22] Pengfei Yu, Zixuan Zhang, Clare Voss, Jonathan May and Heng Ji. Event Extractor with Only a Few Examples. Proc. NAACL2022 workshop on Deep Learning for Low Resource NLP.
- [ZH17] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 35–45.
- [ZH19] Tongtao Zhang, Heng Ji and Avirup Sil. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. 2019. Data Intelligence Vol 1 (2): 99-120
- [ZH19b] Zhong, Ruiqi, Steven Shao, and Kathleen McKeown. "Fine-grained sentiment analysis with faithful attention." arXiv preprint arXiv:1908.06870 (2019).
- [ZH21a] Zixuan Zhang and Heng Ji. Abstract Meaning Representation Guided Graph Encoding and Decoding for Joint Information Extraction. Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2021).
- [ZH21b] Zixuan Zhang, Hongwei Wang, Han Zhao, Hanghang Tong and Heng Ji. EventKE: Event-Enhanced Knowledge Graph Embedding. Proc. The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021) Findings.
- [ZH22] Qi Zeng, Qiusi Zhan and Heng Ji. EA2E: Improving Consistency with Event Awareness for Document-level Argument Extraction. Proc. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT2022) Findings.

[ZH22] Bohui Zhang, Filip Ilievski, Pedro Szekely (2022). Enriching Wikidata with Linked Open Data. Wikidata workshop, co-located with the International Semantic Web Conference (ISWC) 2022.

7 Appendix A - Publications and Presentations

- Akbari, H., Karaman, S., Bhargava, S., Chen, B., Vondrick, C., & Chang, S.-F. (2019). Multi-level multimodal common semantic space for image-phrase grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12476–12486.
- Allaway, E., & McKeown, K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. *ArXiv Preprint ArXiv:2010.03640*.
- Allaway, E., Srikanth, M., & McKeown, K. (2021). Adversarial learning for zero-shot stance detection on social media. *ArXiv Preprint ArXiv:2105.06603*.
- Ananthram, A., Allaway, E., & McKeown, K. (2020). Event guided denoising for multilingual relation learning. *ArXiv Preprint ArXiv:2012.02721*.
- Chan, H. P., Zeng, Q., & Ji, H. (2023). Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization. *ArXiv Preprint ArXiv:2305.14548*.
- Chen, B., Lin, X., Thomas, C., Li, M., Yoshida, S., Chum, L., Ji, H., & Chang, S.-F. (2021). Joint multimedia event extraction from video and article. *ArXiv Preprint ArXiv:2109.12776*.
- Chen, M., Huang, L., Li, M., Zhou, B., Ji, H., & Roth, D. (2022). New frontiers of information extraction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*.
- Chen, M., Zhang, H., Ning, Q., Li, M., Ji, H., McKeown, K., & Roth, D. (2021). Event-centric natural language processing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, 6–14.
- Colas, A., Alvandipour, M., & Wang, D. Z. (2022). GAP: A graph-aware language model framework for knowledge graph-to-text generation. *ArXiv Preprint ArXiv:2204.06674*.
- Colas, A., Sadeghian, A., Wang, Y., & Wang, D. Z. (2021). Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. *ArXiv Preprint ArXiv:2111.00276*.
- Du, X., & Ji, H. (2022). Retrieval-Augmented Generative Question Answering for Event Argument Extraction. *ArXiv Preprint ArXiv:2211.07067*.
- Du, X., Li, S., & Ji, H. (2022). Dynamic global memory for document-level argument extraction. *ArXiv Preprint ArXiv:2209.08679*.
- Du, X., Zhang, Z., Li, S., Yu, P., Wang, H., Lai, T., Lin, X., Wang, Z., Liu, I., Zhou, B., & others. (2022). RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, 54–63.

- Fung, Y. R., Huang, K.-H., Nakov, P., & Ji, H. (2022). The battlefield of combating misinformation and coping with media bias. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4790–4791.
- Fung, Y., Thomas, C., Reddy, R. G., Polisetty, S., Ji, H., Chang, S.-F., McKeown, K., Bansal, M., & Sil, A. (2021). Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1683–1698.
- Gao, H., Shou, Z., Zareian, A., Zhang, H., & Chang, S.-F. (2018). Low-shot learning via covariance-preserving adversarial augmentation networks. *Advances in Neural Information Processing Systems*, 31.
- Han, C., Wang, Z., Zhao, H., & Ji, H. (2023). In-Context Learning of Large Language Models Explained as Kernel Regression. *ArXiv Preprint ArXiv:2305.12766*.
- He, C., Zhu, H., Gao, J., Chen, K., & Nevatia, R. (2020). Cparr: Category-based proposal analysis for referring relationships. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 948–949.
- Hsu, I., Jaiswal, A., Natarajan, P., & others. (2019). Niesr: Nuisance invariant end-to-end speech recognition. *ArXiv Preprint ArXiv:1907.03233*.
- Huang, L., & Ji, H. (2020). Semi-supervised new event type induction and event detection. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 718–724.
- Ji, H., Sil, A., Dang, H. T., Soboroff, I., Nothman, J., & Hub, S. I. (2019). Overview of TAC-KBP 2019 Fine-grained Entity Extraction. *TAC*.
- Ji, H., Sil, A., Huang, S., Dang, H. T., & Soboroff, I. (n.d.). Overview of the TAC-KBP2020 Recognizing Ultra Fine-grained Entities (RUFES) Task.
- Kriman, S., & Ji, H. (2021). Joint detection and coreference resolution of entities and events with document-level context aggregation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 174–179.
- Lai, T., Ji, H., Bui, T., Tran, Q. H., Dernoncourt, F., & Chang, W. (2021). A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. *ArXiv Preprint ArXiv:2104.01697*.
- Lai, T. M., Ji, H., & Zhai, C. (2022). Improving candidate retrieval with entity profile generation for wikidata entity linking. *ArXiv Preprint ArXiv:2202.13404*.
- Li, M., Lin, Y., Hoover, J., Whitehead, S., Voss, C., Dehghani, M., & Ji, H. (2019). Multilingual entity, relation, event and human value extraction. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 110–115.
- Li, M., Reddy, R. G., Wang, Z., Chiang, Y.-S., Lai, T., Yu, P., Zhang, Z., & Ji, H. (2022). Covid-19 claim radar: A structured claim extraction and tracking system. *Proceedings of*

- the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 135–144.
- Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., & Chang, S.-F. (2022). Clip-event: Connecting text and images with event structures. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16420–16429.
- Li, M., Zareian, A., Lin, Y., Pan, X., Whitehead, S., Chen, B., Wu, B., Ji, H., Chang, S.-F., Voss, C., & others. (2020). Gaia: A fine-grained multimedia knowledge extraction system. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 77–86.
- Li, M., Zareian, A., Zeng, Q., Whitehead, S., Lu, D., Ji, H., & Chang, S.-F. (2020). Cross-media structured common space for multimedia event extraction. *ArXiv Preprint ArXiv:2005.02472*.
- Li, M., Zeng, Q., Lin, Y., Cho, K., Ji, H., May, J., Chambers, N., & Voss, C. (2020). Connecting the dots: Event graph schema induction with path language modeling. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 684–695.
- Li, M., Zhang, L., Ji, H., & Radke, R. J. (2019). Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2190–2196.
- Li, S., Ji, H., & Han, J. (2021). Document-level event argument extraction by conditional generation. *ArXiv Preprint ArXiv:2104.05919*.
- Lin, Y., & Ji, H. (2019). An attentive fine-grained entity typing model with latent type representation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6197–6202.
- Lin, Y., Ji, H., Huang, F., & Wu, L. (2020). A joint neural model for information extraction with global features. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7999–8009.
- Lin, Y., Liu, L., Ji, H., Yu, D., & Han, J. (2019). Reliability-aware dynamic feature composition for name tagging. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 165–174.
- Lu, D., Neves, L., Carvalho, V., Zhang, N., & Ji, H. (2018). Visual attention model for name tagging in multimodal social media. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1990–1999.
- Lu, D., Whitehead, S., Huang, L., Ji, H., & Chang, S.-F. (2018). Entity-aware image caption generation. *ArXiv Preprint ArXiv:1804.07889*.
- Ma, X., Kong, X., Wang, S., Zhou, C., May, J., Ma, H., & Zettlemoyer, L. (2021). Luna: Linear unified nested attention. *Advances in Neural Information Processing Systems*, 34, 2441–2453.

- M'hamdi, M., Freedman, M., & May, J. (2019). Contextualized cross-lingual event trigger extraction with minimal resources. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 656–665.
- Reddy, Revant Gangi, Rui, X., Li, M., Lin, X., Wen, H., Cho, J., Huang, L., Bansal, M., Sil, A., Chang, S.-F., & others. (2022). MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 11200–11208.
- Reddy, Revanth Gangi, Chetan, S., Wang, Z., Fung, Y. R., Conger, K., Elsayed, A., Palmer, M., Nakov, P., Hovy, E., Small, K., & others. (2021). Newsclaims: A new benchmark for claim detection from news with attribute knowledge. *ArXiv Preprint ArXiv:2112.08544*.
- Reddy, Revanth Gangi, Chinthakindi, S. C., Fung, Y. R., Small, K., & Ji, H. (2022). A Zero-Shot Claim Detection Framework using Question Answering. *Proceedings of the 29th International Conference on Computational Linguistics*, 6927–6933.
- Reddy, Revanth Gangi, Fung, Y. R., Zeng, Q., Li, M., Wang, Z., Sullivan, P., & others. (2023). SmartBook: AI-Assisted Situation Report Generation. *ArXiv Preprint ArXiv:2303.14337*.
- Reddy, Revanth Gangi, Li, M., Wen, H., & Ji, H. (2020). UIUC TAC 2020 RUFES System Description. *Relation*, 63, 67–5.
- Reddy, Revanth Gangi, Sultan, M. A., Franz, M., Sil, A., & Ji, H. (2022). Entity-conditioned question generation for robust attention distribution in neural information retrieval. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2462–2466.
- Ren, L., Zhang, Z., Wang, H., Voss, C. R., Zhai, C., & Ji, H. (2022). Language Model Pre-Training with Sparse Latent Typing. *ArXiv Preprint ArXiv:2210.12582*.
- Sadeghian, A., Armandpour, M., Colas, A., & Wang, D. Z. (2021). Chronor: Rotation based temporal knowledge graph embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7), 6471–6479.
- Sadeghian, A., Armandpour, M., Ding, P., & Wang, D. Z. (2019). Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems*, 32.
- Shi, G., Feng, C., Huang, L., Zhang, B., Ji, H., Liao, L., & Huang, H.-Y. (2018). Genre separation network with adversarial training for cross-genre relation extraction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1018–1023.
- Shriraj, K. S. K. A. M., & Kumar, S. T. S. (2019). Top K Hypotheses Selection on a Knowledge Graph.
- Spencer Whitehead, H. J., Bansal, M., Chang, S.-F., & Voss, C. R. (2017). Incorporating Background Knowledge into Video Description Generation.
- Sun, K., Maddali, K. A., Salian, S., & Kumar, T. S. (2019). Top K hypotheses selection on a knowledge graph. *The Thirty-Second International Flairs Conference*.

- Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., & others. (2020). COVID-19 literature knowledge graph construction and drug repurposing report generation. ArXiv Preprint ArXiv:2007.00576.
- Wang, Q., Pan, X., Huang, L., Zhang, B., Jiang, Z., Ji, H., & Knight, K. (2018). Describing a knowledge base. ArXiv Preprint ArXiv:1809.01797.
- Wang, X., Wang, H., Ji, H., & Han, J. (2022a). Modern natural language processing techniques for scientific web mining: tasks, data, and tools. Proceedings of the ACM Web Conference, 2022.
- Wang, X., Wang, H., Ji, H., & Han, J. (2022b). New frontiers of scientific text mining: tasks, data, and tools. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 4832–4833.
- Wang, Zhenhailong, Blume, A., Li, S., Liu, G., Cho, J., Tang, Z., Bansal, M., & Ji, H. (2023). Paxion: Patching Action Knowledge in Video-Language Foundation Models. ArXiv Preprint ArXiv:2305.10683.
- Wang, Zhenhailong, Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., Wang, S., Yang, Z., Zhu, C., Hoiem, D., & others. (2022). Language models with image descriptors are strong few-shot video-language learners. Advances in Neural Information Processing Systems, 35, 8483–8497.
- Wang, Ziqi, Han, C., Bao, W., & Ji, H. (2023). Understanding the Effect of Data Augmentation on Knowledge Distillation. ArXiv Preprint ArXiv:2305.12565.
- Wen, H., & Ji, H. (2021). Utilizing relative event time to enhance event-event temporal relation extraction. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 10431–10437.
- Wen, H., Qu, Y., Ji, H., Ning, Q., Han, J., Sil, A., Tong, H., & Roth, D. (2021). Event time extraction and propagation via graph attention networks. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 62–73.
- Whitehead, S., Ji, H., Bansal, M., Chang, S.-F., & Voss, C. (2018). Incorporating background knowledge into video description generation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3992–4001.
- Whitehead, S., Wu, H., Ji, H., Feris, R., & Saenko, K. (2021). Separating skills and concepts for novel visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5632–5641.
- Yu, P., & Ji, H. (2023a). Self Information Update for Large Language Models through Mitigating Exposure Bias. ArXiv Preprint ArXiv:2305.18582.
- Yu, P., & Ji, H. (2023b). Shorten the Long Tail for Rare Entity and Event Extraction. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 1331–1342.
- Yu, P., May, J., & Ji, H. (2023). Bridging the Gap between Native Text and Translated Text through Adversarial Learning: A Case Study on Cross-Lingual Event Extraction. Findings of the Association for Computational Linguistics: EACL 2023, 742–757.

- Yu, P., Zhang, Z., Voss, C., May, J., & Ji, H. (2022). Building an event extractor with only a few examples. *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 102–109.
- Zareian, A., Karaman, S., & Chang, S.-F. (2020). Weakly supervised visual semantic parsing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3736–3745.
- Zeng, Q., Li, M., Lai, T., Ji, H., Bansal, M., & Tong, H. (2021). Gene: Global event network embedding. *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, 42–53.
- Zeng, Q., Zhan, Q., & Ji, H. (2022). EA 2 E: Improving Consistency with Event Awareness for Document-Level Argument Extraction. *ArXiv Preprint ArXiv:2205.14847*.
- Zhang, B., Whitehead, S., Huang, L., & Ji, H. (2020). Global attention for name tagging. *ArXiv Preprint ArXiv:2010.09270*.
- Zhang, T., Ji, H., Chang, S.-F., & Freedman, M. (2020). Training with Streaming Annotation. *ArXiv Preprint ArXiv:2002.04165*.
- Zhang, T., Ji, H., & Sil, A. (2019). Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2), 99–120.
- Zhang, Z., & Ji, H. (2021). Abstract meaning representation guided graph encoding and decoding for joint information extraction. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 39–49.
- Zhang, Z., Wang, H., Zhao, H., Tong, H., & Ji, H. (2021). Eventke: Event-enhanced knowledge graph embedding. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1389–1400.
- Zhou, X., Sadeghian, A., & Wang, D. Z. (2019). Mining rules incrementally over large knowledge bases. *Proceedings of the 2019 SIAM International Conference on Data Mining*, 154–162.
- Zhu, H., Sadhu, A., Zheng, Z., & Nevatia, R. (2021). Utilizing every image object for semi-supervised phrase grounding. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2210–2219.

8 Symbols, Abbreviations, and Acronyms

ACE	Automatic Context Extraction
AFRL	Air Force Research Laboratory
AI	Artificial Intelligence
AIDA	Active Interpretation of Disparate Alternatives

AIF	AIDA Interchange Format
ARC	Argument Reasoning Comprehension
ASR	Automatic Speech Recognition
AWS	Amazon Web Services
BBN	Bolt Beranek and Newman (A different team in the AIDA program)
BERT	Bidirectional Encoder Representations from Transformer
BLIP	Multimodal event model described in [LI22c]
CACI	A different team in the AIDA Program
CDGM	Context Dependent Gated Module
CF	Claim Frame
CLIP	Computer vision model described in [RA21]
COVID	Corona Virus Disease
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CUDA	Compute Unified Device Architecture
DARPA	Defense Advanced Research Projects Agency
DPR	Dense Passage Retrieval
DWD	Darpa Wiki Data
EPGEL	Entity Profile Generation for Entity Linking
ESWC	Extended Semantic Web Conference
EvtGD	Event Guided Denoising
FFNN	Feed Forward Neural Network
GAIA	Generating Alternatives for Interpretation and Analysis. The team described in this report. A collaboration between University of Southern California's Information Sciences Institute (ISI), Columbia University, University of Illinois Urbana Champagne, and the University of Florida
GAIL	Model described in [ZH19]

GB	Gigabytes
GFBF	Good-For-Bad-For
GLIP	Grounded Language-Image Pre-training
GPU	Graphical Processing Unit
GT	Ground Truth
HPCC	High Performance Computing Cluster
IE	Information Extraction
IJCAI	International Joint Conference on Artificial Intelligence
IoU	Intersection over Union
ISI	University of Southern California's Information Sciences Institute
JMEE	Model described in [LI18]
JMMT	Joint Multimodal Transformer
JSON	JavaScript Object Notation
KE	Knowledge Element
KG	Knowledge Graph
KGTK	Knowledge Graph Tool Kit
KL	Kullback-Leibler
LDC	Linguistic Data Consortium
LSTM	Long short-term memory
MERLOT	Multimodal Event Representation Learning Over Time
METEOR	Metric for Evaluaton of Translation with Explicit Ordering
MPQA	Multi-Perspective Question Answering
MTB	Matching the Blanks
NCE	Noise Contrastive Estimation
NIST	National Institute of Standards and Technology
NLI	Natural Language Inference

NLP	Natural Language Processing
OCR	Optical Character Recognition
OPERA	Team from Carnegie Mellon in the AIDA Program
OME	Object Motion-aware Embedding
OSE	Object State Embedding
PDTB	Penn Discourse Treebank 2.0
QA	Question Answering
RDF	Resource Description Frame Work
RoBERTa	Robustly Optimized BERT Pretraining Approach
RTX	Ray Tracing Texel eXtreme
SDP	Shortest Dependency Path
TA	Technical Area 1, 2 or 3, defined by DARPA AIDA Program
TAC KBP	Text Analysis Conference Knowledge Base Population
TFIDF	Term Frequency Inverse Document Frequency
TGA	Topic-Grouped Attention
UIUC	University of Illinois Urbana Champagne
VAST	VARied Stance Topics
VOA	Voice of America
WASE	Model described in [LI20b]
XLM-R	Cross-lingual Language Model - RoBERTa
XNLI	Cross Lingual Natural Language Inference
YAML	Yet Another Markup Language
YOLO	You Only Look Once