AFRL-RI-RS-TR-2023-088

RAMFIS: REPRESENTATION OF VECTORS AND ABSTRACT MEANINGS FOR INFORMATION SYNTHESIS

REGENTS OF THE UNIVERSITY OF COLORADO

MAY 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

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

NOTICE AND SIGNATURE PAGE

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

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

AFRL-RI-RS-TR-2023-088 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
HANNAH M. UVANNI
Work Unit Manager

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

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

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
MAY 2023	FINAL TECHNICAL REPORT	JANUARY 2018	DECEMBER 2022

4. TITLE AND SUBTITLE

RAMFIS: REPRESENTATION OF VECTORS AND ABSTRACT MEANINGS FOR INFORMATION SYNTHESIS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-18-2-0016	N/A	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2FT

6. AUTHOR(S)

Martha Palmer

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

Regents of the University of Colorado
3100 Marine St 572 UCB
Boulder CO 80309-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

DARPA/I2O
675 N Randolph St
Arlington VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/ RI & DARPA/I2O

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-RI-RS-TR-2023-088

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 work was performed under the auspices of the DARPA Active Interpretation of Disparate Alternatives (AIDA) program. The goal was improved multimodal and multilingual information processing aimed at capturing alternative perspectives on significant events. TA1 performers extracted entities and events from individual multimedia documents and passed them along to a TA2 as knowledge elements, usually with reference links to a Knowledge Base (KB). As TA2 performers, our directive was to determine where entities and events in one document were identical with entities and events in another document, and to cluster them together around a single KB link. This facilitated the detection of contradictions and confirmations. In this document, we present our work throughout the program, detailing our evaluation system and how we have measured improvement. We also discuss our efforts in mapping between vector representations and the work we have done on ontology development.

15. SUBJECT TERMS

Elastic search, Longformer, Ontology, PropBank – The Proposition Bank, Roleset, TransE

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	48

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
HANNAH M. UVANNI	**N/A**

Table of Contents

List of Figures

List of Tables

Summary

Research Goals and Technical Approach - As a Technical Area 2 (TA2) performer, the original Ramfis goal was to automatically meld knowledge elements gleaned from different modalities and delivered by multiple Technical Area 1 (TA1) performers into a Common Semantic Representation so that contradictions and confirmations could be recognized. We did this successfully in the first few evaluations and provided the most TA2 knowledge bases built on TA1's from multiple sites of any performer. The original goal was to map from multi-modal vector representations to the Linguistic Data Consortium's (LDC) take on the Active Interpretation of Disparate Alternatives (AIDA) Ontology. Since each TA1 performer was using a unique multidimensional field for vector representations, this also required normalization of vector representations with respect to the ontology, in order to confirm cross-document linking of entities, relations and events. We were successful in this endeavor. As the program evolved, there was very little emphasis on image and video vector representations, so towards the end our focus shifted to text vector representations. Our original plan included progress on AMRs as sentence representations. They proved to be very central to the accurate extraction of entities, events, and relations by TA1's, but they could not be passed to the TA2's so we had no access to them or the original document sets. For the last evaluation TA2s were allowed to integrate their processing with TA1's and TA3's, but with no on-site TA1 we could not take advantage of this opportunity to improve TA1 and TA2 tasks with reciprocal communication and merging. A major shift during the program was the supplantation of the LDC ontology with Wikidata as an Ontology, an effort Colorado led that is continuing under Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS).

TA1 performers extracted entities and events from individual multimedia documents and passed them along to a TA2 as knowledge elements, usually with reference links to a knowledge base (KB). As TA2 performers, our directive was to determine where entities and events in one document were identical with entities and events in another document, and to cluster them together around a single KB link. This facilitated the detection of contradictions and confirmations.

In this document, we present our work throughout the program, detailing our evaluation system and how we have measured improvement. We also discuss our efforts in mapping between vector representations and the work we have done on ontology development. Our primary subgoals, with progress, are discussed below.

Introduction

As mentioned in the Summary, as a TA2 performer, the original Ramfis goal was to automatically meld knowledge elements gleaned from different modalities and delivered by multiple TA1 performers into a common semantic representation (CSR) so that contradictions and confirmations could be recognized.

Our primary subgoals were therefore:

Task 1: Common Semantic Representation (CSR) We initially expected out CSR to be based on AMRs grounded in vector representations. AMRs were expected to provide an appropriate level of abstraction for doing detailed semantic matching of knowledge elements without reference to source texts or images. We did in fact begin with a concerted effort to retrain and update Abstract Meaning Representation (AMR) parsing coverage to include multimodal information. We parsed sample image caption sentences with the transition-based Computational Language and Education Research-Abstract Meaning Representation (CLEAR-AMR) parser. While waiting on TA1 output we acquired the Breaking News dataset and ran Yolo on it to bootstrap our multimodal knowledge base. However, as the program progressed, it became clear that TA2s would never process documents, with or without images, and there would be no role for our AMR parser or the image and image caption processing. Our CSR was initially limited to the LDC Entity, Relation, and Event types in the LDC AIDA Ontology, and therefore was not distinguishable from the information in the Probabilistic Knowledge Base described below. Under the Cross-Program Ontology effort the CSR expanded to encompass most of Wikidata for all performers.

Task 2: Probabilistic Knowledge Base (PKB) The Probabilistic Knowledge Base (PKB) remained a key component for the TA3s throughout the AIDA program. It was initially developed by LDC and handed to the performers. The main TA1 output was an updated version of the PKB based on the documents the TA1 had processed. The documents were expected to contain new information and events (Relations and Events) for known participants (or Entities) as well as introducing new important participants (Entities) with additional Relations and Events. The TA1s were responsible for adding all of this new information as well as within document coreference. In other words, clusters of mentions of the same Entity or the same Event within a single document. The TA2 job was to expand the coreference (the cluster around an Entity or an Event) across documents, i.e., to identify when an additional document was providing information about an Entity or Event that had been mentioned in a previous document. However, the TA2s could not see the original documents, so this cross-document coreference resolution had to be solely based on the TA1 updated entries in the PKB. Our approach to Entity and Event Linking and Resolution is described in more detail in Task 4. Given the atrophy of CSR and the close overlap between PKB and Entity and Event Linking, we ended up merging Tasks 1, 2 and 4 into a single Goal.

Task 3: **Ontology development.** The initial effort by the Ontology Working Group was focused on helping LDC expand their existing inventory of Entities, Relations, and Events to better accommodate the expected AIDA topics. At this stage, Susan Brown acted as Colorado's representative on the Ontology Working Group (OWG) and soon became the OWG's lead in

developing the event portion of the ontology. She also acted as the liaison between the OWG and LDC to help coordinate the development of the ontology with the annotation needs of LDC. When Eduard Hovy left the Carnegie Mellon University (CMU) team to become the Program Manager, Susan became the AIDA OWG overall head. Martha Palmer coordinated University of Colorado's (CU) various AIDA ontology tasks, and then spearheaded the Cross-Program Ontology effort. Numerous Colorado students also contributed to the ontology development. Over time, under the leadership of the Cross-Program Ontology subcommittee, the Ontology effort shifted to focusing on mapping existing LDC types to the Wikidata Ontology, then creating Wikidata based upper models for Entities, Relations and Events, and eventually to expanding the events to include around 5000 PropBank rolesets. Colorado and Brandeis played a central role throughout and is continuing to maintain and expand the Defense Advanced Research Project Agency (DARPA) Wikidata Overlay as a JSON release, available at https://github.com/e-spaulding/xpo .

Initially we had expected the Ontology development to include text and image embeddings, and a main focus of our effort aimed at the PKB and the Entity and Event Linking and Resolution continued to be devoted to processing and aligning vector representations, initially of images and later of text as well. However, this aspect did not end up playing a visible role in AIDA because the TA1s did not pass along vector representations until the very end.

Task 4: Entity and Event Linking and Resolution We had two approaches to Entity and Event Linking and Resolution. One approach included features such as Entities, Relations and Events (ERE) event types, modality labels, entity arguments, and pre-trained lemma trigger word embeddings which were effective for the TA1 task of within-document coreference but not for the TA2 task of cross-document. Our second approach, simple rule-based methods, proved to be a hard baseline to beat for cross-document coreference. After Phase 1 we created a new pipeline using graph embeddings trained on knowledge graphs from TA1 to generate a similarity matrix which could be used for clustering to improve the recall of the system. Given the close reliance on PKB we merged Task 2 and Task 4.

Methods, Assumptions and Procedures

Section 1 PKB, and Entity and Event Linking: Event Trigger and Coreference Annotation Effort

Section 1.1 Cross Document Coreference Resolution

To build initial cross-document event coreference models, we utilized labeled datasets (LDC2016E31, LDC2015E29, LDC2015E68, and LDC2017E24) which were converted into the AIDA interchange format. We followed the AIDA requirements for TA2 evaluation and used features such as ERE event types, modality labels, entity arguments, and pre-trained lemma trigger word embeddings. Using these features, we were able to achieve good results for within document event coreference resolution, however, for cross-document coreference resolution, simple rule based methods proved to be a hard baseline to beat. And so, for the pilot evaluation, we submitted a rule-based two-step coreference system where in the first step, we performed named entity linking by fuzzily matching the name strings. Then, using the entity links and additional event information, such as event types and arguments, we performed an argument matching approach for event linking.

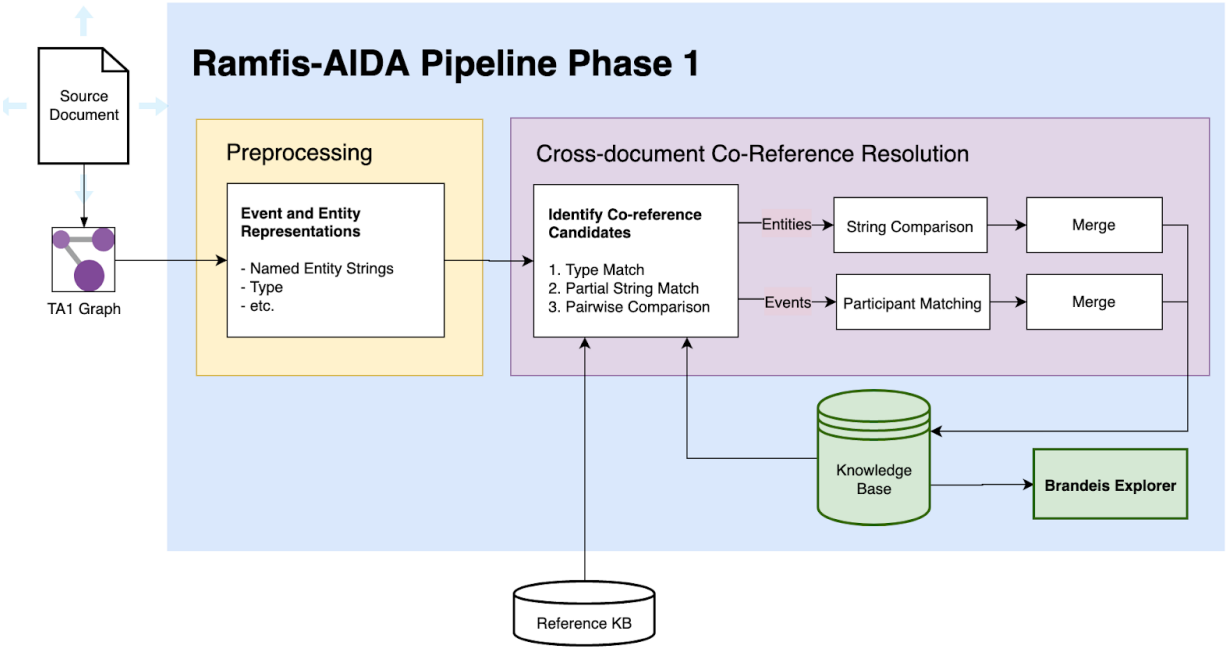

Figure 1. Phase 1 Pipeline Using Type and String Matching to Identify Co-reference Candidates.

Our linking approach was straightforward, but we focused on coreference resolution after merging multiple TA1 outputs. The merging was done as a preprocessing step by matching knowledge elements with the same provenance from different TA1s. We then processed our TA2 system on the merged output, which proved to be beneficial for events and resulted in the ***best event frame recall among TA2 teams in the Month 19 (M19) evaluation.***

After the phase 1 evaluation, LDC released the annotated dataset used by them for evaluating the TA teams. The dataset (LDC2019E77) contained cross-document coreference links based on the Event Hopper guidelines. We used this dataset as a test set for enhancing our coreference system. We improved our system after the M19 evaluation by switching from the brittle rule-based method to an embedding-based method. A new pipeline was created using graph embeddings trained on knowledge graphs from TA1. The TransE architecture was modified to learn embeddings for entities, events, characters, and types to generate the similarity matrix, as shown in **Figure 2**. We performed clustering using this similarity matrix to improve the recall of the system.

The biggest improvement in our system was after incorporating the membrane-breaking information into our pipeline. This information includes the lemma of the event trigger and the sentence in which it was mentioned. We use this information to further rank the event mention pairs for coreference based on sentence similarity and lemma matching. See **Table 1** for event coreference results throughout the program.

Table 1. Cross Document Event Coreference Results on LDC2019E77 Dataset

Project Phase and Method	BCUB F1	MUC F1	Average F1
Ph 1: TA2 system eval version	55.75%	13.76%	34.75
Ph 2: TA2 system joint co-reference	54.21%	20.53%	37.37
Ph 2: TA2 system with Graph Embeddings only	55.53%	43.87%	49.7
Ph 3: TA2 system w/out membrane	60%	61.98	61%

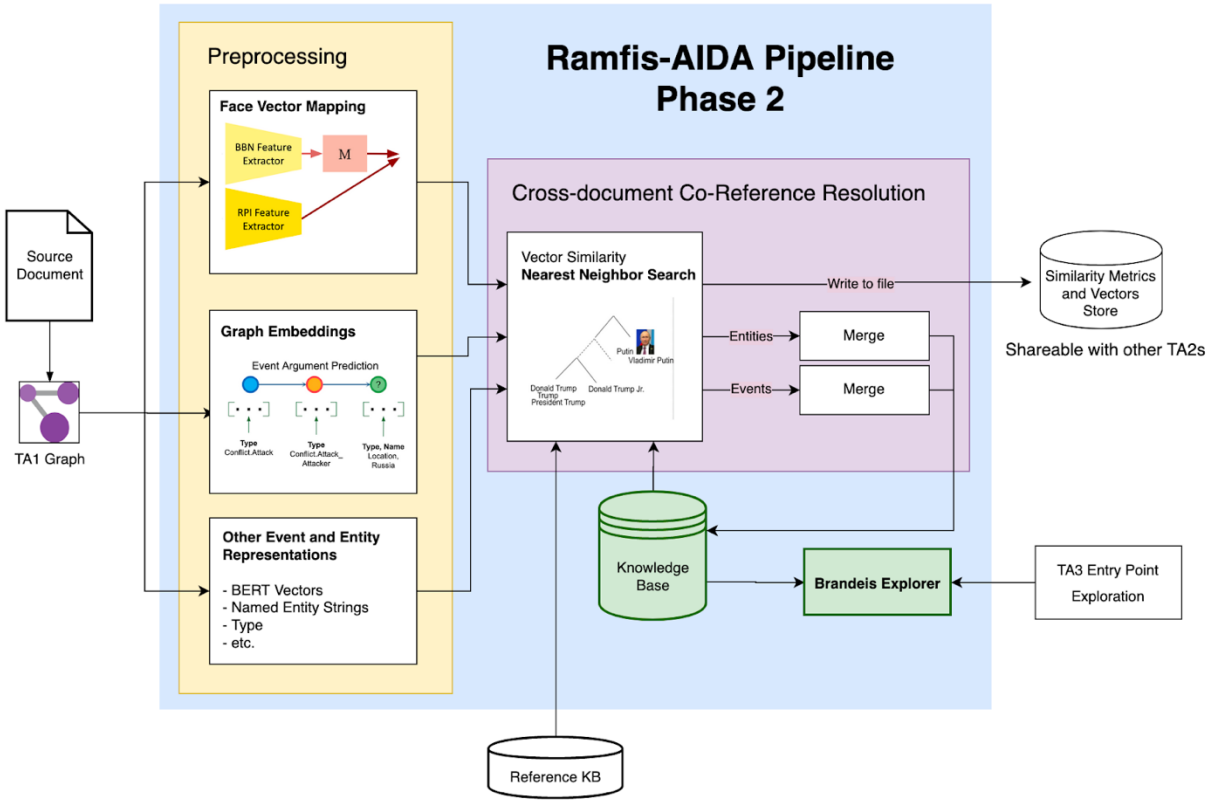

Figure 2. Phase 2 Pipeline Focusing on Integrating Dense Vector Representations of Text, Images, and Graph Embeddings into Co-reference Decisions

In our most recent efforts aimed at event coreference we have used the additional annotation below and the Cross-TA Transformer Mappings developed jointly with Colorado State University (CSU) (see CSU's section below). Abhijnan from CSU assisted Rehan Ahmed at CU Boulder on this effort. Ongoing work includes a two-stage classifier that uses the respective advantages of rule-based and Transformer-based similarity. This approach uses the previously-designed incremental clustering to cluster trivial (easy) vs. non-trivial (hard) examples. This was first completed with the Event Coreference Bank + (ECB+) dataset. We then detect whether an example is trivial using a cross-encoder (based on the Longformer encoder) to encode mentions in a paired fashion. This predicts examples that are lexically trivial or not. We then apply the lexical similarity-based clustering on the trivial examples and use Transformer embedding cosine similarity on the rest. This uses a customized loss function to penalize commutatively identical coreference pairs (i.e., if A and B represent a mention pair, cross-encoder semantic similarity between A and B should be similar to that between B and A). This allows us to improve upon purely lexical or purely embedding-based coreference approaches, and to conduct ablation studies to investigate the contribution of each type of input, and the triviality classifier.

Additional corpora have been parsed and prepared for use in this system: the Gun Violence Corpus (GVC), ECB+ and the LDC data.

Similar coreference work and geometric techniques for manipulation of the embedding space are being used in two papers under review for the ACL conference at time of writing. Ahmed et al. (under review), written in collaboration with CU Boulder, explores a two-pronged strategy for event coreference, where "simple" examples are identified using a lemma-based heuristic, and a long tail of more complicated mention pairs, that require more sophisticated reasoning beyond the surface level that is performed using a Transformer-based classifier. In our latest research, we've made significant progress by combining a Transformer-based coreference classifier with lexical heuristics. This approach achieves comparable results to state-of-the-art methods, while drastically reducing computational costs on two datasets, namely, the ECB+ corpus and the Gun Violence Corpus (GVC), as shown in **Table 2.**

To accomplish this, we first identify synonymous event mention lemma pairs and use sentence token similarity to filter out a vast number of non-coreferent mention pairs. We then focus exclusively on training and inferring the coreference relation on the remaining mention pairs. This results in a balanced set of coreferent and non-coreferent mention pairs for training, and a linear number of mention pairs for inference. This is a marked improvement from previous methods, which trained on a skewed distribution that favored non-coreferent mention pairs and inferred on a quadratic number of mention pairs.

Table 2. Cross Document Event Coreference Results on ECB+ Corpus and GVC Corpus

Method (CoNLL F1 scores)	ECB+	GVC
Bugert et al. (2021)	NA	59.4
Caciularu et al. (2021)	85.6	NA
Held et al. (2021)	85.7	83.7
Our Approach	87.4	75.4

Section 1.2 Event Coreference Annotation Efforts

Rehan Ahmed continued leading Colorado event annotation**,** focusing on event trigger and coreference annotation on the AIDA/LDC datasets. The Venezuela scenario from the Phase 2 evaluation (LDC2020E11) was included, targeting "E203: Assassination attempt of Nicolas Maduro." Seven annotators (3 for Spanish and 4 for English) were trained by **Michael Regan** in using the annotation tool prodi.gy to annotate both English and Spanish documents.
Document Selection. We created three batches of documents (2 English batches and 1 Spanish) English (see **Table 3)**.

Table 3. Documents from the LDC2020E11 Dataset Split between 7 Annotators

Batch1: English	**Batch 2: English**	**Batch 3: Spanish**
IC001VBEZ IC001VGJY KC003AE0U KC003AE39 IC001VBJX JC002Y216 KC003AE20 IC001VGHD KC003ACOV KC003AE2O	IC001VBF1 JC002XZGW JC002YFTP KC003AE1G IC001VBWV JC002Y2LX JC002YGFI IC001VGWW JC002YEO3 KC003A4QB	IC001VBFH IC001VBG6 JC002YFUI KC003ADTY IC001VBFK JC002YEMN KC003ADRQ IC001VBFS JC002YF5N KC003ADRZ
3 annotators	2 annotators	3 annotators

Event Trigger Annotation. We used a modification of prodi.gy annotation tool's named entity recognition annotation recipe for annotating event triggers. The annotators followed a simplified set of guidelines adapted from the Rich Event Description Annotation guidelines and the TimeML specifications. We use a model-in-the-loop style of annotations where at first the model selects spans of text within a sentence that correspond to event triggers. The annotator then corrects the predictions by discarding the incorrect suggestions and selecting the spans missed by the model. The specific workflow (Figure 3) is as follows: For a target mention, the Annotated Event Cluster store presents three potential coreferent candidates. The ranking module (an event coreference resolution (ECR) scorer) then ranks them based on their semantic similarity to the target. The annotator reviews each candidate one-at-a-time and makes decisions on coreference. Candidate 3 is skipped after finding Candidate 2 as coreferent. The cluster store is then updated based on these decisions. See **Table 4** for batch-wise inter-annotator agreement for event spans. **Figure 3** shows the components of the annotation interface, which includes the PropBank website, the current document being annotated, the target event mention, and the form for the rich event descriptors of the target event.

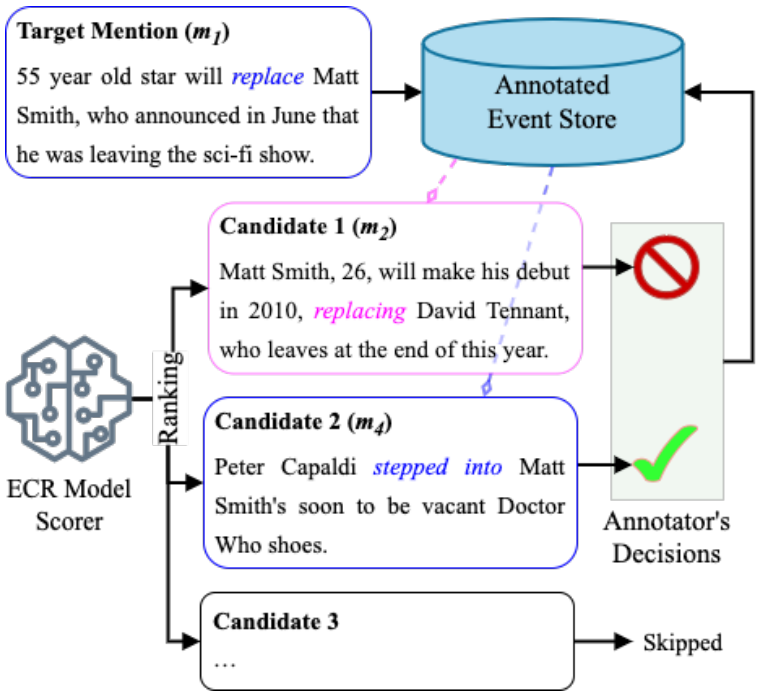

Figure 3. Event Coreference Resolution Annotation Workflow

Table 4. Inter-annotator Agreement for Event Trigger Tagging

Batch 1	Batch 2	Batch 3
0.78	0.84	0.79

Rich Event Descriptors. *With the annotated event triggers, we create descriptors for each trigger by doing word sense disambiguation and semantic role labeling. For this purpose, we use Proposition Bank (PropBank; Palmer et al. (2005); Pradhan et al. (2022)). Each task consists of the event trigger highlighted along with the sentence where it was mentioned. We also provide the PropBank website for the annotators to refer to. The annotators are tasked to create a descriptor for the event trigger of the form: (roleset_id, ARG-0, ARG-1, ARG-LOC, ARG-TIME.*

We use PropBank as a lexical resource to map the trigger's lemma to its correct sense. In a corpus surrounding a particular topic, more often than not, a lemma is used in a single sense. Therefore, we suggest the roleset for a lemma previously tagged by the annotator. Additionally, we provide the PropBank interface in the annotation tool so that the annotator can search for the correct roleset for an event trigger, if there is no roleset associated with the lemma. If the annotator is unable to find a roleset, we ignore that event trigger from the annotations.

To annotate the arguments of an event mention, we adopt the process of manually typing them into the respective input boxes. We input the most informative mention of the arguments by looking through the entire document. For example, for the mention, Nicolas Maduro, we pick the

more informative span: Venezuelan President Nicolas Maduro. When we encounter a pronominal mention as an argument in the target's sentence, we enter the resolved mention in the box.

We successfully completed the annotations for the selected documents in English and Spanish. The next step is to analyze the annotations to assess the agreement between the annotators on the roleset ids and the various Arguments for the event selected by the annotator.

The main research goal from this annotation project is to investigate if the Event Descriptors are sufficient for finding coreference chains. If so, we are able to achieve the task linearly with the event mentions as opposed to the quadratic pairwise method that is followed traditionally.

Section 1.3 Cross-CNN Mappings and Face Embeddings - David McNeely-White, Ben Sattleberg, Shivani Mogullapalli, Ramya Sree Patchava, Bruce Draper, Ross Beveridge, and Nikhil Krishnawamy, Colorado State

The team at Colorado State University (CSU), initially led by **Bruce Draper** and subsequently by **Ross Beveridge,** and finally by **Nikhil Krishnaswamy** (formerly at Brandeis), initially focused on image processing to enable entity linking based on multimedia data. Since features from neural architectures used for image processing may not be directly comparable, CSU solicited system descriptions from multiple TA1 performers which reveal the different models, training methods, and datasets used to extract features from images. CSU demonstrated that the image embeddings generated by convolutional neural networks (CNNs) like those used in TA1 feature extraction can be mapped together, and therefore made directly comparable, by simple affine transformation (McNeely-White et al., 2019). **David McNeely-White** and **Ben Sattleberg** applied this method to AIDA data for cross-TA facial recognition co-reference. While the general approach of the BBN and RPI teams was the same, they used different FaceNet models. FaceNet produces a feature space which allows for comparisons using a simple Euclidean distance threshold for recognition. However, this property may not be preserved when comparing vectors generated using different FaceNet models. The previous work affine vector mapping between CNNs suggested that it might be possible to train a mapping for FaceNet features that would better relate the embeddings provided by the TA1s.

CSU requested face embeddings from the Raytheon BBN Technologies (BBN) and RPI teams generated from the M18 evaluation corpus (i.e., including embeddings which were not identified as a named entity). To pair feature vectors from different TA1s, we compared the bounding boxes of each detection, and took those bounding boxes with the greatest intersection over union (IOU) as the same detection. Negative pairs were then generated by shuffling the members of ground truth pairs. The baseline method consists of comparing the Euclidean distance between two vectors to a threshold. Vector pairs falling below the threshold are predicted to be positive co-reference detections. Using this baseline, feature vectors originating from the same input image are correctly classified about 60% of the time (with a false positive rate of zero). CSU then trained a linear regression model to minimize the Euclidean distance between the vectors from RPI and the corresponding vectors from BBN (or the reverse). Training and test subsets were generated using 6-fold cross validation. These mapped RPI vectors could then be compared

as before using Euclidean distance to their corresponding BBN vectors to determine if they were generated from the same input image. The new model successfully maps over 99% of individual face embeddings from one TA1 embedding space to the other. Using these methods, we learned that we can reliably determine whether two feature vectors come from the same bounding box in the same image.

Subsequently, the CSU team expanded this finding using various large, published datasets and various model architectures. See those models, training datasets, and performance on the popular Labeled Faces in the Wild (LFW) benchmark in **Table 5**. What distinguishes the four models is a combination of the CNN architecture, training loss function and training dataset. The accuracy in the commonly used LFW dataset is provided to demonstrate these are high quality face recognition models. These four models were used to generate embeddings for all faces in the LFW benchmark dataset. By obtaining pairs of LFW face embeddings corresponding to the same individual, mappings can be constructed as in previous work by ridge regression. Those mappings are trained and evaluated with 10-fold cross validation. During each fold, 9 partitions are used to train a mapping and find a distance threshold for nearest-neighbors face verification. Face verification accuracy is then calculated using embedding pairs in the 10th partition, both before and after linear mapping (McNeely-White, 2020). This produces two average accuracies for each combination of the 4 models listed in **Table 5**. See those average accuracies in **Table 6**. A shortened name is provided for easy reference back to each model when showing further results in **Table 6.**

Table 5. Configuration and Accuracy of 4 Face Recognition Models

Name	CNN	Loss function	Training Dataset	Accuracy on LFW
Model-IC	InceptionResNetV1	Softmax + Center	CASIA-WebFace	99.03% ± 0.42%
Model-IV	InceptionResNetV1	Softmax + Center	VGGFace2	99.47% ± 0.37%
Model-MM	MobileNetV2	ArcFace	MS-Celeb-1M	98.70% ± 0.50%
Model-RM	ResNet50V2	ArcFace	MS-Celeb-1M	99.40% ± 0.46%

Table 6. Average Accuracy of Each Network on LFW When Mapped

		To			
		Model-IC	Model-IV	Model-MM	Model-RM
From	Model-IC	**99.03%**	**98.98%** (58.12%)	**98.32%** (51.78%)	**98.67%** (51.00%)
	Model-IV	**98.82%** (57.03%)	**99.47%**	**98.63%** (52.42%)	**98.48%** (52.25%)
	Model-MM	**98.07%** (50.30%)	**98.50%** (51.88%)	**98.70%**	**98.32%** (51.62%)
	Model-RM	**98.60%** (52.45%)	**98.78%** (52.95%)	**98.52%** (52.43%)	**99.40%**

In Table 6, diagonal elements correspond to the unmodified accuracy of models as in **Table 5.** Off-diagonal elements correspond to the accuracy obtained when comparing features across networks, with the "From" model's features mapped by linear transformation to approximate the "To" model's features. The maximum drop in accuracy from any mapping is 1.0%. The number in parentheses indicates the performance when comparing embeddings across networks directly--without mapping. Note that for the task of face verification, 50% corresponds to random chance.

Additionally, the CSU team performed a sensitivity analysis of this linear mapping process. This involved using a random subset of embedding pairs for computing the mapping, observing the effect of fewer examples on mapping quality. In the worst case (i.e., the most sensitive mapping), using only 432 of the original 2,700 pairs produces a mapping which reduces accuracy by only 1% (this is the mapping from Model-IC to Model-RM).

The team also performed dimensionality reduction and found that in all cases, it was possible to reduce the dimensionality of the mapping by at least a factor of eight without losing more than 1% accuracy. These findings made it clear not only that there is a strong underlying similarity between models of varying architecture, but that relatively few corresponding embeddings are needed to solve for the linear transformation that captures the overall relationship between vector encodings.

The above results were demonstrated over an identical set of images that contained verifiably identifiable identities processed using different models. To be truly useful for coreference, it needed to be established whether or not the same identity from different images can be identified using the same process. To accomplish this, first the team at CSU used a small set of labeled face embeddings from BBN and RPI consisting of 295 and 367 identities, respectively. From these sets, there were 67 faces identified and detected by both BBN and RPI. Unfortunately, since face embeddings are expressed as vectors of size 512, 67 paired faces is insufficient to reliably compute the affine mapping between the two embedding spaces.

To overcome this lack of corresponding identity labeled data, CSU went back to our prior work computing the affine mapping between embedding spaces based upon a much larger set of corresponding vectors derived from image co-location within data provided as part of the M18 evaluation corpus. We found roughly 7k paired embeddings based upon co-location in images. From these 7k pairs we were able to compute the affine mapping between the Multi-Task Cascaded Convolutional Network (MTCNN) embedding space and FaceNet embeddings space.

Next, using the 67 common face *identity* embeddings derived from the 295 and 367 labeled faces, this *instance* mapping was evaluated for the purpose of mapping *identities*. While this set was still far too small to draw strong conclusions about this mapping technique we found that After mapping BBN embeddings to RPI's embedding space, a distance-based classifier (nearest neighbor) correctly recognizes 83% of identity samples at a false positive rate of 5%.

Armed with the mapping between the embedding spaces, it now became possible to begin to explore for unidentified people in the data from one TA1, say BBN, where the other TA1 might have also detected (but not labeled) the same person. To prototype this sort of exploration, we used the mapping between embedding spaces to compute the pairwise distance between the 340k BBN embeddings and the 50k RPI embeddings. Next we examined only those pairs which are within a small distance threshold. The result was a small set of filtered pairs deemed likely to be the same person.

Since this linear correspondence apparently allows for reliable comparison across a variety of face recognition systems, the CSU embarked on integrating a dictionary of known faces into our TA2 system for richer inter-TA1 comparison and extraction. Essentially, without having any knowledge of the face recognition systems used by TA1s, our linear correspondence finding allows us to produce face embeddings using our own face recognizer, and linearly map them for comparison to TA1 face embeddings. The dictionary approach (Figure 4) holds the possibility of dramatically expanding the number of identifiable people who may then be coreferenced across TA1 extractions, even when TA1 systems have not explicitly identified those individuals.

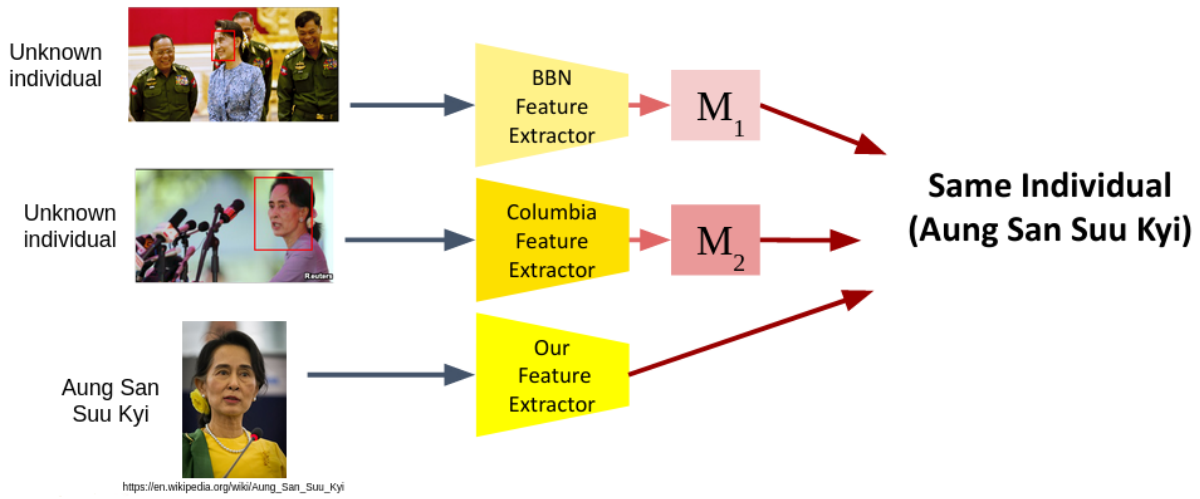

Figure 4. The Face Dictionary Pipeline

Shivani Mogullapalli created a public face set to be used to create a common feature space into which face embeddings from TA1 performers could be mapped. The BBN and RPI datasets were used as initial sources for the faces to be gathered, e.g., from online image sources like Wikipedia, augmented by additional face images of 100 famous people found on Wikipedia whose images were not already in the BBN+RPI superset. A pre-trained model was used to remove images containing multiple faces from the dataset. The remaining images were used to train a common feature space to which other embedding spaces can be mapped. Images were acquired through web scraping, processed through the MTCNN library to isolate, crop, and resize single face images.

This set of images (hereafter referred to as the "CSU dataset") were run through the ArcFace classifier. Embeddings vectors for each face were then extracted from this network. These embeddings (the CSU face embeddings) were then used to compute affine transformations between them and equivalent embeddings from the BBN and RPI datasets, respectively. These transformation matrices form the face dictionary. We then validate the transformation by transforming a face embedding from one feature space to another and computing the cosine similarity between the transformed embedding vector and an embedding of the same person's face from the native feature space.

Ramya Sree Patchava performed evaluation of the face dictionary against new sets of embeddings received from the BBN and Generating Alternative Interpretations for Analysis (GAIA) TA1 teams, which included some identities already present in the face dictionary training data, and some new identities (e.g., in the new BBN data, new identities included Boris Johnson, Xi Jinping, Anthony Fauci, Greg Abbott and Wang Yi). Performance using the previously computed face dictionary transformation matrices tended to be high precision but low recall, particularly on the new GAIA embeddings, which numbered almost 3,500. By updating the transformation matrices using some subsets of the new data, we were able to improve performance and in fact achieved 99% F1 on the BBN data. Low performance on the GAIA data gave rise to some new strategies. In particular, the choice of embedding pairings used to construct the transformation matrices between embedding spaces is particularly important.

In the final evaluation of the face dictionary with the available data, we used four primary methods of choosing embeddings to create cross-embedding space mappings and to calculate cosine similarity relative to for clustering, and then evaluated on a binary classification task where a positive match was one where mapped embeddings of one identity from the source embedding space clustered with embeddings of the same identity in the target embedding space (i.e., fell within the defined cosine similarity threshold with embeddings of that identity). The most successful method was to use the mean source (BBN/RPI) embeddings for identity *I* and calculate cosine similarity relative to the mean target (CSU) embedding for identity *I.* Results are shown in Table . Results from other techniques are given in the appendix.

Table 7. Results of Cross-embedding Space Mappings on Binary Classification Task

	Accuracy	**Precision**	**Recall**	**F1**
RPI → CSU	.99	.55	1.0	.67
BBN → CSU	.97	.89	1.0	.94

Section 1.4 Cross-TA Transformer Mappings - Abhijnan Nath, Shriram Gaddam, and Nikhil Krishnaswamy, Colorado State

The aforementioned findings in the vision domain raised the question of whether the same interchangeability could be applied to the embedding spaces of language models.

Shriram Gaddam developed a pipeline to extract token embeddings from Transformer architectures, focusing on the Bidirectional Encoder Representations from Transformers (BERT) and A Light BERT (ALBERT) models. With Nikhil Krishnaswamy, he explored the properties of constructing linear mappings between Transformer embedding spaces, focusing on named entities to replicate the face dictionary task most closely. Early experiments indicated that sets of contextualized embeddings for a given named entity create subspaces for which said embedding vectors form the spanning set. This suggested that, like facial recognition, co-reference resolution may be possible between different embedding spaces.
Abhijnan Nath and **Huma Jamil** expanded the work to encompass Robustly Optimized BERT-Pretraining Approach (RoBERTa) and Cross-Document Language Model (CDLM) Longformer models. After embedding extraction, preprocessing was carried out to get the CDLM concatenated scores by using two sentences simultaneously (instead of one as was done above) with two mentions/events as the input to the CDLM. The CDLM scores for the events were finally generated within each subtopic from the newly created mention maps. Subsequently, Abhijnan led an experiment on affine transformations between embedding spaces with entity and event mentions, applying the principles of the face embedding work to language models. We created mappings between CDLM, BERT-base, BERT-large, and Coref-BERT embeddings and evaluated the performance of these mapped embeddings compared to the native embeddings on the coreference resolution task against the Phase 1 LDC unsequestered data, which contains gold coreference clusters for entity and event mentions. The three BERT variants are already very similar models, but CDLM is based on Longformer which is based on RoBERTa, and so the family resemblance is more distant. We used precision, recall, and F1 with respect to B-Cubed and MUC score as our primary evaluation metrics (see **Tables 8 and 9**).
We found that while the metrics matter greatly (i.e., B-Cubed and Message Understanding Conference (MUC) score-based evaluations present *very* different pictures of which model performs the best, whether considering mapped or native embeddings, mapping vectors between embedding spaces often produced near comparable performance, bolstering the conclusions from the face embedding work. We also observed many instances where mapped vectors actually *outperformed* native vectors (see blue and red highlights in **Table 8** below). One striking example of this was 768D BERT-base vectors mapped into 1024D BERT-large space outperformed native BERT-large vectors, despite the implicit noisiness in a transformation that goes up in dimensionality (see green highlight). All the BERT variants largely performed similarly across all mappings, and (depending on the metric used), mapped CDLM vectors outperformed native vectors, and mapping other models' vectors into CDLM space achieved near-comparable performance.

Table 8. Event Coref F1 (see Table 1) Performance Using Mapped Vectors: MUC score, with similarity threshold of .6

Source Space → / Target Space ↓	CDLM	BERT-Base	BERT-Large	Coref-BERT
CDLM	80.70	62.52	62.52	58.48
BERT-Base	63.40	60.54	61.01	57.02
BERT-Large	58.73	57.32	52.97	52.32
Coref-BERT	61.60	60.23	59.57	57.84

Table 9. MUC Entity Coref F1 (see Table 1) Performance

Using Threshold of .75 (left) and .9 (right)

Source Space → / Target Space ↓	CDLM	BERT-Base	BERT-Large	Coref-BERT
CDLM	79.89	59.00	59.12	54.81
BERT-Base	69.48	53.24	51.40	48.34
BERT-Large	68.85	49.86	51.36	45.76
Coref-BERT	70.87	56.88	56.00	53.79

Source Space → / Target Space ↓	CDLM	BERT-Base	BERT-Large	Coref-BERT
CDLM	79.89	58.92	39.96	54.55
BERT-Base	57.65	39.39	39.55	36.35
BERT-Large	57.63	39.60	37.08	36.17
Coref-BERT	58.06	39.92	39.96	39.70

Many of the above observations were not seen in the face embedding experiments by McNeely-White et al. (see below). One possible reason for this is that native face recognition performance is already very high, whereas state of the art (SOTA) on coreference is much lower, leaving room for improvement. We also observed similar trends across entity and event coreference tasks. In addition, when varying the cosine similarity threshold cutoff from .6 up to .95, CDLM scores did not change at all, suggesting that CDLM vectors are already clustered together in a tight high-dimensional cone, and coreferents that are detected with a threshold of .95 will of course be detected with a threshold of .6.

Section 1.5 Error Analysis Using the Brandeis Explorer - Peter Anick

The Brandeis Explorer tool was initiated during Phase 2 to make it easier to detect and troubleshoot coreference errors in the knowledge graphs produced by TA2 performers. Version 1

(**Figure 5)** provided a simple web-based user interface for examining events, relationships, and role fillers without requiring knowledge of the underlying graph structure or query language. Version 2 added extensions for examining entities and coreference clusters. Two methods of accessing entity clusters were provided – directly by name and indirectly as fillers of events and relations. Coreferenced mentions could be compared by name, ontological type, and document location.

The tool brought to light a number of cases in which clusters were not combined that should have been, as well as cases where merging clusters introduced false coreferences that propagated during subsequent merges, as illustrated in **Figure 6.**

Version 3 exploited the addition of more temporal and locational information into TA2 knowledge bases by using Explorer functionality to retrieve events by location and sorted by time. This helped to identify new opportunities for clustering events. It also brought to light the risks of using time information, since in some cases the time of the article did not match the time of the events described. Other issues, such as entity clusters with logically incompatible ontology types, cases where accent folding could have enhanced named entity clustering, and proliferation of incompatible names within some entity clusters were also discovered.

Version 4 was a major rewrite of the system to accommodate the migration in AIDA Phase 3 to the use of DARPA Wikidata (DWD) Qnodes and the ability for TA2 performers to see the original document texts. Based on an underlying elasticsearch index of mentions, entities, clusters, Qnodes, and sentential contexts, the system used elasticsearch queries to soft-match mention texts and then organize results by cluster.

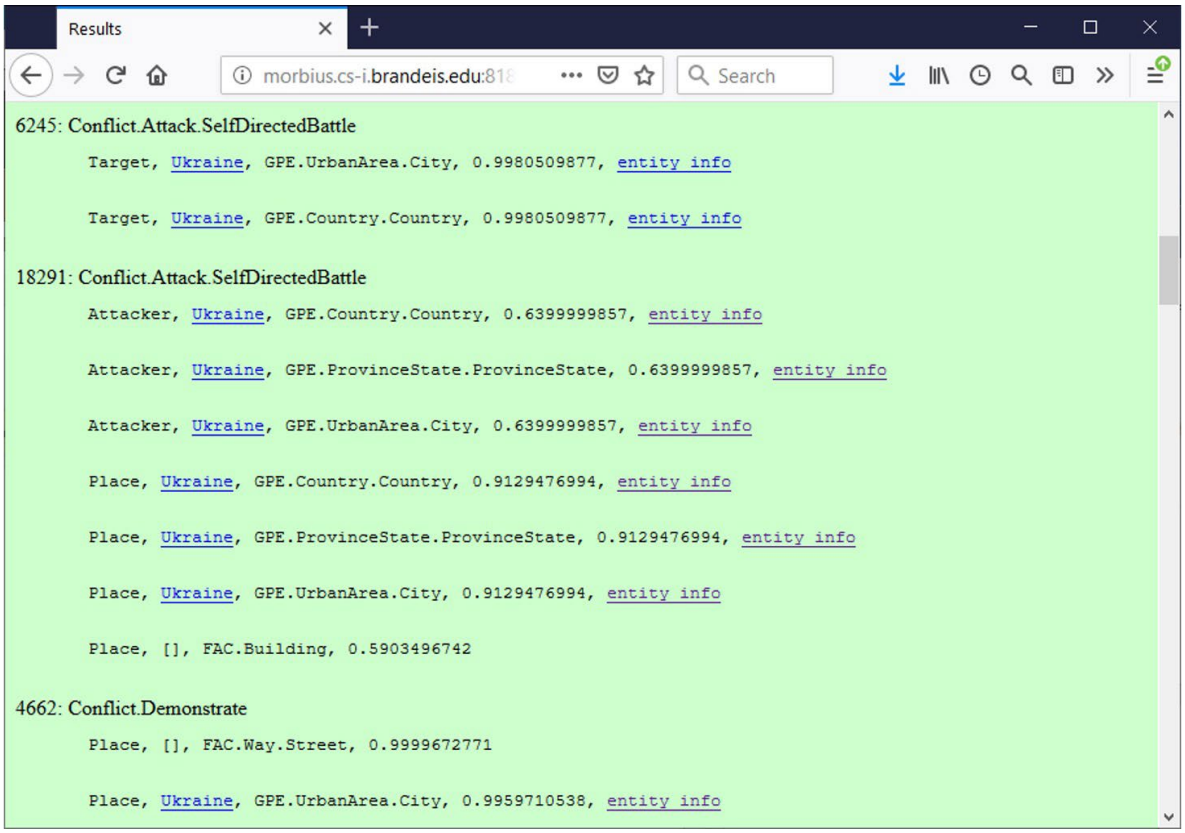

Figure 5. Example of Explorer UI Showing Event Instances with Their Role Fillers

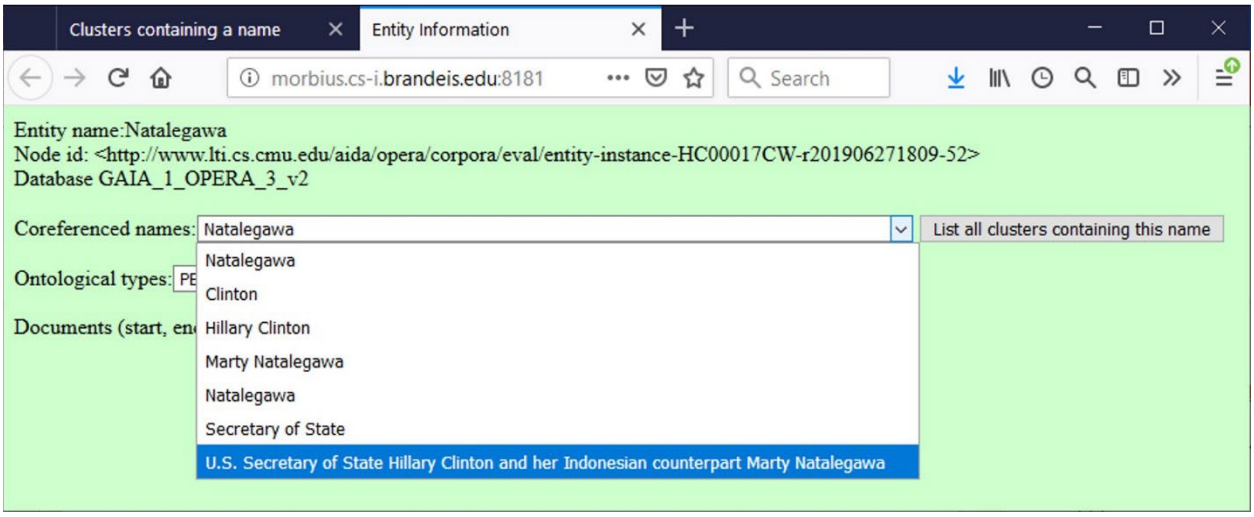

Figure 6. Example of Incorrect Coreferences Propagated by the Improper Handling of Conjunction in the Last Mention Text

Browsing clusters in this manner revealed a number of entity coreference false alarms and misses. Many of these were due to errors in the assignment of textual mentions to Qnode instances or types. For example, “FDA” was correctly assigned to the type/instance of *organization/Food and Drug Administration*, whereas “The FDA” was assigned to *organization/unknown*. “George Soros” was correctly assigned to *person/George Soros*, while last name “Soros” was incorrectly assigned to *city/Soros*.

Colorado’s TA2 clustering rules required that both Qnode instance and type matched in order for entity clusters to be merged. However, after browsing the Explorer results, we determined that this rule resulted in missing many legitimate coreferences when the same instance Qnode was assigned to different but compatible type Qnodes. By surveying the most frequent instance Qnodes which had multiple type assignments, we manually generated a set of compatible types for which our clustering rule could be overridden. In the list below, the bolded term represents the best, or most specific type for the instance. Types preceded by a ? are incorrect type assignments that could still be reasonably clustered with an entity of the bolded type if the mention is an acronym or capitalized single word. Examples of entity instances are shown in parentheses.

broadcaster, organization, corporation, international_organization, ?person (ABC, Amazon)

country, political_territorial_entity (Afghanistan)

continent, geographic_entity (Africa)

military officer, politician, person (Alexander Mishkin)

organization, international_organization, corporation, political_territorial_entity, ?professional, ?politician, ?person (Axios)

professional, person (Alex Jones)

geographic_entity, geographic_region (Arabian Sea)

minister, person, professional, politician (Arauz)

manufacturer, corporation, organization,?political_territorial_entity (AstraZeneca-Oxford)

militia, political_territorial_entity, organization,?person, ?head_of_government (Antifa)

administrative_territorial_entity, political_territorial_entity (Arizona)

Section 2 AIDA Ontology to Cross-Program Ontology (Ontology development with linking to Wikidata)

Early in the AIDA program, Susan Brown acted as Colorado's representative on the Ontology Working Group (OWG) and soon became the OWG's lead in developing the event portion of the ontology. She also acted as the liaison between the OWG and LDC to help coordinate the development of the ontology with the annotation needs of LDC. When Eduard Hovy left the CMU team to become a Program Manager, Susan became the AIDA OWG overall head. Martha Palmer coordinated CU's various AIDA ontology tasks, and then spearheaded the Cross-Program Ontology effort. Numerous Colorado students contributed to the ontology development, as credited below.

Ontology development for AIDA began as an open discussion with representatives from each AIDA performer. This Ontology Working Group was initially led by Eduard Hovy and started by exploring three primary options: (1) developing a custom ontology for the program from scratch, (2) using an existing large-scale, general ontology, or (3) expanding on the DARPA ERE type taxonomy to accommodate new domains. Colorado's initial work for the AIDA ontology focused mainly on establishing the needs the ontology would fulfill and what resources were already available to build upon. We consulted with TA1 and TA3 performers on the granularity and types of ontology classes would be useful for them. To illustrate to them and the ontology working group the types of resources we have and how they could be used, we created easily viewed Rich Event Ontology hierarchies and files showing the linkings and alignments between it, ERE, FrameNet and VerbNet. We also used these linkings to compile argument roles that seemed pertinent to AIDA.

We analyzed the seedling corpus to discover event types of particular importance to this domain, especially those areas not covered by ERE event types. We suggested additions and revisions to the event portion of the draft AIDA ontology in preparation for the Ontology Summit,and continued that work at the summit itself. At the Summit, it was decided to go primarily with option 3) expanding the LDC ERE ontology, with the understanding that a great deal of expansion and re-organizing was needed for the ERE type taxonomy to support the AIDA program.

At this point, Colorado took the lead in developing upper and middle models of the event portion of the AIDA ontology and in proposing more detailed event ontology types (what became known as the "fringe" ontology). Colorado also suggested generalized features that could enhance the available information for multiple event types, such as +legal or +violent. These features obviate the need for multiple inheritance, a goal decided on by the OWG. We also explored developing vector representations for the AIDA ontology types.

Section 2.1 Reconciliation of LDC Ontology and OWG Ontology

As the OWG settled on middle and upper models for event and entity types and continued to expand the "fringe" types, the Linguistic Data Consortium translated the concept types into a

hierarchy and format that conformed more closely to the ERE taxonomy they had been using to annotate data for previous DARPA programs. What was termed the LDC ontology (LO) differed in some significant ways from the OWG ontology (OA). In an effort to bring the two closer together, Susan Brown from Colorado, as the representative from the OWG, and LDC met weekly for about 6 months.

Colorado conducted a detailed comparison of the LO and AO. After consulting with LDC on the meanings of some of the LO types, we resolved hundreds of discrepancies between the two ontologies in the entity and event sections. Many involved missing entities. Some key AO types were added to the LO, and all missing LO types were added to the AO. For the AO types that had not been included in the LO, we indicated a more general LO type that could be considered a "parent" of the AO type. For LO types that were not in the AO, we found appropriate places in the hierarchy and added those types to the AO. Once these simpler differences were resolved, ongoing discussions dealt with types in the AO that had no appropriate slot in the LO, such as abstract entities or enabling or preventing events, and basic representation issues, such as whether to treat crimes as entities or events that allowed arguments. This last issue was resolved by changing Crime entities into events and allowing them to take other LO event types as arguments. Allowing other events (e.g., *learn*) to take other events as argument fillers (e.g., *learn to build,* with *to build* as the topic filler), was put off for the time being.

Colorado annotated finalized LO types and AO fringe types (which were not included in the LO) on AIDA data, jointly with ISI and JHU.

As a baseline for incorporating vector representations, Colorado generated vectors using an auto-encoder approach; specifically, a variant of Rothe and Schutze (2015) simplified to use simple concept embeddings. Since the LO ontology has mappings to PropBank, FrameNet, VerbNet and ERE, and those in turn have mappings into WordNet concepts (synsets) or individual words (lemmas), we had a large set of situations where a concept mapped to many individual English words, and many hierarchical relationships between concept nodes. The auto-encoder approach learned how to weight individual words such that their sum best identified a concept embedding (essentially learning which words are prototypical for that concept), and learned how to combine those concept embeddings to recover the original word embedding, with added constraints to keep hierarchically nearby concepts similar to each other. This provided simple, robust concept embeddings for the LO ontology sub-types and sub-sub-types, which we used as a starting point for TA2 coreference models.

Colorado also trained VerbNet class embeddings on Wikipedia data where verb tokens had been automatically tagged with VerbNet classes. We have demonstrated that these VerbNet embeddings can provide an improvement over Elmo embeddings for a metaphor detection task. Colorado then ran experiments on the seedling data with both sets of embeddings, as well as Elmo, to see if we could associate embeddings with TA1 entities and events that could improve our similarity scores for entity and event coreference.

Section 2.2 Cross-Program Ontology

In early 2020, the decision was made to combine ontology efforts across three DARPA programs: AIDA, KAIROS, and Modeling Adversarial Activity (MAA). Martha Palmer led the new Cross-

Program Ontology (XPO) effort, and Colorado continued coordinating with LDC on the AIDA Phase 2 Annotation Ontology. In April the cross-program ontology working group began calls and quickly started exploring the benefits of mapping from the AIDA Annotation Ontology to the Wikidata Ontology.

At the PI Meeting in June, the XPO subcommittee (Martha Palmer, Susan Brown, Anatole Gershman, Rosario Uceda-Sosa, Pedro Szekely, Sumit Purohit) proposed the Cross Program Ontology (XPO), an expressive representation model that could serve as a core framework for ongoing DARPA language and semantic research across multiple programs. XPO would be an extension and reworking of the AIDA Program Ontology, would also merge in the KAIROS Event Primitives and would be mapped to the Wikidata Ontology.

The LDC AIDA Annotation Ontology is organized around limited sets of annotation tags easily understood by annotators. However, these tag structures can be loosely aligned with more complete ontologies developed by program performers. Our plan was to develop an explicit, skeletal, general purpose upper-middle level ontology that could provide a reasonable superstructure for the existing annotation tags and allow for rational future expansion for all teams. This is the same approach that was used in Deep Exploration and Filtering of Text (DEFT) to develop the Reference Event Ontology (REO). Starting from the LDC Entity, Relation and Event Type definitions, the REO developers defined an upper-level event ontology that supported informative mappings between ERE and rich pre-existing lexical resources such as FrameNet and VerbNet. Much additional work was needed for a more comprehensive, general purpose upper-middle level ontology, as described in more detail below. To achieve our cross-program goals we also needed to merge the KAIROS and AIDA Annotation ontologies more explicitly, revising the over-arching AIDA Program Ontology as needed during the process. KAIROS had adopted some AIDA Event types as-is but had occasionally also added argument slots and slot constraints, and/or generalized to more coarse-grained versions. New Event, Entity and Relation types had been added as well. Overlaps and partial matches had to be clearly marked and reconciled where possible in our merger.

To extend beyond AIDA and KAIROS domains, XPO needed to cover a broad set of topics, ranging from armed conflicts and terrorist activities to elections and epidemics. The key question was which concepts to include in our XPO upper-middle model and at what level of detail, so they could be extended quickly and automatically. By keeping our merged KAIROS/AIDA Annotation tag set in mind while selecting concepts for our general purpose upper-middle level, we ensured a useful mapping between the two. That became the basis for our Cross-Program Ontology (XPO), and we expected to borrow heavily from the AIDA Program Ontology, REO, and IBM's GLO, as well as other resources. We had a program-wide consensus on the use of principled multiple inheritance, perhaps with curation based on attested cross-lingual metonymies. We also agreed that future expansion would greatly benefit from grounding XPO concepts in a very large-scale publicly available ontology such as the Wikidata Ontology, which became the main organizing principle of XPO as time went on.

XPO developed to contain three parts:

- DARPA Wikidata Overlay: an upper-middle ontology that includes a relatively small number of classes that a human can browse and understand, pulled from the broader

DARPA Wikidata Ontology. This included low level ontology Qnode and Pnode mappings for all of the ERE types from the LDC merged AIDA/KAIROS ontology/annotation tag set.

● DARPA Wikidata Ontology (DWD): a more extensive, general ontology that includes many thousands of classes imported from Wikidata, too large to browse and understand. For AIDA, this obeyed the Time Machine rule, allowing nothing post-2010.

● Wikidata: the full Wikidata ontology, available to performers for millions of concepts and entities present in Wikidata.

Section 2.3 Developing the DWD Overlay

The first task in this shift to Wikidata involved mapping the 200+ AIDA/KAIROS ERE types, sub-types and sub-sub-types to Wikidata Ontology Qnodes. Anatole Gershman and Rosario Uceda-Sosa initiated the mapping of entities to Wikidata Qnodes, followed by Ghazaleh Kazeminejad and Elizabeth Spaulding's review and reconciliation of conflicts. Events are more difficult to represent in large-scale, general ontologies, and mapping AIDA events to Wikidata Qnodes was unsurprisingly more difficult than mapping the entities. Colorado's Daniel Chen and Adam Pollins performed a first pass of mapping AIDA events to Wikidata Qnodes, either a single one or a union of several. Ghazaleh Kazeminejad and Susan Brown performed a second pass, resolving disagreements and coming to a consensus, carefully examining Wikidata titles, definitions, and subclasses. In cases of dispute between a Qnode and its superclass, the superclass was selected to ensure wider coverage. Several AIDA event types were found to have no plausible Wikidata Qnode. The first draft of the mapping for entities and events was completed by Colorado in mid-2020 and was then circulated among the Cross-program Ontology Working Group for feedback.

Mapping the AIDA relations to Wikidata required careful manual effort. After comparing the differences between Wikidata Qnodes and Pnodes, the OWG determined that as many relations as possible should be mapped to Pnodes. This mapping was done by Elizabeth Spaulding, James Pustejovsky, and Peter Anick.

- Karan Praharaj implemented algorithms for providing partial credit for types that are near neighbors of LDC annotation types as part of the effort to develop automatic similarity metrics.

JSON Format. Colorado was tasked with the role of consolidating the DWD overlay mapping into a single JSON format. Ghazaleh Kazeminejad led this effort and Elizabeth Spaulding helped, while feedback from the OWG and XPO subcommittee was continuously integrated. The JSON consists of different dictionaries for events, entities, and relations. The key for each dictionary entry is the unique DWD identifier, and each value is another dictionary that provides the mapping information, as well as information about event and relation arguments, similar Wikidata Qnodes, and other information from Wikidata. The similar Qnodes (similarity set (SS) and nearest neighbor (NN)) are determined by human similarity judgment experiments led by Ghazaleh Kazeminejad. The JSON went through several format revisions, with version 5.3 being the latest version. The

overlay is now hosted on a Github repository[1], where program participants can publicly open issues and submit fixes.

For creating the JSON, it was agreed that every Wikidata Qnode or Pnode that represents an event, entity, or relation, should be represented by a DWD node that consists of the string 'DWD' followed by the numerical suffix of the Wikidata Qnode (e.g., Q1527 → DWD1527). This unique identifier maps to a single Wikidata entry, which in turn might map to one or more LDC types. For instance, DWD481609 which represents Q481609 (damage), has two LDC event types mapped to it: ArtifactExistence.DamageDestroyDisableDismantle.Unspecified and ArtifactExistence.DamageDestroyDisableDismantle.Damage.

Each LDC type contains the name of the type, the LDC AnnotIndexID for that type, the list of LDC arguments for each type in events, and the list of PropBank rolesets that have been automatically found and then manually curated for each LDC type. Every member in the list of LDC arguments contains the LDC name, LDC argument output value, as well as LDC constraints for that argument. The DWD node also maps to a set of general arguments in the case of events. These general arguments consist of PropBank numbered arguments, followed by PropBank function tags, followed by VerbNet thematic roles, followed by PropBank descriptions. These were all manually curated. Each general argument is also constrained by a list of general constraints which is automatically extracted from VerbNet selectional preferences (which were manually mapped to Wikidata Qnodes by Anatole Gershman). Each LDC type also contains a DWD arg name field which maps that LDC argument to one of the general arguments, or, in some cases, None.

Development of Event-Event Relations. As expansion of the DWD continued, Colorado worked with LDC and the performers in weekly and bi-weekly calls to define potential Event-Event relations that LDC could annotate. These were ostensibly for KAIROS but were also expected to impact AIDA. The XPO subcommittee had an additional call with DARPA to clarify priorities for these relations. One result was a shift in emphasis away from causation and towards more explicit temporal ordering.

James Pustejovsky and Peter Anick added temporal relations to the v4 overlay, specifically for KAIROS. Two of these mapped to LDC types:

DWD_Q65560376 (partial coincidence) -> Overlap.Unspecified.Unspcified
DWD_Q6014822 (inclusion) -> Overlap.Containment.Unspecified

The other temporal relations (corresponding to two Wikidata Pnodes and five Qnodes) were mapped to a subset of interval relations in James Allen's temporal logic, shown in full in **Figure** 7 below.

[1] https://github.com/e-spaulding/xpo

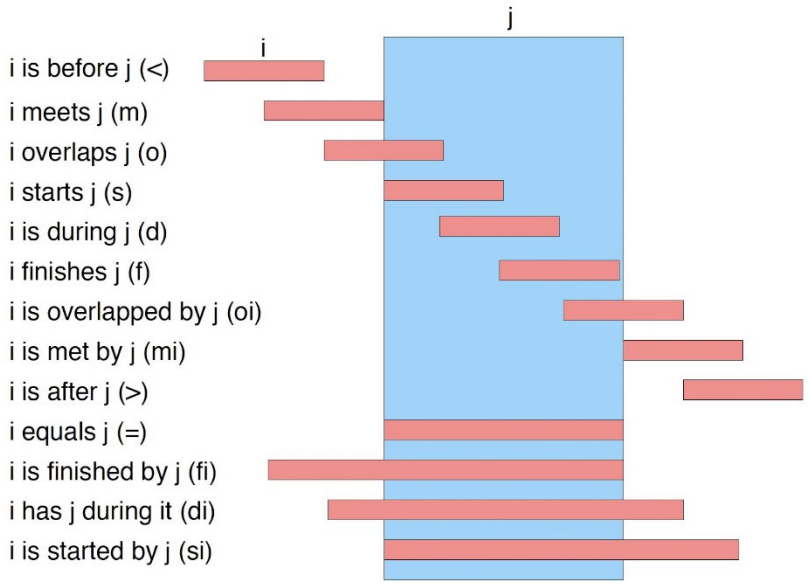

Figure 7. Allen Intervals

If I and j are time intervals, then the temporal relations in the overlay have the following interpretation in Allen's temporal logic:

DWD_Q79030196: I is before j
DWD_P156: I meets (is followed directly by) j
DWD_P155: I is met by (directly follows) j
DWD_Q65560376: I and j partially overlap (where either I overlaps j or I is overlapped by j)
DWD_Q6014822: I occurs within (during) j
DWD_Q79030284: I is after j
DWD_Q842346: I equals (spans the same time interval as) j

It should be noted that these relations are neutral with respect to causation.

Other DWD Relation Expansion. In 2022, over 100 new relations were added to the DWD. Anatole Gershman and Rosario Uceda-Sosa analyzed the existing relations and suggested new relations that were then discussed and approved by the XPO committee. The relations codify types of connections between two entities and establish constraints on the types of entities that can have a particular relation. The Wikidata constraints can be quite general, so it was decided to add PropBank roles that would clarify the types of entities involved in a relation. This innovation was suggested by Rosario and the current additions were developed at IBM. For example, for the new relation "host", defined as "an organism harboring another organism or organisms on or in itself", the two related entities were only constrained to be the Wikidata entity type "organism". By connecting this relation to the PropBank roleset infest.01, we could extract the appropriate PropBank roles that would further define the two organism entities. The first, A0, was mapped to

the infest.01 role A1-gol_thing_becoming_infected, and the second, A1, was mapped to A2_ppt_infectant. This example is illustrated in **Figure 8**, which highlights the new JSON fields for this information.


```
"DWD_P5009": {
    "type": "relation_type",
    "wd_node": "P5009",
    "name": "complies with",
    "wd_description": "the product or work complies with a certain norm or passes a test",
    "curated_by": "xpo",
    "arguments": [
        {
            "name": "A0",
            "constraints": [
                {
                    "name": "product",
                    "wd_node": "Q15401930"
                },
                {
                    "name": "intentional human activity",
                    "wd_node": "Q451967"
                }
            ],
            "pb_mapping": "A0_pag_complier"
        },
        {
            "name": "A1",
            "constraints": [
                {
                    "name": "rule",
                    "wd_node": "Q1151067"
                }
            ],
            "pb_mapping": "A1_ppt_rule"
        }
    ],
    "pb_roleset": "comply.01",
    "related_qnodes": [
        {
            "name": "compliance",
            "wd_node": "Q105476441"
        }
    ]
},
```


Figure 8. Example of Relation Entry in DWD Json

The initial set of DWD relations had been developed from LDC relations by identifying the most appropriate Wikidata Pnodes or Qnodes to represent the LDC relation. In v5.1, any relations that could only be mapped to Wikidata event Qnodes were moved to the events section of DWD. These included the Qnodes for “claim responsibility”, “belief”, and “doubt”. Two relations were remapped to new Wikidata Qnodes and moved to the events section: The LDC relation Evaluate.Sentiment.Negative had been mapped to “misfortune” but was changed to the more appropriate “antipathy”, and Evaluate.Sentiment.Positive had been mapped to “good” but was changed to “praise”.

Section 2.4 Similarity Metrics

When the program shifted from the LDC ontology to using the open-source Wikidata ontology, it was apparent that we needed a way to identify Qnodes that were similar to each other for evaluation purposes. We began by finding similar Qnodes for the types in the DWD overlay. Hans Chalupsky from ISI ran their similarity algorithm on DWD and found, for every given Qnode, the top 20 most similar Qnodes based on different similarity metrics. Combo3 (one of ISI similarity metrics that was a combination of other similarity metrics) was found to be the closest to human judgment. The generated Qnode lists were sent to 4 human judges to annotate every Qnode pair with their judgment: whether the two nodes are similar, and if so, are they near synonyms (therefore SS), or near neighbors (therefore NN). We identified at least one similar node for 121 DWD nodes for events and 189 DWD nodes for entities. These were added to the DWD json and were also used in the development of automatic similarity metrics.

ISI made several Qnode similarity metrics available in a toolkit[2] called Knowledge Graph Toolkit (KGTK). Karan Praharaj at University of Colorado and researchers at PNNL also developed similarity metrics to measure the similarity between Qnodes, but these were not part of the KGTK. Colorado is continuing to explore more advanced similarity metrics. The KGTK documentation gives these descriptions for each metric:

- Class: an ontology-based measure based on Jaccard Similarity of the respective superclass sets of two nodes inversely weighted by the instance counts of the classes
- JC: an ontology-based measure using an interpretation of the Jiang-Conrath ontological distance.
- ComplEx: an embedding-based measure using ComplEx graph embeddings computed over the Wikidata knowledge graph.
- TransE: an embedding-based measure using TransE graph embeddings computed over the Wikidata knowledge graph.
- Text: an embedding-based measure using text embeddings based on automatically generated sentences.
- TopSim: computes top-similar regions for each node by enumerating nearest neighbors from embeddings and from exploring the ontology which are then ranked using an aggregate of the above similarity computations. Once the top-similar regions are available, similarity is computed as a weighted average of the similarities between the two nodes and their top-5 similars.

These similarity metrics were used during the 2021 Hackathons in which "claim frames" were explored as a possible annotation scheme for identifying and co-referring claims made in news

[2] https://kgtk.isi.edu/

text using the Wikidata ontology. Specifically, the similarity metrics were used to create affinity matrices of Qnode pairs in an attempt to find coreferent claims. It quickly became clear that the biggest limitation of these metrics was their inability to be used on anything other than a single pair of Qnodes, especially in the context of a large, structured semantic representation such as a claim frame, which was composed of several different Qnodes and text values. We identified a few research questions, such as: how can we adapt these metrics to work on complex semantic structures, such as instantiated event descriptions? How should we compose similarity values for structures made up of several Qnodes? Should we continue to assume we can summarize the similarity between two events with a single scalar number, or is there a better way to represent the similarity?

Section 2.5 Ongoing ontology work

As a cross-program ontology, the XPO continues to develop under the DARPA KAIROS program, with additional supplemental funding. Three ontology-related tasks that began under AIDA are continuing: 1) improvement to the similarity metrics; 2) manually verifying the PropBank roleset mappings to DWD events and adding them as an appendix to Wikidata so that they are directly incorporated PropBank; 3) Continuing to expand the Relations. Two other tasks have been added: 4) a representation for causal relations; 5) a tool for semi-automatically inducing an ERE ontology with WD mappings from a new document set.

Similarity Metrics. In an effort to address the research questions from past similarity metrics work, Sijia Ge and Elizabeth Spaulding continue Colorado's work on similarity metrics, with a focus on instantiated event descriptions. LDC has provided some sample KAIROS Phase 1 TA2 assessment data with similarity judgements comparing system output to Gold Standard human events, as well as an internal script using the Hungarian algorithm to compute inter-annotator agreement. Colorado has already reformatted the data to ensure suitability for automatic evaluation of event similarity. The dataset has proved to be challenging to work with for similarity judgements because the system output portion of the data is quite noisy, automatically mapping from LDC events to DWD events can sometimes cause slight semantic changes, and similarity judgment classes are imbalanced. After running experiments on the reformatted data using features such as embeddings representations of the argument values and KGTK similarity scores, Colorado has found that the correlation of our automatic metrics and LDC's human judgements is relatively low. The next step is to modify LDC's inter-annotator agreement script, which we have just received at the end of this quarter, to compute similarity on the sample data. Elizabeth Spaulding is leading this effort.

PropBank roleset mappings. Because Wikidata provides no information about the participants or arguments of an event, the OWG added this information semi-automatically to an expansion of the original 132 DWD events that were based on LDC event types. Around 5,000 Qnodes were identified as events and linked to PropBank rolesets using rules created by Anatole Gershman. A Qnode was considered an "event" if it inherited from the top-level Wikidata event Q1190554 "occurrence." In Wikidata, each Qnode may have: a short sentence-length natural language description; possibly several aliases in addition to its label, sometimes in different languages; and a link to the associated Wikipedia articles in each language it is available. All of this information was used in matching PropBank rolesets to Qnodes. The mapping for the expansion proved to be quite noisy, with single PropBank rolesets being mapped to many Wikidata event nodes. Over 50 rolesets each map to 10 or more Wikidata nodes. A few rolesets map to over 100 Wikidata events. A PropBank roleset is designed to apply to all senses of a word that have the same argument structures; therefore, it is not surprising that a single roleset would map to multiple Wikidata events. For example, the PropBank roleset ill.01 maps to Wikidata nodes for general illness and to the dozens of Wikidata nodes representing different diseases.

Although not surprising, these one-to-many mappings presented a problem for performers. To mitigate this problem, the OWG set up a manual review of all of the PropBank-Wikidata mappings, starting with those PropBank rolesets that map to more than 10 Wikidata nodes. This work is expected to be completed following the end of the AIDA program, by continuing under KAIROS. The manual reviewers are working with an initial pass done by University of Illinois Urbana Champaign (UIUC) students. Only those most closely matching the general meaning and granularity of the word will be preserved. The annotators will also be checking for incorrect mappings and will be gathering example sentences of the rolesets being used. Several annotators were hired and trained for this task in August and September 2022, and they have continued their efforts in finding PropBank-Wikidata mappings into 2023.

Incorporation of PropBank roles in Wikidata. The OWG established a new task in 2022 to incorporate PropBank roles into Wikidata itself. This task continues under the DARPA KAIROS program. By moving these roles into Wikidata, performers will eventually be able to use Wikidata directly and repeated updating of the DWD would not be necessary. We anticipate that researchers outside of DARPA will make use of this as well. The OWG contacted the Wikidata organization to get their reaction to this proposal and to get their recommendations on how to proceed. In discussions with Wikidata, it was suggested we hire a Wikidata consultant—someone who is already a frequent contributor to Wikidata—to assist in adding information to Wikidata itself. It was also decided to first release this addition to Wikidata as an appendix. This would allow users to try the enhancement before fully altering the main Wikidata structure.

The OWG first devoted much of its attention to finding the right Wikidata structure for incorporating PropBank roles into Wikidata. The OWG formulated several ways for incorporating these roles into Wikidata, and eventually settled on a format that the Wikidata leadership team was happy with. It provides the desired information while conforming to the Wikidata ontological structure.

Results and Discussion

Section 6: Current Progress and Future Plans

Summary of Recent Accomplishments in AIDA:

1. Effective TA2 system for merging TA1's under original constraints.
2. Released annotation tool for Event Trigger Identification and Coreference resolution using Prodigy
3. Ontology Development
    a. DWD Overlay JSONs through V5.3 completed and released
    b. Structures for incorporating PropBank roles into Wikidata proposed and evaluated
    c. Continued reformatting LDC data for similarity metric evaluation
    d. Began manual assessment and improvement of PropBank-DWD mappings
4. Explorer for error analysis changed its focus from event coreference to claim-frame x-variable (entity) coreference and was migrated to Elasticsearch platform. Used to analyze evaluation results.

Task 3 - Future Plans for Ontology development (Under KAIROS)

1. Continue expanding similarity metric evaluation datasets and evaluate ontology-based algorithms and embedding based algorithms for partial credit for near neighbor types to DWD. Focus on instantiated event structures.
2. Expand the Relations
3. Add templates to all the events for improved accessibility of output
4. Continue manual evaluation and improvement of PropBank role mappings to DWD.
5. Begin adding PropBank roles to a Wikidata Appendix as Qnodes for events in the format agreed upon with Wikidata leadership.
6. Maintain the DWD overlay JSON file, which is now on release version 5.3.
7. Explore induction of salient ERE types in new domains with UIUC.

Table 10. Final Quarter Schedule

***Ramfis* Tasks Final Qtr**	**October**	**November**	**December**
All			
Task 1/2/4: PKB/CSR, Co-ref	Assessed cross-transformer mapping against additional metrics and datasets Report writing on face dictionary results	Incorporated pairwise contextual representations before coreference clusterings Used new annotations for training and testing the system Released updated TA2 system output to Texas UT	Explored cross-model mapping to multimodal representations Explored cross-transformer mapping to decoder-based or generative tasks and models Continued use of the Brandeis Explorer for evaluation purposes. Incorporated embeddings with search results from Elasticsearch entities, and assess impact on evaluation
Task 3: Ontology	Continued w/ similarity metrics; continued manual checking of PB role mappings; explored methods for incorporating PB into Wikidata	Continued similarity metrics and mapping relation entities to PB roles; explored methods for incorporating PB into Wikidata	Continued similarity metrics and mapping relation entities to PB roles; agreed upon methods for incorporating PB into Wikidata
Coreference Annotation Tool	Used Rich Event Descriptors to find coreference chains	Use Richd Event Descriptors to find coreference chains	Use Richd Event Descriptors to find coreference chains

Meetings and Presentations

One or two BiWeekly meetings with LDC to collaborate on ontology revisions
(Martha, Susan, Elizabeth, occasionally Kristin)

BiWeekly/Weekly XPO subcommittee calls
(Martha Palmer, Elizabeth Spaulding, Susan Brown, James Pustejovsky, Peter Anick)

Biweekly XPO calls
(Martha Palmer, Elizabeth Spaulding, Susan Brown, James Pustejovsky, Peter Anick)

Weekly RAMFIS team calls

COLING Keynote, Gyeongjiu, South Korea, Martha Palmer, *Deep Semantics,* Oct, 2022

Conclusion

Ramfis AIDA absorbed the time and energy of a large number of students and faculty at 3 institutions for 5 years. The students all learned a great deal about creating, handling, and merging large Knowledge Bases, as well as clustering of entities and events for coreference purposes, and error analysis, both within documents and across documents. We anticipate that the most enduring benefits will come from our results with processing and aligning vector representations for images and text and from the DWD Overlay for Wikidata developed under the auspices of the Cross-Program Ontology effort.

The embedding alignment work has given rise to several conjectures about the behavior of language models, and potentially machine learning models in general:

1) There exists a canonical language embedding space for Transformer-based models.
2) Each model is uncovering certain information in that space.
3) Mapping models together can augment one model's weaknesses with another's strengths.
4) If two embedding spaces preserve fundamentally the same information, a simple affine transformation can expose this property.

Ongoing work through the team's no-cost extension and after the completion of the AIDA program will include conducting further experiments on other metrics and datasets to test these conjectures. In particular, if conjecture #4, about information equivalence, is found to hold, this opens up new trajectories in diverse model fusion and augmentation using simple methods, e.g., across languages, between modalities, etc. One such experiment is discussed in the appendix.

The ontology-related work continues under KAIROS. There are three tasks that are continuing: 1) improvement to the similarity metrics; 2) manually verifying the PropBank roleset mappings to DWD events and adding them as an appendix to Wikidata so that they are directly incorporated; 3) Continuing to expand the Relations. Two other tasks have been added: 4) a representation for causal relations; 5) a tool for semi-automatically inducing a domain specific ERE ontology with WD mappings from a new document set. With respect to the similarity metrics in particular, we will continue to pursue these research questions: how can we adapt these metrics to work on complex semantic structures, such as instantiated event descriptions? How should we compose similarity values for structures made up of several Qnodes? Should we continue to assume we can summarize the similarity between two events with a single scalar number, or is there a better way to represent complex similarities?

Bibliography

Elasticsearch https://en.wikipedia.org/wiki/Elasticsearch Accessed: April 30, 2023

ECB+ corpus https://paperswithcode.com/paper/using-a-sledgehammer-to-crack-a-nut-lexical Accessed: April 30, 2023

GunViolence Corpus https://paperswithcode.com/dataset/gun-violence-corpus Accessed: April 30, 2023

Longformer
Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: The Long-Document Transformer, arXiv:2004.05150, 2020.

Proposition Bank http://propbank.github.io Accessed: April 30, 2023

RED
O'Gorman, T., K. Wright-Bettner, M. Palmer. (2021). The Richer Event Description Corpus for Event-Event Relations, Caselli, T., Hovy, E., Palmer, M., & Vossen, P. (Eds.), *Computational Analysis of Storylines: Making Sense of Events* (Studies in Natural Language Processing). Cambridge: Cambridge University Press. doi:10.1017/9781108854221

TimeML
Time Markup Language - https://en.wikipedia.org/wiki/TimeML Accessed: April 30, 2023

TransE
Antoine Bordes, Nicolas Usunier Alberto Garcia-Durán, Jason Weston, Oksana Yakhnenko, Translating embeddings for modeling multi-relational data, NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Pages 2787–2795, December 2013,

Wikidata https://www.wikidata.org/wiki/Wikidata:Main_Page Accessed: April 30, 2023

XPO GitHub https://github.com/e-spaulding/xpo Accessed: April 30, 2023

Appendix A: Publications

Ahmed, R., and J. H. Martin. (2021). Within-Document Coreference with BERT-Based Contextualized Representations, **https://arxiv.org/abs/2102.09600**, Accessed: April 30, 2023

Ahmed, S., A. Nath, J. H. Martin, and N. Krishnaswamy. (Submitted). *2*n Is Better Than n^2: Decomposing Event Coreference Resolution into Two Tractable Problems*.

Bonial, C., S. W. Brown, M. Palmer, G. Kazeminejad. The Rich Event Ontology: Ontological Hub for Event Representations. (2021). In Caselli, T., Hovy, E., Palmer, M., & Vossen, P. (Eds.), *Computational Analysis of Storylines: Making Sense of Events* (Studies in Natural Language Processing). Cambridge: Cambridge University Press. doi:10.1017/9781108854221

Brown, S., Bonn, J., Kazeminejad, G., Zaenen, A., Pustejovsky, J. , Palmer, M. (2022). Semantic Representations for NLP Using VerbNet and the Generative Lexicon. *Frontiers in Artificial Intelligence,* April 14, 2022, doi: 10.3389/frai.2022.821697 Accessed: April 30, 2023

Caselli, T., Hovy, E., Palmer, M., & Vossen, P. (Eds.). (2021). Computational Analysis of Storylines: Making Sense of Events (Studies in Natural Language Processing). Cambridge: Cambridge University Press. doi:10.1017/9781108854221

Kazeminejad, G., M. Palmer, T. Li, V. Srikumar. (2021). Automatic Entity State Annotation Using the VerbNet Semantic Parser. In *Proc. of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, held with EMNLP 2021*, Nov 11, 2021.

Kazeminejad, G., Palmer, M., Brown, S. W., & Pustejovsky, J. (2022). Componential Analysis of English Verbs. *Frontiers in Artificial Intelligence*, *5*, May 30. doi: 10.3389/frai.2022.780385 Accessed: April 30, 2023

Martin, M., C. Mauceri, M. Palmer, & C. Heckman. (2020). Leveraging Non-Specialists for Accurate and Time Efficient AMR Annotation. In *Proceedings of the LREC 2020 Workshop on" Citizen Linguistics in Language Resource Development",* May.

Mauceri, C., and C. Heckman. (2020). Robust semantic segmentation in dark environments using RGB-D images. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Mauceri, C., M. Palmer, and C. Heckman. (2019). SUN-Spot: An RGB-D Dataset with Spatial Referring Expressions. In *International Conference on Computer Vision Workshop on Closing the Loop Between Vision and Language*.

McNeely-White, D., B. Sattelberg, N. Blanchard and R. Beveridge. (2022). Canonical Face Embeddings. In *IEEE Transactions on Biometrics, Behavior, and Identity Science,* March.

McNeely-White, D., B. Sattelberg, N. Blanchard, and R. Beveridge. Common CNN-based Face Embedding Spaces are (Almost) Equivalent. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021.

McNeely-White, D., B. Sattelberg, N. Blanchard, R. Beveridge. (2020). Exploring the Interchangeability of CNN Embedding Spaces. *arXiv preprint,* arXiv:2010.02323

McNeely-White, D., Beveridge, J. R., & Draper, B. A. (2020). Inception and ResNet Features Are (Almost) Equivalent. *Cognitive Systems Research*, 59, 312-318. (longer version of BICA 2019 paper).

McNeely-White, D., J. R. Beveridge, and B. A. Draper. (2019). Inception & ResNet: Same Training, Same Features, *2019 Annual International Conference on Biologically Inspired Cognitive Architectures, the 10th Annual Meeting of BICA Society*: Seattle, WA, USA, Springer-Verlag, A. Samsonovich (ed).

McNeely-White, D., Sattelberg, B., Blanchard, N., & Beveridge, R. (2022). Canonical Face Embeddings. *IEEE Transactions on Biometrics, Behavior, and Identity Science,* 4(2), 197-209.

Nath, A. Linear Mappings: Encoding semantic information from transformer models for Cognate Detection and Coreference Resolution (CSU Master's Thesis October 2022): Colorado State University, Fort Collins, CO.

Nath, A., R. Ghosh, and N. Krishnaswamy. Phonetic, Semantic, and Articulatory Features in Assamese-Bengali Cognate Detection. (2022). Ninth Workshop on NLP for Similar Languages, Varieties, and Dialects: Gyeongju, Republic of Korea. At COLING.

Nath, A., S. Mannan, and N. Krishnaswamy. (Under review). *AxomiyaBERTa: A Phonologically-aware Transformer Model for Assamese*.

O'Gorman, T., K. Wright-Bettner, M. Palmer. (2021). The Richer Event Description Corpus for Event-Event Relations, Caselli, T., Hovy, E., Palmer, M., & Vossen, P. (Eds.), *Computational Analysis of Storylines: Making Sense of Events* (Studies in Natural Language Processing). Cambridge: Cambridge University Press. doi:10.1017/9781108854221

Pradhan, S., J. Bonn, S. Myers, K. Conger, T. O'Gorman, J. Gung, M. Palmer. (2022). PropBank Comes of Age - PropBank Comes of Age—Larger, Smarter, and More Diverse, **SEM* 2022. (July, 2022). https://aclanthology.org/2022.starsem-1.24.pdf Accessed: April 30, 2023

Wang, Q., M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, A. Chauhan, Y. Guan, B. Li, R. Li, X. Song, Y. R. Fung, H. Ji, J. Han, S. Chang, J. Pustejovsky, D. Liem, A. Elsayed, M. Palmer, J. Rah, C. Voss, C. Schneider and B. Onyshkevych (2021). 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 [Best Demo Paper Award].*

Appendix B: Face Dictionary details

This section displays results from alternate (less successful) methods of performing the face dictionary evaluation.

Method 1: Use all source (BBN/RPI) embeddings for identity *I* and calculate cosine similarity relative to all target (CSU) embeddings for identity *I* (highest cosine similarity used for clustering).

Table B-1. Method 1 Results of Cross-embedding Space Mappings on Binary Classification Task

	Accuracy	Precision	Recall	F1
RPI → CSU	.96	.79	.20	.23
BBN → CSU	.89	.84	.27	.40

Method 2: Use all source (BBN/RPI) embeddings for identity *I* and calculate cosine similarity relative to one random target (CSU) embedding for identity *I.*

Table B-2. Method 2 Results of Cross-embedding Space Mappings on Binary Classification Task

	Accuracy	Precision	Recall	F1
RPI → CSU	.95	.67	.38	.26
BBN → CSU	.92	.75	.50	.59

Method 3: Use all source (BBN/RPI) embeddings for identity *I* and calculate cosine similarity relative to the mean target (CSU) embedding for identity *I.*

Table B-3. Method 2 Results of Cross-embedding Space Mappings on Binary Classification Task

	Accuracy	Precision	Recall	F1
RPI → CSU	.99	.45	.47	.46
BBN → CSU	.94	1.0	.79	.88

CSU also applied the cross-Transformer mapping technique to a cognate detection task between Assamese and Bengali, two languages of eastern India. While this application used no AIDA data, it does demonstrate the feasibility of this technique to other tasks besides coreference resolution and opens the door to applications on AIDA-relevant data involving similar languages, such as Ukrainian and Russian. The paper on this result appeared at the Workshop on NLP for Similar Languages, Varieties, and Dialects, at COLING 2022 in Gyeongju, Republic of Korea (Nath et al., 2022). NLP in this context refers to natural language processing.

Abhijnan defended his Master's thesis on linear mapping applications to the cross-document coreference and cognate detection tasks on October 20, 2022 (Nath, 2022).

Nath et al. (under review) extends the work from Nath et al. (2022) and presents an updated version of the Assamese-language Transformer model developed therein and evaluates it on a number of existing benchmarks, as well as a novel event coreference benchmark in this language. Notably, we identified that the pretrained embedding space of this language model, due to a smaller data size, is highly anisotropic, which affects downstream performance. Therefore, we optimized the embedding space for performance on long-context tasks like question answering using a combined loss function given by:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log\left(1 - \hat{Y}_i\right)\right)$$

$$L_{COS}(x, y) = \begin{cases} 1 - \cos\left(x_1, x_2\right), & \text{if } y = 1 \\ \max\left(0, \cos\left(x_1, x_2\right) - \mathrm{m}\right), & \text{if } y = -1 \end{cases}$$

$$L_{COMB} = \alpha L_{BCE} + L_{COS}(x, y)$$

Equation B-1. Combined Binary Cross-Entropy and Cosine Embedding Loss" (*LBCE* is Binary Cross-Entropy loss, *LCOS* is Cosine Embedding loss and the results are summed as *LCOMB*)

where the first function is binary cross-entropy loss and the second function takes the 128D output of a cosine embedding layer computation over contextual and candidate representations along with a [CLS]. The combined loss value is fed into an auxiliary discriminator that considers only the candidate and [CLS] representation. The results of this operation show that the embedding space after this dispersal operation during training more closely resembles the embedding space of a large multilingual language model such as XLM-R even though the data size is orders of magnitude smaller.

List of Symbols, Abbreviations and Acronyms

ACL Association for Computational Linguistics
AIDA Active Interpretation of Disparate Alternatives
ALBERT A Light BERT
AMR Abstract Meaning Representation
AO AIDA ontology
BBN Raytheon BBN Technologies
BCUB A coreference evaluation measure
BERT Bidirectional Encoder Representations from Transformers
CDLM Cross-Document Language Model
CLEAR-AMR Computational Language and Education Research-Abstract Meaning Representation
CMU Carnegie Mellon University
CNN convolutional neural net
CoNLL Conference on Computational Natural Language Learning
CSR Common Semantic Representation
CSU Colorado State University
CU University of Colorado
DARPA Defense Advanced Research Project Agency
DEFT Deep Exploration and Filtering of Text
DWD DARPA Wikidata
ECB+ Event Coreference Bank + (sentences annotated with event coreferences)
ECR event coreference resolution
ERE Entities, Relations and Events
GAIA Generating Alternative Interpretations for Analysis
GVC Gun Violence Corpus
IOU intersection over union
ISI Information Sciences Institute
JHU Johns Hopkins University
KAIROS Knowledge-directed Artificial Intelligence Reasoning Over Schemas
KB knowledge base
KGTK Knowledge Graph Toolkit
LDC Linguistic Data Consortium
LFW Labeled Faces in the Wild dataset
Longformer – The Long Document Transformer
LO LDC ontology
M19 Month 19
MAA Modeling Adversarial Activity
MTCNN Multi-Task Cascaded Convolutional Network
MUC Message Understanding Conference
NLP natural language processing

NN near neighbor
OWG Ontology Working Group
PKB Probabilistic Knowledge Base
PropBank – The Proposition Bank (annotated predicate argument structures)
REO Reference Event Ontology
RPI Rensselaer Polytechnic Institute
RoBERTa Robustly Optimized BERT-Pretraining Approach
SOTA state of the art
SS synonym set
TA1 Technical Area 1
TA2 Technical Area 2
TransE – Transferring Embeddings
UI user interface
UIUC University of Illinois Urbana Champaign
XPO Cross-Program Ontology