AFRL-RI-RS-TR-2023-039

MULTIMODAL DECOMPOSITIONAL SEMANTICS

JOHNS HOPKINS UNIVERSITY

MARCH 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-2023-039 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
ALEKSEY V. PANASYUK
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief
Intelligence Systems Division
Information Directorate

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

REPORT DOCUMENTATION PAGE

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

4. TITLE AND SUBTITLE

MULTIMODAL DECOMPOSITIONAL SEMANTICS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-18-2-0015	N/A	62303E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2FB

6. AUTHOR(S)

Benjamin Van Durme, Aaron White, Kyle Rawlins

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

Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218-2683

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505	AFRL/ RI	AFRL-RI-RS-TR-2023-039

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

Johns Hopkins University, partnering with the University of Rochester, pursued research and development of analytics in support of a larger framework for knowledge-driven hypothesis testing. We leveraged our expertise in dataset creation for decompositional semantics, to develop new datasets specifically geared towards the extraction problems of the DARPA's Active Interpretation of Disparate Alternatives (AIDA) program (specifically in event extraction and coreference resolution). Notable examples of results from our team include: the construction of RAMs, the first publicly available multi-sentence event extraction dataset; the development of state of the art multilingual coreference models, including an online variant that handled long documents with a fixed amount of memory, as well as a new multilingual dataset that focused on multi-person dialogues; a new supervised approaches to cross-lingual alignment, supporting the automatic creation of training data through projecting from English to less-resourced languages; a framework for sentence-level paraphrasing and data augmentation; collaborations on the emerging science of "probing" neural language models; and the development of new decompositional resources and analysis across a number of new linguistic dimensions.

In the initial phase of the program, we provide analytic outputs as part of the program-wide evaluation run by NIST (focus on multilingual text and speech). In the second phase we provided fewer components, focusing exclusively on text. In the third phase our focus was on data annotation under a newly proposed "claim frame" task, which exercised our background in crowd-sourcing rich linguistic annotations.

15. SUBJECT TERMS

strategic understanding, mixed modality, mixed domain, multiple modalities, multi-hypothesis semantic engine

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

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
ALEKSEY V. PANASYUK	315-330-3976

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

Johns Hopkins University, partnering with the University of Rochester, pursued research and development of analytics in support of a larger framework for knowledge-driven hypothesis testing. Performers in the program were charged with collaborating on a data processing framework that began with raw unstructured content (text, images, video with audio), converted these into knowledge statements under a shared ontology, merged the results across information sources into a single knowledge graph, then performed inference on this graph to propose additional information that could be derived from what was directly observed. We, the JHU team, were focused on the first step of this process. We proposed a framework that would handle all required input modalities, but were selected to focus on multilingual text and speech (no computer vision). We participated as a stand-alone team in the initial phase of the program, providing analytic outputs as part of the program-wide evaluation run by NIST. In the second phase we provided fewer components, focusing exclusively on text. These components were shared with BBN during program evaluation. In the third phase our main focus was on data annotation under a newly proposed "claim frame" task, which exercised our background in crowd-sourcing rich linguistic annotations.

We proposed a focus on decompositional semantics: a fine-grain multi-valued handling of meaning. Owing to the program's ambitious shared goals and concentration on a single program-wide ontology, we focused between new state of the art language analytic technologies that targeted the shared tasks, and the development of new decompositional resources targeting aspects outside the program ontology. Notable examples of results from our team include: the construction of RAMs, the first publicly available multi-sentence event extraction dataset; the development of state of the art multilingual coreference models, including an online variant that handled long documents with a fixed amount of memory, as well as a new multilingual dataset that focused on multi-person dialogues; a new supervised approaches to cross-lingual alignment, supporting the automatic creation of training data through projecting from English to less-resourced languages; a framework for sentence-level paraphrasing and data augmentation; collaborations on the emerging science of "probing" neural language models; and the development of new decompositional resources and analysis across a number of new linguistic dimensions.

2.0 INTRODUCTION

The state of the art in language analytics has advanced rapidly in the last ten years. DARPA AIDA occurred as neural models for text analysis were rapidly breaking new ground in accuracy. This began with improvements to the statistical NLP pipelines that came before, and then analytics began to be trained "end-to-end": models that no longer required part of speech tagging, syn-tactic parsing, and so on, to power a holistic language understanding process. Rather, models were trained directly on the target outputs with the presumption that sufficient linguistic features were captured inside the parameters of pretrained language models. AIDA ended just as the community began considering another advance in approach, through in-context learning (prompt-hacking) of very large-scale language models (LMs) like GPT3, and a general focus on generative LMs.

JHU with its partner the University of Rochester made contributions to the state of the art in neural models for information extraction, as well as in the new science of probing large language models. We leveraged our expertise in dataset creation for decompositional semantics, to develop new datasets specifically geared towards the extraction problems of the AIDA program (specifically in event extraction and coreference resolution). We developed new resources in decompositional semantics, and in the final phase of the program we were dedicated to the new initiative on understanding how to annotate claims of fact in text (so-called "claim frames").

With regards to program evaluations, we diligently applied ourselves to the ever changing and ambitious requirements of NIST and their partners. We found that we regularly were competitive or superior to other performers in the components that we focused on, especially in multilingual coreference resolution. As the pipeline requirements and knowledge-driven workflow were being developed on the fly during the program, it was observed by everyone to be nontrivial to coordinate across performers, especially in time-sensitive contexts around an evaluation. This unfortunately led to JHU's contributions to the pipeline often being limited: strong analytic com-ponents that were not always fully exercised in the larger prototype framework. Separately from the formal evaluations we built a stand-alone analytic framework that we open-sourced and re-leased to the community. This "LOME" package (Large Ontology Multilingual Extraction) saw adoption in AIDA-related applications outside the program and targeting the mission needs envisioned by the program. Our efforts in AIDA helped lead in part to participation in additional related programs, including DARPA KAIROS and IARPA BETTER, all with a primary focus on the development of enhanced language technologies.

In the following we focus on the key methods and results that emerged from our participation in AIDA. References are provided where available to scientific articles that provide further detail. As stated in our summary, notable examples of our results include: the construction of RAMS, the first publicly available multi-sentence event extraction dataset; the development of state of the art multilingual coreference models, including an online variant that handled long documents with a fixed amount of memory, as well as a new multilingual dataset that focused on multi-person dialogues; a new supervised approaches to cross-lingual alignment, supporting the automatic creation of training data through projecting from English to less-resourced languages; a framework for sentence-level paraphrasing and data augmentation; collaborations on the emerging science of "probing" neural language models; and the development of new decompositional resources and analysis across a number of new linguistic dimensions.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 INFORMATION EXTRACTION

We made contributions in Coreference Resolution, Entity Typing, and Event Extraction.

3.1.1 Coreference Resolution

Coreference resolution is the task of recognizing that different spans of text in a document are used by the author to refer to the same entity. E.g., "***Ben*** *wrote* ***his*** *report".* Recognizing coreference is a critical component in knowledge graph construction. E.g., in "***Ben*** *wrote* ***his*** *report.* ***He*** *works at JHU"* we might wish to extract the fact that **Ben worksFor JHU**. Coreference resolution accuracy increased during the time of AIDA powered largely by new feature extraction technology such as BERT. LMs such as BERT take input text and convert the tokens into contextualized embeddings that lend themselves to performing information extraction tasks. We authored a survey article on encoders such as BERT, organizing the next steps to improving these models [1].

While encoders such as BERT are powerful when used for sentence-level tasks, such as recognizing an event or relation, e.g.,*"He [works for] JHU"*, there is a length limit to how much input these models can take at once: long documents are not able to be contextually encoder in one go. Further, the Transformers that underly BERT are quadratic in the sequence they process. This makes them computationally expensive for processing large numbers of long inputs. Prior approaches to neural coreference resolution would break long documents into smaller contiguous regions that could fit into an encoder such as BERT and encode them in isolation. Then when performing coreference, they would consider all entity mentions in a document jointly, which leads to another quadratic cost, now in the number of entity mentions in an arbitrary document. We developed a new approach that encoded text incrementally from the start to the end of a document, like a human would write and read the information, and as we considered each new portion of a document we would incrementally update clusters of entity mentions; each cluster representing a discovered entity [2]. This incremental approach had minimal impact on performance compared to previous global solutions, while being linear rather than quadratic in the length of the document. This has significant impact to downstream analytic processing frameworks responsible for large scale processing of collections.

We investigated how well coreference models generalized to new domains [3,4], leading us to develop methods for rapidly adapt coreference through active learning [5]. Finally, we performed an investigation on using a sequence to sequence approach to finding entity mentions, inventing a new procedure we called CopyNext [6]. This mechanism supplemented a standard sequence generation model with an operation that explicitly copied a token of input to the output, and then another operation for continuing the operation to copy the next token. In this way, a sequence to sequence model is provided inductive bias to recognize and copy spans of input to the output, a central property of information extraction.

Finally, we developed new models and a large new dataset for multilingual coreference resolution, with a novel focus on multiparty conversations [7].

3.1.2 Entity typing

Entity typing is the task of assigning an entity mention a label according to an ontology. E.g., "***Ben (PERSON)*** *wrote his* ***report (ARTIFACT)****."* The AIDA program ontology for entities and events was organized using a hierarchical structure.

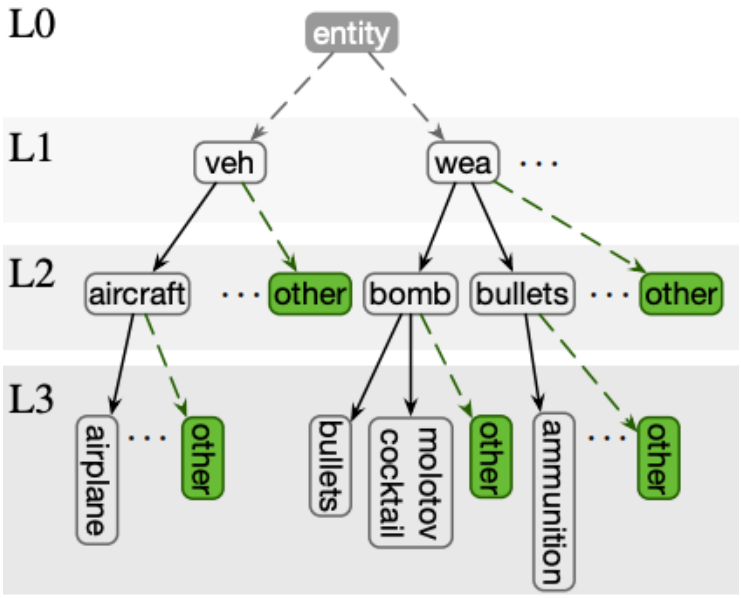

Figure 1: Example of a subset of the AIDA Phase 1 Entity Ontology.

Under a hierarchical ontology entities are assigned labels under a tree structure, including coarse conceptual classes at the higher levels, and finer grain distinctions in lower levels. E.g., *"**Ben (ANIMAL/PERSON/PROFESSOR)** wrote ..."*. Prior work in predicting labels under a hierarchical ontology had largely focused on either predicting labels at each level independently with post hoc constraints ensuring a valid total label, or in treating each full path in a hierarchy as a distinct label. This independence assumption is problematic when there is limited training data: examples that are close in an ontology, e.g., different kinds of PERSON, may appear in similar distributional contexts. When treated as independent then a model must learn a representation for each label based only on the examples for that given label. We developed an approach to neural representation learning of the nodes in a hierarchy such that training examples that were labeled with similar paths would share more information under the model than those examples which were more ontologically distinct. This approach was motivated by the fact that under Phase 1 of AIDA there were very few training examples made available to performers for the new ontology. Another contribution of our approach included inference time recognition of the hierarchy, that predictions began at the root of the ontology and incrementally selected labels one level at a time, conditioned on the current position in the hierarchy. This effort led to state of the art performance on various hierarchical typing datasets [7].

3.1.3 Event extraction

Recognizing events and their arguments is traditionally considered a sentence-level information extraction task. E.g., *"[Ben] **wrote** his [report]"*. We developed a number of new approaches to this task under the course of AIDA.

A regular motivating concern in AIDA was the lack of training data even while new additions were being made to a novel ontology that required analytic support. We recognized that even while there were few or zero examples annotated for some parts of the ontology, there were definitions provided for each of these concept classes, meant for consumption by humans. We explored an

approach that could leverage this kind of information directly to a model, through the use of what we called *bleached* examples.

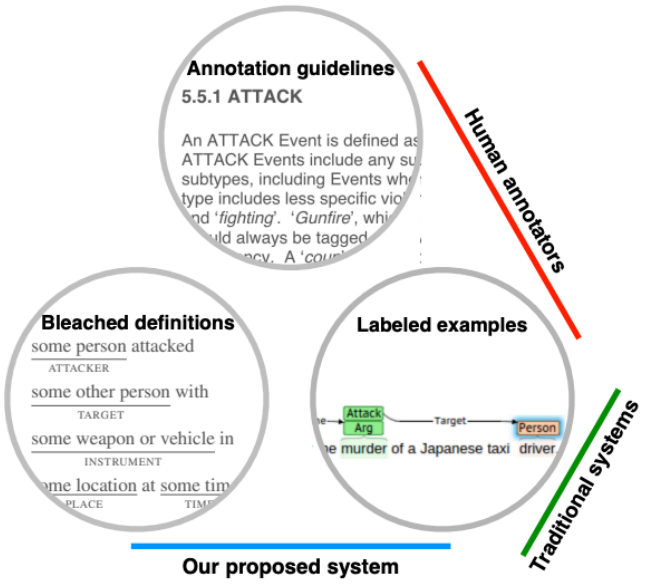

Figure 2: Comparison of data sources for human annotators, traditional information ex-traction sys-tems, and our approach.

In a bleached example, we might define the ATTACK event as:

[some person]:ATTACKER attacked [some other person]:TARGET with [some weapon or vehicle]:INSTRUMENT at [some time]:TIME at [some location]:LOCATION.

Our intuition was that an encoder such as BERT would have significant understanding of a sentence such as: *"Some person attached some other person with some weapon or vehicle …"*, where contextually the spans corresponding to each ontological argument slot for the given event would be "understood" by the model. Even if we have limited training examples for an event, we might still be able to glean information about how to represent each slot, through the contextual encoding of just a single bleached example per event. We constructed a model under this assumption [8], where argument prediction was performed by incrementally replacing the bleached arguments, e.g., "someone", with text spans drawn from an input text. In this way we could construct a knowledge statement in English that could deterministically be mapped into the desired knowledge graph format.

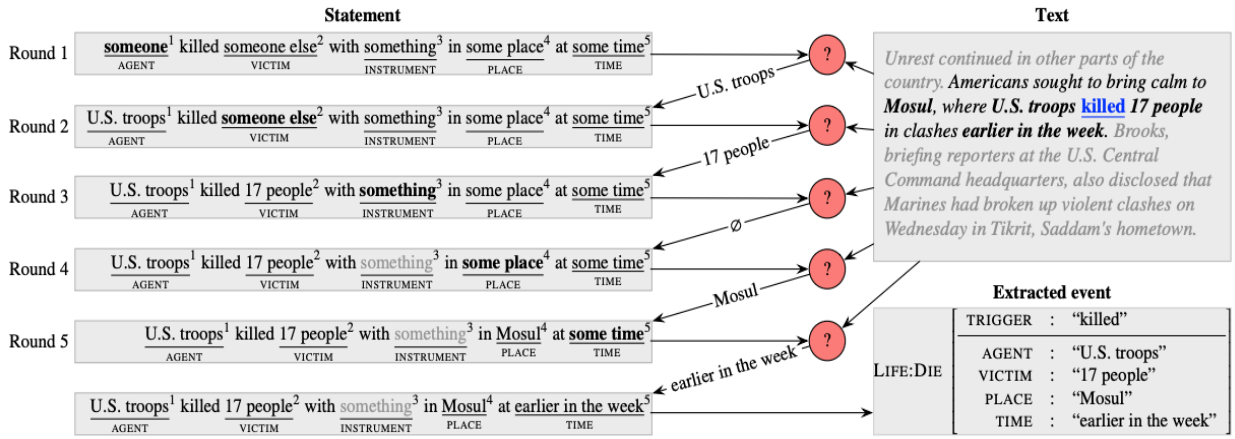

Figure 3: An example of our approach on a sentence for a LIFE:DIE event.

AIDA – in comparison to most previous efforts on document understanding – required performers to extract the arguments for events across multiple sentences. For example, in, *"Ben wrote. The report was eventually complete."*, this describes a WRITING event with arguments, "Ben" and "The report." Limited prior work and data existed for this version of event extraction. This motivated us to create a dataset following the AIDA Phase 1 ontology through crowdsourcing, which we called RAMS (Roles Across Multiple Sentences) [10].

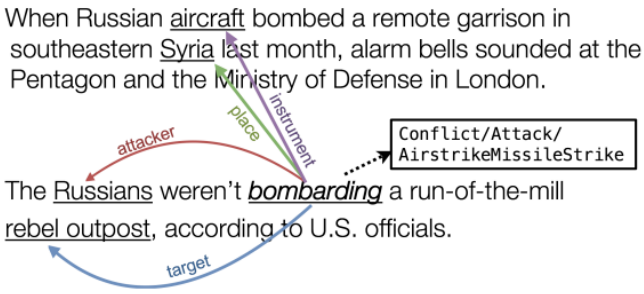

Figure 4: A passage annotated for an event's trigger, type, and arguments.

In RAMS we carefully enumerated examples of lexical triggers for each event type under the AIDA ontology, then selected five contiguous sentences from documents pulled from online materials that matched these terms. Annotators were asked to verify that matching text did refer to an event, that the event was being described as having happened, and that the type of event it was referring to matched the definition under AIDA. Based on locating valid event references, we then asked annotators to highlight spans of text corresponding to arguments that were within two sentences before and after the event mention. This dataset was the first large scale resource of its kind released to the public and has subsequently had significant influence in the information extraction community. In our own work for RAMS we developed a model for performing multisentence argument event linking that drew on our modeling experiences for coreference. In this approach, for each event trigger detected in a text, we implicitly introduced a bleached argument for the event

of each type into the discourse context. These arguments were then used as entity mentions that needed to be "coreferenced" to other mentions in the local context [10].

In subsequent work we developed an improved model that relied on a three step process: (1) event and entity mentions were first located independently; (2) for each event span, all entity mentions were concatenated to that event span, with their respective span-level embeddings derived from an underlying encoder such as BERT; then (3) these span-level representations were re-encoded under a new Transformer model developed just for this task. In this way, the representations of each entity mention were modified based on a given event-mention query. Argument prediction could then be performed atop these modified representations [11].

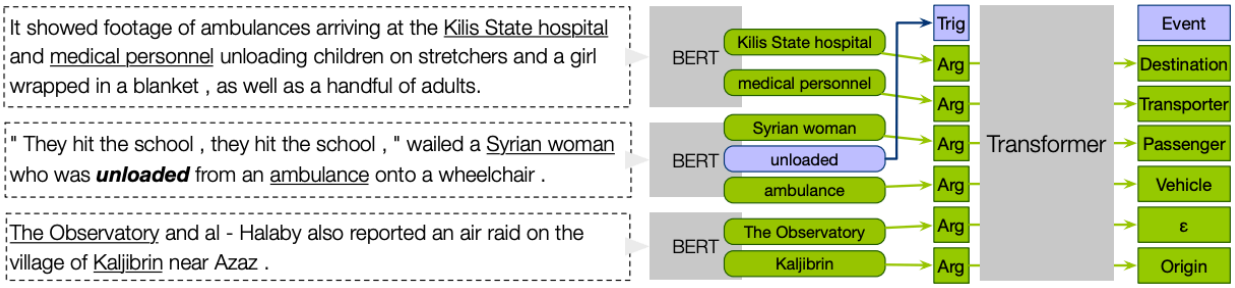

Figure 5: An example of performing joint argument linking in RAMS.

Our models developed for information extraction under AIDA were combined into a single publicly available system we called LOME: Large Ontology Multilingual Extraction [12].

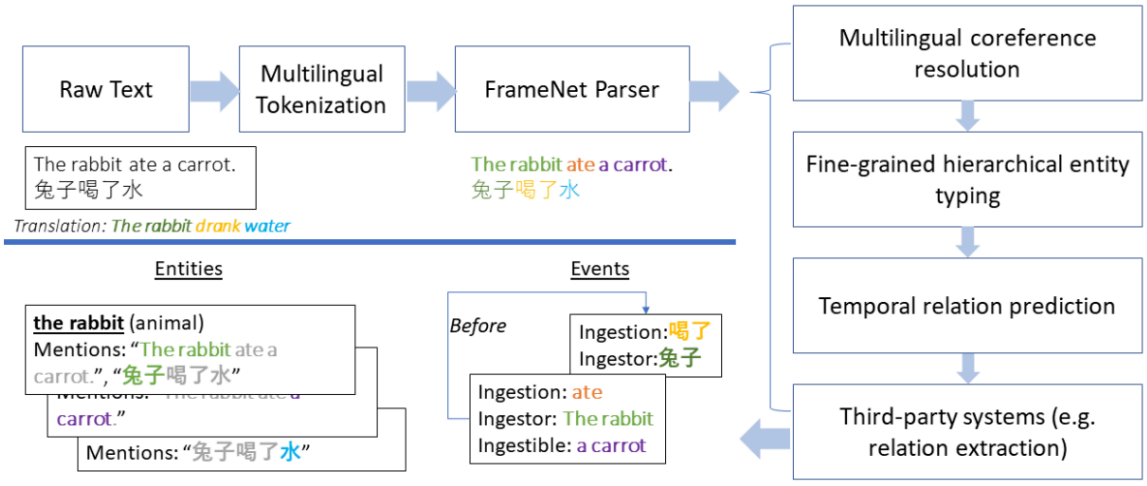

Figure 6: Illustration of the LOME system.

We describe our work in *decompositional* event understanding in Section 3.5.

3.2 METHODS FOR DATA AUGMENTATION

AIDA analytics needed to support an ontology that grew regularly over the program, without a large amount of annotated examples that would drive performance. This drove us to develop new methods for (semi-)automatic generation of additional training examples.

3.2.1 Supervised Alignment and Projection

The majority of annotated IE training data is in English. When building IE systems for non-English languages we would prefer to not have to annotate again for each such language. As Machine Translation quality improves it is a reasonable question to ask whether we can translate training data and use that for training IE systems. In such a process two components are needed: (1) a translation model; and (2) a method for *projecting* the annotations from the source language (usually English) to the translated target side language. The majority of prior work in bitext alignment was unsupervised and meant for aligning all parts of an input to all parts of an output. In IE, usually only certain spans of are interest, e.g., those referring to entities or events. We explored a solution to alignment that was developed for IE data projection, and that assumed access to some amount of time from a bilingual professional that could provide example annotations for alignment [13]. We observed that even modest amounts of supervision provided to our new alignment model could go a long way in driving accurate projections. We employed this supervised alignment model in creating supplemental training data during AIDA, and our work led to interest in the community to continue developing on this idea.

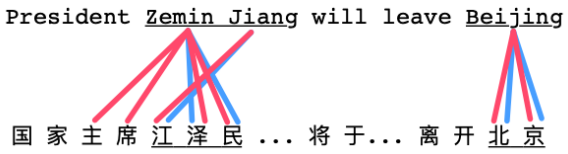

Figure 7: Projecting named entity annotations from English to Chinese.

3.2.2 Paraphrastic Augmentation

The ParaBank line of research explored ways to generate lexically and syntactically diverse paraphrases, with constraints. When combined with our work in annotation projection through alignment this allowed for rapid construction supplemental IE training data.

In our initial work we developed the resource ParaBank, a large-scale English paraphrase dataset that surpassed prior work in both quantity and quality [14]. We trained a Czech-English neural machine translation system to generate novel paraphrases of English reference sentences. By adding lexical constraints to the decoding procedure we were able to produce multiple high-quality sentential paraphrases per source sentence, yielding an English paraphrase resource with more than 4 billion generated tokens and exhibiting significant lexical diversity. Using human judgements we demonstrated that ParaBank's paraphrases improve on prior work on both semantic similarity and fluency. This resource allowed us to train a monolingual sentence rewriter that supported lexically constrained decoding.

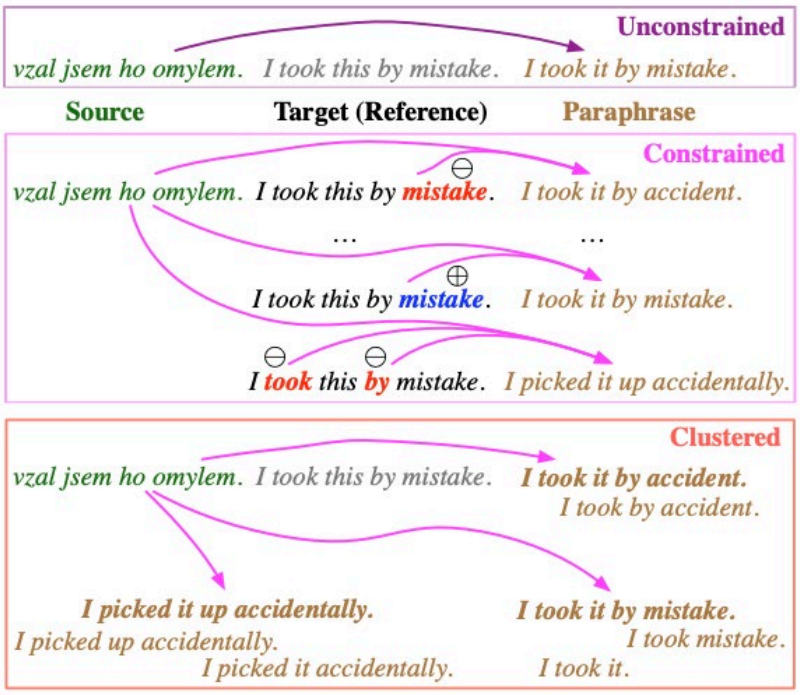

Figure 8: ParaBank explored constrained decoding and then later semantic clustering to ensure greater lexical diversity in paraphrase generation.

We then developed *vectorized dynamic beam allocation*, which extended lexically-constrained decoding to work with batching, leading to a five-fold improvement in throughput when working with positive constraints. Faster decoding enabled faster exploration of constraint strategies: we illustrated this via data augmentation experiments with our monolingual rewriter applied to the tasks of natural language inference, question answering and machine translation, showing improvements in all three tasks as a result of this augmentation [15].

We also explored an approach to diversity in paraphrase construction through sampling many outputs from a decoder and clustering the results, taking cluster centroids as representative of a set of lexically similar options [16]. The result of this effort was ParaBank 2, a larger and more diverse version of ParaBank 1.0.

3.2.3 Combined Projection and Paraphrasing

We combined our efforts in supervised alignment and paraphrasing into a single exploration that was aimed at workflows such as employed in creating data under AIDA. In this work, we began with a few examples of a given event we wished to create data for. We then automatically paraphrased these examples with lexical constraints to ensure specifically that the event trigger from the original example was replaced with a new way to reference the event. Once the paraphrase was generated, we then employed a new neural alignment model to project the event annotation from the initial training example to the appropriate paraphrased span in the newly created example. Through rigorous experiments with humans in the loop, we illustrated how one could rapidly and efficiently expand an IE dataset such as FrameNet, through interacting with model suggestions on

annotated paraphrases of the human's original example [17]. This work also led to a state of the art model for FrameNet prediction, which was included in our released LOME system.

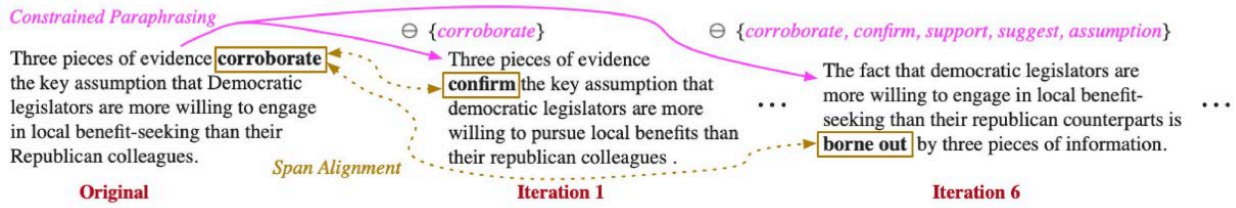

Figure 9: An illustration of our iterated paraphrastic augmentation workflow.

3.3 UNDERSTANDING LANGUAGE MODELS

Neural language models increased in size and performance rapidly during the timespan of the AIDA program. As these models became the workhorse for analytic development it was important to scientifically understand their strengths and weaknesses. This style of research has become known as "probing" the capabilities of an LM, and we were involved in a number of influential studies during this time.

3.3.1 Gender Bias in Coreference Resolution

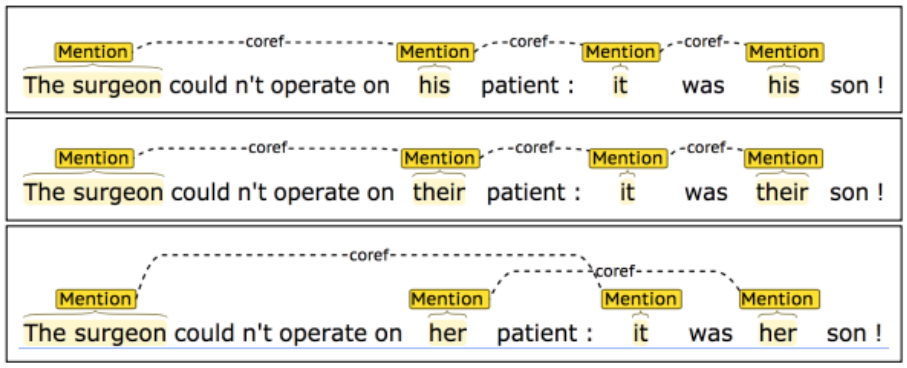

Figure 10: Stanford's CoreNLP rule-based coreference system showing a bias to not associating surgeons with females.

As discussed earlier, coreference resolution is a key component in mapping text into knowledge statements. Coreference models - like any text analytic - are improved through leveraging of large amounts of corpus data. Unfortunately humans possess a number of biases in their worldview which can be reflected in language, that language then used in bulk to improve analytics. With this in mind, we asked whether popular coreference models expressed stereotypical gender biases [18]. As illustrated in the following example, we carefully constructed a series of examples where the semantic context of a sentence strongly preferred a particular coreference interpretation and then we varied the gender of the pronoun that needed to be resolved. This analysis was later adapted into the SuperGLUE benchmark and is considered a key resource in measuring language model

bias. For example, in the article introducing the GPT3 language model those authors noted that our gender bias probing set was one of the few things that their model performed significantly poorly on.

In a followup to exploring gender bias in coreference, the primary author of that work led an effort on understanding the social biases in sentence encoders [19].

3.3.2 Building Natural Language Inference Resources for Probing

Natural Language Inference or Recognizing Textual Entailment is the task of classifying whether a *hypothesis* sentence follows from a *premise* sentence. Since its inception it has been used as an approach to evaluate the underlying capabilities of language understanding tasks, with examples from areas including information extraction rewritten to be in this premise/hypothesis form.

Event Factuality	▶ Find him before he finds the dog food The finding did not happen	✓
	▶ I'll need to ponder The pondering happened	✗
Relation Extraction	▶ Ward joined Tom in their native Perth Ward was born in Perth	✓
	▶ Stefan had visited his son in Bulgaria Stefan was born in Bulgaria	✗
Puns	▶ Kim heard masks have no face value Kim heard a pun	✓
	▶ Tod heard that thrift is better than annuity Tod heard a pun	✗

Figure 11: Examples taken from our Natural Language Inference collection.

To aid in the understanding of ever more capable neural language models, we assembled a diverse collection of Natural Language Inference (NLI) examples from across a wide number of tasks, including named entity recognition and relation extraction [20]. This led to a very large and diverse resource for performing analysis.

3.3.3 Language Model Probing

In the summer of 2018 a number of members of the JHU AIDA team collaborated with a larger group of researchers in a JHU workshop focused on understanding the recently emerging neural language models.

In one effort, we introduced a set of nine challenge tasks that test for the understanding of linguistic function words [21]. These tasks were created by structurally mutating sentences from existing datasets to target the comprehension of specific types of function words (e.g., prepositions, wh-words). Using these probing tasks, we explored the effects of various pretraining objectives for sentence encoders on the learned representations. Our results showed that pretraining on language modeling performs the best on average across our probing tasks, supporting its widespread use for pretraining state-of-the-art NLP models.

Prep.	**With** a single jerk the man's head tore free.	→	The man's head tore free **from** a single jerk.	✓
	With a single jerk the man's head tore free.	→	The man's head tore free **without** a single jerk.	X
Negation	This is a common problem.	→	This is **not** an **uncommon** issue we are facing.	✓
	This is **not** a common problem.	→	This is **not** an **uncommon** issue we are facing.	X
Spatial	To reach . . . turn **left** up a small alleyway	→	do not turn **right** up the alleyway . . .	✓
	To reach . . . turn **left** up a small alleyway	→	Turn **right** up the alleyway . . .	X
Quant.	**all** taken up yeah	→	There are not still **some** left	✓
	all taken up yeah	→	There are still **some** left	X
Comp.	Today there are **more** than 300,000.	→	Today there are not **less** than 300,000.	✓
	Today there are **more** than 300,000.	→	Today there are **less** than 300,000.	X

Figure 12: Examples from our function word probing NLI dataset.

This finding was reinforced through a comprehensive study of pretraining and target-specific fine tuning [22], and since this point language model pretraining is widely acknowledged as the driver of powerful new AI models.

In a highly influential study, alongside the other workshop participants we built on token-level probing to introduce a novel *edge probing* task design and constructed a broad suite of sub-sentence tasks derived from the traditional structured NLP pipeline [23]. We probed word-level contextual representations from four recent models and investigated how they encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena.

3.4 DATA CREATION

3.4.1 Claim Frames

The DARPA AIDA program sought to aggregate information from disparate sources, with an emphasis on conflicting reports and misinformation. The program in later stages transitioned from the development of traditional information extraction capabilities to the development of systems that produce output enriched with information pertaining to the event of the claim itself: who claims the extracted the information, what their sentiment about the underlying content is, what their belief about it is, and other information. The latter pieces of information pertaining to the claiming event itself (the claimer, their sentiment, their epistemic stance, etc.) are contained in a "claim frame".

During the AIDA annotation "hackathons", we developed and debugged a schema for annotating claims made in documents that pertain to certain topics of interest (e.g., the COVID-19 pandemic). After raising various questions and concerns about the task and proposing suggestions that were addressed by DARPA and other performers, we developed protocols for the task of annotating claim frames. We manually annotated 4 documents for claims related to the COVID-19 pandemic. During the last AIDA annotation hackathon, LDC analyzed the annotations produced by each performer and found that our manual annotations most closely matched their own.

We developed an annotation interface for use with Amazon Mechanical Turk in which annotators enter values for the claim frame fields, which also includes highlighting spans of text that provide provenance (textual evidence). We also set up a NIL Cluster Database that allows annotators to

quickly search for NIL clusters found in previous annotation and to add more NIL clusters in real-time, which avoids issues of data coherence across annotators and documents.

Having validated our own annotation of claim frames, we turned to crowdsourcing the annotations. We recruited 12 annotators (from a pool of about 70) on Amazon Mechanical Turk via a paid qualification task. Annotation quality was judged based on similarity to the manual annotations we did on four documents during the AIDA hackathons. Annotators received personalized feedback on their annotations, including ways to correct systematic errors they made. Annotators received instructions through the annotation interface as well as through supplementary documentation, which we revised following the qualification task to address common pitfalls [33,34].

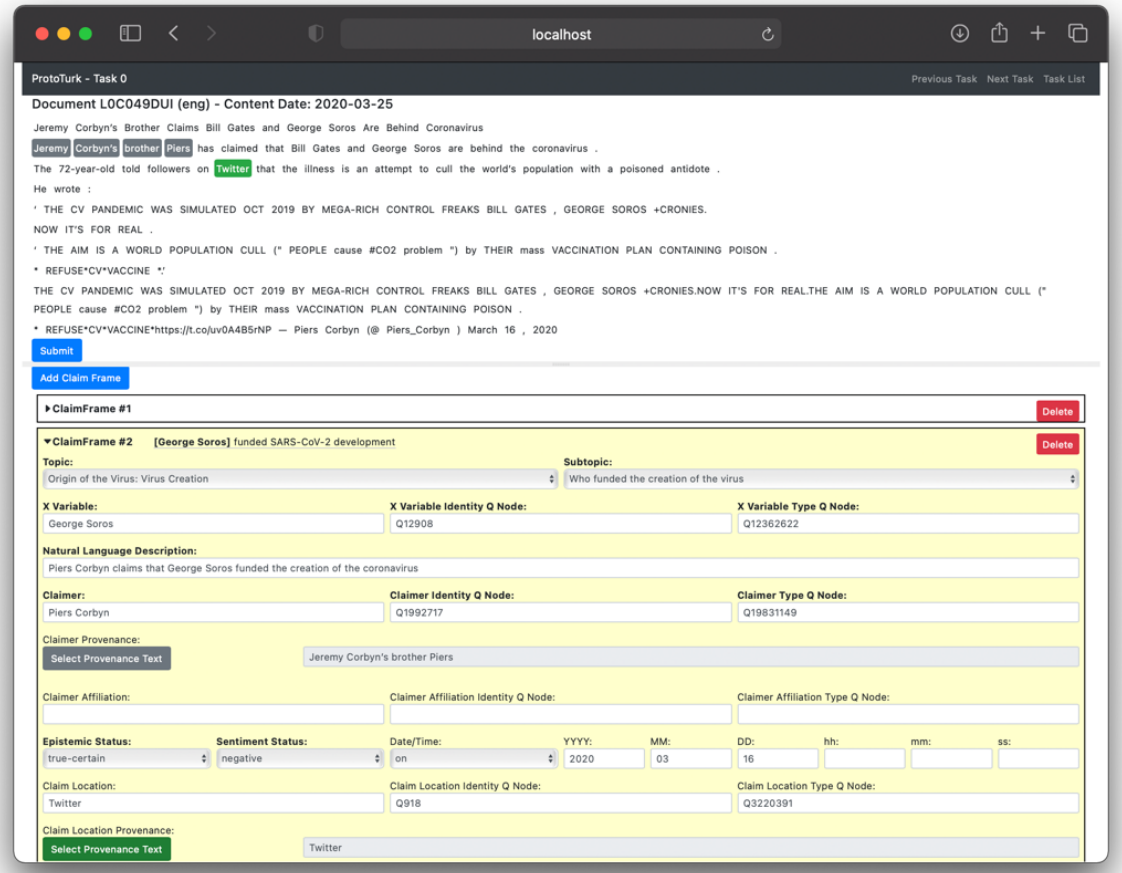

Figure 13: An annotated claim frame in our annotation interface.

Each document was annotated by 3 annotators, which provides broad coverage of annotated claims and various interpretations of the document. We developed post-processing scripts to convert the raw annotations into a format deliverable to LDC as well as to enforce various data quality checks and to facilitate manual deduplication. We obtained approximately 15 claims per document (mean: 16.8, median: 12.5), although there is considerable variance due to document

length. We annotated 60 documents designated by LDC to be used to compute agreement across annotation teams as well as 100 additional documents.

The same pool of annotators gave us annotations for our cross-claim relation pairs. We formulated the cross-claim relation pair as a Natural Language Inference task, giving one claim as a premise (Sentence 1) and the other as a hypothesis (Sentence 2) and asking the annotators the following question: "How likely is Sentence 2 true given that Sentence 1 is true?" The annotators responded on a 5-point Likert scale from "extremely unlikely" (1) to "extremely likely" (5). We map "extremely unlikely" (1) and "very unlikely" (2) to denote the "Refuting" relation. "Even chance" (3) is mapped to the "Related" relation. "Very likely" (4) and "extremely likely" (5) are mapped to the "Supporting" relation.

We used three-way redundancy on annotations in our cross-claims pilot and achieved a Krippendorff's alpha of 0.81. Based on this inter-annotator agreement, we use only one annotator per claim in our bulk task. We created two examples for each claim pair—by reversing the premise and hypothesis—and selected only the pairs where the annotations in both directions are compatible. In our annotation protocol, all pairs were compared only within the same document and the same claim frame topic.

Inner frame annotation of events and relations was done automatically using a FrameNet parser we developed previously [12]. We enumerated the frames that evoke claiming events (45 frames) and the claim frame topics under consideration (44 frames). These frames were also mapped to Wikidata QNodes via program-provided overlay files, Semlink 2.0 mappings, and the KGTK-similarity API [35]. In our last discussions with LDC they relayed a plan to release of the collection of data across all performers engaged in this task.

3.4.2 Frames Across Multiple Sentences (FAMuS)

Most event extraction corpora annotate at the sentence level, meaning they only identify arguments in the same sentence as the event trigger, as seen in the ACE datasets. However, the need for document-level annotation, allowing for arguments outside the sentence containing the event trigger or not requiring an event trigger at all, has been acknowledged since the Message Understanding Conferences (MUCs). To address this need, we previously created the RAMS dataset in the context of AIDA [10]. The RAMS dataset annotates 9,124 events in news articles from Reddit using the AIDA Phase 1 ontology and permits arguments to be anywhere within a five-sentence range of the event trigger.

One downside of having developed this dataset against the AIDA Phase 1 ontology is that it is not always straightforward to update the dataset with every new version of the ontology. The move toward using DARPA WikiData (DWD) could potentially alleviate this issue, but since DWD remains under development, it is similarly a moving target. In this project, we aimed to construct a general-purpose RAMS-like dataset using the broad-coverage FrameNet ontology.

FrameNet has a broad coverage of around 1200 frames. It is based on the concept of Frame Semantics, wherein the essential idea is that meaning of most words can be described by a semantic frame. A frame involves an event type, entity or relation, and the participants involved in the event (known as frame elements). The benefit of using FrameNet is that, in being broad coverage, it is in principle easier to map into any future program-specific ontology or into DWD.

For our project, we considered three categories of frames: events, states, and processes, along with manually selected uncategorized frames. This resulted in a total of roughly 530 frames that we refer to as describing a *situation.*

We take our documents from an ongoing project at Johns Hopkins University that scrapes passages from Wikipedia ensuring a stratified sample of different frames from FrameNet. This corpus has Wikipedia passages parsed with a FrameNet parser we developed under AIDA: LOME [12]. Each passage also comes with the text of its source document. Templatic questions are generated from the passage for each argument in the passage that has an answer in the source. We take the passage and source pairs from this corpus for building our document level role annotations.

For any passage-source pair, we parse the passage through LOME and get the top-5 predicted labels for each trigger in the passage. We first highlight an event trigger from the passage text and give the top-5 predicted labels as options to the annotator (Figure 14). Annotators can click the button for each event type to see its definition and an example in which the frame is evoked (see Figure 15).

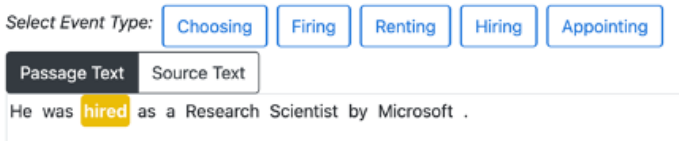

Figure 14: Event typing in our document level role annotation protocol.

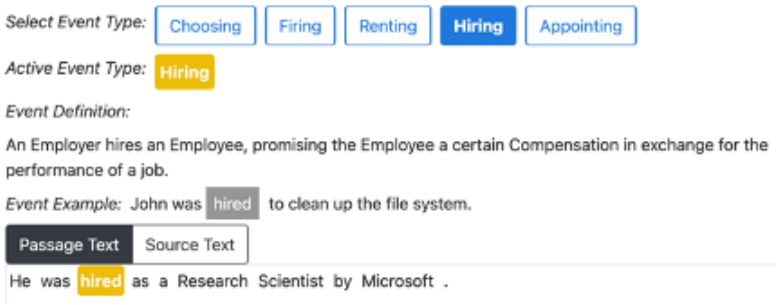

Figure 15: Clicking on a event type button displays the type definition with an example.

The LOME system reports a 91% accuracy on frame identification given the trigger span on the FrameNet v1.7 test set. We compute accuracy for top-k labels considering only cases where the prediction span exactly matches the true span. Figure 16 shows how LOME's frame identification performance improves if the true label is included in the top-k labels as compared to just the top label. We choose top-5 labels (97.1% accuracy) in our annotation to correct the inaccurate event type predictions from LOME's FrameNet parser.

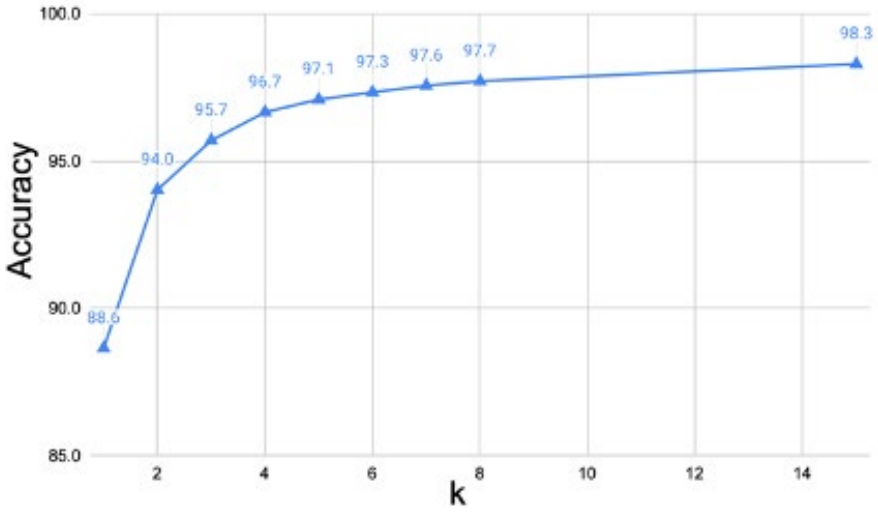

Figure 16: LOME's performance on Framenet v1.7 corpus as k increases.

To achieve a diverse stratified sample for role annotations across frames, we launched a frame identification task where for each frame f_i, we oversample D_i documents (passage-source pairs) based on LOME's precision P_i on the FrameNet test set. This sampling resulted in 3,504 documents for the frame identification task. For each document, we then asked workers on Amazon Mechanical Turk the following two questions:

1. Identify the event type of the highlighted span in the Passage Text
2. Tell us if the situation represented by highlighted span is also present in the Source Text

Identifying whether a source is a valid extraction for the highlighted trigger in the passage is an important task to continue role annotations for the same event in the Source Text. A snapshot from our task is presented below.

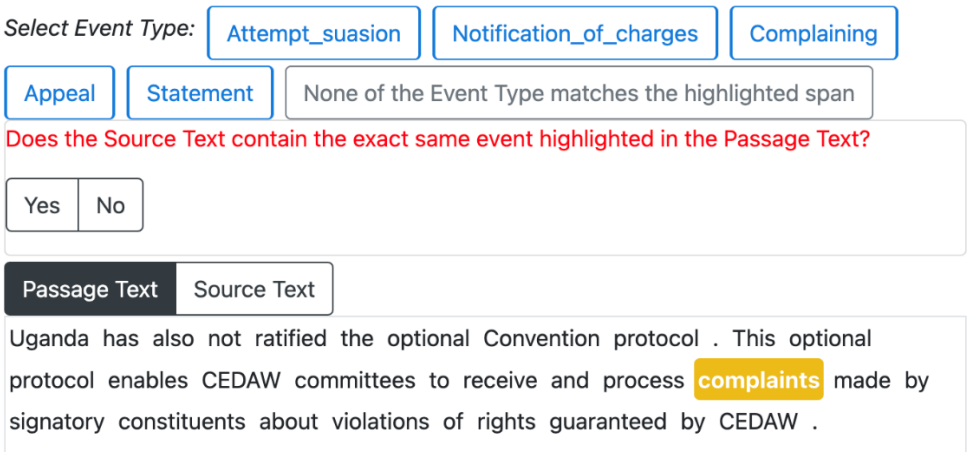

Figure 17: Frame identification task along with the source validation question

We qualified 20 annotators for the frame identification and source validation task by conducting a pilot study of 100 documents with gold annotations. Then, we launched the bulk task on 3,504 sampled documents with our qualified workers, using a three-way redundancy, and achieved a Krippendorf alpha of 0.62, which indicates high agreement between workers for the frame identification task. With the majority vote for source validation, we found that only 48% of all passage-source pairs had a valid source for the highlighted trigger in the passage.

The collected data introduces an interesting task of validating a source for a highlighted trigger in a passage. We intend to use the data collected from the frame identification and source validation task to perform role annotations on both the passage and the source documents.

3.5 DECOMPOSITION

Traditional semantic annotation frameworks generally define complex, often exclusive category systems that require highly trained annotators to build. And in spite of their high quality for the cases they are designed to handle, these frameworks can be brittle to cases that (i) deviate from prototypical instances of a category; (ii) are equally good instances of multiple categories; or (iii) fall under a category that was erroneously excluded from the framework's ontology.

Under AIDA, we developed an alternative approach to semantic annotation that addresses these issues: *decompositional semantics* [24, 25, 26]. In this approach, which is rooted in a long tradition of theoretical approaches to lexical semantics, semantic annotation takes the form of many simple questions about words or phrases (in context) that are easy for naive native speakers to answer, thus allowing annotations to be crowd-sourced while retaining high interannotator agreement.

The decompositional approach can be thought of as a feature-based counterpart to traditional category-based systems, with each question determining a semantic feature. Common feature configurations often correspond to categories in a traditional framework; but unlike such frameworks, a decompositional approach retains the ability to capture configurations that were not considered at design time. Further, unlike a categorical framework, reannotation after an overhaul of the framework's ontology is never necessary, since additional annotations simply accrue to sharpen the framework's ability to capture fine-grained semantic phenomena.

Under AIDA, we developed a variety of semantic annotation datasets and corresponding models that take a decompositional approach, including ones that target event factuality [27], linguistic expressions of generalizations about entities and events [28], and temporal properties and parthood structure of events [29, 30]; and we developed the decomp toolkit for working with these data [26] that we have used in developing our LOME system as well as decompositional semantic parsers [31, 32].

This toolkit unifies the decompositional semantics-aligned annotation sets listed above within the Universal Decompositional Semantics (UDS) semantic graph specification–with graph structures defined by the predicative patterns produced by the PredPatt tool and real-valued node and edge attributes constructed using sophisticated response normalization procedures. It also provides a suite of Python tools that make working with these data seamless, enabling a wide range of queries on these graphs using the SPARQL 1.1 query language.

UDS1.0 consists of three layers of annotations built on top of the English Web Treebank: (i) syntactic graphs built from existing gold Universal Dependencies parses on EWT; (ii) semantic graphs built from the predicate-argument structures deterministically extracted from those parses

using the PredPatt tool [25]; and (iii) semantic types for the predicates, arguments, and their relationships, derived from five decompositional semantics-aligned datasets. The following figure shows an example UDS graph with all three layers of annotation.

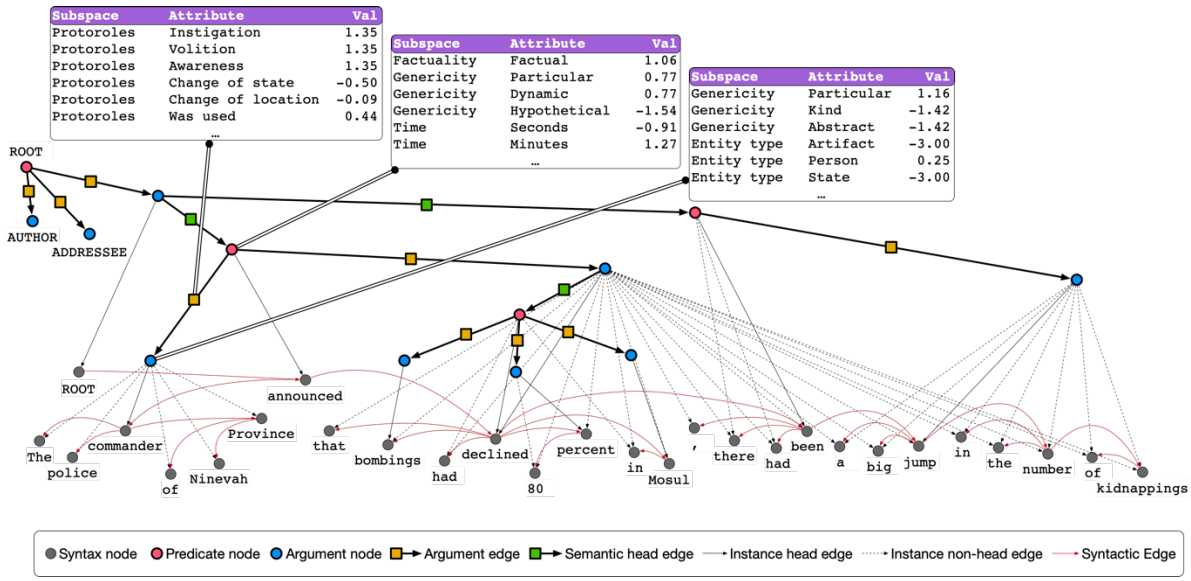

Figure 18: An example Universal Decompositional Semantics graph. Some semantic type information and most syntactic

In the remainder of this section we describe each dataset developed under AIDA, as well as the corresponding models, in more detail.

3.5.1 Factuality

A central function of natural language is to convey information about the properties of events. Perhaps the most fundamental of these properties is *factuality*: whether an event happened or not. A natural language understanding system's ability to accurately predict event factuality is important for supporting downstream inferences that are based on those events. For instance, if we aim to construct a knowledge base of events and their participants, it is crucial that we know which events to include and which ones not to.

The *event factuality prediction* (EFP) task involves labeling event-denoting phrases (or their heads) with the (non)factuality of the events denoted by those phrases. The following figure exemplifies such an annotation for the phrase headed by *leave*, which denotes a factual event (+=factual, – = nonfactual).

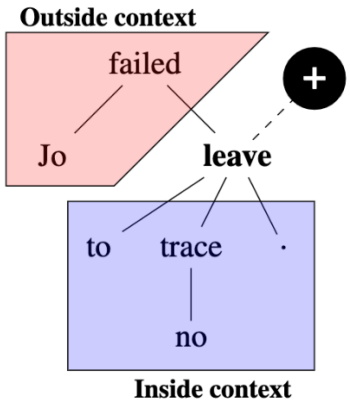

Figure 19: Event factuality (+=factual) and inside v. outside context for leave in the dependency tree.

We collected and release an extension of the UDS-IH1 dataset, which we refer to as UDS-IH2, to cover the entirety of the English Universal Dependencies v1.2 (EUD1.2) treebank, thereby yielding the largest publicly available event factuality dataset up to that point–substantially larger than FactBank, the UW dataset, and MEANTIME.

Table 1. UDS-IH1 dataset size vs other publicly available datasets

Dataset	Train	Dev	Test	Total
FactBank	6636	2462	663	9761
MEANTIME	967	210	218	1395
UW	9422	3358	864	13644
UDS-IH2	22108	2642	2539	27289

The following figure plots the distribution of factuality ratings in the train and dev splits for UDS-IH2, alongside those of FactBank, UW, and MEANTIME. One striking feature of these distributions is that UDS-IH2 displays a much more entropic distribution than the other datasets. This may be due to the fact that, unlike the newswire-heavy corpora that the other datasets annotate, EUD1.2 contains text from genres -- weblogs, newsgroups, email, reviews, and question-answers -- that tend to involve less reporting of raw facts. One consequence of this more entropic distribution is that, unlike the datasets discussed above, it is much harder for systems that always guess 3---i.e. factual with high confidence/likelihood---to perform well.

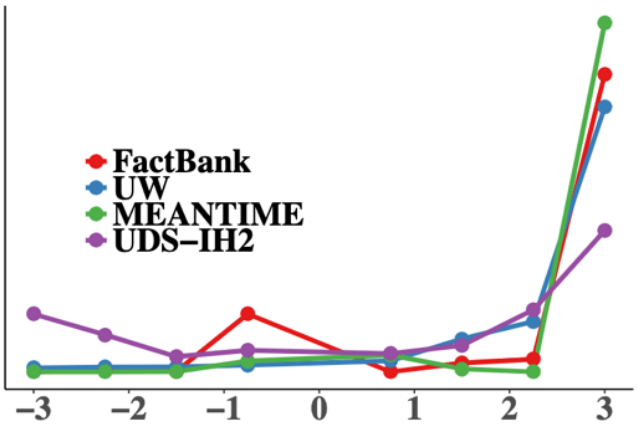

Figure 20: Relative frequency of factuality ratings in training and development sets.

We developed two neural models of event factuality (and several variants thereof), showing that these models significantly outperformed previous systems on four existing event factuality datasets as well as our new dataset.

3.5.2 Expressions of Generalization

Natural language allows us to convey not only information about particular individuals and events, as in *Mary ate oatmeal for breakfast today*, but also generalizations about those individuals and events, as in *Mary eats oatmeal for breakfast every day*.

This capacity for expressing generalization is extremely flexible–allowing for generalizations about the kinds of events that particular individuals are habitually involved in as well as characterizations about kinds of things, as in *bishops move diagonally*.

Such distinctions between *episodic statements*, on the one hand, and *habitual* and *generic* (or characterizing) statements, on the other, have a long history in both the linguistics and artificial intelligence literatures. Nevertheless, few modern semantic parsers make a systematic distinction. This is problematic, because the ability to accurately capture different modes of generalization is likely key to building systems with robust common sense reasoning: such systems need some source for general knowledge about the world and natural language text seems like a prime candidate.

One obstacle to further progress on generalization is that current frameworks tend to take standard descriptive categories as sharp classes–e.g. *episodic*, *generic*, *habitual* for statements and *kind* or *individual* for entities. This may seem reasonable for sentences like the above; but natural text is less forgiving. For instance, it's not clear how to label *client expectations* in *I will manage client expectations* or *the atmosphere* in *the atmosphere may not be for everyone*.

To remedy this, we proposed a novel decompositional framework for capturing linguistic expressions of generalization. In this framework, we decompose categories such as *episodic*, *habitual*, and *generic* into simple referential properties of predicates and their arguments. We deployed this framework to construct a large-scale dataset of annotations covering the entire Universal Dependencies English Web Treebank–yielding the Universal Decompositional Semantics-Generic-

ity (UDS-G) dataset–and we constructed models for predicting expressions of linguistic generalization that combine hand-engineered type and token-level features with static and contextual learned representations, finding that (i) referential properties of arguments are easier to predict than those of predicates; and that (ii) contextual learned representations contain most of the relevant information for both arguments and predicates.

3.5.3 Temporal Relations

Natural languages provide a myriad of formal and lexical devices for conveying the temporal structure of complex events---e.g. tense, aspect, auxiliaries, adverbials, coordinators, subordinators, etc. Yet, these devices are generally insufficient for determining the fine-grained temporal structure of such events. Consider the narrative:

> *At 3pm, a boy **broke** his neighbor's window. He was **running away**, when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn't **arrive** for two hours because they **were still at work***

Most native English speakers would have little difficulty drawing a timeline for these events, likely producing something like that in figure below. But how do we know that the breaking, the running away, the confrontation, and the calling were short, while the parents being at work was not? And why should the first four be in sequence, with the last containing the others?

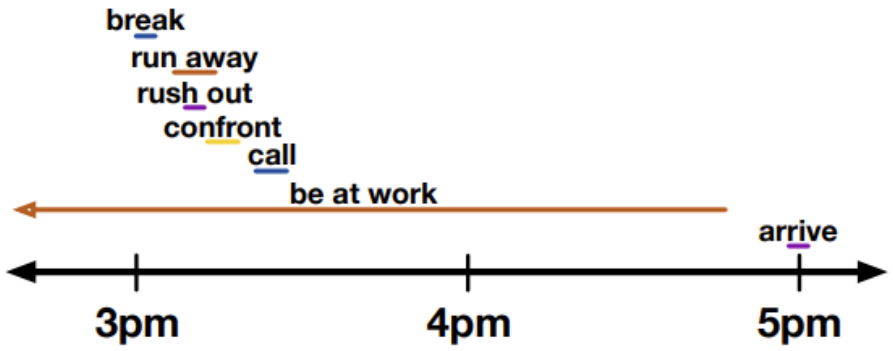

Figure 21: A typical timeline for the narrative in the text.

The answers to these questions likely involve a complex interplay between linguistic information, on the one hand, and common sense knowledge about events and their relationships, on the other. But it remains a question how best to capture this interaction. Prior work in this domain has approached this task as a classification problem, labeling pairs of event-referring expressions---e.g. *broke* or *be at work*---and time-referring expressions---e.g. *3pm* or *two hours*---with categorical temporal relations (often using the TimeML standard or some modified version). The downside of this approach is that time-referring expressions must be relied upon to express duration information. But as the narrative highlights, nearly all temporal duration information can be left implicit without hindering comprehension, meaning these approaches only explicitly encode duration information when that information is linguistically realized.

To address this issue, we developed a novel framework for temporal relation representation that puts event duration front and center. Instead of annotating text for categorical temporal relations as in previous approaches, we map events to their likely durations and event pairs directly to real-valued relative timelines using the interface in the figure below. This change not only supports the goal of giving a more central role to event duration, it also allows us to better reason about the temporal structure of complex events as described by entire documents.

Figure 22: An annotated example using our temporal relations interface.

We collected data for a large subset of predicate pairs found in the English Web Treebank, yielding UDS-T, the largest temporal relations dataset up to the point (see comparison in Table below).

Table 2. Number of total events, and event-event temporal relations captured in various temporal relations corpora, including our own (UDS-T).

Dataset	#Events	#Event-Event Relations
TimeBank	7,935	3,481
TempEval 2010	5,688	3,308
TempEval 2013	11,145	5,272
TimeBank-Dense	1,729	8,130
Hong et al. (2016)	863	25,610
UDS-T	**32,302**	**70,368**

We used this dataset to train a variety of neural models to jointly predict event durations and fine-grained (real-valued) temporal relations, yielding not only strong results on our dataset, but also competitive performance on TimeML-based datasets.

3.5.4 Event Structure

Natural language provides myriad ways of communicating about complex events. For instance, one and the same event can be described at a coarse grain, using a single clause, as in t*he contractors built the house*, or at a finer grain, using an entire document, as in t*hey started by laying the house's foundation. They then framed the house before installing the plumbing. After that...* Further, descriptions of the same event at different granularities can be interleaved within the same document---e.g. the finer description might well directly follow the coarser description as an elaboration on the house-building process.

Consequently, extracting knowledge about complex events from text involves determining the structure of the events being referred to: what their parts are, how those parts are laid out in time, who participates in them and how, etc. Determining this structure requires an event classification whose elements are associated with event structure representations. A number of such classifications and annotated corpora exist.

Similar in spirit to this prior work, but different in method, our work in this domain aimed to develop an empirically derived event structure classification. Where prior work takes a top-down approach---hand-engineering an event classification before deploying it for annotation---we take a bottom-up approach---decomposing event structure into a wide variety of theoretically informed, cross-cutting semantic properties, annotating for those properties, then recomposing an event classification from them by induction. The properties on which our categories rest target (i) the substructure of an event; (ii) the superstructure in which an event takes part; (iii) the relationship between an event and its participants; and (iv) properties of the event's participants. An example of select properties that we add to the existing annotations described above are exemplified below.

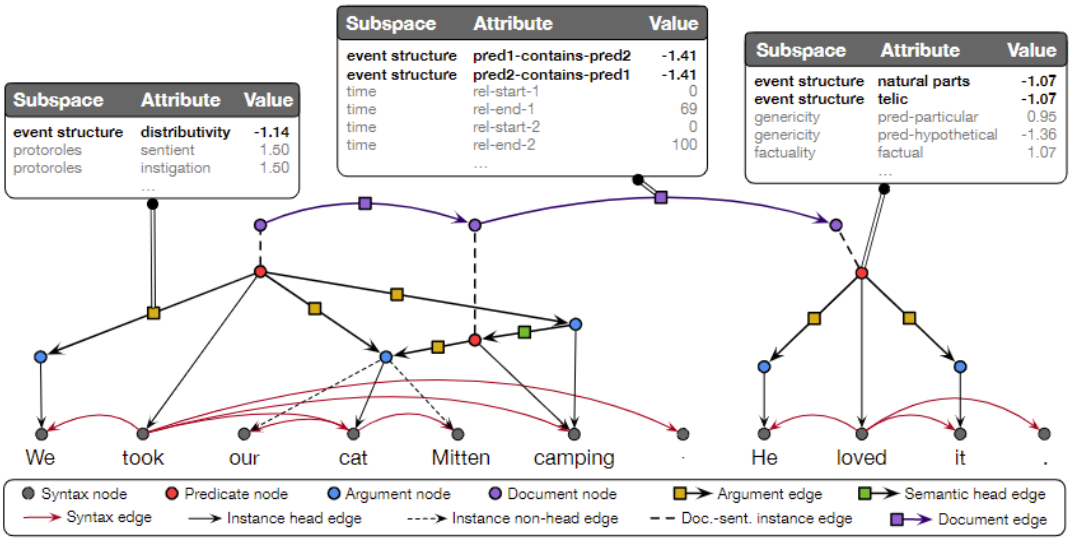

Figure 23: Example UDS semantics and syntax graphs with select properties. Bolded are ones we collected in this project; the document-level UDS graph is also shown in purple.

We collected data for a large subset of predicates and predicates pairs found in the English Web Treebank, yielding UDS-E. To derive an event structure classification from UDS-E and existing UDS annotations, we developed a document-level generative model that jointly induces event, entity, semantic role, and event-event relation types, exemplified below.

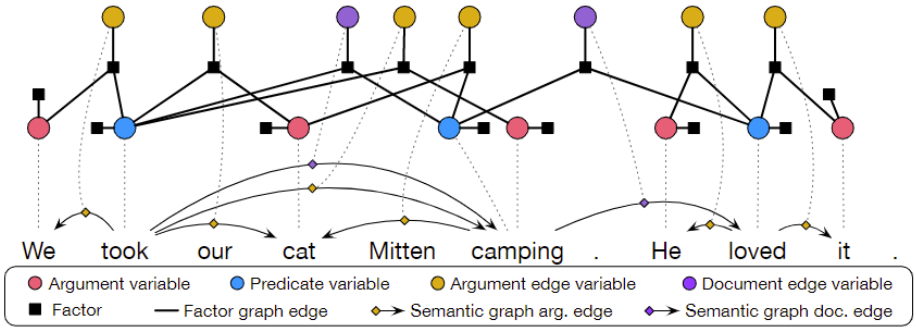

Figure 24: The factor graph assumed by our generative model. Each node or edge annotated in the semantic graphs becomes a variable node in the factor graph, as indicated by the dotted lines. Only factors for the prior distributions over types are shown; the annotation likelihood factors associated with each variable node are omitted for space.

Finally, we compared these types to those found in existing event structure classifications, finding both cases where the classifications converge but also places where they clearly diverge.

4.0 RESULTS AND DISCUSSION

Overall the efforts by the JHU AIDA team resulted in significant advances in fine grained and multilingual understanding models, along with contributing to the community a deeper understanding of the capabilities of neural language models.

5.0 CONCLUSIONS

Moving forward beyond AIDA, the active discussion in the community is whether even larger language models will lead to a move away from custom architectures for information extraction and instead a convergence on prompt-based frameworks that simply "ask" a cognitive architecture for answers in natural language, and these are converted to responses in whatever format the user desires, structured or unstructured. We believe that the advances made by performers such as JHU under AIDA are in some ways robust to these advances: our work in decompositional semantics, FrameNet, claim frames, and so on, these are all advances in data and the development of analytic frameworks for how to think about information extraction. As models improve, they can still be trained and evaluated on these resources. In regards to models, our work such as in incremental coreference resolution was motivated by computational concerns. Even if very large language models grow to be as accurate or more accurate than smaller custom models, it is still the case that high speed bulk analysis of very large collections is expensive: asking very large language models to process each item of a fast moving data stream is not computationally viable. We believe that custom analytics will continue to play an important part in the automatic construction of knowledge bases from text.

6.0 REFERENCES

[1] Which *BERT? A Survey Organizing Contextualized Encoders. Patrick Xia, Shijie Wu, Benjamin Van Durme. EMNLP 2020.

[2] Incremental Neural Coreference Resolution in Constant Memory. Patrick Xia, João Sedoc, Benjamin Van Durme. EMNLP 2020.

[3] On Generalization in Coreference Resolution. Shubham Toshniwal, Patrick Xia, Sam Wiseman, Kevin Gimpel, Karen Livescu. CRAC 2021.

[4] Moving on from OntoNotes: Coreference Resolution Model Transfer. Patrick Xia, Benjamin Van Durme. EMNLP 2021.

[5] Adapting Coreference Resolution Models through Active Learning. Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, Jordan Boyd-Graber. ACL 2022.

[6] CopyNext: Explicit Span Copying and Alignment in Sequence to Sequence Models. Abhinav Singh, Patrick Xia, Guanghui Qin, Mahsa Yarmohammadi, Benjamin Van Durme. SPNLP 2020.

[7] Multilingual Coreference Resolution in Multiparty Dialogue. Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, Benjamin Van Durme. arXiv 2022.

[8] Hierarchical Entity Typing via Multi-level Learning to Rank. Tongfei Chen, Yunmo Chen, Benjamin Van Durme. ACL 2020.

[9] Reading the Manual: Event extraction as definition comprehension. Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, Benjamin Van Durme. SPNLP 2020.

[10] Multi-Sentence Argument Linking. Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, Benjamin Van Durme. ACL 2020.

[11] Joint modeling of arguments for event understanding. Yunmo Chen, Tongfei Chen, Benjamin Van Durme. CODI 2020.

[12] LOME: Large Ontology Multilingual Extraction. Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, Benjamin Van Durme. EACL 2021 Demo.

[13] A Discriminative Neural Model for Cross-Lingual Word Alignment. Elias Stengel-Eskin, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. EMNLP 2019.

[14] ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. J. Edward Hu, Rachel Rudinger, Matt Post, Benjamin Van Durme. AAAI 2019.

[15] Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, Benjamin Van Durme. NAACL 2019.

[16] Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, Benjamin Van Durme. CoNLL 2019.

[17] Iterative Paraphrastic Augmentation with Discriminative Span Alignment. Ryan Culkin, J. Edward Hu, Elias Stengel-Eskin, Guanghui Qin, Benjamin Van Durme. TACL 2021.

[18] Gender Bias in Coreference Resolution. Rachel Rudinger, Jason Naradowsky, Brian Leonard, Benjamin Van Durme. NAACL 2018.

[19] On Measuring Social Biases in Sentence Encoders. Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, Rachel Rudinger. NAACL 2019.

[20] Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, Benjamin Van Durme. EMNLP 2018.

[21] Probing What Different NLP Tasks Teach Machines about Function Word Comprehension.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, Ellie Pavlick. StarSem 2019.

[22] Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling. Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, Samuel R. Bowman. ACL 2019.

[23] What do you learn from context? Probing for sentence structure in contextualized word representations. Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. ICLR 2019.

[24] Dee Ann Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. (2015). Semantic Proto-Roles. Transactions of the Association for Computational Linguistics, 3:475–488.

[25] Aaron Steven White, Dee Ann Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1713–1723. Austin, Texas: Association for Computational Linguistics.

[26] Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2020. The Universal Decompositional Semantics Dataset and Decomp Toolkit. In Proceedings of The 12th Language Resources and Evaluation Conference, 5698–5707. Marseille, France: European Language Resources Association.

[27] Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural Models of Factuality. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 731–744. New Orleans, Louisiana: Association for Computational Linguistics.

[28] Govindarajan, Venkata, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing Generalization: Models of Generic, Habitual, and Episodic Statements. Transactions of the Association for Computational Linguistics 7: 501–517.

[29] Vashishtha, Siddharth, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-Grained Temporal Relation Extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2906–2919. Florence, Italy: Association for Computational Linguistics.

[30] Gantt, William, Lelia Glass, and Aaron Steven White. 2022. Decomposing and Recomposing Event Structure. Transactions of the Association for Computational Linguistics 10: 17–34.

[31] Stengel-Eskin, Elias, Aaron Steven White, Sheng Zhang, and Benjamin Van Durme. 2020. Universal Decompositional Semantic Parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8427–8439. Online: Association for Computational Linguistics.

[32] Stengel-Eskin, Elias, Sheng Zhang, Kenton Murray, Aaron Steven White, and Benjamin Van Durme. 2021. Joint Universal Syntactic and Semantic Parsing. Transactions of the Association for Computational Linguistics 9: 756—773.

[33] Claim Frame Examples: https://docs.google.com/document/d/1t89zE5O8CkwDQpi-dUvCXuEQR95sheB1d7mNoTyyA8Fg/edit?usp=sharing

[34] Claim Frame Annotation Walkthrough: https://docs.google.com/presentation/d/1ljav27OAd5pBzF1xtn3OMZs0beNm8CUkcj6AtoGiREI/edit?usp=sharing

[35] KGTK-similarity API: https://github.com/usc-isi-i2/kgtk-similarity

APPENDIX A – PUBLICATIONS AND PRESENTATIONS

11/20/2020. CopyNext: Explicit Span Copying and Alignment in Sequence to Sequence Models. Fourth Workshop on Structured Prediction for NLP. Abhinav Singh.
[6] CopyNext: Explicit Span Copying and Alignment in Sequence to Sequence Models. Abhinav Singh, Patrick Xia, Guanghui Qin, Mahsa Yarmohammadi, Benjamin Van Durme. SPNLP 2020.

1/27/2019-2/1/2019. ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. AAAI Conference on Artificial Intelligence 2019. Edward Hu.
[14] ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. J. Edward Hu, Rachel Rudinger, Matt Post, Benjamin Van Durme. AAAI 2019.

6/2/2019-6/7/2019. Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting. 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Edward Hu.
[15] Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, Benjamin Van Durme. NAACL 2019.

11/3/2019-11/4/2019. Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. The SIGNLL Conference on Computational Natural Language Learning 2019. Edward Hu.
[16] Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, Benjamin Van Durme. CoNLL 2019.

i11/03/2019-11/07-2019. A Discriminative Neural Model for Cross-Lingual Word Alignment. Conference on Empirical Methods in Natural Language Processing. E. Stengel-Eskin.
[13] A Discriminative Neural Model for Cross-Lingual Word Alignment. Elias Stengel-Eskin, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. EMNLP 2019.

07/06/2020-07/08/2020. Universal Decompositional Semantic Parsing. Annual Conference of the Association of Computational Linguistics. E. Stengel-Eskin.
[31] Stengel-Eskin, Elias, Aaron Steven White, Sheng Zhang, and Benjamin Van Durme. 2020. Universal Decompositional Semantic Parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8427–8439. Online: Association for Computational Linguistics.

11/07/2021-11/11/2021. Joint Universal Syntactic and Semantic Parsing. Conference on Empirical Methods in Natural Language Processing. E. Stengel-Eskin.
[32] Stengel-Eskin, Elias, Sheng Zhang, Kenton Murray, Aaron Steven White, and Benjamin Van Durme. 2021. Joint Universal Syntactic and Semantic Parsing. Transactions of the Association for Computational Linguistics 9: 756—773.

07/28/2019 - 08/02/2019. Fine-Grained Temporal Relation Extraction. Annual Conference of the Association of Computational Linguistics. Siddharth Vashishtha.

[29] Vashishtha, Siddharth, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-Grained Temporal Relation Extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2906–2919. Florence, Italy: Association for Computational Linguistics.

07/06/2020-07/08/2020. Multi-Sentence Argument Linking. Annual Conference of the Association of Computational Linguistics. Seth Ebner.
[10] Multi-Sentence Argument Linking. Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, Benjamin Van Durme. ACL 2020.

11/16/2020-11/18/2020. Which *BERT? A Survey Organizing Contextualized Encoders. Conference on Empirical Methods in Natural Language Processing. Patrick Xia
[1] Which *BERT? A Survey Organizing Contextualized Encoders. Patrick Xia, Shijie Wu, Benjamin Van Durme. EMNLP 2020.

11/16/2020-11/18/2020. Incremental Neural Coreference Resolution in Constant Memory. Conference on Empirical Methods in Natural Language Processing. Patrick Xia
[2] Incremental Neural Coreference Resolution in Constant Memory. Patrick Xia, João Sedoc, Benjamin Van Durme. EMNLP 2020.

11/07/2021-11/11/2021. Moving on from OntoNotes: Coreference Resolution Model Transfer. Conference on Empirical Methods in Natural Language Processing. Patrick Xia
[4] Moving on from OntoNotes: Coreference Resolution Model Transfer. Patrick Xia, Benjamin Van Durme. EMNLP 2021.

11/11/2021. On Generalization in Coreference Resolution. Workshop on Computational Models of Reference, Anaphora and Coreference. Shubham Toshniwal
[3] On Generalization in Coreference Resolution. Shubham Toshniwal, Patrick Xia, Sam Wiseman, Kevin Gimpel, Karen Livescu. CRAC 2021.

04/19/2021-04/23/2021. LOME: Large Ontology Multilingual Extraction. Conference of the European Chapter of the Association for Computational Linguistics. Patrick Xia.
[12] LOME: Large Ontology Multilingual Extraction. Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, Benjamin Van Durme. EACL 2021 Demo.

05/06/2019-05/09/2019 What do you learn from context? Probing for sentence structure in contextualized word representations. International Conference on Learning Representations. Poster presented by multiple authors including Patrick Xia.
[23] What do you learn from context? Probing for sentence structure in contextualized word representations. Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. ICLR 2019.

07/28/2019-08/02/2019. Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling. Annual Conference of the Association of Computational Linguistics. Alex Wang.
[22] Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling. Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, Samuel R. Bowman. ACL 2019.

06/06/2019-06/07/2019. Probing What Different NLP Tasks Teach Machines about Function Word Comprehension. Joint Conference on Lexical and Computational Semantics. Najoung Kim.
[21] Probing What Different NLP Tasks Teach Machines about Function Word Comprehension. Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, Ellie Pavlick. StarSem 2019.

11/16/2020-11/18/2020 Reading the Manual: Event extraction as definition comprehension. Poster presented by Yunmo Chen.
[9] Reading the Manual: Event extraction as definition comprehension. Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, Benjamin Van Durme. SPNLP 2020.

11/16/2020-11/18/2020 Joint modeling of arguments for event understanding. Poster presented by Yunmo Chen.
[11] Joint modeling of arguments for event understanding. Yunmo Chen, Tongfei Chen, Benjamin Van Durme. CODI 2020.

7/6/2020–7/8/2020. Hierarchical entity typing via multi-level learning to rank. Annual Meeting of the Association for Computational Linguistics. Tongfei Chen.
[8] Hierarchical entity typing via multi-level learning to rank. Tongfei Chen, Yunmo Chen, Benjamin Van Durme. ACL 2020.

3/31/2021-4/2-2021. Veridicality and responsivity redux. Invited talk, Workshop on Clause-embedding Predicates, The Ohio State University. Kyle Rawlins.

5/22/2022 - 5/27/222. Adapting Coreference Resolution Models through Active Learning. Annual Conference of the Association of Computational Linguistics. Michelle Yuan.
[5] Adapting Coreference Resolution Models through Active Learning. Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, Jordan Boyd-Graber. ACL 2022.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

DOD	Department of Defense
AIDA	Active Interpretation of Disparate Alternatives
EWT	English Web Treebank
GPT	Generative Pre-Trained
BERT	Bidirectional Encoder Representations from Transformers
IE	Information Extraction
JHU	Johns Hopkins University
LDC	Linguistic Data Consortium
LM	Language Model
LOME	Large Ontology Multilingual Extraction
NIST	National Institute of Standards and Technology
NLI	Natural Language Inference
RAMS	Roles Across Multiple Sentences
UDS	Universal Decompositional Semantics
KGTK	Knowledge Graph Toolkit