Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1831–1838
Marseille, 20-25 June 2022

A Study in Contradiction: Data and Annotation for AIDA Focusing on Informational Conflict in Russia-Ukraine Relations

Jennifer Tracey, Ann Bies, Jeremy Getman, Kira Griffitt, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
{garjen, bies, jgetman, kiragrif, strassel}@ldc.upenn.edu

Abstract

This paper describes data resources created for Phase 1 of the DARPA Active Interpretation of Disparate Alternatives (AIDA) program, which aims to develop language technology that can help humans manage large volumes of sometimes conflicting information to develop a comprehensive understanding of events around the world, even when such events are described in multiple media and languages. Especially important is the need for the technology to be capable of building multiple hypotheses to account for alternative interpretations of data imbued with informational conflict. The corpus described here is designed to support these goals. It focuses on the domain of Russia-Ukraine relations and contains multimedia source data in English, Russian and Ukrainian, annotated to support development and evaluation of systems that perform extraction of entities, events, and relations from individual multimedia documents, aggregate the information across documents and languages, and produce multiple "hypotheses" about what has happened. This paper describes source data collection, annotation, and assessment.

Keywords: information extraction, disinformation, multimedia corpora

1. Introduction

The DARPA Active Interpretation of Disparate Alternatives (AIDA) program aims toward the development of technology that can assist in cultivating and maintaining understanding of events in a world where the truth can be hard to establish given conflicting accounts of what happened (e.g. who did what to whom and/or where and when events occurred). This is made more challenging when information relevant to obtaining a comprehensive understanding of what happened is scattered across different languages and media types. To address this challenge, AIDA systems must extract entities, events, and relations from individual multimedia documents, aggregate that information across documents and languages into a coherent semantic representation, and produce multiple knowledge graph hypotheses that characterize the conflicting accounts about what actually happened that are present in the corpus (DARPA, 2107) .

To support AIDA system development and evaluation, LDC produces corpora of source data and annotations and assesses system responses. Each phase of the program concentrates on a different topic scenario, with a source corpus, annotation tasks and assessment approach reflecting the particular technology goals for that phase. The current paper describes the resources produced for Phase 1 (as well as pilot work leading up to Phase 1), whose scenario is political relations between Russia and Ukraine in the 2010s. Multimedia source data was collected in English, Russian, and Ukrainian, focusing on several subtopics within the Russia-Ukraine domain that were known to have informational conflict; annotations were performed on a subset of the documents, including annotation of entities, relations and events, cross document coreference of annotated entities and events, and annotation of which hypotheses were supported by annotated events and relations. System output was then assessed for accuracy and completeness against the reference. Training data in the traditional sense was not included, but examples of data and annotation were provided for a set of "practice" topics to serve as representative examples of the approach used for evaluation data and topics. Each of these components are described in the following sections of this paper.

1.1 Related Work

The AIDA Phase 1 corpus makes several contributions to linguistic resources in the languages it includes, especially due to the multimedia nature of the corpus, and the focus on conflicting information.

Along with English and Russian, the AIDA Phase 1 corpus contains Ukrainian data, which is a relatively low resource language. Shvedova (2020), in introducing the General Regionally Annotated Corpus (GRAC) of Ukrainian (uacorpus.org), points out the lack of available resources for Ukrainian. The handful of published Ukrainian corpora are text-only resources (e.g. Tracey and Strassel, 2020; Sitchinava, 2012; Grabar and Hamon, 2019; Grabar et al., 2018).

Multimedia data -- especially documents on the web that contain embedded images, video and audio clips, infographics, social media snippets and reader comments interwoven with traditional text -- is gaining importance as sponsors seek robust technologies that can extract key knowledge elements from diverse information streams. Especially for Ukrainian, but for Russian and English as well, most corpora focus on one media type or include multiple media types but treat each media file as a self-contained document. Exceptions are mostly English corpora of subtitled video or images with captions (e.g., Han et al., 2018; Lin et al., 2015). The AIDA corpus presents a new addition to the available resources by providing a multilingual and multimedia data designed to support extraction of complex information in a holistic way.

Finally, the subject of misinformation and conflicting narratives is of great interest and import in the global information age, both generally (Gottlieb, 2020; Dan, 2021; Shu, 2020) and in the Ukraine-Russia relations domain in particular (Szostek, 2018). The AIDA Phase 1 corpus was designed to contain multiple and sometimes opposing narratives about specific topics within the Ukraine-Russia domain, in order to support the development and evaluation of technology capable of building multiple hypotheses to account for alternative interpretations of data imbued with informational conflict.

2. Scenario and Topic Development

The Russia-Ukraine Relations scenario that was the focus of AIDA Phase 1 centered on political and military relations between Russia and Ukraine in the 2014-2015 period, including events leading up to and following the main focal topics. The scenario was chosen because it was rich in opposing narratives and portrayals of the facts, across a variety of media types and languages. Within the broader domain, several subtopics were selected for data collection and annotation.

2.1 Topic Selection and Development

Six scenario-relevant topics were selected for Phase 1 of AIDA, three as practice data for systems and three as data for evaluation, as shown in Table 1 below[1].

Set	Topic
Practice	Who Started the Shooting at Maidan? (February 2014)
Practice	Ukrainian War Ceasefire Violations in Battle of Debaltseve (January-February 2015)
Practice	Donetsk and Luhansk Referendum, aka Donbas Status Referendum (May 2014)
Eval	Suspicious Deaths and Murders in Ukraine (January-April 2015)
Eval	Odessa Tragedy (May 2, 2014)
Eval	Siege of Sloviansk and Battle of Kramatorsk (April-July 2014)

Table 1: AIDA Phase 1 Topics

Topic development began with LDC annotators familiar with the domain; they proposed a number of candidate topics and rated them on criteria such as prevalence of differing assertions about what happened, availability of data in all three target languages, and availability of multimedia data. Their work was informed by documentation produced for AIDA by the MITRE Corporation, which detailed the scenario's key themes, events, and players.

Final selection of topics was made jointly by LDC and MITRE to include those which had good data availability and multiple aspects of informational conflict. We also considered which kinds of informational conflict could be expressed readily in the program ontology and corresponding annotation tag set.

2.2 Creating Topic Models

After topics were selected, a set of topic-based materials was developed to support data collection, annotation and system evaluation. These included a set of queries and hypotheses and a descriptive topic model document.

2.2.1 Queries, Hypotheses, and Prevailing Theories

To highlight each topic's points of informational conflict prior to data collection and annotation, we next developed a set of questions related to different facets of informational conflict for the topic; these were known as queries. Each query represented a specific and narrowly-defined information need in the broader topic, where the answer typically consists of a single entity, relation or event.

For example, the Topic "Who started the shooting at Maidan" has the following queries:

- Who fired on protesters at the Maidan protests?
- Who fired on police at the Maidan protests?
- What groups or organizations were the Maidan shooters affiliated with?
- How many people were shot during the Maidan protests?
- Who owned the weapons used during the Maidan protests?

Queries were used in two ways. First, they helped focus data collection efforts to ensure that the corpus as a whole contained conflicting accounts of key information facets. Second, the queries served as Statements of Information Need (SINs) during the evaluation, in which systems were expected to return hypotheses relevant to the query along with related information from the corpus (NIST, 2019).

Beyond the queries themselves, it was also necessary to create a set of single-facet hypotheses and broader topic-level hypotheses to characterize the kind of information systems might be expected to discover in the corpus. Query or facet-level hypotheses were akin to query answers, and were informed by MITRE scenario documentation and LDC's subsequent topic-specific research. Figure 1 shows an example of a facet-level hypotheses for a single query.

> Query: Who fired on protesters and/or police at the Maidan protests?
>
> H1: Members of the Berkut, a police force loyal to the Yanukovych government, fired on protesters.
>
> H2: Snipers loosely affiliated with the Ukrainian government fired on protesters.

Figure 1: A query and its facet-level hypotheses

In addition to facet-level hypotheses, more complex topic-level hypotheses were also needed to support the evaluation, since AIDA systems must aggregate information from across the whole corpus to produce complex and semantically coherent knowledge graphs that present different hypotheses about what happened with respect to the topic as a whole. Reference topic-level hypotheses took the form of "prevailing theories", which describe domain expert expectations about the differing accounts or perspectives on the topic that are likely to occur in the data based on prior knowledge of the scenario. While

[1] An additional three topics were included only in pilot annotation efforts.

facet-level hypotheses address a single information element, prevailing theories are often responsive to multiple queries and represent a coherent perspective within the larger topic narrative. Prevailing theories were largely informed by MITRE's scenario documentation, but also shaped by LDC's preliminary annotation of the corpus to provide some insight into which of the potential prevailing theories actually appeared in the data.

For each prevailing theory identified, we created a natural language description characterizing this account of the topic, plus a list of the events and relations, along with their arguments (entities), that would be required as part of a knowledge graph that adequately reflects this theory. Some prevailing theories with more narrow scope include only a short list of required events and relations, while some with broader scope include a much longer list.

The figure below illustrates one natural language prevailing theory along with the corresponding list of events and relations that would be required in a system hypothesis that fully captures the content of the prevailing theory.

Natural language prevailing theory

Snipers affiliated with Ukraine's Berkut riot police, under the direction of Aleksandr Yakimenko (who was at the time head of Ukraine's SSU intelligence service) and in collaboration with Russia's intelligence service the FSB, killed at least 53 anti-government activists protesting in Kiev's Independence Square (aka Maidan) on February 20 2014, using AK-47s and sniper rifles.

Required events and their arguments

- **conflict.attack.firearmattack**
  - Victim:protesters
  - Attacker: snipers
  - Weapon: AK-47s and sniper rifles
  - Place: Maidan
  - Temporal: on 2014-02-20
- **contact.collaborate.meet**
  - Participant 1: snipers
  - Participant 2: FSB
  - Place: Maidan
- **life.die.deathbyviolentevents**
  - Victim: at least 53 protesters
- **conflict.demonstrate.marchprotest**
  - Place: Maidan
  - Temporal: in February 2014

Required relations and their arguments

- snipers **affiliated with** Berkut
- Berkut **affiliated with** Ukraine
- Yakimenko **affiliated with** snipers
- Yakimenko **leader of** SSU
- SSU **affiliated with** Ukraine
- FSB **affiliated with** Russia
- protesters **affiliated with** pro-Maidan

Figure 2: A prevailing theory and its corresponding events and relations

2.2.2 Descriptive Topic Model

Once the queries were developed, a topic model was created describing the topic and its main points of informational conflict. The topic model was used as a reference by annotators during data collection and annotation, to help guide their decisions about topic relevance. The topic model includes the topic title, a description of the events in the topic, a summary of the expected informational conflict for this topic, and the list of natural language queries developed for the topic.

3. Source Data

The source data for AIDA Phase 1 corpus consists of multimedia (text, image, and video) data in English, Russian, and Ukrainian. A portion of the documents are directly relevant to the specific topics and queries discussed in section 2, while other documents are generally relevant to the Russia-Ukraine Relations scenario without addressing the particular topics of interest, or are background documents of unspecified topic content.

To ensure that sufficient content to support annotation and evaluation was present in the corpus, we seeded the corpus with documents specifically collected for their relevance to the topics and queries. LDC annotators used the topic model to guide their search for documents on the web, with the goal of locating a diverse set of documents with respect to media types, languages, and conflicting perspectives. Annotators collected information about the URL, language, presence of relevant text, image, and/or video, and presence of query answers for each document. The results of this manual search were then fed into LDC's multimedia data collection pipeline that collects all text, image, and video elements of the specified web page, processes them into separate files, and records metadata that maintains the association between the URL and each element of the page, as well as additional information such as publication date, download date, and relative position of each element on the original web page.

Beyond the manually scouted relevant documents, additional data was collected automatically from websites that were rich in appropriate data to fill out background documents required for the corpus. Keyword ranking was used to ensure that there was sufficient scenario-relevant content on the automatically harvested data to support program requirements.

The collected data was partitioned into three parts: a background corpus of English, Russian, and Ukrainian data from the approximate time period of the topics of interest; a corpus of documents collected specifically for their relevance to the topics designated as practice topics for use in system development, and an evaluation corpus consisting of some documents specifically collected for their relevance to the topics designated as evaluation topis, as well as background data that is from the same approximate time period and similar sources, some of which may also have general relevance to the wider scenario of Russia-Ukraine relations. The table below shows the number of documents in each partition.

	Eng	Rus	Ukr	Mixed	Total
Background	783	1753	1101	6031	9668
Practice	245	518	269	639	1671
Evaluation	752	511	670	63	1996
Total	1780	2782	2040	6733	13335

Table 2: Document counts by language

Table 3 shows the number of individual text, image, and video files present in each partition. Each collected document typically has at most one text element, but may have multiple image and/or video elements, such that the number of files exceeds the number of documents in the corpus.

	Text	Image	Video	Total
Background	8189	13770	943	22902
Practice	1608	10537	479	12624
Evaluation	1996	6194	322	8512
Total	11793	30501	1744	44038

Table 3: File counts by media type

The source data corpora distributed to AIDA include processed versions of the source data, with consistent formatting across sources and languages. All text data is presented in a tokenized and sentence segmented xml format, and all image and video files are presented with a wrapper that encodes standardized metadata in the header for each file. In addition, a table included in the corpus documentation describes the relationship between the "parent" web page (the document) and the "child assets" (the individual text, image, and video files that appeared together as a document on the original live webpage). Finally, all video elements were segmented to provide a stable reference set of shot boundaries for use in evaluation.

4. Annotation

The purpose of AIDA Phase 1 annotation was twofold: to provide a set of practice documents containing examples of the labeled events and relations necessary to characterize AIDA facet- and topic-level hypotheses for practice topics, and to produce a set of reference data that could be used to measure system performance against blind evaluation topics. Annotation proceeded in two stages. To support the AIDA program pilot evaluation, intended to help program performers define fundamental concepts and to establish baseline system performance, LDC produced a set of annotated pilot topics. The larger Phase 1 corpus included the official practice and evaluation topics for the phase evaluation. We discuss both pilot and Phase 1 annotation, noting key differences between the two approaches.

AIDA pilot and Phase 1 annotation consists of labeling topic-relevant events and relations, along with their arguments and attributes, including cross-document (i.e. cross-language/cross-media) coreference and KB linking. While the pilot effort also included explicit annotation of query-level hypotheses, in Phase 1 prevailing theories were incorporated into the system assessment procedure rather than being manually annotated in advance. All annotation was performed by trained annotators who had native fluency in the language of the document.

4.1 Topic Salience

Using information collected during data scouting, documents known to contain query-responsive information were selected for manual annotation. Given AIDA's research goals and evaluation design, for each document annotators labeled events and relations relevant to a single topic, ignoring information about other topics even when it was present in the document. Because annotation was necessarily restricted to topic-relevant information only, it was necessary to create an operational definition of relevance, or "topic salience", that would strike the right balance between completeness and precision. The definition had to balance the need to constrain the inclusion of irrelevant or marginally relevant events and relations in the annotation against the need to allow for the presence of unexpected information in the document, which could lead to discovery of a novel hypothesis about the topic. In addition, the definition of salience needed to be intuitive and promote consistent annotation.

In the pilot effort, topic salience was initially defined as "critical to understanding the topic, and/or frequently associated with the topic." Although annotators found this definition to be reasonably intuitive, it led to over-annotation of irrelevant events and relations. The definition was then revised to "supports one or more facet-level hypotheses," but in practice this definition was too constrained, and prohibited annotation of relevant items that were needed to support construction of semantically coherent topic-level hypotheses. For instance, information about affiliations between entities or supporting events such as the transfer of a weapon from one entity to another were often left out when the annotation was restricted to supporting the hypothesis that a particular entity fired on protesters. This restrictive definition also precluded discovery of new hypotheses that were not part of the a priori set constructed during topic development.

Based on the lessons learned during pilot annotation, for Phase 1 annotation salience was redefined as "relevant to one or more queries". By focusing on relevance to the queries rather than to the hypotheses, this working definition of topic salience sufficiently excluded the presence of irrelevant or marginally-relevant information in the annotations, while allowing annotators to capture the full set of events and relations needed to represent topic-level hypotheses; furthermore it did not preclude annotators from labeling information about potential novel hypotheses that were query-responsive.

4.2 Annotation Decision Points and Non-Exhaustive Annotation

For each salient event or relation subject to annotation, AIDA annotators made a number of decisions. First, each event or relation, and each associated entity argument, was anchored in document-level provenance. Provenance for text instances was recorded as character offsets. In the pilot annotation effort, provenance for image and video instances consisted of a document ID only. In the Phase 1 annotation, image instances were labeled with bounding box provenance, while video instances included keyframe IDs, bounding boxes and/or timestamps as appropriate. Annotators provided a brief text description (word or short phrase) for each event, relation or argument and assigned it a type from the annotation tag set. Arguments were also labeled for the role they play in the event or relation, e.g. the perpetrator vs. victim of an attack event. Text instances of arguments were also labeled as name, nominal or pronoun.

Annotators then specify any attributes associated with the event, relation or argument. The two primary attributes are “not”, indicating negation, and “hedged”, indicating uncertainty. "Not" is used to indicate that the source asserts that the event or relation did not happen, or that the argument did not participate in the event. "Hedged" is used to indicate uncertainty as reported by the source. A mention can combine "hedged" and "not" attributes, indicating that that source asserted that the relation or event possibly or likely did not happen.

Additional attributes were included in the pilot annotation effort. Events could be labeled as "deliberate" or "accidental", capturing explicit assertions by the source about whether the event was intentional or not. Some events could further be labeled as “legitimate” or “illegitimate”, to capture political legitimacy of elections. In Phase 1, these attributes were replaced with new relation types which captured the source of the assertion about deliberateness or legitimacy.

Finally, relations and events are labeled for temporal information. Annotators specify start and/or end dates for the relation or event when that information is present in the document, characterizing the information with as much precision as possible given document information. Dates are characterized as starting (ending) on, before or after a particular date, and the date is expressed in year-month-date format, with partially populated dates possible. For instance, a timestamp of *start on 2015-01-XX* indicates that the event or relation started sometime on or after January 2015, with a more precise date unknown. If the document does not contain any date information for the event or relation, annotators choose a start and/or end date of unknown.

Because the goal of AIDA annotation was to provide examples (in the case of practice topics) and reference annotations (in the case of evaluation topics) of the events and relations plus their arguments required for hypothesis construction, it was not necessary to label every occurrence (i.e. mention) of a given event or relation in the data. Furthermore, by program design the annotated data was not intended to serve as training data for basic information extraction, where exhaustive annotation of all event and relation mentions may be required. Given this context, a pragmatic decision was made to limit annotation to a single occurrence of each event or relation (plus arguments) per each kind of document element, where document elements are the set of individual text, image and video files that make up a complete “parent” document. For example, if the same event is depicted in both the text and an image of a document, it is labeled once in the text and once in the image; if the same event appears 3 separate times in the video portion of the document, it is labeled only once for the video document element.

At the request of AIDA system developers, some additional event and relation instances were labeled in Phase 1, expanding on the approach taken in the pilot. First, audio and video tracks of video document elements were treated as distinct for purposes of annotation, such that if an event appeared in video with both audio and visual evidence for the event, annotators would label one event in the audio track, and a separate event in the video track. Second, arguments that occur in closest proximity to the “trigger” (word or keyframe) where the event or relation first appears in a document were always labeled as base mentions, with more distant argument mentions being optionally labeled as informative mentions. Third, as time permitted annotators labeled additional informative mentions of events/relations and their arguments beyond the single instance per document element, with a strong preference for instances that provided especially helpful instantiations of the event, relation or argument - for instance, a very clear video image or a detailed text description naming the actor in some event. Finally, a handful of documents were exhaustively labeled with all salient events and relations plus their arguments and attributes, to provide a resource for system error analysis.

4.3 Annotation Tag Set

The AIDA annotation tag set started with the adoption of an existing ontology developed the prior DARPA DEFT program (DARPA, 2012) and was enriched for Phase 1 to reflect system and evaluation requirements. The existing DEFT ontology and corresponding annotation tag set included a number of crucial gaps that would be needed to capture important informational conflict in the domain of Russia-Ukraine relation. These gaps were addressed for Phase 1 by incorporating new domain-relevant tags as well as additional types proposed by the AIDA Program Ontology Working Group. The table below shows the number of types annotated in the pilot effort and in Phase 1.

	Pilot	Phase 1
Entity types	20	187
Relation types	23	49
Event types	47	139
Total types	90	375

Table 4: AIDA Program Ontology and Annotation Tag Set Expansion for Phase 1

The motivation for this significant expansion of the ontology was two-fold: it was intended to cover specific areas of information conflict that were absent from the pilot tag set, but also to provide finer grained subtypes and sub-subtypes that AIDA systems might be able to utilize in distinguishing different hypotheses. For example, while the pilot tag set included a single conflict.attack event type, the Phase 1 tag set broke this into 11 different subsubtypes, distinguishing firearm attacks, bombings, and missile strikes among other subsubtypes of attacks. Every event and relation in Phase 1 also included an “unspecified” subsubtype that annotators could use as a backoff category when the more detailed type of attack (or other event/relation) was not clear from the data, or when the detailed subsubtype was not present in the tag set.

4.4 KB Linking and Coreference

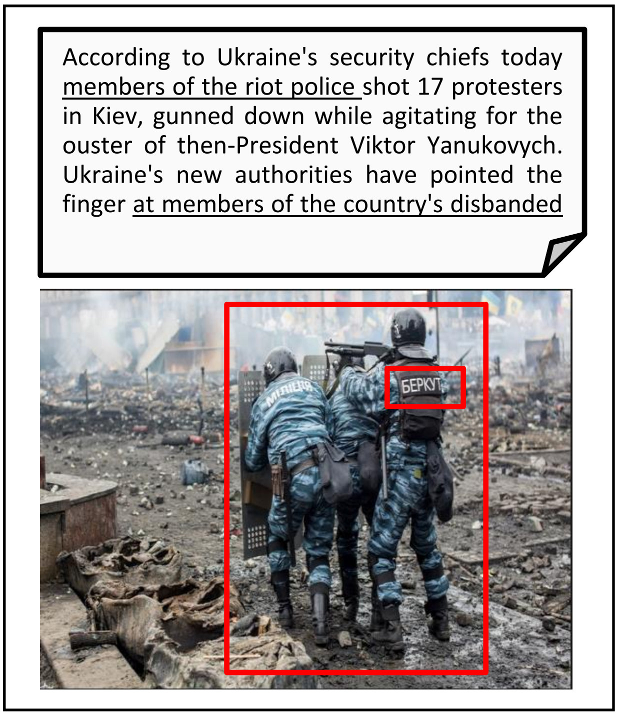
After salient events, relations and their arguments and attributes were labeled in the data, annotators performed coreference. To support creation of corpus-wide hypotheses, cross-document coreference was a requirement; for AIDA, cross-document necessarily means both cross-language and cross-media. Procedurally, coreference was achieved by manually linking individual entity and event instances to a knowledge base (KB).

For the pilot annotation effort, we created individual "mini-KBs" for each topic, containing only the events, relations and entities (arguments) anticipated to be relevant for that topic. Annotators manually linked individual event, relation and entity instances to the KB and flagged any items that could not be linked so that new KB entries could be created.

The Phase 1 evaluation design required a program-wide reference entity KB, constructed by LDC to include salient entities identified during topic development as well as a large number of general domain entities drawn from the LORELEI reference KB (Tracey and Strassel 2020). Coreference annotation for entities consisted of manually linking entity instances to the reference KB. When no match is present in the KB, the entity is marked as NIL; once all KB linking is complete, all NILs are reviewed and clustered, such that multiple mentions of the same NIL entity are assigned a unique NIL ID. Events were also manually clustered and assigned unique NIL IDs. Finally, relations were automatically clustered and assigned unique NIL IDs based on the results of manual entity clustering: relations with the same type, and whose arguments have the same argument role and contain the same entity (KB or NIL) ID, are considered coreferential.

Figure 3 below shows an example of a multimedia document with both text and image. In the annotation shown in the example, first the base mention of the attack event with the members of the riot police as attackers is annotated (shown in italics in the figure, with the attacker underlined), then additional text mentions to the attackers are added (additional underlined mention of the Berkut). Next, bounding boxes are added to the images to provide provenance for additional mentions found in the image. Finally, all of the mentions of the riot police are grouped together under the same KB ID in the KB linking stage of annotation.

For each relation or event labeled in a document, annotators indicated whether or how that item was relevant to the set of hypotheses for that document's topic, specifying whether the event/relation fully or partially supports the hypothesis or contradicts the hypothesis. Each event or relation was marked as relevant to zero, one, or more than one hypothesis. For example, one of the facet-level hypotheses for the query "Who fired on protesters and/or police at the Maidan protests?" is "Members of the Berkut, a police force loyal to the Yanukovych government, fired on protesters." An attack event taking place at Maidan with *Berkut* as the attacker and *protesters* as the victim would be marked as fully supporting the hypothesis. In contrast, an attack event with *security forces* as the attacker and *guns* as the weapon would be marked as partially supporting the hypothesis. An attack event in which *security forces* were the **victim** would be marked as contradicting the hypothesis.



According to Ukraine's security chiefs today members of the riot police shot 17 protesters in Kiev, gunned down while agitating for the ouster of then-President Viktor Yanukovych. Ukraine's new authorities have pointed the finger at members of the country's disbanded

Figure 3: Multimedia document example with annotation

Given the evaluation design for Phase 1, direct annotation of hypotheses was no longer necessary. Instead, the emphasis shifted to development of prevailing theories which characterized the broader topic-level hypotheses expected to be present in the corpus. These prevailing theories were then used as part of the post-hoc assessment of system output, as discussed in Section 5.

4.5 Quality Control

Quality control was performed on the annotation at several points in the process. First, the annotation interface prevents certain types of badly formed annotations from being created (missing required elements, combining incompatible types and subtypes, etc.). Then, after each stage of annotation, a second annotator reviews the annotations and corrects any errors. Once KB linking has been performed, additional corpus-wide quality checks are performed, such as checking all KB links for annotations with identical strings to ensure that they are linked consistently and reviewing the strings and types for all entities linked to the same KB node to check that there are no clear errors.

4.6 Results

The total amount of annotation completed in AIDA Phase 1 is summarized in the table below, which shows the number of documents annotated, as well as the number of entity, event, and relation mentions labeled in each data set.

	docs	event	relation	entity
Pilot	704	1967	569	4529
Phase 1 practice topics	204	1955	1848	8930
Eval topics	248	2456	2029	10508
Total	1156	6378	4446	23967

Table 5: Annotated documents and mentions

In addition to the mention annotations listed above, pilot annotation included 45,240 judgments indicating whether events and relations supported or contradicted facet-level hypotheses.

5. Assessment

Assessment of Phase 1 AIDA systems corresponded to the program's three primary technical areas: extraction of events, relations and entities; coreference of semantically equivalent instances; and hypotheses construction.

The class assessment task asked assessors to judge whether a given system-extracted entity had the correct type assigned; in Phase 1 assessors were also required to complete coreference and KB linking of system responses to facilitate downstream tasks.

In the zero-hop assessment task assessors judged whether a given system entity was assigned by the system to the correct KB node (that is, was it conferential with the reference entity assigned to that KB node).

During graph assessment, assessors examined system-returned relation and/or event arguments and indicated whether the system's predicate justification provided adequate evidence for the assigned type, the argument role, and the entity filling the argument role. This task required close inspection of the multi-media provenance returned by systems to justify their events and relations.

Finally, the most important and complex assessment task was hypothesis assessment. This task consisted of judgements about the relevance of system events, relations and their arguments included in the hypothesis, relative to the Statement of Information Need that prompted the hypothesis. Assessors also judged the semantic coherence of each hypotheses (i.e., is the hypothesis free of illogical or contradictory information), and assessed whether the system adequately covered the events and relations that underlie the prevailing theory associated with this hypothesis. Discussion of evaluation results and scoring is out of scope for this paper, which focuses on the corpus resources produced, rather than the evaluation in which they were used. Information about the evaluation structure and scoring of system responses can be found in NIST's AIDA Phase 1 evaluation plan (NIST, 2019).

Table 6 below shows the number of system responses assessed for each of the assessment tasks. In all assessment judgements, assessors followed a lenient assessment standard, giving credit for system responses that were not perfect (e.g., justification span is inexact but contains an appropriate mention).

Assessment Task	System Responses Assessed
Class	8859
Zero-hop	5978
Graph	16,854
Hypotheses	9949

Table 6: System responses assessed

6. Conclusion

The AIDA Phase 1 corpus makes a unique contribution to the available resources for multilingual, multimedia information extraction, with a particular emphasis on the detection of conflicting information. The corpus is currently available within the AIDA program, and will be published in the LDC catalog once program needs permit it to be shared publicly.

7. Acknowledgments

This material is based upon work supported by Air Force Research Laboratory (AFRL) and the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-18-C-0013. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL or DARPA.

8. Bibliographical References

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., and von Sikorski, C. (2021). Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, 98(3):641-664.

DARPA. (2012). Broad Agency Announcement: Deep Exploration and Filtering of Text (DEFT). Defense

Advanced Research Projects Agency, DARPA-BAA-12-47.

DARPA. (2017). Active Interpretation of Disparate Alternatives (AIDA). Defense

Advanced Research Projects Agency, DARPA BAA HR001117S0026.

Grabar, N. and Hamon, T. (2019). WikiWars-UA: Ukrainian corpus annotated with temporal expressions. In *Proceedings of Computational Linguistics and Intelligent Systems (COLINS 2019)*, Kharkiv, Ukraine, April

Grabar, N., Kanishcheva, O., and Hamon, T. (2018). Multilingual aligned corpus with Ukrainian as the target language. In *Proceedings of SlaviCorp 2018,* Prague, Czech Republic, September

Gottlieb, M., and Dyer, S. (2020). Information and Disinformation: Social Media in the COVID-19 Crisis. *Academic Emergency Medicine*, 27(7):640-641.

Han, Q., John, M., Kurzhals, K., Messner, J., and Ertl, T. (2018). Visual interactive labeling of large multimedia news corpora. In *Proceedings of Leipzig Symposium on Visualization in Applications (LEVIA'18),* Leipzig, Germany, October.

Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in context. https://arxiv.org/abs/1405.0312v3, accessed 1/13/2022.

NIST. (2019). Streaming Multimedia Knowledge Base Population (SM-KBP) 2019. Text Analysis Conference. https://tac.nist.gov/2019/SM-KBP/index.html

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., and Liu, H. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6), e1385.

Shvedova, M., (2020). The general regionally annotated corpus of Ukrainian (GRAC, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020).* Lviv, Ukraine, April, pages 489-506.

Sitchinava, D. (2012). Parallel corpora within the Russian National Corpus. *Philological Studies*, 63: 271-278.

Szostek, J. (2018). Nothing is true? The credibility of news and conflicting narratives during “information war” in Ukraine. *The International Journal of Press/Politics*, 23(1):116–135.

Tracey, J. and Strassel, S. (2020). Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference (LREC 2020)*, pages 277–284, May. European Language Resource Association (ELRA).