GAIA: A Fine-grained Multimedia Knowledge Extraction System

Manling Li[1]*, **Alireza Zareian**[2]*, **Ying Lin**[1], **Xiaoman Pan**[1], **Spencer Whitehead**[1],
Brian Chen[2], **Bo Wu**[2], **Heng Ji**[1], **Shih-Fu Chang**[2]
Clare Voss[3], **Daniel Napierski**[4], **Marjorie Freedman**[4]
[1]University of Illinois at Urbana-Champaign [2]Columbia University
[3]US Army Research Laboratory [4]Information Sciences Institute
{manling2, hengji}@illinois.edu, {az2407, sc250}@columbia.edu

Abstract

We present the first comprehensive, open source multimedia knowledge extraction system that takes a massive stream of unstructured, heterogeneous multimedia data from various sources and languages as input, and creates a coherent, structured knowledge base, indexing entities, relations, and events, following a rich, fine-grained ontology. Our system, GAIA [1], enables seamless search of complex graph queries, and retrieves multimedia evidence including text, images and videos. GAIA achieves top performance at the recent NIST TAC SM-KBP2019 evaluation[2]. The system is publicly available at GitHub[3] and DockerHub[4], with complete documentation[5].

1 Introduction

Knowledge Extraction (KE) aims to find entities, relations and events involving those entities from unstructured data, and link them to existing knowledge bases. Open source KE tools are useful for many real-world applications including disaster monitoring (Zhang et al., 2018a), intelligence analysis (Li et al., 2019a) and scientific knowledge mining (Luan et al., 2017; Wang et al., 2019). Recent years have witnessed the great success and wide usage of open source Natural Language Processing tools (Manning et al., 2014; Fader et al., 2011; Gardner et al., 2018; Daniel Khashabi, 2018; Honnibal and Montani, 2017), but there is no comprehensive open source system for KE. We release

Figure 1: An example of cross-media knowledge fusion and a look inside the visual knowledge extraction.

a new comprehensive KE system, GAIA, that advances the state of the art in two aspects: (1) it extracts and integrates knowledge across multiple languages and modalities, and (2) it classifies knowledge elements into fine-grained types, as shown in Table 1. We also release the pretrained models[6] and provide a script to retrain it for any ontology.

GAIA has been inherently designed for multimedia, which is rapidly replacing text-only data in many domains. We extract complementary knowledge from text as well as related images or video frames, and integrate the knowledge across modalities. Taking Figure 1 as an example, the text entity extraction system extracts the nominal mention *troops*, but is unable to link or relate that due to a vague textual context. From the image, the entity linking system recognizes the flag as Ukrainian and represents it as a *NationalityCitizen* relation in the knowledge base. It can be deduced, although not for sure, that the detected people are Ukrainian. Meanwhile, our cross-media fusion system grounds the *troops* to the people detected in the image. This establishes a connection between the knowledge

*These authors contributed equally to this work.
[1]System page: http://blender.cs.illinois.edu/software/gaia-ie
[2]http://tac.nist.gov/2019/SM-KBP/index.html
[3]GitHub: https://github.com/GAIA-AIDA
[4]DockerHub: text knoweldge extraction components are in https://hub.docker.com/orgs/blendernlp/repositories, visual knowledge extraction components are in https://hub.docker.com/u/dannapierskitoptal
[5]Video: http://blender.cs.illinois.edu/aida/gaia.mp4
[6]Pretrained models: http://blender.cs.illinois.edu/resources/gaia.html

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 77–86
July 5 - July 10, 2020.

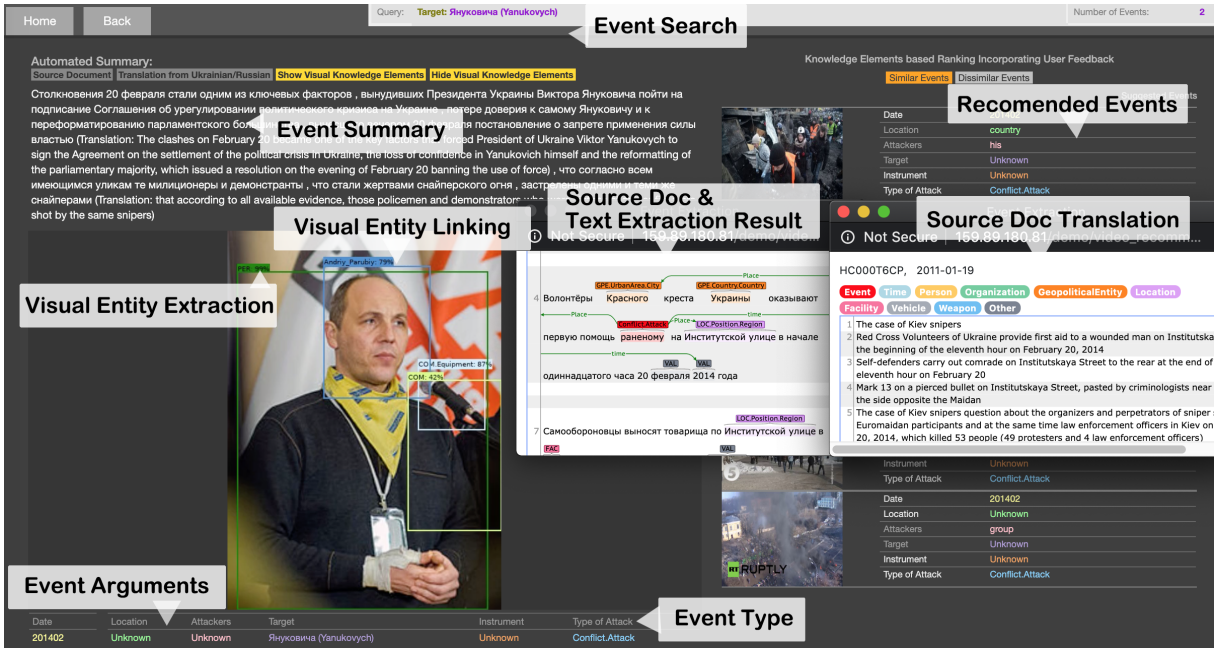

Figure 2: User-facing views of knowledge networks constructed with events automatically extracted from multimedia multilingual news reports. We display the event arguments, type, summary, similar events, as well as visual knowledge extracted from the corresponding image and video.

extracted from the two modalities, allowing to infer that the *troops* are Ukrainian, and *They* refers to the Ukrainian government.

Compared to coarse-grained event types of previous work (Li et al., 2019a), we follow a richer ontology to extract fine-grained types, which are crucial to scenario understanding and event prediction. For example, an event of type *Movement.TransportPerson* involving an entity of type *PER.Politician.HeadOfGovernment* differs in implications from the same event type involving a *PER.Combatant.Sniper* entity (*i.e.*, a political trip versus a military deployment). Similarly, it is far more likely that an event of type *Conflict.Attack.Invade* will lead to a *Contact.Negotiate.Meet* event, while a *Conflict.Attack.Hanging* event is more likely to be followed by an event of type *Contact.FuneralVigil.Meet*.

	Coarse-grained Types	Fine-grained Types
Entity	7	187
Relation	23	61
Event	47	144

Table 1: Compared to the coarse-grained knowledge extraction of previous work, GAIA can support fine-grained entity, relation, and event extraction with types that are a superset of the previous coarse-grained types.

The knowledge base extracted by GAIA can support various applications, such as multimedia news event understanding and recommendation. We use Russia-Ukraine conflicts of 2014-2015 as a case study, and develop a knowledge exploration interface that recommends events related to the user's ongoing search based on previously-selected attribute values and dimensions of events being viewed[7], as shown in Figure 2. Thus, this system automatically provides the user with a more comprehensive exposure to collected events, their importance, and their interconnections. Extensions of this system to real-time applications would be particularly useful for tracking current events, providing alerts, and predicting possible changes, as well as topics related to ongoing incidents.

2 Overview

The architecture of our multimedia knowledge extraction system is illustrated in Figure 3. The system pipeline consists of a Text Knowledge Extraction (TKE) branch and a Visual Knowledge Extraction (VKE) branch (Sections 3 and 4 respectively). Each branch takes the same set of documents as input, and initially creates a separate knowledge base (KB) that encodes the information from its respec-

[7]Event recommendation demo: http://blender.cs.illinois.edu/demo/video_recommendation/index_attack_dark.html

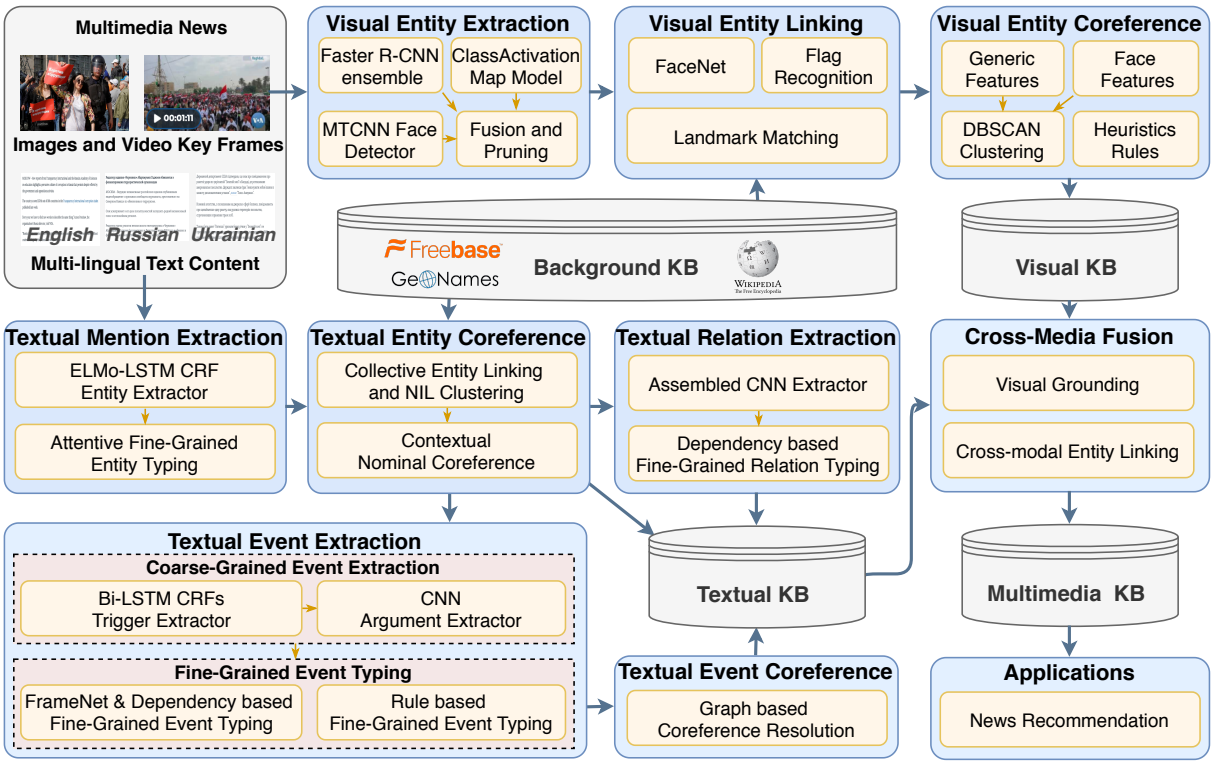

Figure 3: The architecture of GAIA multimedia knowledge extraction.

tive modality. Both output knowledge bases make use of the same types from the DARPA AIDA ontology[8], as referred to in Table 1. Therefore, while the branches both encode their modality-specific extractions into their KBs, they do so with types defined in the same semantic space. This shared space allows us to fuse the two KBs into a single, coherent multimedia KB through the Cross-Media Knowledge Fusion module (Section 5). Our user-facing system demo accesses one such resulting KB, where attack events have been extracted from multi-media documents related to the 2014-2015 Russia-Ukraine conflict scenario. In response to user queries, the system recommends information around a primary event and its connected events from the knowledge graph (screenshot in Figure 2).

3 Text Knowledge Extraction

As shown in Figure 3, the Text Knowledge Extraction (TKE) system extracts entities, relations, and events from input documents. Then it clusters identical entities through entity linking and coreference, and clusters identical events using event coreference.

[8] https://tac.nist.gov/tracks/SM-KBP/2019/ontologies/LDCOntology

3.1 Text Entity Extraction and Coreference

Coarse-grained Mention Extraction We extract coarse-grained named and nominal entity mentions using a LSTM-CRF (Lin et al., 2019) model. We use pretrained ELMo (Peters et al., 2018) word embeddings as input features for English, and pre-train Word2Vec (Le and Mikolov, 2014) models on Wikipedia data to generate Russian and Ukrainian word embeddings.

Entity Linking and Coreference We seek to link the entity mentions to pre-existing entities in the background KBs (Pan et al., 2015), including Freebase (LDC2015E42) and GeoNames (LDC2019E43). For mentions that are linkable to the same Freebase entity, coreference information is added. For name mentions that cannot be linked to the KB, we apply heuristic rules (Li et al., 2019b) to same-named mentions within each document to form NIL clusters. A NIL cluster is a cluster of entity mentions referring to the same entity but do not have corresponding KB entries (Ji et al., 2014).

Fine-grained Entity Typing We develop an attentive fine-grained type classification model with latent type representation (Lin and Ji, 2019). It takes as input a mention with its context sentence and predicts the most likely fine-grained type. We obtain the YAGO (Suchanek et al., 2008) fine-grained

types from the results of Freebase entity linking, and map these types to the DARPA AIDA ontology. For mentions with identified, coarse-grained GPE and LOC types, we further determine their fine-grained types using GeoNames attributes *feature_class* and *feature_code* from the GeoNames entity linking result. Given that most nominal mentions are descriptions and thus do not link to entries in Freebase or GeoNames, we develop a nominal keyword list (Li et al., 2019b) for each type to incorporate these mentions into the entity analyses.

Entity Salience Ranking To better distill the information, we assign each entity a salience score in each document. We rank the entities in terms of the weighted sum of all mentions, with higher weights for name mentions. If one entity appears only in nominal and pronoun mentions, we reduce its salience score so that it is ranked below other entities with name mentions. The salience score is normalized over all entities in each document.

3.2 Text Relation Extraction

For fine-grained relation extraction, we first apply a language-independent CNN based model (Shi et al., 2018) to extract coarse-grained relations from English, Russian and Ukrainian documents. Then we apply entity type constraints and dependency patterns to these detected relations and re-categorize them into fine-grained types (Li et al., 2019b). To extract dependency paths for these relations in the three languages, we run the corresponding language's Universal Dependency parser (Nivre et al., 2016). For types without coarse-grained type training data in ACE/ERE, we design dependency path-based patterns instead and implement a rule-based system to detect their fine-grained relations directly from the text (Li et al., 2019b).

3.3 Text Event Extraction and Coreference

We start by extracting coarse-grained events and arguments using a Bi-LSTM CRF model and a CNN-based model (Zhang et al., 2018b) for three languages, and then detect the fine-grained event types by applying verb-based rules, context-based rules, and argument-based rules (Li et al., 2019b). We also extract FrameNet frames (Chen et al., 2010) in English corpora to enrich the fine-grained events.

We apply a graph-based algorithm (Al-Badrashiny et al., 2017) for our language-independent event coreference resolution. For each event type, we cast the event mentions as nodes in a graph, so that the undirected, weighted edges between these nodes represent coreference confidence scores between their corresponding events. We then apply hierarchical clustering to obtain event clusters and train a Maximum Entropy binary classifier on the cluster features (Li et al., 2019b).

4 Visual Knowledge Extraction

The Visual Knowledge Extraction (VKE) branch of GAIA takes images and video key frames as input and creates a single, coherent (visual) knowledge base, relying on the same ontology as GAIA's Text Knowledge Extraction (TKE) branch. Similar to TKE, the VKE consists of entity extraction, linking, and coreference modules. Our VKE system also extracts some events and relations.

4.1 Visual Entity Extraction

We use an ensemble of visual object detection and concept localization models to extract entities and some events from a given image. To detect generic objects such as person and vehicle, we employ two off-the-shelf Faster R-CNN models (Ren et al., 2015) trained on the Microsoft Common Objects in COntext (MS COCO) (Lin et al., 2014) and Open Images (Kuznetsova et al., 2018) datasets. To detect scenario-specific entities and events, we train a Class Activation Map (CAM) model (Zhou et al., 2016) in a weakly supervised manner using a combination of Open Images with image-level labels and Google image search.

Given an image, each R-CNN model produces a set of labeled bounding boxes, and the CAM model produces a set of labeled heat maps which are then thresholded to produce bounding boxes. The union of all bounding boxes is then post-processed by a set of heuristic rules to remove duplicates and ensure quality. We separately apply a face detector, MTCNN (Zhang et al., 2016), and add the results to the pool of detected objects as additional *person* entities. Finally, we represent each detected bounding box as an entity in the visual knowledge base. Since the CAM model includes some event types, we create event entries (instead of entity entries) for bounding boxes classified as events.

4.2 Visual Entity Linking

Once entities are added into the (visual) knowledge base, we try to link each entity to the real-world entities from a curated background knowledge base. Due to the complexity of this task, we develop distinct models for each coarse-grained entity type.

Face Recognition (a) Landmark Recognition (b) Flag Recognition (c)

Figure 4: Examples of visual entity linking, based on face recognition, landmark recognition and flag recognition.

For the type *person*, we train a FaceNet model (Schroff et al., 2015) that takes each cropped human face (detected by the MTCNN model as mentioned in Section 4.1) and classifies it in one or none of the predetermined identities. We compile a list of recognizable and scenario-relevant identities by automatically searching for each person name in the background KB via Google Image Search, collecting top retrieved results that contain a face, training a binary classifier on half of the results, and evaluating on the other half. If the accuracy is higher than a threshold, we include that person name in our list of recognizable identities. For example, the visual entity in Figure 4 (a) is linked to the Wikipedia entry *Rudy Giuliani* [9].

To recognize *location*, *facility*, and *organization* entities, we use a DELF model (Noh et al., 2017) pre-trained on Google Landmarks, to match each image with detected buildings against a predetermined list. We use a similar approach as mentioned above to create a list of recognizable, scenario-relevant landmarks, such as buildings and other types of structure that identify a specific location, facility, or organization. For example, the visual entity in Figure 4 (b) is linked to the Wikipedia entry *Maidan Square* [10]

Finally, to recognize *geopolitical* entities, we train a CNN to classify flags into a predetermined list of entities, such as all the nations in the world, for detection in our system. Take Figure 4 (c) as an example. The flags of *Ukraine*, *US* and *Russia* are linked to the Wikipedia entries of corresponding countries. Once a flag in an image is recognized, we apply a set of heuristic rules to create a nationality affiliation relationship in the knowledge base between some entities in the scene and the detected country. For instance, a person who is holding a Ukrainian flag would be affiliated with the country Ukraine.

[9] https://en.wikipedia.org/wiki/Rudy_Giuliani

[10] https://en.wikipedia.org/wiki/Maidan_Nezalezhnosti

4.3 Visual Entity Coreference

While we cast each detected bounding box as an entity node in the output knowledge base, we resolve potential coreferential links between them, since one unique real-world entity can be detected multiple times. Cross-image coreference resolution aims to identify the same entity appearing in multiple images, where the entities are in different poses from different angles. Take Figure 5 as an example. The red bounding boxes in these two images refer to the same person, so they are coreferential and are put into the same NIL cluster. Within-image coreference resolution requires the detection of duplicates, such as the duplicates in an collage image. To resolve entity coreference, we train an instance-matching CNN on the Youtube-BB dataset (Real et al., 2017), where we ask the model to match an object bounding box to the same object in a different video frame, rather than to a different object. We use this model to extract features for each detected bounding box and run the DBSCAN (Ester et al., 1996) clustering algorithm on the box features across all images. The entities in the same cluster are coreferential, and are represented using a NIL cluster in the output (visual) KB. Similarly, we use a pretrained FaceNet (Schroff et al., 2015) model followed by DBSCAN to cluster face features.

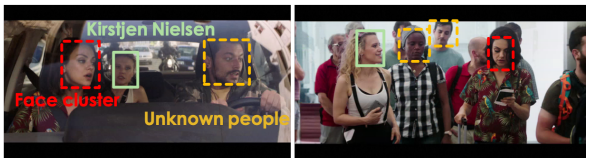

Figure 5: The two green bounding boxes are coreferential since they are both linked to "Kirstjen Nielsen", and two red bounding boxes are coreferential based on face features. The yellow bounding boxes are unlinkable and also not coreferential to other bounding boxes.

We also define heuristic rules to complement the aforementioned procedure in special cases. For example, if in the entity linking process (Section 4.2), some entities are linked to the same real-world entity based on entity linking result, we consider them coreferential. Besides, since we have both face detection and person detection which result in two entities for each person instance, we use their bounding box intersection to merge them into the same entity.

5 Cross-Media Knowledge Fusion

Given a set of multimedia documents which consist of textual data, such as written articles and transcribed speech, as well as visual data, such as images and video key frames, the TKE and VKE branches of the system take their respective modality data as input, extract knowledge elements, and create separate knowledge bases. These textual and visual knowledge bases rely on the same ontology, but contain complementary information. Some knowledge elements in a document may not be explicitly mentioned in the text, but will appear visually, such as the Ukrainian flag in Figure 1. Even coreferential knowledge elements that exist in both knowledge bases are not completely redundant, since each modality has its own unique granularity. For example, the word *troops* in text could be considered coreferential to the individuals with military uniform detected in the image, but the uniforms being worn may provide additional visual features useful in identifying the military ranks, organizations and nationalities of the individuals.

To exploit the complementary nature of the two modalities, we combine the two modality-specific knowledge bases into a single, coherent, multimedia knowledge base, where each knowledge element could be grounded in either or both modalities. To fuse the two bases, we develop a state-of-the-art visual grounding system (Akbari et al., 2019) to resolve entity coreference across modalities. More specifically, for each entity mention extracted from text, we feed its text along with the whole sentence into an ELMo model (Peters et al., 2018) that extracts contextualized features for the entity mention, and then we compare that with CNN feature maps of surrounding images. This leads to a relevance score for each image, as well as a granular relevance map (heatmap) within each image. For images that are relevant enough, we threshold the heatmap to obtain a bounding box, compare that box content with known visual entities, and assign it to the entity with the most overlapping match. If no overlapping entity is found, we create a new visual entity with the heatmap bounding box. Then we link the matching textual and visual entities using a NIL cluster. Additionally, with visual linking (Section 4.2), we corefer cross-modal entities that are linked to the same background KB node.

Component			Benchmark	Metric	Score
Mention Extraction			CoNLL-2003	F_1	91.8%
Relation Extraction		English	ACE&ERE	F_1	65.6%
		Russian	AIDA	F_1	72.4%
		Ukrainian	AIDA	F_1	68.2%
Event Extraction	En	Trigger	ERE	F_1	65.4%
		Argument	ERE	F_1	85.0%
	Ru	Trigger	AIDA	F_1	56.2%
		Argument	AIDA	F_1	58.2%
	Uk	Trigger	AIDA	F_1	59.0%
		Argument	AIDA	F_1	61.1%
Visual Entity Extraction		Objects	MSCOCO	mAP	43.0%
		Faces	FDDB	Acc	95.4%
Visual Entity Linking		Faces	LFW	Acc	99.6%
		Landmarks	Oxf105k	mAP	88.5%
		Flags	AIDA	F_1	72.0%
Visual Entity Coreference			YoutubeBB	Acc	84.9%
Crossmedia Coreference			Flickr30k	Acc	69.2%

Table 2: Performance of each component. The benchmarks references are: CoNLL-2003 (Sang and De Meulder, 2003), ACE (Walker et al., 2006), ERE (Song et al., 2015), AIDA (LDC2018E01:AIDA Seedling Corpus V2.0), MSCOCO (Lin et al., 2014), FDDB (Jain and Learned-Miller, 2010), LFW (Huang et al., 2008), Oxf105k (Philbin et al., 2007), YoutubeBB (Real et al., 2017), and Flickr30k (Plummer et al., 2015).

6 Evaluation

6.1 Quantitative Performance

The performance of each component is shown in Table 2. To evaluate the end-to-end performance, we participated with our system in the TAC SM-KBP 2019 evaluation[11]. The input corpus contains 1999 documents (756 English, 537 Russian, 703 Ukrainian), 6194 images, and 322 videos. We populated a multimedia, multilingual knowledge base with 457,348 entities, 67,577 relations, 38,517 events. The system performance was evaluated based on its responses to *class queries* and *graph queries*[12], and GAIA was awarded first place.

Class queries evaluated cross-lingual, cross-modal, fine-grained entity extraction and coreference, where the query is an entity type, such as *FAC.Building.GovernmentBuilding*, and the result is a ranked list of entities of the given type. Our entity ranking is generated by the entity salience score in Section 3.1. The evaluation metric was

[11] http://tac.nist.gov/2019/SM-KBP/index.html

[12] http://tac.nist.gov/2019/SM-KBP/guidelines.html

Average Precision (AP), where AP-B was the AP score where ties are broken by ranking all Right responses above all Wrong responses, AP-W was the AP score where ties are broken by ranking all Wrong responses above all Right responses, and AP-T was the AP score where ties are broken as in TREC_Eval[13].

Class Queries			Graph Queries		
AP-B	AP-W	AP-T	Precision	Recall	F_1
48.4%	47.4%	47.7%	47.2%	21.6%	29.7%

Table 3: GAIA achieves top performance on Task 1 at the recent NIST TAC SM-KBP2019 evaluation.

Graph queries evaluated cross-lingual, cross-modal, fine-grained relation extraction, event extraction and coreference, where the query is an argument role type of event (e.g., *Victim* of *Life.Die.DeathCausedByViolentEvents*) or relation (e.g., *Parent* of *PartWhole.Subsidiary*) and the result is a list of entities with that role. The evaluation metrics were Precision, Recall and F_1.

6.2 Qualitative Analysis

To demonstrate the system, we have selected Ukraine-Russia Relations in 2014-2015 for a case study to visualize attack events, as extracted from the topic-related corpus released by LDC[14]. The system displays recommended events related to the user's ongoing search based on their previously-selected attribute values and dimensions of events being viewed, such as the fine-grained type, place, time, attacker, target, and instrument. The demo is publicly available[15] with a user interface as shown in Figure 2, displaying extracted text entities and events across languages, visual entities, visual entity linking and coreference results from face, landmark and flag recognition, and the results of grounding text entities to visual entities.

7 Related Work

Existing knowledge extraction systems mainly focus on text (Manning et al., 2014; Fader et al., 2011; Gardner et al., 2018; Daniel Khashabi, 2018; Honnibal and Montani, 2017; Pan et al., 2017; Li et al., 2019a), and do not readily support fine-grained knowledge extraction. Visual knowledge extraction is typically limited to atomic concepts that have distinctive visual features of daily life (Ren et al., 2015; Schroff et al., 2015; Fernández et al., 2017; Gu et al., 2018; Lin et al., 2014), and so lacks more complex concepts, making extracted elements challenging to integrate with text. Existing multimedia systems overlook the connections and distinctions between modalities (Yazici et al., 2018). Our system makes use of a multi-modal ontology with concepts from real-world, newsworthy topics, resulting in a rich cross-modal, as well as intra-modal connectivity.

8 Conclusion

We demonstrate a state-of-the-art multimedia multilingual knowledge extraction and event recommendation system. This system enables the user to readily search a knowledge network of extracted, linked, and summarized complex events from multimedia, multilingual sources (e.g., text, images, videos, speech and OCR).

Acknowledgement

This research is based upon work supported in part by U.S. DARPA AIDA Program No. FA8750-18-2-0014 and KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12476–12486.

Mohamed Al-Badrashiny, Jason Bolton, Arun Tejasvi Chaganty, Kevin Clark, Craig Harman, Lifu Huang, Matthew Lamm, Jinhao Lei, Di Lu, Xiaoman Pan, et al. 2017. Tinkerbell: Cross-lingual cold-start knowledge base construction. In *TAC*.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A Smith. 2010. Semafor: Frame argument resolution with log-linear models. In *Proceedings of*

[13] https://trec.nist.gov/trec_eval/

[14] LDC2018E01, LDC2018E52, LDC2018E63, LDC2018E76, LDC2019E77

[15] http://blender.cs.illinois.edu/demo/video_recommendation/index_attack_dark.html

the 5th international workshop on semantic evaluation, pages 264–267. Association for Computational Linguistics.

Ben Zhou Tom Redman Christos Christodoulopoulos Vivek Srikumar Nicholas Rizzolo Lev Ratinov Guanheng Luo Quang Do Chen-Tse Tsai Subhro Roy Stephen Mayhew Zhili Feng John Wieting Xiaodong Yu Yangqiu Song Shashank Gupta Shyam Upadhyay Naveen Arivazhagan Qiang Ning Shaoshi Ling Dan Roth Daniel Khashabi, Mark Sammons. 2018. Cogcompnlp: Your swiss army knife for nlp. In *11th Language Resources and Evaluation Conference*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK.

Delia Fernández, David Varas, Joan Espadaler, Issey Masuda, Jordi Ferreira, Alejandro Woodward, David Rodríguez, Xavier Giró-i Nieto, Juan Carlos Riveiro, and Elisenda Bou. 2017. Vits: video tagging system from massive web multimedia collections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 337–346.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

Vidit Jain and Erik Learned-Miller. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report.

Heng Ji, Joel Nothman, Ben Hachey, et al. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*, pages 1333–1339.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Manling Li, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. 2019a. Multilingual entity, relation, event and human value extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 110–115.

Manling Li, Ying Lin, Ananya Subburathinam, Spencer Whitehead, Xiaoman Pan, Di Lu, Qingyun Wang, Tongtao Zhang, Lifu Huang, Heng Ji, Alireza Zareian, Hassan Akbari, Brian Chen, Bo Wu, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Yixiang Yao, Jennifer Chen, Eric Berquist, Kexuan Sun, Xujun Peng, Ryan Gabbard, Marjorie Freedman, Pedro Szekely, T.K. Satish Kumar, Arka Sadhu, Ram Nevatia, Miguel Rodriguez, Yifan Wang, Yang Bai, Ali Sadeghian, and Daisy Zhe Wang. 2019b. Gaia at sm-kbp 2019 - a multi-media multi-lingual knowledge extraction and hypothesis generation system. In *Proceedings of TAC KBP 2019, the 26th International Conference on Computational Linguistics: Technical Papers*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ying Lin and Heng Ji. 2019. An attentive fine-grained entity typing model with latent type representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6198–6203.

Ying Lin, Liyuan Liu, Heng Ji, Dong Yu, and Jiawei Han. 2019. Reliability-aware dynamic feature composition for name tagging. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 165–174.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 1130–1139.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc. the 55th Annual Meeting of the Association for Computational Linguistics*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. 2017. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*.

Adnan Yazici, Murat Koyuncu, Turgay Yilmaz, Saeid Sattari, Mustafa Sert, and Elvan Gulen. 2018. An intelligent multimedia information system for multimodal content extraction and querying. *Multimedia Tools and Applications*, 77(2):2225–2260.

Boliang Zhang, Ying Lin, Xiaoman Pan, Di Lu, Jonathan May, Kevin Knight, and Heng Ji. 2018a. Elisa-edl: A cross-lingual entity extraction, linking and localization system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 41–45.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Tongtao Zhang, Ananya Subburathinam, Ge Shi, Lifu Huang, Di Lu, Xiaoman Pan, Manling Li, Boliang Zhang, Qingyun Wang, Spencer Whitehead, Heng

Ji, Alireza Zareian, Hassan Akbari, Brian Chen, Ruiqi Zhong, Steven Shao, Emily Allaway, Shih-Fu Chang, Kathleen McKeown, Dongyu Li, Xin Huang, Kexuan Sun, Xujun Peng, Ryan Gabbard, Marjorie Freedman, Mayank Kejriwal, Ram Nevatia, Pedro Szekely, T.K. Satish Kumar, Ali Sadeghian, Giacomo Bergami, Sourav Dutta, Miguel Rodriguez, and Daisy Zhe Wang. 2018b. Gaia - a multi-media multi-lingual knowledge extraction and hypothesis generation system. In *Proceedings of TAC KBP 2018, the 25th International Conference on Computational Linguistics: Technical Papers*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.