Technical Summary: GAIA: A Fine-grained Multimedia Knowledge Extraction System

Kai-Yu Lu

2025/10/19

1 Research Problem and Motivation

Knowledge extraction aims to identify entities, relations, and events from unstructured data and to link these elements to structured knowledge bases. The paper addresses the absence of a comprehensive open source system that can perform this task in a multimedia and multilingual setting with fine grained semantic types.

Existing open source natural language processing toolkits concentrate on textual inputs and typically provide only coarse grained entity and event types. At the same time, real world information increasingly appears as heterogeneous multimedia streams that combine written articles, transcribed speech, images, and video. Many important details are expressed visually rather than textually, for example national flags, military uniforms, or recognizable buildings. Purely text based systems cannot exploit these signals.

In addition, previous knowledge extraction systems usually rely on relatively small type inventories for entities, relations, and events. Coarse schemas cannot distinguish, for instance, between a head of government and a combatant sniper, or between different subtypes of attack events. For downstream applications such as scenario understanding, event prediction, and news recommendation, this level of granularity is insufficient.

The research problem considered in this work is therefore to design and implement a multimedia, multilingual knowledge extraction system that integrates textual and visual information and assigns fine grained types within a rich ontology, while remaining practical enough to be evaluated in a large scale benchmark and released as open source software.

2 Related Work

Previous work on knowledge extraction has produced a variety of text focused systems. General purpose natural language processing pipelines such as Stanford CoreNLP, Open Information Extraction tools, and broader frameworks for information extraction provide named entity recognition, relation extraction, and event extraction from monolingual text. These systems have enabled applications in disaster monitoring, intelligence analysis, and scientific knowledge mining, but remain limited to textual modalities and usually employ coarse type schemas.

Visual recognition research has produced strong models for object detection, face recognition, and landmark recognition. Representative examples include region based convolutional networks for object detection, face embedding models for face verification and clustering, and local feature based models for landmark retrieval. These methods excel at detecting atomic visual concepts such as people, vehicles, faces, or specific buildings, but generally do not connect visual elements to structured event schemas or textual mentions.

Multimedia information systems have explored extraction and querying across multiple modalities. However, existing systems frequently treat each modality independently and do not emphasize cross modal alignment of entities and events under a shared ontology. They typically focus on low level concepts and do not provide fine grained, ontology driven event and relation types.

The work under review builds upon these strands by combining strong text and vision components into a unified architecture that shares a common ontological space and explicitly targets cross modal grounding and coreference.

3 Dataset Construction

3.1 Multimedia Evaluation Corpus

The central evaluation takes place in the TAC SM–KBP 2019 benchmark. The input corpus consists of multilingual multimedia news documents. The corpus statistics are as follows.

- Text documents: 1999 in total, including 756 English, 537 Russian, and 703 Ukrainian documents.
- Images: 6194 images associated with the documents.
- Videos: 322 videos, processed via key frame extraction.

From this corpus, the system constructs a multimedia knowledge base containing 457,348 entities, 67,577 relations, and 38,517 events. These extracted elements follow the DARPA AIDA ontology, which defines fine grained types for entities, relations, and events.

3.2 Component Training and Benchmark Datasets

To train and evaluate individual components, the system uses a comprehensive set of established benchmarks covering both textual and visual tasks. The main datasets and their roles are summarized in Table 1.

The paper does not report detailed pre processing steps for all resources, but indicates that the visual branch operates on images and video key frames, and that textual inputs include written articles and transcribed speech from the benchmark corpora.

3.3 Pretrained Models and Ontology

The system adopts the DARPA AIDA ontology as a unified type space. The ontology provides the following approximate type counts.

- Entity types: 187 fine grained types, subsuming 7 coarse grained categories.
- Relation types: 61 fine grained types, subsuming 23 coarse grained categories.
- Event types: 144 fine grained types, subsuming 47 coarse grained categories.

Pretrained models, including textual encoders and visual detectors, are made available together with scripts that enable retraining for alternative ontologies. The paper does not introduce explicit mathematical formulas in describing the ontology, and instead relies on textual specification and type mappings.

4 Query Protocol and Task Definitions

4.1 Knowledge Extraction Tasks

The system performs several tightly coupled extraction tasks.

- **Entity extraction and typing**: identification of named and nominal entity mentions in text, recognition of visual entities in images and video frames, and assignment of fine grained entity types.
- **Entity linking and coreference**: linking textual and visual entities to external knowledge bases where possible, clustering unlinked mentions into NIL clusters that represent new entities, and resolving coreference within and across modalities.
- **Relation extraction**: detection and typing of relations between pairs of entities in text, with extension to visual relations in some cases.
- **Event extraction and typing**: identification of event triggers and arguments in text, assignment of fine grained event types, and clustering of co referring event mentions.
- **Cross modal grounding**: alignment of textual mentions with visual regions that refer to the same real world entities, enabling cross modal coreference and joint representation in the multimedia knowledge base.

These tasks are performed under a shared ontology, which guarantees that entities, relations, and events extracted from different languages and modalities are represented in the same semantic space.

4.2 TAC SM–KBP 2019 Evaluation Protocol

The end to end system is evaluated through two types of queries defined in the TAC SM–KBP 2019 evaluation.

Class queries Class queries assess cross lingual, cross modal, fine grained entity extraction and coreference. A class query specifies an entity type, such as a specific subtype of government building. The system returns a ranked list of entities of that type, drawn from the constructed knowledge base. Ranking is based on entity salience scores, which summarize the importance of each entity within the documents.

The primary evaluation metric is Average Precision (AP). The evaluation defines three variants.

- AP–B ranks all correct responses above all incorrect responses in case of ties.
- AP–W ranks all incorrect responses above all correct responses in case of ties.
- AP–T resolves ties using the standard TREC evaluation procedure.

Average Precision measures how well the ranked list places relevant entities near the top. It can be interpreted as a weighted average of precision values at all positions where a correct answer appears, with higher scores indicating better ranking performance.

Graph queries Graph queries evaluate cross lingual, cross modal, fine grained relation and event extraction, and event coreference. A graph query specifies an argument role type for an event or relation. Examples include the victim role of a death event or the parent role in a subsidiary relation. The system must return all entities that play the requested role in the constructed knowledge base.

The evaluation metrics for graph queries are precision, recall, and F1.

- Precision measures the proportion of returned entities that are correct.
- Recall measures the proportion of all correct entities that the system successfully returns.
- F1 is the harmonic mean of precision and recall and summarizes the trade off between them.

The paper does not introduce explicit mathematical formulas for these metrics, and uses their standard definitions.

5 Modeling Approach

The overall architecture consists of two main branches, Text Knowledge Extraction and Visual Knowledge Extraction, followed by a cross media fusion module that integrates both branches into a single multimedia knowledge base.

5.1 Text Knowledge Extraction

The text branch receives multilingual textual inputs and constructs a textual knowledge base. It contains modules for entity extraction and coreference, relation extraction, and event extraction and coreference.

5.1.1 Text entity extraction and typing

Coarse grained entity mention extraction is performed with a sequence labeling model based on LSTM–CRF.

- A bidirectional Long Short Term Memory network encodes each sentence and captures contextual information around every token.
- A Conditional Random Field layer selects the most consistent sequence of entity labels over the sentence.
- For English, contextual word representations from ELMo are used as input features. For Russian and Ukrainian, Word2Vec embeddings trained on Wikipedia corpora are used.

Entity linking maps each extracted mention to entries in background knowledge bases, specifically Freebase and GeoNames. When multiple mentions are linked to the same knowledge base entity, they are treated as coreferential. For named mentions that cannot be linked, heuristic rules within each document group identical surface forms into NIL clusters that represent new entities not present in existing knowledge bases.

Fine grained entity typing uses an attentive classification model with latent type representations. The model takes an entity mention and its context sentence, then predicts one or more fine grained types. Types obtained via Freebase and YAGO are mapped into the AIDA ontology. For geographical entities, attributes such as feature class and feature code from GeoNames are used to refine fine grained types. Nominal mentions that do not link to knowledge bases are handled via curated keyword lists that map noun phrases to ontology types.

Entity salience ranking assigns each entity a document level importance score. The score aggregates contributions from all mentions of the entity, with higher weights for named mentions than for pronouns or nominal descriptions. Entities mentioned only as nominals or pronouns receive reduced scores. Scores are normalized within each document and used for ranking entities in response to class queries.

5.1.2 Text relation extraction

For relations between textual entities, the system first predicts coarse grained relation types using a convolutional neural network that is designed to be language independent. The model is trained on English, Russian, and Ukrainian data.

To obtain fine grained relation types, the system combines the coarse predictions with constraints and patterns.

- Entity type constraints ensure that only entity pairs with compatible types receive certain relation labels.
- Dependency paths are extracted using Universal Dependencies parsers in each language, and serve as patterns for relation subtypes.
- For relation types that lack training data in ACE or ERE, dependency path patterns are manually designed and applied in a rule based fashion.

This procedure converts coarse relation outputs into fine grained relation instances that are consistent with the ontology.

5.1.3 Text event extraction and coreference

Event extraction proceeds in two stages.

First, coarse grained events and their arguments are detected. For each language, a Bi–LSTM–CRF model and a CNN based model identify event triggers and arguments. Triggers are words or phrases that evoke events, while arguments are entities participating in events with particular roles.

Second, fine grained event types are assigned using a combination of rule sets.

- Verb based rules exploit lexical information from event triggers.
- Context based rules examine nearby words and phrases.
- Argument based rules use the types and roles of event participants.
- FrameNet frames are extracted in English corpora to further enrich event typing.

Event coreference resolution is modeled as graph clustering. For each event type, event mentions are treated as nodes in an undirected graph. Edge weights represent coreference confidence scores between pairs of events. A hierarchical clustering algorithm groups related nodes into candidate clusters. A Maximum Entropy classifier then decides which clusters represent true coreference chains. This procedure yields event clusters corresponding to unique real world events discussed in multiple sentences or documents.

5.2 Visual Knowledge Extraction

The visual branch processes images and video key frames to construct a visual knowledge base. It includes visual entity extraction, visual entity linking, and visual entity coreference.

5.2.1 Visual entity extraction

Visual entity extraction employs an ensemble of object detection and concept localization models.

- Two Faster R–CNN detectors trained on MS COCO and Open Images identify generic objects such as persons and vehicles, producing labeled bounding boxes.
- A Class Activation Map model trained in a weakly supervised manner on Open Images and Google image search data produces heatmaps for scenario specific entities and events. Thresholded heatmaps are converted into additional bounding boxes.
- A separate face detector, MTCNN, detects faces and contributes additional person entities.

The union of bounding boxes from these models is post processed by heuristic rules that remove duplicates and low quality detections. Bounding boxes classified as entities become visual entity nodes. Boxes corresponding directly to event types under the ontology are stored as visual event nodes.

5.2.2 Visual entity linking

Due to the heterogeneity of visual entities, different linking strategies are used for different coarse grained types.

For person entities, FaceNet is trained to classify cropped faces into a set of recognizable identities. The identity list is constructed by retrieving images for candidate person names from the background knowledge base using Google Image Search, filtering these images with a binary classifier, and retaining identities that can be recognized reliably. Faces detected in images are then mapped to these identities or left unlinked.

For locations, facilities, and organizations associated with buildings or other landmarks, a DELF based landmark recognition model is used. The model, pretrained on Google Landmarks, retrieves the closest matching landmark from a curated list of scenario relevant structures. Detected buildings can therefore be linked to entries such as major squares or government buildings.

For geopolitical entities, a convolutional neural network classifies flags into a set of countries. Recognized flags are linked to corresponding country entities in the background knowledge base. Heuristic rules then create nationality affiliation relations between people in the same scene and the detected countries. For example, a person holding a Ukrainian flag is associated with Ukraine.

5.2.3 Visual entity coreference

Visual entity coreference addresses both cross image and within image cases. The goal is to determine when multiple bounding boxes across different images or frames refer to the same real world entity.

An instance matching convolutional neural network is trained on the Youtube–BB dataset. The model learns to match object bounding boxes that correspond to the same physical object across frames and to distinguish them from other objects. For each detected bounding box, feature vectors are extracted and clustered across images using the DBSCAN algorithm. Boxes within the same cluster are considered coreferential and represented as NIL clusters in the visual knowledge base.

Similarly, a pretrained FaceNet model is used to obtain feature vectors for detected faces, and DBSCAN clustering groups faces belonging to the same individual. Additional heuristics refine the clusters.

- If multiple entities are linked to the same real world identity during entity linking, they are treated as coreferential.
- Overlapping bounding boxes from face detection and person detection in the same image are merged into a single entity.

5.3 Cross Media Knowledge Fusion

The cross media fusion module combines the textual and visual knowledge bases into a single multimedia knowledge base. The key operation is cross modal entity coreference via visual grounding.

For each textual entity mention, the system extracts contextualized features using ELMo applied to the sentence containing the mention. Simultaneously, convolutional feature maps are computed for images surrounding the mention in the document. A visual grounding model computes a relevance score between the text features and each image, as well as a spatial relevance map over locations in the image.

Images whose overall relevance exceeds a threshold are selected for grounding. For each such image, the relevance map is thresholded to form a heatmap region, which is converted into a candidate bounding box. The content of this box is compared with existing visual entities in the image by measuring overlap.

- If a high overlap is found with an existing visual entity, the textual mention is aligned with that entity.
- If no suitable visual entity exists, a new visual entity is created with the heatmap bounding box.

Aligned textual and visual entities are grouped into cross modal NIL clusters, representing a single real world entity with evidence from both modalities. In addition, if textual and visual entities are independently linked to the same background knowledge base node, they are also treated as cross modal coreferential.

This fusion step allows the system to leverage complementary information. For instance, a textual mention such as “troops” can be grounded to a visual group of people wearing uniforms under a Ukrainian flag. The visual information enables the system to infer relevant attributes such as nationality that might not be explicitly stated in the text.

5.4 Implementation Details

The paper describes the architectures and training datasets for each component, but does not provide explicit numerical hyperparameters such as learning rates, batch sizes, or numbers of training epochs. The software, pretrained models, and retraining scripts are released via public repositories, which implies that precise configurations can be inspected in the implementation.

6 Empirical Results

6.1 Component Level Performance

Table 2 summarizes the reported performance of major components on their respective benchmarks. The table focuses on representative metrics without exhaustively listing all rows in the original component table.

The results indicate that the underlying models achieve strong performance on standard benchmarks. For example, named entity recognition performance on CoNLL 2003 is in the low ninety percent range in terms of F1, and face recognition on LFW reaches almost perfect accuracy. Cross media grounding accuracy on Flickr30k Entities is lower, reflecting the inherent difficulty of fine grained image–text alignment, but remains substantial.

6.2 End to End System Performance

The end to end multimedia knowledge extraction system is evaluated on TAC SM–KBP 2019 tasks. The primary numeric results are given in Table 3.

For class queries, the Average Precision with TREC based tie breaking is 47.7%. This value summarizes both retrieval quality and ranking quality for fine grained entity instances across languages and modalities.

For graph queries, precision is 47.2% while recall is 21.6%, leading to an F1 score of 29.7%. The precision value indicates that nearly half of returned argument role instances are correct, while the lower recall suggests that many correct instances are not retrieved. Overall, the system achieved the top rank in Task 1 of the TAC SM–KBP 2019 evaluation among participating systems.

6.3 Qualitative Case Study

The paper presents a qualitative case study focused on the 2014–2015 Russia–Ukraine conflict. A user interface displays attack events extracted from the scenario specific corpus, along with associated entities and visual evidence.

The interface allows interactive exploration of events by type, place, time, attacker, target, and instrument. Related events are recommended based on shared attributes and graph connectivity. Visual panes display aligned textual entities, visual entities, face and landmark recognition results, and cross modal grounding. This case study demonstrates how the multimedia knowledge base supports rich interactive analysis and emphasizes the importance of fine grained types.

7 Summary

7.1 Research Goals and Contributions

The primary goals of the work are to construct a multimedia, multilingual knowledge extraction system that

- integrates textual and visual information from heterogeneous sources,
- assigns fine grained entity, relation, and event types under a rich ontology,
- performs entity, relation, and event extraction, linking, and coreference across languages and modalities, and
- supports complex graph queries and multimedia evidence retrieval for real world scenarios.

The main contributions relative to previous work are as follows.

- The system is a comprehensive open source platform that operates on text, images, and video key frames, and that is inherently designed for multimedia streams instead of only text.
- A shared ontology with fine grained types for entities, relations, and events is applied consistently to both textual and visual branches, enabling coherent cross modal representations.
- The cross media fusion module performs visual grounding of textual entity mentions to visual regions and resolves cross modal entity coreference, thereby exploiting complementary information such as flags, uniforms, and landmarks.
- The approach is empirically validated at component and system levels, and achieves the best reported performance in the TAC SM–KBP 2019 evaluation, while also being demonstrated in an interactive event exploration interface.

7.2 Innovation versus Previous Methods

Relative to earlier systems that focus on text only knowledge extraction with coarse type schemas, this work introduces several innovations.

First, it explicitly combines a text knowledge extraction branch and a visual knowledge extraction branch within a unified architecture and ontology. This design moves beyond loosely coupled pipelines and provides a principled framework for multimodal knowledge representation.

Second, it targets fine grained types for entities, relations, and events. For example, it distinguishes between different subtypes of attack events and between heads of government and combatant snipers. This level of granularity is not supported by many earlier systems.

Third, it introduces a cross modal grounding mechanism that aligns textual mentions with visual regions using contextual embeddings and convolutional feature maps. This alignment allows information to flow from visual cues to textual representations and vice versa. A concrete example is the inference of troop nationality from a combination of textual mentions and flag recognition.

Fourth, it develops tailored models and heuristics for visual entity linking to background knowledge bases, including face recognition, landmark recognition, and flag recognition, and integrates these signals into the knowledge base.

7.3 Limitations and Future Directions

The paper also implies several limitations and future research directions.

- **Dependence on background knowledge bases**: The quality of entity linking and fine grained typing is limited by the coverage and bias of external knowledge bases such as Freebase, GeoNames, and Wikipedia. Entities or concepts absent from these resources must be handled as NIL clusters, which reduces the ability to integrate them into global graphs.
- **Language and domain coverage**: The current system targets English, Russian, and Ukrainian, together with news style multimedia data. Extending the approach to additional languages, less resourced domains, or non news genres requires new training data and potentially domain adaptation techniques.
- **System complexity and resource demands**: The architecture combines many modules, including multiple neural models, parsers, and clustering algorithms. Training, tuning, and maintaining such a system demands substantial computational resources and engineering effort. End to end deployment in new environments may be challenging.
- **Bias and fairness considerations**: The training and evaluation datasets are subject to selection biases across geography, language, and demographic attributes. These biases propagate into the models and can lead to uneven performance across subpopulations. In addition, reliance on public knowledge bases introduces their own biases in what entities and events are represented.
- **Privacy and dual use risks**: The ability to link entities across documents, modalities, and time, especially when combined with face recognition and cross modal grounding, raises privacy concerns and potential for harmful surveillance applications. The paper highlights the importance of legal frameworks, ethical guidelines, and algorithmic auditing before applying such technology to sensitive data.

Future work suggested by the discussion includes expanding the system to additional languages and domains, improving robustness in the presence of data biases, strengthening the interpretability of cross

modal grounding decisions, and systematically integrating ethical constraints and auditing mechanisms into the development and deployment process.

7.4 Concluding Remarks

This work presents a comprehensive multimedia multilingual knowledge extraction and event recommendation system. By combining strong text and vision components under a shared fine grained ontology and by performing cross modal grounding, the system constructs a rich knowledge network that supports search, exploration, and recommendation of complex events from heterogeneous sources.

The empirical results demonstrate that the proposed architecture is competitive in formal evaluations and practically useful in interactive analysis scenarios. At the same time, the system exposes important challenges in scaling, generalization, fairness, and privacy that motivate further research in multimodal knowledge extraction and responsible deployment.

Dataset	Modality	Task	Role in GAIA
CoNLL 2003	Text	Named entity recognition	Benchmark for mention extraction in English newswire.
ACE 2005, ERE	Text	Entity, relation, event extraction	Training and evaluation for coarse grained English relation and event models.
AIDA Seedling Corpus	Text	Multilingual entity, relation, event	Training and evaluation for Russian and Ukrainian components.
Wikipedia corpora	Text	Word representation learning	Source for Word2Vec models for Russian and Ukrainian.
YAGO types	Knowledge base	Fine grained typing	Source of fine grained types mapped into the AIDA ontology via entity linking.
Freebase, GeoNames	Knowledge bases	Entity linking	Background knowledge bases for linking textual entities and refining geographic types.
MS COCO, Open Images	Image	Object detection	Training for generic object detectors used in visual entity extraction.
FDDB	Image	Face detection	Evaluation of face detection using MTCNN.
LFW	Image	Face recognition	Evaluation of FaceNet based face recognition.
Google Landmarks, Oxf105k	Image	Landmark retrieval	Training and evaluation of landmark recognition with DELF features.
Youtube–BB	Video	Object tracking and detection	Training instance matching networks for visual entity coreference across frames and images.
Flickr30k Entities	Image–text	Region–phrase grounding	Evaluation of cross modal grounding and coreference.
LDC Russia–Ukraine corpora	Multimedia	Scenario specific news	Case study corpus for event visualization and recommendation in the Russia–Ukraine conflict scenario.

Table 1: Datasets and resources used for training and evaluating GAIA components.

Component	Benchmark	Metric	Score
Text mention extraction (English)	CoNLL 2003	F1	91.8%
Text relation extraction (English)	ACE and ERE	F1	65.6%
Text relation extraction (Russian)	AIDA	F1	72.4%
Text relation extraction (Ukrainian)	AIDA	F1	68.2%
Text event trigger extraction (English)	ERE	F1	65.4%
Text event argument extraction (English)	ERE	F1	85.0%
Text event trigger extraction (Russian)	AIDA	F1	56.2%
Text event trigger extraction (Ukrainian)	AIDA	F1	59.0%
Visual object detection	MS COCO	mean Average Precision	43.0%
Face detection	FDDB	Accuracy	95.4%
Face recognition	LFW	Accuracy	99.6%
Landmark recognition	Oxf105k	mean Average Precision	88.5%
Flag recognition	AIDA	F1	72.0%
Visual entity coreference	Youtube–BB	Accuracy	84.9%
Cross media coreference	Flickr30k Entities	Accuracy	69.2%

Table 2: Representative component level results reported in the GAIA system.

Task	AP–B	AP–W	AP–T	Precision	Recall	F1
Class queries	48.4%	47.4%	47.7%	–	–	–
Graph queries	–	–	–	47.2%	21.6%	29.7%

Table 3: End to end performance of GAIA on TAC SM–KBP 2019.