Carnegie Mellon University
Language Technologies Institute

# OPERA

Operations-oriented Probabilistic Extraction, Reasoning, and Analysis

CMU+ISI
TAC 2020 SM-KBP

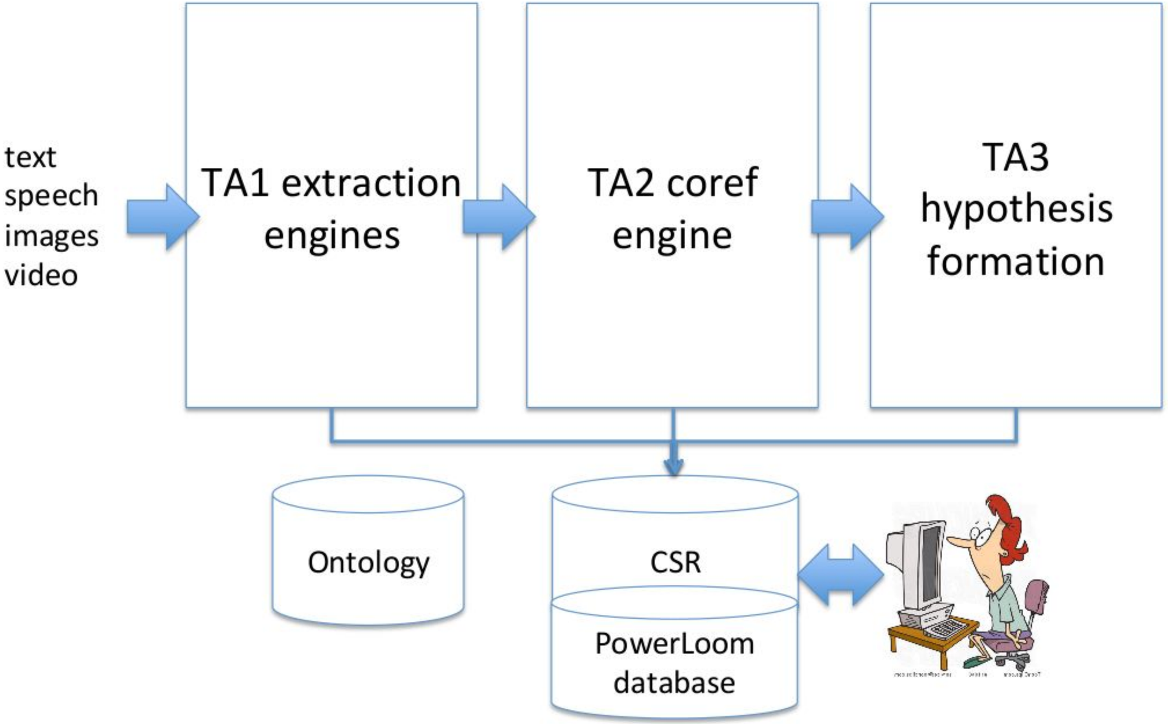
# Team OPERA

- Management:
    - (new) PI: Yonatan Bisk (ybisk@cs.cmu.edu / YonatanBisk.com )
    - Tech leads: Alex Hauptmann, Anatole Gershman, Teruko Mitamura, Hans Chalupsky (ISI)
- TA1
    - Lead: Anatole Gershman, Teruko Mitamura, Alex Hauptmann, Hans Chalupsky
    - Students: Zhisong Zhang, Xiang Kong, Yi-Pei Chen, Siyao Li
    - Knowledge Engineer: Sue Holm
    - Previously: Hector Liu, Vikas Raunak, Tejas Srinivasan, Salvador Medina, Xianyang Chen, Po-Yao (Bernie) Huang, Xuezhe Ma, Ramon Sanabria
- TA2
    - Lead: Hans Chalupsky
    - Students: Varun Gangal
    - Previously: Maria Ryskina
- TA3
    - Lead: Anatole Gershman, Hans Chalupsky
    - Previously: Aditi Chaudhary
- System framework and integration
    - Programmers: **Zaid Sheikh**, Ankit Dangi

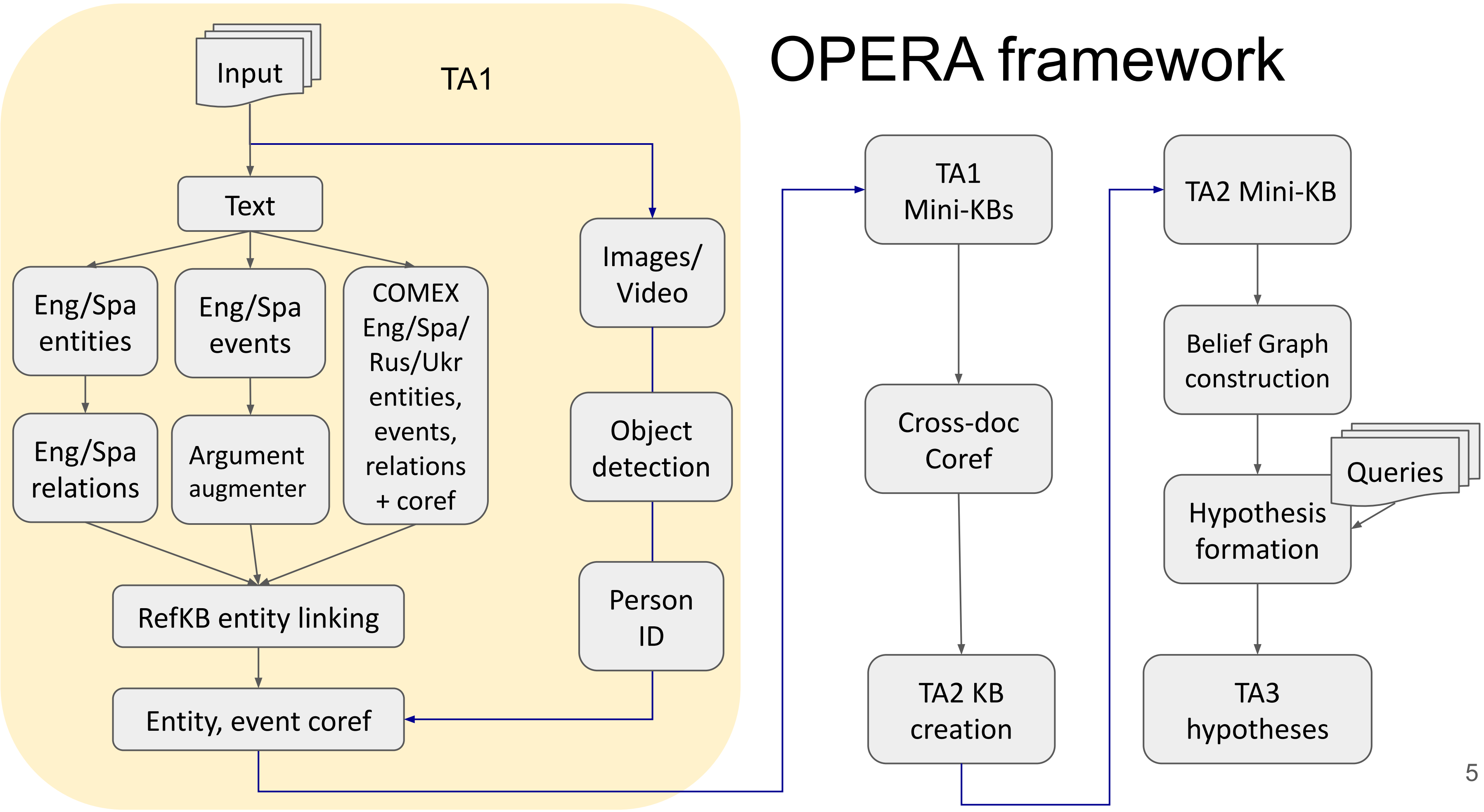
# Overview

- OPERA architecture
- Knowledge-based extraction pipeline (COMEX)
- Entity detection
- Entity linking
- Relation extraction
- Event extraction
- Within-doc coref pipeline (entities and events)
- Visual coreference pipeline
- KB generation (KAgg)
- Dockerization

# OPERA architecture



# OPERA framework



# TAC 2020 SM-KBP TA1 results

| Rank | Team | TypeMetric | ArgumentMetricV1 | ArgumentMetricV2 | CoreferenceMetric | TemporalMetric | FrameMetric |
|---|---|---|---|---|---|---|---|
| 1 | GAIA | 0.2568 | 0.1657 | **0.1256** | **0.3063** | **0.1262** | **0.0933** |
| **2** | **OPERA** | **0.2701** | **0.1698** | 0.1056 | 0.3020 | 0.0552 | 0.0680 |
| 3 | OPERA | 0.2694 | 0.1543 | 0.0991 | 0.2939 | 0.0497 | 0.0645 |
| 4 | OPERA | 0.2659 | 0.1367 | 0.1023 | 0.2757 | 0.0494 | 0.0495 |
| 5 | GAIA | 0.2391 | 0.0758 | 0.0467 | 0.2801 | 0.017 | 0.0478 |
| 6 | BBN | 0.1912 | 0.0743 | 0.0555 | 0.2265 | 0.0049 | 0.0307 |

* consistent improvements over time thanks to the TA1 leaderboard and automatic scoring

# Knowledge-based extraction pipeline

**Universal Dependency Parsing**

**Stanza** **UDPipe** (Spacey)

**Conceptual Mention Extraction (COMEX)**

Ontology, Lexicon, Wikidata, Verbnet

**Linking and Filtering**

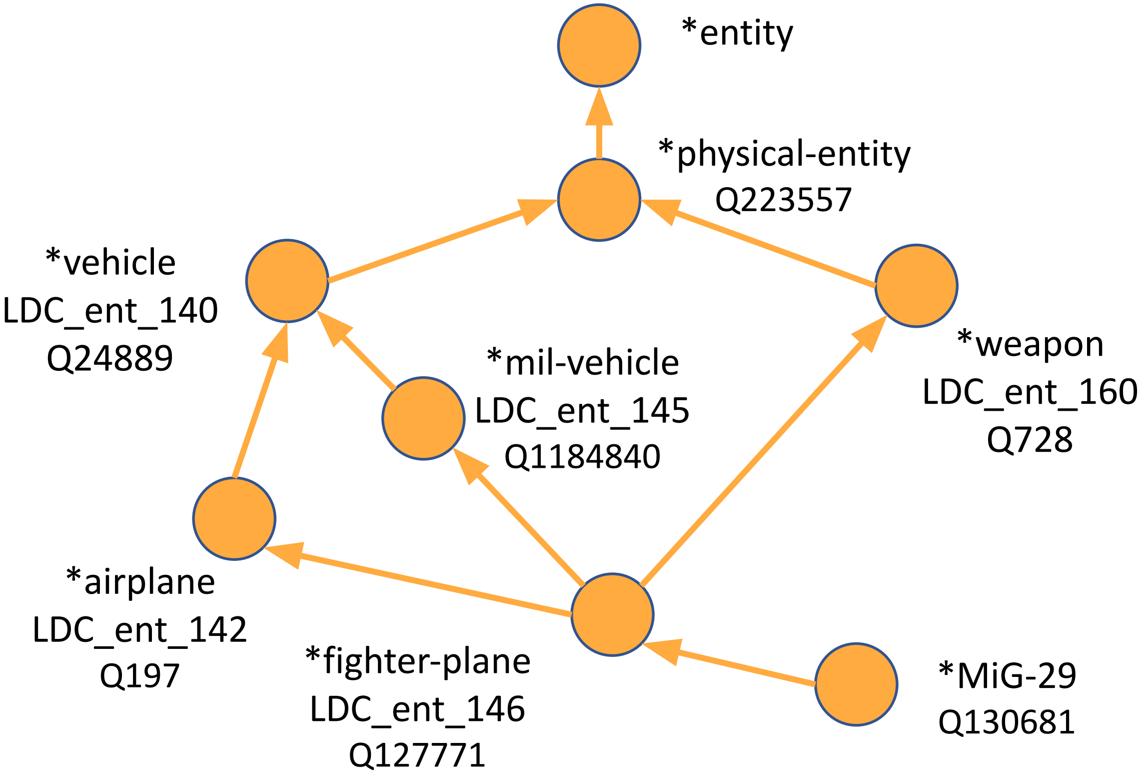
Pronominal coreference, “orphan” deletion

# Conceptual Mentions Extractor (COMEX)

- **Conceptual mention** is a span of text annotated with a concept from the ontology and linked to other mentions via semantic links
- Linked mentions form a mention graph
- COMEX uses lexicons to connect words and phrases in the UD parse tree to the ontological concepts
- Lexical rules are used to connect mentions while semantic constraints come from the ontology
- Lexicons and lexical rules are language-specific

COMEX ontology

- Multiple inheritance
- Superset of LDC AO
- Domain-specific
- Linked to Wikidata
- Compatible with Verbnet selectional restrictions
- Contains instances such as *Maduro and *Caracas




Currently ≈7000 concepts (including ≈6000 entity types imported from Wikidata)

# COMEX lexicons

- English Lexicon ≈ 15,200 items (words and phrases)
- Russian Lexicon ≈ 6,700 items
- Ukrainian Lexicon ≈ 5,000 items
- Spanish Lexicon ≈ 6,600 items

➢ About 1,500 items in each lexicon were manually constructed

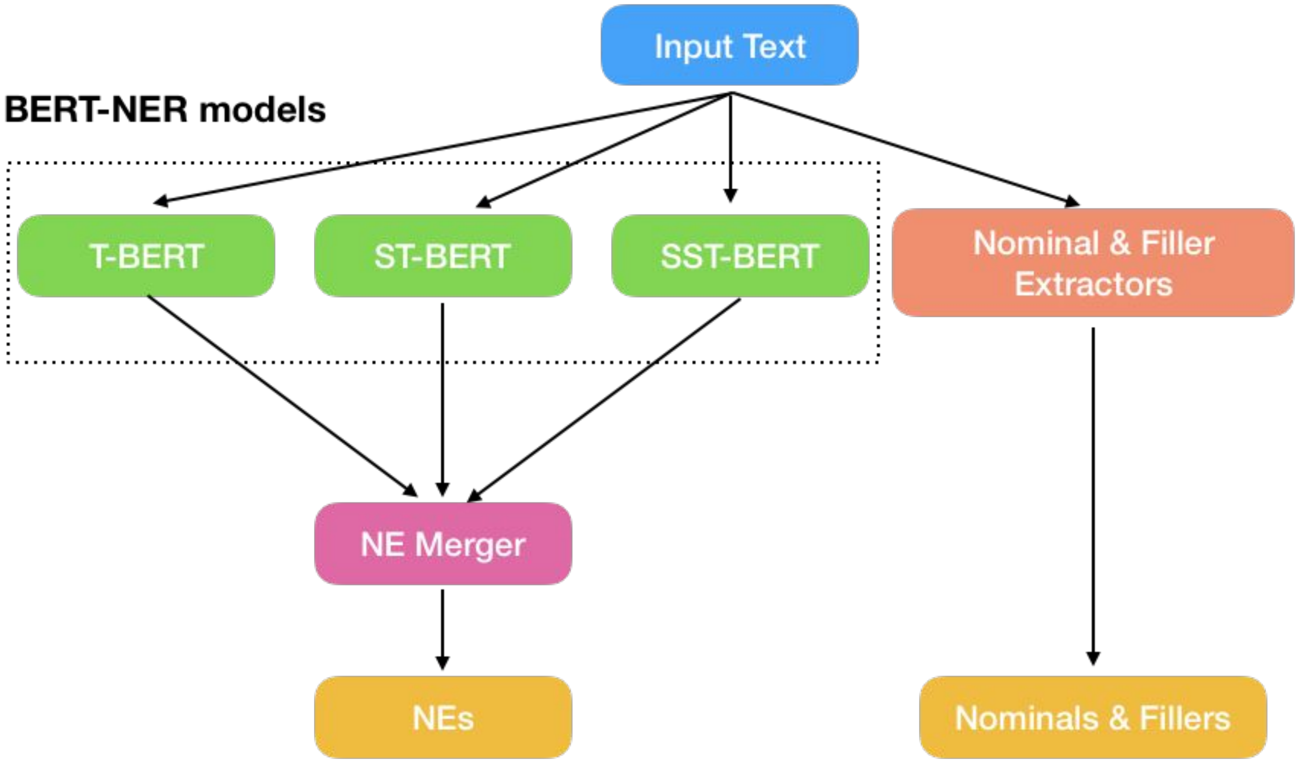
➢ The rest automatically imported from Wikidata

  - Nouns and noun phrases referring to entity types

# Entity detection

- Multi-level BERT-based modules and merger
  - Train separate detectors for type, subtype and subsubtype-level type classification
  - Multilingual BERT + language knowledge transfer to mitigate limited training data
- Type-level training data: KBP NER data and a small amount of self-annotated data
- Sub(sub)type-level training data: YAGO knowledge base (350k+ entity types) obtained from Heng Ji
- Nominal and temporal expression extraction
  - Temporal normalization using SUTime, HeidelTime

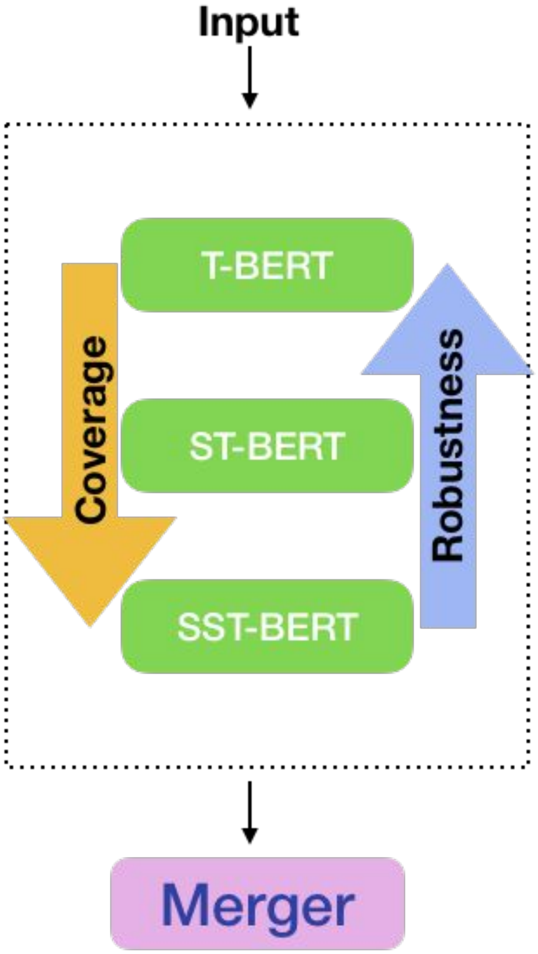
# Entity detection pipeline



# Entity detectors

- Level 1 model is robust
- Level 3 detector knows more types
- Using hierarchical property, Level 1 & 2 models are employed to narrow down the size of candidates for the Level 3 model

| | No check | Consistency check |
|---|---|---|
| Level 3 F1 score | 44 | **53** |



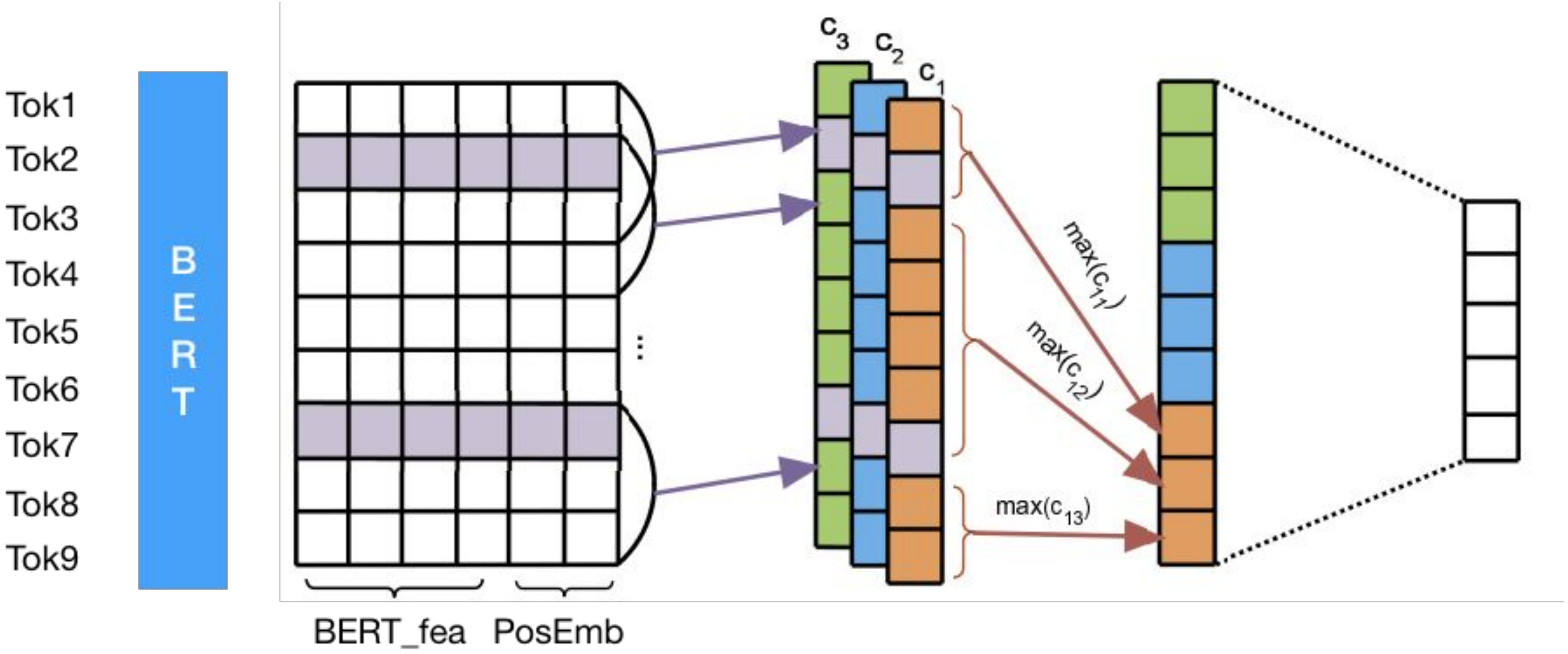
# Entity Linking

- Task: Given NER output mentions, link them to the reference KB
- Challenges: large KB, noisy Geonames
  - Pre-process KB: remove duplicate and unimportant entries
- String match and disambiguate:
  - Search KB for candidates using Lucene
  - Create connectedness graph, with PageRank link strength scores
  - Run disambiguation by pruning graph
- Updates after M36 eval
  - Add transliteration to better handle Spanish and Russian
  - Support entity types other than GPE, LOC, PER and ORG (e.g. FAC)
  - Heuristic to select KB entries with shorter wiki-page links (well-known places)

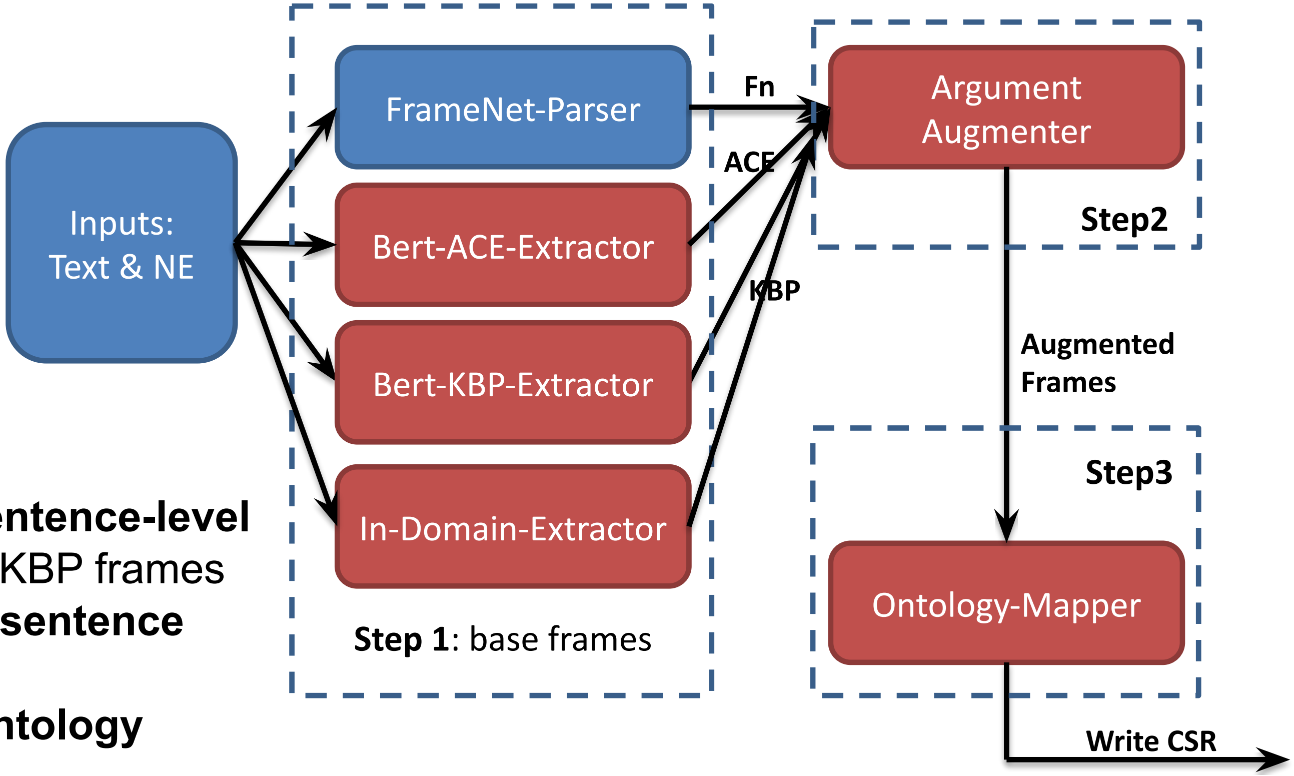
# Relation extraction

- Four step approach:
  - **BERT word embeddings** for features
  - **Convolution**: extract and merge all local features for a sentence
  - **Piecewise max pooling**: input is split into 3 segments according to positions of entity pair, returns max value in each segment
  - **Softmax classifier** to compute the confidence of each relation

# Relation extraction



# Events pipeline




Step 1: Extract **basic sentence-level** FrameNet/ACE/KBP frames

Step 2: Augment **cross-sentence** arguments

Step 3: Map to **target ontology**

# Step 1: Basic Frames

Obtain base frames from three sources:

- **FrameNet** Parser (Re-use our previous Semafor engine)
- Bert-based Model trained on **ACE05** data.
- Bert-based Model trained on **RichERE/KBP** data.

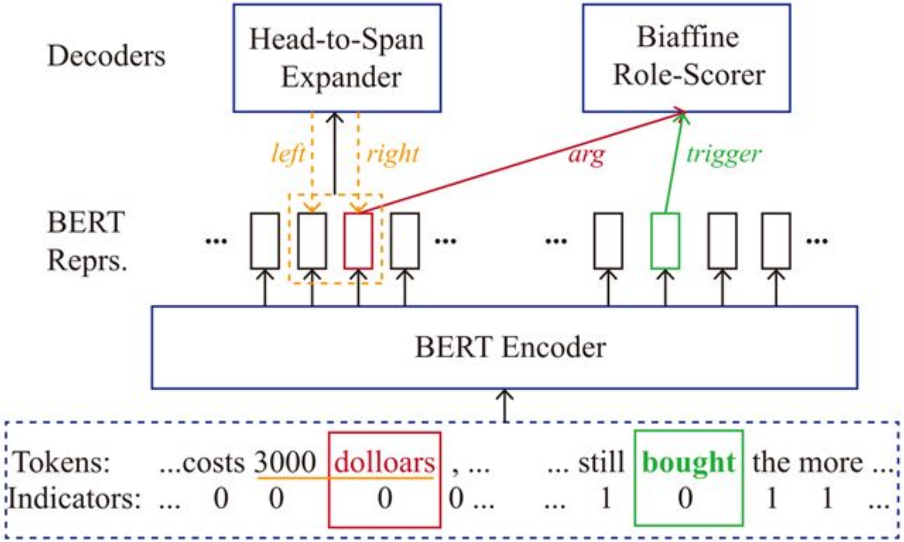
Merge these frames (simple position matching) together for the **basic sentence-level frames**.

# Step 2: Augment Event Arguments

Augment arguments across sentences with models trained on the RAMS dataset

Step 1: **Detect** the (syntactic) head word

Step 2: **Expand** from head words to full spans




Zhang et al. A Two-Step Approach for Implicit Event Argument Detection (ACL 2020)

# Step 3: Ontology Mapping

One of our main challenge is to map to the three-layered **complex target ontology**
We design “rule-constrained” models for the mapping:

- Rules **act as constraints**:
  - -- kbp:Contact.Meet -> aida:Contact.*.Meet
  - -- kbp:Conflict.Attack -> aida:Conflict.Attack.*
- For the remaining ambiguities, use the models (ensembled MLP/KNN classifiers based on BERT embeddings) to classify

# Cross-Lingual Transfer for Event Processing

- Model transfer: Trained a model aware of multiple languages, based on **multilingual BERT.**
- Data transfer: **Translate English training** data to target languages.

# Cross-Lingual Data Transfer by Translation

He **bought** the house with 1M dollars .

Source Annotation + Word Alignment

**Compró** la casa con 1 millón de dólares.

Annotation Projection (Translation):

Src: **bought**(Transaction.TransferOwnership), He(Recipient), the house(Artifact), 1M dollars(Money)

Trg: **Compró**(Transaction.TransferOwnership), la casa(Artifact), 1 millón de dólares(Money)

# Within-doc coref pipeline - Overview

Multilingual mention-pair scorers for within-doc entity/event coref on non-English languages (i.e *es* and *ru*)

**Supervision**: *en* and *es* supervision from TAC KBP 2015-2017

**Prepro**: Since we have non-*en* languages → avoid linguistic preprocessing (e.g dependency edges or pos tags) and input → Lets us easily have shared featurization across langs!

Multilingual formulation → start from already pre-trained multilingual space → use bert-base-multilingual and finetune

- <u>Benefit A</u>: Can generalize to *unseen-during-task-training* langs at test-time (e.g ru for our case)
- <u>Benefit B</u>: Use both en and es supervision together during task training.

# Within-doc coref pipeline - Repn & Training

**Training**: Finetunes a bert-base-multilingual model in paired form on mention pairs in-context.

**Input Structure**: Featurized as [CLS] *<leftContext1>* *<mention1>* *<rightContext1>* [SEP] *<leftContext2>* *<mention2>* *<rightContext2>* [SEP], with contexts capped at 128 subwords each.

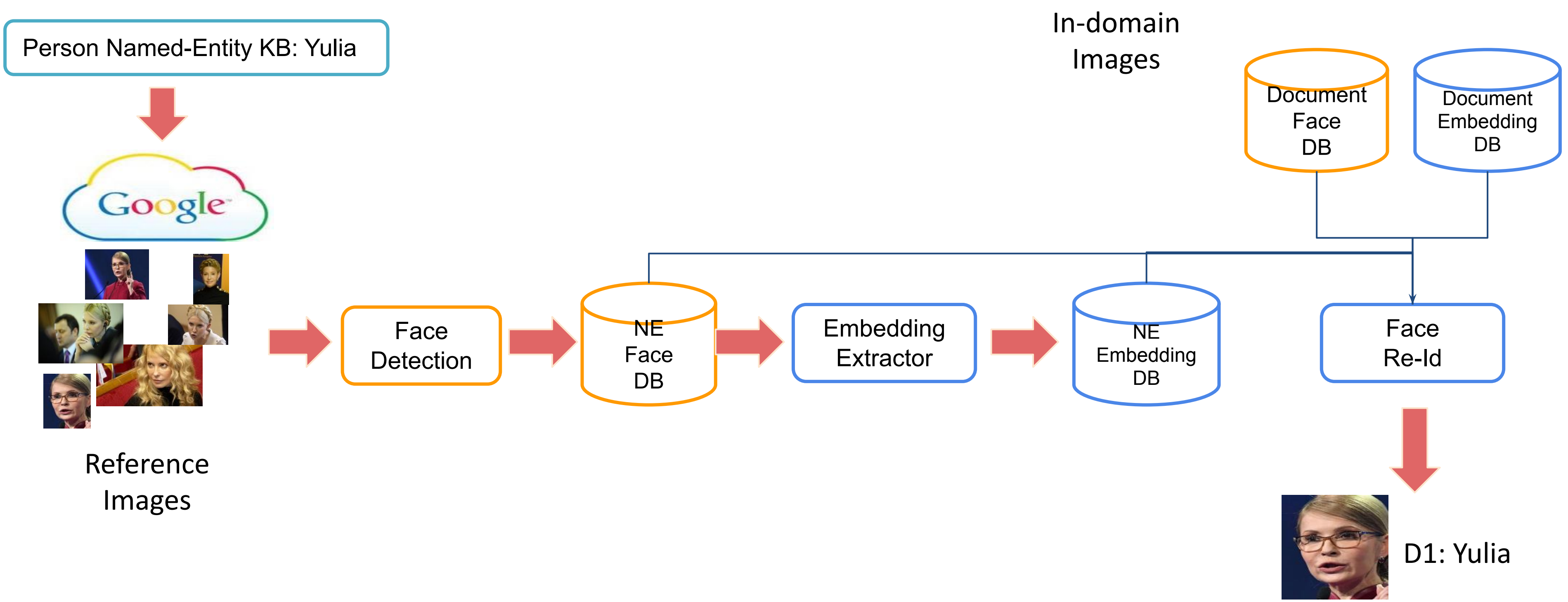
**Task Layers**: Binary classification layer → predict *coreferent* vs *not coreferent*

**Representing non-textual attributes**: Augment around <mention> → Use html tag-enclosed sequences to distinguish from text.

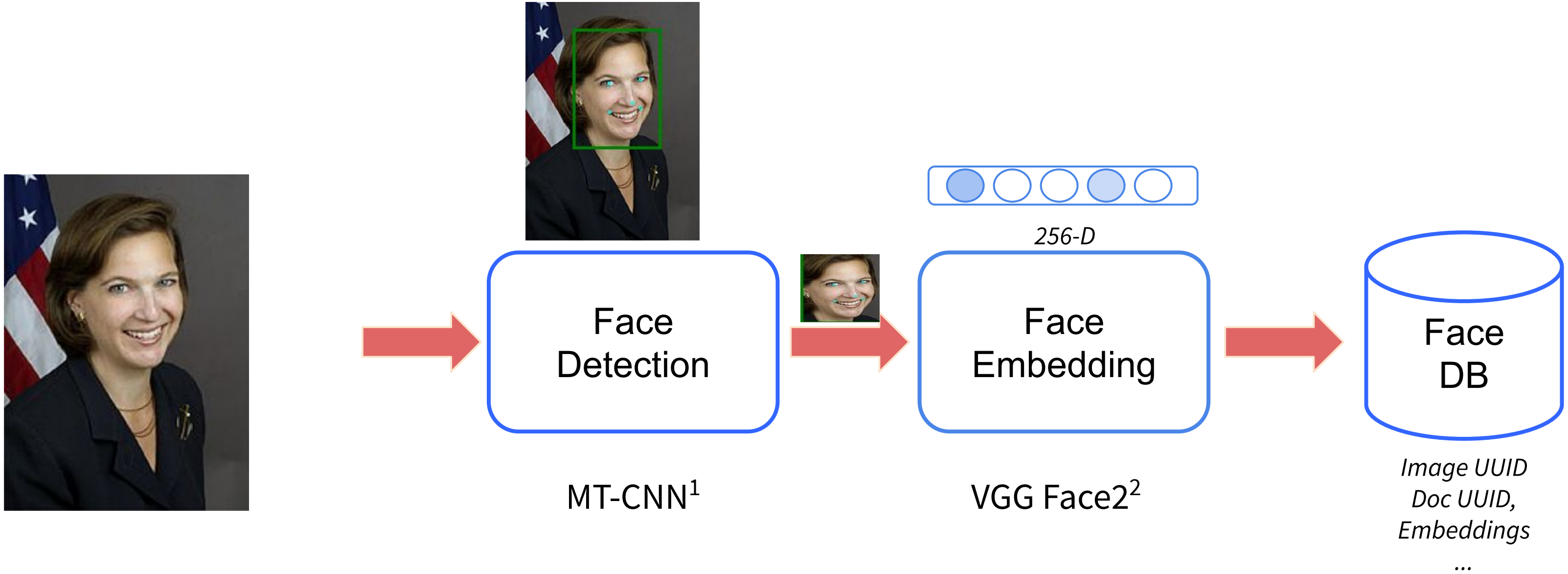
The entity mention *Gamal Abdel Nasser* would be represented as *Gamal Abdel Nasser <typ> per.politician.headOfState </typ>*

For events, there is a lot more information to pack in this way

# Visual coreference pipeline

# Document-level image processing



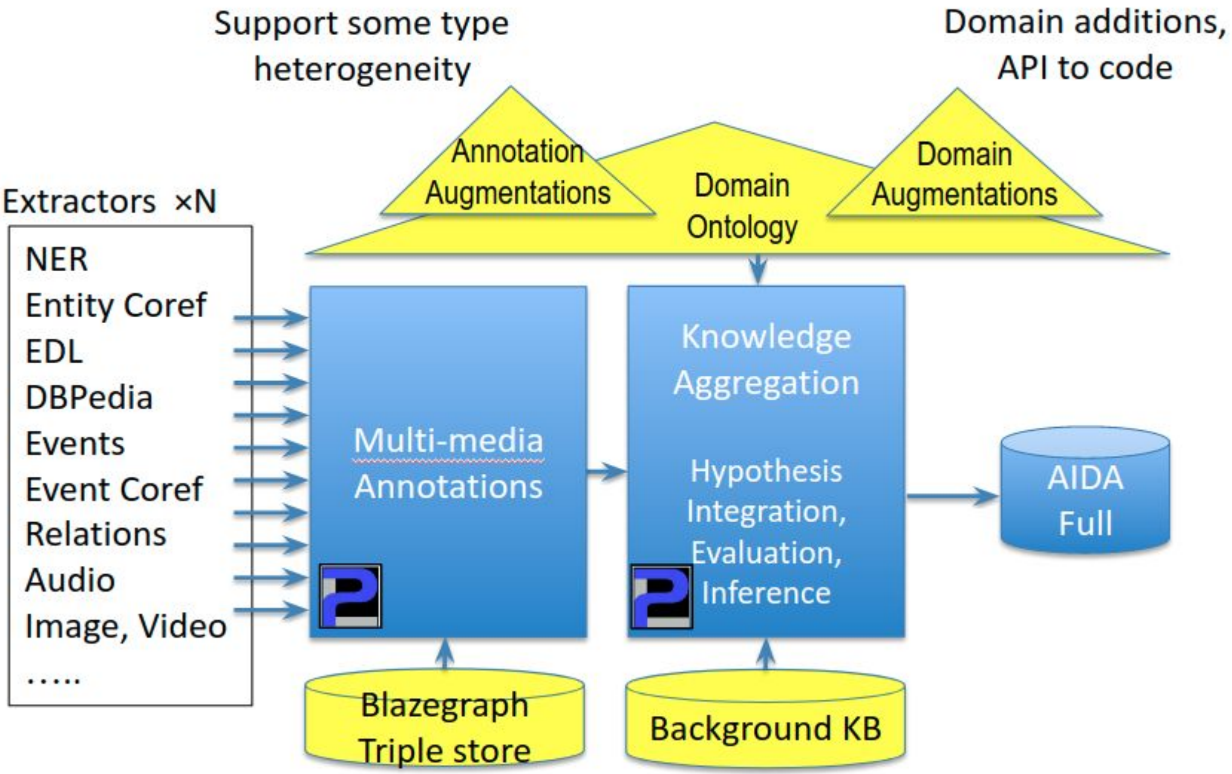

Similarity between image embeddings is L2 distance: $Similarity = \sqrt{\sum_{i=1}^{256} (v_i^{(1)} - v_i^{(2)})}$


1. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016
2. Cao, Qiong, et al. "Vggface2: A dataset for recognising faces across pose and age." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.

# KB generation

- Knowledge Aggregator (**KAgg**) toolkit
  - **PowerLoom** (Chalupsky et al. 2010): Logic-based knowledge representation and reasoning system
  - **Blazegraph**: triple store and graph database to support storage and querying
- Use coref and other equivalence information to connect annotations
  - mention overlap, name links, EDL, within-doc coref, event coref
- Apply inferences, evaluate constraints, detect conflicts, do attribution
  - **Aggregate** input from multiple extraction engines, different developers, third-party and legacy systems
  - **Resolve** missing types, conflicting types once things are linked
  - **Discard** incoherent event arguments, relations based on ontology and domain constraints
- Export to the AIDA interchange format (AIF)

# KAgg architecture



# Dockerization

- All TA1, TA2, TA3 systems are available as docker images from dockerhub
  - TA1: zs12/opera-ta1-m36
  - TA2: dangiankit/opera-ta2
  - TA3: dangiankit/opera-ta3
- Includes updates made after the TAC 2020 SM-KBP evaluation
- Email zsheikh@cs.cmu.edu and adangi@cs.cmu.edu for help

# Papers


- Zhang, Z., X. Kong, Z. Liu, X. Ma, and E.H. Hovy. 2020. A Two-Step Approach for Implicit Event Argument Detection. ACL 2020. Seattle, WA.
- Kong, X., V. Gangal, and E.H. Hovy. 2020. Sentence Cloze Dataset with High Quality Distractors From Examinations. ACL 2020. Seattle, WA.
- Kong X, Zhang Z, Hovy E. Incorporating a Local Translation Mechanism into Non-autoregressive Translation. EMNLP 2020
- Zhang, Z., Kong, X., Levin, L., & Hovy, E. An Empirical Exploration of Local Ordering Pre-training for Structured Learning. EMNLP 2020
- V. Gangal, and E.H. Hovy. "BERTering RAMS: What and How Much does BERT Already Know About Event Arguments?-A Study on the RAMS Dataset." Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020
- Kong, X., Chen, X., & Hovy, E. (2019). Decompressing Knowledge Graph Representations for Link Prediction. *arXiv:1911.04053*.
- Chalupsky, H. 2019. Chameleon 2.0: Integrating Neural and Symbolic Reasoning in PowerLoom. Proceedings of the AKBC Workshop on Neural and Symbolic Representation and Reasoning.
- Chaudhary, C., J. Xie, Z. Sheikh, G. Neubig, J.G. Carbonell. 2019. A Little Annotation does a Lot of Good: A Study in Bootstrapping Low-resource Named Entity Recognizers. Proceedings of the EMNLP conference.
- Hovy,E.H.,J.G.Carbonell,H.Chalupsky,A.Gershman,A.Hauptmann,F.Metze,T.Mitamura,Z. Sheikh, A. Dangi, A. Chaudhary, X. Chen, X. Kong, B. Huang, S. Medina, H. Liu, X. Ma, M. Ryskina, R. Sanabria, V. Gangal. 2019. OPERA: Operations-oriented Probabilistic Extraction, Reasoning, and Analysis. Proceedings of the NIST Text Analysis Conference TAC.
- Huang, P., Hu J., Chang X., Hauptmann A., Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting, ACL 2020.
- Huang, P., Chang X, Hauptmann A.. Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations. EMNLP 2019.
- Huang, P., Kang G., Liu W., Chang X., Hauptmann A., Annotation Efficient Cross-Modal Retrieval with Adversarial Attentive Alignment. ACM MM 2019.
- Huang, P., XChang X, Hauptmann A., Hovy,E., Forward and Backward Multimodal NMT for Improved Monolingual and Multilingual Cross-Modal Retrieval, *Proceedings of the 2020 on ACM International Conference on Multimedia Retrieval (ACM ICMR)*, 2020.
- Kong, X., Tu, Z., Shi, S., Hovy, E., & Zhang, T. Neural machine translation with adequacy-oriented learning. AAAI 2019
- Zhang, Z., X. Ma, and E.H. Hovy. 2019. An Empirical Investigation of Structured Output Modeling for Graph-based Neural Dependency Parsing. Proceedings of the ACL conference.
- Kong, Xiang, et al. "Fast and simple mixture of softmaxes with bpe and hybrid-lightrnn for language generation." AAAI 2019

# Questions?

Thank you