Welcome to the

OPERA

AIDA in 2019 … a challenge

No more training data, only examples that illustrate the evaluation

Increasingly data-intensive neural learners

What do we do???

A range of responses…

- Just make machine learning work! 1
- Learning, augmented with external data 2
- Half-half 3
- Include (some) learning but only if it's easy 4
- Forget machine learning! 5

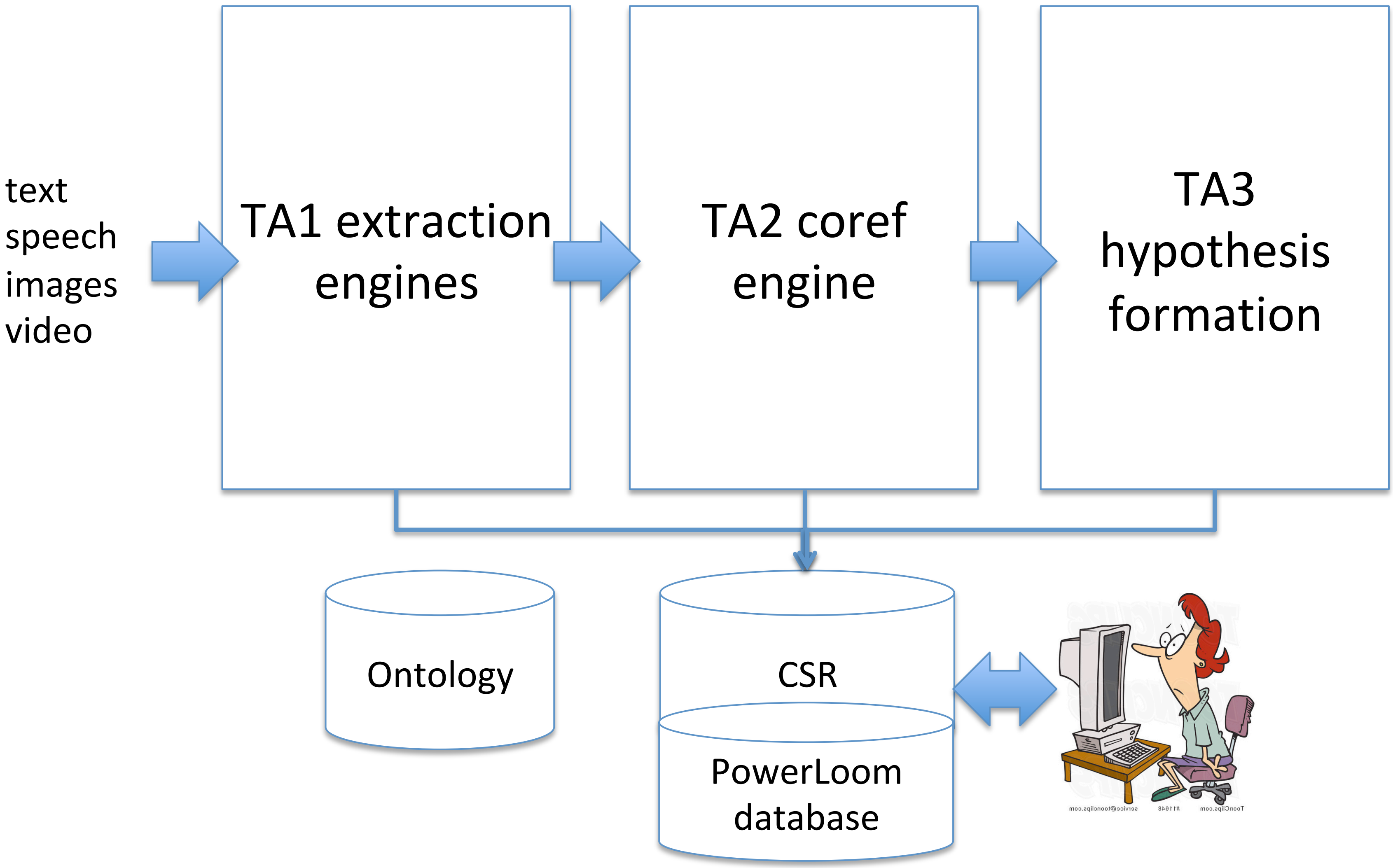
Overview

1. System overview
2. TA1 English entity and relation processing
3. TA1 Rus/Ukr entity and event processing
4. TA1/2 KB construction and validation
5. TA3 Hypotheses

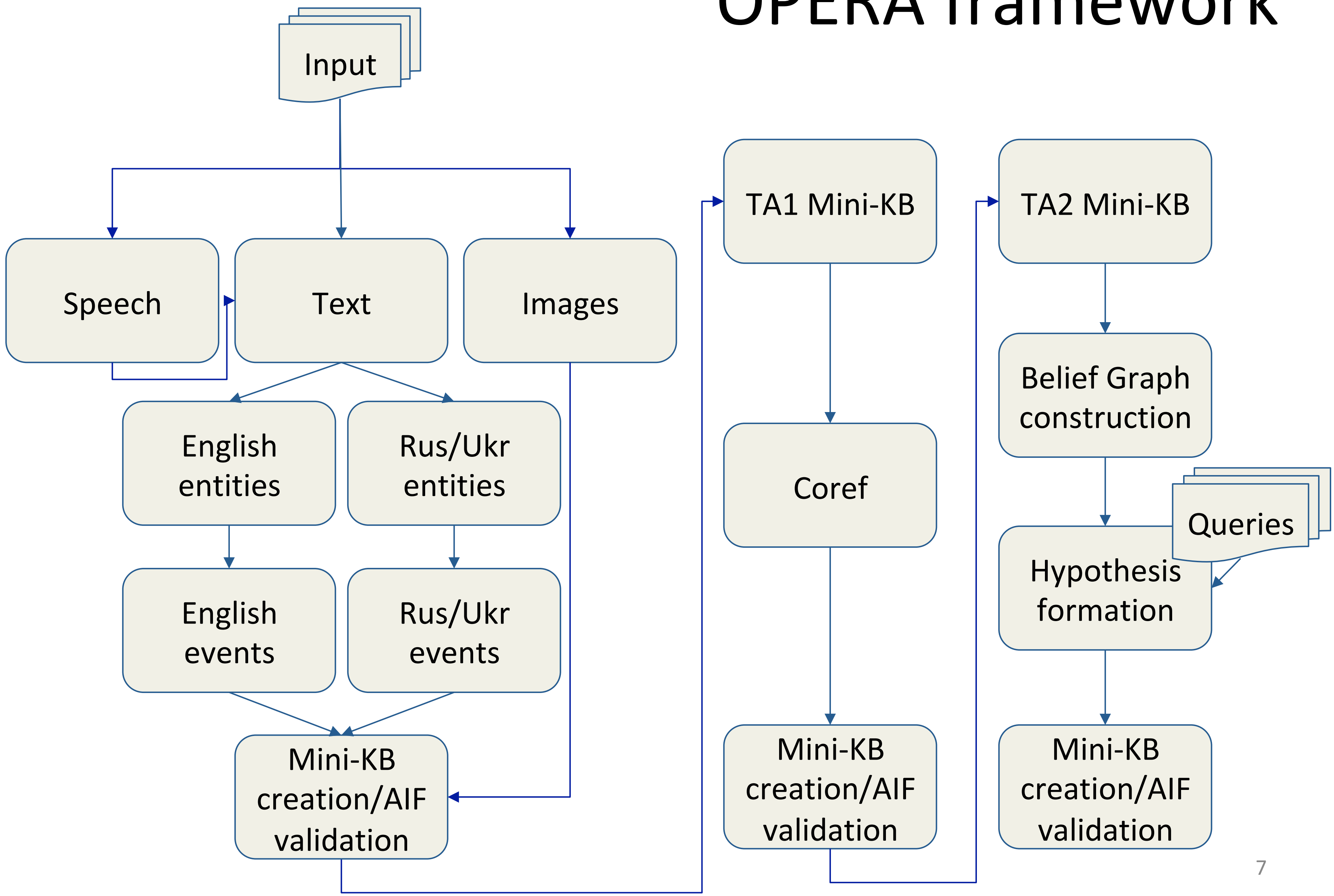
Zaid Sheikh, Ankit Dangi, Eduard Hovy

SYSTEM OVERVIEW

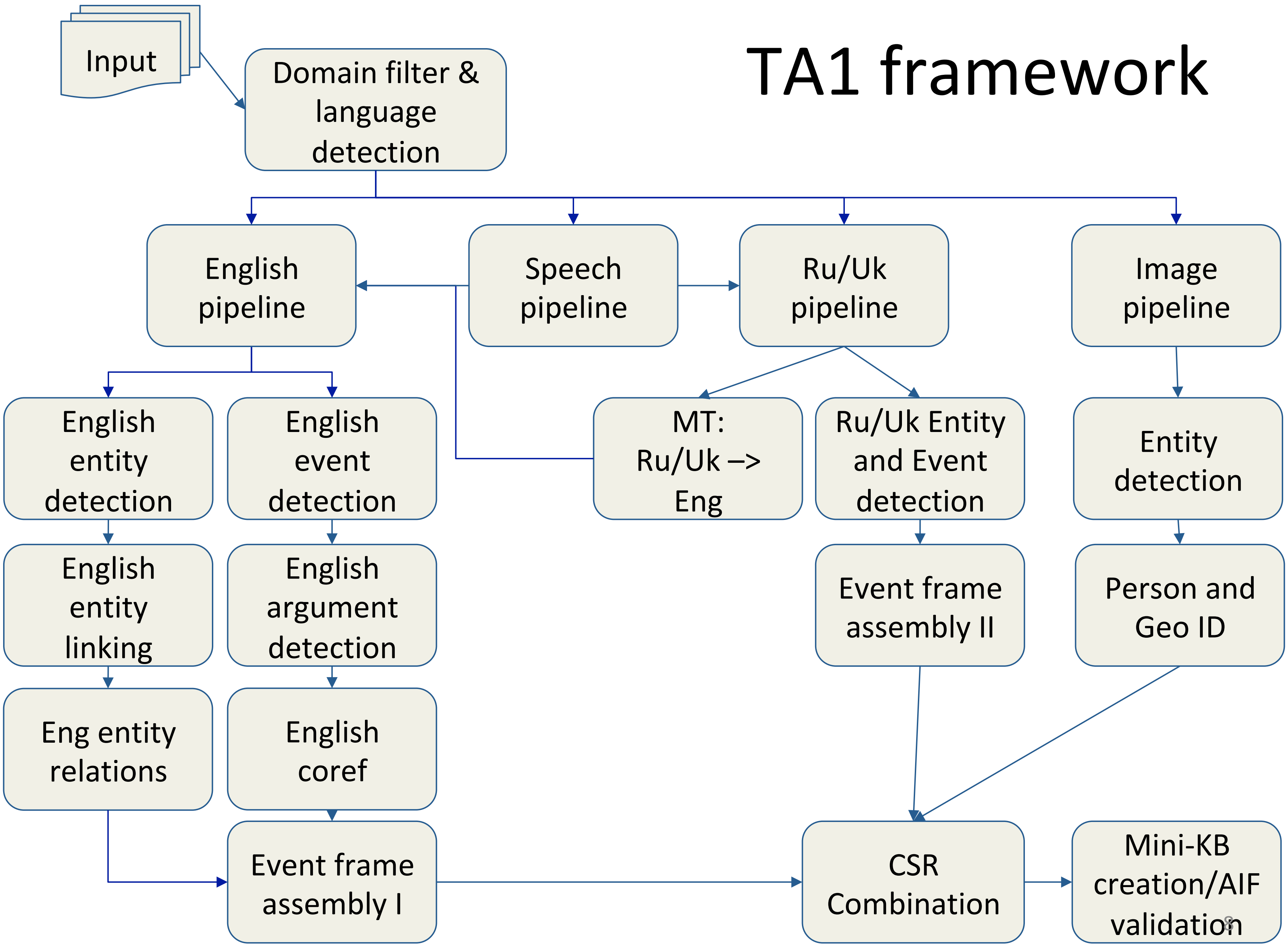
OPERA architecture



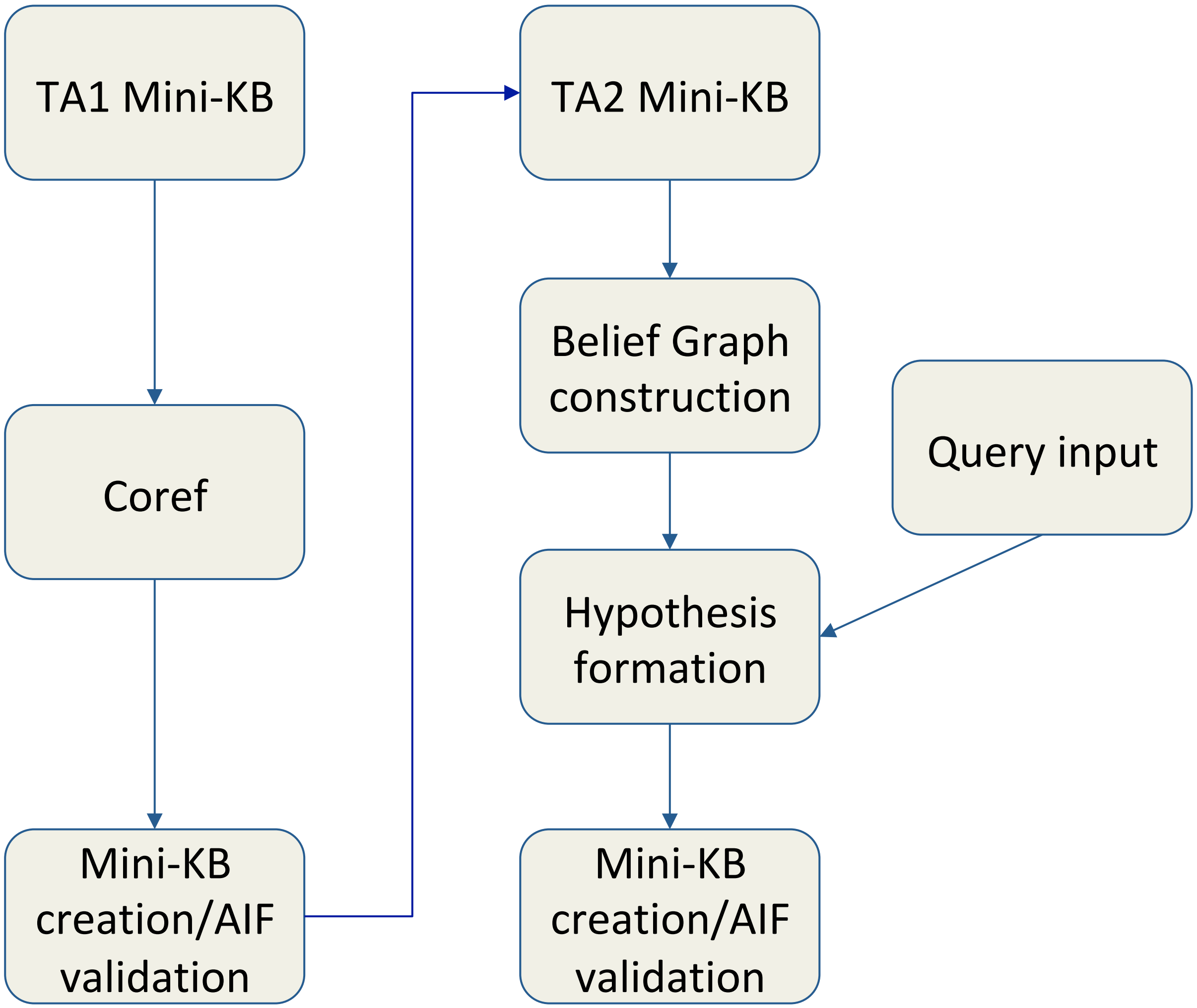
OPERA framework



TA1 framework



OPERA TA2 + TA3 framework



KBs and notations

- All results written in OPERA-internal frame notation (json) and stored in CSR (BlazeGraph)
- Input / output converters from/to AIDA AIF

- Two separate KB creation and validation procedures, for two parallel KBs (gives insurance, coverage, and backup):
  - Chalupsky: uses PowerLoom and Chameleon reasoner
  - Chaudhary: uses specialized rules

4

5

Internal dryruns

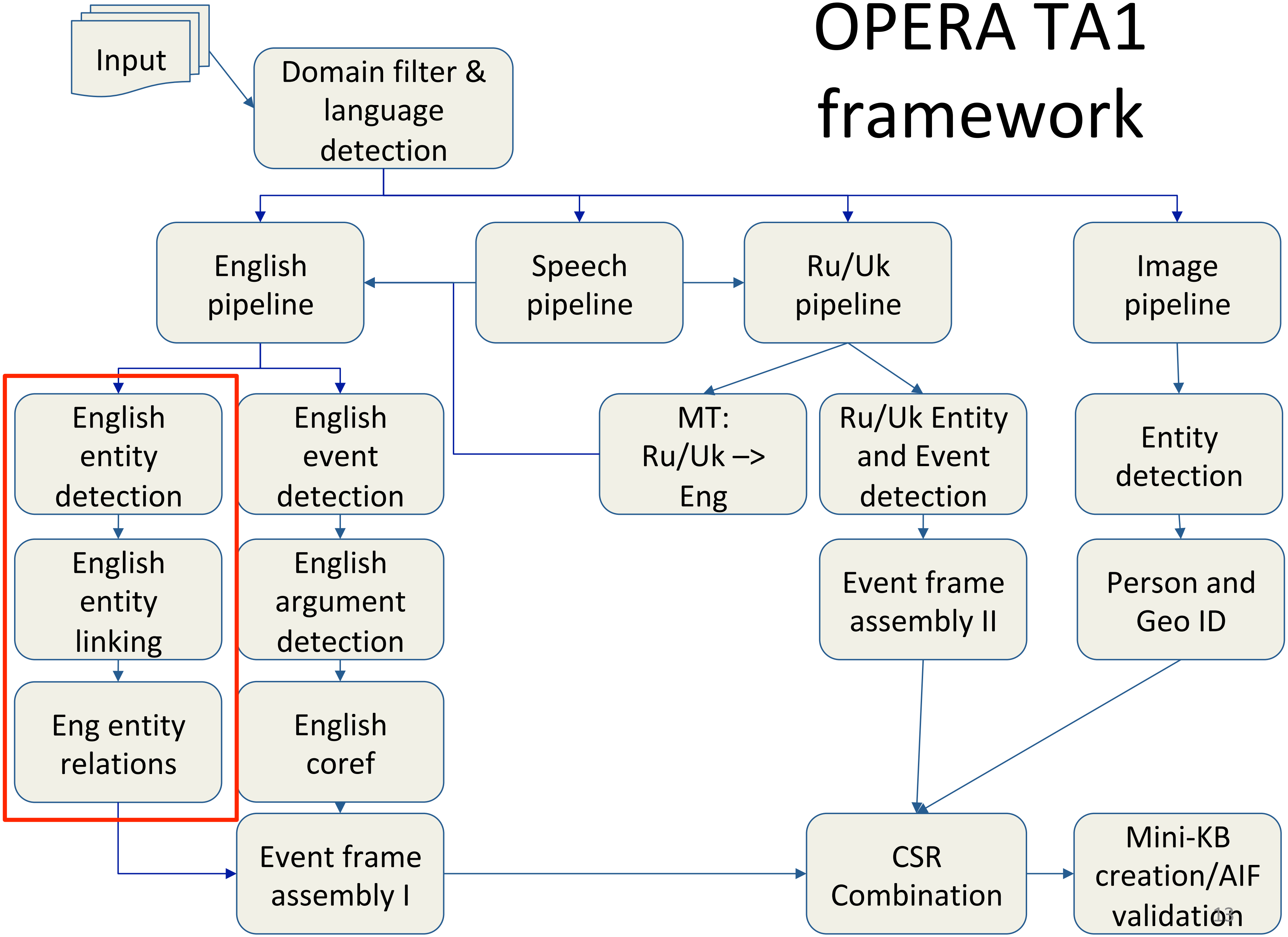
- Internal dry run mini-evals using the practice annotations released by LDC
- Evaluated results manually

- Results look promising, BUT … hard to calculate P/R/F1 for various parts of the TA1 pipeline because LDC does not label all mentions of events, relations and entities, just the “salient” or “informative” ones (so we have to judge them ourselves … laborious and not guaranteed)

Xiang Kong, Xianyang Chen, Eduard Hovy

TA1 TEXT:
ENGLISH ENTITIES AND RELATIONS

OPERA TA1 framework



1. Entity detection: Type-based NER data

- Multi-level learning:
  - Train separate detectors for type, subtype, and subsubtype-level type classification
  - Addresses data imbalance
  - May introduce layer-inconsistent types!

2

- Type-level from LDC ontology:
  - Training data: KBP NER data and a small amount of self-annotated data

3

- Sub(sub)type-level:
  - Training data: YAGO knowledge base (350k+ entity types) obtained from Heng Ji — thanks!

2. Entity linking

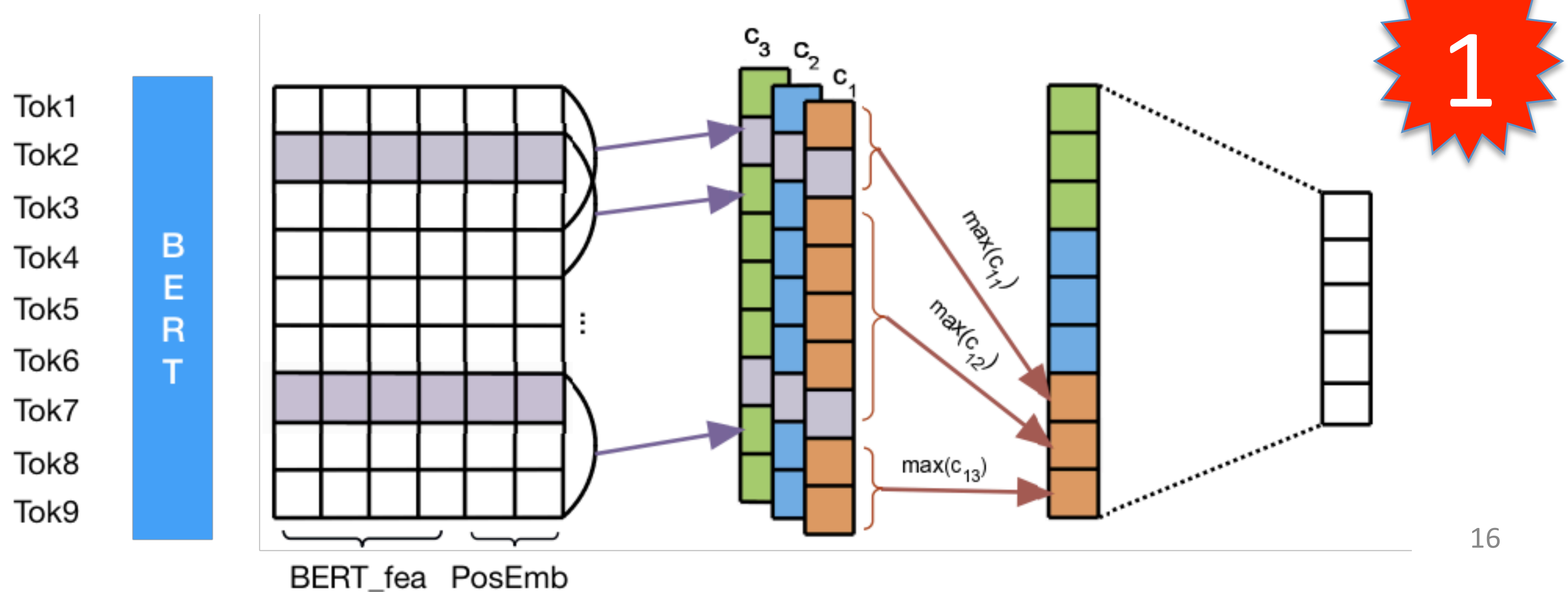
- Task: Given NER output mentions, link them to the reference KB
- Challenges: Over-large KB, noisy Geonames
  - Preprocess KB: Remove duplicated and unimportant entries (i.e., not located in Russia or Ukraine, or no Wikipedia page)

2

- Approach, given an entity:
  - Use Lucene to find all candidates in KB
  - Filter spurious matches
  - Build connectedness graph, with PageRank link strength scores
  - Prune (densify) graph to disambiguate entity

3. Entity relation extraction

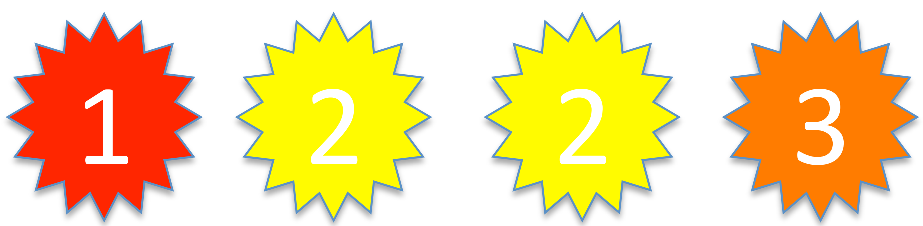
- Task: Extract entity properties and event participants
- Four-step approach:
  1. BERT word embeddings for features
  2. Convolution: extract and merge all local features for a sentence
  3. Piecewise max pooling: split input into three segments (by position) and return max value in each segment, for 2 entities + 1 relation
  4. Softmax classifier to compute confidence of each relation



English entity/relation discussion

- Challenges and problems
  - Subsubtype is super fine-grained; our NER engine is still not robust enough
  - We return both type and subsubtype labels, but in the eval NIST will judge only one of them

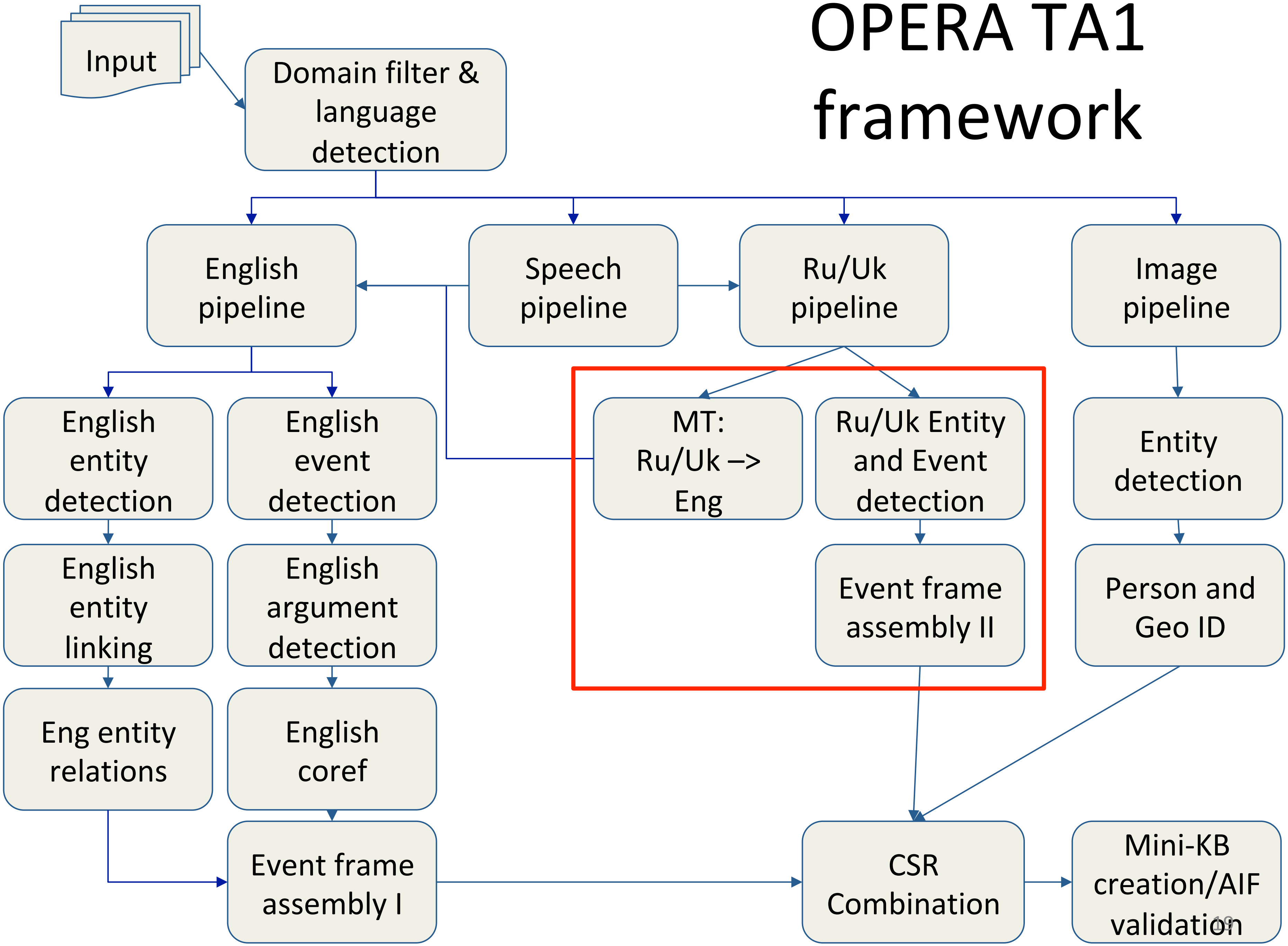
- Mostly learned, but some manual assistance



Mariia Ryskina, Yu-Hsuan Wang, Anatole Gershman

TA1 RUSSIAN AND UKRAINIAN

OPERA TA1 framework



Goals and challenges

- Goal: Extract entity and event mentions from Russian and Ukrainian text, and build frames
- Challenges:
  - Lack of pretrained off-the-shelf extractors
  - Lack of annotated data to train systems
  - Highly specific ontology

- Two pipelines:
  1. Rus and Ukr source text
  2. MT into English

Example input and output

Input: Про-российские сепаратисты атаковали Краматорский аэропорт.

Translation: Pro-Russian separatists attacked Kramatorsk airport.

Output:

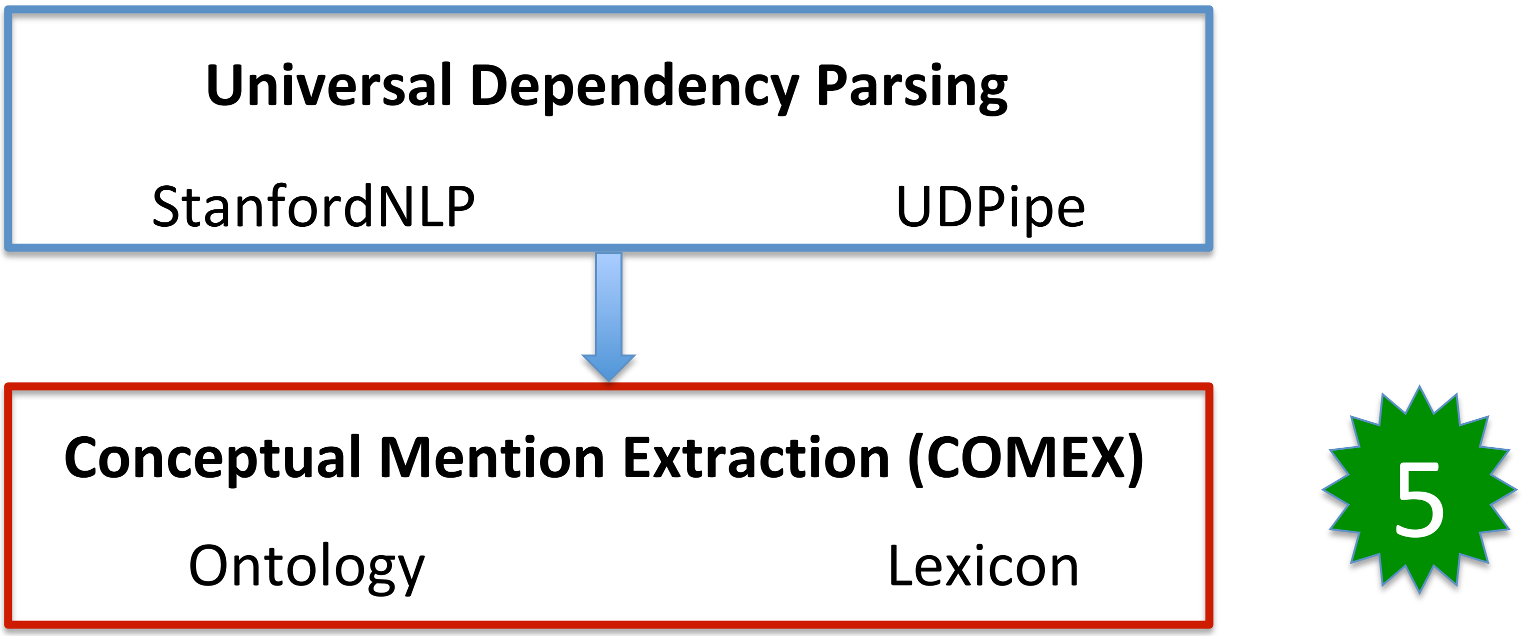
mn0: event *Conflict.Attack*, Attacker: **mn1**, Target: **mn3** — text: атаковали

mn5: relation *GeneralAffiliation.MemberOriginReligionEthnicity* Person: **mn1**, EntityOrFiller: **mn2**, — text: Про-российские сепаратисты

mn6: relation *Physical.LocatedNear*, EntityOrFiller: **mn3**, Place: **mn4** — text: Краматорский аэропорт

mn1: entity *ORG*,	text: Про-российские сепаратисты
mn2: entity *GPE.Country.Country*,	text: Про-российские
mn3: entity *FAC.Installation.Airport*,	text: Краматорский аэропорт
mn4: entity *GPE.UrbanArea.City*,	text: Краматорский

Approach 1: Processing in Rus/Ukr



- Our ontology is a superset of the NIST/LDC ontology
- Lexicons are (semi-)manually created from the training data
- Conceptual extraction using (manual) rule-based inference
- Focus is on high precision

Parsing/tagging/chunking pipeline

- Syntax pipeline:
  - UDPipe 1.2 (Straka & Strakova 2017)
  - Extract head nouns and dependents
  - Not all entities and events needed

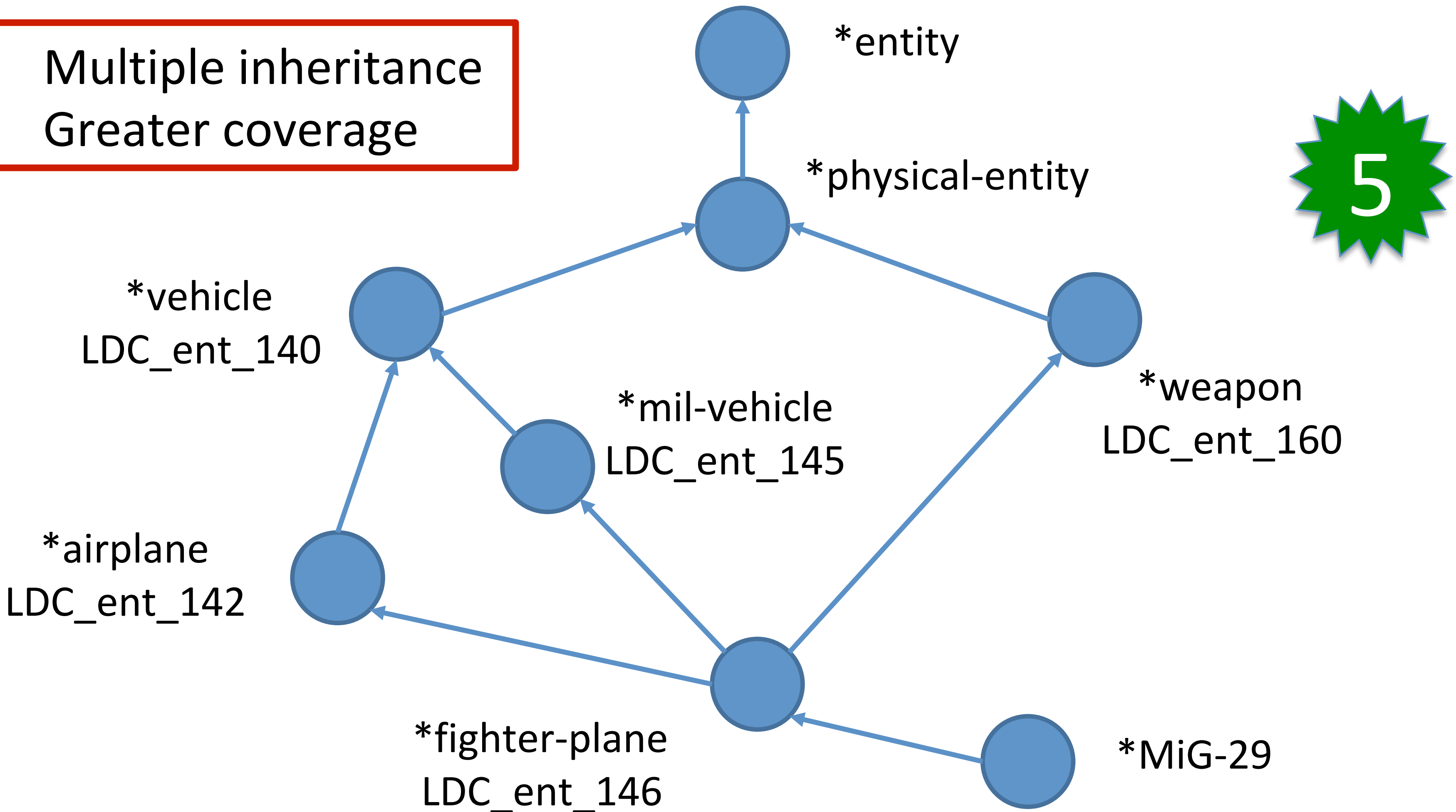
- Event frame construction: COMEX
  - Our ontology is a superset of the AIDA ontology
  - Trigger terms manually mapped to ontology:
    - Direct matching — manually curated list of trigger words
    - English triggers — translation or WordNet/dictionary lookup
  - Analysis guided by annotation:
    - LDC annotations from seedling corpus
    - Own manual annotation as well

5

5

COMEX ontology

- Multiple inheritance
- Greater coverage



COMEX lexicons

- Connect words to ontology concepts via word senses
- Provide rules for connecting concepts into a mention graph
- Semantic requirements for slot fillers are specified in the ontology

```
W, атаковать, WS:attack-physical, WS:attack-verbal
S, WS:attack-physical,  *attack-physical,  VERB
A, WS:attack-physical,      Attacker = Pull:active-subj;     Pull:passive-subj
A, WS:attack-physical,      Target    = Pull:active-dir-obj; Pull:passive-dir-obj
A, WS:attack-physical,      Instr      = Pull:active-subj
A, WS:attack-physical,      Place      = Pull:obl-in
#
R, Pull:active-subj,     nsubj,    Trigger->Voice=Act
R, Pull:passive-subj,    obl,       Trigger->Voice=Pass,   Target->Case=Ins
```

While the lexicons contain hundreds of words, the number of rules is small

Lexicon construction

- Initial vocabulary and the corresponding concepts from the available LDC annotations
- Vocabulary enrichment by extracting all named and nominal entities from the seedling corpus files that contain at least one LDC annotation
- Event trigger enrichment using WordNet
- Cross-language vocabulary enrichment using MT and alignment
- Manual curation of the resulting vocabulary
- Manual addition of attribute rules
- Iterative improvement process:
  1. Extract mentions from a new file
  2. Score results
  3. Add vocabulary, fix rules and do cross-language transfer

5

Sample COMEX performance

	English	Russian	Ukrainian
Precision	0.91–1.0	0.93–1.0	1.0
Recall	0.22–0.56	0.11–0.70	0.07–0.42
F1	0.35–0.70	0.20–0.62	0.13–0.59
Vocabulary	178	1483	1430*
Rules	33	30	13

(This work continues; the numbers change every day)

COMEX is the most ‘manual’ of OPERA’s TA1 extraction modules

Approach 2: Rus/Ukr $\xrightarrow{\text{MT}}$ English

- Pipeline:
  - MT Rus/Ukr –> English using MS Azure
  - Run OPERA TA1 extractors
  - Align source text to extracted mentions in Eng
    - Back-translate from Eng, including XML-like entity/event tags
- Output is generally good (esp when no XML tags)
- Problems in back-translation:
  - Sometimes messes up the XML tags
  - May switch event arguments
  - May mess up proper names (e.g. Slavyansk –> Slavska, Slavovsk, Slavic
  - Things like typos or uncommon words get translated incorrectly into Eng, but may be easy to fix in the source using fuzzy matching

Approaches complementary

- Rus/Ukr: more precise
  - Less noise, better entity typing
- MT: more general
  - Better at names, time/numbers, event typing
- Overlaps and differences:
  - Entity overlap: 84% of Rus/Ukr = 44% of MT output
  - Event overlap: 58% of Rus/Ukr = 49% of MT output
  - Type agreement: 87% of overlap
  - Remaining mentions: 65–70% correct on each side
  - Differences in spans, event vs. entity choices

Rus/Ukr entity/relation discussion

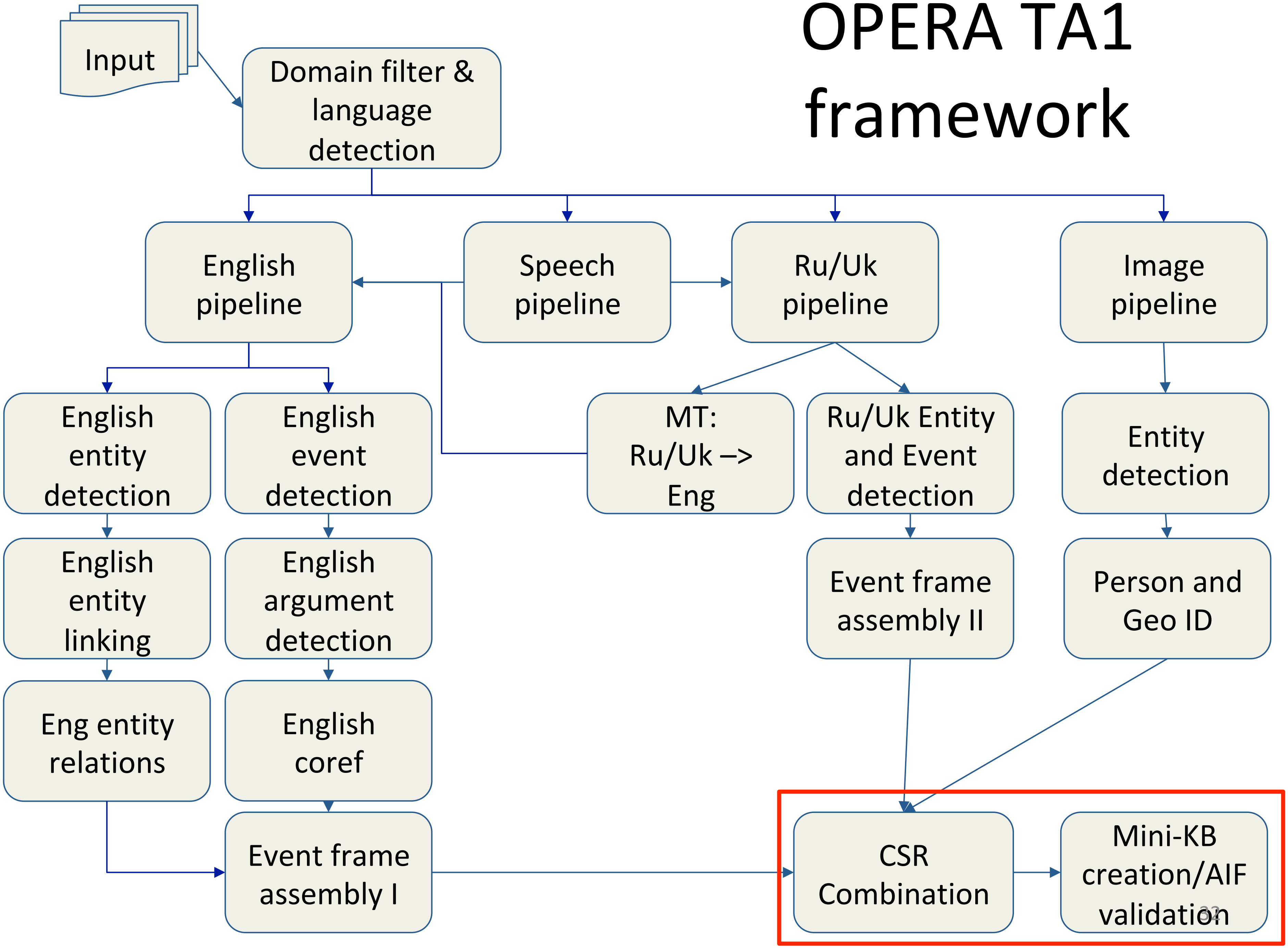
- Challenges and problems
  - Slow manual rule building, limited coverage (but high precision)
  - COMEX<—>AIDA ontology alignment
  - Noise in translation
- Mostly manual

2 5 5 5

Hans Chalupsky

TA1/2 KB CONSTRUCTION AND VALIDATION

OPERA TA1 framework



CSR: PowerLoom-based Common semantic repository

- Contains all KEs
  - Contains discrete term propositions, [structured] distributional vectors/tensors, continuous embeddings
  - Each with vector of scores (e.g., TA1 extraction confidence, source trustworthiness, reasoning implication confidence, cross-KE compatibility, hypothesis-based likelihoods, etc.)

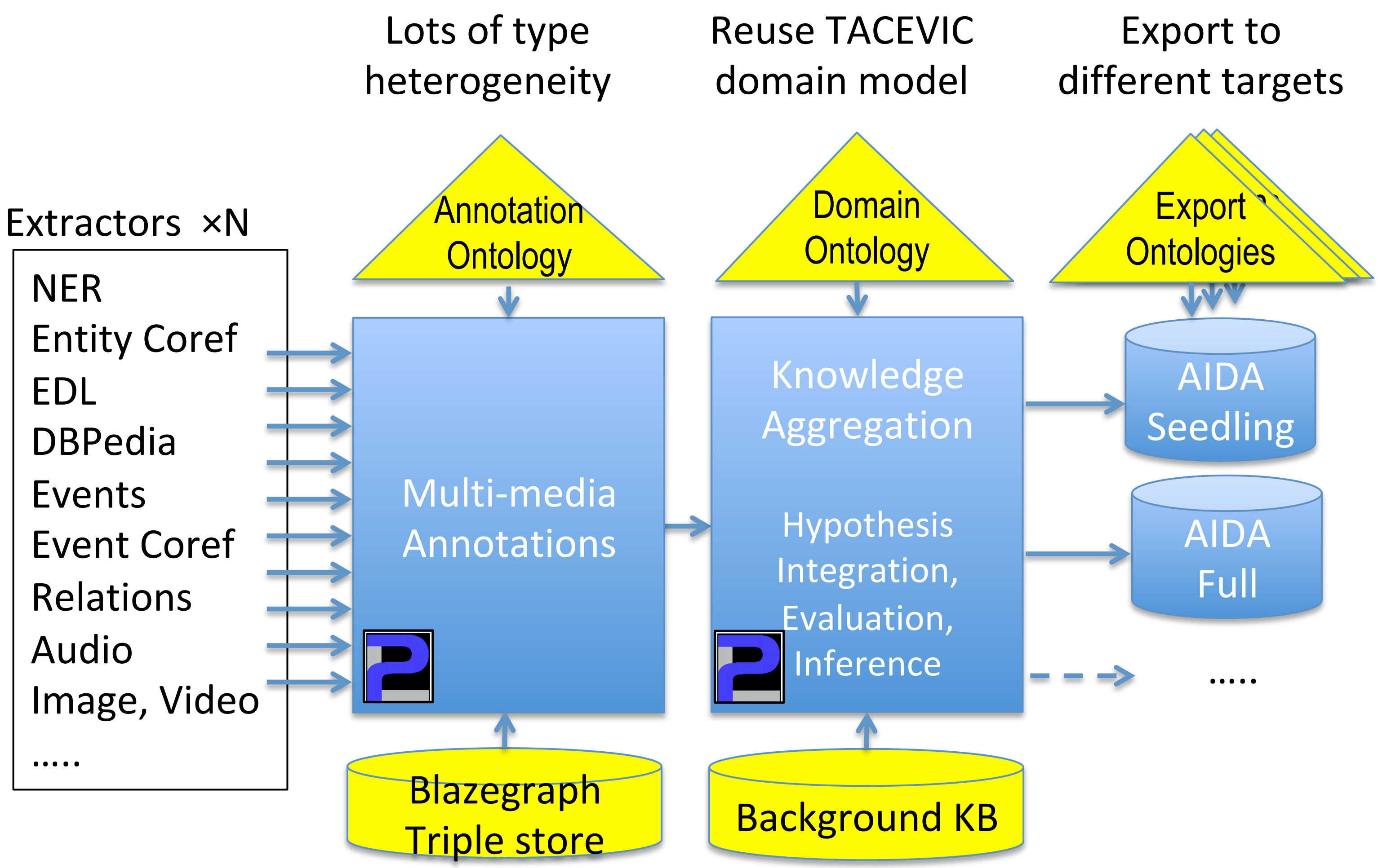
4

- Represented in PowerLoom (Chalupsky et al. 2010)
  - Predicate-logic-based representation based on KIF that is a supported syntax of Common Logic
  - Dynamic, scalable, multi-contextual system to store, manage and reason with information
  - Blazegraph database tech for scalability and integration
  - Represent hypotheses and probabilities via reification

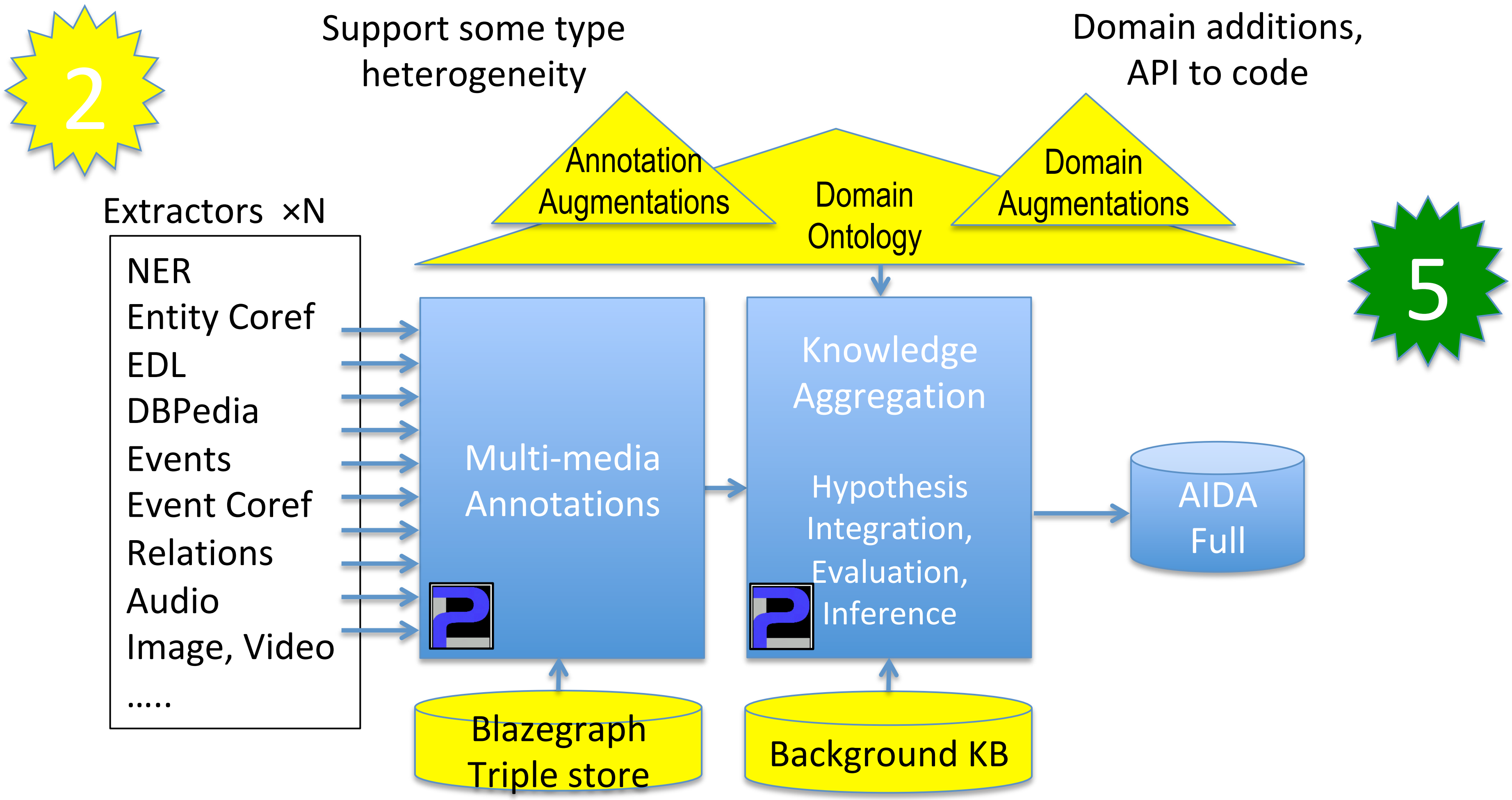
5

- **In the CSR *everything* is a hypothesis**

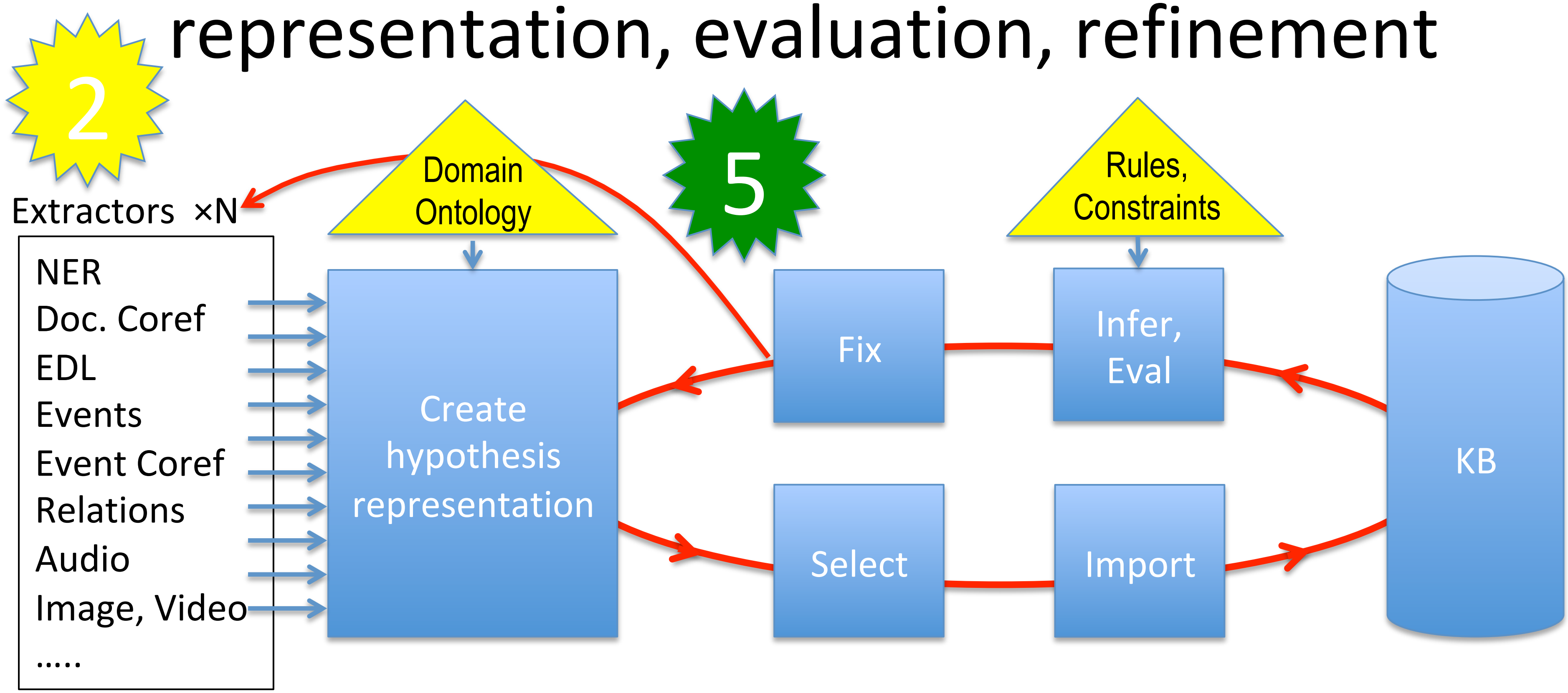
M9 Approach: 3-step decoupling for KB construction and validation



M18 Approach: Single augmented ontology for KB construction and validation



Incremental cycle of hypothesis representation, evaluation, refinement



- Cycle:
  - Use corefs and other identity to connect annotations (mention overlap, name links, EDL, within-doc coref, event coref)
  - Apply inferences, evaluate constraints, detect conflicts, do attribution
  - Fix conflicts "Viktor Yanukovych" ?= "Viktor Viktorovych Yanukovych" — no : irreflexive(parent)

TA1/2 KB integration challenges

- Challenges: Ontological
  - Multiple type systems: NER types, relation types, event types, KB schemas, target schemas…
  - Missing types, conflicting types once things are linked
  - Types, even if fine-grained, primarily useful as constraints, not as equality signal – “Humvee17 *generally-not-equal-to* Humvee42”
  - Inference requirements: “Donechyna” and “Ukraine” are compatible locations of an event but not with respect to having “Donetsk” as their capital
  - Ontological “fluidity” — things change until late in the game

5

- Challenges: Data sparsity and noise
  - Multi-lingual names and cross-lingual matching
  - Language-specific naming schemes (e.g., patronyms)
  - Cross-lingual use of context vectors
  - No fine-grained document, text or media context allowed across documents
  - Linking decisions aggregate support and ontological conflict which propagates

4

5

TA1 scores

TA1 Class queries

Best MAP	Worst MAP	TREC MAP
0.4843	0.4737	0.4773
0.4527	0.3697	0.4020
0.4379	0.2816	0.3278
0.4243	0.1470	0.1957
0.2290	0.0892	0.1244

OPERA

TA1 Graph queries

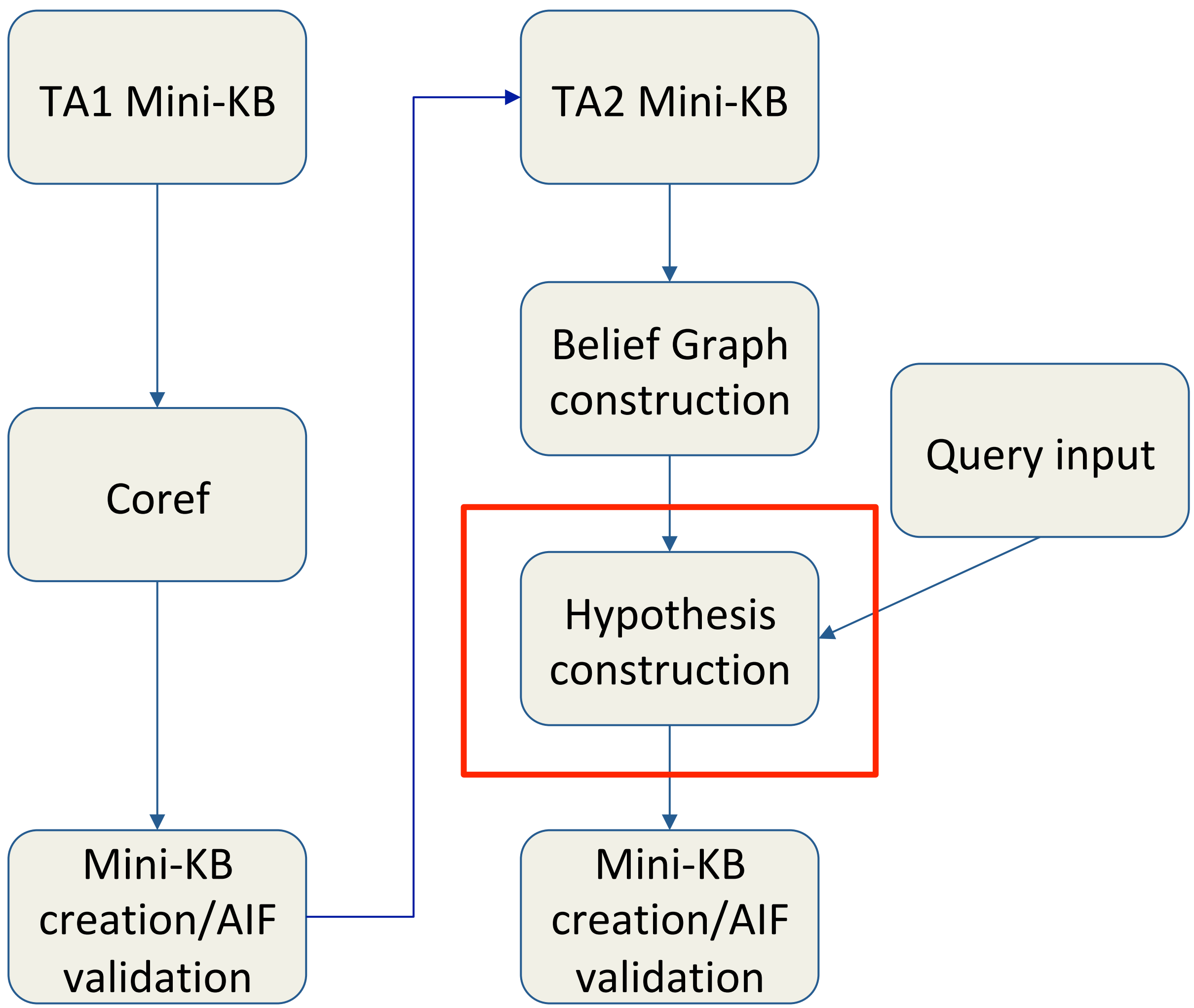
Prec	Recall	F1
0.4715	0.2163	0.2966
0.4944	0.1328	0.2094
0.3605	0.0533	0.0929
0.0398	0.0312	0.0350
0.0138	0.0040	0.0062

Run: TA1a_OPERA_TA1a_aditi_V2

Aditi Chaudhary, Anatole Gershman, Jaime Carbonell

TA3 HYPOTHESIS CONSTRUCTION

OPERA TA2 + TA3 framework



Candidate hypothesis generation

1. (Have completed Belief Graph and belief score propagation throughout)

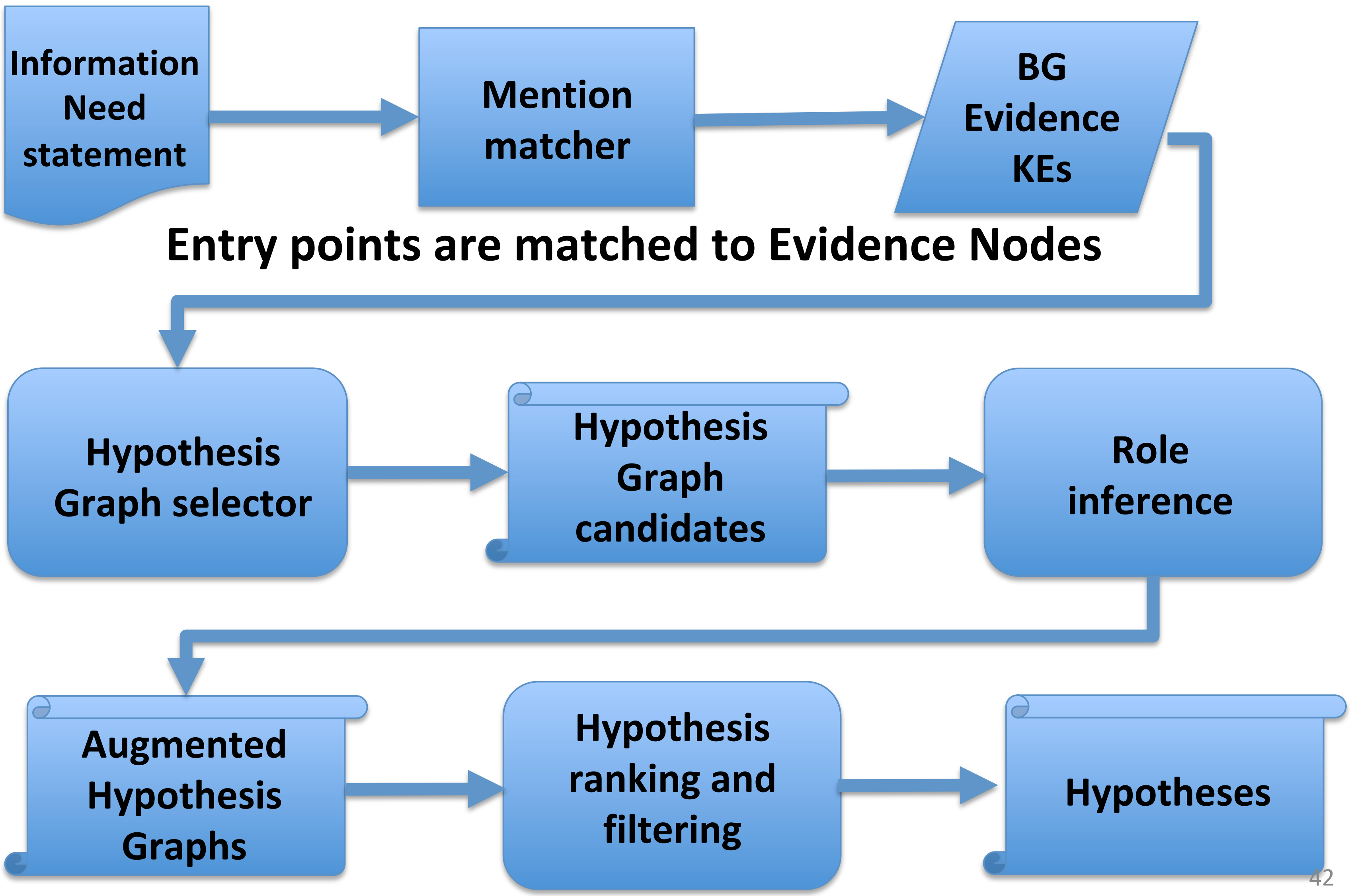
2. Retrieve KEs corresponding to entrypoints

3. Retrieve events E and relations R that match constraints. If:
   - Zero-hop: the retrieved event is one hypothesis
   - One-hop: obtain an event for every role. Prioritize events/relations with maximum overlap with the roles — this may give many permutations

5

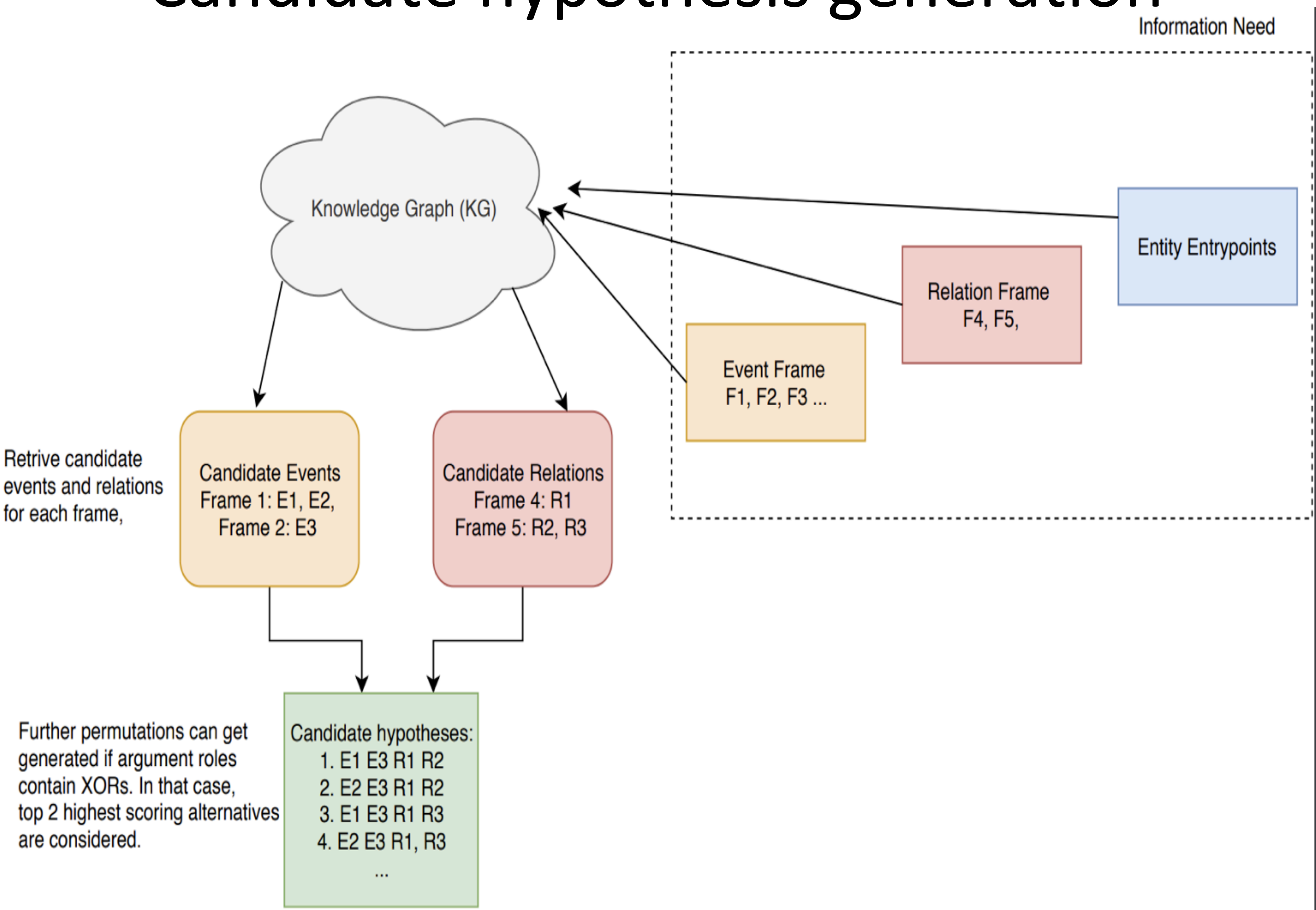
4. Generate hypothesis candidate set H = h_1, h_2 ... h_n from the retrieved E and R

Approach



Information Need statement
Mention matcher
BG Evidence KEs
Entry points are matched to Evidence Nodes
Hypothesis Graph selector
Hypothesis Graph candidates
Role inference
Augmented Hypothesis Graphs
Hypothesis ranking and filtering
Hypotheses

Candidate hypothesis generation



Hypothesis ranking

- Given information need I and candidate set $H = h_1, h_2 \ldots h_n$, we need to rank H based on <u>relevance</u> and <u>diversity</u>

- Maximal marginal relevance:

  MMR =

  $\lambda \ \text{argmax}_{hi,E,H} \ \text{Sim}(h_i, I) \ - \ (1\text{-}\lambda) \ \text{argmax}_{hj,E,H} \ \text{Sim}(h_i, h_j)$

  – $\text{Sim}(h_i, I)$ = similarity score between hypothesis h_i and the information need I — gives <u>relevance</u>
  – $\text{Sim}(h_i, h_j)$ = similarity score between hypotheses h_i and h_j — gives <u>diversity</u>

Relevance and diversity

- $Sim(h_i, I)$ = Measuring relevance … sum of:
  - Percentage of frames covered in I
  - Percentage of events satisfying the event frames
  - Percentage of relations satisfying the relation frames
  - Number of role-entity exact match constraints

- $Sim(h_i, h_j)$ = Measuring diversity (= inverse similarity) between two hypotheses … sum of:
  - Number of overlapping events
  - Number of overlapping relations
  - Number of overlapping entities
  - Number of overlapping arguments for the asked frames

TA3 M18 evaluation

- This was an *extremely* complex task
- We received a lot of numbers
- We’re still analyzing them

- Most of the numbers are not helpful for us
- Many things confuse us
- We wish for more detail about certain aspects

Task 3a (using own/any TA2 KB)

OPERA

Hypos submit-ted	Theories matched	Correct-ness	Edge cohe-rence	KE cohe-rence	Rel strict	Rel lenient	Coverage
24	6	0.4393	0.4834	0.6655	0.2832	0.6554	0.0320
42	4	0.2607	0.2894	0.423	0.1343	0.4192	0.0127
7	1	1.0000	1.0000	1.0000	1.0000	1.0000	0.0035
2	1	0.4167	0.4167	1.0000	0	1.0000	0.0032
42	1	0.3864	0.4475	0.5851	0.3295	0.4836	0.0032

Task 3a (using other TA2 teams' KBs)

Hypos submit-ted	Theories matched	Correct-ness	Edge cohe-rence	KE cohe-rence	Rel strict	Rel lenient	Coverage
34	2	0.1042	0.2178	0.3711	0.1002	0.3512	0.0079
20	2	0.5107	0.6072	0.8145	0.3823	0.7973	0.0061
7	1	1.0000	1.0000	1.0000	1.0000	1.0000	0.0035
2	1	0.4167	0.4167	1.0000	0	1.0000	0.0032
42	1	0.3864	0.4475	0.5851	0.3295	0.4836	0.0032

Task 3b (using LDC KB)

Hypos submit-ted	Theories matched	Correct-ness	Edge cohe-rence	KE cohe-rence	Rel strict	Rel lenient	Coverage
42	2	0.5961	0.6249	0.839	0.3829	0.839	0.0238
45	6	0.8065	0.8069	0.9589	0.6263	0.9589	0.0153
42	4	0.8266	0.8612	0.8979	0.8312	0.9034	0.0100
24	2	0.8207	0.8406	1.0000	0.5905	1.0000	0.0060
29	1	0.8925	0.9063	1.0000	0.7421	1.0000	0.0051

Hypothesis assessment procedure

- Assessment procedure:
  - First assess each edge as correct/incorrect
  - For only correct ones, match against the gold prevailing theory
- How assess/match? Decisions:
  - Type and informative mention of Edge
  - Type and informative mention of Left side
  - Type and informative mention of Right side

Defined in ontology. Small differences.

Undefined. Many differences of opinion

Confusion #1: Initial edge filtering

- Difference in assessed hypothesis scores on **same gold-standard LDC KB input**:

	Edges correct	Edges submitted
OPERA	59.6%	2545
BBN	80.6%	908
GAIA	82.1%	571
UTexas	82.6%	1079
PNNL	89.3%	394
LDC on TA2	0.59 Precision	

- Why the discrepancies? Our SIN-driven hypothesis creation was different. But why does LDC's own Precision not get up to .80?

Confusion #2:

- Why did GAIA do a lot better on GAIA's own KBs than on LDC's KBs?

coverage

0.0320	_version2_QueryTypeBthroughE_GAIA_1.GAIA_2.GAIA_2_v2	GAIA2_v2 running on GAIA KBs
0.0248	_version4_QueryTypeBthroughE_GAIA_1.GAIA_2p.GAIA_2	GAIA2 running on GAIA KBs
0.0238	_version2_QueryTypeBthroughE_LDC_2.LDC_2.OPERA_TA3b_2	OPERA running on LDC KBs
0.0153	_version4_QueryTypeBthroughE_LDC_2.LDC_2.BBN_TA3_v2a	BBN running on LDC KBs
0.0127	_version2_QueryTypeBthroughE_OPERA_TA1a_hans_V3.OPERA_TA2_hans_V5.OPERA_TA3a_2	OPERA running on OPERA KBs
0.0100	_version3_QueryTypeBthroughE_LDC_2.LDC_2.UTexas_3	UTexas running on LDC KBs
0.0079	_version3_QueryTypeBthroughE_GAIA_1.GAIA_2.OPERA_TA3a_1	OPERA running on GAIA KBs
0.0061	_version2_QueryTypeBthroughE_BBN_1.BBN_TA2_v2.GAIA_2	GAIA running on BBN KBs
0.0060	_version2_QueryTypeBthroughE_LDC_2.LDC_2.GAIA_2	GAIA running on LDC KBs
0.0051	_version3_QueryTypeBthroughE_LDC_2.LDC_2.PNNL_sheafbox_10	PNNL running on LDC KBs

Confusion #3: Informative mentions

evt/rel: data:relation-instance-HYP-E102-3-r201907150216-23
type: ldcOnt:**Physical.LocatedNear**
handle: woman of Odessa
edge prov: HC000Q7MI:(5700-0)-(5714-0)
woman of Odessa

----

edge: ldcOnt:**Physical.LocatedNear_EntityOrFiller**
arg: data:entity-instance-HYP-E102-3-r201907150216-0
handle: the strangled woman of Odessa, who for pro-Russians has become a symbol of the Wests partiality in the Ukrainian crisis
assessed: CORRECT

----

edge: ldcOnt:**Physical.LocatedNear_Place**
arg: data:entity-instance-HYP-E102-3-r201907150216-24
handle: Odessa
assessed: WRONG

Arg 1: the woman

edge: ldcOnt:**Physical.LocatedNear_EntityOrFiller**

arg: data:entity-instance-HYP-E102-3-r201907150216-0

handle: the strangled woman of Odessa, who for pro-Russians has become a symbol of the Wests partiality in the Ukrainian crisis

conf: 1.000000

assessed: **CORRECT**

best PT: E102Theory4

arg prov: HC000Q7MI:(5686-0)-(5804-0)

the strangled woman of Odessa, who for pro-Russians has become a symbol of the Wests partiality in the Ukrainian crisis

context:

xactly what happened. This scarcity of information explains why so many rumours have emerged around >>the strangled woman of Odessa, who for pro-Russians has become a symbol of the Wests partiality in the Ukrainian crisis<<. This photo was originally posted here.

Arg 2: Odessa

edge: **ldcOnt:Physical.LocatedNear_Place**

arg: data:entity-instance-HYP-E102-3-r201907150216-24

handle: Odessa

conf: 1.000000

assessed: **WRONG**

arg prov: HC000Q7MI:(5709-0)-(5714-0)

Odessa

context:

his scarcity of information explains why so many rumours have emerged around the strangled woman of >>Odessa<<, who for pro-Russians has become a symbol of the Wests partiality in the Ukrainian crisis. This p

- Why is it wrong? Some theories:
  - Odessa is not LocatedNear Odessa, it **IS** Odessa
  - The woman was **born in** Odessa but now lives in Kiev
  - The woman was only **rumored** to have been strangled
  - …and more…

Challenges

- Hypotheses graphs are too extensive, as events are connected by at least one common argument — need to add restrictions
- Background knowledge sometimes required to link entities (e.g., *SU25 == military jet*) — perhaps pre-populate KB with background knowledge?

FINALE

Finale

- OPERA is an end-to-end system
  - Successful combination of machine learning and manual components and approaches
  - Task 3b (using LDC's KB) submission had the highest coverage
  - Managed with limited data and changing ontology
- Absorbed & processed GAIA TA1&TA2 outputs
  - Got top TA2 graph query F1 (1a) score using GAIA KBs
- Current focus:
  - (Of course) improvements everywhere
  - New domain and ontology
  - Serious integration of component-level scores

Some discussion points

1. Multiple inheritance in the ontology
2. Main role of annotated data is as examples: continuous team–LDC interaction to gt system feedback?
3. It is hard to reconstruct what exactly assessors saw. If the specific textual context that an assessor looked at for a decision is recorded, then we can
   – see the text they based their judgment on
   – maybe also get some finer classification of the error type
4. TA1b bias: Component-based re-processing of same results when given new hypotheses