From Strings to Things:
Populating Knowledge Bases from Text

Tim Finin
University of Maryland, Baltimore County

Joint work with colleagues and students at the
JHU Human Language Technology Center of Excellence
And University of Maryland, Baltimore County

2018-02-09

TL;DR

I’ll describe the NIST TAC **Knowledge Base Population** tasks and the **Kelvin** system we developed to participate in it

NIST Text Analysis Conference

- Annual evaluation workshops since 2008 on natural language processing & related applications with large test collections and common evaluation procedures
- **Knowledge Base Population** (KBP) tracks focus on building KBs from information extracted from text
  - **Cold Start KBP**: construct KB from text w/o using external KBs
  - **Entity discovery & linking**: cluster and link entity mentions
  - Slot filling
  - Slot filler validation
  - Sentiment
  - Events: discover and cluster events in text

http://nist.gov/tac

2017 TAC Cold Start KBP

- Read 90K documents: newswire articles & social media posts in English, Chinese and Spanish
- Find entity mentions, types & relations (optionally plus events & sentiment) using a shared schema
- Cluster entities & events in/across documents, link to reference KB if possible (*which George Bush*)
- Remove errors (*Obama born in Illinois*), draw sound inferences (*Malia and Sasha sisters*)
- Create graph with provenance (*+ optional confidence score*) in TAC format

Cold Start ?

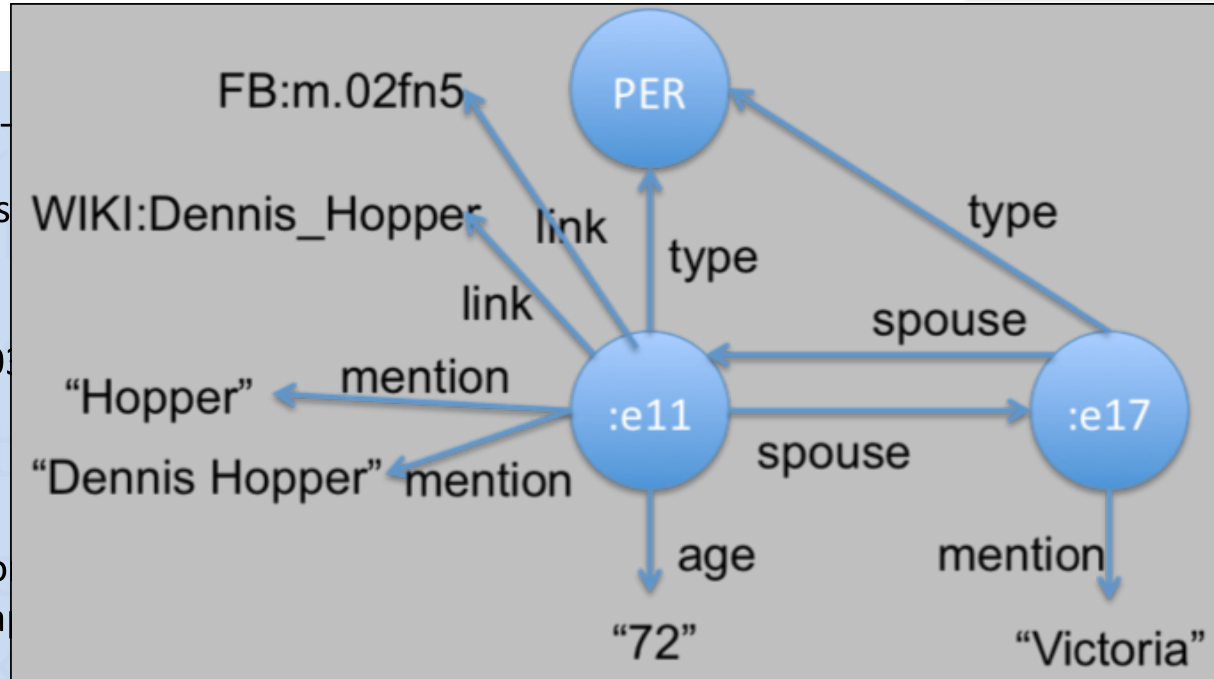
- Goal: reduce focus on popular entities common in newswire
- Start with empty KB
- All facts must be attested in text
- Can’t use external KBs (e.g., Wikidata) or Web searches

2017 TAC Cold Start KBP

```
<DOC id="APW_ENG_20100325.0021" type="story" >
<HEADLINE>
Divorce attorney says Dennis Hopper is dying
</HEADLINE>
<DATELINE>
LOS ANGELES 2010-03-25 00:15:51 UTC
</DATELINE>
<TEXT
<P>
Dennis Hopper's divorce attorney says in a court filing that the actor is dying and can't
undergo chemotherapy as he battles prostate cancer.
</P>
<P>
Attorney Joseph Mannis described the "Easy Rider" star's grave condition in a
declaration filed Wednesday in Los Angeles Superior Court.
</P>
<P>
Mannis and attorneys for Hopper's wife Victoria are fighting over when and whether to
take the actor's deposition.
</P> …
```

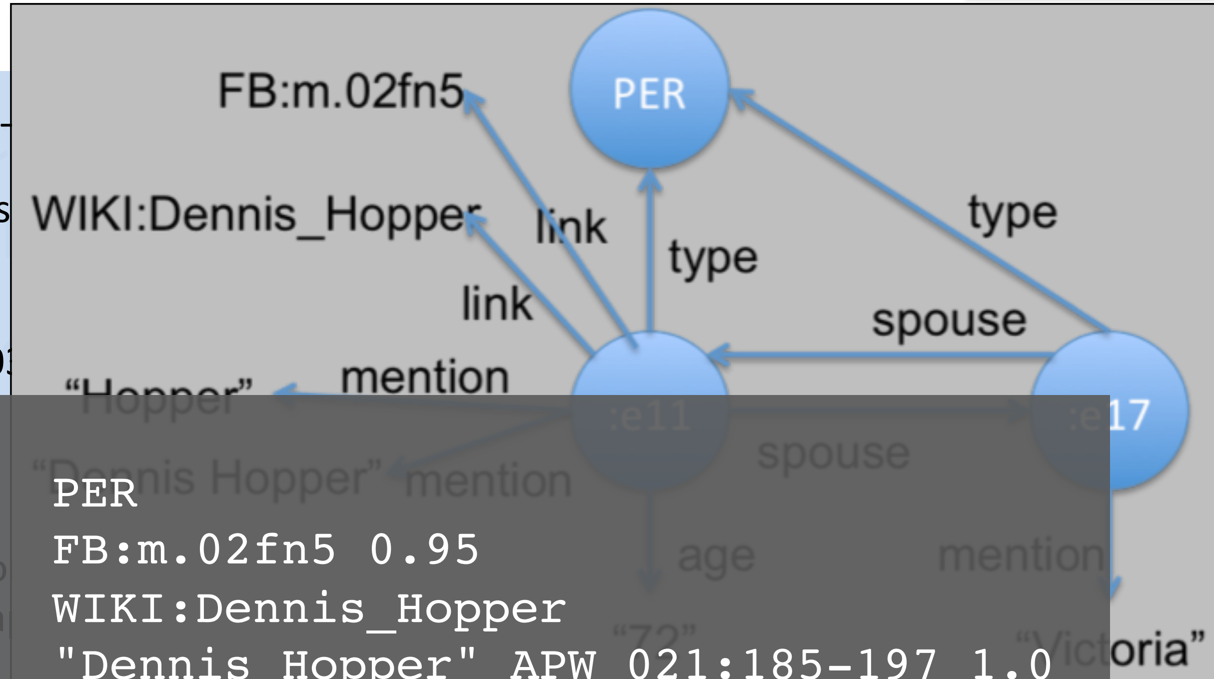
2017 TAC Cold Start KBP

<DOC id="APW_ENG_
<HEADLINE>
Divorce attorney says
</HEADLINE>
<DATELINE>
LOS ANGELES 2010-0
</DATELINE>
<TEXT
<P>
Dennis Hopper's divo
undergo chemothera
</P>
<P>
Attorney Joseph Mannis described the "Easy Rider" star's grave condition in a declaration filed Wednesday in Los Angeles Superior Court.
</P>
<P>
Mannis and attorneys for Hopper's wife Victoria are fighting over when and whether to take the actor's deposition.
</P> …



2017 TAC Cold Start KBP

<DOC id="APW_ENG_
<HEADLINE>
Divorce attorney says
</HEADLINE>
<DATELINE>
LOS ANGELES 2010-0
</DATELINE>
<TEXT>




```
…
:e00211 type        PER
:e00211 link        FB:m.02fn5 0.95
:e00211 link        WIKI:Dennis_Hopper
:e00211 mention     "Dennis Hopper" APW_021:185-197 1.0
:e00211 mention     "Hopper"     APW_021:507-512  1.0
:e00211 mention     "Hopper"     APW_021:618-623  1.0
:e00211 mention     "丹尼斯·霍珀" CMN_011:930-936  1.0
:e00211 per:spouse :e00217      APW_021:521-528  1.0
:e00217 per:spouse :e00211      APW_021:521-528  1.0
:e00211 per:age     "72"        APW_021:521-528  0.9
…
```

Mannis and attorneys for Hopper's wife Victoria are fighting over when and whether to take the actor's deposition.
</P> ...

KB Evaluation Methodology

- Evaluating KBs extracted from 90K documents is non-trivial
- TAC's approach is simplified:
  - **Fix the ontology** of entity types and relations
  - Specify **a serialization** as triples + provenance
  - Sample a KB using a set of **queries** grounded in an *entity mention* found in a document
  - Get ground truth for queries and assess results
- Given a KB, we can then evaluate its **precision and recall** for a set of queries

KB Evaluation Methodology

- **A query:** What schools were attended by children of entity mentioned in document #45611 at characters 401-412
  - That mention is *George Bush* which a system under test identifies as :e629 (i.e., G.H.W. Bush)
  - A query finds answer entities in a test system's graph (e.g., Yale, Harvard, Tulane, UT Austin, UVA …) along with the provenance strings for the two relations
- **Assessors** determine good answers in corpus and check the submitted results' **provenance**

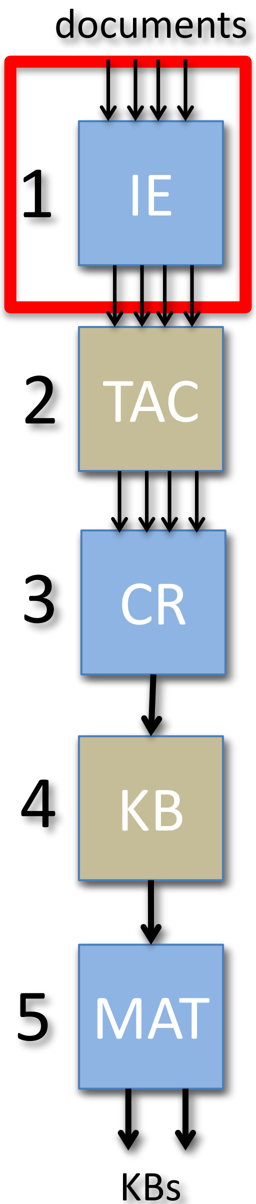
TAC Ontology

- Derived from the Automatic Content Extraction (ACE) ontology
- Entity **types**: PER (people), ORG (organizations), GPE (geopolitical entities), LOC (locations) and FAC (facilities)
- Entity **mentions**: both name & nominal mentions
- 41 **relations** *(plus inverses):* entity to entity/string
- 18 **event types**: plus 85 event argument relations
- 2 **sentiment relations** *(plus inverses):* entity to entity

Kelvin

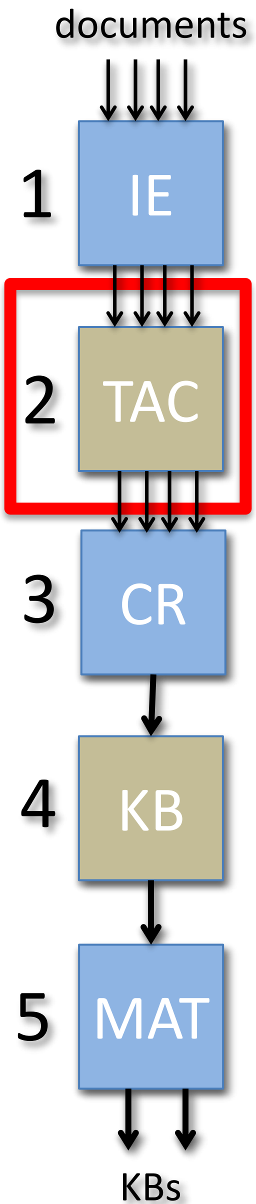
- **KELVIN**: **K**nowledge **E**xtraction, **L**inking, **V**alidation and **In**ference
- Developed at the ***Human Language Technology Center of Excellence*** at JHU and used in TAC KBP (2010-17), EDL (2015-17) and other projects
- Takes English, Chinese & Spanish documents and produce a knowledge graph in several formats
- We'll review its monolingual pipeline, look at the multi-lingual use case

1 Information Extraction



- Process documents in **parallel** on a grid
- Apply an ensemble of NLP tools (e.g., language ID, Serif, CoreNLP, …) to find **document-level** mentions, entities, relations and events
- Produce an **Apache Thrift** object for each document with text and extracted data using **Concrete,** a common NLP schema

2 Integrating NLP data

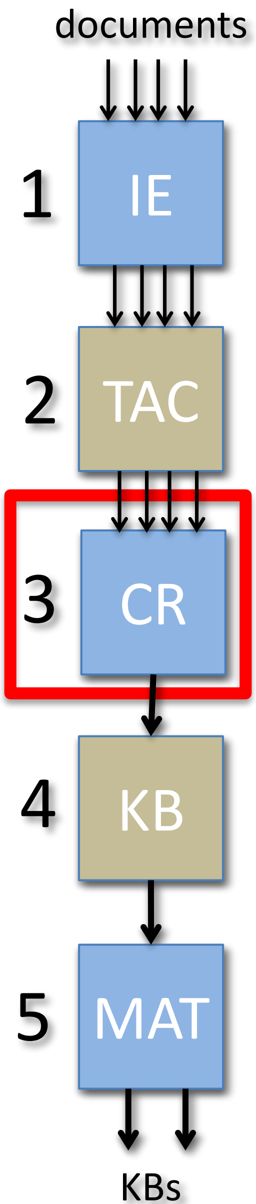


Process Concrete objects in parallel to:

- **Integrate** data from tools (e.g., Serif, CoreNLP, OpenIE)
- **Fix problems**, e.g., trim mentions, find missed mentions, deconflict tangled mention chains, …
- Extract relations from **events** (life.born => date and place of birth)
- Map relations found by open IE systems to TAC ontology (*"is engineer at" => per:employee_of)*
- Map schema to our extended **TAC ontology**

30K ENG: 430K entities; 1.8M relations

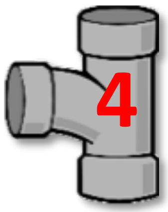
3 Kripke: Cross-Doc Coref

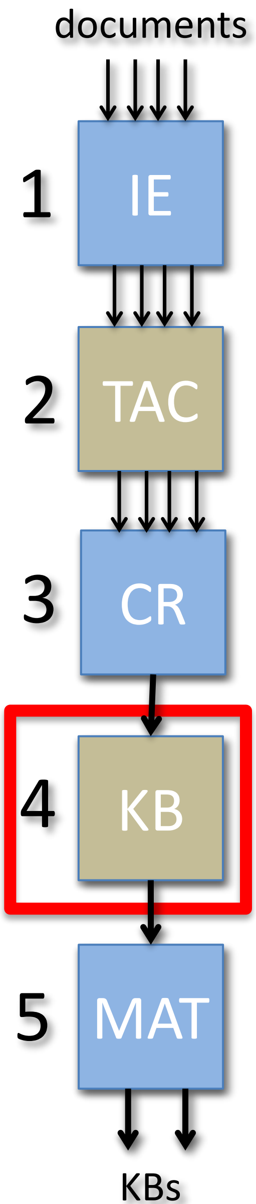


- Cross-document **co-reference** creates initial KB from the document-level data
  - Identify that *Barack Obama* entity in DOC32 is same individual as *Obama* in DOC342, etc.
- **Language agnostic**; works well for ENG, CMN, SPA document collections
- Uses entity **type** and **mention strings** and **context** of co-mentioned entities
- Untrained, agglomerative **clustering**

30K ENG: 210K entities; 1.2M relations

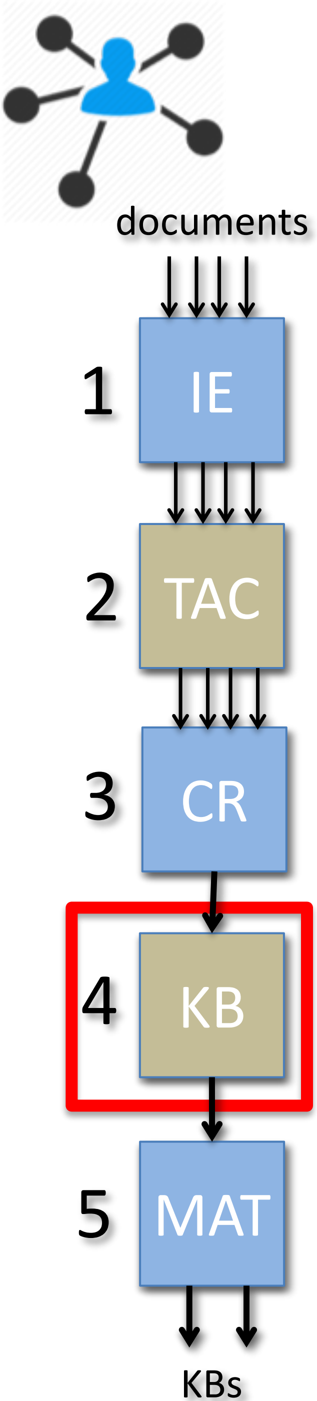
4 Inference & adjudication





Reasoning to

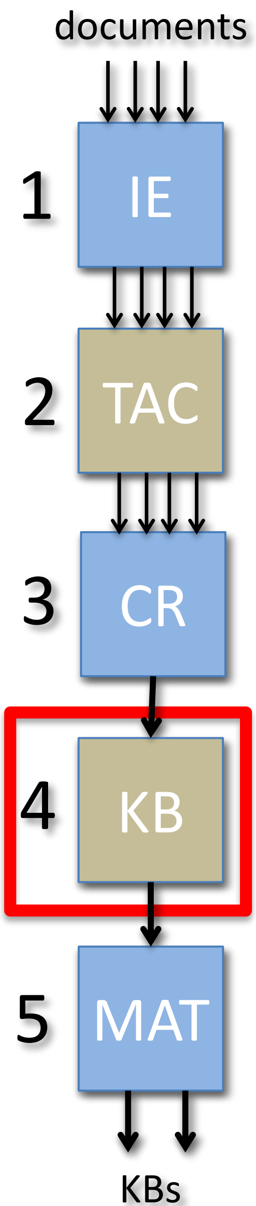
- Delete relations violating ontology constraints
  - *Person can't be born in an organization*
  - *Person can't be her own parent or spouse*
- Infer missing relations
  - *Two people sharing a parent are siblings*
  - *X born in place P_1, P_1 part of P_2 => X born in P_2*
  - *Person probably citizen of their country of birth*
  - *A CFO is a per:top_level_employee*



Entity Linking

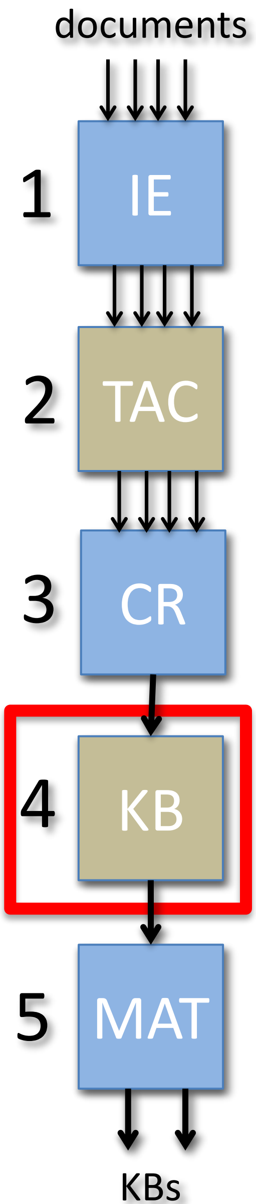
- Try to links entities to reference KB, a subset of Freebase with
  - ~4.5M entities and ~150M triples
  - Names and text in English, Spanish and Chinese
- Don't link if no matches, poor matches or ambiguous matches

KB-level merging rules



- Merge entities linked to same external KB entity
- Merge cities in same region with same name
- Highly discriminative relations give evidence of sameness
  - per:spouse is few to few
  - org:top_level_employee is few to few
- Merge PERs with similar names who were
  - Both married to the same person, or
  - Both CEOs of the same company, or …

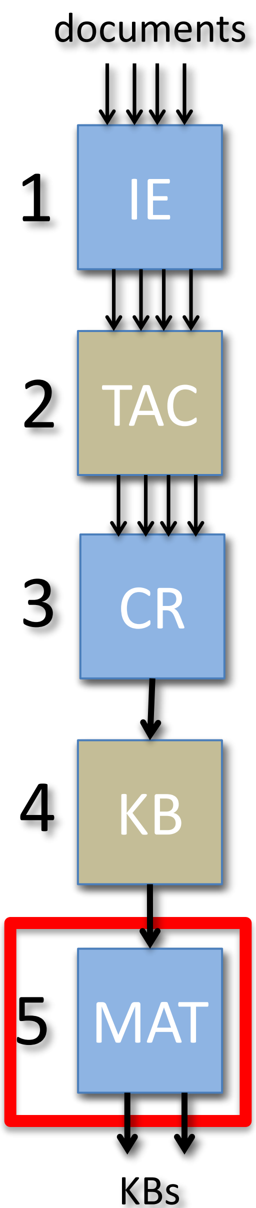
Slot Value Consolidation



- **Problem:** too many values for some slots, especially for ‘popular’ entities, e.g.,
  - An entity with 2 ***per:city_of_birth*** values
  - Obama had ~100 ***per:employee_of*** values
- **Strategy:** rank values and select best
  - Rank values by # of attesting docs and certainty scores
  - Choose best N values depending on relation type and distribution of frequency counts

30K ENG: 183K entities; 2.1M relations

Materialize KB versions



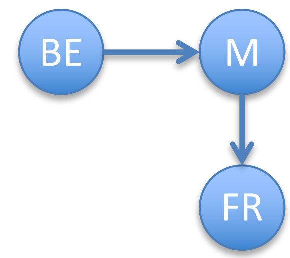
- Generate TAC serialization for scoring
- Also encode KB in a database or graph store, e.g. the RDF/OWL Semantic Web languages or

Multilingual KBP

- Many examples where facts from different languages combine to answer queries or support inference

Q: Who lives in the same city as *Bodo Elleke*?

A: *Frank Ribery* aka *Franck Ribéry* aka *里贝里*

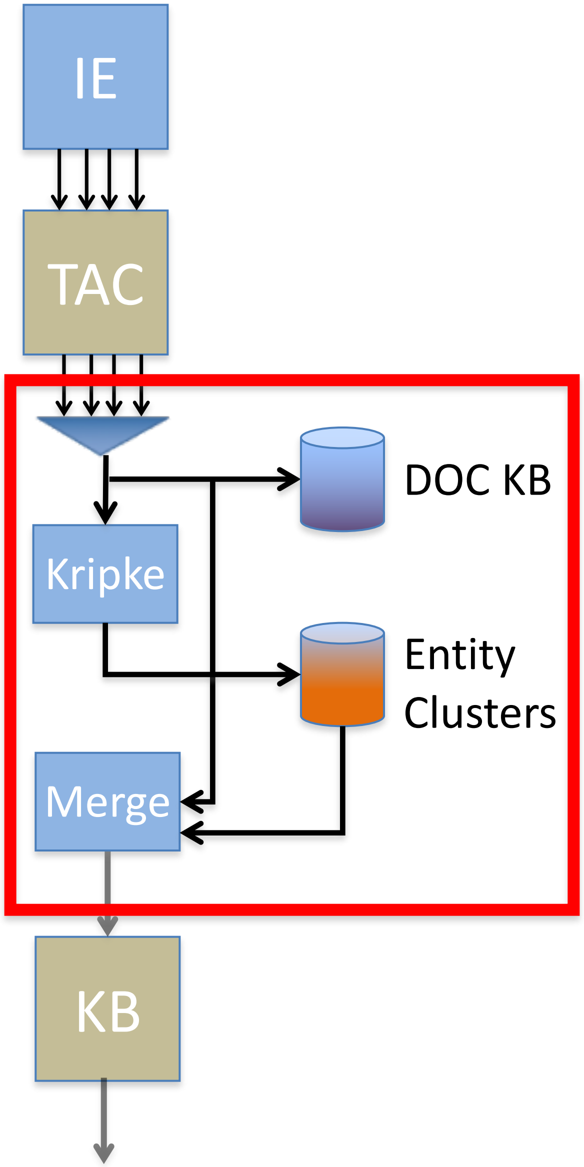


- Why we know both live in Munich:

1. :e8 gpe:residents_of_city :e23 ENG_3:3217-3235
   ...said the younger **Bodo Elleke**, who was born in Schodack in 1930 and is now a retired architect **who lives in Munich**.
2. :e8 gpe:residents_of_city :e25 CMN...0UTJ:292-361
   拉霍伊在接受西班牙国家电台的采访时肯定，今年的三位金球奖热门候选人中，梅西“度过了一个出色的赛季”，而拜仁**慕尼黑球员里贝里**则“赢得了一切”

- Kripke merged entities with mentions *Frank Ribery*, *Franck Ribéry* & *里贝里*

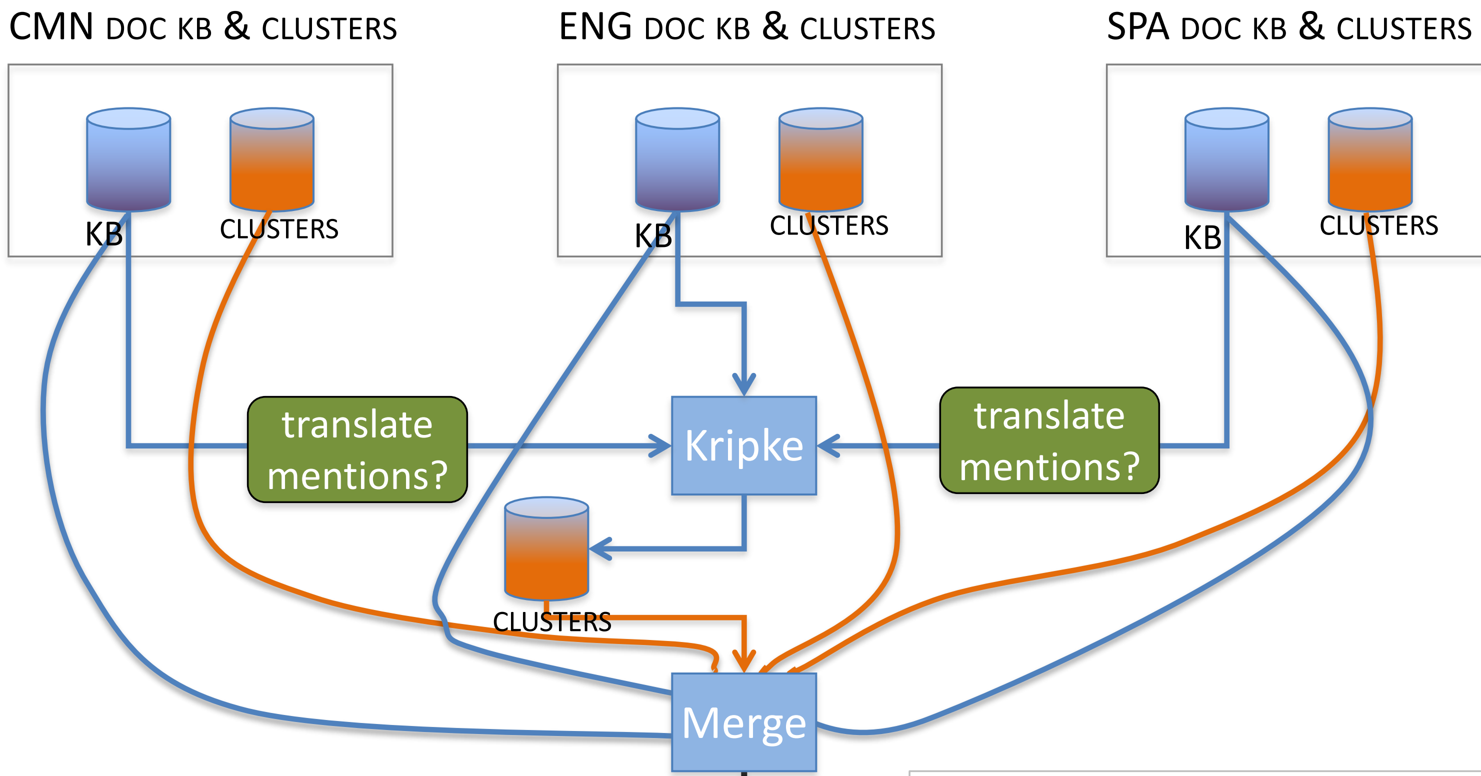
Monolingual to Multilingual Kelvin

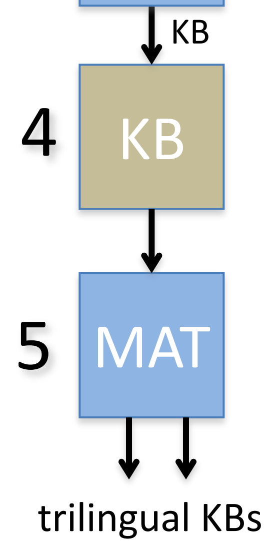


Zoom in on our cross-doc co-ref step

- Concatenate document-level KBs to form a **DOC KB** as input to Kripke
- Kripke outputs a set of **CLUSTERS** defining an equivalence relation
- Merger uses **CLUSTERS** to combine **DOC KB** entities, yielding the initial KB
- We use the **DOC KB** and **CLUSTERS** from each language to create an initial multilingual KB

Trilingual KBP





- **Run Kelvin on monolingual collections**
- **Translate entity mentions into English and recluster**
- **Run results thru rest of pipeline**

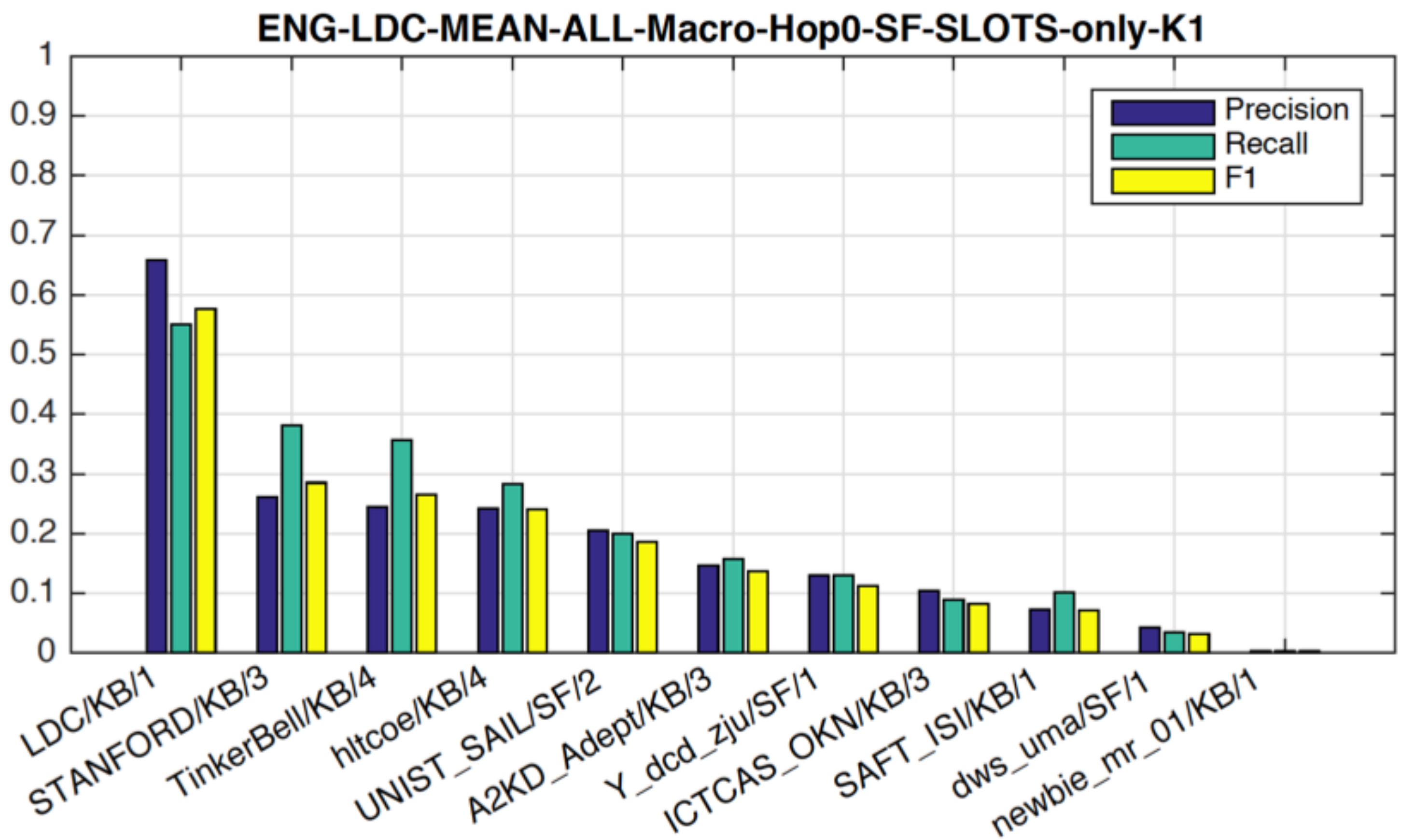
- Kripke computes CLUSTERS for combined monolingual DOC KBs
- Optionally translates non-English mentions
- Use all 4 CLUSTERS to merge entities in 3 DOC KBs

2016 TAC KBP Results

- 2016 KBP results (2017 KBP results similar)
  - 1st or 2nd on XLING
  - 2nd or 4th on ENG depending on metric
  - 1st or 2nd on CMN depending on metric
  - We did poorly on SPA, finding few relations
- Lots of room for improvement for both *precision* and *recall*

The task is hard

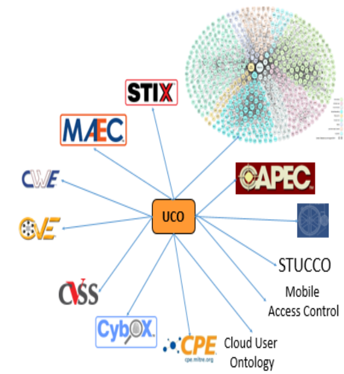
Best 2017 system: F1=0.29 for English hop 0 queries.



Current work 1: improving Kelvin

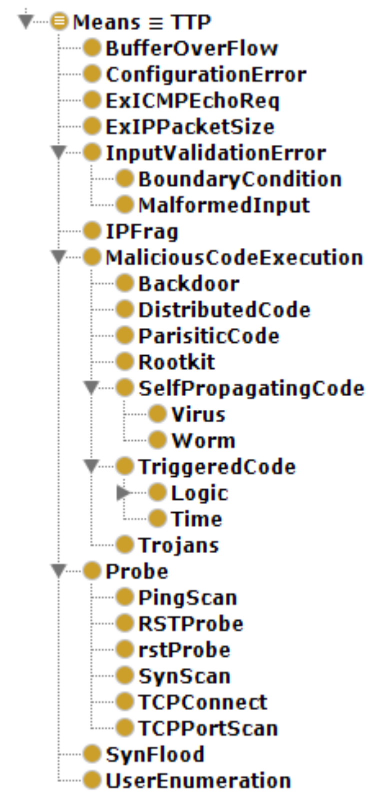
- Upgrade components to use newer machine learning approaches
- Enhance Kripke entity clustering with more data (nominal mentions, embeddings)
- Add tensor-decomposition based learning to identify likely/unlikely relations
- Add other components to detect and fix "dubious facts"

Current work 2: cybersecurity



UMBC is working with IBM on extracting cybersecurity information from text

- Describe entities, relations & events using UCO, the Unified Cybersecurity Ontology
  - Rich schema supports reasoning
  - Better data sharing, interoperability, integration and human understanding
  - Link to background knowledge graphs and common metadata models (CVE, Stix, Cybox…)
- Use graph to enhance analytics and machine learning for intrusion detection systems



http://unifiedCyberOntology.github.io

Lessons Learned

- We always have to mind precision & recall
- Extracting information from text is inherently noisy; reading more text helps both
- Using machine learning at every level is important
- Making more use of probabilities will help
- Extracting information about a events is hard
- Modelling the temporal extent of relations is important, but still a challenge

For more information,
contact finin@umbc.edu