From Strings to Things

Populating Knowledge Bases from Text
The Web is our greatest knowledge source
But it has limitations
It was designed for people, not machines
It was designed for people, not machines
• Its content is mostly text, spoken language, images and videos
• These are easy for people to understand
• But hard for machines

Machines need access to this knowledge too
Access is primarily via information retrieval

Vannevar Bush envisioned a hypertext/IR system in 1945
Access is primarily via information retrieval
• Key-word queries→ranked document list
• We still need to read the documents or watch the videos
• We often want an answer to a question

And so do our machines and apps

Vannevar Bush envisioned a hypertext/IR system in 1945
We need to add knowledge graphs
We need to add knowledge graphs
• High quality semi-structured information about entities, events and relations
• Represented & accessed via standard APIs
• Easily integrated, fused and reasoned with
State of the Art?

Google is a good example, but Microsoft, IBM, Apple and Facebook all have similar capabilities

• 2010 Google acquired MediaWeb and its Freebase KB
• 2014: Freebase: 1.2B facts about 43M entities
• 2015+: Google knowledge graph, updated by text IE

DBpedia open source RDF KB is another

• 800M facts about 4.6M subjects from English Wikipedia, data also available in 21 other languages
• Helps integrate 90B facts from 1000 RDF datasets in the linked data cloud
Ask: When was Tom Sawyer written?

The Adventures of Tom Sawyer / Date written

1876
Apple Pie by Grandma Ople

Recipe by: MOSHASMAMA

“This was my grandmother’s apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well!”

Ingredients

1 recipe pastry for a 9 inch double crust pie
1/2 cup white sugar
1/2 cup unsalted butter
Domino Pure Cane Granulated Sugar

On Sale

What’s on sale near you.
Almost all commercial recipe sites embed semantic data about their recipes in an RDF-compatible form using terms from the schema.org ontology. Search engines read and use this data to better understand the semantics of the page content.
Conversational Bots

Voice-driven conversational systems like Amazon Echo and Google Home use knowledge graphs to help understand our requests.
Where does the knowledge come from?

- Initial knowledge graphs like DBpedia and Freebase started with data from Wikipedia and encoded it in custom ontologies.
- Current focus is on extracting information from text of source documents, e.g., journal articles, Newswire, social media, etc.
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 a KB from text
  • **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
2016 TAC Cold Start KBP

• Read 90K documents: newswire articles & social media posts in English, Chinese and Spanish
• Find entity mentions, types and relations
• Cluster entities within and across documents and link to a reference KB when appropriate
• Remove errors (*Obama born in Illinois*), draw sound inferences (*Malia and Sasha sisters*)
• Create knowledge graph with provenance data for entities, mentions and relations
2016 TAC Cold Start KBP

- Read 90K documents: newswire articles & social media posts in English, Chinese, and Spanish
- Find entity mentions, types, and relations
- Cluster entities within and across documents and link to a reference KB when appropriate
- Remove errors (Obama born in Illinois), draw sound inferences (Malia and Sasha sisters)
- Create knowledge graph with provenance data

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

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.

Attorney Joseph Mannis described the "Easy Rider" star's grave condition in a declaration filed Wednesday in Los Angeles Superior Court.

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

...
Kelvin

• **KELVIN**: Knowledge Extraction, Linking, Validation and Inference

• Developed at the *Human Language Technology Center of Excellence* at JHU and used in TAC KBP (2010-16), EDL (2015-16) and other projects

• Takes English, Chinese & Spanish documents and produce a knowledge graph in several formats

• We’ll review its monolingual processing, look at the multi-lingual use case
1 Information Extraction

- Process documents in **parallel** on a grid, applying information extraction tools to find mentions, entities, relations and events.

- Produce an **Apache Thrift** object for each document with text and relevant data produced by tools using a common **Concrete** schema for NLP data.
2 Integrating NLP data

Process Concrete objects in parallel to:

• **Integrate** data from tools (e.g., Stanford, Serif)

• **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 schema to extended **TAC ontology**

**30K ENG: 430K entities; 1.8M relations**
3 Kripke: Cross-Doc Coref

- Cross-document co-reference creates initial KB from a set of single-document KBs
  - 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
- Only uses entity mention strings
- Untrained, agglomerative clustering

30K ENG: 210K entities; 1.2M relations
4 Inference and 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 in 2016 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 of same type linked to same 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 four different *per:age* values
  – Obama has ~100 *per:employee_of* values

• **Strategy:** rank values and select best
  – Rank values by # of attesting docs and probability
  – Choose best N value depending on relation type

30K ENG: 183K entities; 2.1M relations
Materialize KB versions

- Encode KB in your favorite database or graph store
- We like the RDF/OWL Semantic Web technology stack
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:

     ...said the younger Bodo Elleke, who was born in Schodack in 1930 and is now a retired architect who lives in Munich.

     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
2016 TAC KBP Results

For the 2016 KBP submissions, depending on metric, we placed
- 1\textsuperscript{st} or 2\textsuperscript{nd} on XLING and were the only team to do all three languages
- 2\textsuperscript{nd} or 4\textsuperscript{th} on ENG depending on metric
- 1\textsuperscript{st} or 2\textsuperscript{nd} on CMN depending on metric
- We did poorly on SPA, finding few relations

Lots of room for improvement for both \textit{precision} and \textit{recall}
An application:
Cybersecurity strings to things

UMBC is working with IBM to develop systems to extract cybersecurity information from text

• **Find** entities and their properties, relations & events
• **Encode** as knowledge graphs with *evidence & certainty*
• **Recognize** entities & events referring to same things and **link** to background knowledge graphs if possible
• **Reason** over graphs to improve and assess accuracy, coherence and trustworthiness
• **Support** analytics and machine learning systems
Cyber situational awareness

• Most IDS systems are point-based & driven by known signatures
• Our situationally-aware system maps multiple sensors to a common ontology,
• Reasons over the resulting knowledge,
• Detecting possible intrusions missed by standard systems
Approach

• **Leverage** existing and new tools for information extraction, semantic similarity, inference, etc.

• Evolve our **Unified Cyber Ontology** as the underlying semantic model

• Develop, curate & annotate **cybersecurity corpora** from alerts, newswire, social media & chatrooms

• **Train systems** for knowledge graph population, concept spotting, entity recognition, relation and event extraction, word embeddings, topic modeling, etc.
Unified Cybersecurity Ontology

• Common semantic model for cybersecurity domain
  – Data sharing, interoperability, integration and human understanding
  – Links to background knowledge graphs
  – Maps to common metadata schemas like Stix and Cybox

• Uses semantically rich representation
  – Grounded in formal semantics

• Supports reasoning
  – Infer/retrieve new information & detect dubious facts
Information extraction from text

We use information extraction techniques to identify entities, relations and concepts in security related text.

These are mapped to terms in our ontology and the DBpedia knowledge base extracted from Wikipedia.

http://dbpedia.org/resource/Buffer_overflow

http://dbpedia.org/resource/Arbitrary_code_execution

http://dbpedia.org/resource/Microsoft_Windows_Vista


http://dbpedia.org/resource/Windows_7

http://ebiq.org/p/540
“John Smith lives in Baltimore, Maryland. He is married to Mary Jones. She works at Loyola University where she is a professor. The university is in Baltimore.”
John Smith lives in Baltimore, Maryland. He is married to Mary Jones. She works at Loyola University where she is a professor. The university is in Baltimore.

docid: "text1.txt",
corefs: {
  "9": [
    {"endIndex": 6, 
      "animacy": "INANIMATE", 
      "text": "Baltimore", 
      "isRepresentativeMention": true, 
      "number": "SINGULAR", 
      "startIndex": 5, 
      "sentNum": 1, 
      "gender": "NEUTRAL", 
      "position": [1, 2q], 
      "headIndex": 5, 
      "type": "PROPER", 
      "id": 1 }
  ],
  { ....

##### :e_text1_1 LOCATION "Baltimore" #####
:e_text1_1 type LOCATION
:e_text1_1 canonical_mention "Baltimore" text1:20-29
:e_text1_1 mention "Baltimore" text1:20-29
:e_text1_1 mention "Baltimore" text1:151-160

##### :e_text1_2 ORGANIZATION "Loyola University" #####
:e_text1_2 type ORGANIZATION
:e_text1_2 canonical_mention "Loyola University" text1:85-102
:e_text1_2 mention "Loyola University" text1:85-102
:e_text1_2 mention "The university" text1:130-144

##### :e_text1_3 PERSON "John Smith" #####
:e_text1_3 type PERSON
:e_text1_3 canonical_mention "John Smith" text1:0-10
:e_text1_3 mention "John Smith" text1:0-10
:e_text1_3 mention "He" text1:42-44
:e_text1_3 mention "She" text1:72-75
:e_text1_3 mention "she" text1:109-112
:e_text1_3 openie:lives_in :e_text1_1 text1:0-3
:e_text1_3 per:spouse :e_text1_5 text1:42-43
:e_text1_3 openie:is_married_to :e_text1_5 text1:42-43
:e_text1_3 per:employee_of :e_text1_2 text1:72-74
Part-of-Speech:

1. John Smith lives in Baltimore, Maryland.
2. He is married to Mary Jones.
3. She works at Loyola University where she is a professor.
4. The university is in Baltimore.

Named Entity Recognition:

1. John Smith lives in Baltimore, Maryland.
2. He is married to Mary Jones.
3. She works at Loyola University where she is a professor.
4. The university is in Baltimore.
Basic Dependencies:

1. John Smith lives in Baltimore, Maryland.

2. He is married to Mary Jones.

3. She works at Loyola University where she is a professor.

4. The university is in Baltimore.

Enhanced Dependencies:
1. John Smith lives in Baltimore, Maryland.
2. He is married to Mary Jones.
3. She works at Loyola University where she is a professor.
4. The university is in Baltimore.
Coreference:

1. John Smith lives in Baltimore, Maryland.
2. He is married to Mary Jones.
3. She works at Loyola University where she is a professor.
4. The university is in Baltimore.
John Smith lives in Baltimore, Maryland.

He is married to Mary Jones.

She works at Loyola University where she is a professor.

The university is in Baltimore.
Conclusion

• KGs help in extracting information from text
• The information extracted can update the KGs
• The KGs provide support for new tasks, such as question answering, speech interfaces and produce data useful in applications, like IDSs
• There use will grow and evolve in the future
• New machine learning frameworks will result in better accuracy