NLP* from Strings to Things

*natural language processing
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
- Large knowledge graph with 1B statements about ~72M items
- Fine-grained ontology: ~2M types; ~5K properties
- Multilingual, strings tagged with language id
- Links to entity’s Wikimedia pages
- Entities have a canonical **name** and **aliases** in multiple languages and multiple claims
- UMBC=Q64780099, with type University, 569 statements
- Editable by humans and bots
- Can query with SPARQL query language
Ask: When was Tom Sawyer written?

1875
Apple Pie by Grandma Ople

9K made it | 6969 reviews

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](http://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.
  ...</TEXT>
</DOC>
Information extraction from text

- Information extraction techniques identify entities, relations and concepts in security related text
- Map to terms in our ontology and DBpedia knowledge graph
- Also map them to terms in the Wikidata knowledge graph

CVE-2012-0150
Buffer overflow in msvcrt.dll in Microsoft Windows Vista SP2, Windows Server 2008 SP2, R2, and R2 SP1, and Windows 7 Gold and SP1 allows remote attackers to execute arbitrary code via a crafted media file, aka "Msvcrt.dll Buffer Overflow Vulnerability."

http://dbpedia.org/resource/Arbitrary_code_execution
http://dbpedia.org/resource/Windows_7
http://dbpedia.org/resource/Buffer_overflow
http://ebiq.org/p/540
NLP Tools

• There is a rich and growing collection of open-source NLP tools

• Comprehensive pipelines:
  – Stanford CorNLP tools
  – Spacy
  – NLTK

• Word embeddings
  – Word2vec, BERT, Semsim
Stanford CoreNLP Tools

CoreNLP

version 4.3.2

Text to annotate:
Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland.

Annotations:

Part-of-Speech:

<table>
<thead>
<tr>
<th>1</th>
<th>NNP</th>
<th>NNP</th>
<th>VBZ</th>
<th>DT</th>
<th>NN</th>
<th>IN</th>
<th>NNP</th>
<th>DT</th>
<th>NN</th>
<th>VBN</th>
<th>IN</th>
<th>NNP</th>
<th>NNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Named Entity Recognition:

<table>
<thead>
<tr>
<th>1</th>
<th>PERSON</th>
<th>TITLE</th>
<th>ORGANIZATION</th>
<th>CITY</th>
<th>STATE_OR_PROVINCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

KBP Relations:

Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland.
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.

---

**JSON/XML => KG triples**

```json
{
  "text": "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."
}
```

---

**Coreference Resolution**

- **John Smith**: type PERSON, canonical mention "John Smith", mention "He", openie:lives_in :e_text1_1, per:spouse :e_text1_5, per:employee_of :e_text1_2
- **Mary Jones**: canonical mention "Mary Jones", mention "She", openie:is_married_to :e_text1_5, per:employee_of :e_text1_2
- **Baltimore**: type LOCATION, canonical mention "Baltimore", mention "Baltimore", openie:lives_in :e_text1_1
- **Loyola University**: type ORGANIZATION, canonical mention "Loyola University", mention "The university"
Industrial-Strength Natural Language Processing

Get things done
spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and it's easy to install, and its API is simple and productive.

Blazing fast
spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

Awesome ecosystem
In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

GET STARTED
FACTS & FIGURES
READ MORE
Learning word meaning?

• How can we learn what a word means?

• The linguist John Rupert Firth famously write in 1957
  “You shall know a word by the company it keeps”

• A way to recognize that two words have similar meanings is to note that they occur in similar contexts
  – E.g., physician & doctor, nurse & doctor, love & hate
Word Embeddings

- **Latent Semantic Analysis (LSA)** learns a vector (e.g., 300 reals 0..1) for each unique word in a corpus to represent its meaning
  - LSA also used for document *topic modelling*

- An example of **dimensionality reduction**

  - Frequency of co-occurrence of words in a 5-word window in a huge corpus
  - Each row is a vector of 300 reals for degree a word has of that topic
Sentence similarity

How similar are the two sentences semantically on a scale of 0-5?

The mouse ate some cheese
Cheddar was eaten by a rat

It’s a 4!

Pearson’s Correlation

3.824

Close enough!

We used this approach in 2013 to win in a sentence similarity task
UMBC semantic similarity service

The input word: 
Part of Speech:  ○ Noun  ○ Verb  ○ Adjective  ○ Adverb
The value of N:  ○ 10  ○ 20  ○ 30  ○ 40  ○ 50  ○ 100
Type:  ○ Concept Similarity  ○ Relation Similarity
Corpus:  ✔ Refined Stanford WebBase corpus  □ LDC English Gigawords Corpus (American newswire services only)

Get Top-N Most Similar Words
word2vec

Uses a shallow neural network to map words to a vector space where words with similar contexts have close vectors.

Chris Albon
Word2Vec

• Developed by Google also in 2013 using a neural network approach
• Two models: CBOW and skip grams
• Trained on a much larger corpus from the Web
• Models can be downloaded and are still used today
  – E.g., the spaCy NLP system includes word2vec to measure similarity
Word2vec demo

Models
Select one of the available models

English GoogleNews Negative300

Nearest words
Given a word, this demo shows a list of other words that are similar to it, i.e. nearby in the vector space.

Type in a word  Show nearest  Case sensitive:  Top N: 10

Similarity of two words
Given two words, this demo gives the similarity value between 1 and -1.

Type in a word  Type in a word  Show similarity

Word analogy
This demo computes word analogy: the first word is to the second word like the third word is to which word? Try for example ilma - lintu - vesi (air - bird - water) which would expect to return kala (fish) because fish is to water like birds is to air. Other cases could be for example sammakko - hyppää - kala. This is however only a toy to show what is possible - most of the time the analogy does not work particularly well (at least for the Finnish data).

Type in a word  Type in a word  Type in a word  Show  Top N: 2
Scientists using fMRI to measure brain activity find locations associated with similar concepts – brain embeddings!
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