Knowledge Graphs: Extracting and Leveraging Meaning from Unstructured Text

Bang Liu, Haochen Shi
April 3rd, 2023
Introduction about the teacher and teaching assistant

- **Bang Liu (刘邦)** is an Assistant Professor in the Department of Computer Science and Operations Research (DIRO) at the University of Montreal. He is a core member of the RALI laboratory (Applied Research in Computer Linguistics) of DIRO, an associate member of Mila – Quebec Artificial Intelligence Institute, and a Canada CIFAR AI (CCAI) Chair. His research interests primarily lie in the areas of natural language processing, multimodal & embodied learning, theory and techniques for AGI (e.g., understanding and improving large language models), and AI for science (e.g., health, material science, XR).

- **Haochen Shi (史皓琛)** is a Ph.D. candidate in the Department of Computer Science and Operations Research (DIRO) at the University of Montreal, supervised by Prof. Bang Liu. He received his master's degree in 2021 from Zhejiang University, as well as his bachelor's degree from Huazhong University of Science and Technology. His research interests primarily lie in the areas of natural language processing (NLP), multimodal & embodied learning, and resource-economical deep learning.
Tutorial outline

๏ Part I: Introduction to knowledge graph
  • What is a knowledge graph?
  • Why are knowledge graphs important?
  • Success stories of knowledge graphs

๏ Part II: Semantic Web Knowledge Graph
  • W3C Semantic Web
  • RDF, SPARQL, etc.

๏ Part III: Knowledge graph construction
  • A pipeline: KG source -> KG construction -> KG maintenance -> KG application

๏ Part IV: Knowledge graph applications
  • Emphasize on KG-based question answering

๏ Part V: Active Research about KG

๏ Part VI: ChatGPT and KG
Introduction to KG
Suppose you are searching…

Query: “Honda Civic, Hyundai Elantra”

What he is interested in?
Is-A relationship between things

Concept: a collection of things that share some common attributes

- Fuel-efficient cars
- Economy cars
- Marvel heroes
- Revengers
7 Suppose you are asking…

Question: “How tall is the Eiffel Tower?”

How does Google know?
# Properties of things

## The Eiffel Tower at a glance

An object of discord, desire and fascination, the Eiffel Tower never fails to impress. Enriched by a history full of new developments, here you can discover all of its key information.

### Main figures

<table>
<thead>
<tr>
<th>Description</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current height</td>
<td>1063 feet</td>
</tr>
<tr>
<td>Original height without antenna</td>
<td>1024 feet</td>
</tr>
<tr>
<td>Total width</td>
<td>410 feet (on the ground)</td>
</tr>
<tr>
<td>Width of a pillar</td>
<td>82 feet (on the ground)</td>
</tr>
<tr>
<td>First floor</td>
<td>187 feet, 14,485 square feet</td>
</tr>
<tr>
<td>Second floor</td>
<td>377 feet, 4,692 square feet</td>
</tr>
<tr>
<td>Third floor</td>
<td>906 feet, 820 square feet</td>
</tr>
<tr>
<td>Lifts</td>
<td>5 lifts from the esplanade to second floor, 2 x 2 dualifts from second floor to the top</td>
</tr>
<tr>
<td>Weight of the metal frame</td>
<td>7,300 tons</td>
</tr>
</tbody>
</table>

[https://www.toureiffel.paris/en/the-monument/key-figures](https://www.toureiffel.paris/en/the-monument/key-figures)
Suppose you are shopping...

How does Amazon provide the information?

<table>
<thead>
<tr>
<th>Brand</th>
<th>Cetaphil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingredients</td>
<td>Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butyricarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.</td>
</tr>
<tr>
<td>Scent</td>
<td>Fragrance free</td>
</tr>
<tr>
<td>Additional Item Information</td>
<td>Non-Comedogenic, Fragrance-free, Natural</td>
</tr>
<tr>
<td>Skin Type</td>
<td>Sensitive</td>
</tr>
</tbody>
</table>

About this item
- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean
- Mild & non-irritating: Soap free formulation won't strip skin of its natural protective oils and emollients
Suppose you are shopping...

How does Amazon provide choices?
Suppose you are shopping...

How does Amazon improve searching and recommendation?
Suppose you are diagnosing…

How to better diagnose and answer questions with EHR?
Suppose you would like to analyze COVID-19…

- You would like to build an application to visualize and analyze COVID-19

What would you primarily need?

Jobie Budd et al., Digital technologies in the public-health response to COVID-19, nature medicine
Data

Closed vs. Open Data

Tabular vs. Graph

Human Readable vs. Machine Readable

Separated vs. Integrated/Linked
**5-Star Open Data**

- **Tim Berners-Lee**, the inventor of the Web and Linked Data initiator, suggested a **5-star deployment scheme** for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.

1. **Make your stuff available on the Web** (whatever format) under an open license.
2. **Make your stuff available as structured data** (e.g., CSV instead of Excel).
3. **Make it available in a non-proprietary open format** (e.g., CSV instead of Excel).
4. **Use URIs to denote things**, so that people can point at your stuff.
5. **Link your data to other data to provide context**.

[https://5stardata.info/en/](https://5stardata.info/en/)
What is Knowledge

- Plato’s definition: Justified true belief

Merriam-Webster

SINCE 1828

Dictionary  Thesaurus

Essential Meaning of knowledge

1. information, understanding, or skill that you get from experience or education

   // She has little/no/some knowledge of fashion/history/cooking.
   // He has devoted himself to the pursuit of knowledge.

   See More Examples

2. awareness of something: the state of being aware of something

   // He claimed to have no knowledge of the plan. = He denied all/any knowledge of the plan.
   // Did you have any knowledge of her intentions?
   // The decision was made without my knowledge. [=I did not know about the decision]
What is Graph
So a knowledge graph is...

- Knowledge represented as entities, edges and attributes

**Personal entity** showing that Tom watched *Ghosts of the Abyss*

**Key concepts**

<table>
<thead>
<tr>
<th>Entity</th>
<th>Represent something in the real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge</td>
<td>Represent relationship</td>
</tr>
<tr>
<td>Attribute</td>
<td>Represent something about an entity</td>
</tr>
<tr>
<td>Ontology</td>
<td>Definition of possible types of entities, relationships and attributes</td>
</tr>
</tbody>
</table>

**Edge (i.e. relationship)** showing that "Ghosts of the Abyss" was ‘directed_by’ and ‘produced_by’ James Cameron
Understanding the Role of Taxonomies, Ontologies, Schemas and Knowledge Graphs
What is a Taxonomy?

- A data taxonomy is the classification of data into categories and sub-categories.

- Taxonomy identifies hierarchical relationships within a category.
What is an Ontology?

 Ontologies are semantic data models that define the types of things that exist in our domain and the properties that can be used to describe them. There are three main components to an ontology, which are usually described as follows:

- **Classes**: the distinct types of things that exist in our data.
- **Relationships**: properties that connect two classes.
- **Attributes**: properties that describe an individual class.

Ontologies are generalized data models, meaning that they only model general types of things that share certain properties, but don’t include information about specific individuals in our domain.

The main difference between Taxonomy and Ontology is that taxonomy is simpler in nature than ontology. Taxonomy takes into consideration one type of relationship, whereas ontology takes into account many different complex relationships between the concepts.
What is a Schema?

A schema is the organization or structure for a database.

Ontologies focus more on the semantic relationships whereas schemas focus more on the data structure (e.g., the data class Elephant that has the data properties: name and livesIn).
What is a Knowledge Graph?

Using our ontology as a framework, we can add in real data about individual books, authors, publishers, and locations to create a knowledge graph.

ontology + data = knowledge graph
Why Knowledge Graphs?

- **Graphs are a natural way** to represent entities and their relationships.
- Graphs can **capture a broad spectrum of data** (structured / unstructured).
- Graphs can be managed **efficiently**.

**Game-changing data integration**

**Robust data quality assurance**

**Intuitive domain modelling**

**Flexibility & performance**

**Low up-front investment**
Minimum set of characteristics of knowledge graphs

- Mainly describes real world entities and their interrelations, organized in a graph.
- Defines possible classes and relations of entities in a schema.
- Allows for potentially interrelating arbitrary entities with each other.
- Covers various topical domains.
Open Knowledge Graphs

- Open knowledge graphs are published online, making their content accessible for the public good.
  - DBpedia
  - Freebase
  - Wikidata
  - YAGO
  - ……

- Cover many domains and are either extracted from Wikipedia, or built by communities of volunteers.
Cyc is a long-term artificial intelligence project that aims to assemble a comprehensive ontology and knowledge base that spans the basic concepts and rules about how the world works. Hoping to capture common sense knowledge, Cyc focuses on implicit knowledge that other AI platforms may take for granted.

Freebase was a large collaborative knowledge base consisting of data composed mainly by its community members. It was an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions. [3][2] Freebase aimed to create a global resource that allowed people (and machines) to access common information more effectively.

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.
State of the art knowledge graphs

DBpedia is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web.

YAGO is an open source knowledge base developed at the Max Planck Institute for Computer Science in Saarbrücken. It is automatically extracted from Wikipedia and other sources.

Never-Ending Language Learning system (NELL) is a semantic machine learning system developed by a research team at Carnegie Mellon University, and supported by grants from DARPA, Google, NSF, and CNPq with portions of the system running on a supercomputing cluster provided by Yahoo!
## Recently-Learned Facts

<table>
<thead>
<tr>
<th>instance</th>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>high_end_clock is an item found on a wall</td>
<td>1111</td>
<td>06-jul-2018</td>
<td>99.6</td>
</tr>
<tr>
<td>movie_edward_scissorhands is a movie</td>
<td>1111</td>
<td>06-jul-2018</td>
<td>100.0</td>
</tr>
<tr>
<td>n1999_toyota_camry is a model of automobile</td>
<td>1111</td>
<td>06-jul-2018</td>
<td>94.7</td>
</tr>
<tr>
<td>michael_i is a monarch</td>
<td>1111</td>
<td>06-jul-2018</td>
<td>96.7</td>
</tr>
<tr>
<td>gonzales_music_festival_1968 is a music festival</td>
<td>1111</td>
<td>06-jul-2018</td>
<td>100.0</td>
</tr>
<tr>
<td>roof is a building feature made from the building material steel</td>
<td>1115</td>
<td>03-sep-2018</td>
<td>92.7</td>
</tr>
<tr>
<td>tyco_healthcare hired dennis_kozlowski</td>
<td>1112</td>
<td>24-jul-2018</td>
<td>100.0</td>
</tr>
<tr>
<td>air_pacific is a company that has an office in the city los_angeles_ca</td>
<td>1112</td>
<td>24-jul-2018</td>
<td>99.2</td>
</tr>
<tr>
<td>the sports team yankees was the winner of n1962_world_series</td>
<td>1116</td>
<td>12-sep-2018</td>
<td>100.0</td>
</tr>
<tr>
<td>silvestre_herrera died in the city glendale</td>
<td>1116</td>
<td>12-sep-2018</td>
<td>100.0</td>
</tr>
</tbody>
</table>

[http://rtw.ml.cmu.edu/rtw/]
**Input:** Wikipedia infoboxes, WordNet and GeoNames

**Output:** KG with 350K entity types, 10M entities, 120M facts

**Temporal and spatial information**

[Link to YAGO](https://yago-knowledge.org/)
ConceptNet

ConceptNet
An open, multilingual knowledge graph

beetles
An English term in ConceptNet 5.5

Derived terms
- beatle
- beatledom
- beatlemania
- beatlesque
- fourth beatle

beetles is a type of...
- a British band
- man (n)
- band (n)
- musician (n)
- album (n)

Links to other sites
- dbpedia.org The Beatles
- sw.opencyc.org Beatles
- umbel.org Beatles
- wordnet-rdf.princeton.edu 400520405-N
- wordnet-rdf.princeton.edu 108386847-n
- wikidata.dbpedia.org Q1299
- en.wiktionary.org Beatles
- dbpedia.org The Beatles (No. 1)
- wikidata.dbpedia.org Q738260
- fr.wiktionary.org Beatles
- dbpedia.org The Beatles (The Original Studio Recordings)
- wikidata.dbpedia.org Q603122

https://conceptnet.io/
Enterprise Knowledge Graphs

- Enterprise knowledge graphs are typically **internal to a company** and applied for **commercial use-cases**
  - **Web search:** Bing, Google, Airbnb, Amazon, eBay, Uber, …
  - **Social network:** Facebook, LinkedIn, …
  - **Finance:** Accenture, Bloomberg, Capital One, Wells Fargo, …

- Applications include search, recommendations, personal agents, advertising, business analytics, risk assessment, automation, and more besides.
Recap: suppose you are searching…

Query: “Honda Civic, Hyundai Elantra”

What he is interested in?
Recap: Is-A relationship between things

- Honda Civic
- Hyundai Elantra
- Fuel-efficient cars
- Economy cars
- Marvel heroes
- Revengers

**Concept:** a collection of things that share some common attributes
Create a web-scale ontology to represent user interests and document topics.

Liu et al., SIGMOD 2020, GIANT: Scalable Creation of a Web-scale Ontology
Microsoft Concept Graph

Our goal is to enable machines to better understand human communication. An important question is, what does the word “understand” mean here? Consider the following example. For human beings, when we see “25 Oct 1881”, we recognize it as a date, although most of us do not know what it is about. However, if we are given a little more context, say the date is embedded in the following piece of short text “Pablo Picasso, 25 Oct 1881, Spain”, most of us would have guessed (correctly) that the date represents Pablo Picasso’s birthday. We are able to do this because we possess certain knowledge, and in this case, “one of the most important dates associated with a person is his birthday.”

As another example, consider a problem in natural language processing. Humans do not find sentences such as “animals other than dogs such as cats” ambiguous, but machine parsing can lead to two possible understandings: “cats are animals” or “cats are dogs.” Common sense tells us that cats cannot be dogs, which renders the second parsing improbable.
Recap: suppose you are asking…

**Question:** “How tall is the Eiffel Tower?”

How does Google know?
Google Knowledge Graph

https://blog.google/products/search/introducing-knowledge-graph-things-not/
Recap: suppose you are shopping...

How does Amazon provide the information?

### Cetaphil Gentle Skin Cleansing Cloths

- **Brand**: Cetaphil
- **Ingredients**: Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butyrcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.
- **Scent**: Fragrance free
- **Additional Information**: Non-Comedogenic, Fragrance-free, Natural
- **Skin Type**: Sensitive

**About this item**
- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use.
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean.
- Mild & non-irritating: Soap free formulation won't strip skin of its natural protective oils and emollients.
Amazon Product Knowledge Graph

Product Graph Example for 2 Products

https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/
Product Graph vs. Knowledge Graph

Generic KG

- Movie,
- Music,
- Book,
- Aetc.

Product KG

- (Hardline, softline, consumables, etc.)

https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/
Semantic Web
Knowledge Graph
W3C semantic web or linked data standards are widely used in the development of knowledge graphs. JSON-LD, in fact, is one of those standards.

“Semantic web” is a term that originated with web pioneer and World Wide Web Consortium founder Tim Berners-Lee (TBL) in an article in Scientific American in 1999, one that described a vision of a relationship-rich, scalable, contextual data web, a web of graph-oriented and well-described data that could live side by side with existing web content.

Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data.

Linked data are empowered by technologies such as RDF, SPARQL, JSON-LD, OWL, SHACL and SKOS.
Semantic Web vs. KG

- Semantic Web is based on RDF (Resource Description Framework) and OWL (Web Ontology Language).

- Google Knowledge Graph is based on its API, which in turn is based on Scheme in JSON-LD format and which in turn is also based on RDF or RDF Schema. JSON-LD is a lightweight Linked Data format. It is easy for humans to read and write. It is based on the already successful JSON format and provides a way to help JSON data interoperate at Web-scale.

- There are different technologies to build and operate a knowledge graph. A knowledge graph that employs Semantic Web standards and technologies is called a Semantic Web Knowledge Graph.
KG Standards

- W3C® RDF: the data part
- W3C® OWL: the schema part
- W3C® SPARQL: the query part
Resource Description Framework (RDF)

- Data model, based on S-P-O triple structure (Subject, Predicate, Object)

- Used for describing things, yes, every, single, thing And anyways, **RDF = Resource Description Framework**

- Four Key Principles for RDF Graphs
  - Uniform Resource Identifiers (URIs) as identifiers
  - HTTP URI as information access protocol
  - RDF and SPARQL as knowledge representation and querying languages
  - Link information from one source to other
RDF Schema (RDFS)

- RDFS is a vocabulary for modeling schemas used in RDF data.
- RDFS provides mechanisms for defining classes and properties.
- Via RDFS, e.g., one can provide the domains and ranges of properties.
- The official document of RDFS is available: https://www.w3.org/TR/rdf-schema/
Simple Knowledge Organization System (SKOS)

- SKOS is short for **Simple Knowledge Organization System**.
- It is a vocabulary for expressing the basic structure and content of concept schemes such as thesauri and taxonomies.
- SKOS allows concepts and relationships among concepts to be composed and published on the World Wide Web.

**SKOS & RDF**

SKOS provides a standard way to represent knowledge organization systems using the **Resource Description Framework (RDF)**. Encoding this information in RDF allows it to be passed between computer applications in an interoperable way.

Using RDF also allows knowledge organization systems to be used in distributed, decentralised metadata applications. Decentralised metadata is becoming a typical scenario, where service providers want to add value to metadata harvested from multiple sources.
If RDF captures knowledge, then SPARQL retrieves knowledge, querying knowledge captured by RDF!

Short for: SPARQL Protocol and RDF Query Language
SPARQL documentation: https://www.w3.org/TR/rdf-sparql-query/

Example:

```sparql
@prefix ex: <https://example.org/resource/> .
@prefix prop: <https://example.org/property/> .

ex:matt prop:knows ex:scott .

SELECT ?twitter WHERE {
  ?person prop:knows ex:scott
}
```

*What are the Twitter accounts of people knowing Scott?*
Web Ontology Language (WOL? No, it is OWL!)

- Short for: Web Ontology Language. OWL is a language to create ontologies.

- OWL is more expressive/heavyweight compared to RDFS.

- Key features:
  - Vocabulary description
  - Reasoning: You can conclude new things based on existing facts!

  For example: owl subClassOf bird + bird subClassOf animal
  Now, if Bobi is an owl, do you think Bobi is an animal?

- Documentation:
  https://www.w3.org/TR/owl-primer/
More on Semantic Web

- [https://www.w3.org/standards/semanticweb/](https://www.w3.org/standards/semanticweb/)

Semantic Web

In addition to the classic “Web of documents” W3C is helping to build a technology stack to support a “Web of data,” the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term “Semantic Web” refers to W3C’s vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

**Linked Data**

The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files.

**Vocabularies**

At times it may be important or valuable to organize data. Using OWL (to build vocabularies, or “ontologies”) and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

**Query**

Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data. SPARQL is the query language for the Semantic Web.

**Inference**

Near the top of the Semantic Web stack one

**Vertical Applications**

W3C is working with different industries —
KG Construction
KG Workflow: How to build a KG from Scratch

**KG Sources**
- Unstructured Sources
- Structured Sources
- Domain Experts

**KG Construction**
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

**Knowledge Graph**

**KG Maintenance**
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

**KG Applications**
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
KG Workflow: How to build a KG from Scratch

KG Sources
- Unstructured Sources
- Structured Sources
- Domain Experts

KG Construction
- Information Extraction
- Preprocessing
- Named Entity Recognition
- Entity Linking
- Relation Extraction
- Coreference Resolution
- Quality Checking
- KG Publication & Hosting

Knowledge Graph

KG Maintenance
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

KG Applications
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
We know that a knowledge graph is when an ontology is applied to a set of individual data. All of which are based on RDF triples.

The construction of KG will involve how do we get triples from data of different sources.
The right figure shows the sub-KG that covers the 3-hop neighbors of SARS-COV-2 with disease type nodes. We sampled 10 paths from the complete KG-Covid-19.

Data source: https://github.com/Knowledge-Graph-Hub/kg-covid-19
Knowledge Sources

- Structured Sources
  - Relational Databases
  - Feeds
  - XML
  - CSV
  - ...

- Unstructured Sources
  - Web Craws
  - Scientific Articles
  - Social Media
  - Emails
  - ...
Structured Knowledge Sources to KGS

Structured information can be mapped to KGs through a semantic integration process.

Common strategy:
- First adopting reference ontologies as global schemas.
- Then construct mappings between the global schema and the local schema of the target data source.
How to excavate structured knowledge from unstructured data automatically?
How to excavate structured knowledge from unstructured data automatically?
How to excavate structured knowledge from unstructured data automatically?
Information Extraction

KG Sources
- Unstructured Sources
- Structured Sources
- Domain Experts

KG Construction
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Relation Extraction
  - Entity Linking
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

KG Maintenance
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

Knowledge Graph

KG Applications
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources. (From Wikipedia)
How does IE work?

Raw Text → Preprocessing → Named Entity Recognition (NER) → Entity Linking → Relation Extraction → Coreference Resolution → Structured Knowledge

Robert was stuck at the airport because of the snow storm. He missed the wedding of his daughter.

Structured Knowledge

(Robert, PHYS, the airport), (Robert, PER-SOC, daughter)

https://huggingface.co/hmtl/#emd
The pre-processing usually involves varied procedures (depends on followed algorithms):
- Language Detection
- Sentence Segmentation
- Tokenization
- Part-of-Speech (POS) Tagging
- Dependency Parsing
- …
Tokenization

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters.

Tools:
- NLTK Word Tokenize
- Stanford CoreNLP
- Gensim word tokenizer
- spaCy Tokenizer
- TextBlob Word Tokenize
- Tokenization with Keras
- BERT Tokenizer
- ...
POS Tagging

POS Tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

Tools:
- spaCy Tagger
- NLTK pos_tag
- ...

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
<th>English Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>new, good, high, special, big, local</td>
</tr>
<tr>
<td>ADP</td>
<td>adposition</td>
<td>on, of, at, with, by, into, under</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>really, already, still, early, now</td>
</tr>
<tr>
<td>CONJ</td>
<td>conjunction</td>
<td>and, or, but, if, while, although</td>
</tr>
<tr>
<td>DET</td>
<td>determiner, article</td>
<td>the, a, some, most, every, no, which</td>
</tr>
<tr>
<td>NOUN</td>
<td>noun</td>
<td>year, home, costs, time, Africa</td>
</tr>
<tr>
<td>NUM</td>
<td>numeral</td>
<td>twenty-four, fourth, 1991, 14:24</td>
</tr>
<tr>
<td>PRT</td>
<td>particle</td>
<td>at, on, out, over per, that, up, with</td>
</tr>
<tr>
<td>PRON</td>
<td>pronoun</td>
<td>he, their, her, its, my, I, us</td>
</tr>
<tr>
<td>VERB</td>
<td>verb</td>
<td>is, say, told, given, playing, would</td>
</tr>
<tr>
<td>.</td>
<td>punctuation marks</td>
<td>. , ; ; !</td>
</tr>
<tr>
<td>X</td>
<td>other</td>
<td>ersatz, esprit, dunno, gr8, university</td>
</tr>
</tbody>
</table>
Dependency Parsing

- Dependency Parsing is the process to analyze the **grammatical structure** in a sentence and find out related words as well as the type of the relationship between them.

- **Tools:**
  - spaCy Parser
  - NLTK StanfordDependencyParser
  - Stanza
  - …

![Diagram of Dependency Parsing and Constituency Parsing]
**NLTK** is a leading platform for building Python programs to work with human language data. ([https://www.nltk.org/](https://www.nltk.org/))

<table>
<thead>
<tr>
<th>NLTK Documentation</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>API Reference</td>
<td>Natural Language Toolkit</td>
</tr>
<tr>
<td>Example Usage</td>
<td>NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.</td>
</tr>
<tr>
<td>Module Index</td>
<td>Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.</td>
</tr>
<tr>
<td>Wiki</td>
<td>NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”</td>
</tr>
<tr>
<td>FAQ</td>
<td>Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at <a href="https://www.nltk.org/book_1ed.">https://www.nltk.org/book_1ed.</a>)</td>
</tr>
<tr>
<td>Open Issues</td>
<td>Some simple things you can do with NLTK</td>
</tr>
<tr>
<td>NLTK on GitHub</td>
<td></td>
</tr>
<tr>
<td>Installation</td>
<td></td>
</tr>
<tr>
<td>Installing NLTK</td>
<td></td>
</tr>
<tr>
<td>Installing NLTK Data</td>
<td></td>
</tr>
<tr>
<td>More</td>
<td></td>
</tr>
<tr>
<td>Release Notes</td>
<td></td>
</tr>
<tr>
<td>Contributing to NLTK</td>
<td></td>
</tr>
<tr>
<td>NLTK Team</td>
<td></td>
</tr>
</tbody>
</table>
CoreNLP is your one stop shop for natural language processing in Java and currently supports 8 languages: Arabic, Chinese, English, French, German, Hungarian, Italian, and Spanish. (https://stanfordnlp.github.io/CoreNLP/)

What's new: The v4.3 releases (late 2021) added support for Italian and Hungarian.
Gensim is a Python library for *topic modelling*, *document indexing* and *similarity retrieval* with large corpora. (https://radimrehurek.com/gensim/)
spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. (https://spacy.io/)
Named Entity Recognition (NER)

- NER seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.
Named Entity Recognition (NER)

NER Approaches:

- **Rule-based Approaches** rely on hand-crafted rules. The Rules can be designed based on domain-specific gazetteers and syntactic-lexical patterns.

- **Unsupervised Learning Approaches** typically leverage clustering based on lexical resources, patterns, and statistics on a large corpus to infer mentions of named entities.

- **Feature-based Supervised Learning Approaches** cast NER to a multi-class classification or sequence labeling task, with annotated data samples and carefully designed features.

- **Deep Learning-based Approaches** discovering hidden features automatically.
Named Entity Recognition (NER)

- **Rule-based Approaches**

  - **Pros:**
    The domain-specific rules usually brings high precision.

  - **Cons:**
    Low recall resulted by incomplete dictionaries.
    Construction of rules can be expensive.
    Cannot be transferred to other domains.

  - **Rule-based systems:**
    ProMiner, LaSIE-II, NetOwl, Facile, SAR, LTG, …
Named Entity Recognition (NER)

- Unsupervised Learning Approaches:
  - **Pros:**
    - Low required resources.
  - **Cons:**
    - Low performance.
  - Unsupervised NER systems:
    - KNOWITALL, Collins et al., Nadeau et al., Zhang and Elhadad, …
**Named Entity Recognition (NER)**

**Feature-based Supervised Learning Approaches:**

- **Pros:**
  High in-domain performance.

- **Cons:**
  The designing of features may involve experts.
  The labeling of data may be expensive.

- **Systems:**
  IdentiFinder, MENE, McNamee and Mayfield, …
**Deep Learning-based Approaches**

- **Pros:**
  - End-to-end
  - High performance
  - Automatic feature discovery

- **Cons:**
  - Generally requires abundant labeled data
  - Requires higher computational resources

- **Systems:**
  - CNN-based, LSTM-based, ELMo-based, BERT-based, …
### NER Datasets

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Year</th>
<th>Text Source</th>
<th>#Tags</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-6 Plus</td>
<td>1995</td>
<td>Additional news to MUC-6</td>
<td>7</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC96T10">https://catalog.ldc.upenn.edu/LDC96T10</a></td>
</tr>
<tr>
<td>ACE</td>
<td>2000 - 2008</td>
<td>Transcripts, news</td>
<td>7</td>
<td><a href="https://www.ldc.upenn.edu/collaborations/past-projects/ace">https://www.ldc.upenn.edu/collaborations/past-projects/ace</a></td>
</tr>
<tr>
<td>OntoNotes</td>
<td>2007 - 2012</td>
<td>Magazine, news, web, etc.</td>
<td>18</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2013T19">https://catalog.ldc.upenn.edu/LDC2013T19</a></td>
</tr>
<tr>
<td>BBN</td>
<td>2005</td>
<td>Wall Street Journal</td>
<td>64</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2005T33">https://catalog.ldc.upenn.edu/LDC2005T33</a></td>
</tr>
<tr>
<td>N³³</td>
<td>2014</td>
<td>News</td>
<td>3</td>
<td><a href="http://aksw.org/Projects/N3NERNEDNIF.html">http://aksw.org/Projects/N3NERNEDNIF.html</a></td>
</tr>
<tr>
<td>Gillick</td>
<td>2016</td>
<td>Magazine, news, web, etc.</td>
<td>89</td>
<td><a href="https://arxiv.org/abs/1412.1820v2">https://arxiv.org/abs/1412.1820v2</a></td>
</tr>
<tr>
<td>FG-NER</td>
<td>2018</td>
<td>Various</td>
<td>200</td>
<td><a href="https://figner.alt.iai/">https://figner.alt.iai/</a></td>
</tr>
<tr>
<td>NNE</td>
<td>2019</td>
<td>Newswire</td>
<td>114</td>
<td><a href="https://github.com/nickyringland/nested_named_entities">https://github.com/nickyringland/nested_named_entities</a></td>
</tr>
<tr>
<td>GENIA</td>
<td>2004</td>
<td>Biology and clinical text</td>
<td>36</td>
<td><a href="http://www.geniaproject.org/home">http://www.geniaproject.org/home</a></td>
</tr>
<tr>
<td>FSU-PRGE</td>
<td>2010</td>
<td>PubMed and MEDLINE</td>
<td>5</td>
<td><a href="https://julielab.de/Resources/FSU_PRGE.html">https://julielab.de/Resources/FSU_PRGE.html</a></td>
</tr>
<tr>
<td>DFKI</td>
<td>2018</td>
<td>Business news and social media</td>
<td>7</td>
<td><a href="https://dfki-lt-re-group.bitbucket.io/product-corpus/">https://dfki-lt-re-group.bitbucket.io/product-corpus/</a></td>
</tr>
</tbody>
</table>
Recent works on neural NER

<table>
<thead>
<tr>
<th>Work</th>
<th>Character</th>
<th>Input representation</th>
<th>Model</th>
<th>Context encoder</th>
<th>Tag decoder</th>
<th>Performance (F-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[94]</td>
<td>-</td>
<td>Trained on PubMed</td>
<td>POS</td>
<td>CRF</td>
<td>GENIA</td>
<td>71.61%</td>
</tr>
<tr>
<td>[95]</td>
<td>-</td>
<td>Trained on Cogword</td>
<td>GRU</td>
<td>CRF</td>
<td>ACT2008:80.60%</td>
<td></td>
</tr>
<tr>
<td>[96]</td>
<td>-</td>
<td>Random</td>
<td>LSTM</td>
<td>CRF</td>
<td>Polter Network:86.86%</td>
<td></td>
</tr>
<tr>
<td>[97]</td>
<td>-</td>
<td>Trained on NYT</td>
<td>LSTM</td>
<td>CRF</td>
<td>NYT:49.50%</td>
<td></td>
</tr>
<tr>
<td>[98]</td>
<td>-</td>
<td>SIF;NN - Wave shape</td>
<td>LSTM</td>
<td>CRF</td>
<td>CoNLL03:90.86%</td>
<td></td>
</tr>
<tr>
<td>[99]</td>
<td>LSTM</td>
<td>Google went2Vec</td>
<td>CRF</td>
<td>CRF</td>
<td>OriNote04:90.64%</td>
<td></td>
</tr>
<tr>
<td>[100]</td>
<td>CNN</td>
<td>GloVe</td>
<td>CRF</td>
<td>CRF</td>
<td>CoNLL04:75.00%</td>
<td></td>
</tr>
<tr>
<td>[101]</td>
<td>LSTM</td>
<td>Google went2Vec</td>
<td>CRF</td>
<td>CRF</td>
<td>CoNLL05:80.32%</td>
<td></td>
</tr>
<tr>
<td>[102]</td>
<td>CNN</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>CoNLL05:93.29%</td>
<td></td>
</tr>
<tr>
<td>[103]</td>
<td>LSTM</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>CoNLL05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[104]</td>
<td>LSTM</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>CoNLL05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[105]</td>
<td>CNN</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>SenNA:87.31%</td>
<td></td>
</tr>
<tr>
<td>[106]</td>
<td>LSTM-LM</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[107]</td>
<td>CNN-LSTM-LM</td>
<td>SenNA</td>
<td>CRF</td>
<td>CRF</td>
<td>OriNote05:93.64%</td>
<td></td>
</tr>
<tr>
<td>[108]</td>
<td>-</td>
<td>Random</td>
<td>CNN</td>
<td>CRF</td>
<td>OriNote06:89.86%</td>
<td></td>
</tr>
<tr>
<td>[109]</td>
<td>-</td>
<td>SenNA</td>
<td>CNN</td>
<td>CRF</td>
<td>OriNote06:90.83%</td>
<td></td>
</tr>
<tr>
<td>[110]</td>
<td>-</td>
<td>SenNA</td>
<td>CNN</td>
<td>CRF</td>
<td>OriNote06:91.62%</td>
<td></td>
</tr>
<tr>
<td>[111]</td>
<td>-</td>
<td>SenNA</td>
<td>CNN</td>
<td>CRF</td>
<td>OriNote06:96.54%</td>
<td></td>
</tr>
<tr>
<td>[112]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>MLP</td>
<td>CRF</td>
<td>OriNote05:91.17%</td>
<td></td>
</tr>
<tr>
<td>[113]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.07%</td>
<td></td>
</tr>
<tr>
<td>[114]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>W-NLU17:46.62%</td>
<td></td>
</tr>
<tr>
<td>[115]</td>
<td>CNN</td>
<td>SenNA</td>
<td>LSTM</td>
<td>CRF</td>
<td>W-NLU17:43.66%</td>
<td></td>
</tr>
<tr>
<td>[116]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>W-NLU17:41.81%</td>
<td></td>
</tr>
<tr>
<td>[117]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>SnapCaptions:32.4%</td>
<td></td>
</tr>
<tr>
<td>[118]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.75%</td>
<td></td>
</tr>
<tr>
<td>[119]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:87.95%</td>
<td></td>
</tr>
<tr>
<td>[120]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[121]</td>
<td>LSTM</td>
<td>SenNA</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.48%</td>
<td></td>
</tr>
<tr>
<td>[122]</td>
<td>LSTM</td>
<td>Google went2Vec</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.34%</td>
<td></td>
</tr>
<tr>
<td>[123]</td>
<td>LSTM</td>
<td>Google went2Vec</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.67%</td>
<td></td>
</tr>
<tr>
<td>[124]</td>
<td>LSTM-LM</td>
<td>SenNA</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:91.16%</td>
<td></td>
</tr>
<tr>
<td>[125]</td>
<td>CNN-LSTM-LM</td>
<td>SenNA</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[126]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[127]</td>
<td>LSTM-LM</td>
<td>GloVe, POS</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[128]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[129]</td>
<td>LSTM-LM</td>
<td>GloVe, POS</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[130]</td>
<td>LSTM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[131]</td>
<td>LSTM-LM</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[132]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[133]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[134]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[135]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[136]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[137]</td>
<td>CNN</td>
<td>GloVe, POS</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[138]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[139]</td>
<td>CNN</td>
<td>GloVe, POS</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[140]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[141]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[142]</td>
<td>CNN</td>
<td>GloVe, POS</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[143]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[144]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[145]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[146]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[147]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[148]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[149]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
<tr>
<td>[150]</td>
<td>CNN</td>
<td>GloVe</td>
<td>LSTM</td>
<td>CRF</td>
<td>OriNote05:90.96%</td>
<td></td>
</tr>
</tbody>
</table>

## Off-the-shelf NER tools

<table>
<thead>
<tr>
<th>NER System</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>StanfordCoreNLP</td>
<td><a href="https://stanfordnlp.github.io/CoreNLP/">https://stanfordnlp.github.io/CoreNLP/</a></td>
</tr>
<tr>
<td>OSU Twitter NLP</td>
<td><a href="https://github.com/aritter/twitter_nlp">https://github.com/aritter/twitter_nlp</a></td>
</tr>
<tr>
<td>Illinois NLP</td>
<td><a href="http://cogcomp.org/page/software/">http://cogcomp.org/page/software/</a></td>
</tr>
<tr>
<td>NeuroNER</td>
<td><a href="http://neuroner.com/">http://neuroner.com/</a></td>
</tr>
<tr>
<td>NERSuite</td>
<td><a href="http://nersuite.nlplab.org/">http://nersuite.nlplab.org/</a></td>
</tr>
<tr>
<td>Polyglot</td>
<td><a href="https://polyglot.readthedocs.io">https://polyglot.readthedocs.io</a></td>
</tr>
<tr>
<td>Glimli</td>
<td><a href="http://bioinformatics.ua.pt/glimli">http://bioinformatics.ua.pt/glimli</a></td>
</tr>
<tr>
<td>spaCy</td>
<td><a href="https://spacy.io/api/entityrecognizer">https://spacy.io/api/entityrecognizer</a></td>
</tr>
<tr>
<td>NLTK</td>
<td><a href="https://www.nltk.org">https://www.nltk.org</a></td>
</tr>
<tr>
<td>OpenNLP</td>
<td><a href="https://opennlp.apache.org/">https://opennlp.apache.org/</a></td>
</tr>
<tr>
<td>LingPipe</td>
<td><a href="http://alias-i.com/lingpipe-3.9.3/">http://alias-i.com/lingpipe-3.9.3/</a></td>
</tr>
<tr>
<td>AllenNLP</td>
<td><a href="https://demo.alenlp.org/">https://demo.alenlp.org/</a></td>
</tr>
<tr>
<td>IBM Watson</td>
<td><a href="https://natural-language-understanding-demo.ng.bluemix.net">https://natural-language-understanding-demo.ng.bluemix.net</a></td>
</tr>
<tr>
<td>FG-NER</td>
<td><a href="https://fgner.alt.ai/extractor/">https://fgner.alt.ai/extractor/</a></td>
</tr>
<tr>
<td>Intellexer</td>
<td><a href="http://demo.intellexer.com/">http://demo.intellexer.com/</a></td>
</tr>
<tr>
<td>Repustate</td>
<td><a href="https://reputate.com/named-entity-recognition-api-demo">https://reputate.com/named-entity-recognition-api-demo</a></td>
</tr>
<tr>
<td>AYLIEN</td>
<td><a href="https://developer.aylien.com/text-api-demo">https://developer.aylien.com/text-api-demo</a></td>
</tr>
<tr>
<td>Dandelion API</td>
<td><a href="https://dandelion.eu/semantic-text/entity-extraction-demo">https://dandelion.eu/semantic-text/entity-extraction-demo</a></td>
</tr>
<tr>
<td>displaCy</td>
<td><a href="https://explosion.ai/demos/displacy-ent">https://explosion.ai/demos/displacy-ent</a></td>
</tr>
<tr>
<td>ParallelDots</td>
<td><a href="https://www.paralleldots.com/named-entity-recognition">https://www.paralleldots.com/named-entity-recognition</a></td>
</tr>
<tr>
<td>TextRazor</td>
<td><a href="https://www.textrazor.com/named_entity_recognition">https://www.textrazor.com/named_entity_recognition</a></td>
</tr>
</tbody>
</table>

Entity Linking (EL)

Entity Linking (or Named-Entity Disambiguation) assigns a unique identity in target knowledge bases (KBs) to entity mentions in text.

We know ‘Sebastian Thrun’ is a person but do we know which person exactly?

When Sebastian Thrun started at Google in 2007, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode earlier this week.

A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

About: Sebastian Thrun

Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur, educator, and computer scientist from Germany. He was CEO and co-founder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech.

http://dbpedia.org/page/Sebastian_Thrun
Entity Linking (EL)

- A formal definition of EL consist of:
  - A definition of the mentions (concepts, entities)
  - Determining the target KB
  - Defining the unique identity in the target KBs

Yann LeCun was a proponent of deep learning and neural networks with his work on CNN.

1. Define the mentions
2. Determine target KBs
3. Find the unique identity

Named Entities
- Yann_LeCun

Generic Concepts
- Deep_learning
- Artificial_neural_network
- Convolutional_neural_network
Entity Linking: Define the mentions

- **A mention**: a phrase used to refer to something in the world
  - Named Entity
  - Object
  - Substance
  - Event
  - Philosophy
  - Mental State
  - Rule
  - ...

- Task definitions of **EL** vary across the definition of **mentions**
  - All N-grams
  - Dictionary-based selection
  - Data-driven controlled vocabulary (e.g., all Wikipedia titles)
  - **Named entities** (by NER in our demo).

- Ideally, the definition of mentions should adapt to the application
Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Some task definitions insist on dealing only with mentions that are named entities.

How about: **Hosni Mubarak’s wife**?
Both entities have a Wikipedia page.

Blumenthal (D) is a candidate for the **U.S. Senate** seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Chris Dodd
The New York Times
Connecticut
The definition of which mentions to highlight should depend on the expertise and users' interests.

https://cogcomp.seas.upenn.edu/page/demo_view/EDLEnglish
In principle, multiple KBs can be used as the target KB.
Find the unique identity in the target KB (Disambiguation)

Which one to link?

This is a bass

This is also a bass

James Craig (police chief) - Wikipedia
James E. Craig (born 1956) is the former head of the Detroit Police Department. He served in this capacity from 2013 to 2021, after previously serving as...
Political party: Republican (2011–present) Mayor: Dave Bing; Mike Duggan
Other political affiliations: Democratic (before 2011) Education: West Coast University (BS); University of... Policing career: 2022 Michigan gubernatorial...
Find the unique identity in the target KB (Disambiguation)

What if multiple mentions correspond to the same concept is outside KB?

Marc-Alexandre Côté is an author of Textworld.

First cluster relevant mentions as representing a single concept
Then map the cluster to NIL.
Entity Linking (EL)

What are the main issues with EL?

- **Ambiguity**
  - James Craig (police chief)
  - James Craig (1st Viscount Craigavon)
  - James Craig (actor)
  - 
  
- **Variability**
  - CT
  - The Nutmeg State
  - Connecticut
  - 
  
- **Concepts outside of KB**
  - (NIL)
  - Marc-Alexandre Côté is an author of Textworld.

- **Scale**
  - Millions of labels
Entity Linking (EL)

Other challenges

- Popularity Bias

Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Michael Jordan (born 1963) is an American businessman and former professional basketball player.

Michael Jordan or Mike Jordan may also refer to:

People [edit]

Sports [edit]

- Michael Jordan (footballer) (born 1998), English goalkeeper
- Mike Jordan (racing driver) (born 1958), English racing driver
- Mike Jordan (baseball, born 1953) (1869–1940), American baseball player
- Mike Jordan (cornerback) (born 1992), American football cornerback
- Michael Jordan (offensive lineman) (born 1998), American football offensive lineman
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordan (born 1990), Czech ice hockey player

Other people [edit]

- Michael E. Jordan (born 1987), American actor
- Michael J. Jordan (born 1957), American researcher in machine learning and artificial intelligence
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael H. Jordan (1930–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael Jordan [mycologist], English mycologist

But if you search for Michael Jordan
Entity Linking requires addressing several subtasks:

- Mention Identification
- Candidate KB Entities Generation
- Candidate Entity Ranking
- NIL Detection and Clustering

**General Architecture**

- Mention Identification
  - Each n-gram
  - Surface form based filtering
  - Classification and statistics based filtering

- Candidate Generation
  - Surface form matching
  - Expansion using aliases
  - Prior probability computation

- Candidate Ranking
  - Local Inference
  - Global Inference

- NIL Detection & Clustering
  - NIL Detection
  - NIL Clustering
Mentions Identification

- Highest recall: Each n-gram is a potential concept mention
  - Intractable for large documents

- Surface form based filtering
  - Shallow parsing (especially NP chunks)
  - NP’s augmented with surrounding tokens
  - Capitalized words
  - …

- Classification and statistics based filtering
  - Named entity recognition (our demo)
  - Key phrase extraction
  - …
Methods used by some famous systems

- CCG @ U. Penn. Entity Detection and Linking (2019)
  - spaCy NER model
- Illinois Wikifier (Cheng and Roth, 2013)
  - NP chunks and substrings, NER (+nesting), prior anchor text
- DBPedia Spotlight (Mendes et al., 2011)
  - Dictionary-based chunking with string matching (via DBpedia lexicalization dataset)
- RPI Wikifier (Huang et al., 2014)
  - Mention Extraction (Li and Ji, 2014)
Candidate Generation

- Surface form matching
  - A candidate list is composed of entities that match various surface forms of mentions in the text. (Our demo)

- Expansion using aliases
  - A dictionary of additional aliases is constructed using KG metadata or a dictionary of aliases and synonyms
  - e.g. NYC -> New York City

- Prior probability computation
  - Based on pre-calculated prior probabilities of correspondence between certain mentions and entities, p(e|m).
  - e.g. Compute p(e|m) based on Wikipedia entity hyperlinks (e) and anchor text (m).
  - Based on CrossWikis, leveraging the frequency of mention-entity links in web crawl data.
## Candidate Generation

- **Example**

<table>
<thead>
<tr>
<th>Method</th>
<th>5 sample candidate entities for the example mention “Big Blue”</th>
</tr>
</thead>
<tbody>
<tr>
<td>surface form matched based on DBpedia names</td>
<td>Big_Blue_Sky, Big_Blue_(Old_Dominion), Big_Blue_CraneCollapse, Dexter_Bexley_and_the_Big_Blue_Beastie, Big_Bluegrass</td>
</tr>
<tr>
<td>expansion using aliases on YAGO means</td>
<td>Big_Blue_River_(Indiana), Big_Blue_River_(Kansas), Big_Blue_(crane), Big_Red_(drink), IBM</td>
</tr>
<tr>
<td>probability + expansion using aliases on [44]:Anchor prob. + CrossWikis + YAGO</td>
<td>IBM, Big_Blue_River_(Kansas), The_Big_Blue, Big_Blue_River_(Indiana), Big_Blue_(crane)</td>
</tr>
</tbody>
</table>
Candidate Ranking

- For each mention $m_i$ in document $d$, rank the corresponding candidates $e_j \in T(m_i)$.
  - Local Inference
    - For each $m_i$ in $d$:
      - Generate corresponding candidate set $T(m_i)$
      - Rank entities $e_i \in T(m_i)$.
        (i.e., consider local statistics of edges)
  - Global Inference
    - For each document $d$,
      - Consider all $m_i \in d$, and all $e_i \in T(m_i)$
      - Re-rank entities $e_i \in T(m_i)$
        (i.e., the corresponding $e_i$ of related $m_i$ may also be related)
Local Ranking

- Basic idea:
  - Use similarity measure between candidates $e_i \in T(m_i)$ and mention $m_i$ with its context $c_i$. 

---

```
Scott Young (American football)
Scott Young (writer)
Scott Young (Welsh footballer)
```
Global Ranking

- **Basic idea:**
  - Resolve all mentions simultaneously based on entity coherence and relatedness.
Global Ranking: Illustration

Northern Ireland

James Craig

Catholics

James Craig

James Craig (actor)

James Craig

JC, 1st Viscount Craigavon

Catholic Church

Roman Catholic Church

Catholic Church

American Catholic Church
Global Ranking: Formulation

- A Combinatorial Optimization Problem

- Problem Formulation

\[
\Gamma^* = \arg \max_{\Gamma} \sum_{i=1}^{N} [\phi(m_i, e_i) + \sum_{e_i \in \Gamma_i, e_j \in \Gamma_j} \psi(e_i, e_j)]
\]

- How to define the coherence and relatedness between two candidate entities? (What is \(\psi\)?)
Co-occurrence

- Typical approach: generate a graph containing candidate entities $e$ of mentions $m$ in $d$ and perform random walk algorithms (e.g. PageRank) over it to select highly consistent entities.
- Calculate $\psi(e_i, e_j)$ as Pointwise Mutual Information of $e_i, e_j$ based on their incoming sets.
- ...
NIL Detection

- Binary classification (Within KB vs. NIL)
- Select NIL cutoff by tuning confidence threshold

Augment KB with NIL entry and treat it like any other entry

Marc-Alexandre Côté is an author of *Textworld*.

James Craig was born in Nashville, Tennessee and studied at the Rice Institute.

James Craig attempted to announce his bid for Governor on Belle Isle,...
NIL Clustering

“All in one”

Simple string matching

“One in one”

Often difficult to beat!

Collaborative Clustering

Most effective when ambiguity is high
NIL Detection & Clustering

- Collaborative Clustering
EL Tools

Tools (Target KB: Wikipedia):

- FacebookResearch BLINK  
  - SOTA of zero-shot EL  
  - Well-documented

- OpenAI DeepType  
  - SOTA of EL

- Spacy Entity Linker  
  - Easy to use

```python
import spacy  # version 3.0.6

# initialize language model
nlp = spacy.load("en_core_web_md")

# add pipeline (declared through entry_points in setup.py)
nlp.add_pipe("entityLinker", last=True)

doc = nlp("I watched the Pirates of the Caribbean last silvester")

# returns all entities in the whole document
all_linked_entities = doc._.linkedEntities
# iterates over sentences and prints linked entities
for sent in doc.sents:
    sent._.linkedEntities.pretty_print()

# OUTPUT:
# https://www.wikidata.org/wiki/Q194318  Pirates of the Caribbean
# https://www.wikidata.org/wiki/Q12525597  Silvester
```

https://github.com/egerber/spaCy-entity-linker
EL Tools

OpenAI DeepType

Figuring out which object is meant by a word by having a neural network decide if the word belongs to each of about 100 automatically-discovered “types”.

Rather than trying to reason directly whether *jaguar* means the *car*, *animal*, or something else, the system plays “20 questions” with a pre-chosen set of categories.

https://openai.com/blog/discovering-types-for-entity-disambiguation/
Relation Extraction

**KG Sources**
- Unstructured Sources
- Structured Sources
- Domain Experts

**KG Construction**
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

**Knowledge Graph**

**KG Maintenance**
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

**KG Applications**
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
What is relation extraction (RE)?

- **Company report:** “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co.(C-T-R) …”

- Complex relation in the report:

```
<table>
<thead>
<tr>
<th>Relation</th>
<th>Company</th>
<th>Location</th>
<th>Date</th>
<th>Original-Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company-Founding</td>
<td>IBM</td>
<td>New York</td>
<td>June 16, 1911</td>
<td>Computing-Tabulating-Recording Co.</td>
</tr>
<tr>
<td>Founding-year</td>
<td>IBM, 1991</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Founding-location</td>
<td>IBM, New York</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

- But we will focus on the simpler task of extracting relation **triples:**

  - **Founding-year**(IBM, 1991)
  - **Founding-location**(IBM, New York)
  - ...
Extracting relation triples from text

Textual abstract: Summary for human

Structured knowledge extraction: Summary for machine

<table>
<thead>
<tr>
<th>Subject</th>
<th>Relation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53</td>
<td>is_a</td>
<td>protein</td>
</tr>
<tr>
<td>Bax</td>
<td>is_a</td>
<td>protein</td>
</tr>
<tr>
<td>p53</td>
<td>has_function</td>
<td>apoptosis</td>
</tr>
<tr>
<td>Bax</td>
<td>has_function</td>
<td>induction</td>
</tr>
<tr>
<td>apoptosis</td>
<td>involved_in</td>
<td>cell_death</td>
</tr>
<tr>
<td>Bax</td>
<td>is_in</td>
<td>mitochondrial outer membrane</td>
</tr>
<tr>
<td>Bax</td>
<td>is_in</td>
<td>cytoplasm</td>
</tr>
<tr>
<td>apoptosis</td>
<td>related_to</td>
<td>caspase activation</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Wiki Text

Policing career [ edit ]

Craig first worked at the Detroit Police Department from 1977 until he was laid off in 1981. He went to Los Angeles to work for the LAPD for nearly three decades. From 2009 to 2011, he was police chief for the Portland Police Department in Portland, Maine. From 2011 to 2013, he was chief of police for the Cincinnati Police Department in Cincinnati, Ohio. During his tenure with the Cincinnati police, the crime rate was its lowest since 2000, though it was part of an overall trend of decreasing crime.

Craig first worked at the Detroit Police Department from 1977 until he was laid off in 1981. He went to Los Angeles to work for the LAPD for nearly three decades.
Why Relation Extraction?

- Building & extending knowledge bases and ontologies
- Scholarly literature databases: Google Scholar, CiteSeerX
- People directories: Rapleaf, Spoke, Naymz
- Shopping engines & product search
- Bioinformatics: clinical outcomes, gene interactions, …
- Stock analysis: deals, acquisitions, earnings, hirings & firings
- Intelligence analysis for business & government
- …
But what relations should we extract?

- Depends on the **domain of interest**.
  - For generic news texts
    - Freebase relation types
      - 23 million entities, thousands of relations
      - | Relation name | Size | Example |
        |----------------|------|---------|
        | /people/person/nationality | 281,107 | John Dugard, South Africa |
        | /location/location/contains | 253,223 | Belgium, Nijlen |
        | /people/person/profession | 208,888 | Dusa McDuff, Mathematician |
        | /people/person/place_of_birth | 105,799 | Edwin Hubble, Marshfield |
        | /dining/restaurant/cuisine | 86,213 | MacAyo’s Mexican Kitchen, Mexican |
        | /business/business_chain/location | 66,529 | Apple Inc., Apple Inc., South Park, NC |
        | /biology/organism_classification_rank | 42,806 | Scorpaeniformes, Order |
        | /film/film/genre | 40,658 | Where the Sidewalk Ends, Film noir |
        | /film/film/language | 31,103 | Enter the Phoenix, Cantonese |
        | /biology/organism_higher_classification | 30,052 | Calopteryx, Calopterygidae |
        | /film/country | 27,217 | Turtle Diary, United States |
  - Geographical relation types
  - Protein interactions

![Diagram of Geographical Entity (GE)](image)

> “We show that **CBF-A** and **CBF-C** interact with each other to form a **CBF-A-CBF-C complex** and that **CBF-B** does not interact with **CBF-A** or **CBF-C** individually but that it associates with the **CBF-A-CBF-C complex**.”
Relation extraction approaches

- Rule-based RE
- Weakly supervised RE
- Supervised RE
- Distantly supervised RE
- Unsupervised RE
Many instances of relations can be identified through hand-crafted patterns, looking for triples \((X, r, Y)\) where \(X, Y\) are entities and \(r\) are words in between.

- A simple rule

\[
\text{Paris is in France} \quad \overset{(e_{\text{subj}}, r, e_{\text{obj}})}{\rightarrow} \quad (\text{Paris, is_in, France})
\]

- Consider entity type and part-of-speech (POS) tags to remove additional false positive

\[
\text{Paris is in France} \quad \overset{(\text{CITY, } r, \text{ COUNTRY})}{\rightarrow} \quad (\text{Paris, is_in, France})
\]

- Use dependency paths for longer-range patterns and increase coverage.

\[
\text{Fred and Mary got married} \quad \overset{\text{Dependency-based rules}}{\rightarrow} \quad (\text{Fred, got_married, Mary})
\]
Rule-based RE

Pro
- Humans created patterns tend to lead high precision
- Can be tailored to specific domains

Cons
- Human patterns suffer from low-recall
- Creation of patterns can be expensive or even intractable
  - Hard to write and maintain the rules
  - Have to create rules for every relation type

Examples of the Hearst patterns

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$ and other $Y$</td>
<td>temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>$X$ or other $Y$</td>
<td>bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>$Y$ such as $X$</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>such $Y$ as $X$</td>
<td>...such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>$Y$ including $X$</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>$Y$, especially $X$</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
Weakly Supervised RE

**Basic idea:**
- **Bootstrapping:** start out with a set of hand-crafted rules and automatically find new ones from the *unlabeled text data* through an iterative process.
Weakly Supervised RE: Example

- **Target relation:** burial place

- **Seed tuple:** [Mark Twain, Elmira]

- Grep/Google for “Mark Twain” and “Elmira”
  - “Mark Twain is buried in Elmira, NY”  
    - X is buried in Y
  - “The grave of Mark Twain is in Elmira”  
    - The grave of X is in Y
  - “Elmira is Mark Twain’s final resting place”  
    - Y is X’s final resting place

- Use those patterns to search for new tuples
Pros

- **Higher recall**: more relations can be discovered
- **Less cost**: only require a high quality seed

Cons

- **Sensitive to** original set of **seeds**
- Problem of **semantic drift** at each iteration
- **Precision** tends to be **not that high**
- **Still hard for relation type extension**: require manually labeled new seeds for new relation types

A classical work about weakly supervised RE

**Snowball (Agichtein & Gravano 2000)**

Require that X and Y be named entities of particular types

```
<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>'s, 0.4</th>
<th>headquarters, 0.4</th>
<th>in, 0.1</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCATION</td>
<td>-0.75</td>
<td>based, 0.75</td>
<td>ORGANIZATION</td>
<td></td>
</tr>
</tbody>
</table>
```
The supervised approach requires:

- Defining an inventory of output **labels**
  - Relation detection: **True/False**
  - Relation classification: **located_in, employee_of, be_born_in**, …

- Collection sufficient **labeled training data** (Hard Part)
  - Public benchmarks: MUC-7, ACE-2005, SemEval, NYT10, FewRel, …

- Designing architecture of neural networks for **DL-based methods**
  (or designing feature representation & choosing classifier for **ML-based methods**)
  - CNNs, RNNs, Transformers, …

- **Training and evaluation**
How are neural networks being used to do RE?

[Zeng et al. 2014] apply word embeddings and Convolutional Neural Network (CNN) to relation classification.

(a) Neural Network Architecture for Relation Classification.

(b) Extracting Sentence-level Features using CNN.

lexical features: specific to entity pairs

Word Feature

Position Feature

https://aclanthology.org/C14-1220/
[Liu et al. 2015] propose a **recursive neural network** (RNN) to model the **dependency subtrees**, and a CNN to capture the most important features on the shortest path.

1. Get subtree representations by Recursive NN
2. Get sentence level representation by CNN through the shortest path on dependency tree from subject to object.

https://aclanthology.org/P15-2047/
How are neural networks being used to do RE?

[Wu and He 2019] apply BERT pre-trained language model to relation extraction. (R-BERT)

Supervised RE

**Pros:**
- High accuracy
  - At least for some seen relations
  - If we have sufficient training data

**Cons:**
- Performance is limited by the amount of expensive labeling samples
- Does not generalize well to new domains

**Beyond supervised RE?**
- Collect sufficient labeled data at low cost: Distantly supervised RE
- Get rid of the reliance on labeled data: Unsupervised RE
**Basic idea:**

- Combine the idea of **using seed data** (as for Weakly Supervised RE) with **training a classifier** (as for Supervised RE).
- Instead of providing seed tuples ourselves, we can take it from KBs.
For each relation
- For each tuple in KB
  - Find sentences in large corpus with both entities
  - Assume these sentences are positive examples for this relation type
- Train a supervised classifier on this distantly-labeled dataset

Born_In

<(Edwin Hubble, Born_In, Marshfield)>
<(Albert Einstein, Born_In, Ulm)>

S1: Hubble was born in Marshfield
S2: Einstein, born (1879), Ulm
S3: Hubble … Steven Tyler … Marshfield
Distantly Supervised RE

**Pros**
- Less manual effort
- Can **scale** to large amount of corpus and many relations
- **No iterations required** (unlike Weakly Supervised RE)

**Cons**
- Generated **labels** can be **noisy**
- Be **restricted to** the referred KBs

[Mintz et al. 2009] propose distant supervision to reduce the cost of manually labeling corpus.

Distant supervision for relation extraction without labeled data

Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky
Stanford University / Stanford, CA 94305
{mikemintz,sbills, rion, jurafsky}@cs.stanford.edu

https://aclanthology.org/P09-1113/
Unsupervised RE for Open Information Extraction

**Basic idea:**
- Extract relations from text based on a set of very **general constraints and heuristics**, thus get rid of the reliance on labeled data, seed tuples, and hand write rules.

---

**Constraints**
1. There exists a dependency chain between \( e_1 \) and \( e_2 \) that is not longer than a certain length.
2. The chain should contain some words of the reasoner (usually the main verb).
3. The path from \( e_1 \) to \( e_2 \) along the syntactic tree doesn’t cross the sentence-like boundary (e.g., relative clauses). This means that this path can contain some (NP, ADJ, etc.) constituents only at the common ancestor position.
4. Entities do not consist solely of the pronoun.
5. \( e_2 \) should contain at least one VP tag.
6. \( e_1 \) and \( e_2 \) should have at least one VP tag as a common ancestor.

---

**TextRunner**
[Bach, Nguyen, and Sameer Badaskar 2007]

**Constraints are not relation-specific patterns.**
Unsupervised RE

Influential work by Allen Institute for AI (Oren Etzioni’s group)

- 2005: KnowItAll
  - Generalizes Hearst patterns to other relations
  - Requires zillions of search queries; very slow
- 2007: TextRunner
  - No predefined relations; highly scalable; imprecise
- 2011: ReVerb
  - Improves precision using simple heuristics
- 2012: Ollie
  - Operates on Stanford dependencies, not just tokens
- 2013: OpenIE 4.0
  - Use heuristics from semantic role labeling (SRL) frames
  - Extends the definition of Open IE extractions to include n-ary extractions
- 2018: OpenIE 5.1
  - Improves upon extractions from numerical sentences
  - Improves upon conjunctive sentences by breaking conjunctions in arguments to generate multiple extractions.
- 2020: OpenIE6
  - Iterative Grid Labeling and Coordination Analysis
Unsupervised RE

**Pros:**
- No/Almost none labeled training data required
- Without the require of manually pre-specifying all the relation of interest, it considers all possible relation types

**Cons:**
- Performance depends a lot on how well constructed the constraints and heuristics are
- Relations are not normalized as pre-specified relation types

---

<table>
<thead>
<tr>
<th>Argument 1</th>
<th>Relations</th>
<th>Argument 2</th>
<th>Compara</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is bacteria?</td>
<td>kids</td>
<td>what contains antibiotics?</td>
<td>All</td>
</tr>
</tbody>
</table>

Example Queries:
- What is bacteria?
- What are the Pyramids?
- What did Thomas Edison invent?
- What contains antibiotics?

Typed Example Queries:
- What countries are located in Africa?
- What actors starred in which film?
- What is the symbol of which country?
- What foods are grown in which country?
- What drug is approved by the FDA approved?

To learn more about Open IE, watch our YouTube video!

Powered by Ro Web; our Open Information Extraction, yielding over 5 billion extractions from over a billion web pages.

Open IE 4.0, the successor to ReWeb and Cite, has been released. Download it from GitLab!

Public resources based on Open IE:
- 1.4 million question-processing (Linder et al. ACL, 2013)
- 3.8B Replays for named entities (Bick and <unknown> 2013)
- 1.0M Replays for relations (Linder et al. EMNLP 2013)
- 1.0M Replays for constraints (Linder et al. EMNLP 2013)
- 1.0M Replays for relations (Linder et al. ACL 2013)
- 1.0M Replays for relations (Linder et al. ACL 2013)

Codes for extracting sentences from CiteWeb.
Relation Extraction: Tools

- **Domain-dependent:**
  - spaCy Kindred (Biomedical)
  - OpenNRE (NYT10m and Wiki20m)
  - ...

- **Domain-independent:**
  - Stanford OpenIE
    - About 4x faster than AllenAI OpenIE, while brings lower performance.
  - AllenAI OpenIE 5.1
    - Utilize pre-trained language model
    - Higher performance, with higher computational cost.
  - ...

Open Information Extraction

Example Queries:
- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

Typed Example Queries:
- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?

To learn more about Open IE, watch our YouTube video!

Powered by ReVerb, our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

Open IE 4.0, the successor to ReVerb and Ollie, has been released. Download it from GitHub!

https://openie.allenai.org/
Coreference Resolution

What is Coreference Resolution?
- Coreference Resolution is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity.

Why Coreference Resolution?
- Coreference resolution could increase the overall efficiency and the coverage of the information extraction.

Tools:
- spaCy·neuralcoref, NLTK·coref, BART, …
Quality Checking

KG Sources
- Unstructured Sources
- Structured Sources
- Domain Experts

KG Construction
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

KG Maintenance
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

Knowledge Graph

KG Applications
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
Quality Checking

Not everything accurately extracted is fact

Raw Text

```
"I voted for Trump because he was most aligned with my values", John said.
```

Information Extraction

Structured Knowledge

(Subject, predicate, Object)

(John, vote_for, Trump)

Is this a fact?

The claim is with the context “John said”.

We need to further check the quality of the extracted knowledge.
Quality Checking

- Not everything accurately extracted is fact

- Many recent efforts on assessing truth and finding supports
  - Knowledge-based Trust [Dong et al. 2015]
  - Multilingual answer validation [Rodrigo et al. 2009] [Kobayashi et al. 2017]
  - FactChecker [Nakashole & Mitchell 2014]
  - PolitiFact [Vlachos & Riedel 2014], [Wang 2017]
  - Fake News challenge [Pomerleau & Rao 2017]
  - Fake news detection via crowd signals [Tschiatschek et al. 2018]
  - Fact Verification competition [Thorne et al. 2018]
Quality Checking

- Fact Verification competition (FEVER) [Thorne et.al 2018]

- Goal:
  - Given a claim, label claim SUPPORTS, REFUTES, or NOT-ENOUGH-INFO
  - For the first two classes, select relevant sentences from Wikipedia intro sections.

- Largest annotated fact sets
  - 185,445 annotated claims.
  - Claims generated by mutating Wikipedia sentences: paraphrasing, negation, substitution of entity/relation, generalize/specialize claims

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los Angeles Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots, were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los Angeles County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported
Quality Checking

○ FEVER baseline – sentence classification [Thorne et al. 2018]

• Basic idea: **align parts** of the text in sentences **a** and **b** and then aggregate info to predict the label

• Example

Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.  
Bob is awake.  
It is sunny outside.
### FEVER Progress

#### 2021 Shared Task

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>FEVEROUS score</th>
<th>Accuracy</th>
<th>Evidence F1</th>
<th>Evidence Precision</th>
<th>Evidence Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bust a move!</td>
<td>0.2701</td>
<td>0.5607</td>
<td>0.1308</td>
<td>0.0773</td>
<td>0.4258</td>
</tr>
<tr>
<td>2</td>
<td>Papelo</td>
<td>0.2592</td>
<td>0.5757</td>
<td>0.1187</td>
<td>0.0716</td>
<td>0.3460</td>
</tr>
<tr>
<td>3</td>
<td>NCU</td>
<td>0.2514</td>
<td>0.5229</td>
<td>0.1581</td>
<td>0.0991</td>
<td>0.3907</td>
</tr>
<tr>
<td>4</td>
<td>Z team</td>
<td>0.2251</td>
<td>0.4901</td>
<td>0.1312</td>
<td>0.0776</td>
<td>0.4264</td>
</tr>
<tr>
<td>5</td>
<td>EURECOM_Fever</td>
<td>0.2001</td>
<td>0.4779</td>
<td>0.1952</td>
<td>0.1373</td>
<td>0.3373</td>
</tr>
<tr>
<td>6</td>
<td>FEVEROUS Baseline</td>
<td>0.1773</td>
<td>0.4548</td>
<td>0.1503</td>
<td>0.1017</td>
<td>0.2878</td>
</tr>
<tr>
<td>7</td>
<td>Saturday_Night_Fever</td>
<td>0.1763</td>
<td>0.4804</td>
<td>0.1618</td>
<td>0.1122</td>
<td>0.2900</td>
</tr>
<tr>
<td>8</td>
<td>Martin Funkquist</td>
<td>0.1261</td>
<td>0.4302</td>
<td>0.1045</td>
<td>0.0642</td>
<td>0.2789</td>
</tr>
<tr>
<td>9</td>
<td>Albatross</td>
<td>0.1159</td>
<td>0.4035</td>
<td>0.0963</td>
<td>0.0644</td>
<td>0.1902</td>
</tr>
<tr>
<td>10</td>
<td>METUIS</td>
<td>0.0636</td>
<td>0.3897</td>
<td>0.0634</td>
<td>0.0462</td>
<td>0.1011</td>
</tr>
<tr>
<td>11</td>
<td>ChaCha</td>
<td>0.0389</td>
<td>0.4194</td>
<td>0.0398</td>
<td>0.0251</td>
<td>0.0969</td>
</tr>
<tr>
<td>12</td>
<td>seda_kaist</td>
<td>0.0362</td>
<td>0.4140</td>
<td>0.0384</td>
<td>0.0242</td>
<td>0.0920</td>
</tr>
<tr>
<td>13</td>
<td>qmul_uou_iith</td>
<td>0.0223</td>
<td>0.3999</td>
<td>0.0282</td>
<td>0.0245</td>
<td>0.0330</td>
</tr>
</tbody>
</table>

[https://fever.ai/task.html](https://fever.ai/task.html)
KG Hosting

KG Sources
- Unstructured Sources
- Structured Sources
- Domain Experts

KG Construction
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

KG Maintenance
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

Knowledge Graph

KG Applications
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
After checking the quality of extracted knowledge, we can store the RDF triples as knowledge graph by:

- Storing them into RDF Triple Stores. ([https://www.w3.org/wiki/LargeTripleStores](https://www.w3.org/wiki/LargeTripleStores))

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th># triples tested with</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oracle Spatial and Graph with Oracle Database 12c</td>
<td>1.08 T</td>
</tr>
<tr>
<td>2</td>
<td>AnzoGraph DB by Cambridge Semantics</td>
<td>1.065 T</td>
</tr>
<tr>
<td>3</td>
<td>AllegroGraph</td>
<td>1+ T</td>
</tr>
<tr>
<td>4</td>
<td>Stordog</td>
<td>50 B</td>
</tr>
<tr>
<td>5</td>
<td>OpenLink Virtuoso v7+</td>
<td>39.8 B</td>
</tr>
<tr>
<td>6</td>
<td>GraphDB™ by Ontotext</td>
<td>17 B</td>
</tr>
</tbody>
</table>

- Or converting them to nodes and edges and then storing them into Graph Databases.
Graph Databases vs. RDF Triple Stores

- **How They Are Similar**
  - Both two focus on the relationships between the data, often referred to as “linked data”.

- **How They Are Different**
  - Graph databases are more versatile with query languages: Neo4J can run an RDF triple store and use SPARQL but generally focuses on its own proprietary language, Cypher; RDF triple stores only use SPARQL as the query language.
  - Graph databases can store various types of graphs; RDF triple stores focus solely on storing rows of RDF triples.
  - Graph databases are node centric, whereas RDF triple stores are edge-centric.
  - Graph databases are better optimized for graph traversals, while the cost of traversing an edge tends to be logarithmic with RDF triple stores.
  - Graph databases do not provide inferences on data, but RDF triple stores do.
Neo4j, a graph database for high-performance graph queries on large datasets. More on https://neo4j.com/
Storing as **Knowledge Graph**:

- **Query**:
  ```
  1 MATCH p=shortestPath((n {name: "SARS-CoV-2"})-[*1..3]-(m:Disease))
  2 RETURN p LIMIT 10;
  ```

- **Cypher**

- **Python Driver**
  [https://neo4j.com/developer/python/](https://neo4j.com/developer/python/)

- **Others**

- **Summary**:
  high performance graph store, multiple programming language driver supported robust graph database
**KG Maintenance**

**KG Sources**
- Unstructured Sources
- Structured Sources
- Domain Experts

**KG Construction**
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

**KG Maintenance**
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

**Knowledge Graph**

**KG Applications**
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
The extracted facts can be **incomplete**, **conflicting**, and **incorrect**:
- Missing Data
- Data error due to IE tech or human errors
- Abbreviations and truncation
- …

Mentions in digital world is biased and various. We need to de-bias and handle the diversity.
KG Maintenance

- Knowledge Assessment

- Knowledge Cleaning
  - Error detection
  - Error correction

- Knowledge Enrichment
  - Identifying new knowledge sources
  - Data integration
First step to improve the quality of a KG: **assess the situation**

Closely related to **data quality** literature

Various dimensions for data quality assessment introduced [Batini & Scannapieco, 2006], [Färber et al., 2018], [Pipino et al., 2002], [Wang, 1998], [Wang & Strong, 1996], [Cao et al., 2001], [Zaveri et al., 2016]
KG Assessment

1. accessibility
2. accuracy (veracity)
3. appropriate amount
4. believability
5. completeness
6. concise representation
7. consistent representation
8. cost-effectiveness
9. easy of manipulating
10. easy of operation
11. easy of understanding
12. flexibility
13. free-of-error
14. interoperability
15. objectivity
16. relevancy
17. reputation
18. security
19. timeliness (velocity)
20. traceability
21. understandability
22. value-added
23. variety
24. fitness for use

......
Several core assessment dimensions for KG

Coverage
Have we got the information we need?

Freshness
Is information up to date?

Correctness
Is our information accurate?

Three forces in constant conflict:

- Increased freshness and coverage → Harder to ensure correctness
- Increased correctness → Harder to ensure freshness and coverage
- Correctness is always hard – what is true and correct? Particularly critical in today’s world

Will Smith: Single entity, 108K facts assembled from 41 web sites. There are 200 Will Smiths on Wikipedia alone.
Knowledge Assessment Tools

- LINK-QA [Guéret et al., 2012]
  - Benefits from network features to assess data quality (e.g. counting open chains to find wrongly asserted isSameAs relationships)

- Sieve [Mendes et al., 2012]
  - Uses data quality indicators, scoring functions and assessment metrics

- SDType [Paulheim & Bizer, 2013]
  - Uses statistical distributions to predict the types of instances. Incoming and outgoing properties are used as indicators for the types of resources.

- RDFUnit [Kontokostas et al., 2014]
  - A framework that assesses linked data quality based on test cases defined in various ways (e.g. RDFS/OWL axioms can be converted into constraints)

- Validata [Hansen et al., 2015]
  - An online tool check the conformance of RDF graphs against ShEx (Shape Expressions)

- Luzzu [Debattista et al., 2016]
  - Allows declarative definitions of quality metrics and produces machine-readable assessment reports based on Dataset Quality Vocabulary
Knowledge Cleaning

- The actions taken to **improve the correctness** of a knowledge graph.

- Two major steps:
  - Error **detection**
  - Error **correction**
Knowledge Cleaning Tools

- **DValidate [Paulheim & Bizer, 2014]**
  - Uses statistical distribution to detect erroneous statements that connect two resources. The statements with less frequent predicate-object pairs are selected as candidates for being wrong.

- **KATARA [Chu et al., 2015]**
  - Learns the relationships between data columns and validate the learn patterns with the help of existing Knowledge Bases and crowd, in order to detect errors in the data. Afterwards it also suggests possible repairs.

- **HoloClean [Rekatsinas et al., 2017]**
  - An error detection and correction tool based on integrity constraints to identify conflicting and invalid values, external information to support the constraints, and quantitative statistics to detect outliers.

- **TISCO [Rula et al., 2019]**
  - A framework that tries to identify the time interval where a statement was correct. It uses external knowledge bases and the web content to extract evidence to assess the validity of a statement for a time interval.
Knowledge Enrichment

- Improve the *completeness* of a knowledge graph by adding new statements.
Knowledge enrichment: the property level
Knowledge enrichment: the instance level (ABox)

ABox statements: "assertion component" – facts associated with the TBox's conceptual model or ontologies.
Knowledge enrichment: the terminological level (TBox)

TBox statements: "terminology component" describe a domain of interest by defining classes and properties as a domain vocabulary.
Knowledge Enrichment Tools

Duplication detection and resolution tools

- **Dedupe**
  - A Python library that uses machine learning to find duplicates in a dataset and to link two datasets.

- **LIMES** [Ngomo & Auer, 2011]
  - A link discovery approach that benefits from the metric spaces (in particular triangle inequality) to reduce the amount of comparisons between source and target dataset.

- **Duke** [Garshol & Borge, 2013]
  - Uses various similarity metrics to detect duplicates in a dataset or link records between two datasets based on a given configuration.

- **Legato** [Achichi et al., 2017]
  - A recording linkage tool that utilizes Concise Bounded Description* of resources for comparison.
Demo part 1
Constructing KG
KG Applications
KG Applications

KG Sources
- Unstructured Sources
- Structured Sources
- Domain Experts

KG Construction
- Information Extraction
  - Preprocessing
  - Named Entity Recognition
  - Entity Linking
  - Relation Extraction
  - Coreference Resolution
  - Quality Checking
  - KG Publication & Hosting

Knowledge Graph

KG Maintenance
- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

KG Applications
- Question Answering
- Dialog System
- Recommender System
- Information Retrieval
- Domain Specific Apps
- Other Applications
Knowledge graph has a tremendous impact on Bing impressions
KG Applications: Office

- Enriching the Office experience with knowledge graph.
KG-based Question Answering

- **Given:**
  - Knowledge graph ingested from unstructured, structured, and semi-structured data sources

- **Input:**
  - Natural language query

- **Output:**
  - Answer in the form of knowledge
Pete Carroll

Pete Carroll has no worries about job status: ‘I’m in great shape’ On a day in which some NFL head coaches might see their tenures end, Pete Carroll isn’t fretting his future in Seattle following Sunday’s 38-30 season-ending victory over the Arizona Cardinals. Jan 10, 2022

https://www.nfl.com/news/seahawks-coach-pete-carroll-

Seahawks coach Pete Carroll has no worries about job status

associated researches

Bill Belichick, Russell Wilson, Gena Goranson, Brennan Carroll, Andy Reid, Urban Meyer, Mike Tomlin
KG-based Question Answering

• Challenges:
  • Matching natural language relations to formal relations in KGs
  • There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, …etc
  • Scalable entity linking
  • Word sense disambiguation
  • Semantic roles and relationships extraction

• Large search space
  • Every entity can have hundreds of edges and every entity instance can have hundreds of millions of edges/facts

• Compositionality: multi-hop reasoning
  • {Movies starring the first wife of tom hanks}, {movies directed by the director of titanic}
Approaches for KG-based Question Answering

- Semantic parsing approaches:
  - Generic semantic parsing followed by ontology grounding

- Knowledge embedding

- Information extraction approaches:
  - Information retrieval methods with semantic enrichment
Semantic Parsing

2017 movies starring the actor that played Batman in Batman


Spider-Man: Homecoming
American Assassin

The Founder

Character Name
“Batman”

Film:Character Entity 3c_b
Character:Performance
Role:Relationship 7r
Film:Actor Entity 7a
Actor:Performance
Film:Actor

year(2017)

Film:Release_Date
Film Entity 7f
Film Entity 7f_b
Film:Name
“Batman”
In this approach as in the example provided by [Kwiatkoski 13], we:
1. Perform a generic semantic parsing of the utterances
2. Perform ontology matching on relationships

For example: \{**who is Donald Trump’s Daughter**\}
\[
\lambda x. \text{daughter\_of\ Donald\ Trump, } x
\]
\[
\lambda x. \text{child\_of\ Donald\ Trump, } x \wedge \text{gender\ (x, female)}
\]

This semantic expression can be then compiled into a knowledge graph database query e.g. Cypher/SPARQL and executed to return the results.
Ontology Matching on Relationships using DSSM

- Input is mapped into two $k$ dimensional vectors.
- Probability is determined by \textit{Softmax} of their cosine similarity.

$$P(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$$

![Diagram](https://aclanthology.org/D14-1002.pdf)
Knowledge Embedding Approach

Question $q$

"Who did Clooney marry in 1987?"

Detection of Freebase entity in the question

Freebase subgraph

G. Clooney

K. Preston

Honolulu

Model

1987

J. Travolta

Subgraph of a candidate answer $a$ (here K. Preston)

Information Extraction Approaches

- Extracting and answers on the fly.

- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.
Answer Type Detection

- Who first landed on the moon => Person
- Where is the headquarters of Microsoft => Location
- What is the largest country in population => Country
- Highest flying bird => Animal/Bird
Answer Type Detection

- **Rules**
  - Grammar for e.g. who be/... => Person
  - Head word for e.g. which city is the largest
  - ...

- **Learned type classifier**
  e.g. SVM utilizing features like question words, phrases, POS tags, headwords, mentioned entities, ...etc [Dan Jurafsky]
Passage Retrieval

- Retrieve documents using expanded query terms + search engine
- Segment the documents into smaller units e.g. passages/paragraphs
- Rank passages using learned model utilizing features like:
  - Number of named entities of the right type in the passage
  - Number of query words in the passage
  - Number of question n-grams in the passage
  - Proximity of query words in the passage
  - Longest sequence of question words
  - Rank of document containing passage,...etc
Detect answer entity by running NER on the passage

Mark the answer entity in the passage

- How many bones in an adult human body? (Number)
- The human skeleton is the internal framework of the body. It is composed of 270 bones at birth – this total decreases to 206 bones by adulthood after some bones have fused together.

Question Answering [Dan Jurafsky, Stanford]
Answer Semantic Enrichment using KB

- Question:
  Who was the first American in space?

- Sentence Collection:
  1. On May 5, 1961, Shepard piloted ...
  2. Alan Shepard became the first American ...
  3. ...

- Top-K Answers:
  1. Alan Shepard
  2. Sally Ride
  3. John Glenn
  4. ...

- Answer Candidate Pool:
  1. Freedom 7; 2. Alan Shepard
  3. Sally Ride; 4. ...

- Feature Generation & Ranking

- Candidate Generation:
  Via Entity Linking

- Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

- Freebase

- 5-20% mean reciprocal rank (MRR) improvement

https://dl.acm.org/doi/10.1145/2736277.2741651
KG-based Dialogs

Approaches:

- E2E Seq2seq (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015)
- Knowledge based ontological slot filling (Dai+ 2017)
- Knowledge grounded neural approaches (Ghazvininejad+ 2018)
- ...
E2E Dialog Systems

- Suitable for chitchat kind of bots.
- Predicted target sequences are usually free from facts.
Suitable for chitchat kind of bots.

Predicted target sequences are usually free from facts.
Knowledge Grounded Neural Approaches

![Diagram showing a flow of conversation with facts encoded and decoded to generate a response.](https://arxiv.org/pdf/1702.01932.pdf)
Enterprise Scenarios

All the challenges mentioned previously plus the following:

- Compliance

- Different data formats: databases, emails, chat logs, discussion forums, web blogs, pdfs, PowerPoint/Word/Excel documents etc.

- Different schemas: schema mapping and merging, and new schema discovery.

- Consumption via dialog systems, search interface, mobile devices or other modalities, API.

- Highly domain-specific models required, bootstrapped by pre-trained models. Need on-prem domain-adaptation.
Exciting Active Research
**Richpedia**: provide a comprehensive multi-modal knowledge graph by distributing sufficient and diverse images to textual entities in Wikidata.
Choosing an intuitive follow-up scene

Q: On stage, a woman takes a seat at the piano. She:
1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
4. nervously sets her fingers on the keys.

Answering this question requires knowledge that humans possess and apply, but machines cannot distill directly in communication.

- **ConceptNet**: pianos have keys and are used to perform music
- **WordNet**: pianos are played by pressing keys
- **ATOMIC**: before a person plays piano, they need to sit at it, be on stage, and reach for the keys.
- **FrameNet**: frame of a performance contains two separate roles for the performer and the audience, meaning that these two are distinct entities, which can be seen as evidence against answer
**Causal Knowledge Graph**

- **CausalKG**: Causal Knowledge Graph - Explainability using interventional and counterfactual reasoning

![Diagram of Causal Knowledge Graph Framework](image)

**Figure 2.** Causal Knowledge Graph Framework consists of three main steps, i) a Causal Bayesian Network and a domain-specific observational dataset, ii) Causal Ontology creation and enriching the domain ontology with causal relationships, and iii) Estimating the causal effects of the treatment, mediator, and outcome variable in the domain for a given context.
More research on knowledge graphs

- Point-wise - Manifold - Complex - Gaussian - Discrete
- Distance - Semantic Matching - Others

- Linear/Bilinear - Factorization - Neural Nets - CNN - RNN - Transformers - GCN
- Recognition - Typing - Disambiguation - Alignment

- Neural Nets - Attention - GCN - GAN - RL - Others
- Embedding-based Ranking - Path-based Reasoning - Rule-based Reasoning - Meta Relational Learning - Triple Classification

Knowledge Graph

Representation Space

Natural Language Understanding

Question Answering

Knowledge-Aware Applications

Dialogue Systems

Recommender Systems

Others Applications

- Single-fact QA - Multi-hop Reasoning
- Question Generation - Search Engine - Medical Applications - Mental Healthcare - Zero-shot Image Classification - Text Generation - Sentiment Analysis

Temporal Knowledge Graph

Temporal Embedding

Entity Dynamics

Temporal Relational Dependency

Temporal Logical Reasoning

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called chatGPT, which interacts in a conversational way. The dialogue format makes it possible for chatGPT to answer followup questions, admit mistakes, correct premises, and reject inappropriate suggestions. chatGPT is a sibling model to InstructGPT, which is a parameterized instruction in a prompt and provide a
Hi ChatGPT, can you briefly yourself and GPT-4?

Hello! I'm ChatGPT, a large language model designed by OpenAI. I was trained using the GPT-3.5 architecture, which is a variant of the GPT-3 architecture. As a language model, I'm capable of generating human-like responses to text-based inputs, thanks to my extensive training on a diverse range of text data.

As for GPT-4, I'm not sure about its specific details as it has not been officially announced or released yet. However, based on OpenAI's previous developments with GPT models, it's expected to be an even more powerful language model than its predecessors. It may have a larger training dataset, improved training algorithms, and potentially new features that enhance its language understanding and generation capabilities. Nonetheless, this is all speculation, and we'll have to wait for official announcements from OpenAI for more accurate information.

https://chat.openai.com/chat
Large Language Models: ChatGPT and GPT-4

ChatGPT
ChatGPT, or Chatbot Generalized Pre-trained Transformer, is an advanced language model developed by OpenAI. It builds on the success of the GPT series of language models and is specifically designed for generating human-like conversational responses. ChatGPT is capable of understanding context, answering questions, and providing relevant information in a coherent and fluent manner. It is widely used in various applications, including customer support, virtual assistants, content creation, and more. The model is trained on diverse text data from the internet, which allows it to generate contextually appropriate responses in a wide range of domains.

GPT-4
GPT-4, the fourth iteration of the Generalized Pre-trained Transformer series, is a hypothetical and more advanced version of the GPT-3 model. As GPT-3 is the state-of-the-art language model at the time of this response's knowledge cutoff in September 2021, GPT-4 represents a potential future development in the field. Like its predecessors, GPT-4 would be expected to showcase even greater performance and capabilities in natural language understanding, generation, and other language-related tasks. With each iteration of the GPT series, the models typically become more powerful, showcasing better context understanding, improved coherence, and a wider range of applications.

Ask GPT-4

Ask Bing

Can you briefly introduce yourself, ChatGPT, and GPT-4?

Searching for: introduction to ChatGPT and GPT-4
Searching for: introduction to ChatGPT and GPT-4 wiki
Searching for: introduction to ChatGPT and GPT-4 openai
Generating answers for you...

ChatGPT is a conversational AI model developed by OpenAI that can generate human-like responses to text-based inputs. It is a member of the generative pre-trained transformer (GPT) family of language models. ChatGPT Plus is a version of ChatGPT based on GPT-4, which was released on March 14, 2023. GPT-4 is a large multimodal model that accepts image and text inputs and emits text outputs. It exhibits human-level performance on various professional and academic benchmarks. It was created by OpenAI and is the fourth in its GPT series.

Learn more:
1. en.wiki/pedia.org
2. en.wiki/pedia.org
+11 more
Do we still need knowledge graph?

Yes, Knowledge Graphs can still be valuable even with the advancements of language models like ChatGPT and future models like GPT-4. While language models like ChatGPT and GPT-4 are capable of generating responses to a wide range of questions and inputs, they may still lack the depth and breadth of knowledge that a well-structured Knowledge Graph can provide.

Knowledge Graphs can organize and represent complex information in a structured and machine-readable format, which can be used to augment the language models' understanding of certain concepts and relationships between them. Knowledge Graphs can also help address issues with ambiguity and context that may arise in language-based interactions.

Moreover, Knowledge Graphs can be used for other applications beyond just question-answering, such as semantic search, recommendation systems, and data integration across different domains and sources.

In summary, while ChatGPT and GPT-4 are powerful language models, Knowledge Graphs can still play a valuable role in enhancing their understanding of complex information and improving the accuracy of their responses.
Do we still need knowledge graph?

- Knowledge graph is a structured, explicit, stable representation and storage format for information/knowledge.

- Large language models like ChatGPT stores implicit knowledge with their weights.

- Knowledge graph and ChatGPT play different roles
  - Like memory and CPU
Knowledge graph for large language models

KG can help address a few of the major shortcomings of language models

- **Memory**
  - Knowledge graphs can provide accurate information about any specific subject, which ChatGPT can incorporate into its text generation, resulting in more informative and accurate output. It helps reducing the *hallucination* problem of language models.

- **Logic and reasoning**
  - Knowledge graphs provide a structured way of linking concepts together, allowing ChatGPT to make inferences and draw conclusions based on the information it has been provided, thus *improving the reasoning capabilities of the model*.

- **Interpretability**
  - By observing what memories are accessed, we can *understand sources of resulting statements* and information flow.
  - Alternatively, by observing how the language model processes and categorizes information, we can understand the inherent structure of the information learned by the raw language model.

- **Computational resource use for training**
  - Knowledge Graphs can store vast amount of data, which allows chatbot providers to *scale* their chatbot to handle large numbers of users and queries, without continuously retraining the model parameters.

- **Personalization**
  - By using knowledge graph, chatbot providers can create *personalized chatbot experiences* for their customers, as the chatbot can use the user’s browsing history, interests, and demographic data to provide personalized recommendations and responses.
Zero-Shot Information Extraction via Chatting with ChatGPT

Xiang Wei¹, Xingyu Cui¹, Ning Cheng¹, Xiaobin Wang², Xin Zhang, Shen Huang², Pengjun Xie², Jinan Xu¹, Yufeng Chen¹, Meishan Zhang, Yong Jiang², and Wenjuan Han¹

¹ Beijing Jiaotong University, Beijing, China
² DAMO Academy, Alibaba Group, China

Large language models for knowledge graph

IE with prompt engineering

sentence: 《我的爱情日记》是1990年在北京上映的中国剧情电视剧，由吴天戈执导，苏瑾、孙恩鸽等主演

“My Love Diary” is a Chinese TV series released in Beijing in 1990, directed by Wu Tiange and starred by Su Jin and Sun Shilan.

The given sentence is: "sentence" in the list of entity/relation/event types: [...] in this sentence, what entity/relation/event types might be included in the sentence? ...

According to the given sentence, the type of two entities are (Film-TV, Works, Date) and the relation between them is release-time, please find the two entities...

Where is My Love Diary released? ...

relation: release-time, subject: My Love Diary, subject_type: Film-TV works, object: <1990, Beijing, object_type: <Date, Location>
Symbolic Knowledge Distillation: from General Language Models to Commonsense Models

Peter West†, Chandra Bhagavatula‡, Jack Hessel†, Jena D. Hwang†
Liwel Jiang†, Ronan Le Bras‡, Ximing Lu‡, Sean Welleck†, Yejin Choi‡,§
†Paul G. Allen School of Computer Science & Engineering, University of Washington
‡Allen Institute for Artificial Intelligence


Distill knowledge from LLM to KG
Demo part 3
ChatGPT for KG
Demo of our tutorial

- Demo part 1: Pipeline of Knowledge Graph Construction
  - [https://colab.research.google.com/drive/1FEmL6hE9UK_Do7dOqrgbKowVaG4WYdUO?usp=sharing](https://colab.research.google.com/drive/1FEmL6hE9UK_Do7dOqrgbKowVaG4WYdUO?usp=sharing)

- Demo part 2: Knowledge-Graph for Question Answering
  - [https://github.com/MrBearWithHisSword/KGQA-COVID](https://github.com/MrBearWithHisSword/KGQA-COVID)

- Demo part 3: ChatGPT for Knowledge Graph
  - [https://colab.research.google.com/drive/1hSGcFQ2QCQSzhBYFfejJuyq0hue1GbYw?usp=sharing](https://colab.research.google.com/drive/1hSGcFQ2QCQSzhBYFfejJuyq0hue1GbYw?usp=sharing)
References

- Fariz Darari. Knowledge graphs mini handbook.
- Marta Maślankowska, Paweł Mielniczuk. Intro to coreference resolution in NLP.
- Sundar V. Entity Linking: A primary NLP task for Information Extraction.
- Xiang Ren. Entity Linking and Coreference Resolution.
- Laura Dietz. Entity Linking.
- https://github.com/tomhartke/knowledge-graph-from-GPT-3
Please cite with the following BibTex entry if these slides are helpful to your research works

@misc{Liu2022KGTutorial,  
  author={Bang Liu and Haochen Shi},  
  title = {Knowledge Graphs: Extracting and Leveraging Meaning from Unstructured Text},  
  howpublished = \url{http://www-labs.iro.umontreal.ca/~liubang/files/KG_tutorial_ivado22.pdf},  
  year = {2022}  
}
Thanks! Q&A

Bang Liu
Email: bang.liu@umontreal.ca
Homepage: http://www-labs.iro.umontreal.ca/~liubang/
Github: https://github.com/BangLiu/