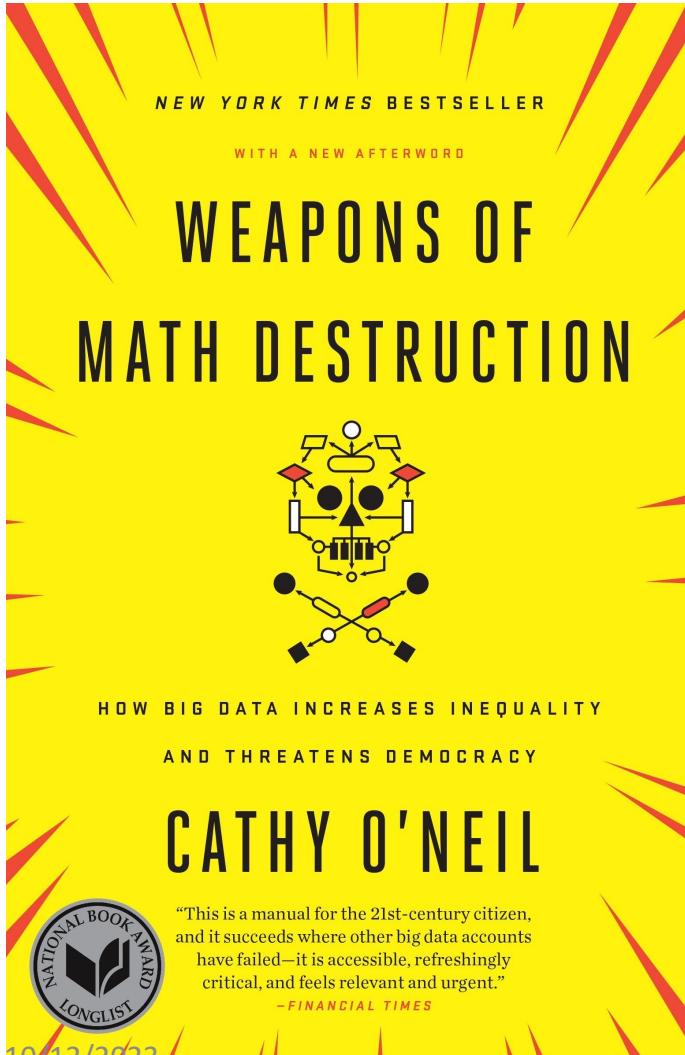


CIKM tutorial

Work in progress

Teaser: History of Irresponsible AI Risk (5 years ago) Product gets canceled



Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via [The Guardian](#) | Source [TayandYou \(Twitter\)](#) | 68 comments

f t SHARE



Microsoft sued for 'racist' application

Microsoft says it fixed the problem -- long before the litigation.



Written by **Matthew Broersma**, Contributor on June 29, 1999

Medium – Knowledge graphs

Introduction

- Knowledge graph (KG) describes objects of interest and connections
- Organizing data as nodes and edges
- Examples
 - Microsoft Satori, Google Knowledge Graph, Amazon Product Graph
 - Knowledge bases (KBs): Yago, Freebase
- Knowledge graph and knowledge base terms are used interchangeably

Introduction - II

- How to identify nodes and derives edges
- So far, most research on KGs/KBs use Wikipedia as source
 - Benefits: easy to read, easy to parse, Wikipedians
 - Drawbacks: coverage, outdated content, bias
- What do we do when there is no Wikipedia?
- ... and there are plenty of examples in the real world
- More specifically
 - KG is a repository of entities, types and relationships
 - KG defines entities, types, attributes, relations, provenance
 - KG is data
 - KG evolves and needs maintenance

KG vs DB

- What's the difference?
- DB
 - Store data for a specific application purpose
 - Semantics understood by those who built it and use it
 - Tables, columns, attributes
- KG
 - All we know about a slice of the world
 - Semantics understood and agreed by all stakeholders
 - Application independent
 - Things, not strings

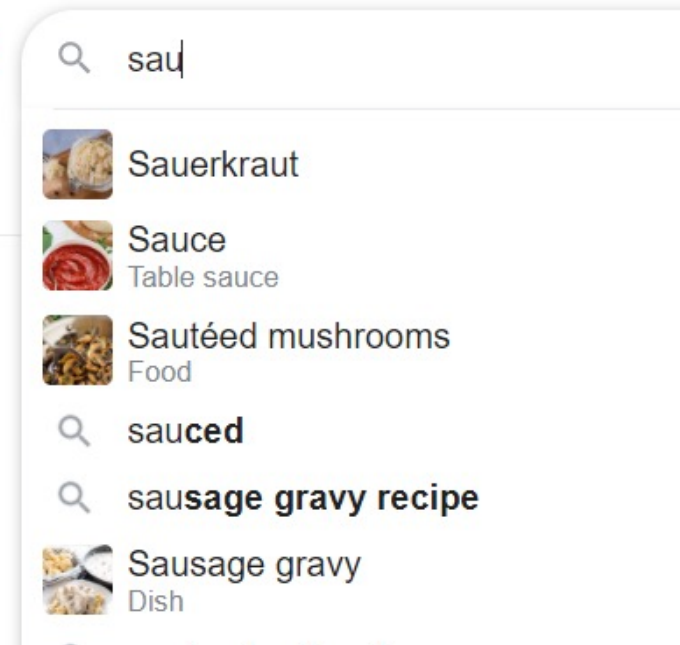
Why we care?

- Machine readable facts about a domain
- Data can be used for different use cases
- Separation of concerns


KGs in action

- Semantic search
 - Going beyond 10-blue links
 - Understanding queries and documents
- Document retrieval
 - Expansion
 - Language modeling
- Entity retrieval
- Recommendations
- Question-answering
- Data cleaning

Example - Autocomplete



Example - Entity cards



Pesto

Pesto, or pesto alla genovese, is a sauce originating in Genoa, the capital city of Liguria, Italy. It traditionally consists of crushed garlic, European pine nuts, coarse salt, basil leaves, and hard cheese such as Parmigiano-Reggiano or Pecorino Sardo, all blended with olive oil. [Wikipedia](#)

Place of origin: [Italy](#)

Main ingredients: Basil, garlic, olive oil, grated hard cheese, pine nuts

Alternative names: Pesto alla genovese

What kind of pasta goes with pesto [View 1+ more](#)



Penne



Fusilli



Cavatappi



Rotini



Linguine

People also search for [View 15+ more](#)



Basil



Pine nut




Pasta



Parmigia...



Bolognese
sauce



Pesto

Sauce

Pesto, or pesto alla genovese, is a sauce originating in Genoa, the capital city of Liguria, Italy. It traditionally consists of crushed garlic, European pine nuts, coarse salt, basil leaves, and hard cheese such as Parmigiano-Reggiano or Pecorino Sardo, all blended with olive oil.

[Wikipedia](#)

Main ingredients: Basil, garlic, olive oil, grated hard cheese, pine nuts

Place of origin: [Italy](#)

Course: Sauce

People also search for [See all \(20+\)](#)



Basil



Chimichurri



Bolognese
sauce



Italian food






Carbonara


Example - answers


does tiramisu have alcohol


✕





 All

 Images

 Shopping

 Videos

 News

 More

Settings

Tools

About 11,000,000 results (0.56 seconds)



View all

DOES TIRAMISU CONTAIN ALCOHOL? Traditionally, **tiramisu** is made with Marsala wine in the filling, and the ladyfingers are soaked in a boozy coffee mixture. ... If you enjoy a boozy treat once in a while, you can use any kind of liqueur that complements coffee well!




Apr 12, 2018

bakingamoment.com › classic-tiramisu-recipe


Classic Tiramisu Recipe: fluffy, rich, & irresistible! -Baking a ...

 About featured snippets •  Feedback

People also ask

- Does tiramisu contain alcohol? 
- Why does tiramisu have alcohol? 
- Does Costco tiramisu have alcohol? 

does tiramisu have alcohol



Sign in

ALL WORK IMAGES VIDEOS MAPS NEWS SHOPPING





3,050,000 Results Any time

Tiramisu is an Italian dessert made with ladyfingers, **mascarpone**, eggs, cream, sugar, coffee and cocoa powder. This sweet treat may also contain **alcohol** in some cases, although this ingredient is not required to make this dessert.

[What Kind of Liquor Is in Tiramisu? | LEAFtv](#)
 www.leaf.tv/articles/what-kind-of-liquor-is-in-tiramisu/

Was this helpful?  



- PEOPLE ALSO ASK
- Does Tiramisu contain alcohol? 
- What kind of liquor is in Tiramisu? 
- What do you drink with Tiramisu? 
- Is there caffeine in my Tiramisu? 

Feedback

[What Kind of Liquor Is in Tiramisu? | LEAFtv](#)

Tiramisu

Dessert



Tiramisu is a coffee-flavoured Italian dessert. It is made of ladyfingers dipped in coffee, layered with a whipped mixture of eggs, sugar, and mascarpone cheese, flavoured with cocoa. The recipe has been adapted into many varieties of cakes and other ... +

 Wikipedia

Main ingredients: Savoiardi, egg yolks, mascarpone, cocoa, coffee

Serving temperature: Cold

Place of origin: Italy

Course: Dessert

People also search for

See all (20+)



Main concepts

- Entity
 - Object or concept in the real world that can be identified
 - Uniquely characterized by its name(s), type(s), attributes, and relationships to other entities
- Named-entity
 - Specific entity for which one or many designators or proper names can be used to refer to it
 - Examples: Red Cross (Organization), California (Location), February (Date)
- Unique identifier
 - An identifier for an entity is a string of chars that uniquely denotes this entity
 - one-to-one correspondence between each entity identifier (ID) and the object it represents

Main concepts - II

- Types
 - Entities may be categorized into multiple entity types
 - Types can also be thought of as containers that group together entities with similar properties
- Ontology
 - The process of describing the kinds, properties of and relationships on things in the world
 - Use the tools of logic to formalize this description
- Taxonomies
 - Taxonomy is a directed acyclic graph, where the nodes are classes and there is an edge from class X to class Y

Main concepts - III

- Names
 - Multiple entities may share the same name
 - These alternative names are called surface forms or aliases
- Attributes
 - Different types of entities are typically characterized by different sets of attributes
- Relationships
 - Relationships describe how two entities are associated to each other

Relationships and attributes

- SPO

- Subject-Predicate-Object

- Example

<Tom Brady, place of birth, San Mateo>

<Tom Brady, member of sports team, Tampa Bay>

<Tom Brady, occupation, American football player>

<fettuccine, subclass of, pasta>

<fusilli, subclass of, pasta>

<linguine, subclass of, pasta>

<paella, country of origin, Spain>

<paella, has ingredient, chicken>

<paella, has ingredient, rice>

Data models

- Direct edge-labeled graphs
 - RDF is an example
- Graph dataset
 - Set of named graphs. Each named graph is a pair (graph id, graph)
- Property graphs
 - Allows a set of (property, value) pairs and a label to be associated with nodes/edges
 - Common in graph databases

Data access

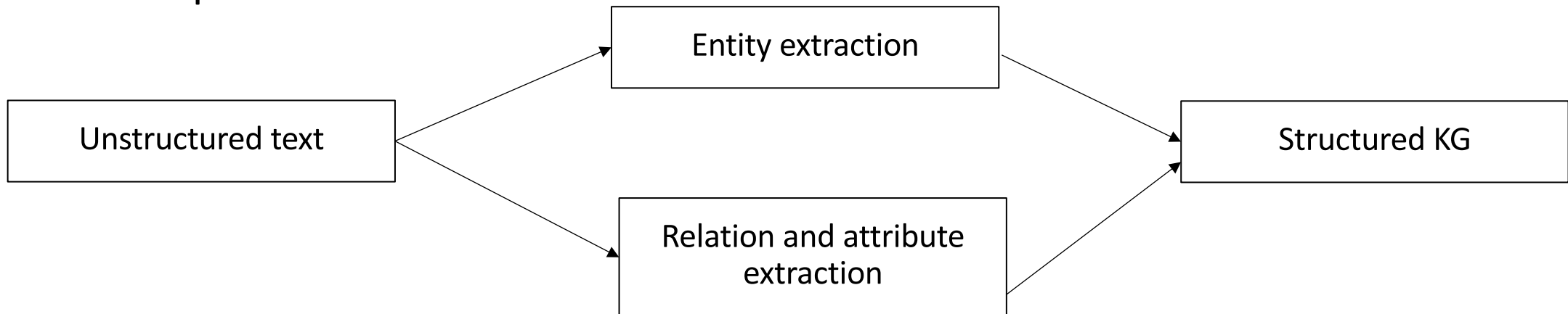
- Querying
 - SPARQL
 - SQL
- Raw data
- Materialization
 - Publish high quality data
- Search & Browse UI

Information needs

- Keyword queries
 - Free text queries
- Structured queries
 - SQL, SPARQL
- Keyword++
 - Queries with filters or facets
- Natural language
 - Natural language queries, questions
- Zero queries
 - You are the query

How do we build one

- Manually
 - Experts, crowdsourcing
- Automation
 - Generated from input sources
 - Information extraction
 - Representation



Many choices

- Semantic Web-like
 - RDF, OWL, SPARQL
- Key-values
 - Text files, JSON
- RDBMS
 - Tables, columns, SQL
- Hybrid
 - Your favorite combination

Input sources

- Data
 - Wikipedia, catalogs, web pages, query logs, etc.
- Importance of top-tier sources
 - Authoritative content, high coverage, clean representation
 - Domain specific
- Pre-existing categorization
 - Potentially useful
 - Alignment

Entity discovery

- NER detects mentions of entities and assigns types
 - Part of your typical NLP toolkit (e.g., NLTK, Stanza, GATE)
 - Popular types: People, places, organizations, date, etc.
- Dictionaries
 - Abbreviations (Apple, Apple Inc.)
 - Acronyms (MS, Microsoft)
 - Stage name (Lady Gaga, Stefani Joanne Angelina Germanotta)
- Pattern-based methods
 - Hearst patterns; co-occurrence
 - “such as”, “X like Y”, “X and other Y”, “X including Y”

Entity discovery - II

- ML
 - CRF
 - LSTM models
- Embeddings
 - Word embeddings are computed from co-occurrences and neighborhoods of words in large corpora
 - Words are highly related if they are used in similar context
- Taxonomies from catalogs, networks and user behavior
 - Wikipedia categories
 - Tagging systems
 - Query logs and clicks

Entity linking

- Recognizing entity mentions in text and linking them to the corresponding entries in a KG
 - Assume a KG with existing entities
- What's the difference between NER and EL?
 - NER recognizes entities and assigns an entity type
 - EL recognizes entities and assigns an entity id
 - Importance of having an identifier management system

Entity linking- example

Tom Brady to Tampa Bay Buccaneers before Super Bowl ...
<https://www.nbcsports.com/washington/football-team/...>

Tom Brady (player)
Tom Brady (director)

Tampa (Florida)
Tampa (Kansas)

Tampa Bay (Location)
Tampa Bay (Buccaneers - NFL)
Tampa Bay (Lightning - NHL)
Tampa Bay (Ray - MLB)

Super Bowl 2021
Super Bowl 2020
...

Tampa Bay Buccaneers (NFL team)

Entity linking – components

- Mention detection
 - Identification of text snippets that can potentially be linked to entities
 - Using dictionary of entity names and variations
- Candidate selection
 - Ranked list of candidate entities is generated for each mention
- Disambiguation
 - The best entity (or none) is selected for each mention using context (if available)
 - Ranking problem
- Entity annotations

Entity matching

- Compute equivalence class
 - Also known as duplicate detection or record linkage
- Name similarity
 - String similarity
- Context similarity
 - Importance of proximity
- Mention-entity popularity
 - Popularity of an entity
- NEMO (Named Entities Made Obvious)
 - The best evidence for entity disambiguation is provided by the set of co-occurring entities

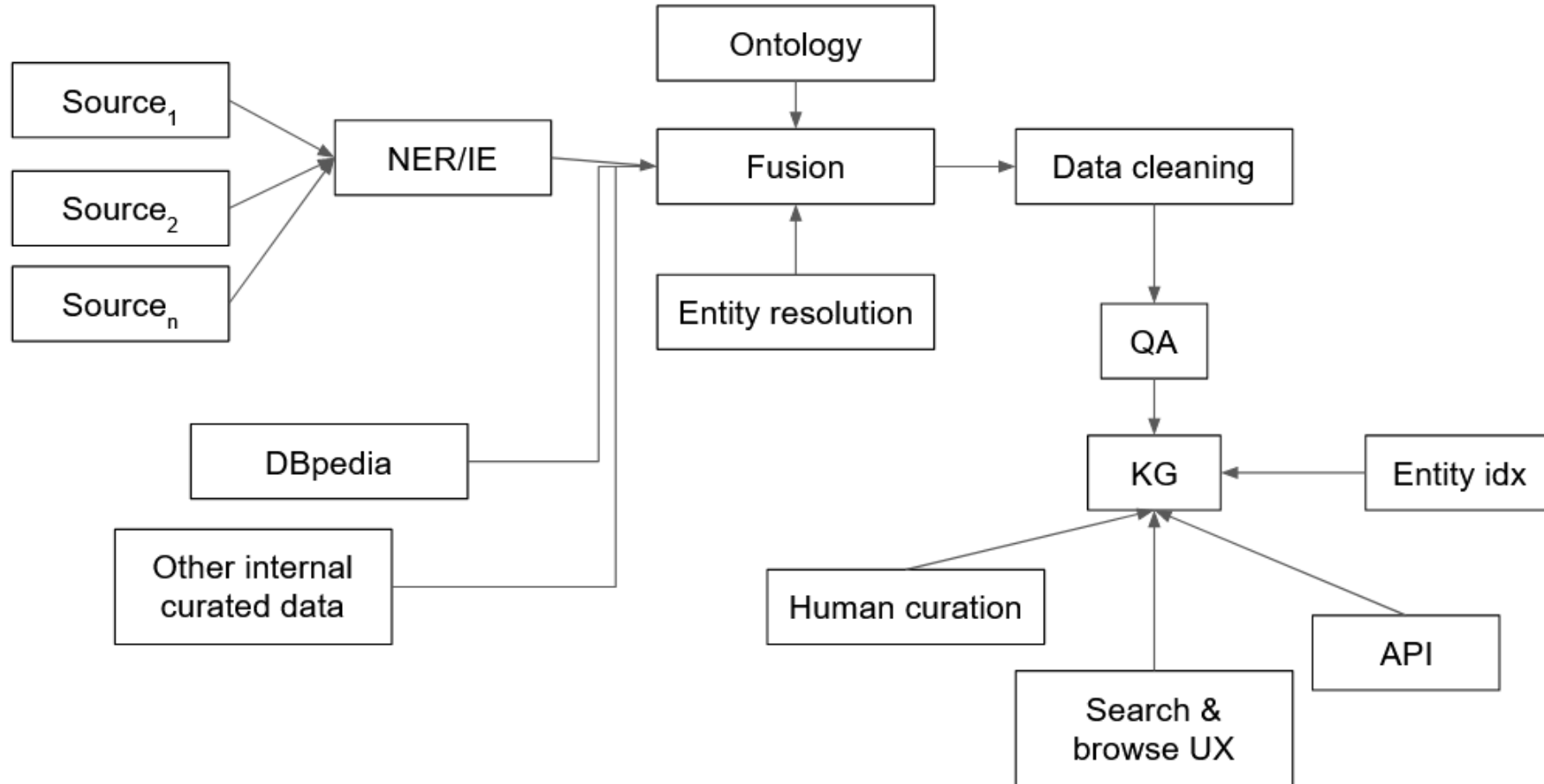
Attributes and relationships

- Pattern-based
 - Regex
 - Rule-base extraction
- Extraction from semi-structured content
 - DOM trees
 - Web tables
- Information extraction
 - Extract (entity-type, alpha, entity-type)
- Infoboxes are great if you have them

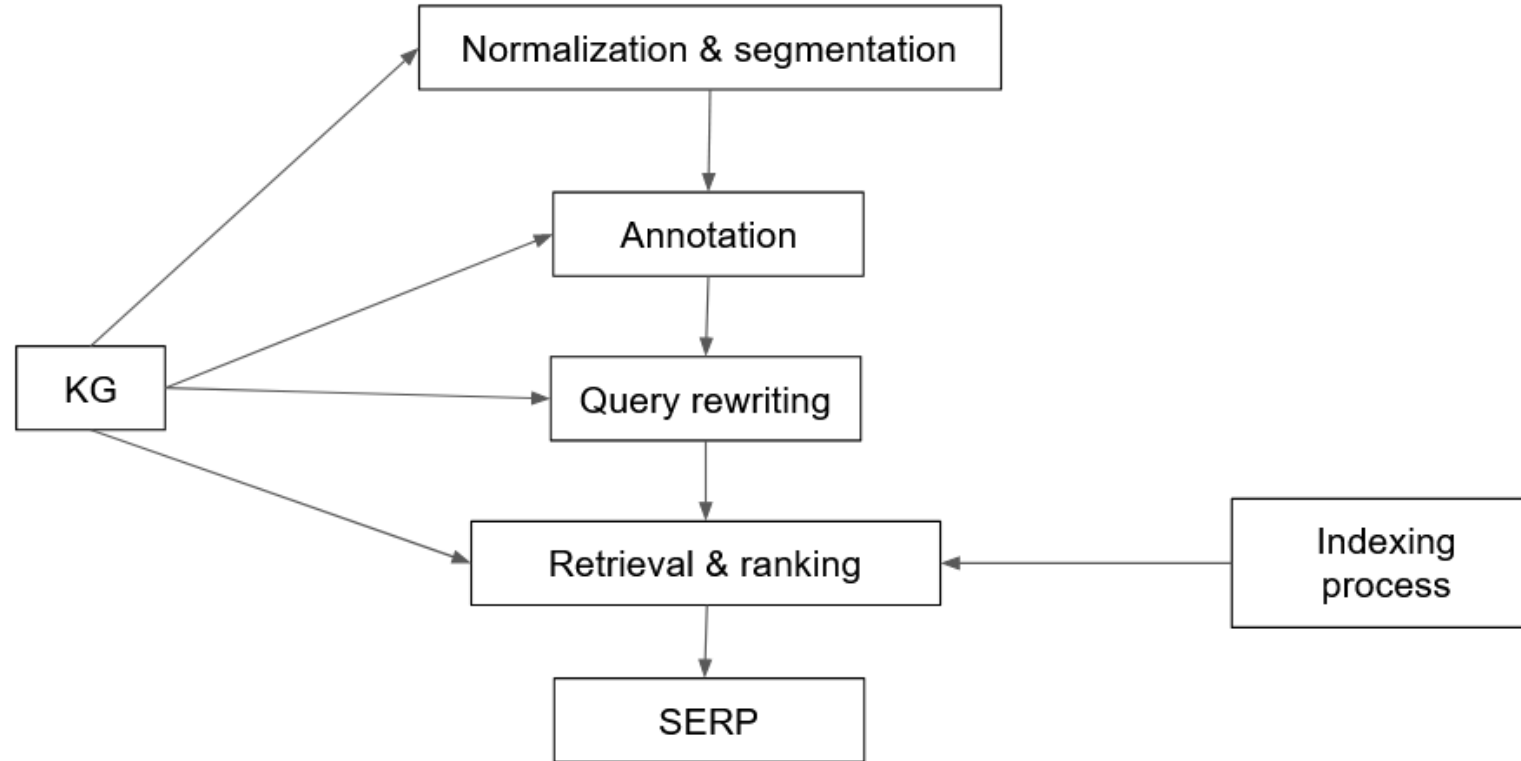
KG Curation

- Quality
 - Precision, recall
- Crowdsourcing
 - Human in the loop
- KG life cycle
 - Provenance metadata (source, timestamp, extraction method)
 - Versioning
 - Twitter CEO (Jack -> Parag -> Elon)
- Maintenance

Architecture for KG construction



Where is the KG?



Representation

- Unique problem
 - Entities in KG have no textual representation, apart from their names
 - We can run SPARQL queries but how do we add the IR part?
- Predicate folding
 - Build a textual representation for each entity by considering all triples
 - Grouping predicates together into a small set of predefined categories
 - From SPOs triples to a structured document

Predicate folding - example

<spaghetti carbonara, instance_of, recipe>
<spaghetti carbonara, has_ingredient, spaghetti>
<spaghetti carbonara, has_ingredient, pancetta>
<spaghetti carbonara, has_ingredient, eggs>
<spaghetti carbonara, has_ingredient, parmesan>
<spaghetti carbonara, recipe_cuisine, italian cuisine>
<spaghetti carbonara, serving_size, 4>
<spaghetti carbonara, calories, 510>
<spaghetti carbonara, cook_time, 25min>

Name	spaghetti carbonara
Ingredients	Spaghetti, pancetta, eggs, parmesan
Attributes	italian cuisine, serves 4, calories 510, cook time 25min
Related entities	spaghetti aglio e olio, fettuccine alfredo

Entity retrieval

- Field search retrieval
- Linear combination of matching functions
- Can use LTR to learn weights

$$\text{score} = w_1 * \text{match}(f_1, q) + w_2 * \text{match}(f_2, q) + \dots + w_i * \text{match}(f_i, q)$$

Document retrieval

- Preprocessing
 - Documents are preprocessed with EL + additional information obtained from KG
- Query annotation
 - Query processed with EL
- Expansion
 - KG feedback: query is issued against an index of a KG in order to retrieve related entities
 - Corpus-based feedback

Semantic search

- Understanding information needs
- Query classification
 - Assign a query to one or multiple pre-defined categories
 - Query intent classification (Broder)
- Query annotation
 - Generate semantic markup for a query
 - Query segmentation: group terms into phrases
 - Query tagging (POS, NER)

EL in queries

- Problems
 - Queries are very short
 - Limited context, or none at all
 - Online process under time constraints
- Components
 - Mention detection
 - Candidate ranking
 - Interpretations: query may have more than 1 interpretation

Using entities for search

- Query assistance
 - Auto-complete
 - Specific subset of queries that can be decomposed into entity and refiner components
 - Query recommendations
- Entity cards
 - Summaries and facts
- Entity recommendation
 - Entity based
 - Query based
 - Explanations

Domain specific KGs: healthcare domain

- Scientific medical knowledge
- Existing taxonomies and data sources
- Examples
 - SNOMED (Systematized Nomenclature of Medicine)
 - RxNorm (medications available on the US market)
 - MeSH (Medical Subject Headings)

Challenges

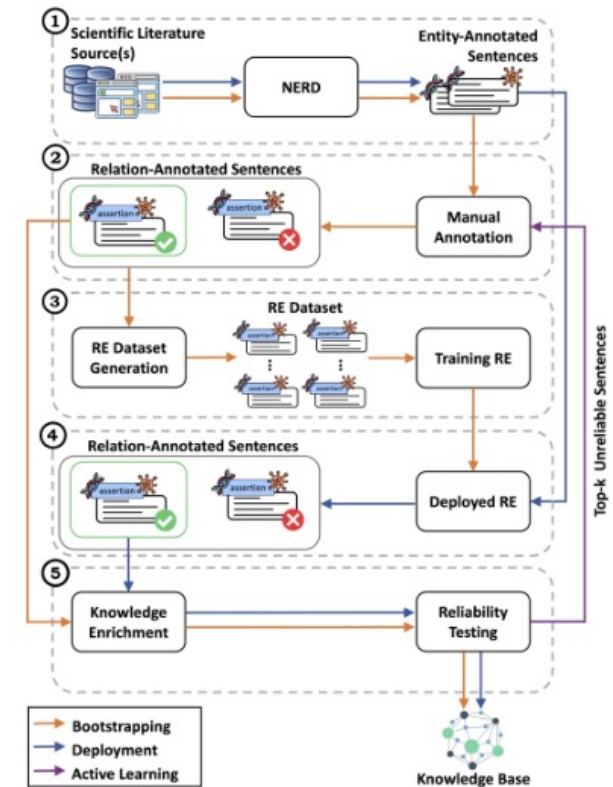
- Very sensitive data
- EMR (Electronic Medical Records)
- Clinical relevance
- Vocabulary mismatch
 - Patient describing a symptom
 - MDs describing a diagnosis
- Data labeling and curation

How to bootstrap?

- No single approach to build a KG
- Research and engineering problems
- Iterative development cycle
- Content
- Infrastructure
- Applications

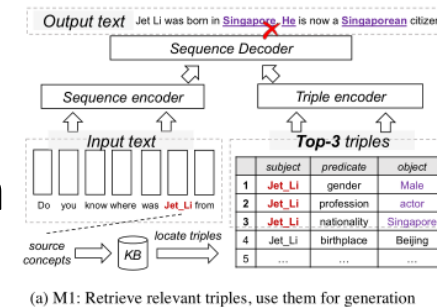
CORE (Collaborative Oriented Relation Extraction)

- KG generation system
- Combination of automated ML-based methods and domain experts
- Modular architecture that can be easily modified
- Reliability tests
- Active learning process make the system suited to iterative
- Versioning

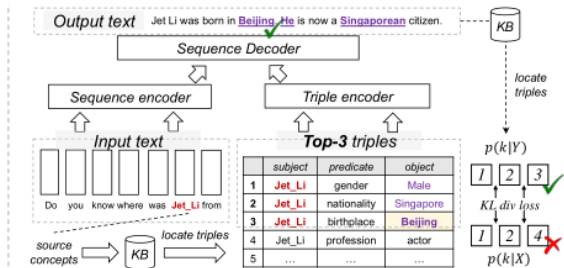


Knowledge-enhanced generation with KGs

- Design Supervised Tasks around KG
 - Discover the dependencies of elements within a sequence
 - Retrieve relevant triples, then using them for generation
 - Using KL to measure the proximity between prior and posterior distribution
- Selecting KG or facts in a KG



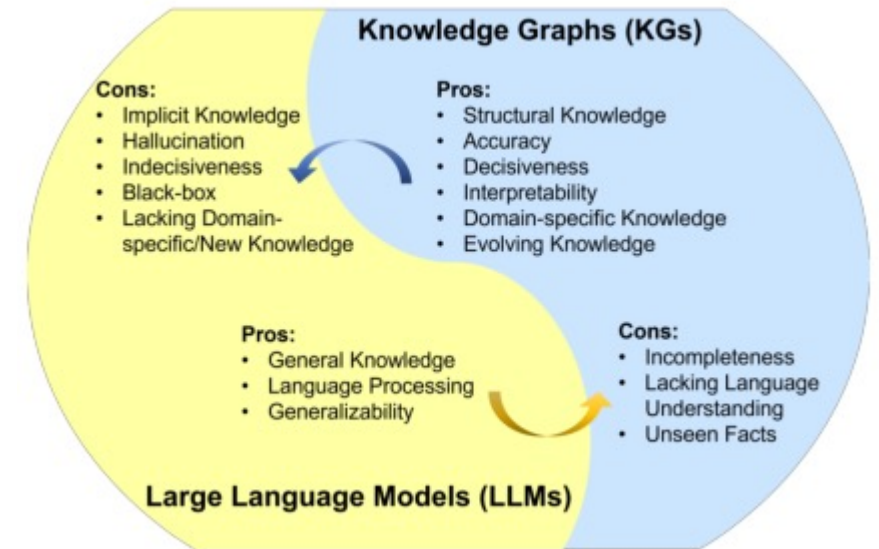
(a) M1: Retrieve relevant triples, use them for generation



(b) M2: Use KL to measure the proximity between prior and posterior

Combining KGs & LLMs

- LLMs lack of factual knowledge
- LLMs memorize and knowledge in a training set
- No interpretability
- KGs are difficult to construct and maintain
- KGs are domain specific



KG-enhanced LLMs

- KG-enhanced LLM pre-training
 - Training objective
 - LLM inputs
 - Fusion models
- KG-enhanced LLM inference
 - Dynamic fusion
 - Retrieval augmented
- KG-enhanced LLM interpretability
 - Probing and analysis

LLM-augmented KGs

- Embedding
 - Text encoders
 - Joint text and KG embeddings
- Completion
- Construction
 - Entity discovery
 - Relation extraction
 - Coreference
 - Distilling KGs from LLMs
- KG-to-text generation
- LLM-augmented KG question answering