Medium

- Query processing
- Aggregated search
- Knowledge graphs

Query processing

- The front door of the search engine
- In a nutshell
  - Query intent correctly identified and re-written in internal representation
  - Use re-written query to retrieve and match relevant results
  - Use all info to generate SERP
- Lots of annotations and classifiers are used in this step
- NLP components (POS, NER, stemming, keyphrases, geotagging, etc.)
- Performance and matching (not covered in this tutorial)

Query intent

- Broder’s taxonomy
- Web search evolves
- New users and behaviors
- Emerging new types of search intents

Question intent

- Question taxonomy compared to queries
- Multi-faceted question taxonomy
- Detailed methodology
- Agreement analysis and tie-breaking

Type	Examples
Advice	how can I be successful in life? how should I invest my salary?
Attribute	what is pristine edge's real name what is senegal's official language
Calculation	4,146.70+700+11900 1/2 cups in tbsp
Description	what is propylene kit what is oracle vpd functionality
Entity	who replaced ted kennedy in the senate who produced transformers
Language	what is puppy in swahili what is the common name for jade
List	types of aircraft southampton to guernsey types ant poison
Location	where are protists most abundant in humans what is oklahoma's absolute location
Opinion	is donald trump a good president? is ronaldo or messi a better player?
Process	what is needed to get home insurance how to check warranty of sd card sandisk
Quantity	how long is csus transfer orientation cost of an ice cream truck
Reason	why do knees swell up why do lipomas grow back
Resource	python temperature converter code tum mile love reprise lyrics english
Temporal	when do the oscar awards start when does daylight saving time return?
Verification	is tomorrow Monday? is donald trump the 34th president?
Weather	5 day weather forecast for york tybee island weather in march

B. Cambazoglu "An Intent Taxonomy for Questions Asked in Web Search", CHIIR 2021

NFQ intent taxonomy

- Very little understanding of Non-Factoid questions
- Lots of research on factoid QA (SQuAD, MS MARCO) not on NFQ
- Detailed taxonomy creation methodology
- Data set and model available https://github.com/Lurunchik/NF-CATS

Category	Description	Expected Answer Structure	Patterns
INSTRUCTION	You want to understand the **procedure/method** of doing/achieving something.	Instructions/guidelines provided in a step-by-step manner.	How to ...? How can I do ...? What is the process for ...? What is the best way to ...?
REASON	You want to find out **reasons** of/for something.	A list of reasons with evidence.	Why does ...? What is the reason for ...? What causes ...? How come ... happened?
EVIDENCE-BASED	You want to learn about the **features/description/definition** of a concept/idea/object/event.	Wikipedia-like passage describing/defining an event/object or its properties based only on facts.	What is ...? How does/do ... work? What are the properties of ...? What is the meaning of ...? How do you describe ...?
COMPARISON	You want to **compare/contrast** two or more things, understand their differences/similarities.	A list of key differences and/or similarities of something compared to another thing.	How is X ... to/from Y? What are the ... of X over Y? How does X ... against Y?
EXPERIENCE	You want to **get advice** or **recommendations** on a particular topic.	Advantages, disadvantages, and main features of an entity (product, event, person, etc) summarised from personal experiences.	Would you recommend ...? How do you like ...? What do you think about ...? Should I ...?
DEBATE	You want to **debate on a hypothetical question** (is someone right or wrong, is some event perceived positively or negatively?).	Arguments on a debatable topic consisting of different opinions on something supported or weakened by pros and cons of the topic in the question.	Does ... exist? Can ... be successful? Do you think ... are ...? Is ... really a ...?

V. Bolotova et al. “A Non-Factoid Question-Answering Taxonomy” SIGIR 2022

Problems

- Taxonomies and categorization mechanisms
- Better understanding of user intent
- Domain specific cases
- Subjective queries and questions
- Near-duplicate questions

Aggregated search

- Integrated search across different sources
- Common presentation
- Content is retrieved from verticals for a given query
- What is a vertical?
  - A specialized search service that focuses on a specific type of information need
  - Examples: maps, news, images, etc.
- Aggregated search system decides what information to display and where

Google

when the levee breaks

All Videos News Images Books More Tools

About 1,860,000 results (0.46 seconds)

When the Levee Breaks

Song by Led Zeppelin

Overview Lyrics Videos Listen Other recordings Artists Composers

Listen

Spotify YouTube Music Apple Music iHeart

About

Artist: Led Zeppelin

Album: Led Zeppelin IV

Released: 1971

Microsoft Bing

tom brady

Sign in 35

ALL NEWS IMAGES VIDEOS MAPS SHOPPING MORE

Add Bing Firefox extension

On this page:

News

Videos

Related Search

187,000,000 Results Any time Open links in new tab

News about Tom Brady

bing.com/news

Ryan Fitzpatrick Responds To Tom Brady's 'Motherf----- Theory'

Ryan Fitzpatrick says Tom Brady showed him 'zero respect'

Former NFL quarterback Ryan Fitzpatrick isn't a big fan of Tom Brady. He says the ex-New …

NBC Sports · 12m

Ryan Fitzpatrick Thinks He Was 'That Motherf--ker' Tom Brady Referenced on 'The…

Ryan Fitzpatrick Responds to Tom Brady 'That Motherf-----' QB Theory

Thomas Edward Patrick Brady Jr. is an American football quarterback for the Tampa Bay Buccaneers of the National Football League. He spent his first 20 seasons with the New England Patriots, where … +

Wikipedia

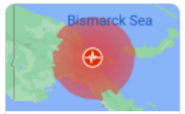
Google

earthquake

All News Images Videos Books More Tools

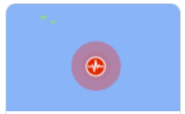
About 404,000,000 results (0.61 seconds)

Recent earthquakes



M 7.6
48 miles from Lae, Papua New Guinea
Sep 10, 4:46 PM

M 7.0
129 miles from Vanuatu
Sep 14, 4:04 AM



M 5.8
245 miles from Samoa Islands
Sep 13, 10:08 PM

M 4.4
2 miles from Santa Rosa, CA
Sep 13, 6:39 PM

Note: Official reports of new earthquakes may be delayed
All times are in Pacific Time · Source: U.S. Geological Survey

Feedback

Google

pesto

All Shopping Images Videos Maps More Tools

About 178,000,000 results (0.69 seconds)

Recipes

Basic Pesto
NYT Cooking - The New York…
5.0 ★★★★★ (5.8K)
15 min
Pine nuts, parmesan cheese, olive oil, garlic, fresh basil

Fresh Basil Pesto
Simply Recipes
4.9 ★★★★★ (237)
15 min
Pine nuts, baby spinach, olive oil, black pepper

Basil Pesto
Love and Lemons
5.0 ★★★★★ (173)
5 min
Pine nuts, parmesan cheese, olive oil, lemon juice, garlic

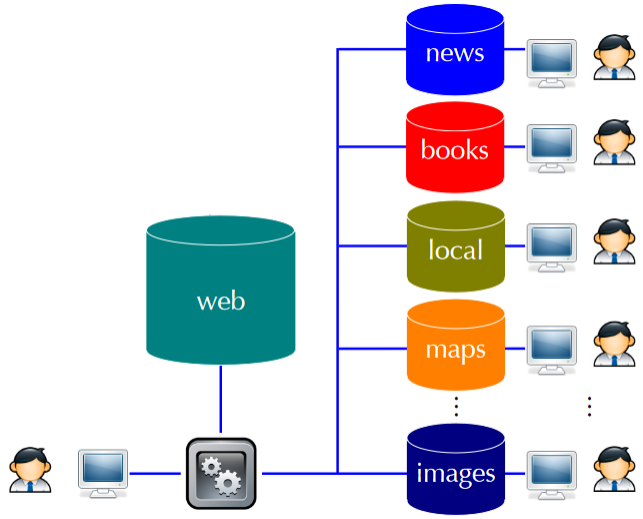
Show more

Pesto

Food

Aggregated search task

- Combine results from multiple verticals into a single representation
- Goal is to satisfy user with aggregated results
- Better than to search on each vertical separately
- Motivation:
  - Single point of entry
  - User may not know that a particular vertical exists
  - Results from different verticals at once



J. Arguello, Aggregated Search. Found. Trends Inf. Retr. 10(5), 2017

Sources of evidence

- Predict which vertical is relevant for a given query
- Predict where to present the information
- Given a particular context, predict which vertical to present and where on the page
- Context = core content, information need, user profile

Types of features

- Query features
  - Properties of the input query
  - Independent of the vertical
  - Example: query contains “news”
- Vertical features
  - Properties of the vertical
  - Independent of the query
  - Example: number of queries issued to the vertical
- Query-vertical features
  - Properties of the vertical and query
  - Example: number of results from vertical to a query

Pre-retrieval features

- Query keywords: presence of certain keywords (e.g., news, shopping)
- Query characteristics: NER
- Query class: navigational, transactional, informational
- Query category: affinity to pre-defined topical categories
- Co-occurrence with vertical triggers: query terms and trigger terms
- Vertical corpus: similarity between query and sampled vertical results
- Vertical query log: similarity between query and vertical query traffic
- Query-vertical CTR
- Vertical CTR

Post-retrieval features

- Top vertical similarity: pairwise similarity between top vertical results
- Recency
- Vertical scores
- Clarity: divergence of top results from a background language model
- Metadata (e.g., reviews, location, publication date, etc.)

Post-presentation features

- CTR
- Average dwell time
- Other behavioral signals like mouse movement, scrolling, etc.

Vertical selection

- Decide which verticals to present in response to a query
- Classification task
- Single evidence
  - Compute a measure for each candidate vertical and predict
  - Comparable across verticals
- Multiple evidence
  - Use ML to combine multiple sources
- Two main problems when combining evidence
  - Certain features may not be available for some verticals
  - Learn a vertical-specific relationship between certain features and the relevance of a particular vertical
- Adaptivity

Vertical presentation

- Where to present relative to web results
- Whole page layout constraints
- Block or answer is sequence of results that are presented together
- Pointwise: predict the relevance of each block to the query
  - Present at specific slots within web results
  - Simple and intuitive
- Pairwise: predict the relevance between each candidate pairs
  - Voting
- LTR
  - Requires common feature representation

Composite retrieval

- Aggregated search anchored around entities
- Alternative to a list of results
- Bundles
  - Results from verticals are grouped into bundles
  - Goal is to help with exploratory search tasks
  - Criteria (relevance, cohesion, diversity)
- Techniques for composite
- Knowledge graph assumption
  - KG contains entities and relationships
  - Query and traverse the KG to extract relevant elements
- Query to categories: bundle around the best category
- Result categories: bundling based on categories shared by the results

Events

Tue, Sep 13 2:00 PM	Kupuna Cards Kona West Hawai'i Civic Center
Tue, Sep 13 5:00 PM	TACO TUESDAYS COMEDY! Kona Elks Lodge
Tue, Sep 13 7:00 PM	HILO - Trivia Night Ola Brew Co
Wed, Sep 14	Hulihe'e Palace – 2:30 pm Do… Hulihe'e Palace

View 25+ more

People also search for View 10+ more

Problems

- Research
  - Quantify utility of new and existing verticals/data sources
  - Algorithms and ML techniques to generate automatically aggregations
  - Relevance models
  - Optimizations for different configurations
  - Slotting
- Engineering
  - Data flow architecture
  - Extract and publish the aggregations/bundles
  - Refresh content/keywords
- What’s the best CX?
  - Layout
  - Relevance cues

Knowledge graphs

- 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
- Similar ideas have been around for quite some time
  - Ontologies and common sense knowledge
  - Cyc

What is it?

- 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
- Why we care?
  - Machine readable facts about a domain
  - Data can be used for different use cases
- 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

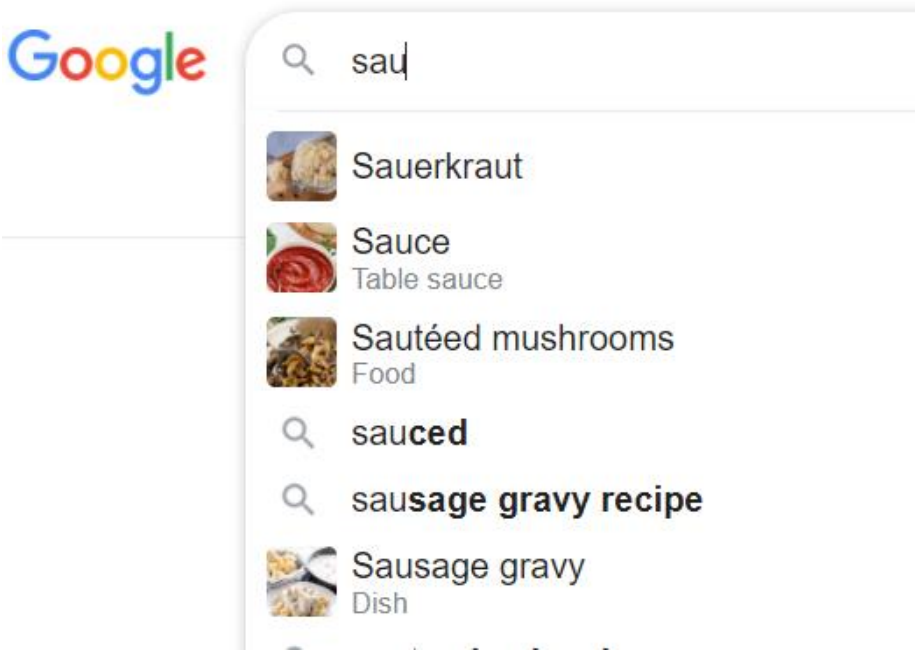
KG vs DB

- 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

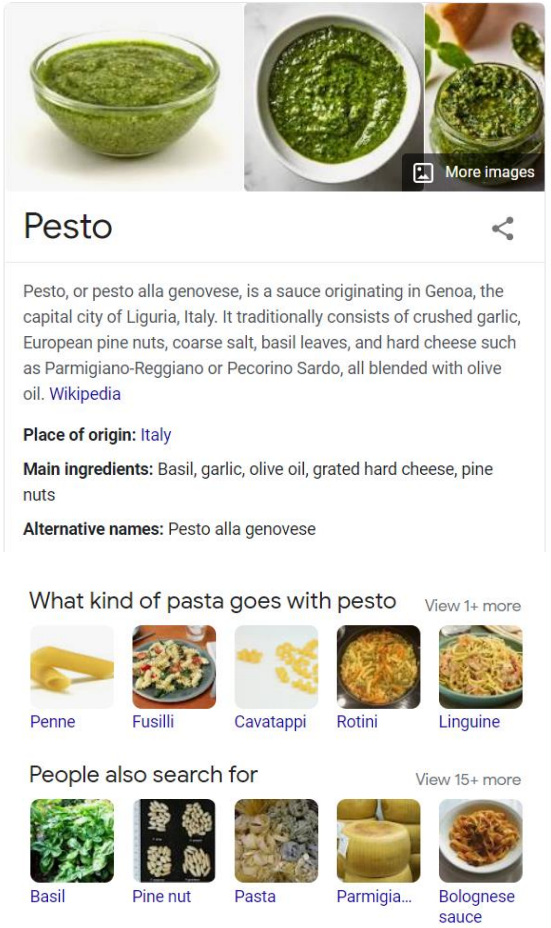
KGs in action

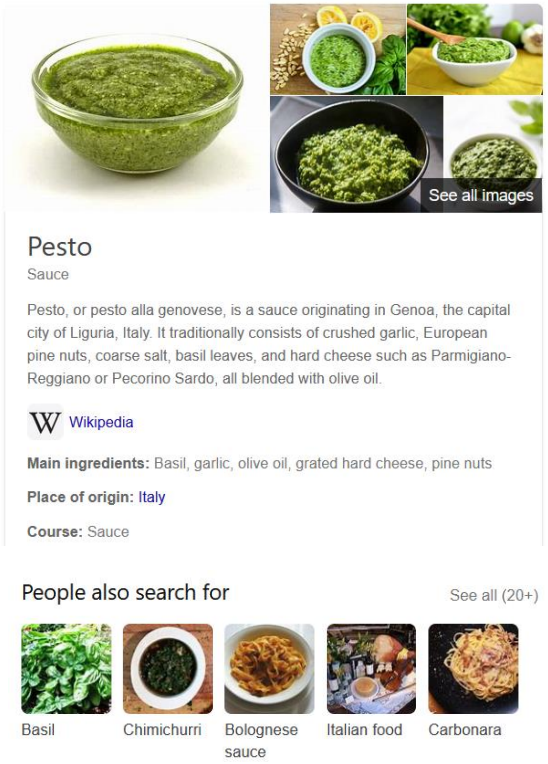
- Semantic search
  - Going beyond 10-blue links
  - Understanding queries and documents
- Document retrieval
  - Expansion
  - Language modeling
- Entity retrieval
- Recommendations
- Question-answering
- Data cleaning

Example - Autocomplete



Example - Entity cards





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

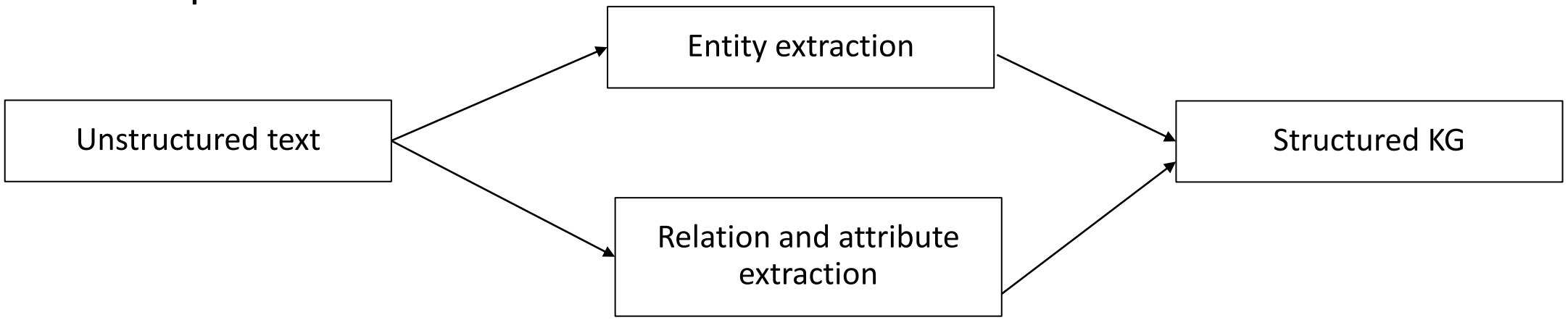
- Direct edge-labeled graphs
  - RDF is an example
- Graph dataset
  - Set of named graphs. Each named graph is a pair (graph id, graph)
- Property graphs
  - Allows a set of (property, value) pairs and a label to be associated with nodes/edges
  - Common in graph databases

Data access

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

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”

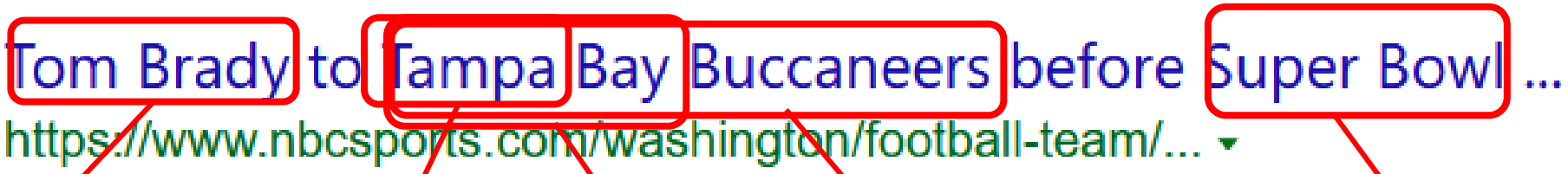
Entity discovery - II

- 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 (player)

Tom Brady (director)

Tampa (Florida)

Tampa (Kansas)

Tampa Bay (Location)

Tampa Bay (Buccaneers - NFL)

Tampa Bay (Lightning - NHL)

Tampa Bay (Ray - MLB)

Tampa Bay Buccaneers (NFL team)

Super Bowl 2021

Super Bowl 2020

…

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

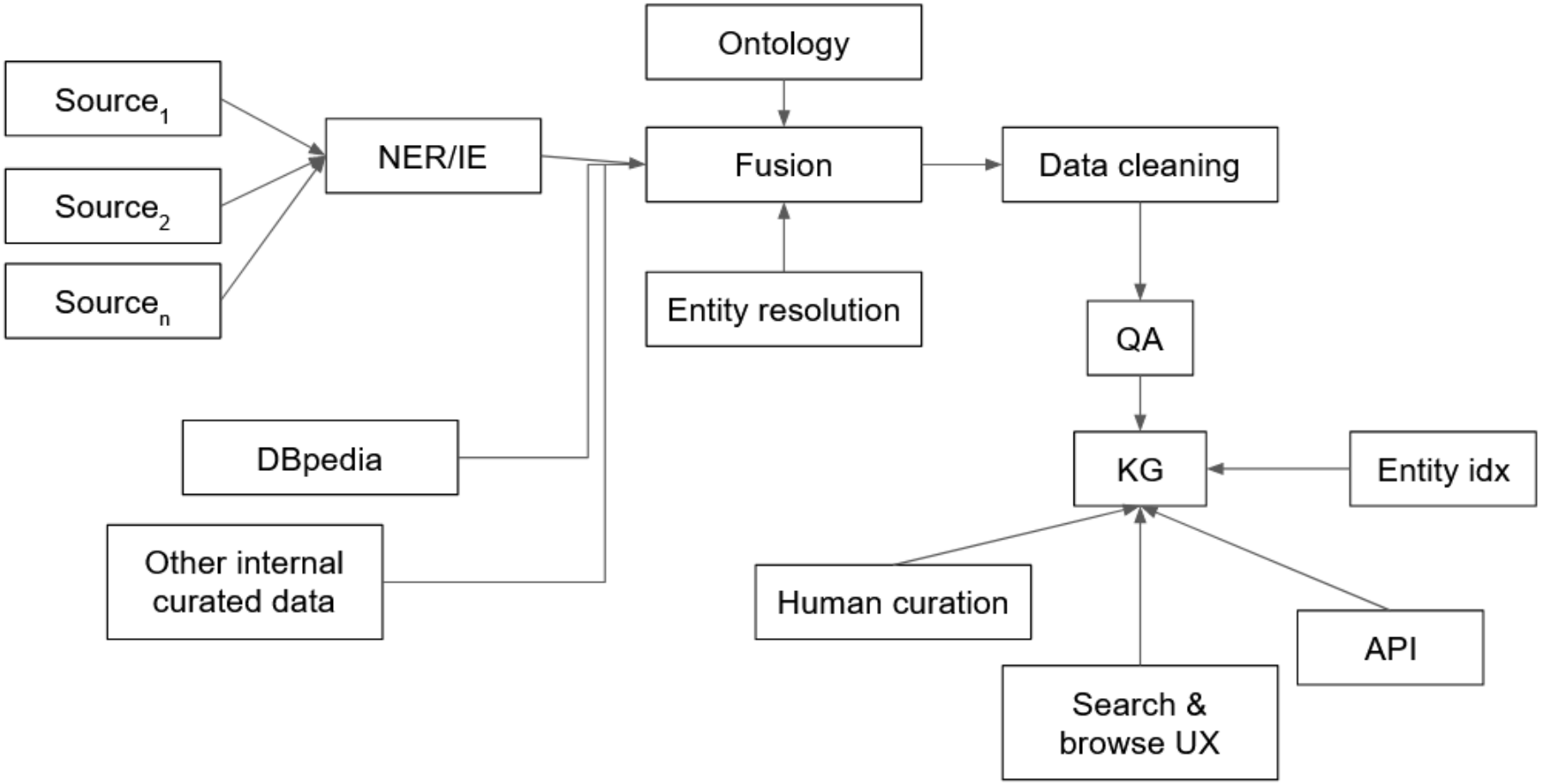
Attributes and relationships

- 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

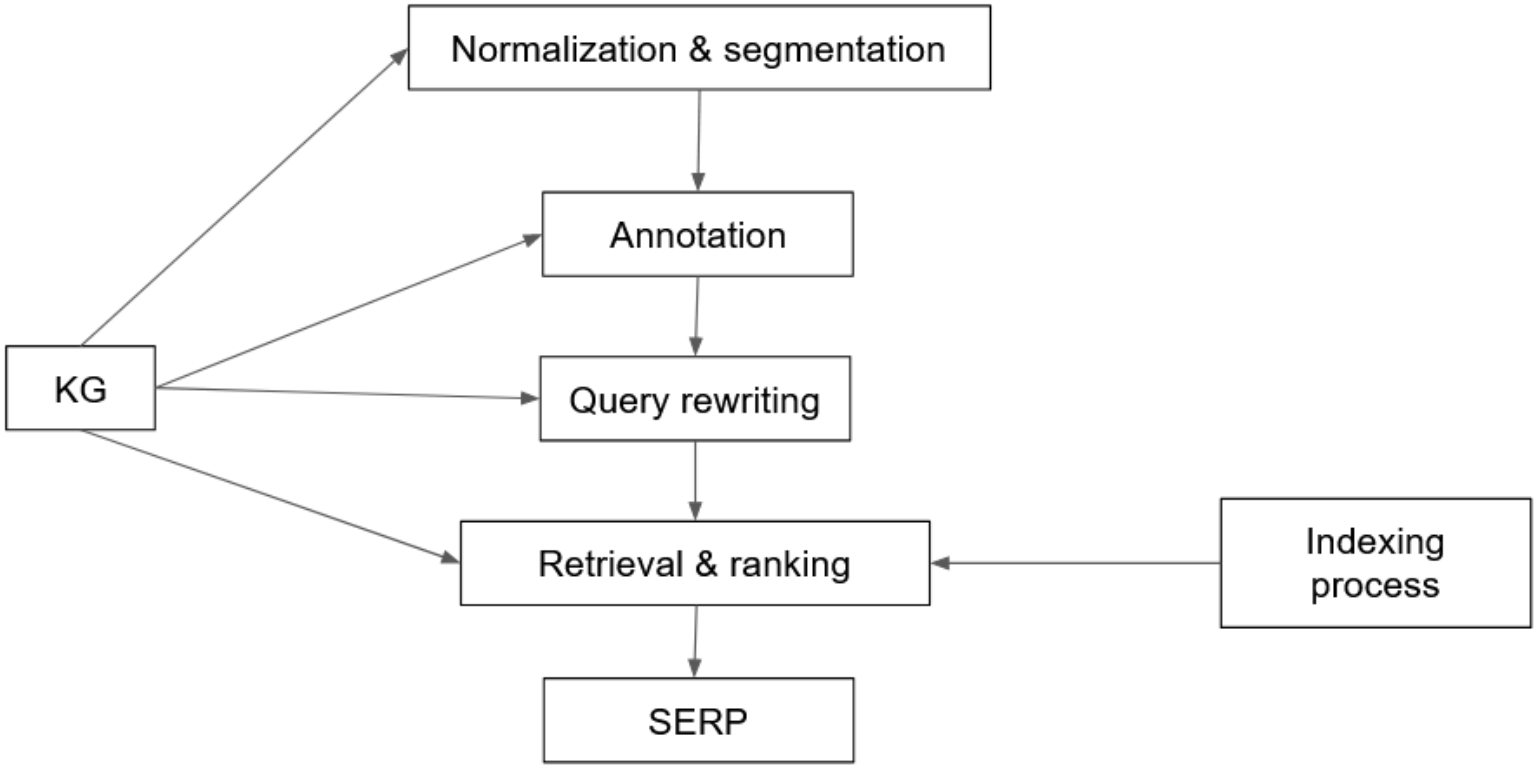
KG Curation

- 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

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

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

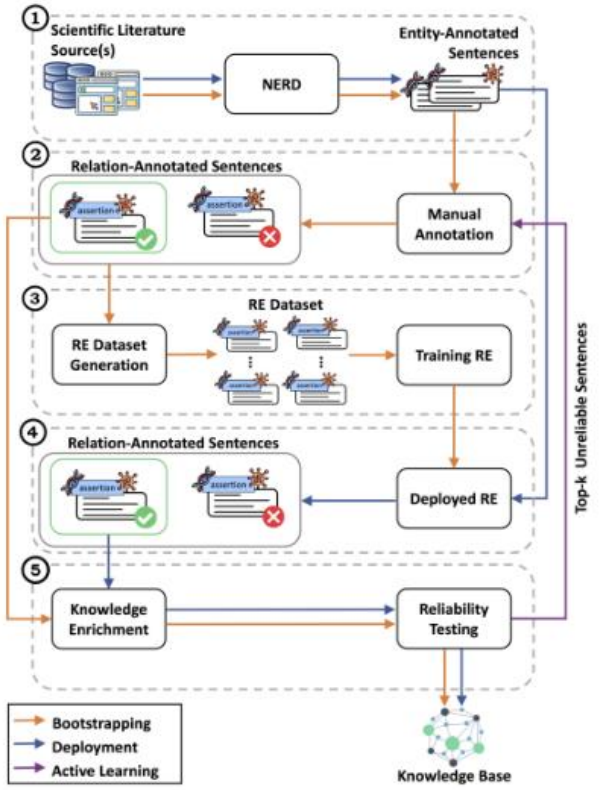
- Very sensitive data
- EMR (Electronic Medical Records)
- Clinical relevance
- Vocabulary mismatch
  - Patient describing a symptom
  - MDs describing a diagnosis
- Data labeling and curation

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)

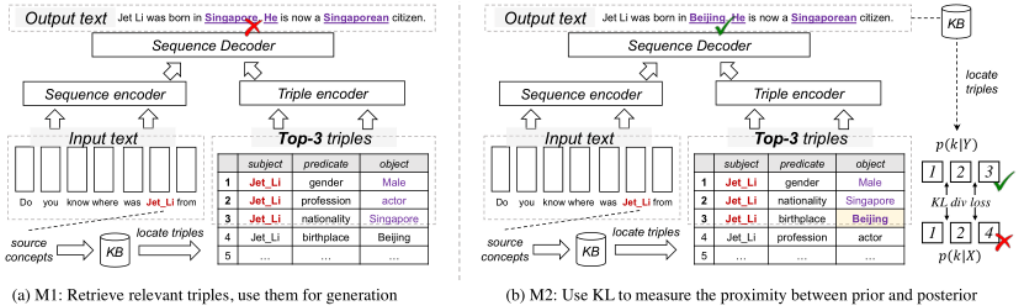
- KG generation system
- Combination of automated ML-based methods and domain experts
- Modular architecture that can be easily modified
- Reliability tests
- Active learning process make the system suited to iterative
- Versioning



S. Marchesin et al. "Building a Large Gene Expression-Cancer Knowledge Base with Limited Human Annotations" Database, 2023

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



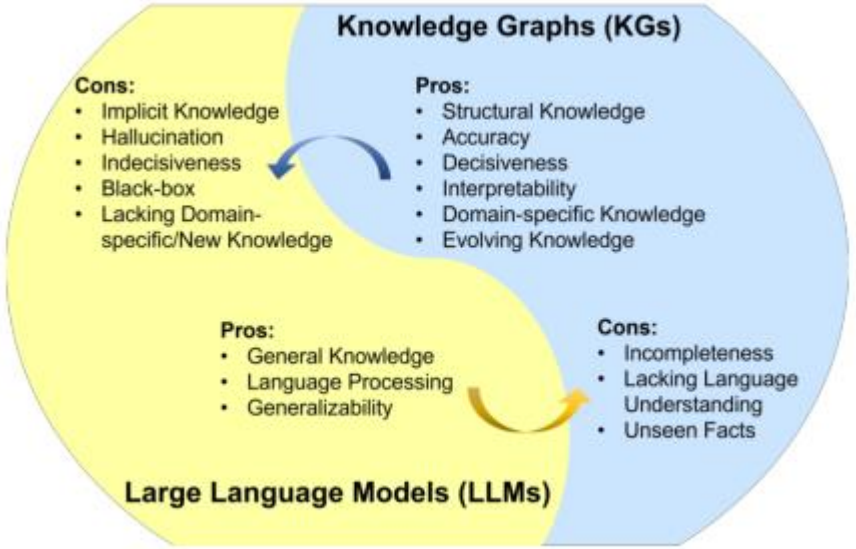
(a) M1: Retrieve relevant triples, use them for generation

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

W. Yu et al. "A Survey of Knowledge-Enhanced Text Generation" ACM Computing Surveys 2022

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



S. Pan et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap" arxiv.org/abs/2306.08302

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

Problems

- Data quality for KGs
- Use of LLMs for KG construction
- Redundancy and completeness
- Maintenance
- LLM answer validation with KGs