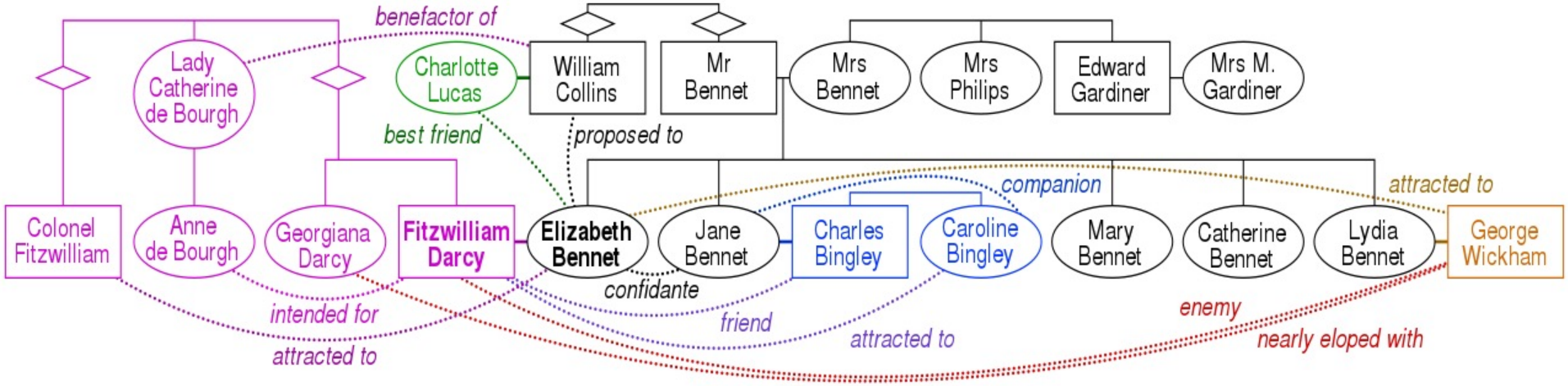
Natural Language Processing

09: Information Extraction

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-06-10

Technology
Arts Sciences
TH Köln



parent(Mr. Bennet, Jane Bennet)

Kamelle zum Schneewalzer

Frost kühlt die Feierlaune. Fangesang und Schunkellieder begleiten Zug im Stop-and-go-Verkehr

Aufmarsch im Gewölk: Grün-weiß sind die Reiser farben. In Erle kamen die Garde-Mädels damit gut an. Fotos: Caroline Seidel

Jörn Stender

Winter pur von Ückendorf bis Buer. Da haben es Narren schwer, heißblütig zu feiern und das Kultur-Motto der Session mit Leben zu füllen. Stoffarme Kostüme verbieten sich Rosenmontag quasi zwangsläufig. Und so sind viele Verummmte zu sehen. Dazu dick gepolsterte Häftlinge, Bienen und Teufel, jede Menge Cowboys und Mönche, einige Bart Simpsons, entzückende Engelchen und ganz viele, die höchstens ein wenig Konfetti im Haar zur Skijacke tragen.

Vereint tanzen sie den „Schnee, Schnee-, Schneewalzer" an der Cranger Straße, wärmen sich mit Glühwein zum Bierchen, erhitzen sich beim verschärften Kamelle-Klauben und Stadion-erprob-

Ganz in blau hatten die beiden Damen mächtig Spaß.

«Komm auf ein Bierchen nach Gelsenkirchen. Das ist Karnevalsmusik.»

ten Schlachtrufen. „Oh, wie ist das schön" tönt es an der Aral-Tankstelle, dem Jugendtreff zum Zug. Dazu passt, das viele auf Fantasiekostüme setzen: Schalke-Schal, alternativ ein Trikot und dazu alles, was die Kehle ölt. Glasverbot hätte sich hier bezahlt gemacht. So freuen sich die Pfandsammler.

Die Verbal bilden Bryzel-Wagen. Die Versorgung langs der jecken Piste ist gut. Alle paar Meter ein Würstchen- oder Bierstand, hie und da ein Mobil-Klo. Paul und André haben ihr Grillzelt aufgebaut. Würstchen und Koteletts brutzeln auf dem Rost. Das wärmt, aber nicht die Fülle. „Wir sind seit 9 Uhr hier, das langt." Weniger als sonst los sei diesmal, findet André. Frost kühlt das Geschäft merklich ab.

Um 14.30 Uhr ist der Zug gestartet. Von der Sparkasse aus sieht man eine halbe Stunde später gerade mal ein paar Blinklichter in der Ferne. Die Zeit vertreibt sich eine zuntere Marinetruppe mit der richtigen Einstimmung. Sechs Mann hoch sind die Energiebündel der KG Narrenzunft unterwegs mit Bollerwagen und Bierkästen. Die Herren sind gestählt durch Auftritte bei diversen Damensitzungen. „Wir sind die Showtanzgruppe", lacht einer. Und schon ziehen die Blauen Jungs weiter.

Der mittlerweile 40. Rosenmontagszug rollt endlich ein. Was heißt: rollt. Es geht im Stop-and-go-Verkehr voran im karnevalistischen Drei-Kampf aus Helau, Schunkeln und

Gut gelaunt unterwegs: Björn I. und Jessica I. hoch oben auf ihrer rollenden Narrenburg. Das Prinzenpaar der Erler Funken hatte ordentlich Wurfgut an Bord.

Auf Matrosen ... überschwappende Laune beim Zuch in Erle haben diese beiden Herren.

Wurfgut abschmeißen. Da fliegen Bonbons, Bärchen, Bälle und (arg bedrohlich) belegte Brötchen tief, da werden Schalke-Poster und Popcorntüten in die Menge geschossen. Und die Piccolo-Kanone feuert Konfettiregen hinterher. 58 Programmpunkte hat der Zug auf seiner 2,5 km langen Fahrt zu bieten: Zwischen Lok und Disco-Wagen der Alten Hütte fahren die Zoom-Erlebniswelt des KC Asteria, hebt der Reiterverein Gelsenkirchen musikalisch mit dem „Ha-, Ha-, Hubschrauber" ab und sind die Prinzenpaare auf ihren rollenden Punkenburgen unterwegs. Ein AUF-Trüppchen gibt politisch die Haifischjäger. Potenzielle Beute folgt zwei Wagen später: Die Eishockey-Sharks geben sich die Ehre.

Dabeisein ist für Silke Glüsing mit Laura und Marco, Nico und Lenny geradezu Pflicht. Die gemischte Schalke-Tiergruppe steht an der Ecke Friedenstraße. „Unser Stammplatz", sagt die Mutter. „Wir wohnen ganz in der Nähe." Die Kinder haben die Tüten voll Wurfgut. Nachbarin Nicole Lindner hat sich zur Feier des Tages erstmals verkleidet. Als flauschiger rosa Esel. Hauptsache warm.

Bericht Seite 3

ZUGBILANZ

„Es ist alles super gelaufen"

„Es ist alles super gelaufen. Keiner verletzt, keine Probleme. Auch Publikum war ausreichend da", freut sich Zugleiter Werner Preißler. Er glaubt, dass 100 000 nach Erle strömten. Die Polizei rechnet etwas zurückhaltender mit 80 000 Zuschauern. 111 Beamte waren neben Feuerwehr und Sanitätern vor Ort. „Es ist alles ruhig verlaufen", heißt es nach dem Zug auf der Leitstelle. Sie blieb es bis in den Abend. Der Frost kühlte auch so manches Mütchen.

bützen(Ralf, Jessica)

Information extraction

- Named entity recognition
- Relation extraction
- Entity linking

Named entity recognition

[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

- Identifying spans of text that correspond to typed entities

Main entities according to ACE (Automatic Content Extraction):

- Person (PER)
- Organization (ORG)
- Geo-political Entity (GPE)
- Location (LOC)
- Facility (FAC)
- Vehicle (VEH)
- Weapon (WEA)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf

Named entity recognition

- GENIA corpus of MEDLINE abstracts (biomedical content)

We have shown that [interleukin-1]PROTEIN ([IL-1]PROTEIN) and [IL-2]PROTEIN control [IL-2 receptor alpha (IL-2R alpha) gene]DNA transcription in [CD4- CD8- murine T lymphocyte precursors]CELL LINE

protein
cell line
cell type
DNA
RNA

http://www.aclweb.org/anthology/W04-1213

Named entity recognition

- In Academia Neural or Conditional-Random-Field models are the norm to NER

- Commercial approaches to NER are often based on pragmatic combinations of lists and rules
  - One common approach: repeated rule-based passes over a text
  - Starting with rules with very high precision but low recall, and, in subsequent stages, using machine learning methods that take the output of the first pass into account.

- Other approaches: Human-annotations, like crowd-sourcing

Relation extraction

The Big Sleep (1946 film)

From Wikipedia, the free encyclopedia

The Big Sleep is a 1946 American film noir directed by Howard Hawks,[2][3] the first film version of the 1939 novel of the same name by Raymond Chandler. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results".[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.[5][6]

subject	predicate	object
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall

Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

Relation extraction – ACE

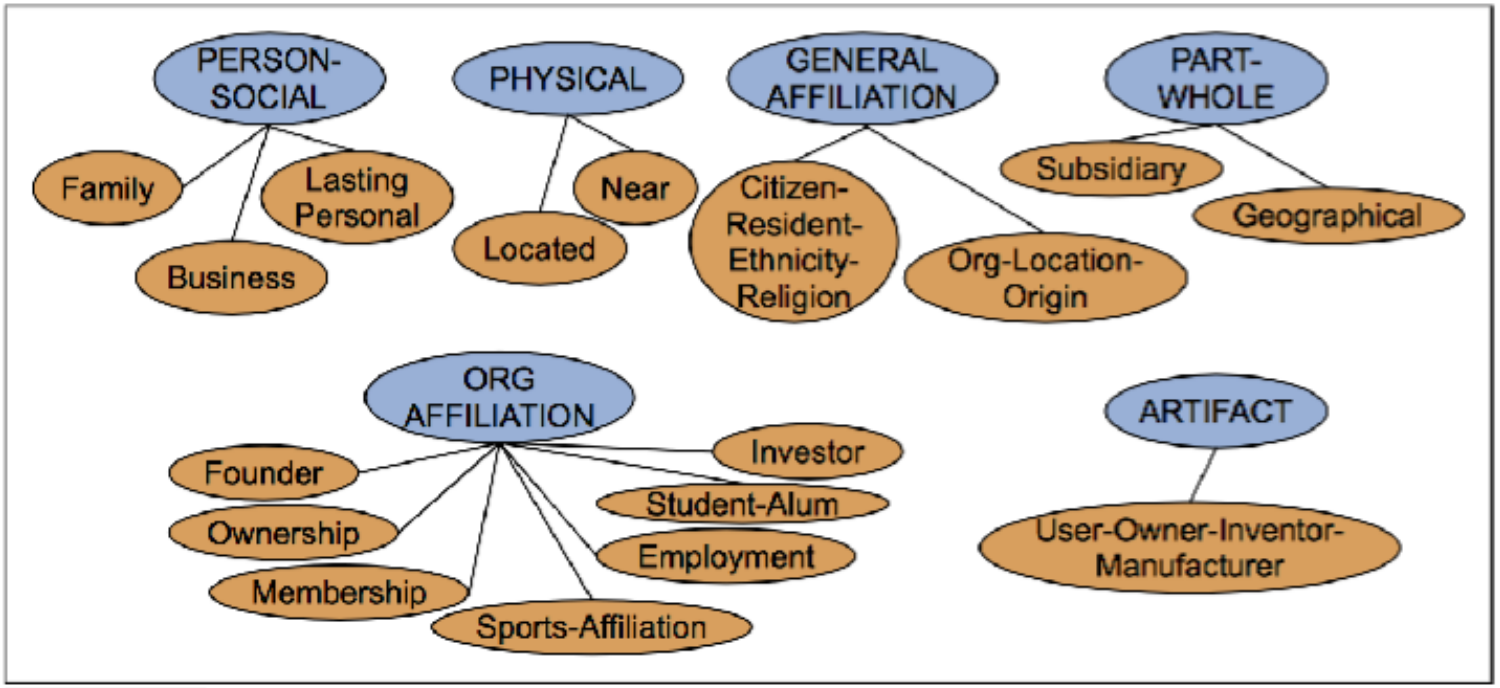


Figure 17.9 The 17 relations used in the ACE relation extraction task.

Jurafsky, Martin (2020): Speech and Language Processing, p. 334

Relation extraction – UMLS

Entity	Relation	Entity
Injury	disrupts	Physiological Function
Bodily Location	location-of	Biological Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Unified Medical Language System (UMLS)

Regular expressions

Regular expressions are precise ways of extracting high-precision relations

- "NP_1 is a film directed by NP_2" → directed_by(NP_1, NP_2)
- "NP_1 was the director of NP_2" → directed_by(NP_2, NP_1)

Hearst patterns

pattern	sentence
NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP,}* {(or|and)} NP	European countries, especially France, England, and Spain

NP_H is the parent or hyponym

Hearst (1992): Automatic Acquisition of Hyponyms from Large Text Corpora - https://www.aclweb.org/anthology/C92-2082.pdf

Hearst patterns

Hand-built patterns have

- the advantage of high-precision and they can be tailored to specific domains.

On the other hand,

- they are often low-recall, and it's a lot of work to create them for all possible patterns.

Alternative: Use supervised relation classifiers

- But: These classifiers need a lot of labeled training data

Distant supervision

- It's uncommon to have labeled data in the form of <sentence, relation> pairs
- More common to have knowledge base data about entities and their relations that's separate from text.
- We know the text likely expresses the relations somewhere, but not exactly where.

Distant supervision

Start with seed relation: chancelor(Angela Merkel, Germany) and search in a large dataset (like the Web or Wikipedia)

- **Angela Merkel** (née Kasner; born 17 July 1954) is a **German** politician serving as the **chancellor** of **Germany** since 2005.
- **Chancellor Angela Merkel**, who hosted the online meeting, pledged further ...
- Biography of **German** politician **Angela Merkel**, who in 2005 became the first female **chancellor** of **Germany**.
- Annalena Baerbock, Greens ... The only woman in the race to succeed **Angela Merkel**, she is the Greens' first ever candidate for **chancellor**, as ...

Distant supervision

- Training instances can now be extracted from the data
- We can then apply feature-based classification.
  - Features are: The named entity labels of the two mentions, the words and paths in between the mentions, neighboring words, …
  - Each tuple will have features collected from many training instances; the feature vector for a single training instance will have lexical and syntactic features from many different sentences that mention Merkel and Germany.
- Because distant supervision has very large training sets, it is also able to use very rich features that are conjunctions of these individual features. So we will extract thousands of patterns that conjoin the entity types with the intervening words.

Wikipedia Infoboxes

The Big Sleep (1946 film)

From Wikipedia, the free encyclopedia

The Big Sleep is a 1946 American film noir directed by Howard Hawks,[2][3] the first film version of the 1939 novel of the same name by Raymond Chandler. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results".[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.[5][6]

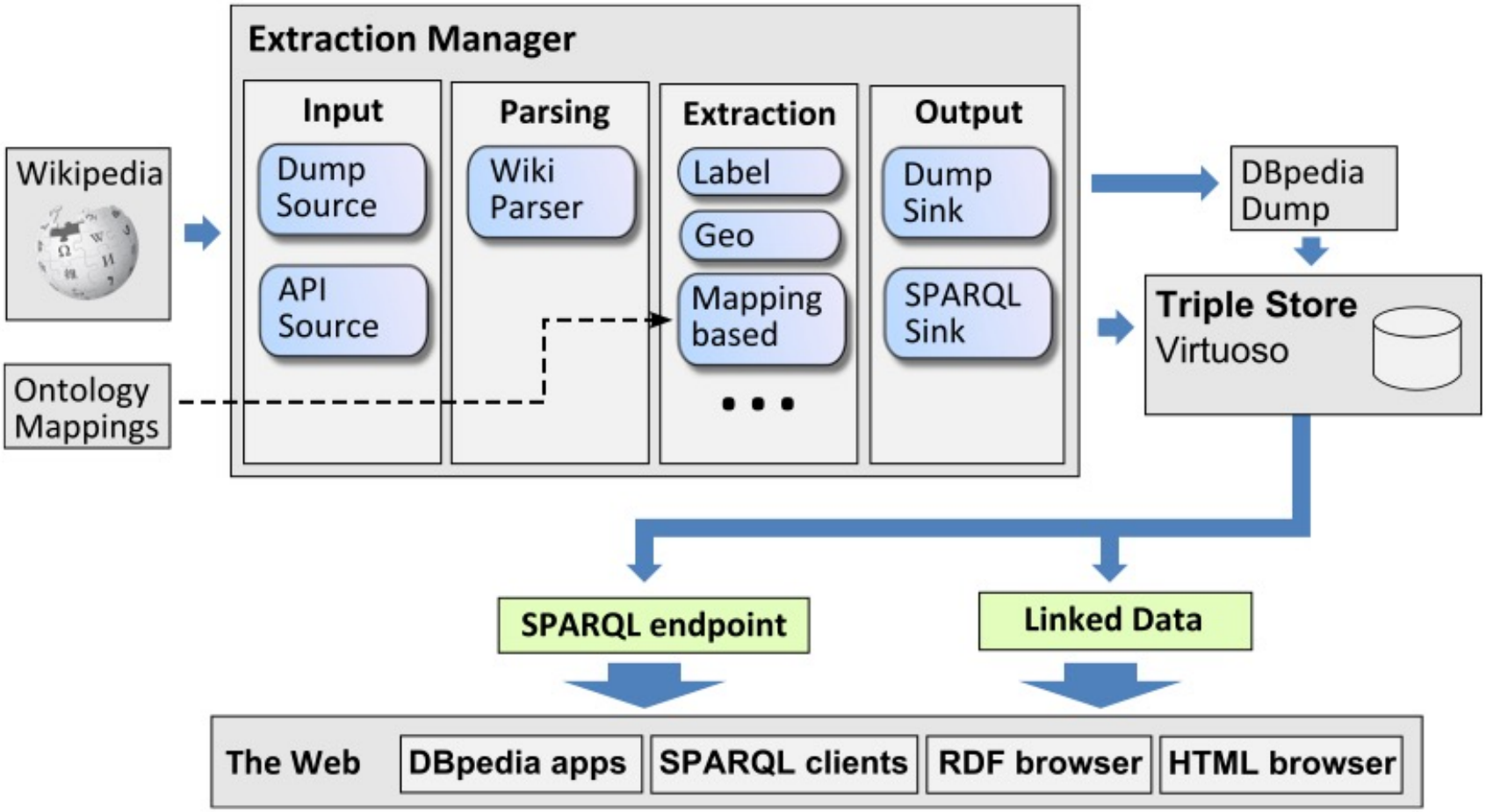
The Big Sleep

Theatrical release lobby card

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	*The Big Sleep* by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall Martha Vickers Dorothy Malone
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby

https://en.wikipedia.org/wiki/The_Big_Sleep_(1946_film)

DBpedia



Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

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/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

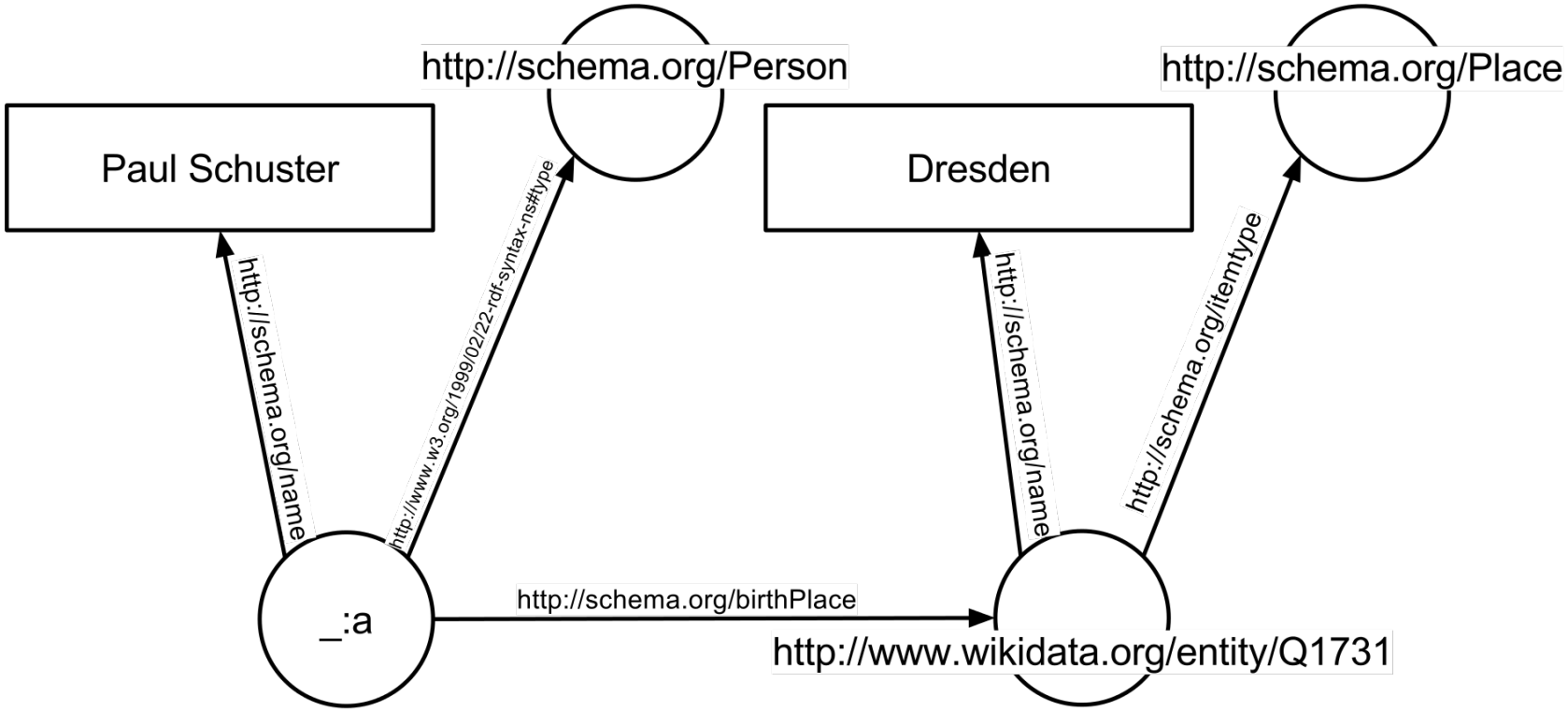
Table 2: The 23 largest Freebase relations we use, with their size and an instance of each relation.

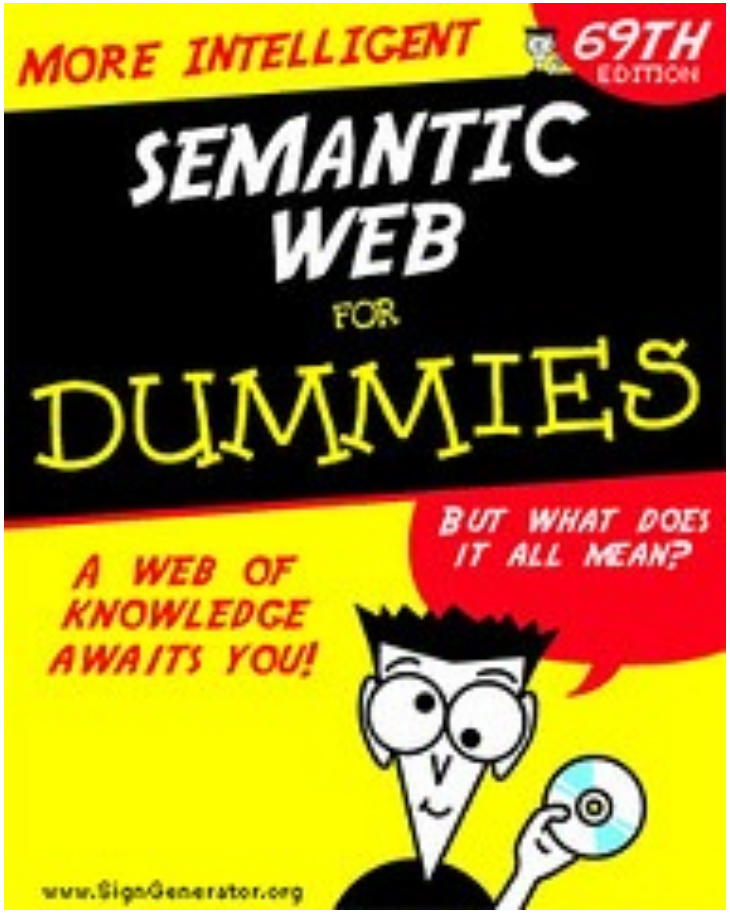
Mintz et al. (2009): Distant supervision for relation extraction without labeled data - https://www.aclweb.org/anthology/P09-1113.pdf

RDF – Resource Description Framework

```
<div vocab="http://schema.org/" typeof="Person">
    <span property="name">Paul Schuster</span> wurde in
    <span property="birthPlace" typeof="Place
    href="http://www.wikidata.org/entity/Q1731">
        <span property="name">Dresden</span>
    </span> geboren.
</div>
```

RDF – Resource Description Framework





From HTML to Knowledge Graphs

Classic Web

- HTML
- URL

- human generated free text
- addressing documents

Metadata

- RSS
- URI

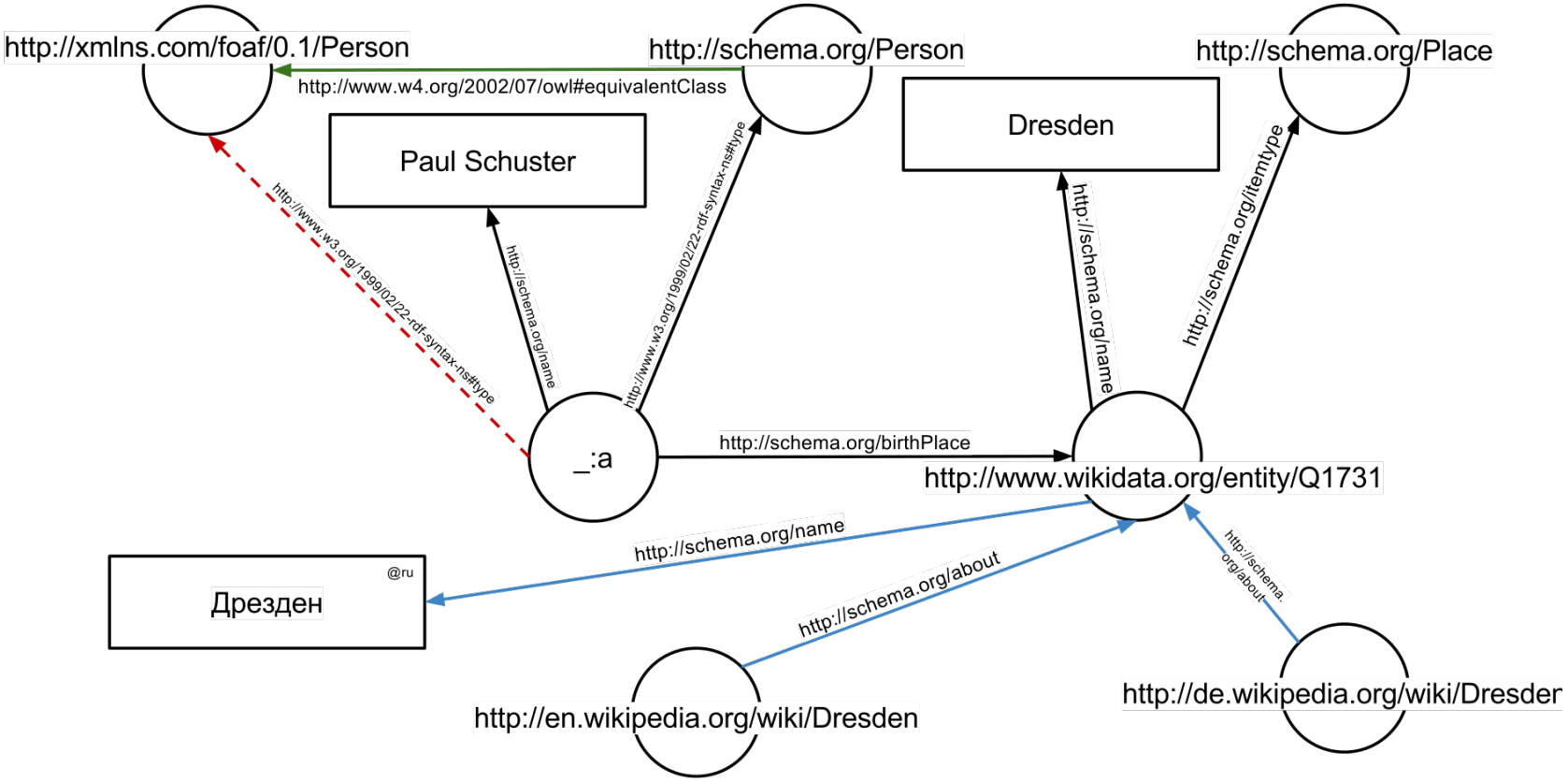
- description on document level
- addressing all ressources

Knowledge Represenation

- RDF
- LOD

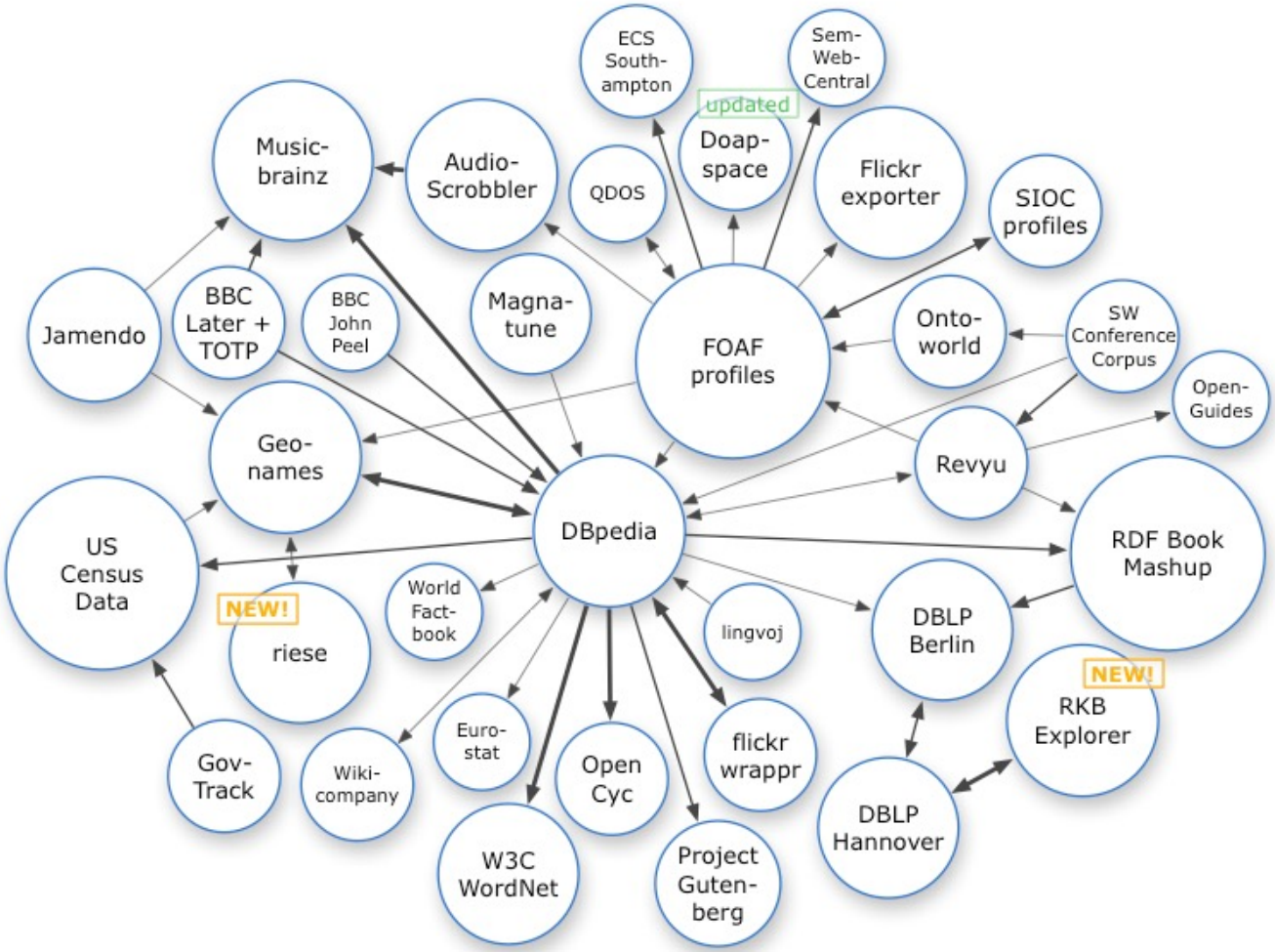
- description on entity level
- linking between entities

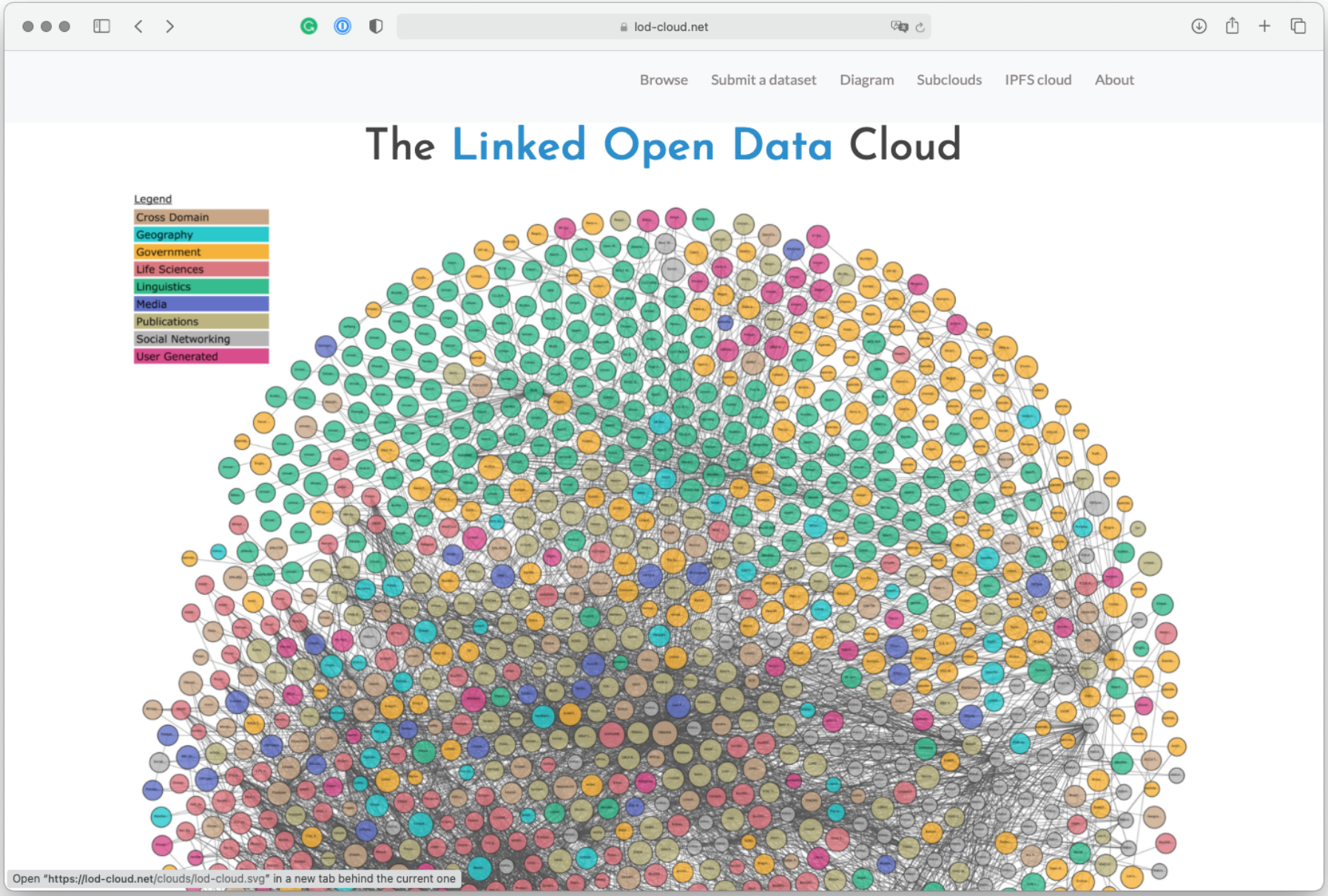
Linked (Open) Data



https://de.wikipedia.org/wiki/Semantic_Web#/media/Datei:RDF_example_extended.svg – CC0 by Denny Vrandečić

Putting it all together… ~ 2009





https://lod-cloud.net

Fun fact

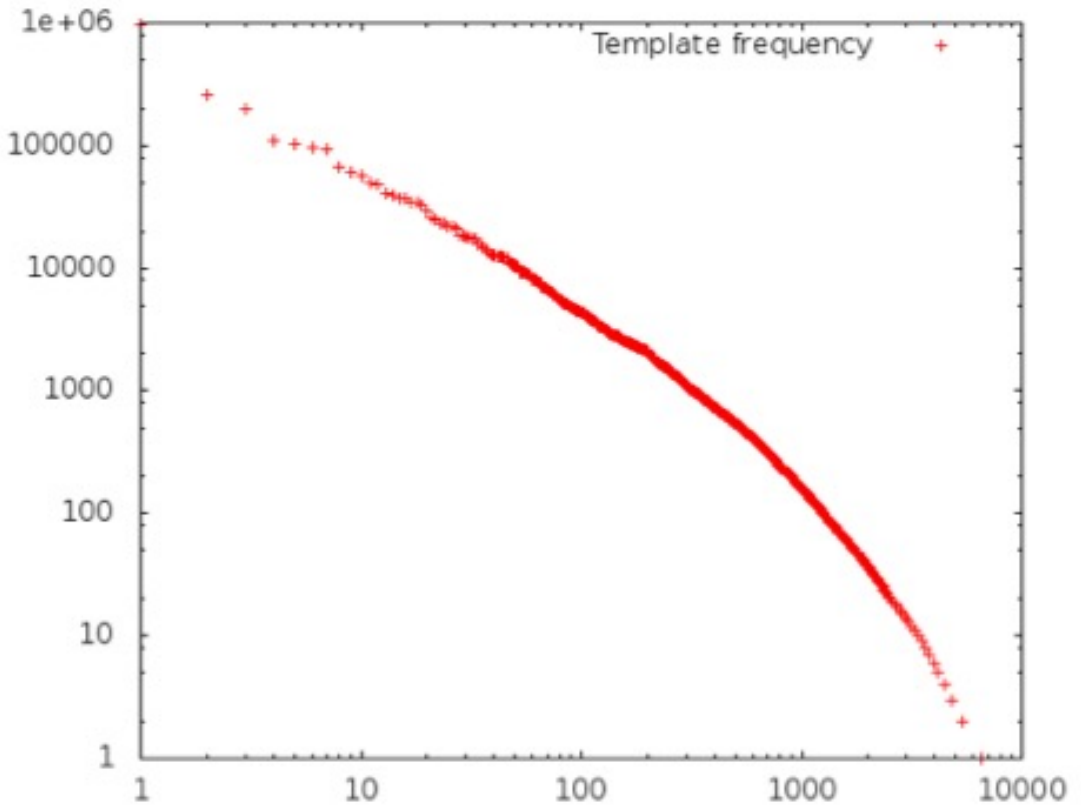
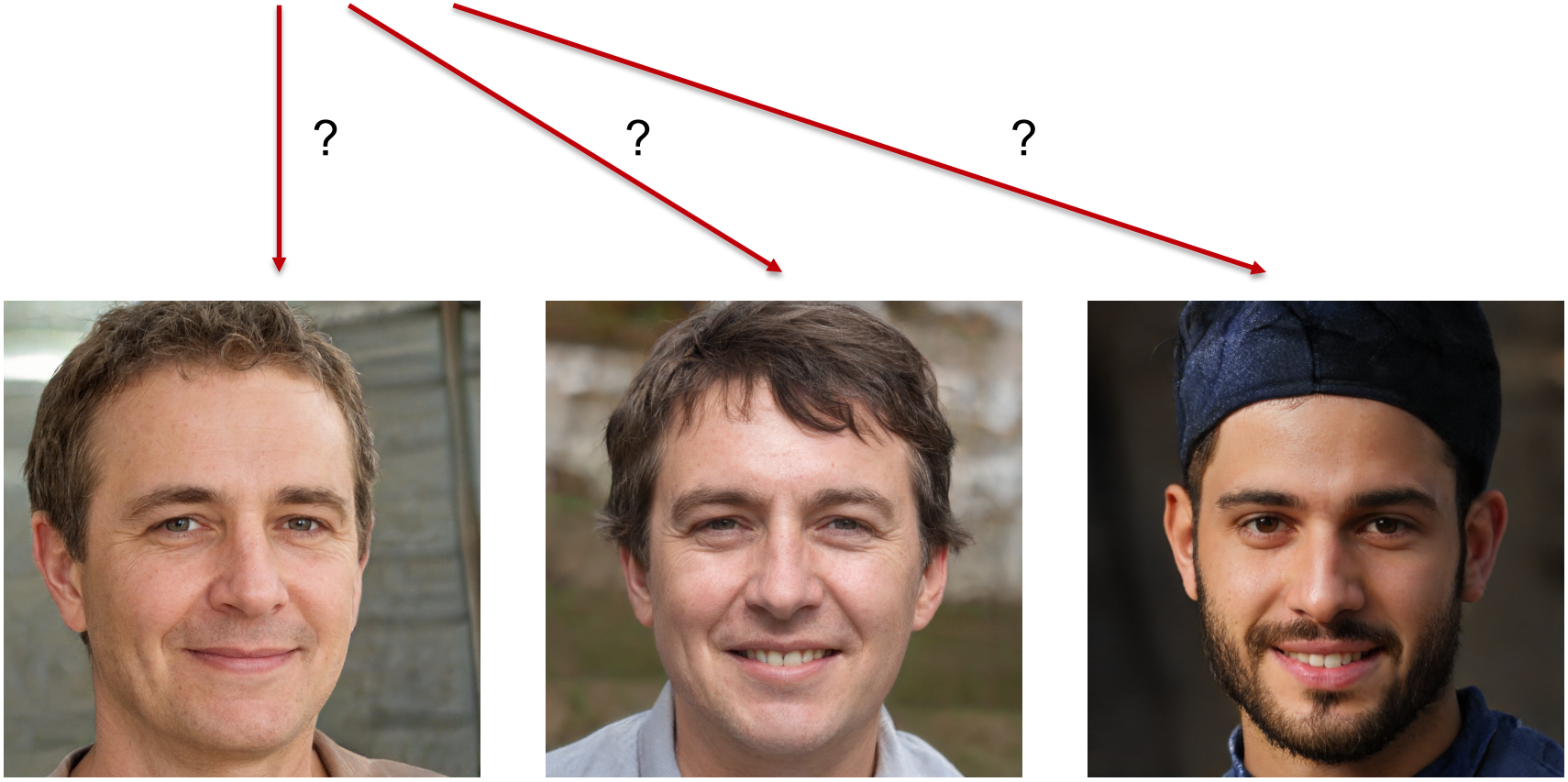


Fig. 7. English property mappings occurrence frequency (both axes are in log scale)

Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

Entitiy Linking

- The Fluxcompensator engine was discovered in 2023 be "Michael de Blaunche"



https://thispersondoesnotexist.com – Karras et al. (2019)

Entity Linking aka Wikification

- Entity linking is the task of associating a mention in text with the representation of some real-world entity in an ontology

- Wikipedia is often the source of the text that answers the question. Each unique Wikipedia page acts as the unique id for a particular entity. This task of deciding which Wikipedia page corresponding to an individual is being referred to by a text mention has its own name: Wikification.

- Different approaches:
  - Use of anchor texts (TAGME): https://tagme.d4science.org/tagme/
  - Neural Graph-based linking

Ferragina, P. and Scaiella, U. (2011). Fast and accurate annotation of short texts with wikipedia pages

Fresh from the Research Front

TREC News Track

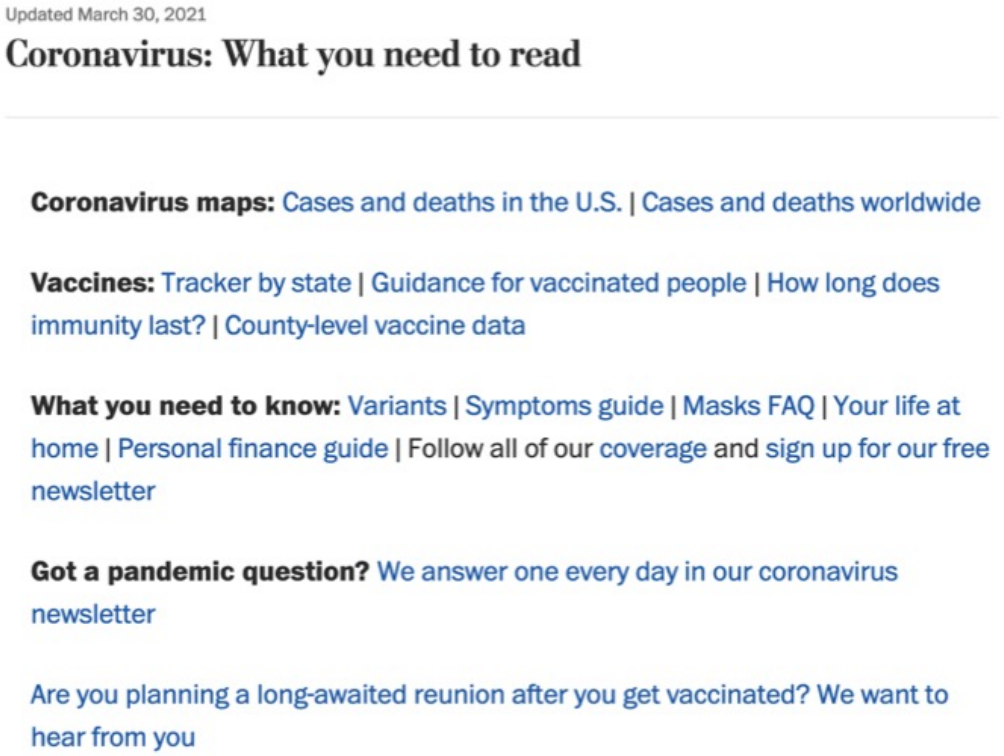
Task 1: Background linking

- Given a news story, retrieve other news articles that provide important context or background information. Build these links into an "explainer" box where links are grouped by what kind of context they offer. The goal is to help a reader understand or learn more about the story or main issues in the current article using the best possible sources.

Task 2: Wikification

- Given a news story, identify short passages in the article that should be hyperlinked to either another article, or a Wikipedia article, in order to provide in-context access to information that would help contextualize or fill in background on the story being read.

Background linkling



Updated March 30, 2021

Coronavirus: What you need to read

Coronavirus maps: Cases and deaths in the U.S. | Cases and deaths worldwide

Vaccines: Tracker by state | Guidance for vaccinated people | How long does immunity last? | County-level vaccine data

What you need to know: Variants | Symptoms guide | Masks FAQ | Your life at home | Personal finance guide | Follow all of our coverage and sign up for our free newsletter

Got a pandemic question? We answer one every day in our coronavirus newsletter

Are you planning a long-awaited reunion after you get vaccinated? We want to hear from you

The Washington Post implements manually linked "explainer boxes". This explainer box appeared at the end of most articles about COVID during the pandemic.

http://trec-news.org/guidelines-2021.pdf

TREC News Track

Task 1: Background linking

- Corpus with 728,626 newspaper articles from the Washington Post
- 51 Topics consisting of: Article, Description, Narrative

```
<top>
<num> Number: xxx </num>
<docid> f30b7db4-cc51-11e6-a747-d03044780a02</docid>
<url> https://www.washingtonpost.com/local/public-safety/homicides-remain-st eady-in-the-washington-region/2016/12/31/f30b7db4-cc51-11e6-a747-d03044780a 02_story.html</url>
<title> Topic title </title>
<desc> I would like to learn more about this topic </desc>
<narr> A traditional TREC narrative paragraph on the topic </narr>
<subtopics> <sub num="1">This is the first subtopic.</sub> <sub num="2">And this is the second one.</sub> </subtopics>
</top>
```

http://trec-news.org/guidelines-2021.pdf

Judging Relevance of Background Links

The relevance scale used by the NIST assessors was:

- 0: The linked document provides little or no useful background information.
- 1: The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.
- 2: The document provides significantly useful background . . .
- 3: The document provides essential useful background . . .
- 4: The document MUST appear in the sidebar otherwise critical context is missing.

TREC News Track

Example

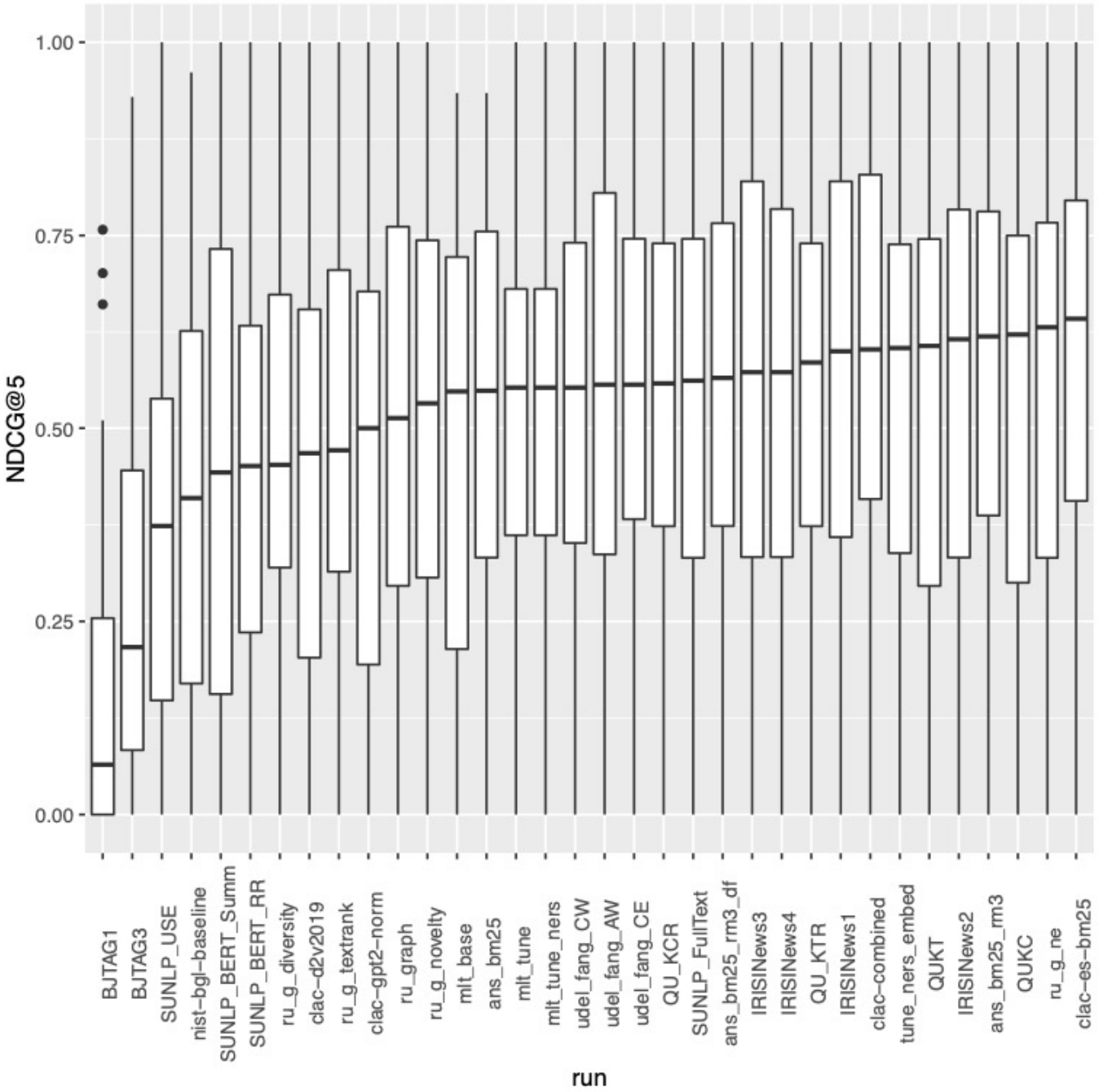
Query article: Love in the time of climate change: Grizzlies and polar bears are now mating (May 23, 2016)

This article describes and analyzes a phenomenon where grizzlies and polar bears are mating to create a new species known as pizzlies or grolars. It explains why this is happening and points out that it happens (or has happened) to other species as well. Articles along these lines are good background links. For example:

- Coywolves, coyote-wolf hybrids, are prowling Rock Creek Park and D.C. suburbs (July 1, 2014)
- Humans and Neanderthals may have interbred 50,000 years earlier than previously thought (February 17, 2016)

However, the following article is of less relevance and should be ranked lower because it's not about interbreeding.

- Why do seals keep trying to have sex with penguins? (November 18, 2014)


NDCG@5
1.00
0.75
0.50
0.25
0.00
run
BJTAG1
BJTAG3
SUNLP_USE
nist-bgl-baseline
SUNLP_BERT_Summ
SUNLP_BERT_RR
ru_g_diversity
clac-d2v2019
ru_g_textrank
clac-gpt2-norm
ru_graph
ru_g_novelty
mlt_base
ans_bm25
mlt_tune
mlt_tune_ners
udel_fang_CW
udel_fang_AW
udel_fang_CE
QU_KCR
SUNLP_FullText
ans_bm25_rm3_df
IRISINews3
IRISINews4
QU_KTR
IRISINews1
clac-combined
tune_ners_embed
QUKT
IRISINews2
ans_bm25_rm3
QUKC
ru_g_ne
clac-es-bm25

Our approach

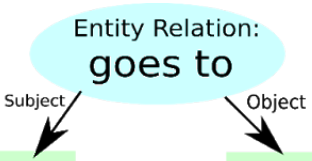
- Large corpus makes costly NLP techniques for all documents infeasible.
  - Use index and TF-IDF to select (approx. 200 documents/topic) interesting articles.
  - Extract entities and relations for selected documents.

- Assumption: Articles are more likely to be relevant if both articles share many common relations.

Thx @Björn

Our approach

- Entities often occur by chance, possibly in different contexts, in both articles:
  - **Nixon** *goes to* **China.**
  - **Nixon** was born in 1913. … In 2010, China became the world's second largest economy.



The shorthand is "Nixon goes to China," meaning a moment in which a leader reverses his past positions to do something that is shocking but beneficial.

Thx @Björn

Our approach

Extract all entities for each of the selected documents (spacy):

- New York, Trump, Labor Department, Nasa, Air Force

Find relations between the entities:

- Nixon visits China, Obama pass healthcare

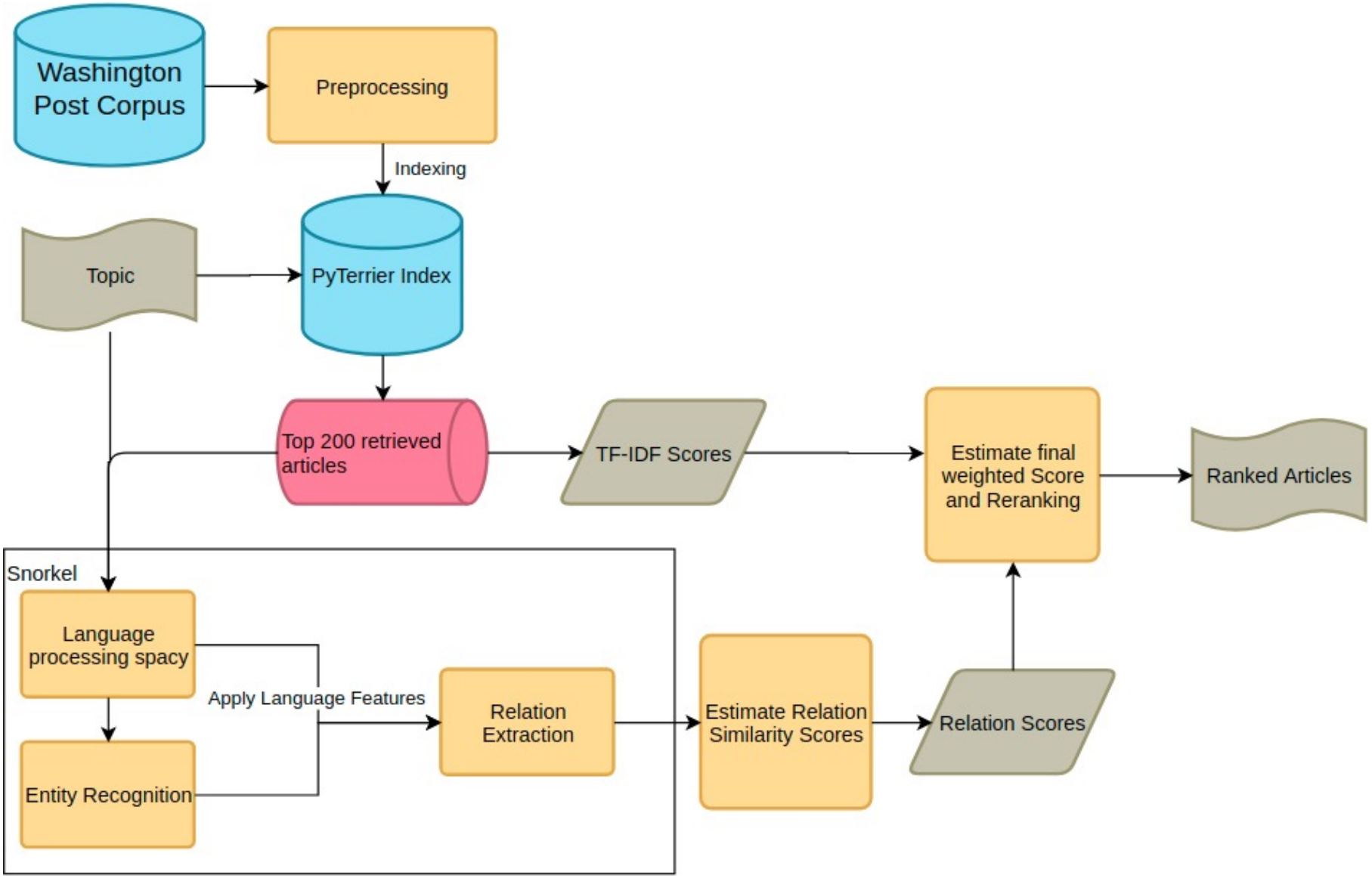
Simple recognition of relations via linguistic features (Snorkel):

- Do entities appear in the same sentence?
- Is there a subject-object relationship between entities?
- Which verb connects the entities?

Give more weight to rare entities (IDF).

Final score for document calculated from TF-IDF score and weighted amount of shared relations.

Thx @Björn


Washington Post Corpus
Preprocessing
Indexing
Topic
PyTerrier Index
Top 200 retrieved articles
TF-IDF Scores
Estimate final weighted Score and Reranking
Ranked Articles
Snorkel
Language processing spacy
Apply Language Features
Entity Recognition
Relation Extraction
Estimate Relation Similarity Scores
Relation Scores

Thx @Björn

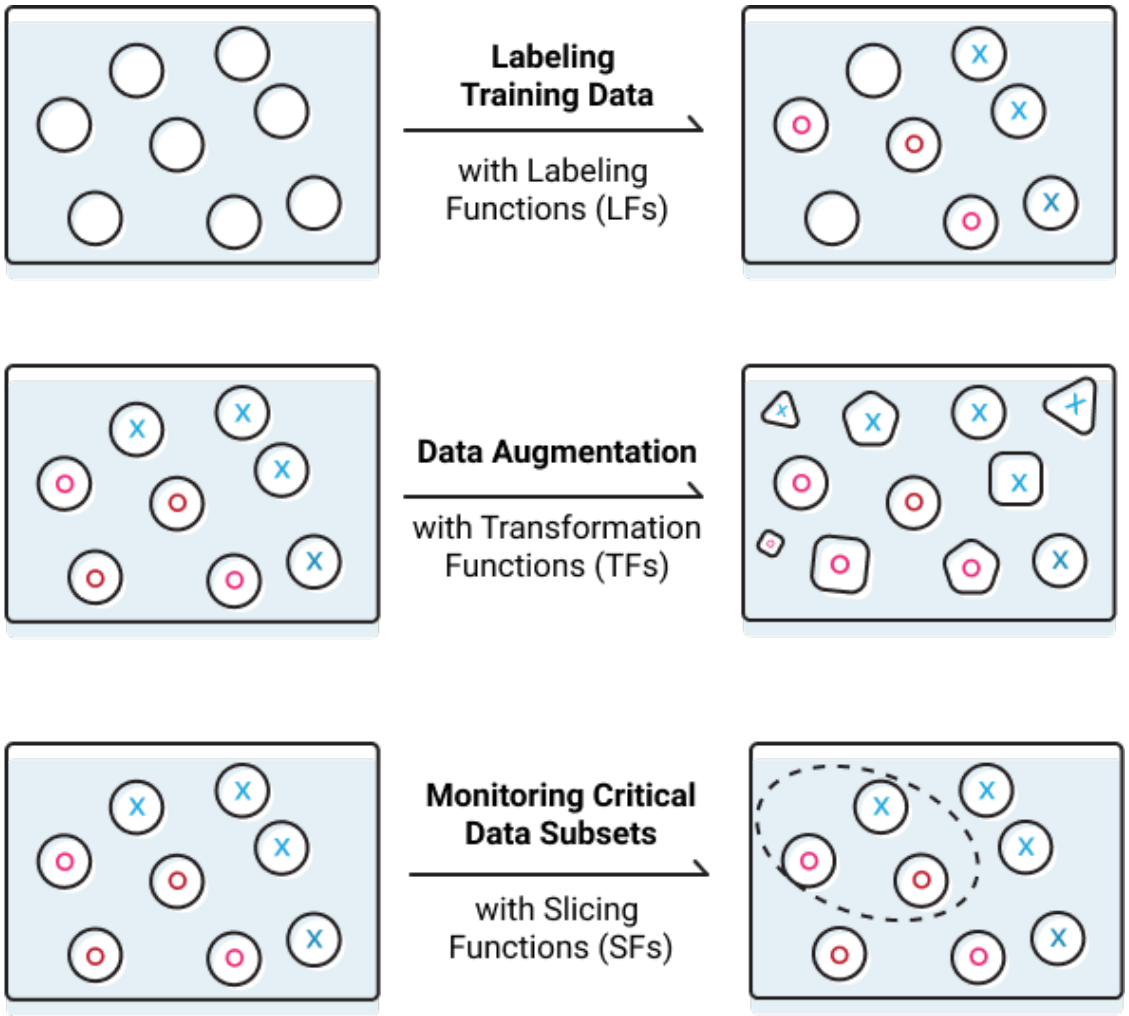
Real-world Example

Retrieved Example:

- **Topic:** Olympics 2020
- **Topic Title:** For 2020 Olympic hopefuls, postponement is another challenge to overcome.
- **Title of retrieved Article:** For finely tuned Olympic athletes, a one-year postponement changes everything
- **Shared Relation Entities:**
  - 'U.S.' - 'Tokyo'
  - 'British' - 'Adam Peaty'
  - 'Helen Maroulis' - 'Olympic'
  - 'American' - 'Emma Coburn'

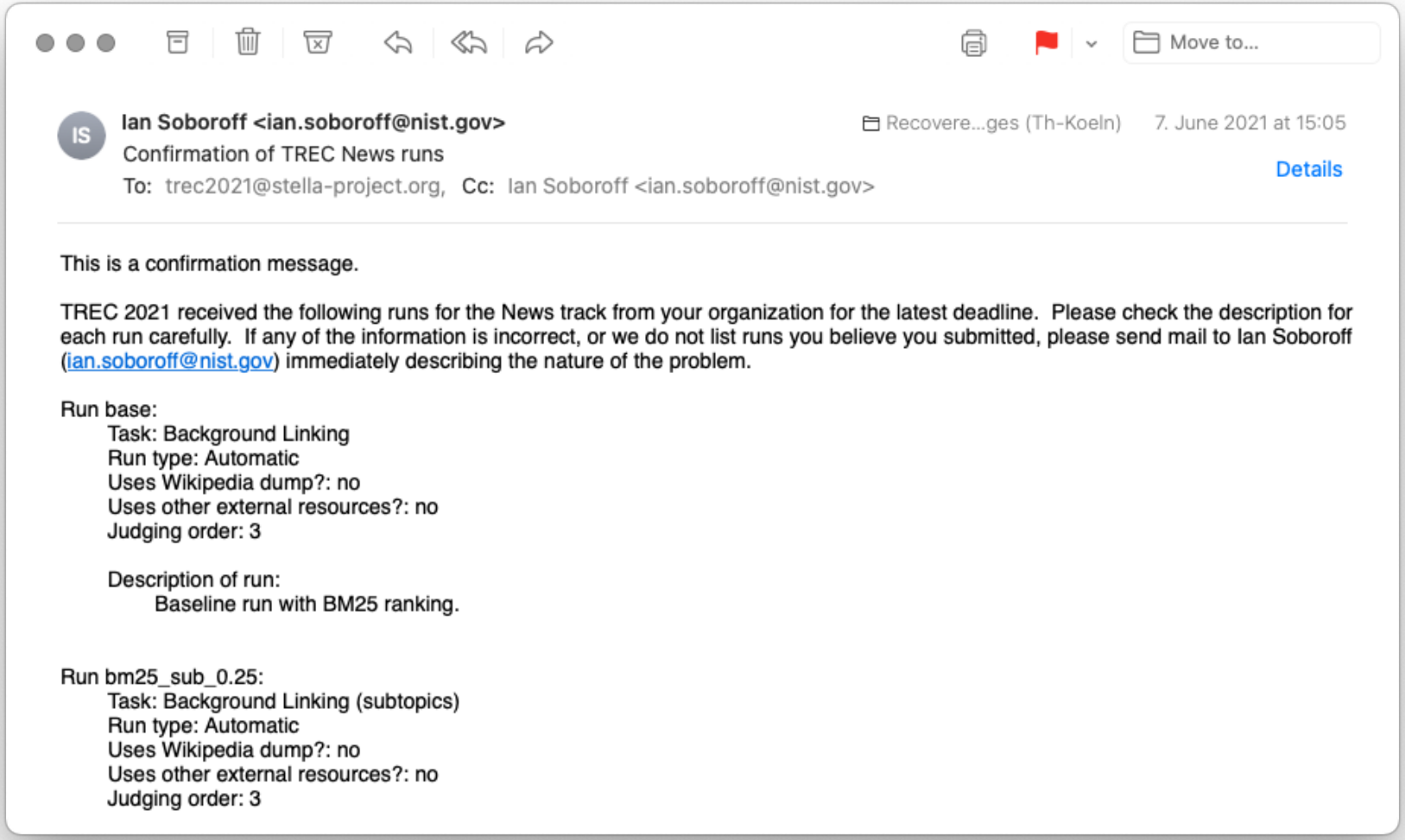
Thx @Björn

Outlook: Snorkel



https://www.snorkel.org/get-started/

Let's see …



Move to...

IS **Ian Soboroff <ian.soboroff@nist.gov>** Recovere...ges (Th-Koeln) 7. June 2021 at 15:05

Confirmation of TREC News runs

To: trec2021@stella-project.org, Cc: Ian Soboroff <ian.soboroff@nist.gov>

Details

This is a confirmation message.

TREC 2021 received the following runs for the News track from your organization for the latest deadline. Please check the description for each run carefully. If any of the information is incorrect, or we do not list runs you believe you submitted, please send mail to Ian Soboroff (ian.soboroff@nist.gov) immediately describing the nature of the problem.

Run base:
Task: Background Linking
Run type: Automatic
Uses Wikipedia dump?: no
Uses other external resources?: no
Judging order: 3

Description of run:
Baseline run with BM25 ranking.

Run bm25_sub_0.25:
Task: Background Linking (subtopics)
Run type: Automatic
Uses Wikipedia dump?: no
Uses other external resources?: no
Judging order: 3