

Natural Language Processing

KG and Entity Linking
Instructor: Yangqiu Song

Motivation



ChatGPT >

Tell me about Yangqiu Song.

Yangqiu Song is a computer scientist and an associate professor at the Hong Kong University of Science and Technology. He is known for his research in [natural language processing](#) (NLP) and machine learning.

Song received his Ph.D. in computer science from the [University of California, Berkeley](#) in [2010](#). After completing his Ph.D., he worked as a [research scientist](#) at [Google](#), where he contributed to the development of [Google Translate](#).

At [HKUST](#), Song leads the HKUST-NExT Research Centre, which focuses on [NLP](#) and [social media analysis](#). He has published numerous research papers on topics such as [sentiment analysis](#), machine translation, and social media analysis. His research has been recognized with several awards, including the [Best Paper Award](#) at the Conference on [Empirical Methods](#) in Natural Language Processing (EMNLP) in [2016](#).

Song is also an active member of the research community. He has served as an area chair for top-tier [NLP conferences](#) such as ACL and EMNLP and has been a program committee member for many other conferences. He is also an associate editor of the [Journal of Artificial Intelligence Research](#) (JAIR).

Overall, [Yangqiu Song](#) is a well-respected researcher in the field of NLP and [machine learning](#), and his work has contributed significantly to the development of these fields.



Tell me about Yangqiu Song.

Yangqiu Song is a computer scientist and researcher whose work primarily focuses on [natural language processing](#) (NLP), [machine learning](#), and [artificial intelligence](#) (AI). As of my knowledge cutoff date in September 2021, Song was an Assistant Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign (UIUC).

His research interests include designing new algorithms and models for [NLP tasks](#), such as question answering, [sentiment analysis](#), and [text generation](#). He has worked on various NLP and AI projects, including developing methods to improve the performance of deep learning models in understanding and generating human-like text.

Song has authored numerous research papers, and his work has been published in prestigious conferences and journals within the field of AI and NLP. He has also been a part of collaborative projects involving both academia and industry partners.

Please note that the information provided is accurate as of September 2021, and there may have been updates or changes in [Yangqiu Song's](#) career and research since then.

**Author**

ChatGPT-4o-Latest

Tell me about Yangqiu Song.



ChatGPT-4o-Latest

Poe

As of my knowledge cutoff in October 2023, **Yangqiu Song** is an academic and researcher in the fields of **artificial intelligence (AI)**, **machine learning**, **natural language processing (NLP)**, and **knowledge representation**. He has worked on various topics, including machine learning algorithms, information retrieval, and knowledge graphs.

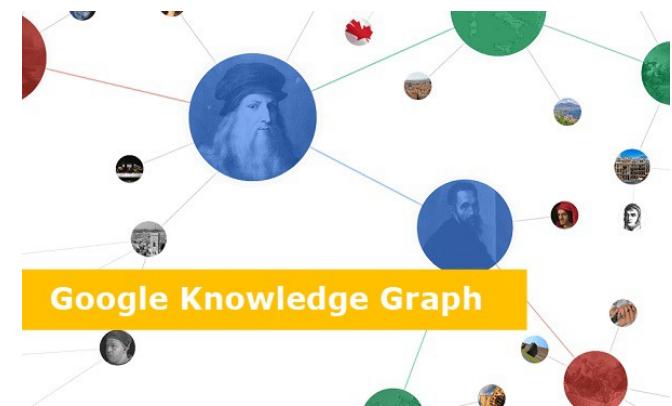
Key Highlights:

- **Position:** Yangqiu Song has been an associate professor in the **Department of Computer Science and Engineering** at **Hong Kong University of Science and Technology (HKUST)**.
- **Research Interests:** His research focuses on **machine learning**, **knowledge graphs**, **natural language processing**, **AI-driven systems**, and **information retrieval**. He aims to develop models and techniques that enhance machine understanding of text, reasoning, and knowledge-based systems.
- **Publications:** He has contributed to numerous academic papers in top-tier conferences and journals, including **ACL**, **EMNLP**, **ICML**, **NeurIPS**, and **AAAI**. His work often revolves around using structured knowledge (like knowledge graphs) to improve machine learning models' reasoning capabilities.
- **Collaborations:** Song has collaborated with various international research institutions and has been involved in projects that bridge the gap between academia and industry in AI and NLP.
- **Education:** He holds a PhD from **Tsinghua University** in China and has conducted postdoctoral research at prestigious institutions, including **Carnegie Mellon University** (CMU) in the United States.

We need knowledge to
ground the generated
texts!

What's a Knowledge Graph?

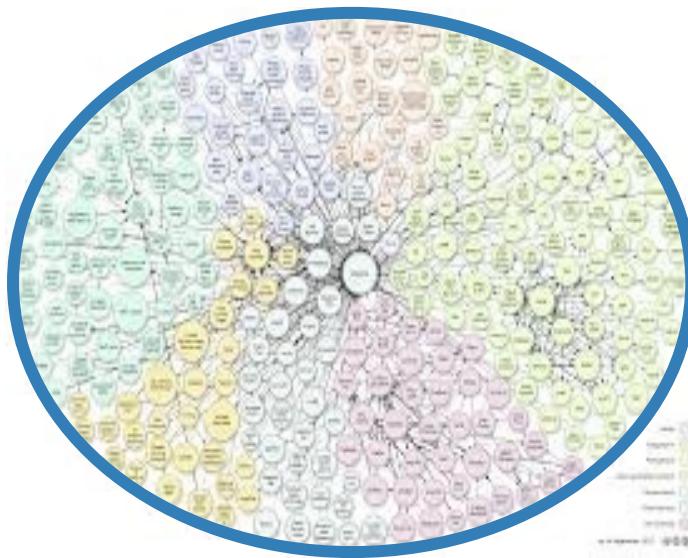
- A knowledge graph has many names in the history
 - Semantic networks, knowledge base, ontology, ...
- In 2012, Google released its project “Google Knowledge Graph”
 - A graph-based knowledge representation connecting real-world entities to support search
 - Landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more
 - Get information instantly relevant to a query



Knowledge graphs



Facebook's
Entity Graph



OpenIE
(Reverb, OLLIE)

Microsoft's Satori
Google's Knowledge Graph

State-of-the-art Enterprise-level KGs

	Data Model	Size of Nodes	Size of Edges	Development Stage
Google	Strongly typed entities, relations with domain and range inference	~1 Billion	~70 Billions	Actively used in products
Microsoft	The types of entities, relations, and attributes in the graph are defined in an ontology	~2 Billions	~55 Billions	Actively used in products
Facebook	All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal.	~50 Millions	~500 Millions	Actively used in products
eBay	Entities and relation, well-structured and strongly typed	~100 Million	~1 Billion	Early stages of development and deployment
IBM	Entities and relations with evidence information associated with them	~100 Millions	~5 Billions	Actively used in products and by clients

Surfacing structured results in web search

Google pisa italy

All Maps Images News Videos More ▾ Search tools

About 39,200,000 results (0.53 seconds)

Pisa - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Pisa> ▾ Wikipedia

Pisa is a city in Tuscany, Central Italy, straddling the River Arno just before it empties into the Tyrrhenian Sea. It is the capital city of the Province of Pisa.

Leaning Tower of Pisa · Piazza dei Miracoli · Baptistry · San Francesco (Pisa)

Images for pisa italy

Report images

PISA

More images for pisa italy

Virtual tour of Pisa Italy - History, facts, top attractions & things to do ...

www.italyguides.it/en/tuscany/pisa/ ▾

★★★☆★ Rating: 4 - 5,520 votes

Travel guide of Pisa Italy. Maps, articles, photos and destination guides about Pisa major attractions.

Leaning Tower of Pisa · Piazza dei Miracoli · Interactive map of Pisa

Pisa, Italy: Tourist Guide to Visiting the Leaning Tower of Pisa and ...

<https://www.discovertuscany.com/pisa/> ▾ Discover Tuscany

Pisa: City of the Leaning Tower and More. ... Buy Tickets for the Leaning Tower. Visit the Leaning Tower in Pisa: do not miss the chance to discover the secrets of one of the most famous monuments in the world.

Pisa 2016: Best of Pisa, Italy Tourism - TripAdvisor

https://www.tripadvisor.com/Tourism-g187899-Pisa_Province_of_Pisa_T... ▾ TripAdvisor

Pisa Tourism: TripAdvisor has 144925 reviews of Pisa Hotels, Attractions, and Restaurants making it your best Pisa resource.

Things to do in Pisa · Pisa Hotels · Restaurants · Pisa Hotel Deals

Travel Guide to Pisa in Tuscany, Italy - Italy Travel - About.com

goitaly.about.com/od/pisa/p/pisa.htm ▾

Pisa, Italy: Pisa is best known for its leaning tower but there is much more to see in this Tuscan town. The area around the cathedral and tower, Piazza dei ...

Pisa - Lonely Planet

<https://www.lonelyplanet.com/italy/tuscany/pisa> ▾ Lonely Planet

Once a maritime power to rival Genoa and Venice, Pisa now draws its fame ... economy since the 1400s, and students from across Italy compete for places in its ...

SIGIR 2016 | July 17-21 2016 – Pisa, Tuscany, Italy

sigir.org/sigir2016/ ▾ Special Interest Group on Information Retrieval

We welcome the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, exactly 30 years since SIGIR 1986, also ...

You've visited this page many times. Last visit: 6/21/16





Pisa

City in Italy

Pisa is a city in central Italy's Tuscany region best known for its iconic Leaning Tower. Already tilting when it was completed in 1372, the 56m white-marble cylinder is the bell tower of the Romanesque, striped-marble cathedral that rises next to it in the Piazza dei Miracoli, a grassy, walled square.

Hotels: 3-star averaging \$90, 5-star averaging \$140. [View hotels](#)

Getting there: 14 h 10 min flight, around \$2,560. [View flights](#)

Weather: 75°F (24°C), Wind NW at 3 mph (5 km/h), 83% Humidity

Local time: Thursday 10:09 PM

Province: Province of Pisa

Points of interest

[View 15+ more](#)


Leaning Tower of Pisa


Piazza dei Miracoli


Camposanto Monumentale


Santa Maria della Spina


Knights' Square

Colleges and Universities


UNIVERSITÀ DI PISA
University of Pisa


SANT'ANNA SCHOOL OF ADVANCED STUDIES
Sant'Anna School of Advance...


SCUOLA NORMALE SUPERIORE
Scuola Normale Superiore...

Augmenting the presentation with relevant facts

Surfacing facts proactively

Google **italy population**

All Images News Maps Books More ▾ Search tools

About 10,000,000 results (0.73 seconds)

Italy / Population

59.83 million

80M
60M
40M
20M
0
1960 1970

Explore more

Sources include: World Bank

Italy Population (2013)
www.worldometers.info
Italy Population (LIVE) Th based on the latest United Nations population. Italy ranks number 41 in the world.

Chris Wood: "It Will Take A Political Genius To Hold The EU Together", Italy Is The Flash Point
Zero Hedge - 1 hour ago
This Italian issue was discussed in more detail here a few months ago (see GREED & fear ...)

Germany Vs Italy: Heavyweights Collide at Euro 2016
ABC News - 10 hours ago

Thomas Muller confident Germany can breach Italy wall in Euro 2016 quarter-final
Daily Mail - 9 hours ago

More news for italy

Images for italy Report images

ROMA CAPITALE Colosseum

Lake Garda

Leaning Tower of Pisa

Amalfi Coast

Lake Como

View 45+ more

Google **italy**

All Images News Videos More ▾ Search tools

About 819,000,000 results (0.55 seconds)

[Italian Tourism Official Website](http://www.italia.it/en/home.html)
www.italia.it/en/home.html ▾ 10 hours ago
Italian tourism official website: vacations, art and culture, history, events, nature, lakes, mountains, golf, sci, boating, thermal spas, sports and adventure.

[Italy - Wikipedia, the free encyclopedia](http://https://en.wikipedia.org/wiki/Italy)
https://en.wikipedia.org/wiki/Italy ▾ Wikipedia
Location of Italy (dark green) – In Europe – in the European Union (light green) – [Legend]. Capital and largest city, Rome - 41°54'N 12°29'E / 41.900°N ...
Sergio Mattarella - History of Italy · Italian language · Demographics of Italy

In the news

Chris Wood: "It Will Take A Political Genius To Hold The EU Together", Italy Is The Flash Point
Zero Hedge - 1 hour ago
This Italian issue was discussed in more detail here a few months ago (see GREED & fear ...)

Germany Vs Italy: Heavyweights Collide at Euro 2016
ABC News - 10 hours ago

Thomas Muller confident Germany can breach Italy wall in Euro 2016 quarter-final
Daily Mail - 9 hours ago

More news for italy

Points of interest

View 45+ more

Italy

Country in Europe

Italy, commanding a long Mediterranean coastline, has left a powerful mark on Western culture and cuisine. Its capital, Rome, is home to the Vatican as well as landmark art and ancient ruins. Other major cities include Florence, with Renaissance treasures such as Michelangelo's "David" and its leather and paper artisans; Venice, the sinking city of canals; and Milan, Italy's fashion capital.

Capital: Rome

Currency: Euro

Population: 59.83 million (2013) World Bank

President: Sergio Mattarella

Prime minister: Matteo Renzi



FREEBASE AS AN EXAMPLE OF A LARGE SCALE KNOWLEDGE REPOSITORY

- **Freebase** is an open, [Creative Commons](#) licensed repository of [structured data](#)
- **Typed entities** rather than



Person Type

Key: /people/person Includes: Topic

A person is a human being (man, woman or child) known to have actually existed. Living persons, celebrities and politicians are persons

Table Diagram

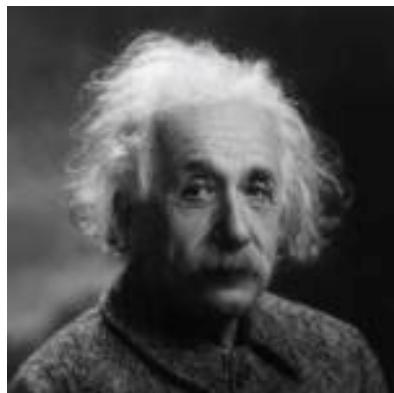
Properties

Property	ID	Expected Type
Date of birth	/people/person/date_of_birth	/type/datetime
Place of birth	/people/person/place_of_birth	/location/location
Country of nationality	/people/person/nationality	/location/country
Gender	/people/person/gender	/people/gender enumerated
Profession	/people/person/profession	/people/profession
Religion	/people/person/religion	/religion/religion
Ethnicity	/people/person/ethnicity	/people/ethnicity
Parents	/people/person/parents	/people/person
Children	/people/person/children	/people/person
Siblings	/people/person/sibling_s	/people/sibling_relationship
Spouse (or domestic partner)	/people/person/spouse_s	/people/marriage
Employment history	/people/person/employment_history	/business/employment_tenu
Education	/people/person/education	/education/education

Relations are typed too!

Subject-Predicate-Object (SPO) Freebase™

</m/0jcx,



/m/04m8,

Place of birth

/m/019xz9>



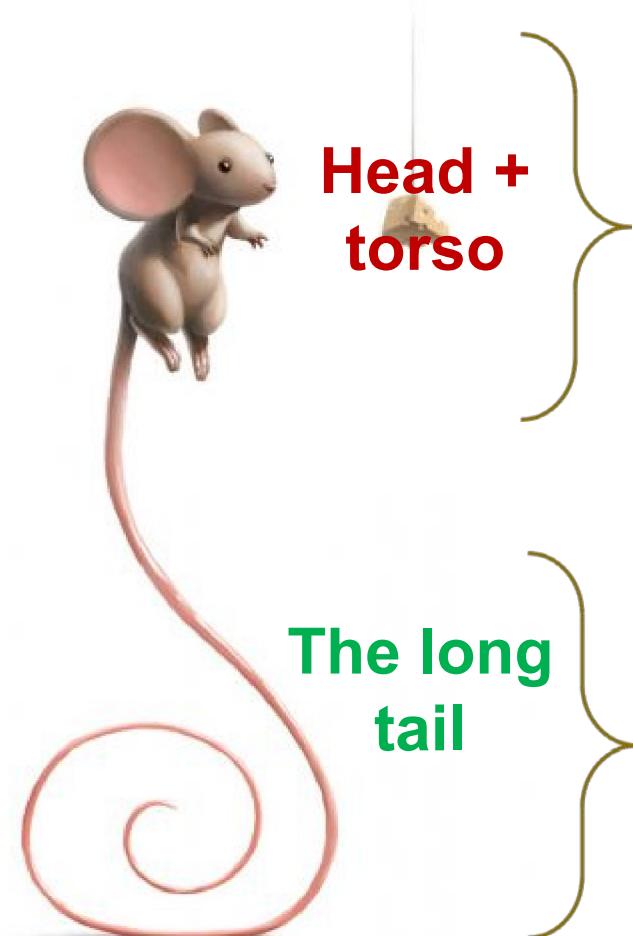
/en/albert_einstein

Albert Einstein

/en/ulm

Ulm

Google: AUTOMATIC KNOWLEDGE ACQUISITION

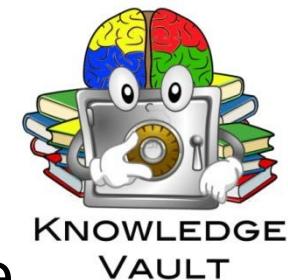


Relying on humans  **Freebase™**

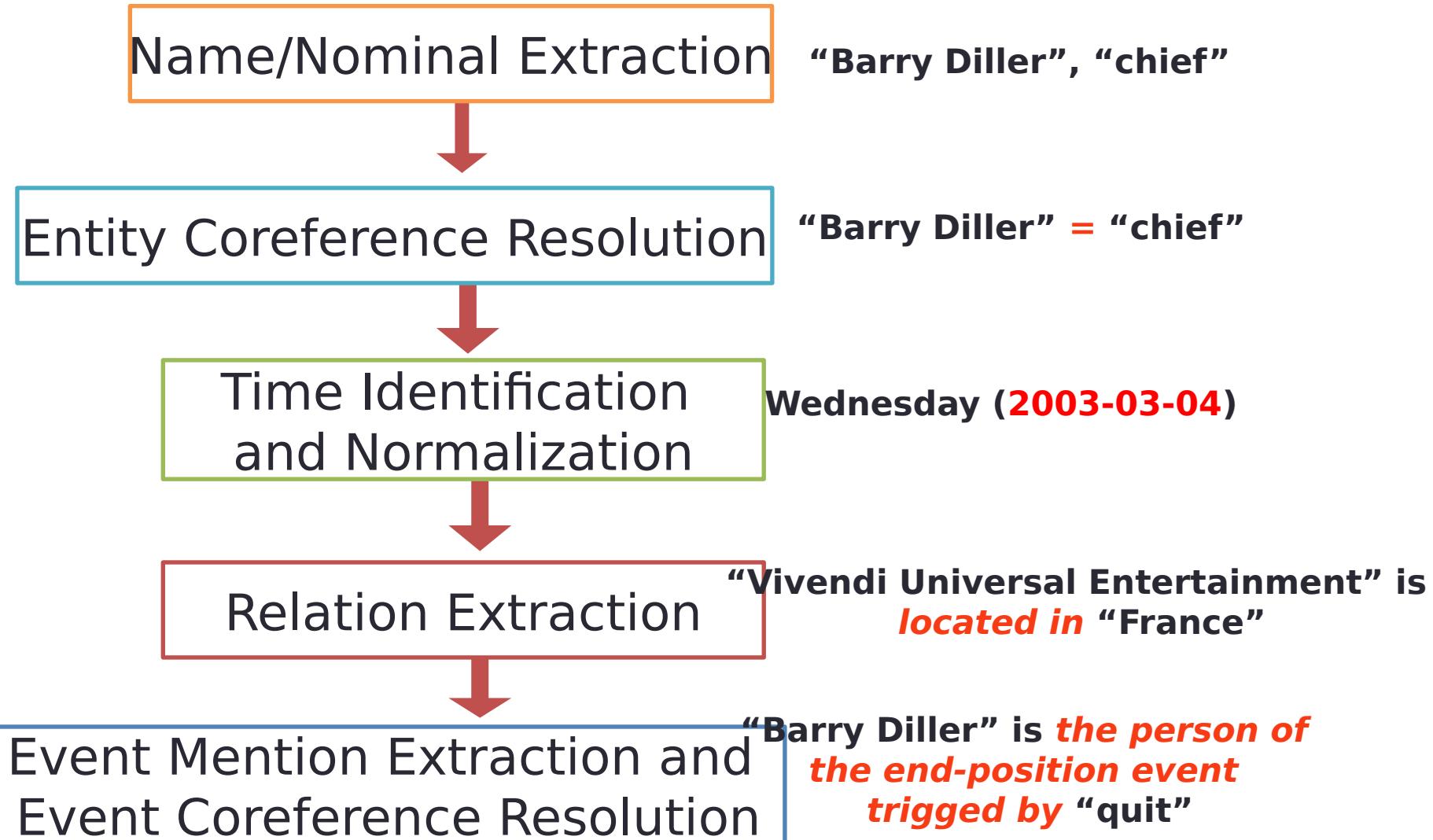
- Volunteer contributions at Freebase.com
- Import of large datasets (e.g., EDGAR, MusicBrainz)

Automatic extraction

- Extraction from web page
- Using an existing knowledge base !



Traditional IE Components



Knowledge Fusion



Freebase™

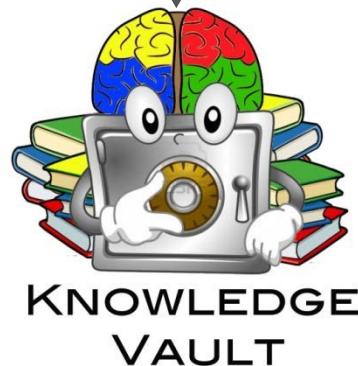


Extractors

Graph priors

Fusion

Cf. Knowledge Graph
Identification
[Pujara et al., ISWC '13]





Extract named entities and their types in the following text: John Winston Ono Lennon[nb 1] (born John Winston Lennon; 9 October 1940 – 8 December 1980) was an English singer, songwriter, musician and peace activist who achieved worldwide fame as founder, co-songwriter, co-lead vocalist and rhythm guitarist of the Beatles. Lennon's work was characterised by the rebellious nature and acerbic wit of his music, writing and drawings, on film, and in interviews. His songwriting partnership with Paul McCartney remains the most successful in history.[2] Born in Liverpool, Lennon became involved in the skiffle craze as a teenager. In 1956, he formed the Quarrymen, which evolved into the Beatles in 1960. Sometimes called "the smart Beatle", he was initially the group's de facto leader, a role gradually ceded to McCartney. Through his songwriting in the Beatles, Lennon embraced a myriad of musical influences, initially writing and co-writing rock and pop-oriented hit songs in the band's early years, then later incorporating experimental elements into his compositions in the latter half of the Beatles' career as his songs became known for their increasing innovation. Lennon soon expanded his work into other media by participating in numerous films, including How I Won the War, and authoring In His Own Write and A Spaniard in the Works, both collections of nonsense writings and line drawings. Starting with "All You Need Is Love", his songs were adopted as anthems by the anti-war movement and the larger counterculture of the 1960s. In 1969, he started the Plastic Ono Band with his second wife, the multimedia artist Yoko Ono, held the two-week-long anti-war demonstration Bed-ins for Peace, and left the Beatles to embark on a solo career.



Peace, and left the Beatles to embark on a solo career.

Named entities and their types:

- John Winston Ono Lennon - PERSON
- John Winston Lennon - PERSON
- October 9, 1940 - DATE
- December 8, 1980 - DATE
- English - NORG
- the Beatles - ORG
- Paul McCartney - PERSON
- Liverpool - GPE
- Quarrymen - ORG
- How I Won the War - WORK_OF_ART
- In His Own Write - WORK_OF_ART
- A Spaniard in the Works - WORK_OF_ART
- Plastic Ono Band - ORG
- Yoko Ono - PERSON
- Bed-ins for Peace - EVENT
- 1960s - DATE

Like

Dislike

Share



Extract the relations of provided named entities in the following text:

John Winston Ono Lennon[nb 1] (born John Winston Lennon; 9 October 1940 – 8 December 1980) was an English singer, songwriter, musician and peace activist who achieved worldwide fame as founder, co-songwriter, co-lead vocalist and rhythm guitarist of the Beatles. Lennon's work was characterised by the rebellious nature and acerbic wit of his music, writing and drawings, on film, and in interviews. His songwriting partnership with Paul McCartney remains the most successful in history.[2] Born in Liverpool, Lennon became involved in the skiffle craze as a teenager. In 1956, he formed the Quarrymen, which evolved into the Beatles in 1960. Sometimes called "the smart Beatle", he was initially the group's de facto leader, a role gradually ceded to McCartney. Through his songwriting in the Beatles, Lennon embraced a myriad of musical influences, initially writing and co-writing rock and pop-oriented hit songs in the band's early years, then later incorporating experimental elements into his compositions in the latter half of the Beatles' career as his songs became known for their increasing innovation. Lennon soon expanded his work into other media by participating in numerous films, including How I Won the War, and authoring In His Own Write and A Spaniard in the Works, both collections of nonsense writings and line drawings. Starting with "All You Need Is Love", his songs were adopted as anthems by the anti-war movement and the larger counterculture of the 1960s. In 1969, he started the Plastic Ono Band with his second wife, the multimedia artist Yoko Ono, held the two-week-long anti-war demonstration Bed-ins for Peace, and left the Beatles to embark on a solo career. Named entities and their types:

John Winston Ono Lennon - PERSON

John Winston Lennon - PERSON

October 9, 1940 - DATE

December 8, 1980 - DATE

English - NORG

the Beatles - ORG

Paul McCartney - PERSON

Liverpool - GPE

Quarrymen - ORG

How I Won the War - WORK_OF_ART

In His Own Write - WORK_OF_ART

A Spaniard in the Works - WORK_OF_ART



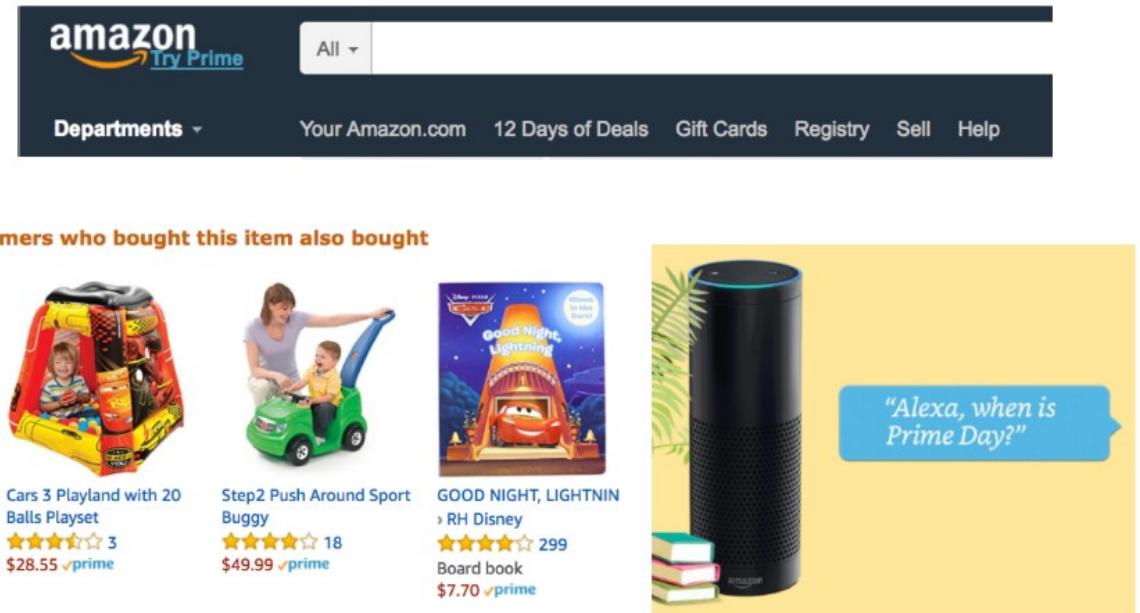
13888 - DATE

Some possible relations between the named entities in the text are:

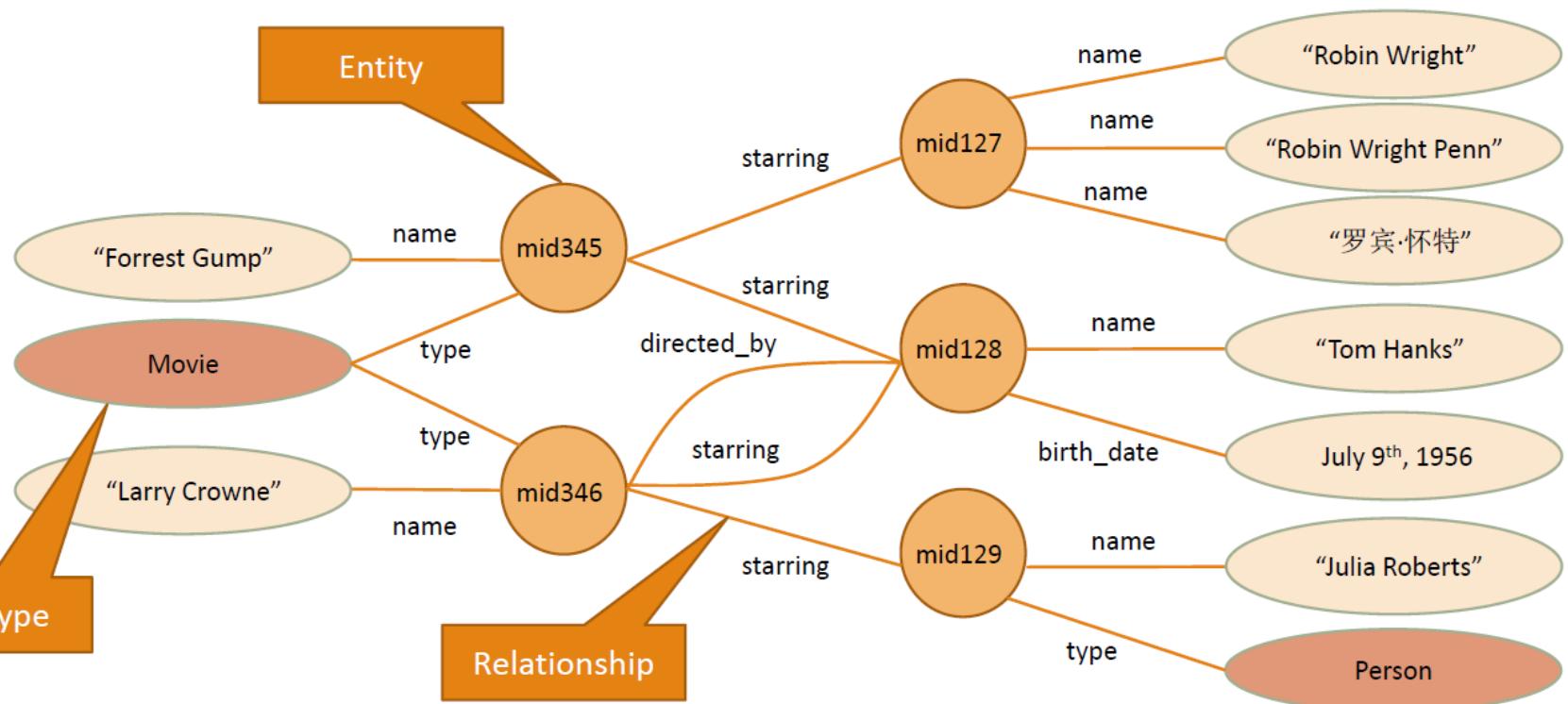
- John Winston Ono Lennon was the founder, co-songwriter, co-lead vocalist, and rhythm guitarist of the Beatles.
- Lennon had a songwriting partnership with Paul McCartney.
- Lennon was born in Liverpool and became involved in the skiffle craze there as a teenager.
- Lennon formed the Quarrymen, which evolved into the Beatles in 1960.
- Lennon embraced a myriad of musical influences through his songwriting in the Beatles, and his songs became known for their increasing innovation.
- Lennon participated in numerous films, including How I Won the War, and authored In His Own Write and A Spaniard in the Works.
- Lennon's songs, starting with "All You Need Is Love," were adopted as anthems by the anti-war movement and the larger counterculture of the 1960s.
- Lennon started the Plastic Ono Band with his second wife, Yoko Ono, and held the two-week-long anti-war demonstration Bed-ins for Peace.
- Lennon left the Beatles to embark on a solo career.

Amazon Product Graph

- Mission: To answer any question about products and related knowledge in the world

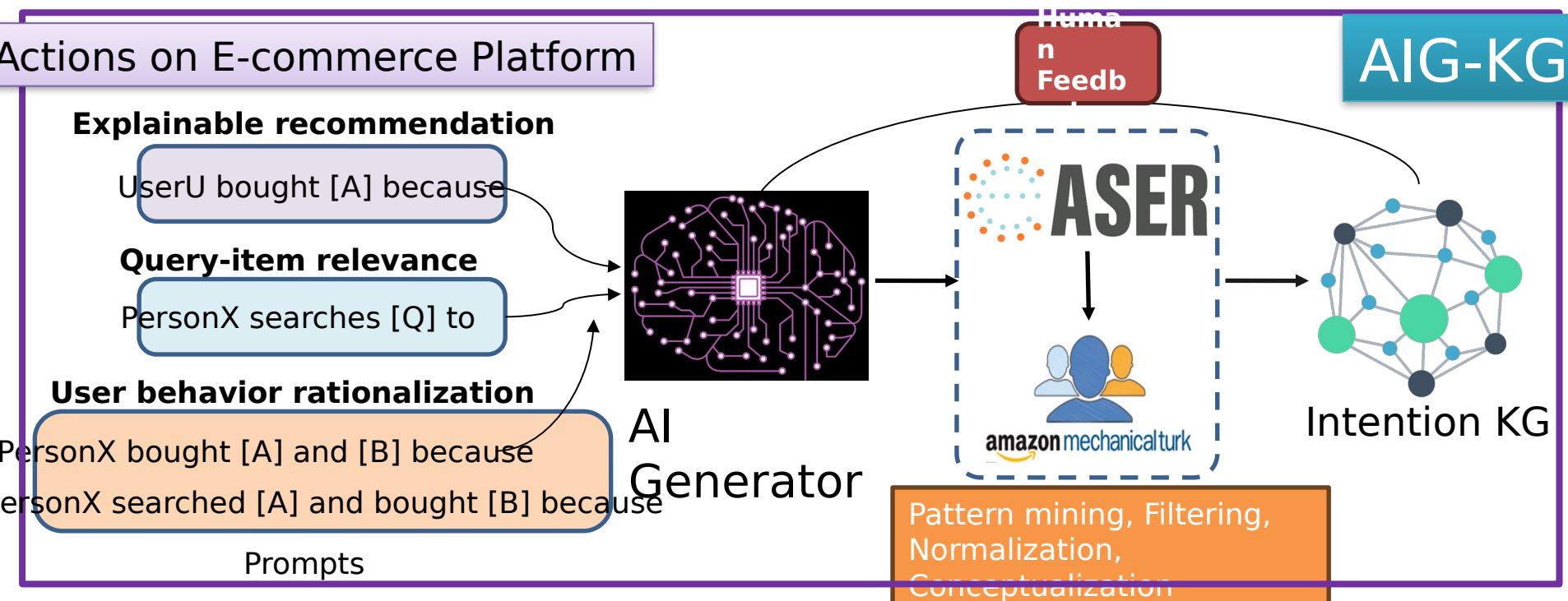


Knowledge Graph Example for 2 Movies

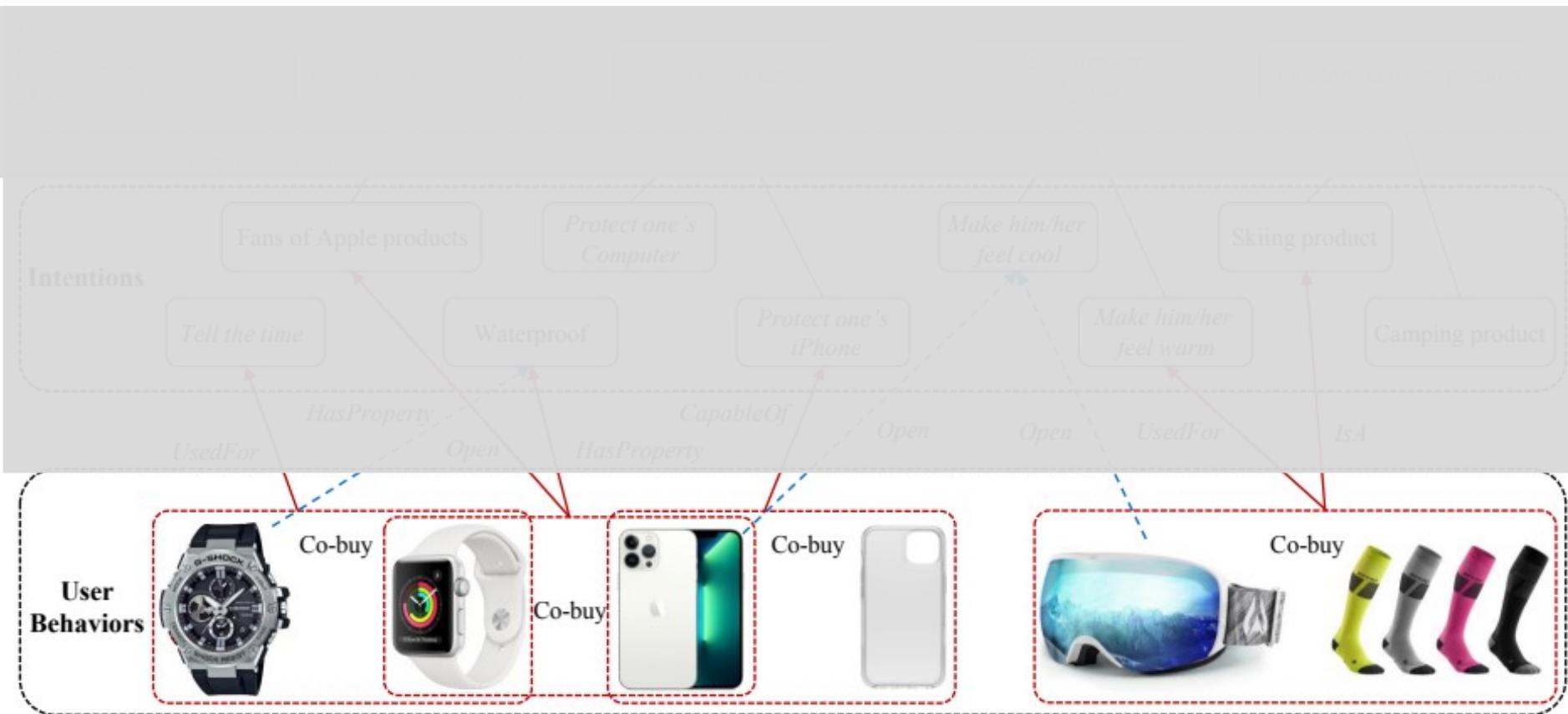


FolkScope and COSMO: Collective Intention KGs

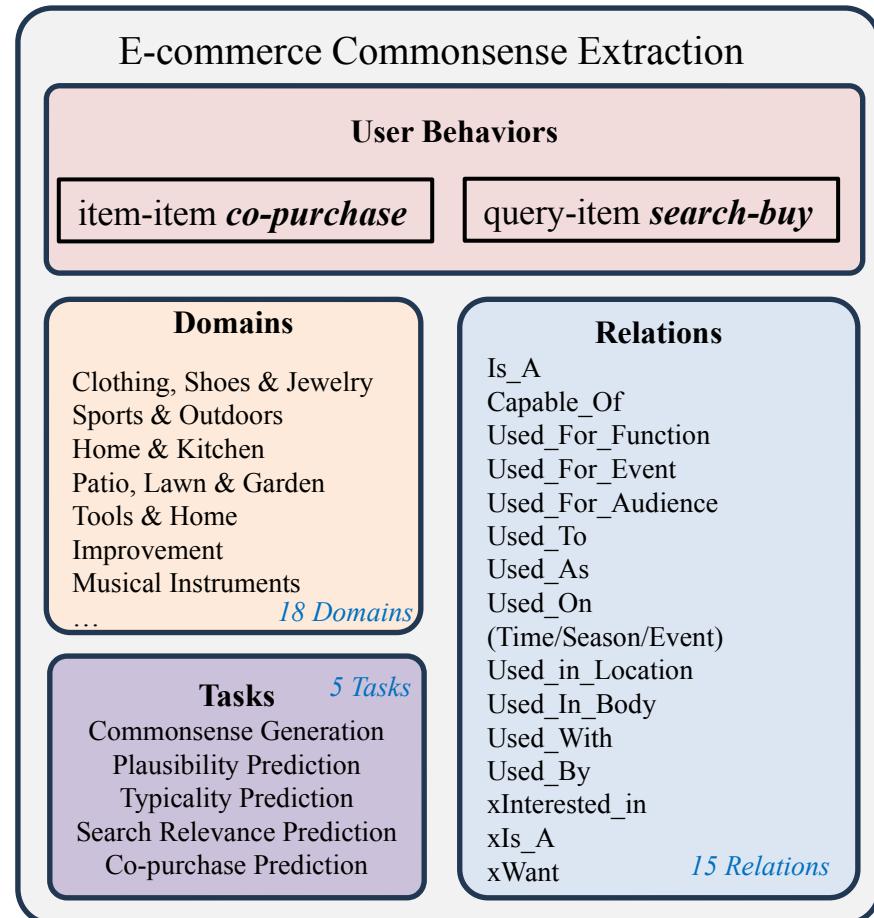
- AI generated knowledge graph construction framework



FolkScope and COSMO: Collective Intention KGs



COSMO: Collective Intention KG for Search-Buys



So here we focus on
grounding/linking

Entity Linking

- Motivation and Definition
- A Skeletal View of a Wikification System
 - High Level Algorithmic Approach

Organizing knowledge

It's a version of **Chicago**
- the standard classic
Macintosh menu font,
with that distinctive thick
diagonal in the "N".

Chicago was used by
default for Mac menus
through MacOS 7.6,
and OS 8 was released
mid-1997..

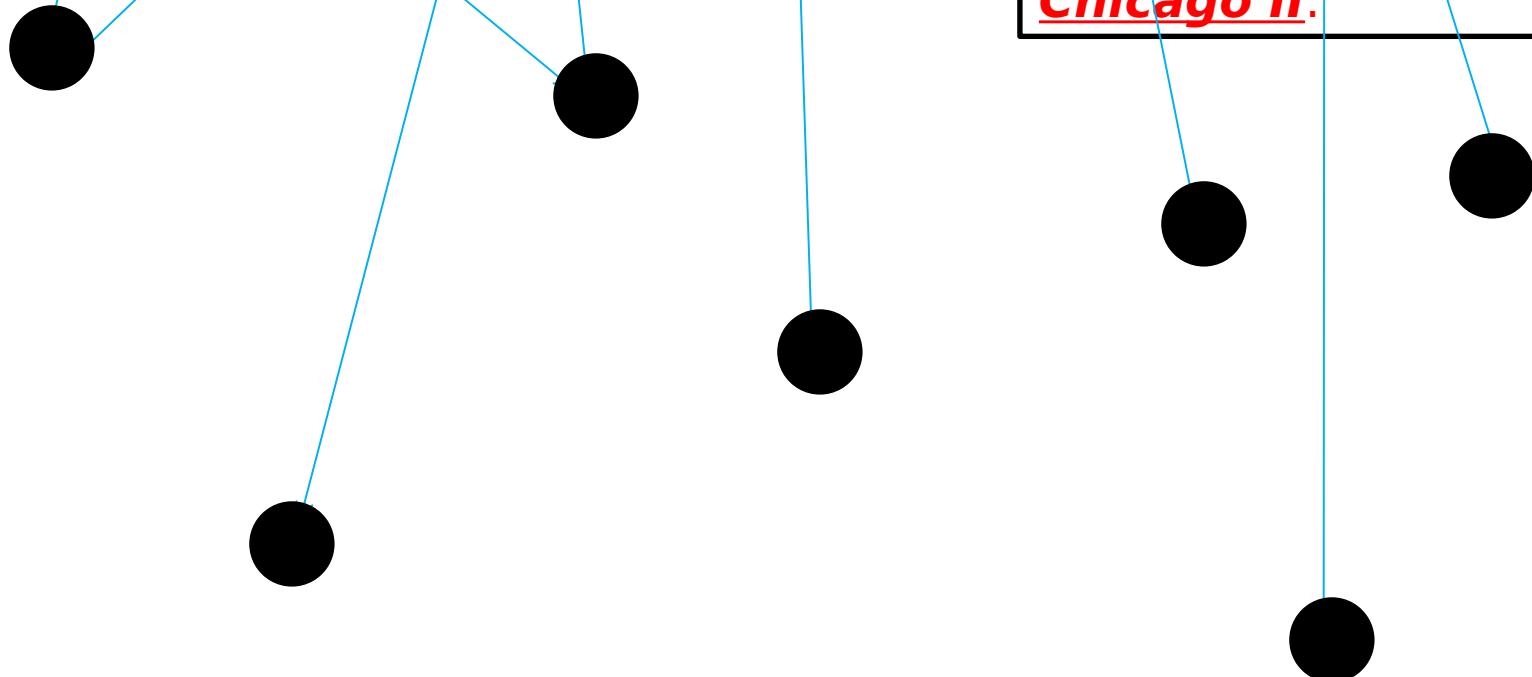
Chicago VIII was one
of the early 70s-era
Chicago albums to
catch my
ear, along with
Chicago II.

Cross-document co-reference resolution

It's a version of ***Chicago***
- the standard classic
Macintosh menu font,
with that distinctive thick
diagonal in the "N".

Chicago was used by
default for ***Mac*** menus
through ***MacOS 7.6***,
and ***OS 8*** was released
mid-1997..

Chicago VIII was one
of the early 70s-era
Chicago albums to
catch my
ear, along with
Chicago II.



Reference resolution: (disambiguation to Wikipedia)

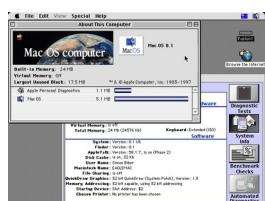
It's a version of ***Chicago*** - the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".



Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..



Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.

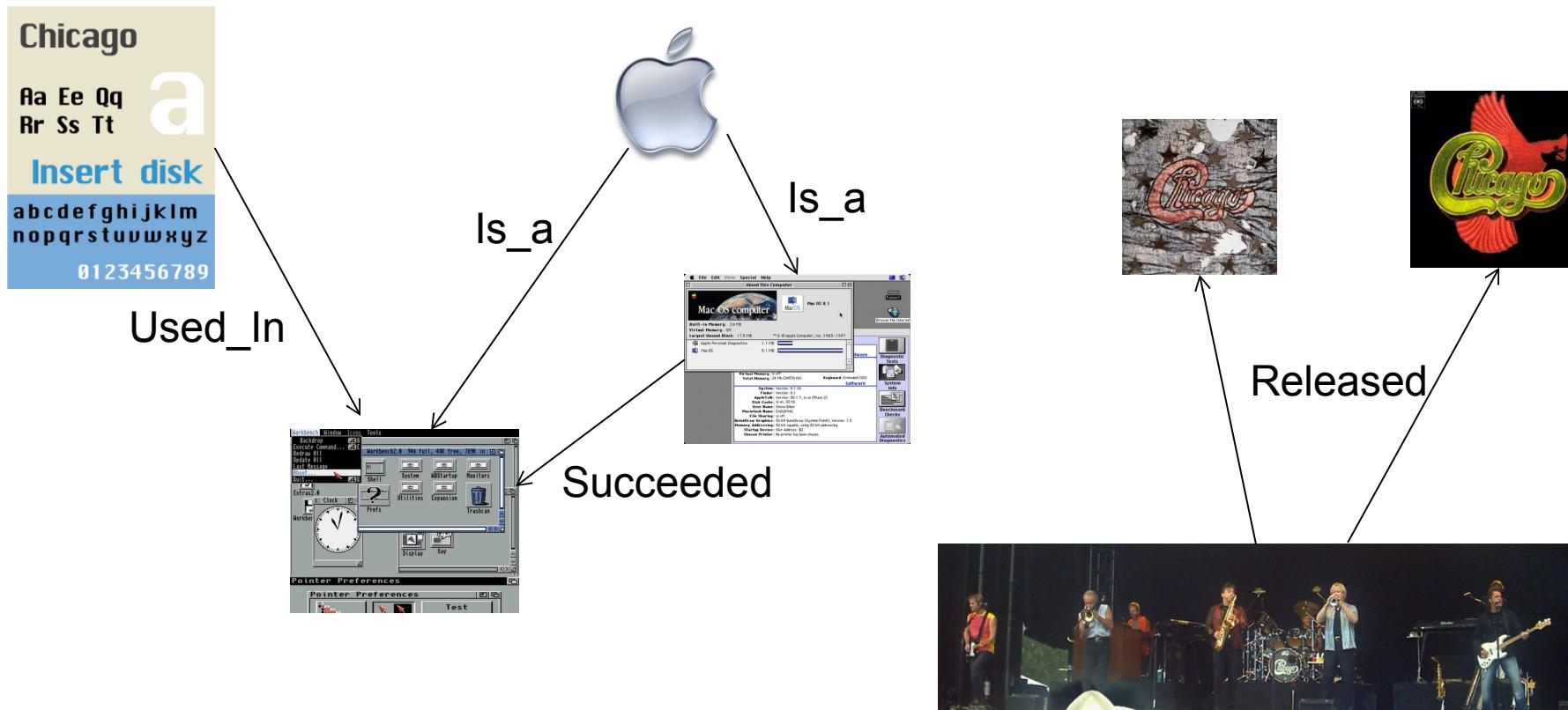


The “Reference” Collection has Structure

It's a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the "N".

Chicago was used by default for **Mac** menus through **MacOS 7.6**, and **OS 8** was released mid-1997..

Chicago VIII was one of the early 70s-era **Chicago** albums to catch my ear, along with **Chicago II**.

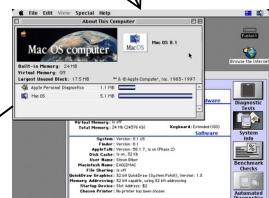


Analysis of Information Networks

It's a version of ***Chicago*** - the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".



Chicago was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..



Chicago VIII was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago II***.



Here - Wikipedia as a knowledge resource but we can use other resources



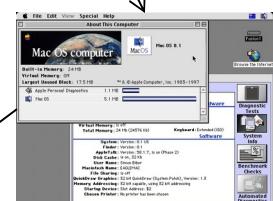
Used_In



ls_a



ls_a



Succeeded



Released



Navigating Unfamiliar Domains

The screenshot shows a web interface for the "Wikifier Demo" from the Cognitive Computation Group. At the top, there is a navigation bar with "Cognitive Computation Group" followed by arrows pointing to "Demos" and "Wikifier". Below the navigation is a logo featuring a stylized orange letter "I" inside a blue globe-like shape. To the right of the logo is the text "Wikifier Demo". A horizontal slider is positioned next to the text "fewer concepts" and "more concepts", with the slider currently set towards "fewer concepts". Below the slider are two buttons: an orange "wikify!" button with a magnifying glass icon and a blue "clear" button with a cross icon. A note at the bottom of the main content area reads: "* If you wish to cite this work, please cite the following publications: (1) Retinov et. al. and (2) Cheng and Roth." The main content area contains a detailed paragraph about Human Immunodeficiency Virus (HIV). The text discusses the construction of a fusion protein between the gp41 membrane-proximal external region (MPER) and a variable-length Gly4Ser linker (where x is 4 or 8), expressed in BL21(DE3)pLysS cells and purified by metal affinity chromatography followed by gel filtration. It notes the nanomolar efficacy in blocking BaL-pseudotyped HIV-1 infection of HOS.T4.R5 cells, its specificity against HIV-1, and its ability to release intraviral p24 protein from host cells. The text also mentions the chimera's requirement for both MPER and CVN components for virolytic effect and its potential to destabilize the virus and lead to inactivation.

Human immunodeficiency virus (HIV) is the primary **etiological** agent responsible for the AIDS pandemic. We constructed a **fusion** of the **gp41** membrane-proximal external region (MPER) peptide along with a variable-length (Gly4Ser) x linker (where x is 4 or 8) between the **C terminus** of the former and **N terminus** of the latter. The **His-tagged** recombinant proteins, expressed in **BL21(DE3)pLysS** cells and purified by immobilized metal affinity chromatography followed by **gel filtration**, were found to display a nanomolar efficacy in blocking BaL-pseudotyped **HIV-1 infection** of **HOS.T4.R5 cells**. This antiviral activity was **HIV-1** specific, since it did not inhibit **cell infection** by **vesicular stomatitis virus (VSV)**. The **chimeric proteins** were found to release intraviral p24 protein from both BaL-pseudotyped **HIV-1** and fully infectious BaL **HIV-1** in a dose-dependent manner in the absence of **host cells**. The addition of either **MPER** or **CVN** was found to outcompete this virolytic effect, indicating that both components of the **chimera** are required for virolysis. The **finding** that engaging the **Env protein** spike and membrane using a **chimeric ligand** can destabilize the **virus** and lead to inactivation opens up a means to investigate **virus particle metastability** and to evaluate this approach for inactivation at the earliest stages of exposure to **virus** and before **host cell** encounter.

Navigating Unfamiliar Domains

Chimeric Cyano Engineered Protein HIV-1

Mark Contarino^a, Arangam Ramalingam Venkat Kar^b, Varshini Gangupomu^d, I

+ Author Affiliations

ABSTRACT

Human immunodeficiency virus (HIV) is the causative agent of the AIDS pandemic. In this study, we have used a chimeric cyano engineered protein (MPER) peptide along with a recombinant protein to test the possibility of simultaneous binding between the C terminus of the MPER peptide and the C terminus of recombinant proteins, expressed in E. coli. We used immobilized metal affinity chromatography (IMAC) to display a nanomolar concentration of the chimeric protein on HOS.T4.R5 cells. This article describes the cell infection by vesicular stomatitis virus (VSV). Importantly, the chimeric protein from both BaL-1 and CVN was found to outcompete the components of the chimeric protein spike and lead to inactivation of the virus and to evaluate the metastability and to evaluate the exposure to virus and b

Cognitive Computation Group
Wikifier Demo

wikify! clear

* If you wish to cite this work, please

Human immunodeficiency virus (HIV) is the causative agent of the AIDS pandemic. We constructed a chimeric protein consisting of the MPER peptide along with a recombinant protein. We used immobilized metal affinity chromatography (IMAC) to display a nanomolar concentration of the chimeric protein on HOS.T4.R5 cells. This article describes the cell infection by vesicular stomatitis virus (VSV). Importantly, the chimeric protein from both BaL-1 and CVN was found to outcompete the components of the chimeric protein spike and lead to inactivation of the virus and to evaluate the metastability and to evaluate the exposure to virus and b

Fusion protein
From Wikipedia, the free encyclopedia

This article is about chimeric fusion proteins. For proteins involved in membrane fusion, see [membrane fusion protein](#). For proteins involved in gene fusion, see [fusion gene](#).

Gp41
From Wikipedia, the free encyclopedia

Gp41 also known as glycoprotein 41 is a subunit of the envelope protein complex of retroviruses. Human immunodeficiency virus (HIV). Gp41 is a transmembrane protein that contains several site-directed antibodies that are required for infection of host cells.

Affinity chromatography
From Wikipedia, the free encyclopedia

Affinity chromatography is a method of separating biochemical mixtures based on antibody, enzyme and substrate or receptor and ligand.

Educational Applications: Unfamiliar domains may contain terms unknown to a reader. The Wikifier can supply the necessary background knowledge even when the relevant article titles are not identical to what appears in the text, dealing with both ambiguity and variability.

Task Definition

- A formal definition of the task consists of:
 1. A definition of the **mentions** (concepts, entities) to highlight
 2. Determining the target encyclopedic resource (**KB**)
 3. Defining what to point to in the KB (**title**)

1. Mentions

- A **mention**: a phrase used to refer to something in the world
 - Named entity (person, organization), object, substance, event, philosophy, mental state, rule ...
- Task definitions vary across the definition of **mentions**
 - All N-grams (up to a certain size); Dictionary-based selection; Data-driven controlled vocabulary (e.g., all Wikipedia titles); **only named entities**.
- Ideally, one would like to have a mention definition that **adapts** to the application/user

Examples of Mentions (1)

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.



Examples of Mentions (2)



The screenshot shows the Wikifier Demo interface from the Cognitive Computation Group. The top navigation bar includes links to 'Demos' and 'Wikifier'. Below the navigation is a logo featuring a blue globe with a white 'I' and the text 'Wikifier Demo'. A slider is present with the labels 'fewer concepts' on the left and 'more concepts' on the right. At the bottom of the interface are two buttons: 'wikify!' (orange) and 'clear' (blue). The main content area displays a text snippet about Alex Smith. Several words in the text are highlighted with red circles: 'Alex Smith', 'offseason', 'turnover', and 'feet'. The text itself discusses the Chiefs trading for Alex Smith, his role as a game manager, and his performance in the season.

* If you wish to cite this work, please reference the following publications: (1) R. Levy et. al. and (2) Cheng and Roth.

The **Chiefs** didn't trade for **Alex Smith** this season because they wanted a smart game manager who wouldn't kill their offense. They acquired him because they needed a **quarterback** who knows how to **turnover**. It requires him to do what he's done all season: throw the ball well, make the key play when necessary and keep the chains moving when his arm can't get the job done. These days it means **Smith** has to show people more of what he revealed in Sunday's 41-38 loss to **San Diego** -- that he can elevate his game when his team is in dire straits.

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

Examples of Mentions (3)

Perhaps the definition of which **mentions** to highlight should depend on the **expertise** and **interests** of the users?

The screenshot shows a web-based application titled "Wikifier Demo" from the "Cognitive Computation Group". The interface includes a navigation bar with "Demos" and "Wikifier", a search bar with a magnifying glass icon, and buttons for "wikify!" and "clear". A slider allows users to choose between "fewer concepts" and "more concepts". Below the search bar, a note says: "* If you wish to cite this work, please cite the following publications: (1) Retinov et. al. and (2) Cheng and Roth." The main content area displays a scientific text about HIV. Three specific words are highlighted with red ovals: "HIV", "gp41", and "virus". The text discusses the construction of chimeric proteins by fusing the gp41 membrane proximal external region peptide of HIV with the C-terminus of MPER, and how these chimeric proteins inhibit HIV infection.

Human immunodeficiency virus (HIV) is the agent responsible for the epidemic. We constructed chimeric proteins by fusing the membrane proximal external region peptide along with a variable-length (Gly_{45er})_x linker (where x is 4 or 8) to the C terminus of the former and N terminus of the latter. The His-tagged chimeric proteins, expressed in BL21(DE3)pLysS cells and purified by immobilized metal ion chromatography followed by gel filtration, were found to display a similar efficacy in blocking BaL-pseudotyped HIV-1 infection of HOS.T4.R5 cells. This activity was HIV-1 specific and it did not inhibit cell infection by vesicular stomatitis virus (VSV). It was also found to release intraviral p24 protein from BaL-pseudotyped HIV-1 and many infectious BaL HIV-1 in a dose-dependent manner in the absence of host cells. The addition of either MPER or CVN was found to outcompete this virolytic effect, indicating that both components of the chimera are required for virolysis. The finding that engaging the Env protein spike and membrane using a chimeric ligand can destabilize the virus and lead to inactivation opens up a means to investigate virus particle instability and to evaluate this approach for inactivation at the earliest stages of infection before host cell encounter.

2. Concept Inventory (KB)

- Multiple KBs can be used, in principle, as the target KB.
- Wikipedia has the advantage of a broad coverage, regularly maintained KB, with significant amount of text associated with each title.
 - All type of pages?
 - Content pages
 - Disambiguation pages
 - List pages
- What should happen to mentions that do not have entries in the target KB?



3. Null Links

- Often, there are multiple sensible links.

Dorothy Byrne, a state coordinator for the Florida Green Party,....

- How to capture the fact that Dorothy Byrne does not refer to any concept in Wikipedia?
- **Wikification:** Simply map Dorothy Byrne ↲ Null
- **Entity Linking:** If multiple mentions in the given document(s) correspond to the same concept, which is outside KB
 - First cluster relevant mentions as representing a single concept
 - Map the cluster to Null

Naming Convention

- **Wikification:**
 - Map Mentions to KB Titles (particularly, Wikipedia)
 - Map Mentions that are not in the KB to NIL
- **Entity Linking:**
 - Map Mentions to KB Titles
 - If multiple mentions correspond to the same Title, which is outside KB:
 - First **cluster relevant mentions** as representing a single Title
 - Map the cluster to **Null**
- If the set of target mentions only consists of **named entities** we call the task: **Named Entity [Wikification, Linking]**

Outline

- Motivation and Definition
- A Skeletal View of a Wikification System
 - High Level Algorithmic Approach

Wikification: Subtasks

- Wikification and Entity Linking requires addressing several sub-tasks:
 - Identifying Target Mentions
 - Mentions in the input text that should be Wikified
 - Identifying Candidate Titles
 - Candidate Wikipedia titles that could correspond to each mention
 - Candidate Title Ranking
 - Rank the candidate titles for a given mention
 - NIL Detection and Clustering
 - Identify mentions that do not correspond to a Wikipedia title
 - Entity Linking: cluster NIL mentions that represent the same entity.

High-level Algorithmic Approach.

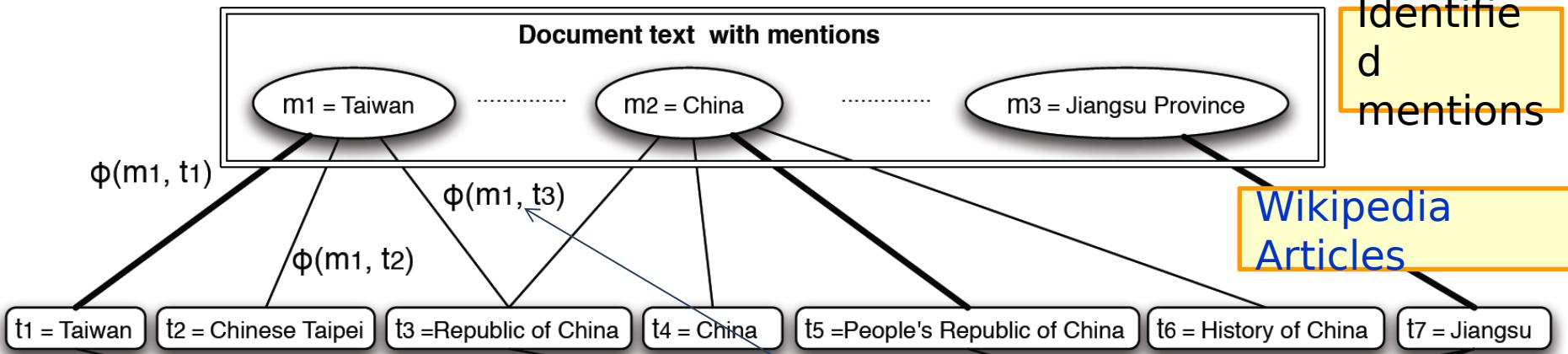
- **Input:** A text document d ; **Output:** a set of pairs (m_i, t_i)
 - m_i are mentions in d ; $t_j(m_i)$ are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
 - For each m_i in d :
 - Identify a set of relevant titles $T(m_i)$
 - Rank titles $t_i \in T(m_i)$
[E.g., consider local statistics of edges (m_i, t_i) , $(m_i, *)$, and $(*, t_i)$ occurrences in the Wikipedia graph]
- (3) Global Inference
 - For each document d :
 - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
 - Re-rank titles $t_i \in T(m_i)$
[E.g., if m, m' are related by virtue of being in d , their corresponding titles t, t' may also be related]

Local approach

A text Document

Identified mentions

Wikipedia Articles

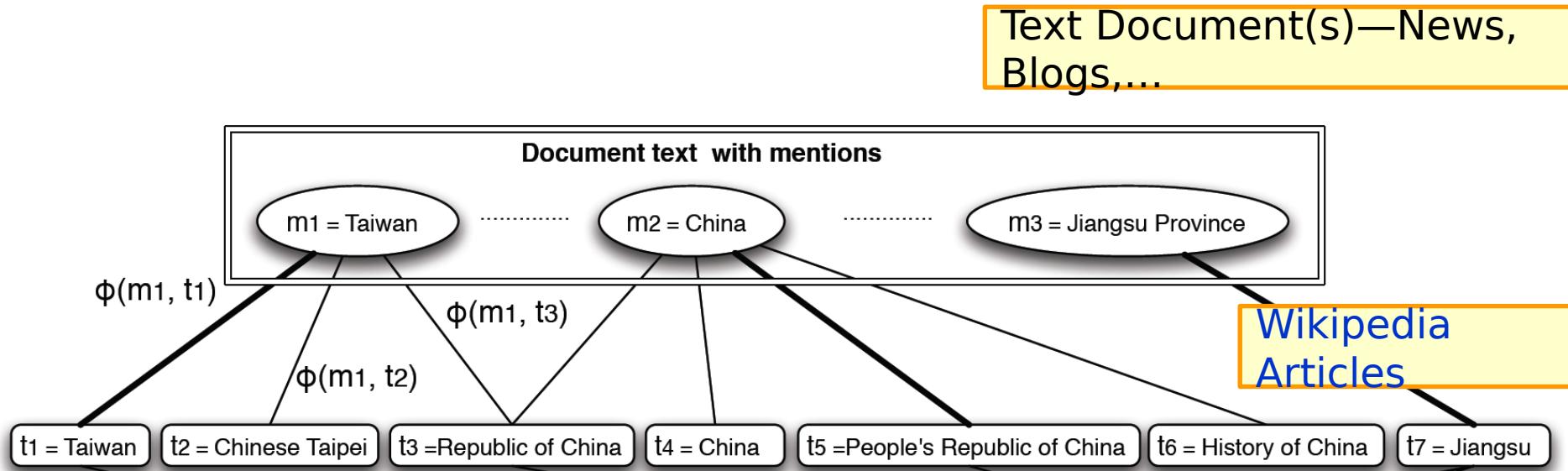


- Γ is a solution to the problem
 - A set of pairs (m, t)
 - m : a mention in the document
 - t : the matched Wikipedia Title

Local score of matching the mention to the title (decomposed by m_i)

$$\Gamma_{\text{local}}^* = \arg \max_{\Gamma} \sum_{i=1}^N \phi(m_i, t_i) \quad (1)$$

Global Approach: Using Additional Structure



$$\Gamma^* \approx \arg \max_{\Gamma} \sum_{i=1}^N [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)]$$

- Adding a “global” term to evaluate how good the **structure** of the solution is.
- Use the local solutions Γ' (each mention considered independently).
 - Evaluate the structure based on pairwise coherence scores $\Psi(t_i, t_j)$
 - Choose those that satisfy **document** coherence conditions.

High-level Algorithmic Approach

- **Input:** A text document d ; **Output:** a set of pairs (m_i, t_i)
 - m_i are mentions in d ; t_i are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
 - For each m_i in d :
 - Identify a set of relevant titles $T(m_i)$
 - Rank titles $t_i \in T(m_i)$

[E.g., consider local statistics of edges (m_i, t_i) , $(m_i, *)$, and $(*, t_i)$ occurrences in the Wikipedia graph]
- (3) Global Inference
 - For each document d :
 - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
 - Re-rank titles $t_i \in T(m_i)$

[E.g., if m, m' are related by virtue of being in d , their corresponding titles t, t' should also be related]

Mention Identification

- Highest recall: Each n-gram is a potential concept mention
 - Intractable for larger documents
- Surface form based filtering
 - Shallow parsing (especially NP chunks), NP's augmented with surrounding tokens, capitalized words
 - Remove: single characters, “stop words”, punctuation, etc.
- Classification and statistics based filtering
 - Name tagging (Finkel et al., 2005; Ratinov and Roth, 2009; Li et al., 2012)
 - Mention extraction (Florian et al., 2006, Li and Ji, 2014)
 - Key phrase extraction, independence tests (Mihalcea and Csomai, 2007), common word removal (Mendes et al., 2012;)

High-level Algorithmic Approach

- **Input:** A text document d ; **Output:** a set of pairs (m_i, t_i)
 - m_i are mentions in d ; t_i are corresponding Wikipedia titles, or NIL.
- **→ (1) Identify mentions m_i in d**
- **(2) Local Inference**
 - For each m_i in d :
 - Identify a set of relevant titles $T(m_i)$
 - Rank titles $t_i \in T(m_i)$
[E.g., consider local statistics of edges (m_i, t_i) , $(m_i, *)$, and $(*, t_i)$ occurrences in the Wikipedia graph]
- **(3) Global Inference**
 - For each document d :
 - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
 - Re-rank titles $t_i \in T(m_i)$
[E.g., if m, m' are related by virtue of being in d , their corresponding titles t, t' should also be related]

Generating Candidate Titles

- 1. Based on canonical names (e.g. Wikipedia page title)
 - Titles that are a super or substring of the mention
 - Michael Jordan is a candidate for “Jordan”
 - Titles that overlap with the mention
 - “William Jefferson Clinton” ↳ Bill Clinton;
 - “non-alcoholic drink” ↳ Soft Drink
- 2. Based on previously attested references
 - All Titles ever referred to by a given string in training data
 - Using, e.g., Wikipedia-internal hyperlink index
 - More Comprehensive Cross-lingual resource (Spitkovsky & Chang, 2012)

Initial Ranking of Candidate Titles

- Initially rank titles according to...
 - Wikipedia article length
 - Incoming Wikipedia Links (from other titles)
 - Number of people or the largest area (for geo-location titles)
- More sophisticated measures of prominence
 - Prior link probability
 - Graph based methods

P(t|m): “Commonness”

$$\text{Commonness}(m \Rightarrow t) = \frac{\text{count}(m \rightarrow t)}{\sum_{t' \in W} \text{count}(m \rightarrow t')}$$



Typography

By default, a font called [Charcoal](#) is used to replace the similar [Chicago](#) typeface. Additional system fonts are also provided including [Capitals](#), [Gadget](#), [Sand](#), [Terminal](#). An operating system need to be provided, such as the [Command key](#) symbol, . I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States ([Boston](#), [Chicago](#), [Minneapolis](#), [New York](#), [Orlando](#), [Seattle](#), and [Washington](#)), three in Canada ([Halifax](#), [Toronto](#) and [Winnipeg](#)) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

P>Title|"Chicago")

The Greatest Show on Earth were a [British rock](#) band, who recorded two albums for [Harvest Records](#) in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as [Blood Sweat & Tears](#) or [Chicago](#).^[1]

$P(t|m)$: “Commonness”

Rank	t	$P(t “Chicago”)$
1	Chicago	.76
2	Chicago (band)	.041
3	Chicago (2002_film)	.022
20	Chicago Maroons Football	.00186
100	1985 Chicago Whitesox Season	.00023448
505	Chicago Cougars	.0000528
999	Kimbell Art Museum	.00000586

- First used by Medelyan et al. (2008)
- Most popular method for initial candidate ranking

Note on Domain Dependence

- “Commonness” Not robust across domains

Formal Genre

Corpus	Recall
ACE	86.85%
MSNBC	88.67%
AQUAINT	97.83%
Wiki	98.59%

Ratinov et al. (2011)

Tweets

Metric	Score
P1	60.21%
R-Prec	52.71%
Recall	77.75%
MRR	70.80%
MAP	58.53%

Meij et al. (2012)

Graph Based Initial Ranking

- Centrality (Hachey et al., 2011; Hakimov et al., 2012)

$$Centrality(a) = \frac{\partial_a}{\sum_{b \in W} s(a,b)} * in_links(a) * out_links(a)$$

- ∂_a : the number of all reachable nodes from a
- $s(a,b)$: the distance between a and b (all reachable nodes)
- Importance of the title with respect to Wikipedia - Similar to PageRank (Brin & Page, 1998)
 - Hachey et al. (2011) showed that centrality works slightly better than PageRank

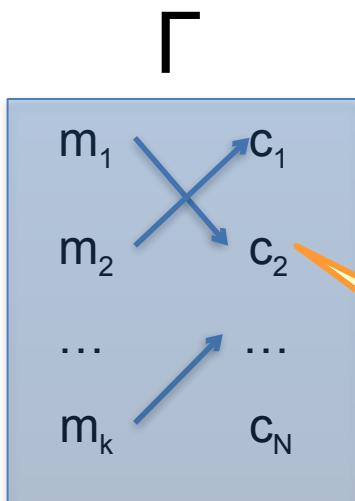
Basic Ranking Methods

- Local: Mention-Concept Context Similarity
 - Use **similarity measure** to compare the **context of the mention** with the **text associated with a candidate title** (the text in the corresponding page)
- Global: Document-wide Conceptual Coherence
 - Use topical/semantic **coherence** measures between the set of referent concepts for all mentions in a document

Context Similarity Measures

Determine assignment
that maximizes pairwise
similarity

$$\Gamma^* = \operatorname{argmax}_{\Gamma} \sum_i \varphi(m_i, t_i)$$



Mapping from
mentions to
titles

Feature vector to
capture degree of
contextual similarity

Mention, Title

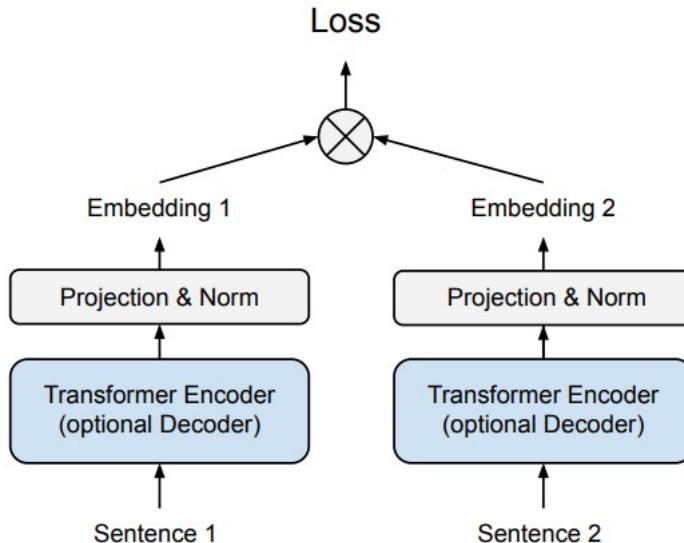
φ

Typical Features for Ranking

Mention/Concept Attribute		Description
Name	Spelling match	Exact string match, acronym match, alias match, string matching...
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical	Words in KB facts, KB text, mention name, mention text.
		Tf.idf of words and ngrams
	Position	Mention name appears early in KB text
	Genre	Genre of the mention text (newswire, blog, ...)
	Local Context	Lexical and part-of-speech tags of context words
Entity Context	Type	Mention concept type, subtype
	Relation/Event	Concepts co-occurred, attributes/relations/events with mention
	Coreference	Co-reference links between the source document and the KB text
Profiling		Slot fills of the mention, concept attributes stored in KB infobox
Concept		Ontology extracted from KB text
Topic		Topics (identity and lexical similarity) for the mention text and KB text
KB Link Mining		Attributes extracted from hyperlink graphs of the KB text
Popularity	Web	Top KB text ranked by search engine and its length
	Frequency	Frequency in KB texts

- (Ji et al., 2011; Zheng et al., 2010; Dredze et al., 2010; Anastacio et al., 2011; Zhou et al., 2014)

Sentence Embeddings



	Transfer	STS
ST5-EncDec (11B params)	90.46	84.94
ST5-Enc (11B params)	91.63	84.96
SimCSE-RoBERTa (large) (Gao et al., 2021)	90.23 ²	83.76
SBERT (large) (Reimers and Gurevych, 2019)	87.69	76.55
USE (Cer et al., 2018)	85.10	71.22
InferSent (Conneau et al., 2017)	85.59	65.01

Table 1: ST5 versus notable sentence embedding models on SentEval tasks. The reported numbers are the average of transfer tasks and STS tasks

Putting it All Together

	Score Baseline	Score Context	Score Text
Chicago_cit y	0.99	0.01	0.03
Chicago_fo nt	0.0001	0.2	0.01
Chicago_ba nd	0.001	0.001	0.02

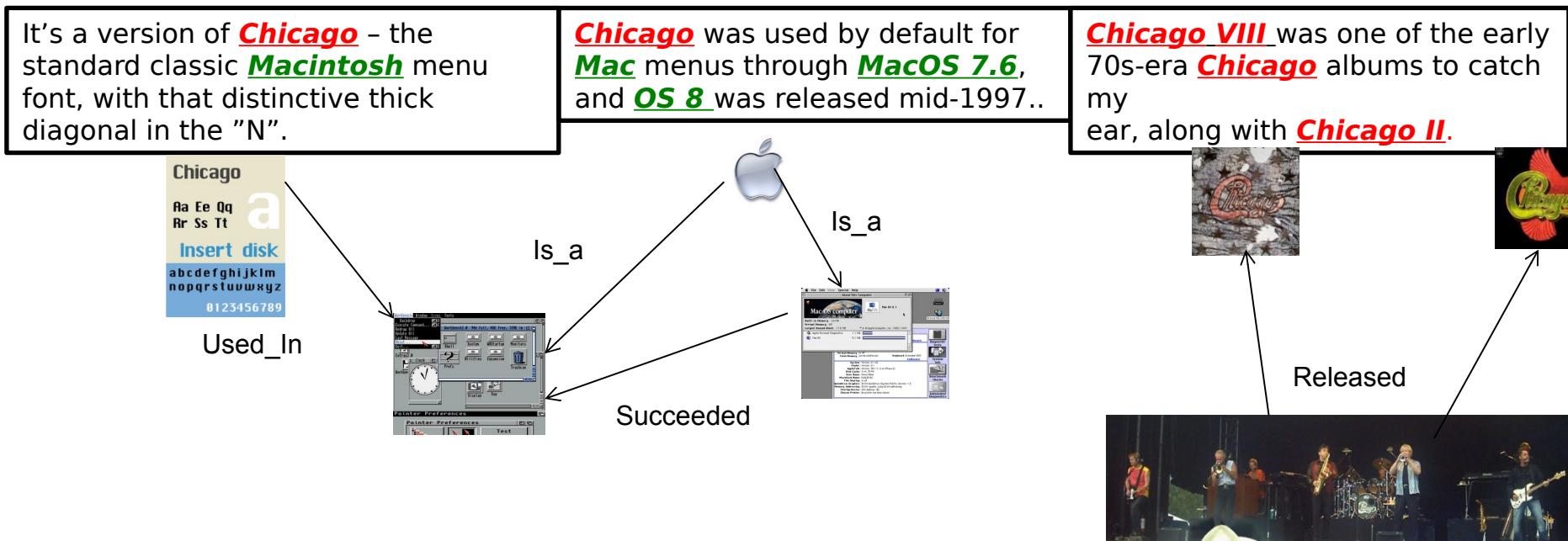
- Learning to Rank:
 - Consider all pairs of title candidates
 - Supervision is provided by Wikipedia
 - Train a ranker on the pairs (learn to prefer the correct solution)
 - A Collaborative Ranking approach: outperforms many other learning approaches (Chen and Ji, 2011)

High-level Algorithmic Approach

- **Input:** A text document d ; **Output:** a set of pairs (m_i, t_i)
 - m_i are mentions in d ; t_i are corresponding Wikipedia titles, or NIL.
- (1) Identify mentions m_i in d
- (2) Local Inference
 - For each m_i in d :
 - Identify a set of relevant titles $T(m_i)$
 - Rank titles $t_i \in T(m_i)$
- [E.g., consider local statistics of edges (m_i, t_i) , $(m_i, *)$, and $(*, t_i)$ occurrences of in the Wikipedia graph]
- (3) Global Inference
 - For each document d :
 - Consider all $m_i \in d$; and all $t_i \in T(m_i)$
 - Re-rank titles $t_i \in T(m_i)$
- [E.g., if m, m' are related by virtue of being in d , their corresponding titles t, t' should also be related]

Conceptual Coherence

- Recall: The reference collection (might) have structure.



- Hypothesis:
 - Textual co-occurrence of concepts is reflected in the KB (Wikipedia)
- Incite:
 - Preferred disambiguation Γ contains structurally coherent concepts

Co-occurrence (title 1, title 2)

Typography

By default, a font called [Charcoal](#) is used to replace the similar [Chicago](#) typeface. Additional system fonts are also provided including [Capitals](#), [Gadget](#), [Sand](#), [Terminal](#). An operating system need to be provided, such as the [Command key](#) symbol, $\ddot{\wedge}$.



Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States ([Boston](#), [Chicago](#), [Minneapolis](#), [New York](#), [Orlando](#), [Seattle](#), and [Washington](#)), three in Canada ([Halifax](#), [Toronto](#) and [Winnipeg](#)) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

The city senses of Boston and Chicago appear together often.

The Greatest Show on Earth were a British rock band, who recorded two albums for Harvest Records in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as [Blood Sweat & Tears](#) or [Chicago](#).^[1]

Co-occurrence(Title1, Title2)



Typography

By default, a font called [Charcoal](#) is used to replace the similar [Chicago](#) typeface. Additional system fonts are also provided including [Capitals](#), [Gadget](#), [Sand](#), [Terminal](#) and [Times New Roman](#). The operating system need to be provided, such as the [Command key](#) symbol, $\ddot{\wedge}$. I

Airlines and destinations

Although the population of Iceland is only about 300,000, there are scheduled flights to and from seven locations in the United States ([Boston](#), [Chicago](#), [Minneapolis](#), [New York](#), [Orlando](#), [Seattle](#), and [Washington](#)), three in Canada ([Halifax](#), [Toronto](#) and [Winnipeg](#)) and 30 cities across Europe. The largest carriers at Keflavík are Icelandair and Iceland Express.

Rock music and albums appear together often

The Greatest Show on Earth were a British [rock](#) band, who recorded two [albums](#) for [Harvest Records](#) in 1970.

The band had been conceived by Harvest Records in an attempt to create a horn-based rock combo, such as [Blood Sweat & Tears](#) or [Chicago](#).^[1]

Global Ranking

$$\Gamma^* \approx \arg \max_{\Gamma} \sum_{i=1}^N [\phi(m_i, t_i) + \sum_{t_i \in \Gamma, t_j \in \Gamma'} \psi(t_i, t_j)]$$

- How to define relatedness between two titles? (What is Ψ ?)

Title Coherence & Relatedness

- Let C, D be a pair of titles ...
- Let C and D be their sets of incoming (or outgoing) links
 - Unlabeled, directed link structure

$$\text{relatedness}(c,d) = \frac{\log(\max(|C|, |D|)) - \log(\min(|C|, |D|))}{\log(W) - \log(\min(|C|, |D|))},$$

See García et al.
(JAIR2014) for
variational
details

$$PMI(c,d) = \frac{|C \cap D| / |W|}{(|C| / |W|) * (D / |W|)}$$

$$\text{relatedness}(c,d) = \langle C, D \rangle$$

Introduced by Milne & Witten (2008)
Used by Kulkarni et al. (2009), Ratinov et al (2011), Hoffart et al (2011)

Relatedness
Outperforms Pointwise Mutual Information
(Ratinov et al., 2011)

Category based similarity introduced by Cucerzan (2007)

- Let C and D $\in \{0,1\}^K$, where K is the set of all categories

NIL Detection and Clustering

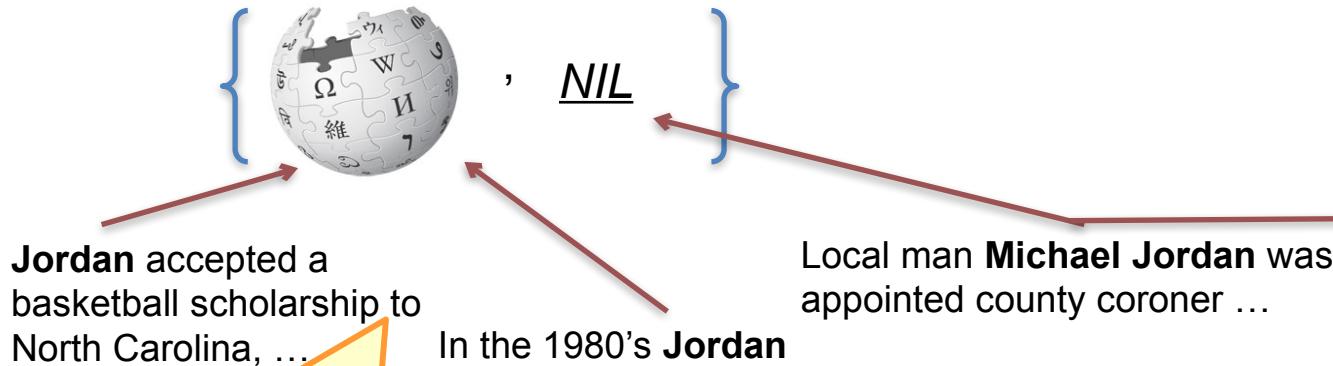
- The key difference between Wikification and Entity Linking is the way NIL are treated.
- In **Wikification**:
 - Local Processing
 - Each mention m_i that does not correspond to title t_i is mapped to NIL.
- In **Entity Linking**:
 - Global Processing
 - Cluster all mentions m_i that represent the same concept
 - If this cluster does not correspond to a title t_i , map it to NIL.
- Mapping to NIL is challenging in both cases

NIL

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

2. Include general NIL-indicating features

Is it in the KB?



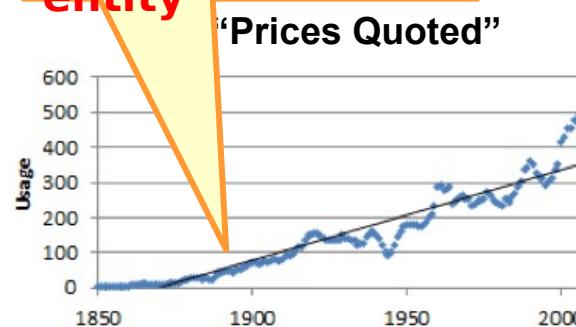
1. Binary classification (Within KB vs. NIL)

2. Select NIL cutoff by tuning confidence threshold

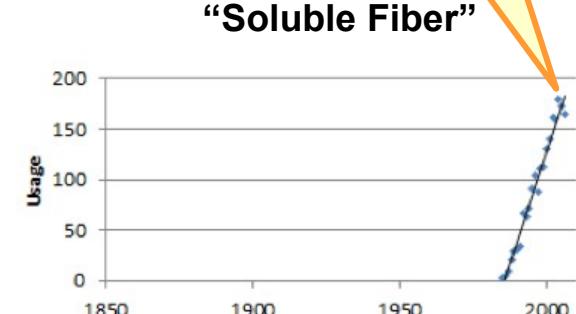
Is it an entity?

- Concept Mention Identification (above)
- Not all NP's are linkable

No spike: Not an entity



Sudden Google Books frequency spike: Entity

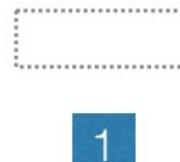
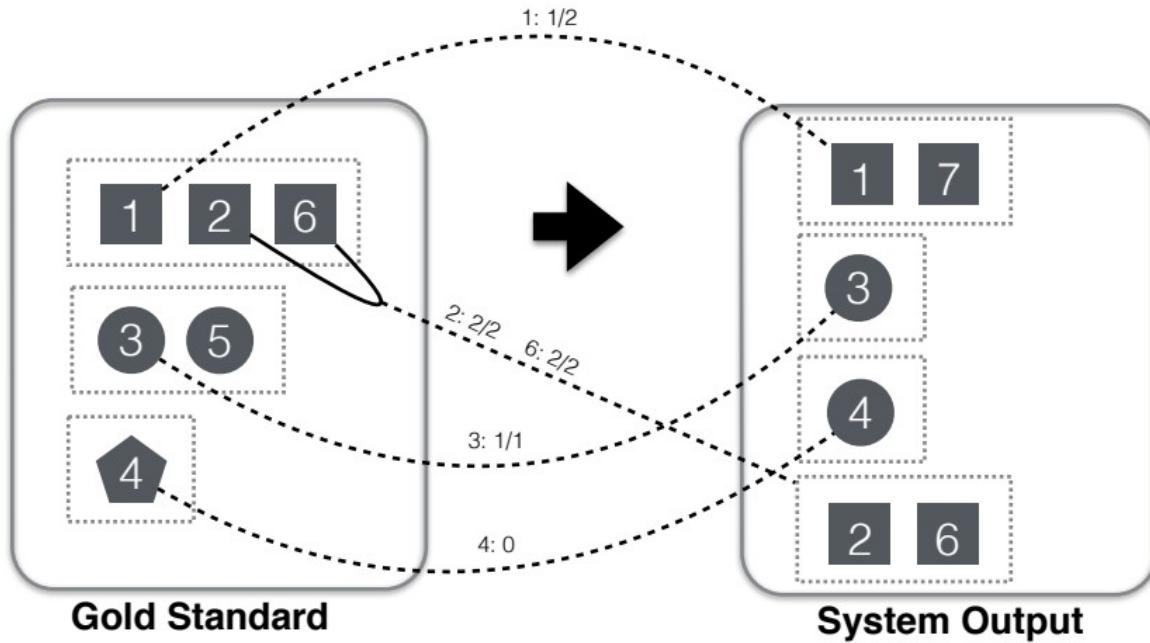


Evaluation Metrics

- Concept/Entity Extraction
 - F-Measure, Clustering
- Linking
 - Accuracy @ K (K=1, 5, 10...)
- End-to-end Concept/Entity Extraction
 - + Linking + NIL Clustering
 - B-cubed
 - CEAFF
 - Graph Edit Distance

B³: Precision

- Precision = sum mention credits / #system-output-mentions
 $= (1/2 + 2/2 + 2/2 + 1/1 + 0)/6 = 0.583$

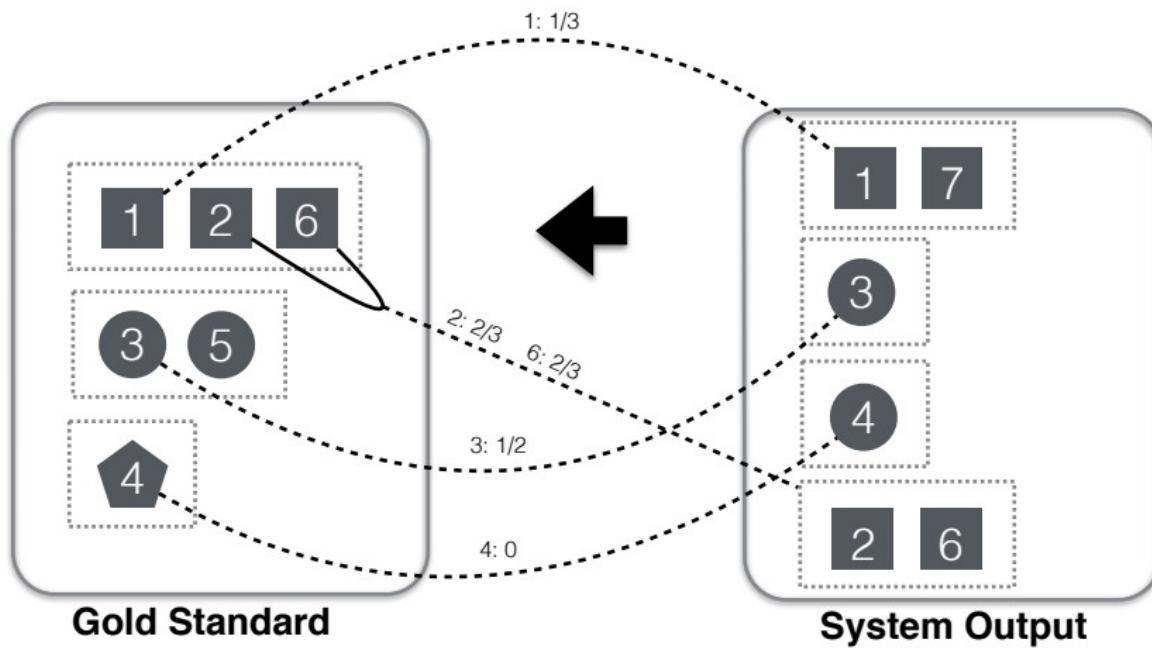


cluster mentions together

color refer to kb_id
shape refer to entity type
number refer to doc_id + offset

B³: Recall

- Recall = sum mention credits / #gold-standard-mentions
 $= (1/3 + 2/3 + 2/3 + 1/2)/6 = 0.361$



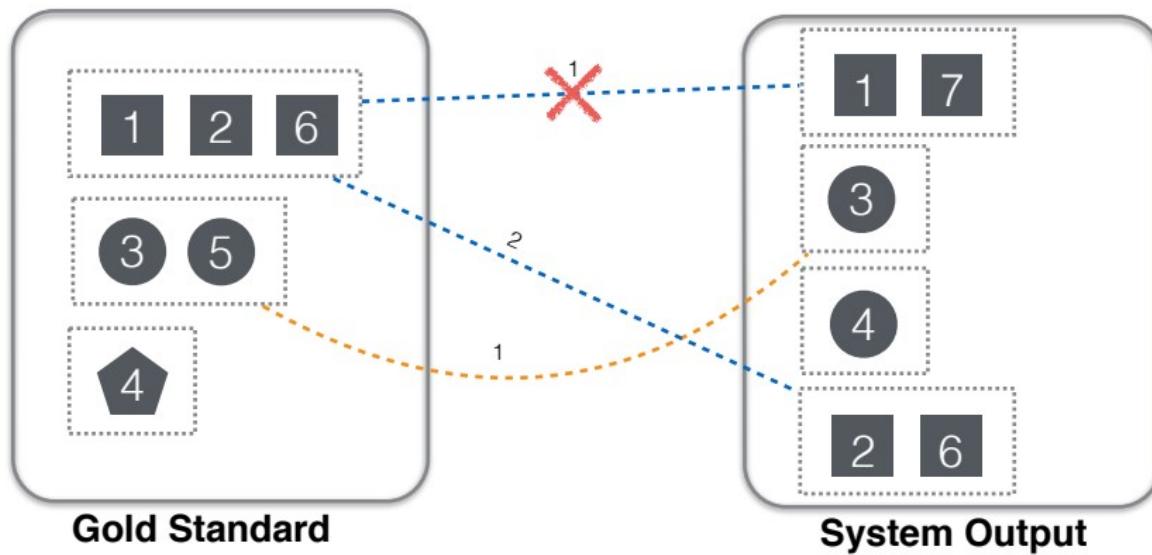
cluster mentions together

CEAF (Luo, 2005)

- Idea: a mention or entity **should not be** credited more than once
- Formulated as a bipartite matching problem
 - A special ILP problem
 - Efficient algorithm: Kuhn-Munkres

CEAFm: Example

- Solid: best 1-1 alignment
- $\phi(G_i, S_i) = |G_i \cap S_i|$
- Recall = #common / #mentions-in-key = $(2+1)/6 = 1/2$
- Precision = #common / #mentions-in-response = $(2+1)/6 = 1/2$



cluster mentions together



color refer to kb_id
shape refer to entity type
number refer to doc_id + offset