SIC Saarland Informatics Campus

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

Question Answering Systems

Named entity recognition and disambiguation

Rishiraj Saha Roy
Max Planck Institute for Informatics, Germany

4

Saarland University, Summer Semester 2020

Lecture overview

- Logistics
- Named entity basics
- NERD System 1: TAGME
- NERD System 2: AIDA

NED

Logistics: Exam

→ clarify doubts

- All 21 papers → Syllabus
- 10 – 20
- ~15 minutes per student
- ~3-5 topics → questions
- Question test your understanding of QA syss. → content covered in class
- Relative grading → Difficult

Logistics: Assignments

6 ECTS

- Self-assessment, no tutorials
  - Grades available every week
  - Sample good reviews (try)
  - Extensive guidelines subtle factors
- Assignments contribute to final grade
- Exact formula only at end
- Our decision is final, vetted by two

Logistics: Material

- Reading material
  - https://www.mpi-inf.mpg.de/question-answering-systems/
- Slides and recordings
  - https://drive.google.com/drive/folders/1Z0IjVSjymCD6IX_TCn4dNInuz8Kb42BL
- Assignment ~~grades~~ shared folder
  - https://docs.google.com/spreadsheets/d/e/2PACX-1vRN0qyr-ooE1JGLfLoPM89ipdPRBprwRUAKLkaRXPqCmDcj0Ht9T5LfGlqEe3gLk3sHS9YIHndDjRLl/pubhtml?gid=1036917168&single=true&urp=gmail_link

→ shared spreadsheet

Question of the day

How can we disambiguate named entities present in questions?

You'll find this covered in

① ▪ Fast and Accurate Annotation of Short Texts with Wikipedia Pages TAGME

  - Ferragina and Scaiella
  - CIKM 2010 + IEEE Software 2011 CIKM, SIGIR, WSDM, WWW : IR+Web
  - Links → https://arxiv.org/pdf/1006.3498.pdf

② ▪ Robust Disambiguation of Named Entities in Text AIDA

  - Hoffart et al.
  - EMNLP 2011 ACL, EMNLP, NAACL, EACL, COLING : NLP
  - https://www.aclweb.org/anthology/D11-1072.pdf

    official

Entities: Basics

Obj. facts

NE Michelle is married to Barack NE (R, V)

NE India is the birthplace of Gandhi (N, R)

NE? Pandas eat bamboo (R, V)

verb-med R
noun-med R
noun E

- Entities and relationships (E, R)
- Entities and named entities
  - People
  - Organizations
  - Locations
  - More (ever expanding: movies, food, animals...)

consensus

E, P/R, C, L

~~classes~~?

think about generalizing entities: marriage, marriage laws, bank investment

concepts

- Concepts and classes

lists → category

- Course Here: What is there in Wikidata/Wikipedia ☺

Entities: Tasks

NLP + IR

Stanford corenlp.run

- Named entity recognition : NER in text
- Named entity disambiguation / linking
  NED/NEL
- Named entity typing : NET
  - Fine : Association football club
  - Coarse : PER

- Main ideas: Similarity and coherence

in NERD

Michelle is married..

NER

Barcelona FC is the current champ in UEFA

NER

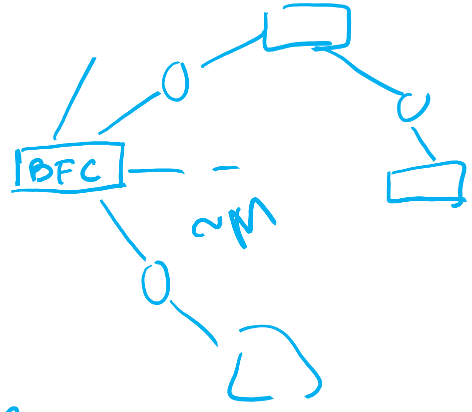
coherence

points

to KG / Wikipedia

Film club

Football club



Answer typing

Not exactly NET

Where does Messi play?

type of ans

Nolan directed film won...

Entities: Applications

- 1 Question understanding
- 2 News readability
- 3 Information extraction
- Many many more in IR + NLP
    - Can you think of some..?

Which Nolan film ...

correct Nolan ⇒ right Nolan.

Mexico FC

Mexico defeated Sudan ...

FC

Research paper 1

Fast and Accurate Annotation of
Short Texts with Wikipedia Pages

[Fast and accurate annotation of short texts with wikipedia pages](#)
P Ferragina, U Scaiella
IEEE software 29 (1), 70-75

815 * 2011

versions

Try it out!

tunable knob

URL

https://tagme.d4science.org/tagme/

Thanks to Prof. Paolo Ferragina for the slides

Tagged text | Topics

On this day 24 years ago Maradona scored his infamous "Hand of God" goal against England in the quarter-final of the 1986

Wikipedia

Less links

Tagged text | Topics

On this day 24 years ago Maradona scored his infamous "Hand of God" goal against England in the quarter-final of the 1986

API

More links

One issue: Synonymy

classical approach:
word vectors

challenges

He is using Microsoft's browser

He is a fan of Internet Explorer

→ same

Another issue: Polysemy

celebrity

the paparazzi photographed the star

the astronomer photographed the star

astronomical
body

Wikipedia is a rich source of instances

Title

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (book).

anchors

Steven Paul "Steve" Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[5][6] was an Arab-American[7] entrepreneur[8] and inventor,[9] who was the co-founder, chairman, and CEO of Apple Inc. Through Apple, he was widely recognized as a charismatic pioneer of the personal computer revolution and for his influential career in the computer and consumer electronics fields, transforming "one industry after another, from computers and smartphones to music and movies..."[12] Jobs also co-founded and served as chief executive of Pixar Animation Studios; he became a member of the board of directors of The Walt Disney Company in 2006, when Disney acquired Pixar. Jobs was among the first to see the commercial potential of Xerox PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa and, one year later, the Macintosh. He also played a role in introducing the LaserWriter, one of the first widely available laser printers, to the market.[13]

After a power struggle with the board of directors in 1985, Jobs left Apple and founded NeXT, a computer platform development company specializing in the higher-education and business markets. In 1986, he acquired the computer graphics division of Lucasfilm, which was spun off as Pixar.[14] He was credited in *Toy Story* (1995) as an executive producer. He served as CEO and majority shareholder until Disney's purchase of Pixar in 2006.[15] In 1996, after Apple had failed to deliver its operating system, Copland, Gil Amelio turned to NeXT Computer, and the NeXTSTEP platform became the foundation for the Mac OS X.[16] Jobs returned to Apple as an advisor, and took control of the company as an interim CEO. Jobs brought Apple from near bankruptcy to profitability by 1998.[17][18][19]

Steve Jobs

Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

Born	Steven Paul Jobs

Infobox

PARC (company)

From Wikipedia, the free encyclopedia

(Redirected from PARC User Interface)

Wikipedia's categories contain classes

research prototypes

not class

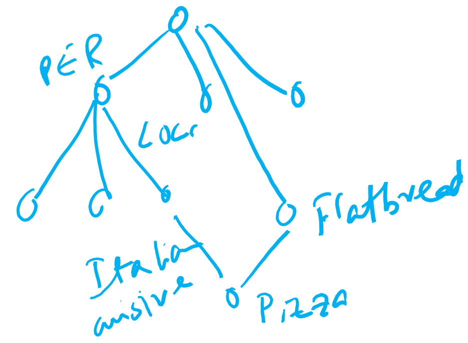
Categories: Steve Jobs | 1955 births | 2011 deaths | American adoptees | American billionaires | American chief executives | American computer businesspeople | American industrial designers | American inventors | American people of German descent | American people of Swiss descent | American people of Syrian descent | American technology company founders | American Zen Buddhists | Apple Inc. | Apple Inc. employees | Businesspeople from California | Businesspeople in software | Cancer deaths in California | Computer designers | Computer pioneers | Deaths from pancreatic cancer | Disney people | Internet pioneers | National Medal of Technology recipients | NeXT | Organ transplant recipients | People from the San Francisco Bay Area | Pescetarians | Reed College alumni

classes

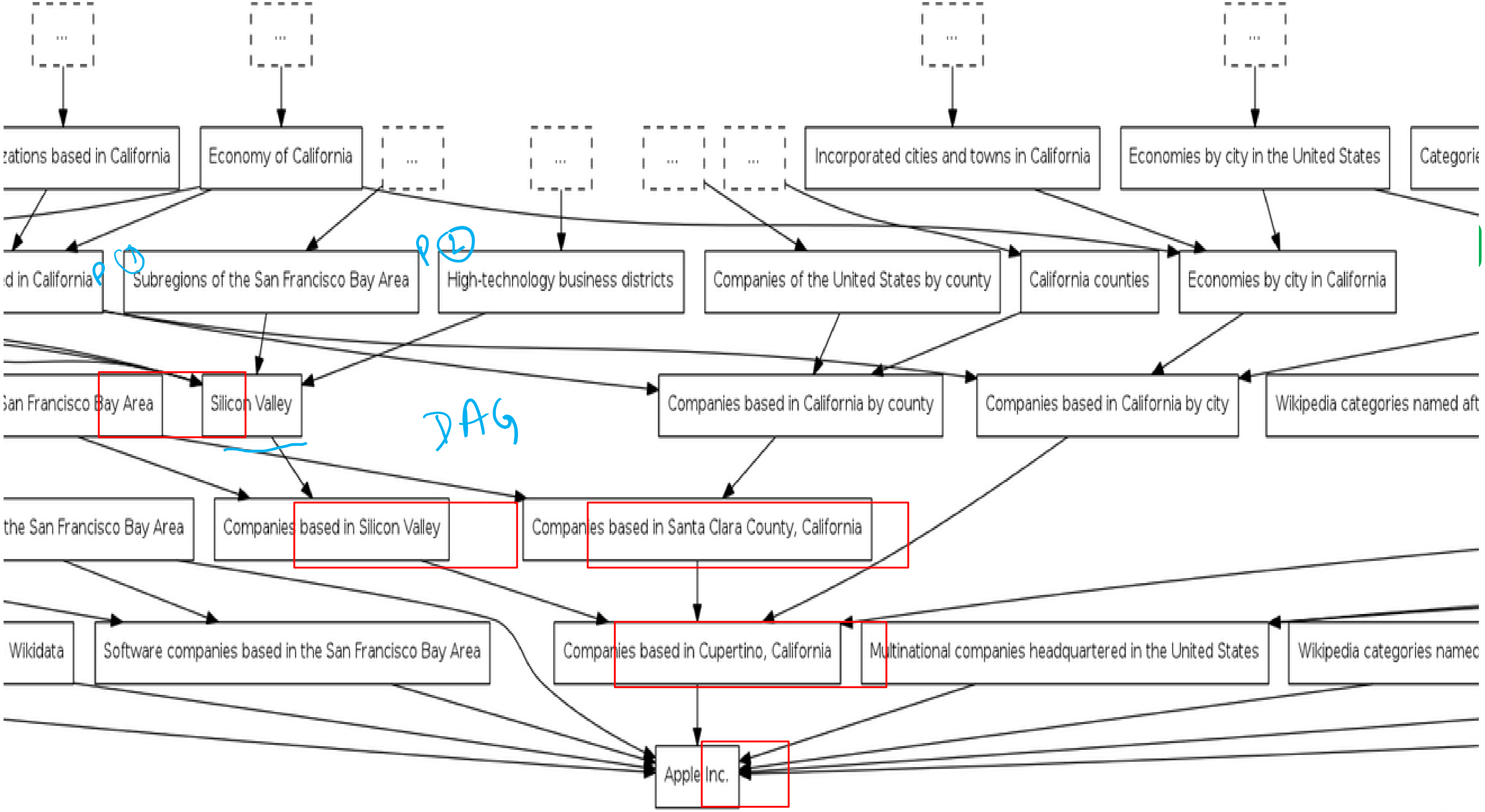
not class

ontology

Categories typically form a taxonomic DAG



DAG of categories



Topic-based annotation

↳ wikipedia articles

multiple meanings

text span
Mention
↳ spot

"Diego Maradona won against Mexico"

Mexico's football team

*Ex-**Argentina's player***

→ Diego Maradona

Diego played

anchor

Entity
↓
canonicalized
KG, Wikipedia

General Task →

Find anchors and annotate them
with articles drawn from Wikipedia!

specific

Synonymy

Internet Explorer

From Wikipedia, the free encyclopedia

Internet Explorer[a] (formerly **Microsoft Internet Explorer**[b]
Microsoft and included in the Microsoft Windows line of opera
were available as free downloads, or in-service packs, and in
browser is discontinued, but still maintained.[4]

He is using <u>Microsoft's browser</u>

She plays with <u>Internet Explorer</u>

Polysemy

Celebrity *is a person who is famously recognized ...*

the paparazzi photographed the star

the astronomer photographed the star

Star *is a massive, luminous ball of plasma ...*

Why is it a difficult problem?

NER + NED

Trump

Jab We Met

"Diego Maradona won against Mexico"

- Ex-Argentina's ~~coach~~
- ~~His nephew~~
- ~~Maradona Stadium~~
- ~~Maradona Movie~~
- …

- ~~Mexico nation~~
- ~~Mexico state~~
- Mexico football team
- ~~Mexico baseball team~~
- …

NED

Don't annotate!

The literature

- TagMe (Univ. Pisa)
- DBPedia Spotlight (Univ. Berlin)
- Illinois Wikifier (Univ. Illinois)
- AIDA (Max Planck Institute for Informatics)
- CNMS (Univ. Amsterdam)
- Wikipedia Miner (Univ. Waikato)

Many commercial software: AlchemyAPI, DBpedia Spotlight, Extractiv, Lupedia, OpenCalais, Saplo, SemiTags, TextRazor, Wikimeta, Yahoo! Content Analysis, Zemanta.

The TAGME system

- Designed for short texts → unique
  - news, blogs, search-results snippets, tweets, ads, etc
  - competitive on long texts too → what are long texts ??

new problem!

- Achieves high accuracy
  - Massive experimental test on millions of short texts
- Fast
  - More than 10x faster than others
  - 100% Java

Good research problem
- everything done ??
- look deeper!
- Problem + quality of results

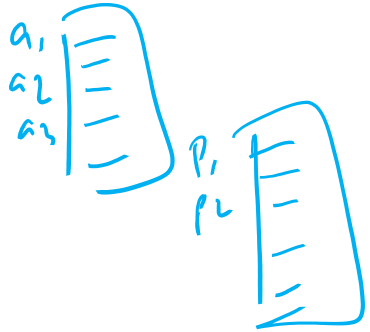
find out → speed ?
→ memory ?

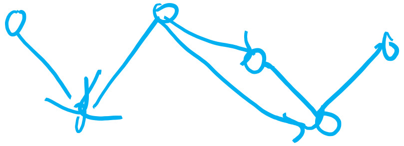
often unsaid papers

TAGME: Distilled information

Lucene/Solr/Indri

- Anchor dictionary
- Page catalog
- In-link graph





clean notation

page, anchor

$p, a, Pg(a)$

$link(a), freq(a)$ $\quad a = cat$

$Pr(p \mid a)$

✓ $l_p(a) = \frac{link(a)}{freq(a)}$ $\quad p_a \in Pg(a)$

$a \mapsto p$

TAGME: Overview / Anatomy

- Anchor parsing
- Anchor disambiguation
- Anchor pruning

TagMe is NERD → NER + NED

find mention → parsing

| w_1 | w_2 | w_3 | w_4 |

nested strings

$a_1 \in a_2$

Trojan War

Exception: the act vs act

overlapping mention?

tom cruise ship scene

$lp(a_1) > lp(a_2) \Rightarrow link(a_1) >> link(a_2)$

Features used to link a ➔ p

prior

Commonness of a page p wrt an anchor a

$$\Pr(p \mid a) = \frac{\#\ a \text{ linked to } p}{\#\ a \text{ as anchor}}$$

X

Context of *a* around the mention

T = ...w_1 w_2 w_3 a w_4 w_5 w_6 ...

left right

and the content of a *page/entity*

Page p = z_1 z_2 z_3 z_4 z_5 z_6 ...

link node

p

a

Link probability of an anchor a

$$lp(a) = \frac{\text{freq of } a \text{ as anchor}}{\text{freq of } a \text{ in the text}}$$

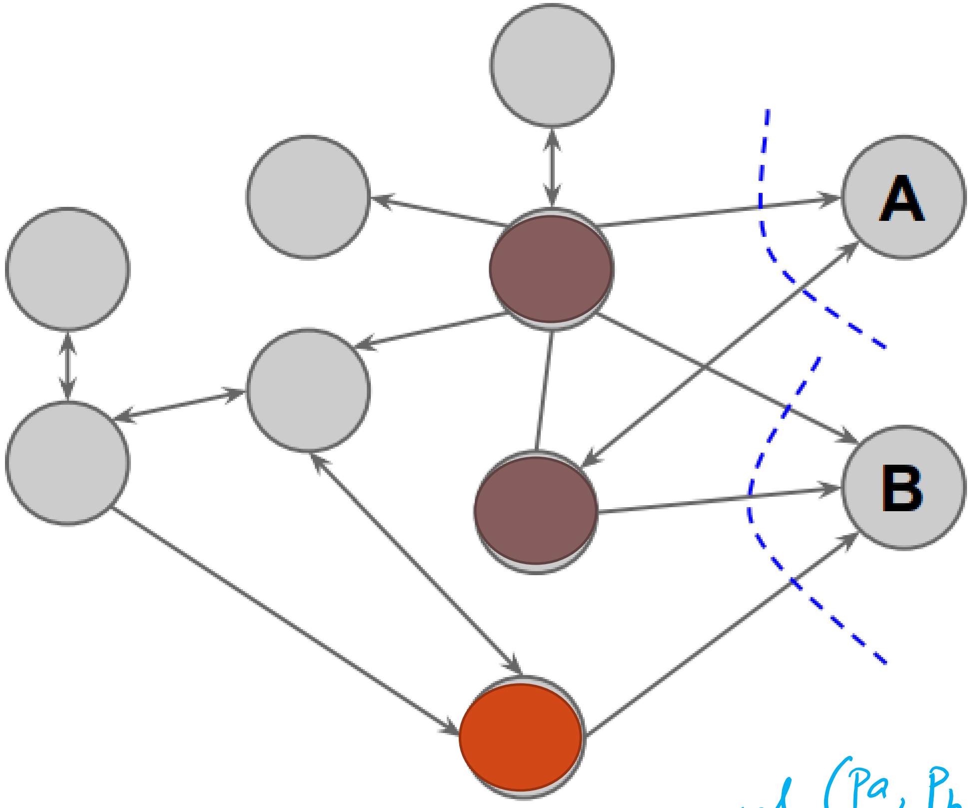
Graph-based features

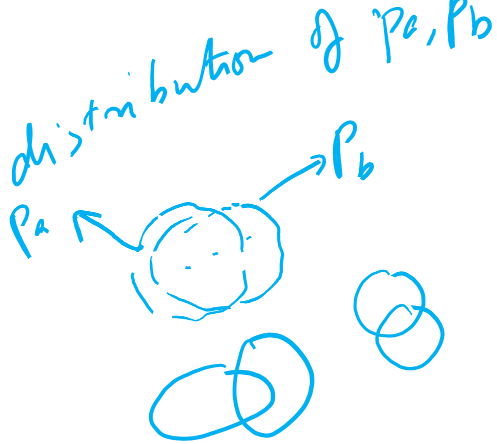
a is a mention-node

p is an entity-node

Links a → p and paths between pages

Relatedness between pages





relatedness

in (p_a)

in (p_b)

rel (p_a, p_b)

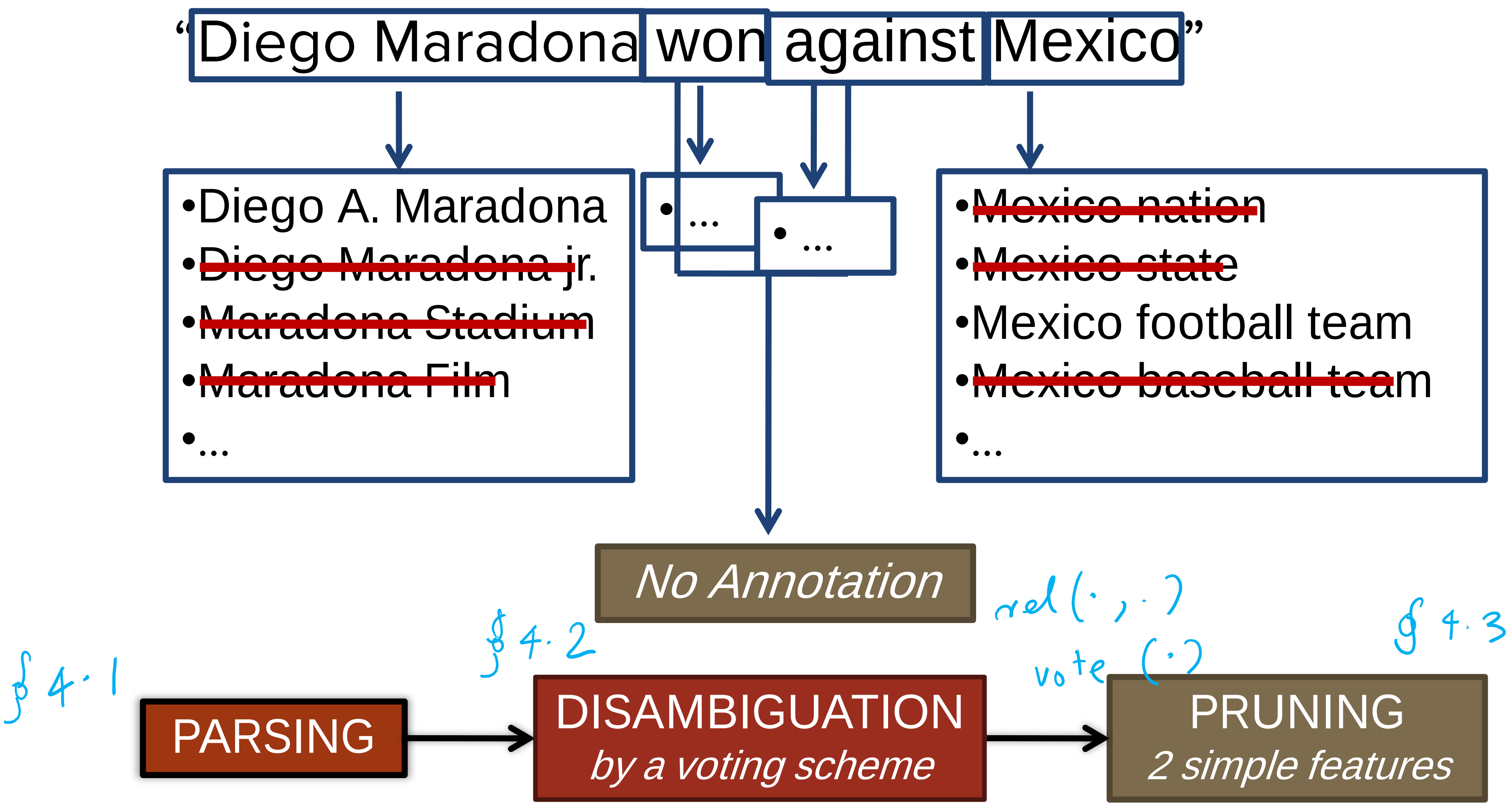
$$sr(a,b)=\frac{\log(\max(|A|,|B|))-\log(|A\cap B|)}{\log(|W|)-\log(\min(|A|,|B|))}$$

log (10^6) ~> 6 constant

Milne & Witten

✓ https://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-005.pdf

How TAGME works



Disambiguation: The Voting Scheme

4. 2

Collective agreement among topics via voting

b "Diego Maradona won against Mexico" a

p_b

- Diego A. Maradona
- Diego Maradona jr.
- Maradona Stadium
- Maradona Film
- ...

$\in Pg(b)$

expression

$$vote_b(p_a) = \frac{\sum_{p_b \in Pg(b)} rel(p_b, p_a) \cdot \Pr(p_b|b)}{|Pg(b)|}$$

$$rel_a(p_a) = \sum_{b \in A_T \setminus \{a\}}$$

p_a

$\in Pg(a)$

Wiki-articles
Mexico
State of Mexico
Mexico National Football Team
Mexico

so-called highest vote

Disambiguation: All steps

4.2 τ

τ = trade-off
speed vs. recall

$\varepsilon, \tau, \rho, \ldots, \alpha, \beta, \gamma$

Pruning by
commonness < **τ**

τ

Commonness of a page
p wrt an anchor a

$$\Pr(p \mid a) = \frac{\#\, a \text{ linked to } p}{\#\, a \text{ as anchor}}$$

Voting scheme

Select top-**ε** pages

Select the best in
commonness

Pruning §4.3

- Use 2 features:
  1. link probability
  2. coherence wrt context

  v. imp. in NED

- Compute a ρ score via
  - 3 classifiers, avg, linear combination

training step

- if $\rho < \rho_{NA}$ then prune!

threshold

Link probability of an anchor *a*

$$lp(a) = \frac{\text{freq of } a \text{ as anchor}}{\text{freq of } a \text{ in the text}}$$

Avg. relatedness btw the assigned concept to the others

$$coh(a \mapsto p_a) = \frac{1}{|S|-1} \sum_{p_b \in S \setminus \{p_a\}} rel(p_b, p_a)$$

balance precision vs. recall

Research paper 2

Robust Disambiguation of Named Entities in Text

Robust disambiguation of named entities in text J Hoffart, MA Yosef, I Bordino, H Fürstenau, M Pinkal, M Spaniol, ... Proceedings of the Conference on Empirical Methods in Natural Language …	833	2011
	+	
Aida: An online tool for accurate disambiguation of named entities in text and tables MA Yosef, J Hoffart, I Bordino, M Spaniol, G Weikum Proceedings of the VLDB Endowment 4 (12), 1450-1453	158	2011

Disambiguating to KG entities

- TAGME is fast and effective
- Works well for short texts
- Does not go all the way!
- Wikipedia more general, but we need KG-linking!
- Lookup KG entities using Wikipedia links?
- Harness KG properties! Enter AIDA.

mention → wikipedia → KG (Wikidata)

search KG

Christopher Nolan (Wikipedia URL)

Nolan in KG

Chr.

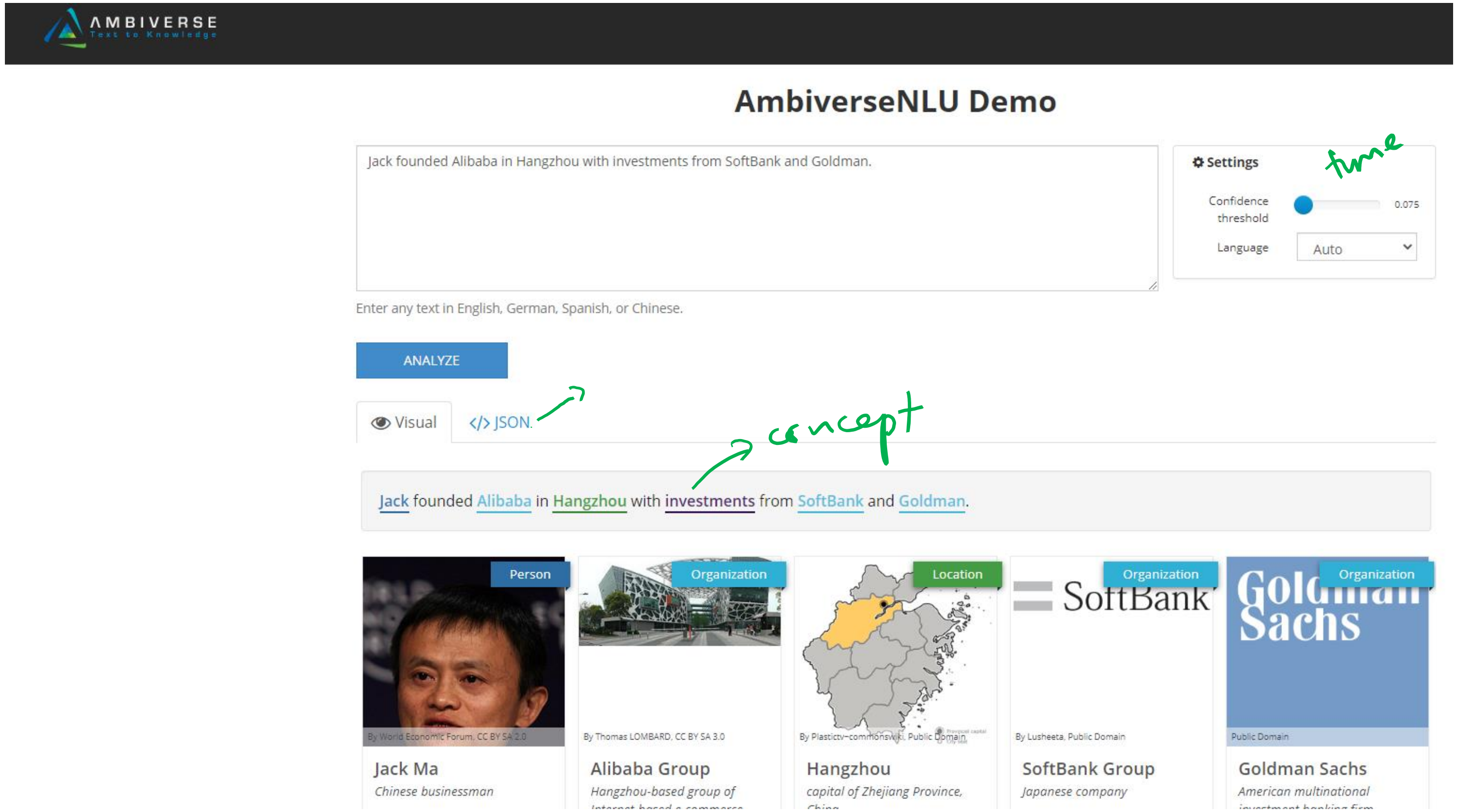
Try it out!

AIDA → Ambiverse startup

AIDA ↓

Acc Onl Disa NE

→ 86.1



tune

concept

https://ambiversenlu.mpi-inf.mpg.de/

Disambiguating names to entities

AIDA Framework

Mention & ambiguity

Entity candidates

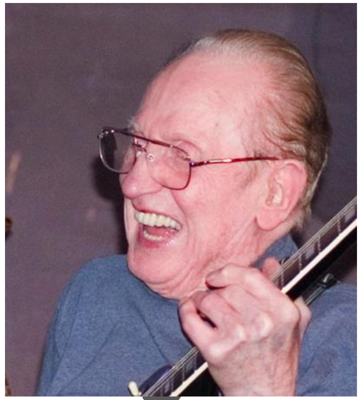
Popularity prior

Context sim (m, e)

Coherence

Overall obj. fn

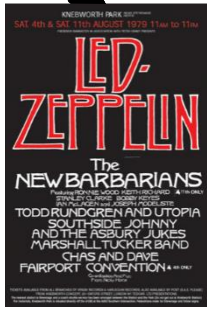
mentions



NED is non-trivial

When Page played Kashmir at Knebworth, his Les Paul was uniquely tuned.

Images taken from Wikipedia under CC BY-SA 3.0

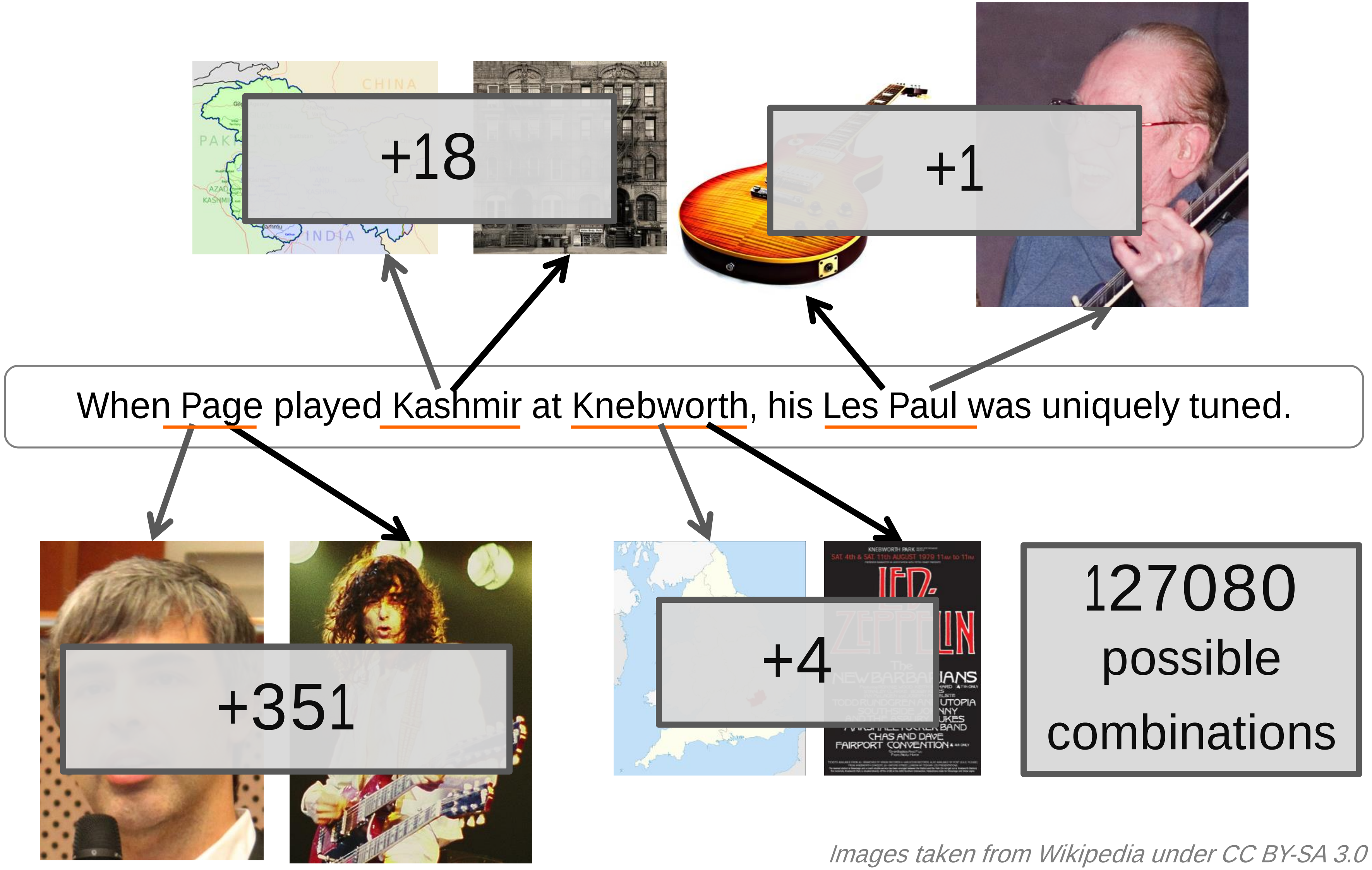


AIDA uses (Stanford) NER

→ NED

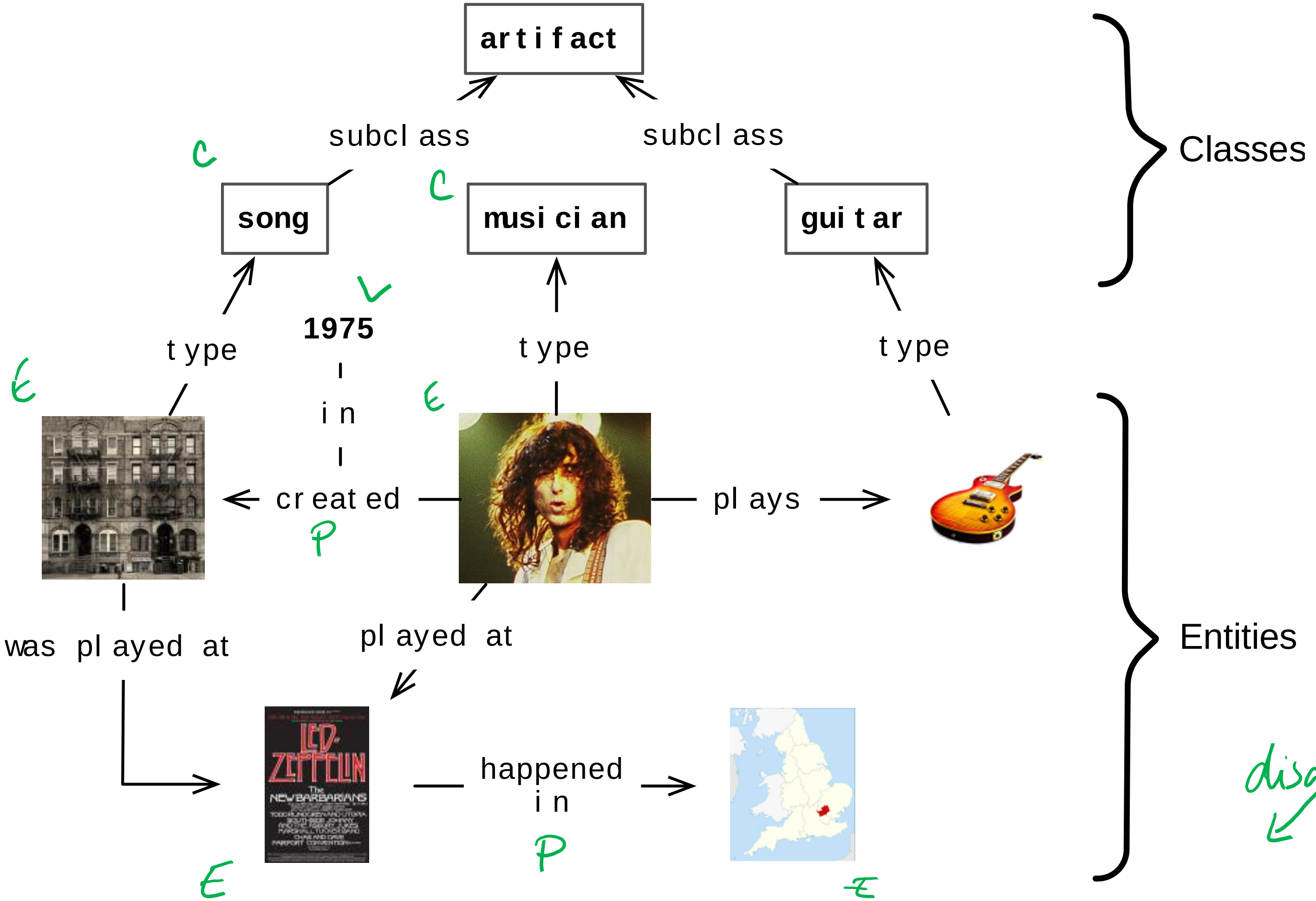
Thanks to Dr. Johannes Hoffart for the slides

Disambiguating names to entities



Images taken from Wikipedia under CC BY-SA 3.0

Entities in knowledge bases



AIDA
disambiguates

built from

AIDA features for disambiguation

Images taken from Wikipedia under CC BY-SA 3.0

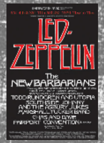
- India
- Pakistan
- Pashmina
- ...

0.0

When Page played Kashmir at Knebworth, his Les Paul was uniquely tuned.

2.4

- Led Zeppelin
- Jimmy Page
- Knebworth Festival
- ...

Prior	Context	Coherence
91%	0.0	  
5%	2.4	  

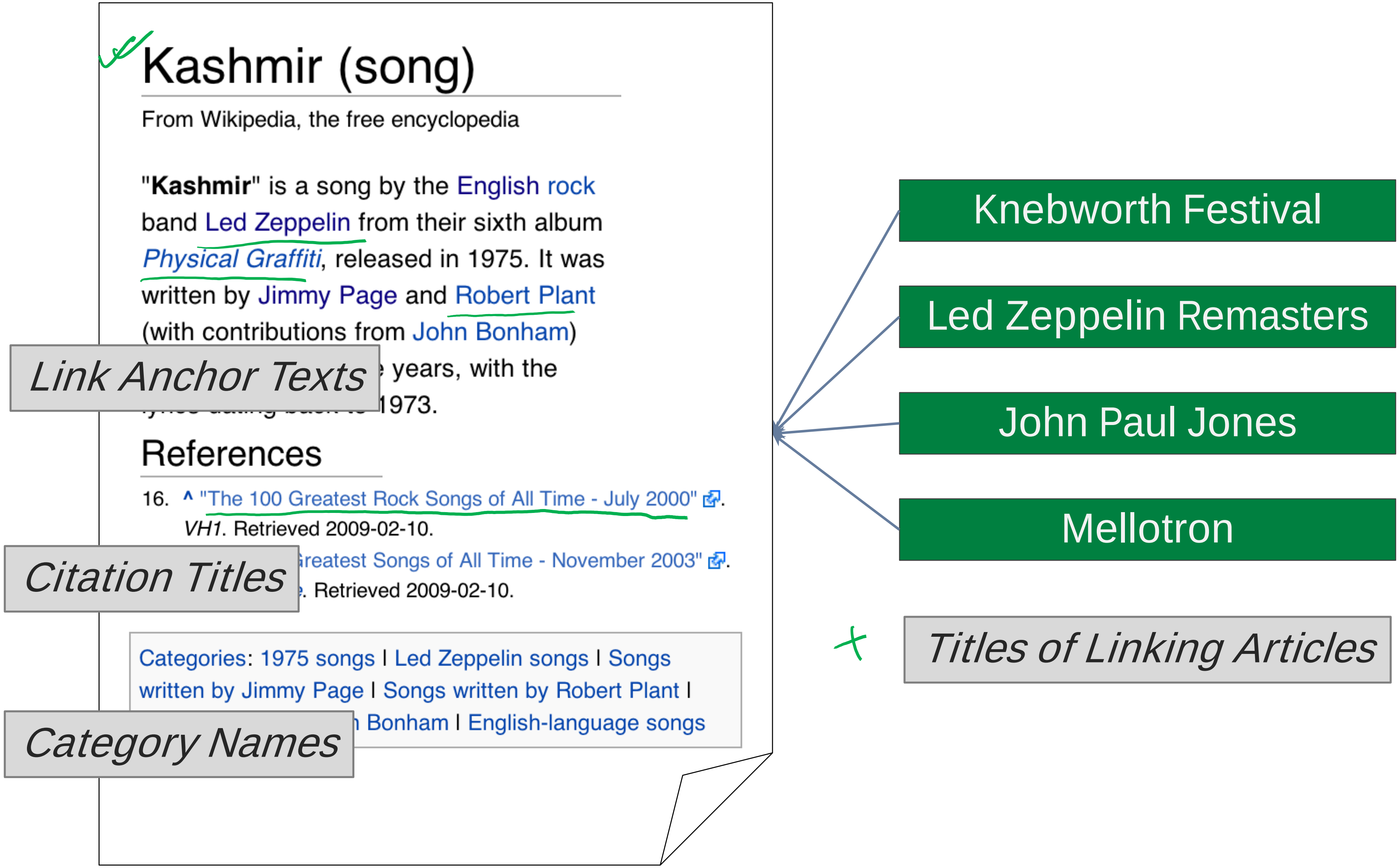
How often did “Kashmir” link to this entity in Wikipedia?

Are the disambiguated entities related?

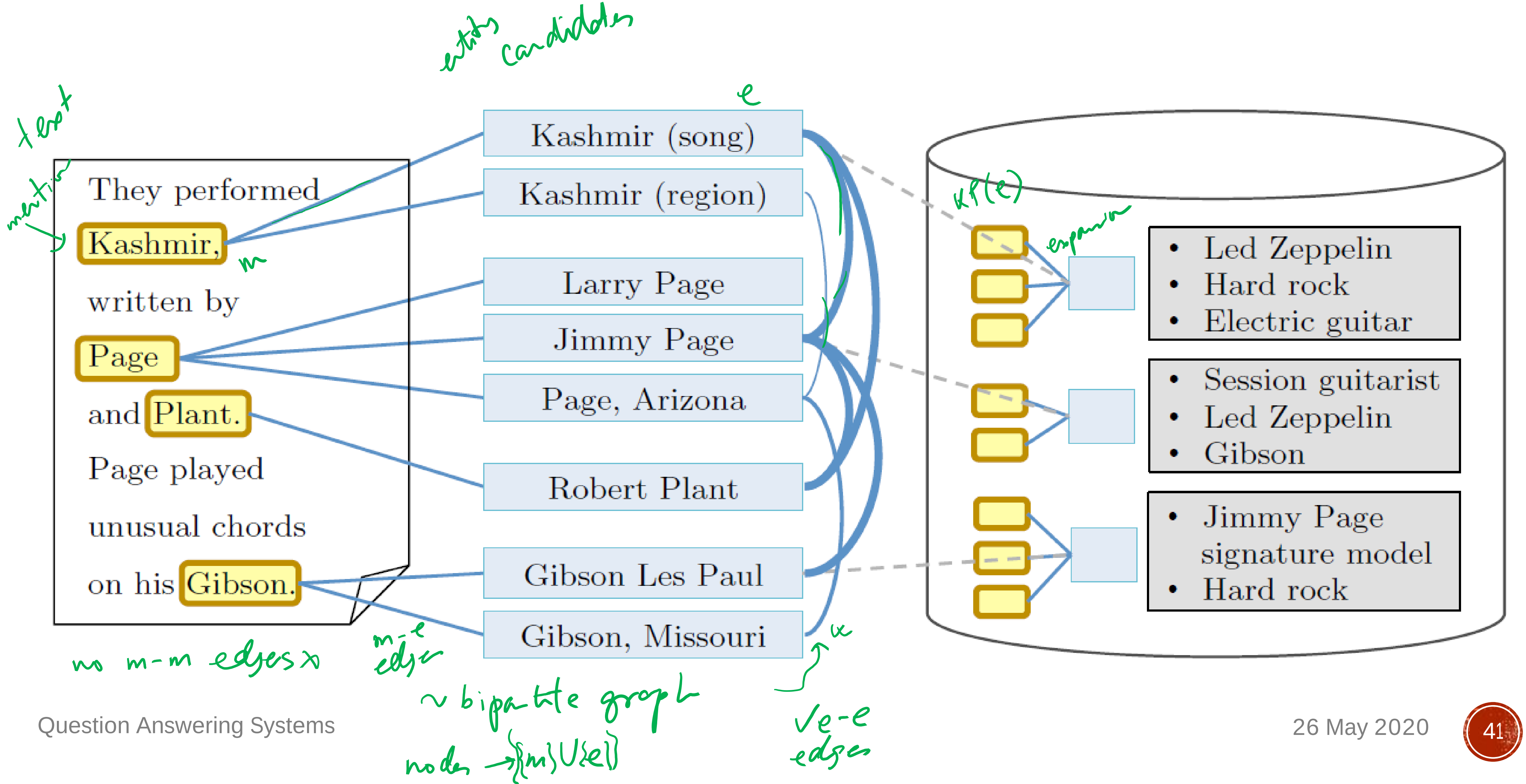
Entity keyphrases

§4.2

kp(e)

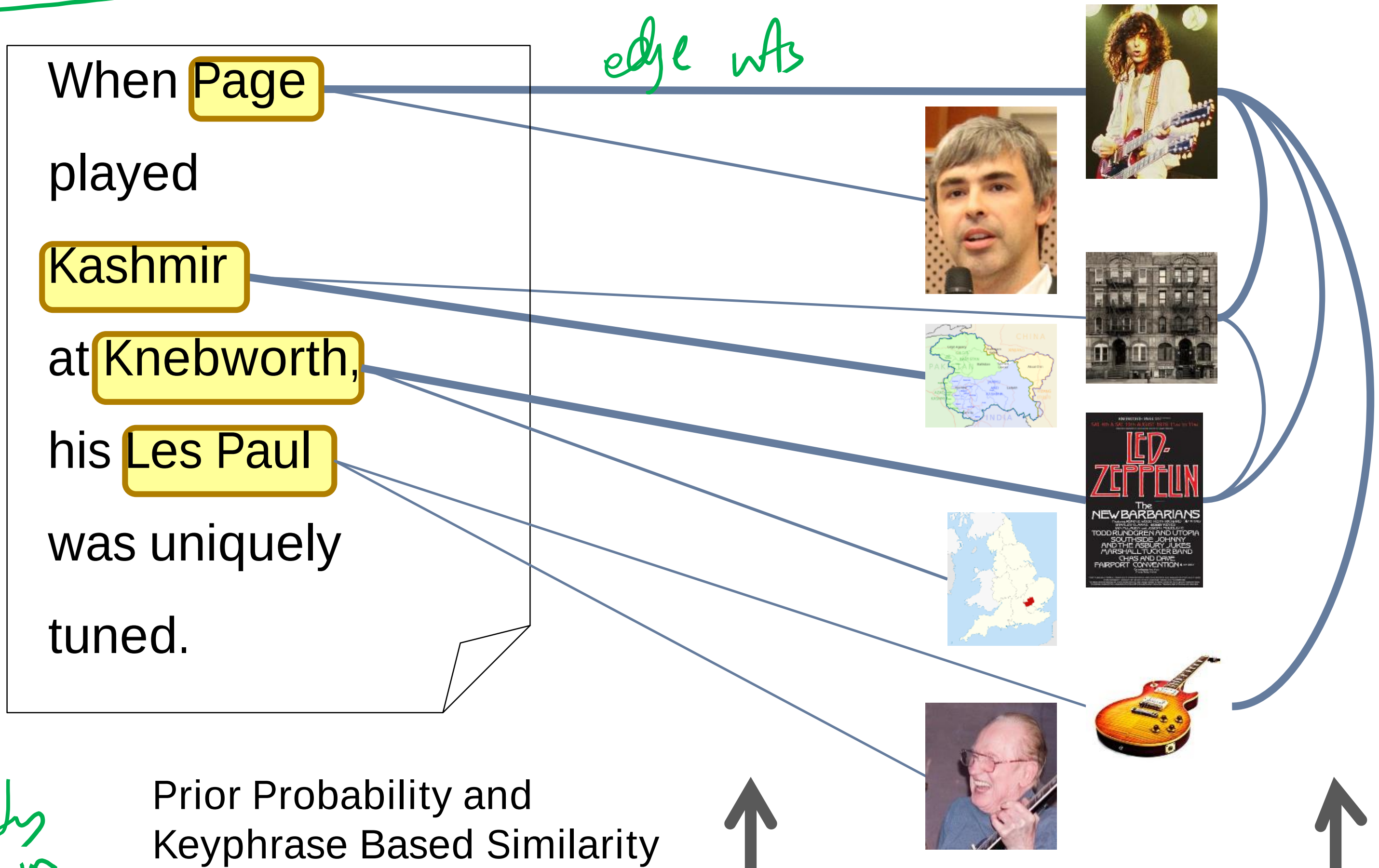


Mention-entity graph example



AIDA: Joint disambiguation as graph problem

together, not pipeline
page, then kashmir, then kneb X

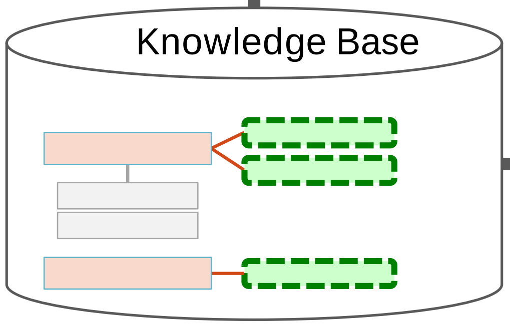


ref to GST

density

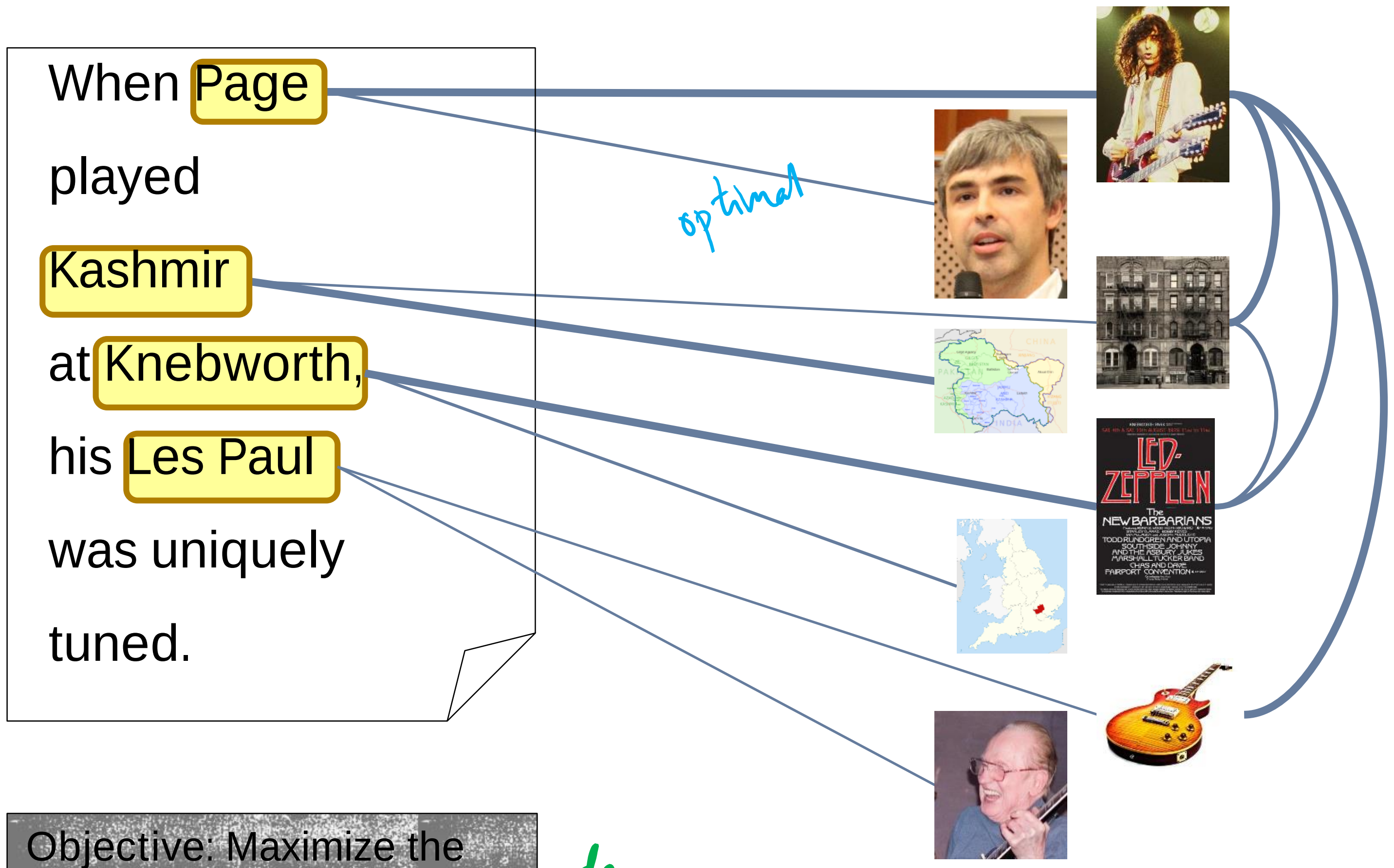
cf QUEST

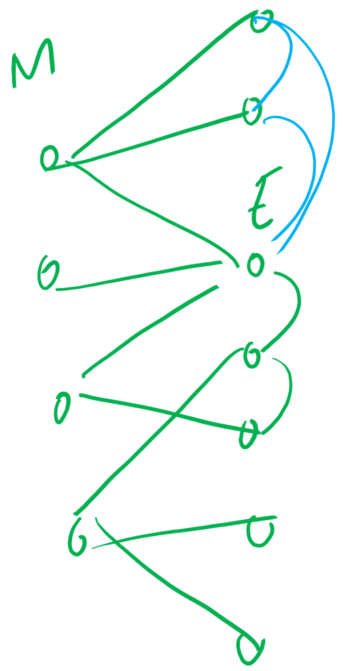
Goal: Dense subgraph with one mention-entity edge per mention



AIDA don't use GST

AIDA: Joint disambiguation as graph problem





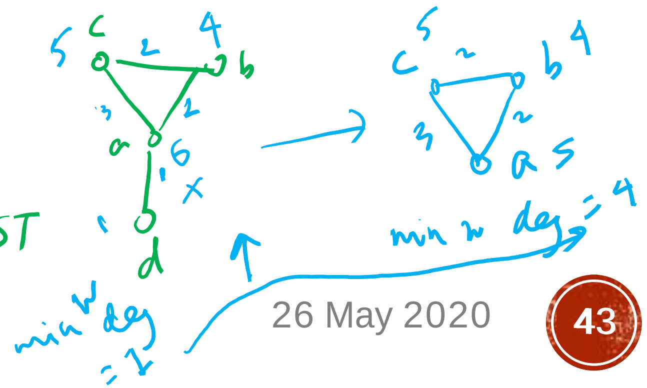
$|M| \ll |E|$

Objective: Maximize the minimum weighted degree

✓ Constraint: Keep at least one entity per mention

dense → min. wt. ed degree

recall NEGA, QUEST



Keyphrase-based similarity

- Keyphrases (*kp*) commonly occur only partially

t1 t2

$$score(kp) = \underbrace{\frac{\#\ matching\ words}{length\ of\ cover(kp)}}_{\text{Account for partial matches}} \underbrace{\left(\frac{\sum_{w \in cover} weight(w)}{\sum_{w \in kp} weight(w)}\right)^2}_{\text{Weight of contained tokens } w}$$

§ 4.2

score (q) = z

Kashmir was $\underbrace{\text{written by Page and Plant.}}_{cover}$

q ∈ KP(e)

wt a func.

"Songs written by Robert Plant"

cover → max span

- To score an entity, all keyphrase scores are summed

$$sim\ score(m, e) = \sum_{q \in KP(e)} score(q)$$

Keyword weighting

$\swarrow z$ wt. (w)

$$score(kp) = \frac{\#\ matching\ words}{length\ of\ cover(kp)} \left(\frac{\sum_{w \in cover} weight(w)}{\sum_{w \in kp} weight(w)} \right)^2$$

wt (w)

- Global IDF of a keyphrase token w in Wikipedia ~ IRDM

- Mutual Information of a token w and an associated entity word association measure
  - How often does the token occur in the keyphrase set of an entity?

https://en.wikipedia.org/wiki/Mutual_information

MI
PMI
nPMI

Disambiguation by joint inference

Input

- Mentions
  - context of mention *cxt*(m)
  - entity candidates *e + cxt(e)*
- Features

Prior	*prior(m,e)*
Similarity	*sim(cxt(m),cxt(e))*
Coherence	*coh(e1,e2)*

Goal

$$\alpha \cdot \sum_{i=1}^{k} \mathrm{prior}(m_i, e_{j_i})$$

not used in TAGME

$$+\beta \cdot \sum_{i=1}^{k} \mathrm{sim}(\mathrm{cxt}(m_i), \mathrm{cxt}(e_{j_i}))$$

$$+\gamma \cdot \mathrm{coh}(e_{j_1}, e_{j_2}, \ldots e_{j_k})$$

$$= \max!$$

obj. fn.

notation

$m_i \leftrightarrow e_{j_i}$ $\arg\max(\cdot)_{\text{score}}$

over what?

Greedy graph algorithm

- Input: weighted graph of mentions and entities
- Output: result graph with maximum density

- Objective: maximize the minimum weighted degree
- Constraint: keep at least one entity per mention

Key idea

1. Prune entities that are too distant from all mentions
2. While an entity can be removed, remove the one with the lowest weighted degree
   - Keep graph with best minimum weighted degree

Final steps

- Find subgraph maximizing total edge weight
  - If graph is small enough, enumerate all potential mention-entity mappings
  - Otherwise do local search, randomly switching mention-entity mappings for a fixed number of times

Robustness issues

§5.3

Prior may be misleading
Given the prior probability for all entity candidates, only use prior when very good indicator for one single entity (>90%)

Coherence can get hooked to a wrong subgraph
Given prior probability and similarity distribution for all entity candidates. If they are reasonably similar, fix entity for mention before running the graph algorithm.

Dataset is available at http://www.mpi-inf.mpg.de/yago-naga/aida

Conclusions

- Understanding entities is vital to QA
- NERD is a vital (first) cog in the QA wheel
- Often used off-the-shelf
- Many, many innovative techniques
- Mainly based on priors, similarity and coherence m - e
- Applications span beyond QA!