Semantičko indeksiranje (semantic annotation / entity linking)

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

PREPOZNAVANJE ENTITETA U TEKSTU

- *Named Entity Recognition* (*NER*)
- Entiteti mogu biti različitog tipa: osoba, organizacija, lokacija, datum, valuta sl.
- Primer:

> Peter Norvig presents as part of the UBC Department of Computer Science's Distinguished Lecture Series, September 23, 2010.

> Peter Norvig [PER] presents as part of the UBC Department of Computer Science's [ORG] Distinguished Lecture Series, September 23, 2010 [DATE].

Semantičko indeksiranje

- *Semantic indexing* = *NER* + *Disambiguation*
- *Disambiguation* = jedinstveno identifikovanje prepoznatog entiteta

Primer koristi TagMe servis:
https://sobigdata.d4science.org/web/tagme/

PRIMERI SERVISA ZA SEM. INDEKSIRANJE

- Alchemy API Language Services
  - http://www.alchemyapi.com/products/alchemylanguage
- TextRazor
  - http://www.textrazor.com/
- OpenCalais
  - http://www.opencalais.com/
- Dandelion API
  - https://dandelion.eu/
- TagMe (open source)
  - https://sobigdata.d4science.org/web/tagme/
- DBpedia Spotlight (open source)
  - https://github.com/dbpedia-spotlight/dbpedia-spotlight

Semantičko indeksiranje

- Kombinacija m. učenja i znanja sadržanog u bazama znanja na Web-u
- Najčešće korišćene baze znanja: Wikipedia, DBpedia, WikiData

OSNOVNI KORACI U PROCESU SEM. INDEKSIRANJA

1) *Entity spotting* & *candidate selection* – identifikacija termina koji bi mogli označavati entitete (*entity-mentions*) i selekcija mogućih entiteta iz baze znanja za svaki *entity-mention*
2) *Disambiguation* – izbor "najboljeg" entiteta za svaki *entity-mention,* tj. izbor entiteta koji najbolje odražava semantiku datog termina u datom kontekstu
3) *Filtering* – filtriranje rezultata u cilju eliminacije irelevantnih entiteta

ENTITY SPOTTING & CANDIDATE SELECTION

- Ciljevi prve faza procesa prepoznavanja entiteta su:
  - identifikovati tzv. *entity-mentions* u ulaznom tekstu, tj, delove teksta (pojedinačne reči i izraze) koji označavaju entitete;
  - identifikovati u bazi znanja (npr., Wikipedia ili DBpedia) skup mogućih entiteta za svaki *entity-mention*

ENTITY SPOTTING & CANDIDATE SELECTION

- Primer

"They performed Kashmir, written by Page and Plant.
Page played unusual chords on his Gibson."

Kandidati:
dbpedia:Kashmir – a valley between Pakistan, India and Ladakh
dbpedia:Kashmir_(band) – a Danish rock band
dbpedia:Kashmir_(song) – 1975 song by rock band Led Zeppelin
dbpedia:Kashmir,_Iran – a village in Iran
…

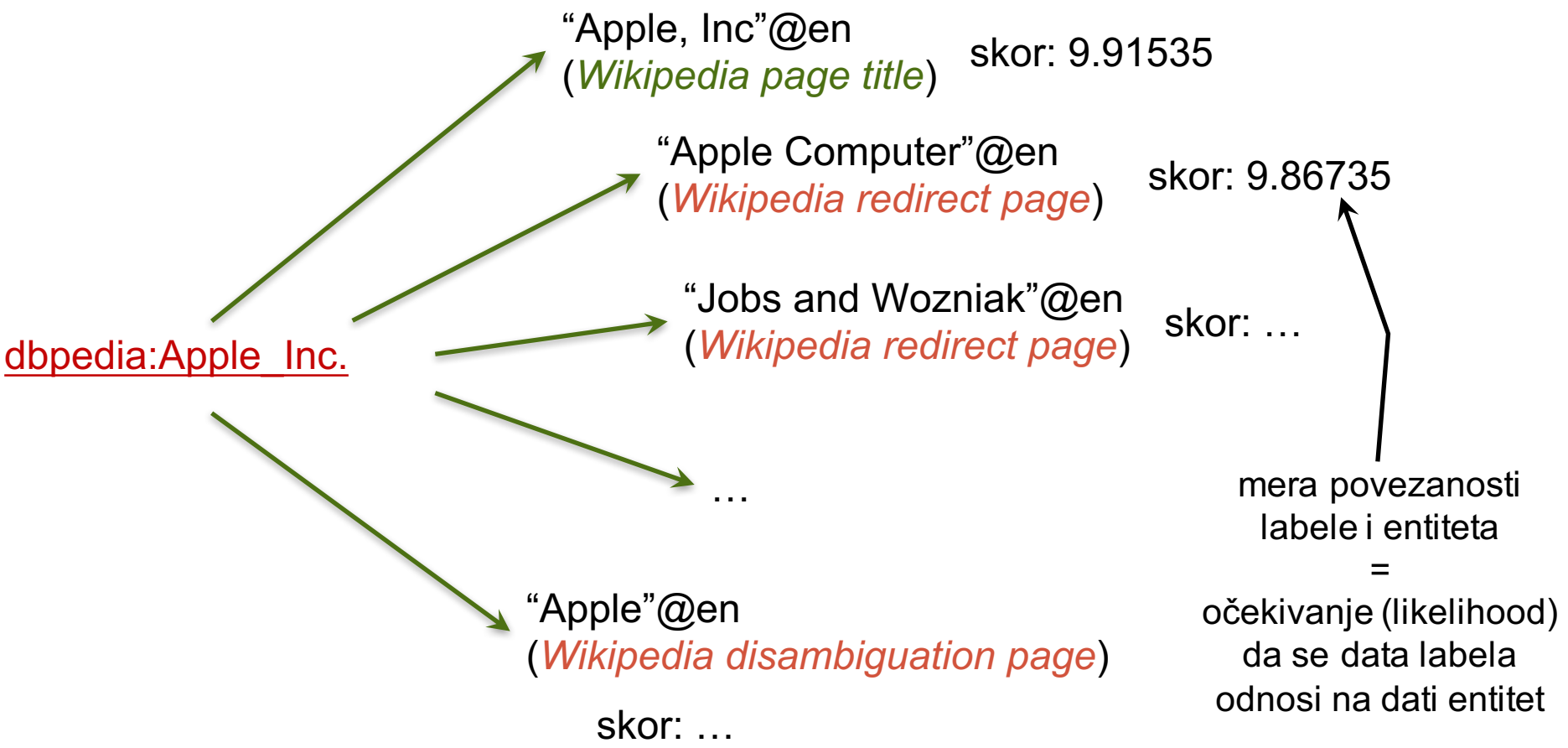
Izvor kandidata: Wikipedia/DBpedia disambiguation stranica za pojam Kashmir

Primer preuzet iz: Hoffart J. et al. (2011). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), 782-792.

ENTITY SPOTTING & CANDIDATE SELECTION

- Ova faza se tipično realizuje kao *dictionary look-up task*
  - Formira se rečnik putem ekstrakcije labela i opisa svih entiteta sadržanih u izabranoj bazi znanja
  - Wikipedia i DBpedia se najčešće koriste kao baze znanja, odnosno kao izvori iz kojih se estrahuju labele i opisi entiteta
  - Rečnik može sadržati, za svaki entitet, i različite statistike
    - npr. relevantnost određene labele za određeni entitet

Primer: DBpedia Lexicalization dataset

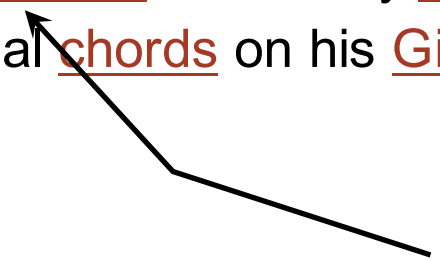


Dataset je raspoloživ na adresi: http://wiki.dbpedia.org/lexicalizations#

DISAMBIGUATION

- Cilj ove faze: za svaki *entity-mention*, selektovati jedan ili više entiteta koji mu po svom značenju (semantici) najviše odgovaraju
  - selekcija se radi iz, obično povećeg, skupa kandidata identifikovanih u prethodnoj fazi procesa
- Nastavljajući sa istim primerom:

"They performed Kashmir, written by Page and Plant.
Page played unusual chords on his Gibson."



dbpedia:Kashmir – a valley between Pakistan, India and Ladakh
dbpedia:Kashmir_(band) – a Danish rock band
dbpedia:Kashmir_(song) – 1975 song by rock band Led Zeppelin
dbpedia:Kashmir,_Iran – a village in Iran
…

DISAMBIGUATION

Postoji više različitih pristupa za realizaciju ove faze; neki od najčešće primenjivanih:

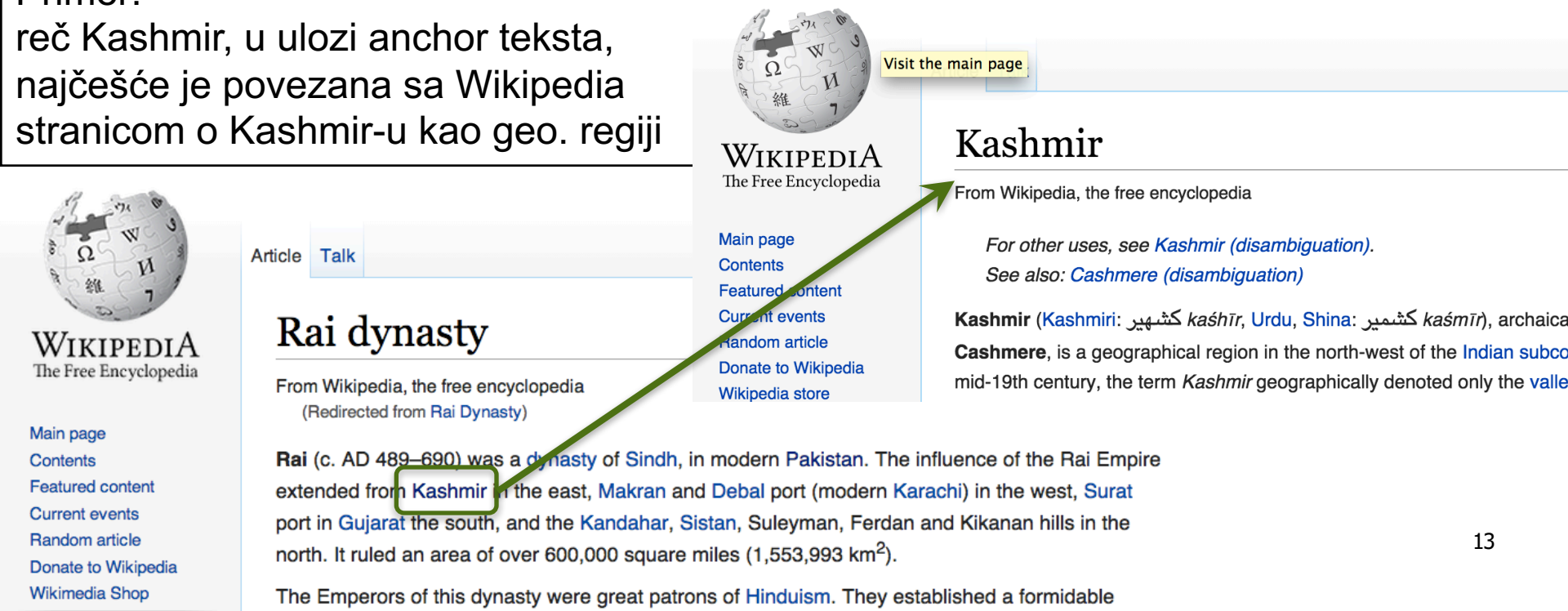
- *Popularity-based (mention-entity) prior*
- *Context-based approach*
- *Collective disambiguation*

DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

Ovaj pristup se sastoji u izboru najistaknutijeg entiteta za datu reč / frazu (*entity mention*)

- Npr., entitet sa kojim je data reč / izraz najčešće povezana kad se (ta reč/izraz) pojavljuje kao *anchor* tekst u Wikipedia-i

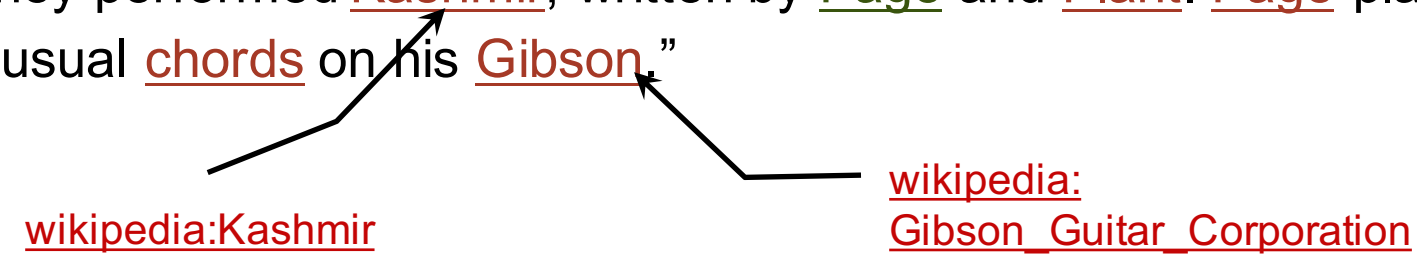
Primer:
reč Kashmir, u ulozi anchor teksta, najčešće je povezana sa Wikipedia stranicom o Kashmir-u kao geo. regiji



DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

Nastavljajući sa istim primerom:

"They performed Kashmir, written by Page and Plant. Page played unusual chords on his Gibson."



U Wikipedia-i,

- "Gibson" je primarno povezan sa Gibson guitar corporation entitetom, dok je samo marginalno povezan sa 24 preostala moguća značenja ovog termina
- "Kashmir" je primarno povezan sa Kashmir region entitetom (90.91% svih pojavljivanja ovog termina), dok se samo retko odnosi na pesmu grupe Led Zeppelin (5.45%)

DISAMBIGUATION: POPULARITY-BASED (MENTION-ENTITY) PRIOR

- Ovo je jednostavan pristup, ali često podložan greškama; zbog toga se koristi u kombinaciji sa drugim pristupima
- Greške se javljaju usled toga što se ne pridaje pažnja
  - kontekstu u kome se reč/izraz (mention) pojavljuje
  - generalnoj temi teksta

Ilustracija greške koja se obično javlja ukoliko se samo popularnost tj učestanost mention-entity konekcije uzima u razmatranje

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

- Jedan od često korišćenih pristupa za realizaciju ove faze
- Zasniva se na poređenju konteksta određenog *entity-mention*-a i konteksta svih entiteta koji su selektovani kao kandidati za taj *entity-mention*
- Za kontekst *entity-mention*-a se tipično uzima rečenica u kojoj se pojavljuje, dok se za kontekst entiteta uzima njegov opis iz baze znanja

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

- Kontekst se obično predstavlja kao prost skup reči tj. koristi se bag-of-words pristup za predstavljanje teksta
- Poređenje konteksta se vrši primenom neke od metrika za računanje sličnosti vektora
- Često korišćene metrike:
  - Cosine similarity,
  - (weighted) Jaccard coefficient,
  - Wikipedia link-based measure*

* Witten, I.H. & Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence, Chicago, USA, July, 2008. (pp. 25-30).

DISAMBIGUATION: PRISTUP ZASNOVAN NA KONTEKSTU

"They performed Kashmir, written by Page and Plant. Page played unusual chords on his Gibson."

bag-of-words →

perform
Kashmir
write
Page
Plant
play
chord
…

http://en.wikipedia.org/wiki/Kashmir_(song)
…was written by Jimmy Page and Robert Plant…
…performed by the band at almost every concert…

bag-of-words →

write
Jimmy
Page
Robert
Plant
perform
band
concert
…

similarity

similarity

+ 15 more candidate entities

http://en.wikipedia.org/wiki/Kashmir
…northwestern region of the Indian subcontinent.…
…became an important center of Hinduism and later of Buddhism…

bag-of-words →

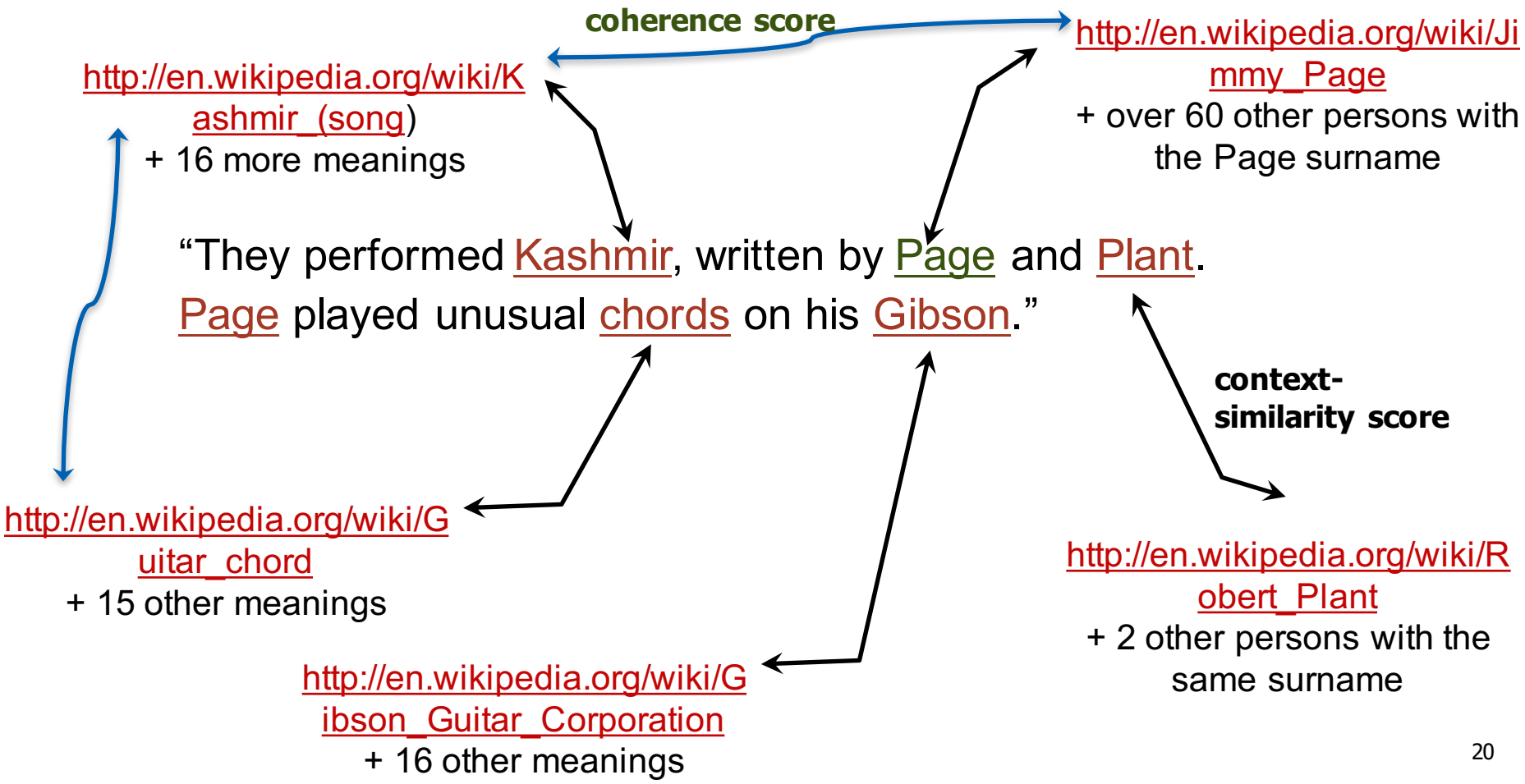
northwest
region
India
subcontinent
center
Hinduism
Buddhism
…

DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

- Sastoji se u istovremenom određivanju značenja (*disambiguation*) više reči/izraza (*mentions*) u tekstu koji je predmet analize
- Predstavlja proširenje pristupa zasnovanog na kontekstu:
  - pored računanja skora za kontekstualnu sličnost *(context similarity)* svakog *mention-entity* para, računa se takođe i koherentnost (*coherence score*) za sve parove entiteta
  - koherentnost je u ovom slučaju definisana kao semantička povezanost (*semantic relatedness*) razmatranih entiteta

DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

Nastavljajući sa istim primerom:



DISAMBIGUATION: COLLECTIVE DISAMBIGUATION APPROACH

- Ovaj pristup daje dobre rezultate ukoliko
  - postoji dovoljno veliki broj entiteta pomenutih u tekstu, i
  - pomenuti entiteti čine tematski homogen skup
- Greške se najčešće javljaju u slučaju da
  - tekst razmatra više nepovezanih ili slabo povezanih tema
  - entiteti sa kojima reči/izrazi (*mentions*) iz teksta mogu biti povezane, mogu formirati više tematski koherentnih grupa; na primer:

"Real Madrid and Barcelona edge out Manchester and Chelsea to secure trials for Argentine wonder-kid"

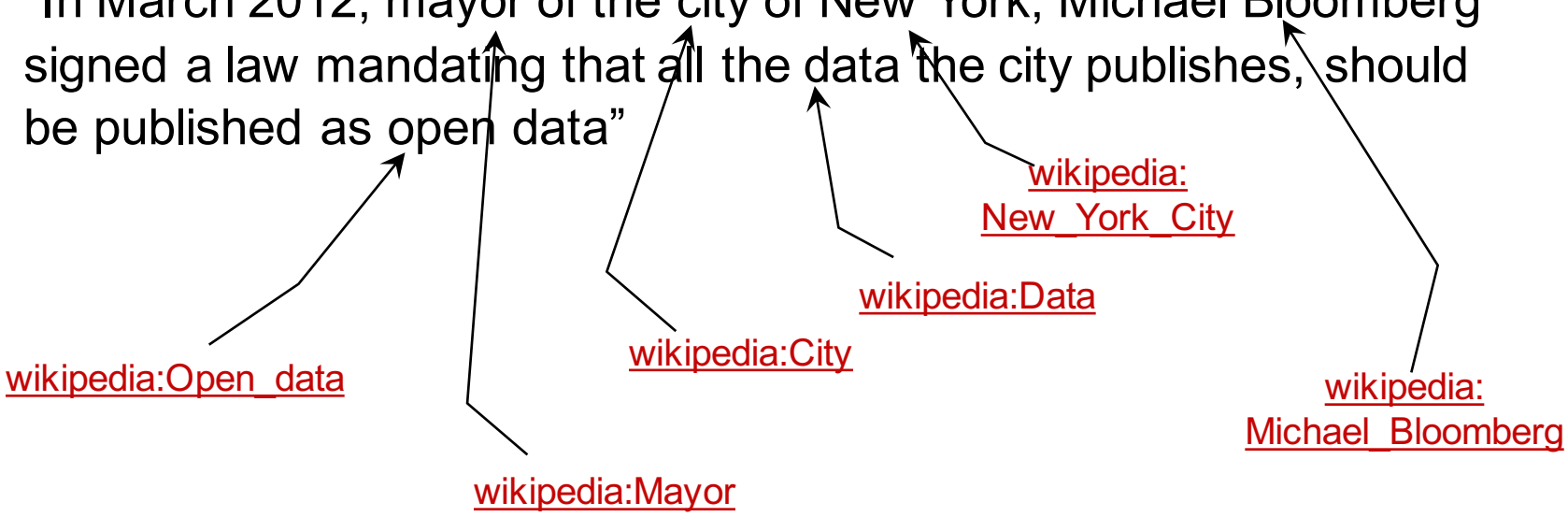
Ovde imamo potencijalno dve tematski koherentne grupe entiteta: lokacije (gradovi) i fudbalski klubovi

FILTERING

- Cilj ove faze je da se iz skupa rezultata uklone oni entiteti koji najverovatnije ne bi bili relevatni korisniku
  - npr., entiteti koji se odnose na neke opšte koncepte ili oni koji su samo marginalno povezani sa glavnom temom teksta

- Primer

"In March 2012, mayor of the city of New York, Michael Bloomberg signed a law mandating that all the data the city publishes, should be published as open data"



Performanse današnjih alata za semantičko indeksiranje

Tip teksta: novinski članci

	AlchemyAPI	dataTXT	DBpedia	Lupedia	Textrazor	THD	Yahoo!	Zemanta
p	**70.63**	39.20	26.93	57.98	49.21	32.50	61.24	35.58
r	14.05	**54.93**	42.21	29.90	51.66	40.10	9.65	7.78
f	23.43	45.75	32.88	39.45	**50.41**	35.90	16.68	12.77

p – precision; r – recall; f – F1 measure

(napomena: dataTXT je prerastao u komercijalni servis DandelionAPI)

Izvor: Rizzo et al. 2014. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web, 9th Edition of the Language Resources and Evaluation Conference (LREC'14)

Performanse današnjih alata za semantičko indeksiranje

Tip teksta: poruke sa Twitter-a

	AlchemyAPI	dataTXT	DBpedia	Lupedia	Textrazor	THD	Yahoo!	Zemanta
p	**72.22**	22.11	13.99	37.37	30.69	23.54	60.68	35.54
r	3.91	34.74	29.70	11.13	**34.89**	23.98	10.68	10.08
f	7.42	27.02	19.02	17.15	**32.65**	23.76	18.16	15.70

p – precision; r – recall; f – F1 measure

Izvor: Rizzo et al. 2014. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web, 9th Edition of the Language Resources and Evaluation Conference (LREC'14)

WIKILINKS CORPUS

- Najveći javno dostupan dataset za obuku algoritama nadgledanog m. učenja za semantičko indeksiranje, konkretno, prepoznavanje Wikipedia entiteta u tekstu
- URL: http://www.iesl.cs.umass.edu/data/wiki-links
- Osnovni podaci o dataset-u:
  - 10 miliona Web stranica
  - 3 miliona Wikipedia entiteta
  - 40 miliona jedinstveno identifikovanih pominjanja entiteta
  - publikovan 08.03.2013. od strane Google Research-a
- Više informacija u članku: Learning from Big Data: 40 Million Entities in Context

FREEBASE ANNOTATIONS OF SOCIAL MEDIA CONTENT

- Google Freebase Annotations of TREC KBA 2014 Stream Corpus
  - TREC – Text Retrieval Conference
  - KBA – Knowledge Base Acceleration
- URL: http://trec-kba.org/data/fakba1/index.shtml
- Osnovni podaci o korpusu:
  - 394M dokumenata sa bar jednim anotiranim Freebase entitetom
  - 9.4 milijarde reči/izraza (*mentions*) povezanih sa Freebase entitetima
  - anotacije su urađene automatski i samim tim nisu savršene
  - na osnovu ručno analiziranog slučajnog uzorka, procenjeno je:
    - ~9% reči/izraza (*mentions*) je povezano sa pogrešnim Freebase entitetima
    - ~8% reči/izraza (*mentions*) koji predstavljaju entitete nisu povezani sa odgovarajućim Freebase entitetima

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I