PREPOZNAVANJE ENTITETA U TEKSTU

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Pregled predavanja

- Glavni izazovi u domenu prepoznavanja entiteta
- Osnovne vrste pristupa za prepoznavanje entiteta
  - List lookup pristupi
  - Pristupi zasnovani na pravilima
  - Pristupi zasnovani na mašinskom učenju
  - Pristupi zasnovani na m. učenju i bazama znanja
- Korisni Web resursi vezani za ovu temu

Glavni izazovi u domenu prepoznavanja entiteta

- Pravilna identifikacija segmenata teksta kojima su entiteti predstavljeni (tzv. chunking)
  - entiteti mogu biti predstavljeni jednom reči (npr. *MIT*) ili nizom reči (*Massachusetts Institute of Technology*)

- Zaključivanje da određeni segment teksta stvarno predstavlja entitet
  - posebno nezgodno u slučajevima kad se višeznačne reči nađu na početku rečenice (npr., May, Galaxy, …)

Glavni izazovi u domenu prepoznavanja entiteta

- Određivanje tipa entiteta

Group (Team) vs. Location:

"**England** won the World Cup" vs.

"The World Cup took place in **England**"

Company vs. Artefact:

"having shares in **BBC**" vs. "watching **BBC**"

"staying in **Hotel California**" vs. "listening to **Hotel California**"

Location vs. Organisation:

"she met him at **Heathrow**" vs. "the **Heathrow** authorities"

Glavni izazovi u domenu prepoznavanja entiteta

- Prepoznavanje segmenata teksta koji se odnose na *isti* entitet
  - Problem: različiti načini referenciranja na isti entitet; primeri:
    - John Smith; Mr Smith; John
    - *UMBC*; *University of Maryland Baltimore County*

- Održavanje ažurnim lista/rečnika koji sadrže nazive entiteta (potrebni za većinu aktuelnih sistema)

Osnovne vrste pristupa prepoznavanju entiteta

- *List lookup* pristupi
  - zasnovani na korišćenju rečnika ili lista imena entiteta
- Pristupi zasnovani na pravilima
  - *shallow parsing* pristupi
  - pristupi zasnovani na regularnim izrazima
- Pristupi zasnovani na mašinskom učenju
- Pristupi zasnovani na m. učenju i bazama znanja
- Hibridni pristupi
  - kombinuju dva ili više navedenih pristupa
  - najčešće se primenjuju u praksi

LIST LOOKUP PRISTUPI

List lookup pristupi

- Primenjuju se kad imamo liste (ili rečnik) imena relevantnih entiteta
  - npr., liste kompanija i/ili eksperata iz određene branše
- Prepoznaju samo one entitete čija imena su prisutna u listama/rečniku

List lookup pristupi

- Dva pristupa mečiranju imena:
  - *Exact matching*: zahteva potpuno poklapanje reči/termina iz teksta i imena datih u listama/rečniku
  - *Approximate matching*: proširuje exact matching tehnikama za približno poređenje stringova
    - Primer: *Levenshtein distance* (edit distance) - min. broj izmena potreban da bi se jedna reč transformisala u drugu; dozvoljene izmene su umetanje, brisanje ili zamena jednog karaktera

      Lev (machine, marine) = 2
      - brisanje karaktera ‘c’
      - zamena karaktera ‘h’ -> ‘r’

List lookup pristupi

Gazetteer

- Komponenta velike većine NLP framework-a
- Koristi liste imena da bi prepoznao entitete u tekstu
  - gazetteer liste su obični tekstualni fajlovi, sa jednim podatkom (imenom) u svakoj liniji
  - Svaka lista sadrži skup imena,
    kao što su imena gradova, organizacija, dana u nedelji,...
  - index fajl se koristi za pristup ovim listama;
  - Termin koji odgovara imenu iz neke od ovih lista biće anotiran glavnim (major type) tipom i podtipom (minor type) liste kojoj to ime pripada.
  - Primer: “Belgrade”

    Anotacija: majorType = location, minorType = city

List lookup pristupi

- Prednosti:
  - Jednostavnost
  - Brzina (brži u odnosu na ostale pristupe)
  - Nezavisnost od jezika
  - Mogućnost jednostavne adaptacije na nove vrste teksta
- Nedostaci:
  - Kreiranje i održavanje lista imena (rečnika)
  - Ne mogu da prepoznaju entitete u slučaju slabog preklapanja imena u listama i u tekstu
  - Nemaju mogućnost razumevanja entiteta u kontekstu i razrešavanja dvosmislenosti (ili višeznačnosti)

PRISTUPI ZASNOVANI NA PRAVILIMA

Pristupi zasnovani na pravilima: *shallow parsing*

- Zasniva se na identifikaciji i primeni heurističkih, iskustvenih ‘pravila’ koja se tiču
  - strukture rečenice,
  - tipično korišćenih izraza i fraza u tekstu

- Osnovni koraci ovog pristupa:
  - identifikovati uobičajene jezičke formulacije i tipove entiteta na koje se odnose
  - predstaviti ih u formi jezičnih paterna
  - formalizovati identifikovane paterne korišćenjem jezika za modelovanje pravila

Pristupi zasnovani na pravilima: *shallow parsing*

Primer – prepoznavanje lokacije:

CapWord + {City, Forest, Center}

e.g. Sherwood Forest

CapWord + {Street, Boulevard, Avenue, Crescent, Road}

e.g. Portobello Street

“to the” COMPASS “of” CapWord

e.g. *to the south of Boston*

“based in” CapWord

e.g. *based in Boston*

CapWord “is a” (ADJ)? GeoWord

e.g. *Boston is a friendly city*

Pristupi zasnovani na pravilima: *shallow parsing*

- Primer: poznati Hearst paterni za prepoznavanje entiteta

such NP as {NP,} {or | and} NP*

... works by such authors as Herrick, Goldsmith, and Shakespeare

NP {,} including {NP,} {or | and} NP*

All common-law countries, including Canada and England ...

NP {,} especially {NP,} {or | and} NP*

... most European countries, especially France, England, and Spain.

M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proc. of the 14th Int'l Conference on Computational Linguistics, Nantes France, 1992 (link).

Pristupi zasnovani na pravilima: *shallow parsing*

- Korišćenje pravila za formalizaciju jezičkih paterna
- Na primer, JAPE* omogućuje definisanje pravila oblika:

  templejt => akcija

  - Leva strana pravila sastoji se od jednog ili više templejta koji se mečiraju sa tekstom koji se analizira
  - Desna strana pravila sastoji se od iskaza za anotaciju mečiranih segmenata teksta i manipulisanje tim anotacijama

*JAPE je sastavni deo GATE Java framework-a za analizu teksta

Pristupi zasnovani na pravilima: *shallow parsing*

- Primer JAPE pravila

```
Rule: Location_1 //CapWord + {City, Forest, Center}
(
  {Token.kind == word, Token.category == NP,
                              Token.orth == ″upperInitial″}
  {Token.kind == ″space″}
  ( {Token.string == ″City″} |
    {Token.string == ″Forest″} |
    {Token.string == ″Center″}
   )
):loc
 -->
 :loc.Location = {rule = ″Location_1"}
```

Pristupi zasnovani na pravilima: *shallow parsing*

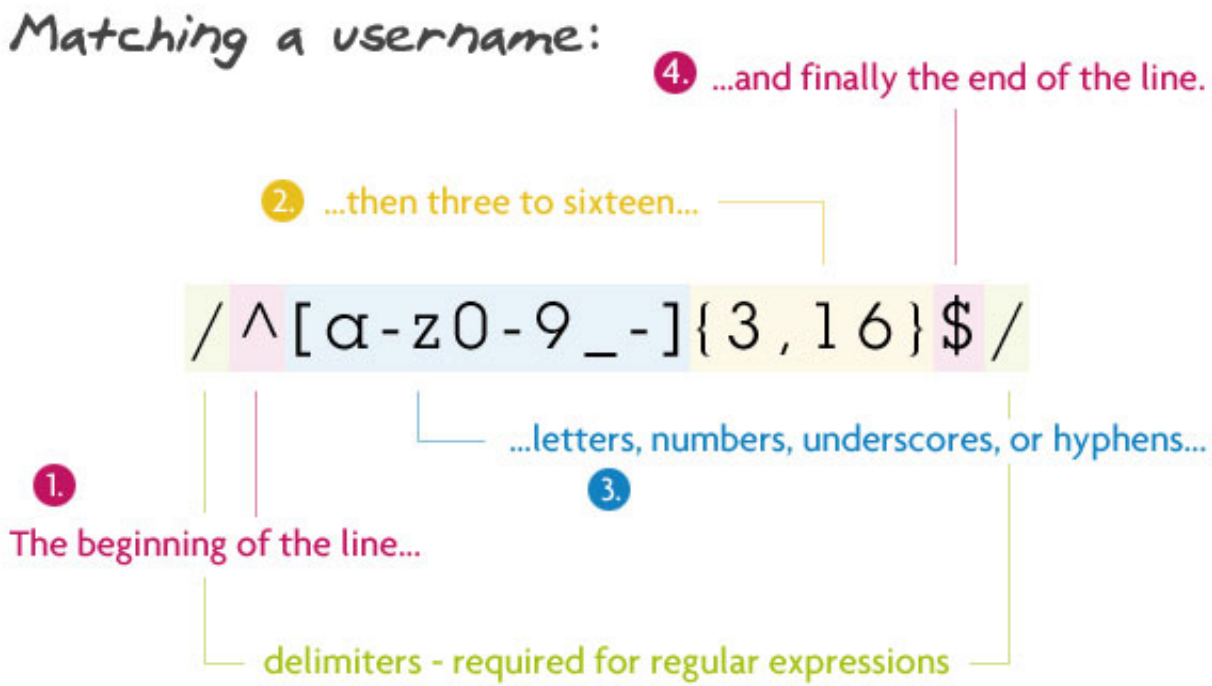
Problemi sa *Shallow Parsing* pristupom

- Identifikacija pouzdanih jezičkih paterna
- Neodređenost značenja početnog prvog slova u rečenici
  - Da li označava samo početak rečenice ili je takođe sastavni deo naziva entiteta?
  - Npr. [All American Bank] vs. All [State Police]
- Strukturna višesmislenost
  - Npr. [Cable and Wireless] vs. [Microsoft] and [Dell]
  - Npr. [Center for Computational Linguistics] vs.

    message from [City Hospital] for [John Smith]

Pristupi zasnovani na pravilima: regularni izrazi

- Pogodni za entitete čija tekstualna reprezentacija mora da sledi neku dobro definisanu strukturu i formu
- Primer: regularni izraz za prepoznavanje username-a



Izvor: http://net.tutsplus.com/tutorials/other/8-regular-expressions-you-should-know/

PRISTUPI ZASNOVANI NA MAŠINSKOM UČENJU

Pristupi zasnovani na mašinskom učenju

- Najčešće zasnovani na primeni *nadgledanog* m. učenja, odnosno klasifikaciji
- Osnovna ideja:
  - Program uči karakteristike/osobine koje odlikuju entitete određenog tipa
  - Osobine entiteta se određuju na osnovu termina kojima su entiteti predstavljeni u tekstu, kao i termina koji čine njihov kontekst/okruženje
- Preduslov:
  - Postojanje dovoljno velikog skupa podataka za trening tj. korpusa anotiranih/obeleženih dokumenata

Primena nadgledanog m. učenja

Razmotrićemo osnovne elemente procesa nadgledanog m. učenja pri prepoznavanju entiteta u tekstu:

- (obeleženi) podaci za treniranje modela
- atributi (features) za formiranje modela
- selekcija algoritma m. učenja
- evaluacija modela

Primena nadgledanog m. učenja: podaci za trening

Primer teksta anotiranog za potrebe “obuke” algoritma nadgledanog učenja:

Unlike <PERSON>Robert</PERSON>, <PERSON>John Briggs Jr</PERSON> contacted <ORGANIZATION>Wonderful Stockbrockers Inc</ORGANIZATION> in <LOCATION>New York</LOCATION> and instructed them to sell his <NUMBER>100</NUMBER> shares in <ORGANIZATION>Acme</ORGANIZATION>

Očigledno, priprema podataka za obučavanje algoritma je prilično zahtevna…

Primena nadgledanog m. učenja: podaci za trening

- Izvori gotovih "obeleženih" podataka tj. anotiranih tekstova za obuku algoritama
  - Stručne konferencije na temu analize i razumevanja teksta
    - Message Understanding Conference (MUC): MUC06 i MUC07 datasets
    - Conference on Computational Natural Language Learning (CoNLL): CoNNL-2002 i CoNNL-2003 datasets
  - Stručne asocijacije u domenu obrade prirodnog jezika
    - Linguistic Data Consortium održava katalog lingvističkih dataset-ova
  - Istraživačke grupe i pojedinci
    - Twitter NER – dataset korišćen za treniranje modela za prepoznavanje entiteta u tweet porukama (više o tome ovde)
    - GeneTag dataset – publikovan od strane US National Center for Biotechnology Information

Primena nadgledanog m. učenja: atributi

- Širok spektar atributa koji se mogu koristiti
  - Atributi koji se odnose na pojedinačne reči:
    - dužina reči;
    - prisutnost velikih slova;
    - vrsta reči;
    - učestanost pojavljivanja reči u dok. za trening;
    - prisutnost znakova interpunkcije;
    - pozicija reči u rečenici,…
  - Atributi koji se odnose na okruženje reči:
    - opseg okruženja;
    - vrsta reči u okruženju i sl.

Primena nadgledanog m. učenja: atributi

- Izbor atributa
  - zavisi od vrste teksta koji je predmet analize (npr. tweet poruke vs. novinski članci vs. stručni tekstovi)
  - Primer: atributi korišćeni za prepoznavanje entiteta u Twitter porukama, predstavljeni u radu (str. 4):

    Edgar Meij, Wouter Weerkamp, Maarten de Rijke. (2012). Adding semantics to microblog posts. In Proceedings of the 5th International Conference on Web Search and Web Data Mining, pp. 563-572.

Primena nadgledanog m. učenja: atributi

Izabrani skup atributa se koristi za predstavljanje pojedinačnih reči i/ili izraza od kojih je tekst sačinjen

Jednostavan primer

Pretpostavimo da smo izabrali sledeće atribute:

- Boolean atribut koji ukazuje da li reč počinje velikim slovom
- Numerički atribut koji predstavlja dužinu reči
- Nominalni atribut koji predstavlja reč napisanu malim slovima

Rečenica: “The apple sign makes Apple laptops easily recognizable.”

će imati sledeću reprezentaciju:

<true, 3, "the">, <false, 5, "apple">, <false, 4, "sign">, <false, 5, ”makes">, <true, 5, ”apple">, … , <false, 12, ”recognizable">

Primena nadgledanog m. učenja: izbor modela

- Najčešće korišćeni modeli/algoritmi*
  - Stabla odlučivanja (Decision trees)
  - Hidden Markov Models (HMM)
  - Maximum Entropy classification
  - Support Vector Machines (SVM)
  - Conditional Random Fields (CRF)

*samo informativno, ovo su dosta složeniji modeli od onih koje smo radili

Primena nadgledanog m. učenja: evaluacija modela

- Evaluacija je zasnovana na metrikama tipičnim za zadatak klasifikacije:
  - Preciznost (Precision), Odziv (Recall), F mera (F measure)

- Softverski okviri (frameworks) za poređenje različitih alata/servisa za prepoznavanje entiteta (tzv. benchmarking):
  - NERD (Named Entity Recognition and Disambiguation): http://nerd.eurecom.fr/
  - GERBIL (General Entity Annotator Benchmark): http://gerbil.aksw.org/gerbil/

Alternativni oblici m. učenja

- Problem: priprema dovoljno velikog skupa anotiranih dokumenata (korpusa) potrebnog za trening, je prilično zahtevan zadatak

- Usled toga, polu-nadgledano i nenadgledano m. učenje se često nameću kao alternative
  - ovi pristupi ne zahtevaju anotirani skup dokumenata
  - tradicionalno su imali slabije performanse u odnosu na pristupe nadgledanog m. učenja, ali su nova rešenja sve bolja

Polu-nadgledano m. učenje

- *Bootstrapping* je popularna tehnika polu-nadgledanog m. učenja
  - Podrazumeva mali stepen “nadgledanja”, tipično u formi inicijalno zadatog skupa primera, potrebnog za pokretanje procesa učenja
- Ilustracije radi, razmotrimo primer sistema namenjenog prepoznavanju proizvoda koji se pominju u tekstu
  - inicijalno, korisnik zadaje mali broj primera tj. naziva različitih proizvoda;
  - sistem analizira tekst i identifikuje elemente koji karakterišu kontekst zadatih primera; zatim, identifikuje druga pojavljivanja proizvoda na osnovu identifikovanih karakteristika konteksta;
  - proces učenja se ponovo primenjuje polazeći od novo-otkrivenih instanci (proizvoda), što vodi otkrivanju novih relevantnih konteksta;
  - ponavljajući ovaj proces, veliki broj proizvoda i konteksta u kojima se oni pojavljuju će biti otkriven.

Polu-nadgledano m. učenje

Preporuka:

Predavanje Tom Mitchell-a pod nazivom

Semisupervised Learning Approaches

održano u okviru

Autumn School 2006: Machine Learning over Text and Images

URL: http://videolectures.net/mlas06_mitchell_sla/

PRISTUPI ZASNOVANI NA M. UČENJU I BAZAMA ZNANJA

Pristupi zasnovani na m. učenju i bazama znanja

- Kombinacija nadgledanog m. učenja (klasifikacija) i znanja sadržanog u bazama znanja na Web-u
- Najčešće korišćene baze znanja: Wikipedia, Freebase, DBpedia
- Dodatne, napredne mogućnosti ovog pristupa:

  pored prepoznavanja tipa entiteta, omogućuju i

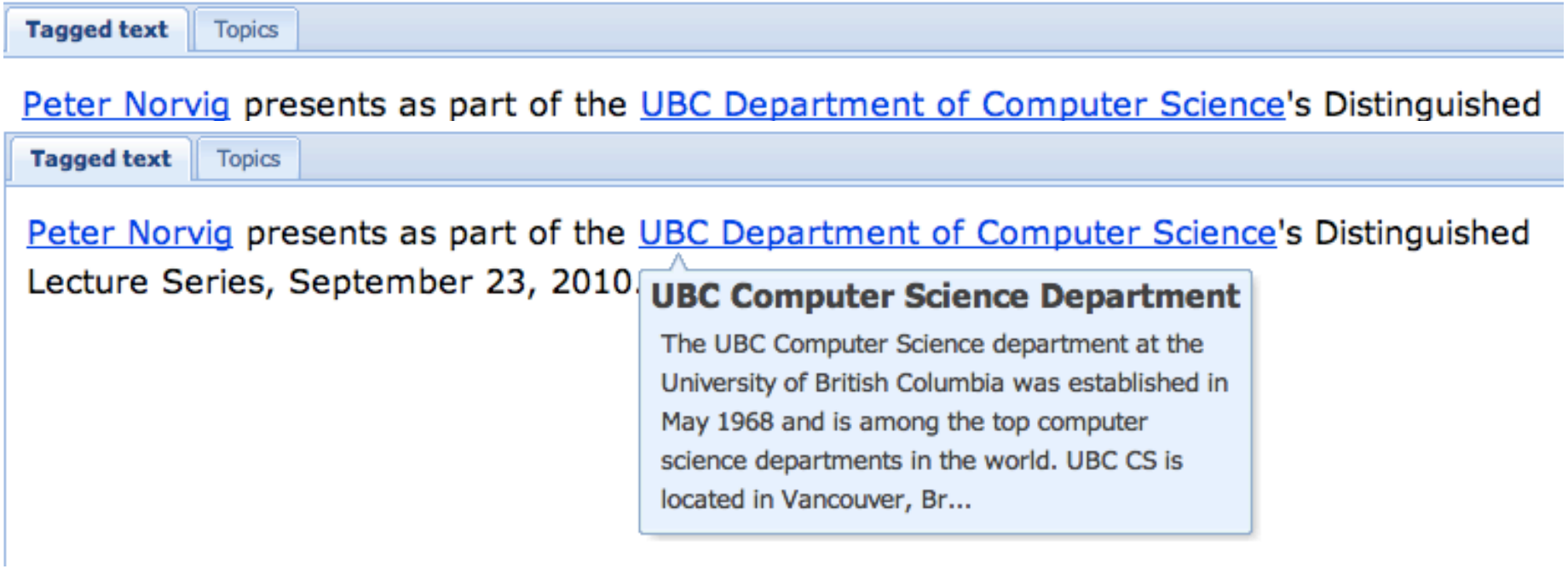
  jedinstveno identifikovanje entiteta (disambiguation)

Pristupi zasnovani na m. učenju i bazama znanja

- Klasično prepoznavanje entiteta u tekstu:

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

- Prepoznavanje entiteta uz korišćenje baze znanja:



Pristupi zasnovani na m. učenju i bazama znanja

- Dodatna pogodnost ovih pristupa je jednostavnije kreiranje skupa podataka za obučavanje algoritma
- Na primer, u slučaju Wikipedia-e:
  - Svaki termin koji predstavlja interni link u Wikipedia-i – zvaćemo ga *anchor* – tretira se kao potencijalni entitet
  - Svaki *anchor* obezbeđuje nekoliko trening instanci:
    - pozitivni primer: destinacija linka (Wikipedia stranica), odnosno "pravo" značenje datog anchor termina u datom kontekstu
    - negativni primeri: sve ostale moguće destinacije linka, odnosno ostala moguća značenja datog anchor termina

Kreiranje dataset-a za obuku algoritma korišćenjem internih Wikipedia linkova – ilustracija pristupa

Za termin (anchor) *tree* postoji 26 mogućih destinacija (tj. značenja), što daje 1 poz. primer i 25 neg. primera za trening algoritma

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 expanded nodes are added to a LIFO stack for exploration.

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%
…		

Ovim pristupom se, npr., od svega 500 Wikipedia članaka, može dobiti dataset od > 50,000 instanci

Pristupi zasnovani na m. učenju i bazama znanja

Osnovni koraci u procesu prepoznavanja entiteta u tekstu:

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 (2)

- Primer

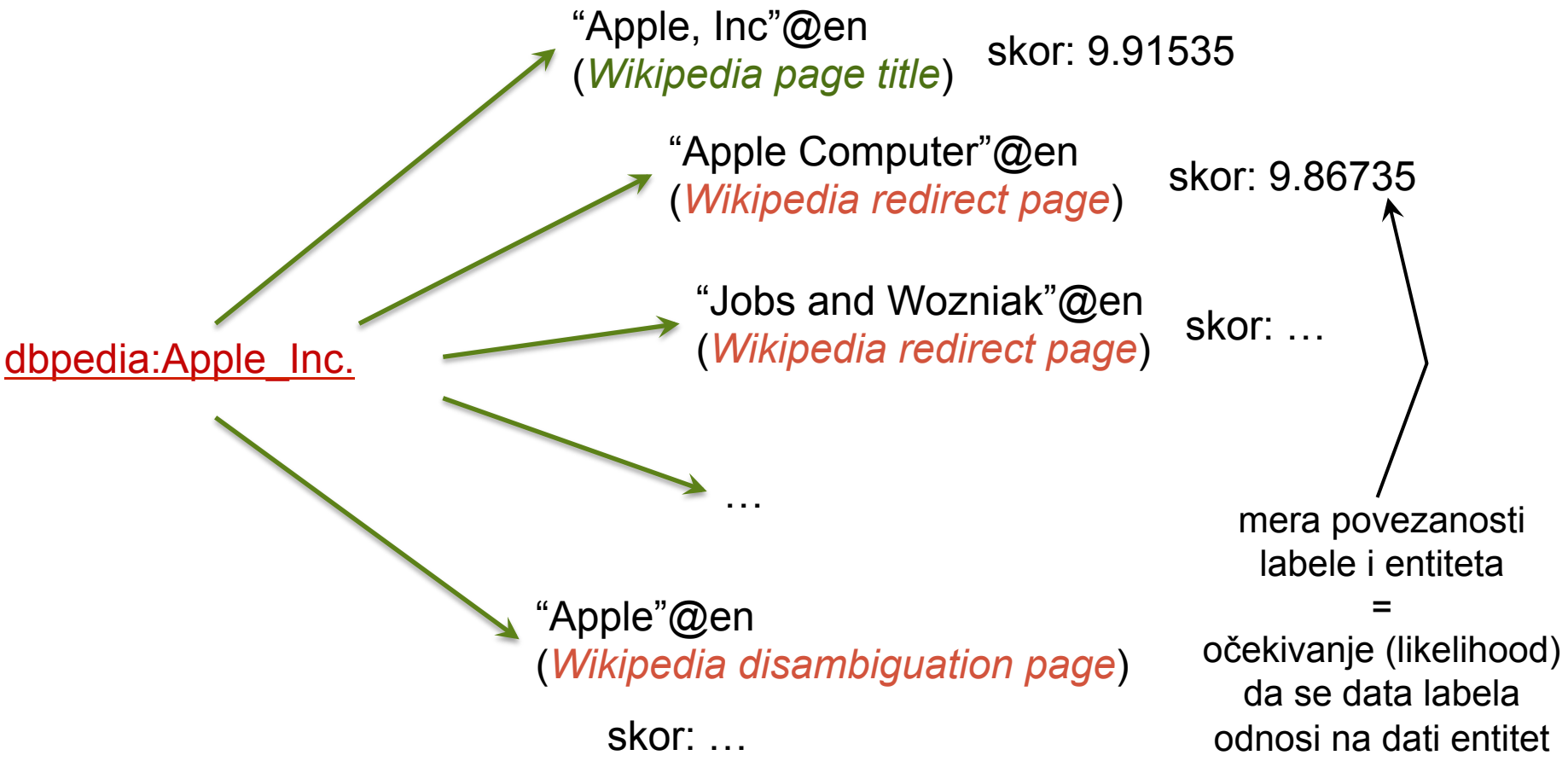
"They performed Kashmir, written by Page and Plant.
Page played unusual chords on his Gibson."

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
…

Entity spotting & candidate selection (3)

- Ova faze 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://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: 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*
- Kontekst se obično predstavlja kao skup reči (bag-of-words) i poređenje se vrši primenom neke metrike za računanje sličnosti
  - često korišćene metrike: Cosine similarity, weighted Jaccard coefficient, Wikipedia links-based measure

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

+ 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 →

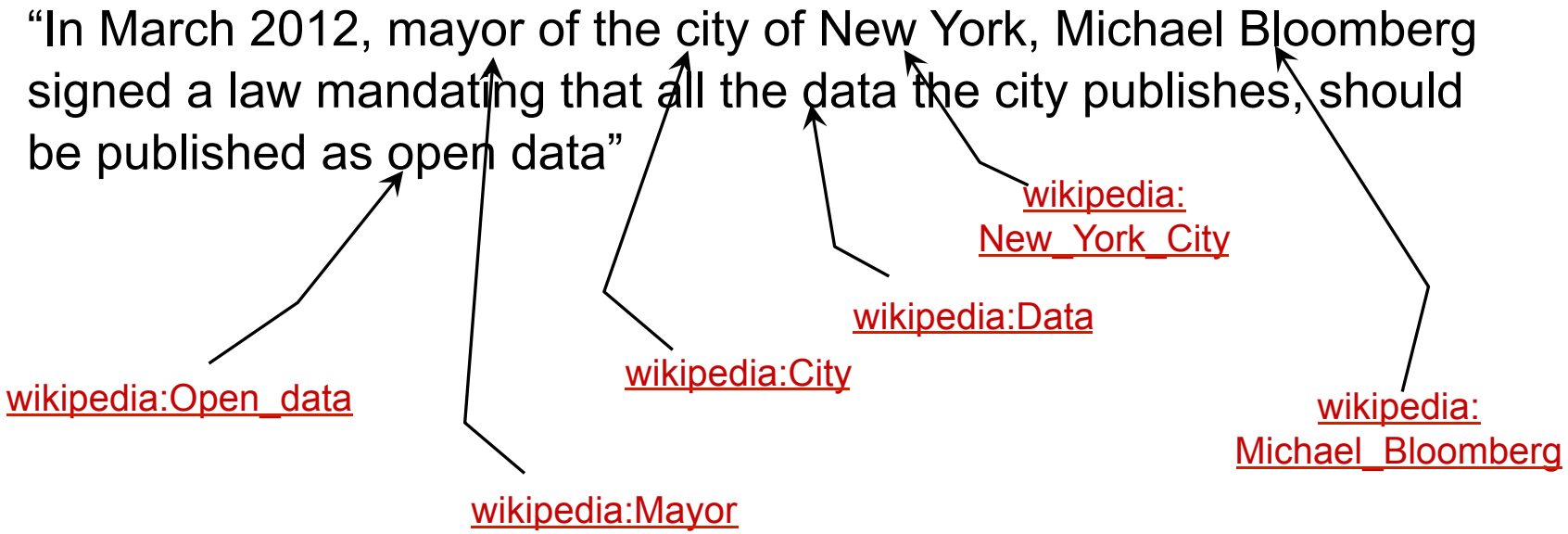
northwest
region
India
subcontinent
center
Hinduism
Buddhism
…

similarity

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



Alati koji implementiraju opisani pristup

- Wikipedia Miner - obezbeđuje niz servisa:
  - *wikify* – identifikuje entitete iz Wikipedia-e u zadatom tekstu
  - *compare* – utvrđuje i objašnjava povezanost između dva Wikipedia entiteta
  - *suggest* – predlaže entitete koji su semantički slični/srodni zadatim entitetima
- TagMe – obezbeđuje sledeće servise:
  - *tagging* – identifikuje entitete iz Wikipedia-e u zadatom tekstu
  - *spotting* – detektuje relevantne termine u tekstu (ne povezuje ih sa Wikipedia entitetima)
  - *relating* – određuje semantičku povezanost dva zadata entiteta

KORISNI WEB RESURSI

Korisni Web resursi

Wikilinks Corpus

- Najveći javno dostupan dataset za obuku algoritama nadgledanog m. učenja za prepoznavanje (Wikipedia) entiteta u tekstu
- URL: http://www.iesl.cs.umass.edu/data/wiki-links
- Osnovni podaci:
  - 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

Korisni Web resursi

Softverski alati za prepoznavanje entiteta u tekstu

- AlchemyAPI: http://www.alchemyapi.com/tools/
- Open Amplify: http://www.openamplify.com/quickstart
- Text Razor: http://www.textrazor.com/
- TextWise: http://www.textwise.com/
- TagMe: http://tagme.di.unipi.it/
- Wikipedia Miner: http://wikipedia-miner.cms.waikato.ac.nz/
- Denote: http://denote.io/

Korisni Web resursi

- Aktuelni Java okviri (frameworks) za analizu i razumevanje teksta
  - Stanford CoreNLP: http://nlp.stanford.edu/software/corenlp.shtml
  - Apache OpenNLP: http://opennlp.apache.org/
  - Apache Stanbol: http://stanbol.apache.org/
  - GATE: http://gate.ac.uk/
  - LingPIPE: http://alias-i.com/lingpipe/

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

http://goo.gl/cqdp3I