EKSTRAKCIJA INFORMACIJA IZ TEKSTA

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Pregled izlaganja

Ekstrakcija informacija (EI) iz teksta

- Osnovni pojmovi
- Primeri primene
- Osnovni tipova zadataka EI
- Činioci koji utiču na performanse procesa EI
- Karakteristike i performanse osnovnih tipova zadataka EI

Ekstrakcija informacija

Tehnologija zasnovana na

analizi prirodnog jezika

sa ciljem ekstrakcije informacija o

predefinisanim tipovima

entiteta, relacija i/ili događaja

Ekstrakcija informacija

Primer:

"Novak Djokovic extended his unbeaten record at the 2012 China Open in Beijing on Tuesday, but needed three sets to defeat qualifier Michael Berrer of Germany.

The Serb progressed to the second round with a 6-1, 6-7 (3/7), 6-2 definitive victory…"

Informacije koje bi se mogle estrahovati iz ovog teksta:

- Novak Djokovic; 2012 China Open; Beijing; Tuesday; …; the Serb.
- *Novak Djokovic* i *the Serb* se odnose na isti entitet
- unbeaten (record); definitive (victory)
- (Novak Djokovic) opponent (Michael Berrer)
- qualify for the 2nd round of 2012 China Open

EI vs. Information Retrieval (IR)

- EI & IR: dva slična, ali dosta različita procesa:
  - IR sistem pronalazi (potencijalno) relevantne tekstove i prezentuje ih korisniku
    - Primer: Web pretraživači kao što su Google i Bing
  - EI sistem analizira tekstove i prezentuje *samo segmente informacija* (izvučene iz teksta) za koje korisnik može biti zainteresovan
    - Primer: IBM Watson

Primeri primene

Napredno pretraživanje Web-a

"*Google is undergoing a major, long-term overhaul of its search-engine, using what's called semantic Web search…*

*…people familiar with the initiative say that Google users will able to browse through the company's '**knowledge graph**,' or its ever-expanding **database** of information **about 'entities'** – people, places and things – **the 'attributes' of those entities** and **how** different **entities are connected** to one another*."

What Google's Search Changes Might Mean for You

Wall Street Journal, March 14, 2012

Primeri primene

Napredno pretraživanje Web-a

*"At the core of Microsoft's work to create a state-of-the-art Bing digital assistant is **Satori**, a **knowledge repository** of more than a billion objects digested in the past 3.5 years…*

*…Satori catalogs **entities** and the **associated data and relationships** among them...*

*…Satori is a **self-learning system** that is running every day and learning more, adding 28,000 DVDs of content every day…*

…Bing search and Windows already are using Satori's knowledge repository…"

Microsoft's Bing seeks enlightenment with Satori
CNET News, July 30, 2013

Primeri primene

Napredno pretraživanje društvenih mreža

"Facebook is building a rich stock of knowledge that could make its software smarter and boost the usefulness of its search engine…

*…Entities such as colleges and employers are **learned from data** typed **in profile pages**; businesses, movies, fictional characters, and other concepts are **learned from fan pages** created by Facebook users. … **analyzing** many **employment histories** on the site allows Facebook's search engine to know that a search for "software engineers" should also return people who say they are "coders."*

Facebook Nudges Users to Catalog the Real World

MIT Technology Review, February 27, 2013

Primeri primene

Poslovna analitika

- Ekstrakcija informacija od interesa za poslovno odlučivanje, poput informacija o
  - geo-političkim i makro-ekonomskim dogadjajima, i/ili
  - dogadjajima relevantnim za pojedine kompanije i brendove
- Estrahovane informacije predstavljaju ulaz za analitičke sisteme tipično zasnovane na Business Rules Engine
- Primeri:
  - RavenPack News Analytic (http://www.ravenpack.com/)

Primeri primene

Marketing & PR

- *social media monitoring*
- *reputation management*
- *social customer relationship management*
- primeri:
  - Safesforce Marketing Cloud (ex. Radian6; link)
  - Lithium Social Intelligence product (link)

Primeri primene

- Marketing & PR (nastavak)

"*It's fine to have a Facebook fan page or count 'Likes' but really, that's looking at a tiny, narrow portion of a social strategy*"

"There are millions of conversations going on related to retailers, their products and their brands. They need to be aware of and participate in that breadth of the social web."

Izvor: http://bit.ly/yg7hTO

Primeri primene

Online reklamiranje

- analiza sadržaja web stranice radi ekstrakcije:
  - entiteta (osoba, lokacija, kompanija, brendova,…),
  - tipa teksta,
  - emocija sadržanih u tekstu,
  - poruke koju tekst nastoji da iskomunicira
- estrahovane informacije se koriste za preporuku reklama za datu web stranu
- primer: ADmantX (http://www.admantx.com/)

Osnovni tipovi IE zadataka

- Prepoznavanje imenovanih entiteta (*Named Entity recognition*)
  - može se odnositi na različite vrste entiteta (ljudi, organizacije, datumi, valute i sl)
- Razrešavanje koreferenci *(Co-reference resolution)* – obuhvata:
  - *anaphoric resolution*
    - Npr. utvrditi da se u tekstu: "*Tom* is my best friend. I know *him* since we were kids." zamenica 'him' odnosi na imenicu 'Tom';
  - *proper noun resolution*
    - Npr. utvrditi da sledeće imenice označavaju isti entitet: 'IBM', 'IBM Europe', 'International Business Machines Ltd.', . . .

Osnovni tipovi IE zadataka

- Prepoznavanje opisa entiteta (*Descriptions resolution*)
  - Koje atribute entiteti imaju?
- Prepoznavanje relacija (*Relations resolution*)
  - Koje relacije postoje među entitetima?
- Prepoznavanje događaja (*Events resolution)*
  - Identifikacija događaja u kojima entiteti učestvuju

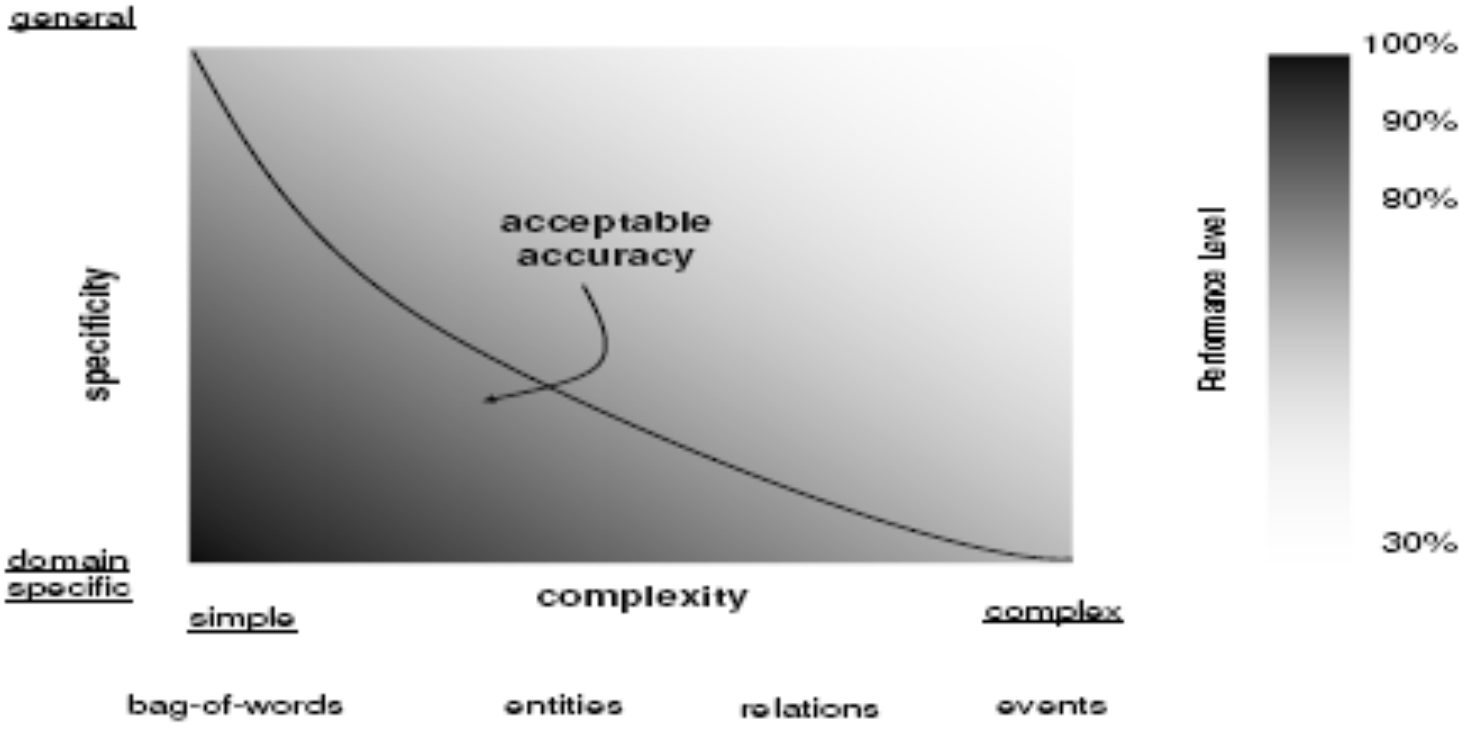
EI: Determinante performansi

Specifičnosti konkretnog EI zadatka

- Tip teksta – vrsta teksta sa kojim se radi; npr. novinski članci ili email poruke ili poslovni izveštaji ili naučni radovi i sl.
- Tema (ili domen) – šire definisan opseg tema (domen) kome sadržaj teksta pripada
- Stil pisanja – nivo formalnosti jezika, korišćenje stručne terminologije i sl.
- Konkretni tipovi informacija za koje je korisnik zainteresovan
  - Npr., osobe, kompanije, akvizacija neke kompanije,...

EI: Determinante performansi

Zavisnost performansi od specifičnosti i kompleksnosti EI zadatka



Preuzeto iz: H. Cunningham, Information Extraction, Automatic. Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier. 2005.

EI: Procena performansi

Najčešće korišćene mere za procenu performansi EI (ili IR) sistema:

- ***Preciznost (precision)*** – Da li su svi estrahovani segmenti informacija relevantni?
- ***Odziv (recall)*** – Da li su svi relevantni segmenti informacija prepoznati?

	Tačno	Pogrešno
Estrahovani	A	B
Nisu estrahovani	C	D

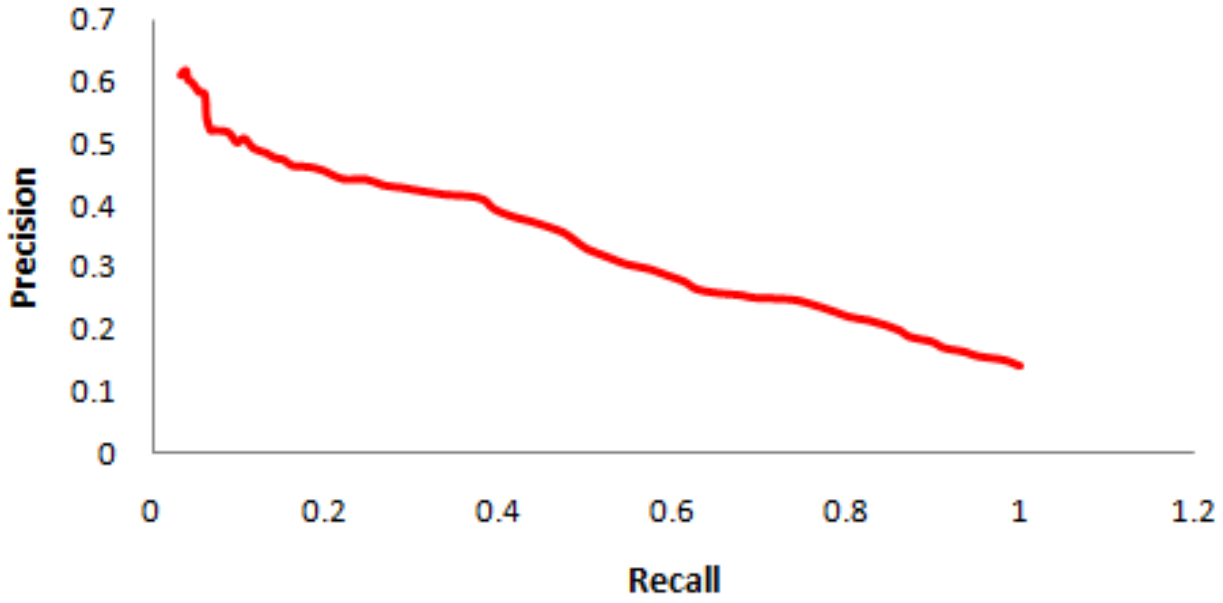
Precision = A / (A U B)

Recall = A / (A U C)

EI: Evaluacija performansi

Preciznost i odziv su često u “konfliktu”:

- Možemo razviti sistem koji neće praviti mnogo grešaka (visoka preciznost), ali će propustiti da prepozna puno relevantnih informacija (nizak odziv);
- Alternativno, možemo staviti akcenat na odziv i propustiti manje relevantnih informacija, ali po ceni pravljenja više grešaka.



Izvor: http://groups.csail.mit.edu/cb/struct2net/webserver/images/prec-v-recall-v2.png

KARAKTERISTIKE I PERFORMANSE OSNOVNIH ZADATAKA EI

Prepoznavanje imenovanih entiteta

- Funkcioniše na nivou ljudskih performansi
  - Do ~ 95% preciznosti
  - Ljudi ne obavljaju ovaj zadatak sa 100% tačnosti
- Slabo domenski zavistan zadatak
  - Npr., promena fokusa sa finansijskih vesti na vesti iz neke druge oblasti, bi zahtevala samo sitnije izmene
- Veoma zavisi od tipa teksta koji se analizira
  - Npr., promena fokusa sa vesti na naučne radove bi zahtevala prilično velike izmene

Razrešavanje koreferenci

- Glavni domen primene:
  - pridruživanje deskriptivnih informacija “rasutih” po tekstu entitetima na koje se odnose
- Performanse:
  - Neprecizan proces
  - Rezultati značajno variraju od domena do domena (domenski zavistan zadatak)
  - Zavisno od domena, preciznost je na nivou 50-60%

Prepoznavanje opisa entiteta

- Formalni naziv: *Template Element construction*
- Koristi rezultate prethodna dva zadatka
- Dodaje deskriptivne informacije entitetima
- Performanse
  - Dobri rezultati se kreću oko 80%
  - Na zadacima ovog tipa, performanse ljudi se kreću oko 95%
- Portabilnost
  - Slabo domenski zavistan zadatak
  - Veoma zavisi od tipa teksta

Prepoznavanje relacija

- Formalni naziv: *Template Relation construction*
- Odnosi se na identifikaciju relacija između imenovanih entiteta,
  - Na primer,

    relacija zaposleni između osobe i kompanije,

    porodične relacije između dve ili više osoba,

    relacija filijala između dveju kompanija, …
- Performanse
  - Dobri rezultati se kreću oko 75%
- Portabilnost
  - Slabo domenski zavistan

Prepoznavanje događaja

- Formalni naziv: *Scenario Templates recognition*
- Povezuje entitete, njihove opise i relacije u opis događaja
- Performanse
  - Najbolji sistemi dostižu ~ 60%
- Portabilnost
  - Domenski zavistan
  - Usko vezan za scenarije od interesa za datog korisnika

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

http://goo.gl/cqdp3I