Big Text: from Names and Phrases to Entities and Relations

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
http://www.mpi-inf.mpg.de/~weikum/
From Natural-Language Text to Knowledge

more knowledge, analytics, insight

Web Contents

Knowledge

knowledge acquisition

intelligent interpretation
Web of Data & Knowledge (Linked Open Data)

> 50 Bio. subject-predicate-object triples from > 1000 sources

Web of Data & Knowledge

> 50 Bio. subject-predicate-object triples from > 1000 sources

- 10M entities in 350K classes
- 120M facts for 100 relations
- 100 languages
- 95% accuracy

- 4M entities in 250 classes
- 500M facts for 6000 properties
- live updates

- 600M entities in 15000 topics
- 20B facts

- 40M entities in 15000 topics
- 1B facts for 4000 properties
- core of Google Knowledge Graph

Web of Data & Knowledge

> 50 Bio. subject-predicate-object triples from > 1000 sources

Yimou_Zhang type movie_director
Yimou_Zhang type olympic_games_participant
movie_director subclassOf artist
Yimou_Zhang directed Shanghai Triad
Li Gong actedIn Shanghai Triad
Yimou_Zhang memberOf Beijing_film_academy
validDuring [1978, 1982]
Yimou_Zhang „was classmate of“ Kaige_Chen
Yimou_Zhang „had love affair with“ Li_Gong
Li_Gong knownAs „China‘s most beautiful“
Knowledge for Intelligent Applications

Enabling technology for:

• **disambiguation**
  in written & spoken natural language

• **deep reasoning**
  (e.g. QA to win quiz game)

• **machine reading**
  (e.g. to summarize book or corpus)

• **semantic search**
  in terms of entities&relations (not keywords&pages)

• **entity-level linkage**
  for Big Data & Big Text analytics
Big Text Analytics: Who Covered Whom?

in different language, country, key, ... with more sales, awards, media buzz, ...

1000‘s of Databases
100 Mio‘s of Web Tables
100 Bio‘s of Web & Social Media Pages

<table>
<thead>
<tr>
<th>Musician</th>
<th>Original</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis Presley</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Robbie Williams</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Sex Pistols</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Frank Sinatra</td>
<td>Claude Francois</td>
<td>Comme d‘Habitude</td>
</tr>
</tbody>
</table>

.....
Big Text Analytics: Who Covered Whom?

in different language, country, key, … with more sales, awards, media buzz, …

1000‘s of Databases
100 Mio‘s of Web Tables
100 Bio‘s of Web & Social Media Pages

Musician | Original | Title
--- | --- | ---
Claudia Leitte Only Won | Bruno Mars | Billionaire
Yip Sin-Man | Bruno Mars | I wanna be an engineer
Norah Jones | Celine Dion | 我心永恒
Beyonce | Bob Dylan | Forever Young
Jay-Z | Alphaville | Forever Young
Big Text Analytics: Who Covered Whom?

in different language, country, key, …
with more sales, awards, media buzz, …
…..

Musician | PerformedTitle
----------|----------------
Sex Pistols | My Way
Frank Sinatra | My Way
Only Won | I wanna be an engineer
Beyonce | Forever Young

Musician | CreatedTitle
----------|----------------
Francis Sinatra | My Way
Paul Anka | My Way
Bruno Mars | Billionaire
Bob Dylan | Forever Young

Name | Event
-----|--------
Beyonce | Grammy
Norah J. | Jobs Mem.
Claudia L. | FIFA 2014

Name | Group
-----|--------
Sid Vicious | Sex Pistols
Bono | U2

1000’s of Databases
100 Mio’s of Web Tables
100 Bio’s of Web & Social Media Pages
Big Text Analytics: Who Covered Whom?

in different language, country, key, …
with more sales, awards, media buzz, …

1000’s of Databases
100 Mio’s of Web Tables
100 Bio’s of Web & Social Media Pages

Musician

Sex Pistols
Frank Sinatra
Claudia Leitte

Beyonce

Only Won

I wanna be an engineer

Travie McCoy
Bob Dylan

Billionaire

Billionaire

Forever Young

I Wanna Be An Engineer - (Billionaire Geeked Out Mix) by: Only Won - the Lyrical Engineer

Claudia Leitte feat. Travie McCoy e Bruno Mars - Famosa (Billionaire)

Jennifer Lopez Pitbull & Claudia Perform at FIFA Opening

Billionaire space pioneer announces plan to colonise Mars with 80,000 people in two decades

- Elon Musk will take passengers to Mars for the princely sum of $600,000

Volume

Velocity

Variety

Veracity
### Big Data & Big Text Analytics

| **Entertainment:** | Who covered which other singer?  
|                   | Who influenced which other musicians? |
| **Health:**       | Drugs (combinations) and their side effects |
| **Politics:**     | Politicians’ positions on controversial topics and their involvement with industry |
| **Business:**     | Customer opinions on small-company products, gathered from social media |
| **Culturomics:**  | Trends in society, cultural factors, etc. |

**General Design Pattern:**

- Identify relevant **contents sources**
- Identify **entities** of interest & their **relationships**
- Position in **time & space**
- Group and **aggregate**
- Find insightful **patterns** & predict **trends**
Outline

✓ Introduction

🌟 Lovely NERD

🌟 The New Chocolate

🌟 The Dark Side

🌟 Conclusion
Lovely NERD
Named Entity Recognition & Disambiguation (NERD)

contextual similarity: mention vs. entity (bag-of-words, language model)

prior popularity of name-entity pairs

Zhang played in Lee's epic masterpiece, with Ma's score.

Ma also covered the Ecstasy.
Zhang played in Lee's epic masterpiece, with Ma's score. Ma also covered the Ecstasy.
Zhang played in Lee's epic masterpiece, with Ma's score. Ma also covered the Ecstasy.
Zhang played in Lee's epic masterpiece, with Ma's score. Ma also covered the Ecstasy.

NED algorithms compute mention-to-entity mapping over weighted graph of candidates by popularity & similarity & coherence.

KB provides building blocks:
- name-entity dictionary,
- relationships, types,
- text descriptions, keyphrases,
- statistics for weights
Joint Mapping

- Build mention-entity graph or probabilistic factor graph from knowledge and statistics in KB
- Compute high-likelihood mapping (ML or MAP) or dense subgraph (with high total edge weight) such that: each m is connected to exactly one e (or at most one e)
Coherence Graph Algorithm

- Compute dense subgraph to maximize min weighted degree among entity nodes such that:
  - each m is connected to exactly one e (or at most one e)
- Approx. algorithms (greedy, randomized, …), hash sketches, …
- 82% precision on CoNLL‘03 benchmark
- Open-source software & online service AIDA

[J. Hoffart et al.: EMNLP‘11, CIKM‘12, WWW’14]

http://www.mpi-inf.mpg.de/yago-naga/aida/
NERD Online Tools

J. Hoffart et al.: EMNLP 2011, VLDB 2011
https://d5gate.ag5.mpi-sb.mpg.de/webaida/

P. Ferragina, U. Scaella: CIKM 2010
http://tagme.di.unipi.it/

P.N. Mendes, C. Bizer et al.: I-Semantics 2011, WWW 2012
http://spotlight.dbpedia.org/demo/index.html

D. Milne, I. Witten: CIKM 2008
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

L. Ratinov, D. Roth, D. Downey, M. Anderson: ACL 2011
http://cogcomp.cs.illinois.edu/page/demo_view/Wikifier

D. Odijk, E. Meij, M. de Rijke: OAIR 2013
http://semanticize.uva.nl

Reuters Open Calais:  http://viewer.opencalais.com/
Alchemy API:  http://www.alchemyapi.com/api/demo.html
Zhang played the Geisha and in Lee's masterpiece, with Ma's score. Ma also covered the Ecstasy.
Research Challenges & Opportunities

- Efficient interactive & high-throughput batch NERD
  a day’s news, a month’s publications, a decade’s archive

- Entity name disambiguation in difficult situations
  Short and noisy texts about long-tail entities in social media

- Handling long-tail and emerging entities
  to complement and continuously update KB
  key for KB life-cycle management

- Web-scale entity linkage with high quality
  across text sources, linked data, KB‘s, Web tables, …
Outline

✓ Introduction

✓ Lovely NERD

★ The New Chocolate

★ The Dark Side

★ Conclusion
Big Text: the New Chocolate

CHOCOLATE
Food of the Gods

When Money Grew on Trees
Semantic Search over News

stics.mpi-inf.mpg.de

Searching with Strings, Things, and Cats

see also J. Hoffart et al.: demo @ CIKM'14
Semantic Search over News

stics.mpi-inf.mpg.de
Semantic Search over News

China's Alipay offers mobile wallet for US sites

... unit in 2011 into a company controlled by Alibaba CEO Jack Ma.
... Because it was spun off, Alipay was not part of Alibaba's
mammoth $25 billion ... Chinese shoppers pay in yuan using their
Alipay account and Alipay transmits the payment to merchants in 14
currencies via 180 ...

Forbes
Alibaba's Potential Arbitrage With Buying Yahoo Now

... the headache of some US politician raising a stink that Jack Ma
is going to start reading the email of his ... So, for all this work, Jack
Ma gets to reduce his debt from $10 billion ... Lynch analyst Justin
Post said for 1% move in Alibaba upwards (or downwards), there
should be a ... But if you believe BAML that Alibaba is going to trade
Beijing Repeats As Top City

... List members (living there, famous denizens) Baidu CEO Robin Li 2. ... 38, Fosun International Chairman Guo Guangchang Alibaba Group Chairman Jack Ma 5. ... Chairman Hui Wing Mau Netease CEO William Ding 7. ...

Entities in this article

Beijing, Tencent, Shanghai, Jack Ma, Hong, Chongqing, Ding Lei, Baidu, Hangzhou, Guangzhou

Brazil, Indonesia, Mexico, Turkey, India Powering Next Wave Of App Growth, Says App Annie

... in China will have access to high-speed mobile networks from China Mobile, China Telecom and China Unicom. ... access to high-speed mobile networks from China Mobile, China Telecom and China Unicom. It will be interesting to see the impacts of ... will have access to high-speed mobile networks from China Mobile, China Telecom and China Unicom. ...

China's Internet Giants Are Going Global In Search Of Competitive Gains
Are patients with primary hypothyroid observational study.

AIM: To assess the percentage of primary hypothyroid patients with abnormal thyroid function despite being prescribed levothyroxine for at least 2 m.

MATERIALS AND METHODS: A cross-sectional, single visit observational study in adult patients with primary hypothyroidism on treatment with levothyroxine for at least 2 m was undertaken across 10 cities in India. Compliance to Thyroxine therapy was assessed by interviewing the subjects and their quality of life was assessed by administering the SF-36 questionnaire. TSH levels were correlated with the current dose of levothyroxine. A random blood sample (5ml) was drawn from the study subjects during the same visit for assessing serum TSH levels. A total of 1950 subjects (mean age 41.4 ± 11.17 years, female 81.2 %, male 18.8 %) with primary hypothyroidism were enrolled in the study.

RESULTS:

The mean dose of thyroxine in this study was 1.23 g/k/day (±0.35). Of the 1925 subjects in whom TSH values were available, 605 (41.87 %) were under-treated (TSH > 4 mIU/L) and 243 (12.62 %) were over-treated (TSH < 0.4 mIU/L). Age and autoimmune hypothyroidism were the factors that had significant impact on serum TSH. Majority of subjects (90.79 %) were compliant/moderately compliant to thyroxine therapy.

CONCLUSION:

Subjects with abnormal TSH had significantly lower scores for role limitation due to emotional problems (P = 0.0273) and due to physical health (P = 0.0763). The mean daily dose of thyroxine (1.23 g/kg ± 0.85) was less than the recommended full replacement dose. This study concluded that around half (54 %) of known hypothyroid subjects had out-of-range serum TSH despite being treated with levothyroxine for at least 2 m.
Side effects of levothyroxine + supplements?

Hi, I have very stupidly been taking iodine supplements as well as levothyroxine, which I’ve just found out from the internet that you’re absolutely not supposed to do. I had been having weird side effects, like palpitations and a burning sensation in my gullet, which seemed odd, given that I’m on a very low dose of levothyroxine (25 mcg) and have only been on it for a few weeks, but hadn’t even thought that the over-the-counter supplements might be the source of the problem. I did ask my GP before taking them and she said she doubted they’d be of much benefit, but didn’t think they’d do any harm, either! I’ve been taking a 300mg keip tablet once or twice a week and a multivitamin including iodine on days when I don’t take the keip. According to the packaging, this is within the recommended daily dose. Although I’ve recently started eating dairy again, I was a vegan for 8 and a half years and I suspect that my hypothyroidism might have been caused by iodine deficiency, which is why I wanted to try supplements. However, I’ve scared myself a bit by reading on the internet that a burning sensation in the gullet is a sign of iodine poisoning (although I’m guessing I’d have other symptoms, too, if I’d consumed a life-threatening amount). Obviously I’m going to stop taking the supplements immediately, but I’m worried I might have harmed myself. I’m reluctant to ask my GP, as I’ve bothered them a lot lately and, anyway, while they know a lot about prescription drugs, I get the sense that they aren’t that clued up on dietary issues/food supplements, so probably couldn’t give me much reassurance. Has anyone else experienced similar side effects? And has anyone else combined levothyroxine with supplements?
Big Data & Text Analytics:
Side Effects of Drug Combinations

Deeper insight from both expert data & social media:
- actual side effects of drugs
- ... and drug combinations
- risk factors and complications of (wide-spread) diseases
- alternative therapies
- aggregation & comparison by age, gender, life style, etc.

Structured Expert Data


http://www.patient.co.uk
The Maestro from Rome wrote scores for westerns.

Ma played his version of the Ecstasy.

Maestro Card
Leonard Bernstein
Ennio Morricone
Rome (Italy)
AS Roma
Lazio Roma
Jack Ma
Yo-Yo Ma
Massachusetts
Western Digital

born in

plays for

goal in football

plays sport

plays music

cover of

story about

western movie

film music

music

massachusetts
Paraphrases of Relations

composed: musician × song
covered: musician × song

Dylan wrote a sad song Knockin‘ on Heaven‘s Door, a cover song by the Dead
Morricone‘s masterpiece is the Ecstasy of Gold, covered by Yo-Yo Ma
Amy‘s souly interpretation of Cupid, a classic piece of Sam Cooke
Nina Simone‘s singing of Don‘t Explain revived Holiday‘s old song
Cat Power‘s voice is haunting in her version of Don‘t Explain
Cale performed Hallelujah written by L. Cohen

SOL patterns over words, wildcards, POS tags, semantic types:
<musician> wrote * ADJ piece <song>

Relational phrases are typed:
<singer> covered <song>
<book> covered <event>

Relational synsets (and subsumptions):
covered: cover song, interpretation of, singing of, voice in * version, …
composed: wrote, classic piece of, ‘s old song, written by, composed, …

350 000 SOL patterns from Wikipedia: http://www.mpi-inf.mpg.de/yago-naga/patty/
Disambiguation for Entities, Classes & Relations
(M. Yahya et al.: EMNLP’12, CIKM’13)

Maestro
- e: MaestroCard
- e: Ennio Morricone
- c: conductor
- c: musician
- r: actedIn
- r: bornIn

from
- e: Rome (Italy)
- e: Lazio Roma
- r: composed
- r: giveExam
- c: soundtrack
- r: soundtrackFor
- r: shootsGoalFor
- c: western movie
- e: Western Digital

wrote scores

scores for

westerns

ILP optimizers like Gurobi solve this in seconds

Combinatorial Optimization by ILP (with type constraints etc.)

weighted edges (coherence, similarity, etc.)
Research Challenges & Opportunities

Context-aware “strings, things, cats“:
suggested entities & categories should consider input prefix

Comprehensive repository

- binary-predicate counterparts
- despite Patty, Wiseman

Managing text & data

- text with SPO/ER markup linked to DB/KB records
- DB/KB records linked to provenance, context, fact spottings

Query & analytics language – API?

Sparql Full-Text, extended / relaxed Sparql, data&text cube, …

User-friendly search & exploration – UI?

natural language, visual, multimodal, …
Research Challenges & Opportunities

Context-aware “strings, things, cats“:
suggested entities & categories should consider input prefix

Comprehensive repository of relational paraphrases:
binary-predicate counterpart of WordNet - still elusive
despite Patty, WiseNet, Probase, ReVerb, ReNoun, Biperpedia, …

Managing text & data on par: DB&IR integration, finally
- text with SPO/ER markup linked to DB/KB records
- DB/KB records linked to provenance, context, fact spottings

Query & analytics language – API ?
Sparql Full-Text, extended / relaxed Sparql, data&text cube, …

User-friendly search & exploration – UI ?
natural language speech, visual, multimodal, …
Outline

- Introduction
- Lovely NERD
- The New Chocolate
- The Dark Side
- Conclusion
The Dark Side of Big Data

Nobody interested in your research? We read your papers!
Entity Linking: Privacy at Stake

Zoe

publish & recommend

discuss & seek help

search

female 29y Jamame

female 25-30 Somalia

Cry Freedom
Nive Nielsen

social network

Synthroid tremble
Addison disorder

online forum

Levothroid shaking
Addison’s disease
Greenland singers
Somalia elections
Steve Biko

search engine

Internet
Privacy Adversaries

Linkability Threats:
- Weak cues: profiles, friends, etc.
- Semantic cues: health, taste, queries
- Statistical cues: correlations

Social network: Facebook

Online forum: Patient.co.uk

Internet

Cry Freedom

Nive Nielsen

Female 25-30 Somalia

Female 29y Jamame

Synthroid tremble
Addison disorder

Levothroid shaking
Addison’s disease

Nive concert
Greenland singers
Somalia elections
Steve Biko

Search engine
Goal: Automated Privacy Advisor

Privacy Adviser (PA):
Software tool that
- analyses risk
- alerts user
- advises user
  - explains consequences
  - recommends policy changes

Your queries may lead to linking your identities in Facebook and patient.co.uk!

Would you like to use an anonymization tool for your search requests?

---

ERC Project **imPACT**
(M. Backes, P. Druschel, R. Majumdar, G. Weikum)
see also: J. Biega et al.: PSBD Workshop @ CIKM‘14
## Research Challenges & Opportunities

<table>
<thead>
<tr>
<th>Question</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which data is (or can be linked to become) privacy-critical?</td>
<td>highly user-specific, but needs a global perspective</td>
</tr>
<tr>
<td>How are privacy risks building up over time?</td>
<td>where is my data, who has seen it, who can copy and accumulate it</td>
</tr>
<tr>
<td>Who are the adversaries? How powerful? At which cost?</td>
<td>role of background knowledge &amp; statistical learning</td>
</tr>
<tr>
<td>Explain risks, advise on consequences, recommend counter-measures and</td>
<td>mitigation steps</td>
</tr>
<tr>
<td>Long-term privacy management:</td>
<td>policies, risks, privacy-utility trade-offs</td>
</tr>
</tbody>
</table>
Outline

✓ Introduction
✓ Lovely NERD
✓ The New Chocolate
✓ The Dark Side

🌟 Conclusion
Big Text & Big Data

**Big Text & NERD:**
valuable content about entities
lifted towards knowledge & analytic insight

**Machine Reading:**
discover and interpret names & phrases as
entities, classes, relations,
spatio-temporal modifiers, sentiments, beliefs, ....

**Big Data:**
interlink natural-language text, social media,
structured data & knowledge bases, images, videos
and help users coping with privacy risks
Take-Home Message

more knowledge, analytics, insight

Web Contents

Knowledge

knowledge acquisition

intelligent interpretation

„Who Covered Whom?“ and More!
(Entities, Classes, Relations)