

Information Extraction

(~20 slides from E. Agichtein)

Steffen Staab
Maciej Janik

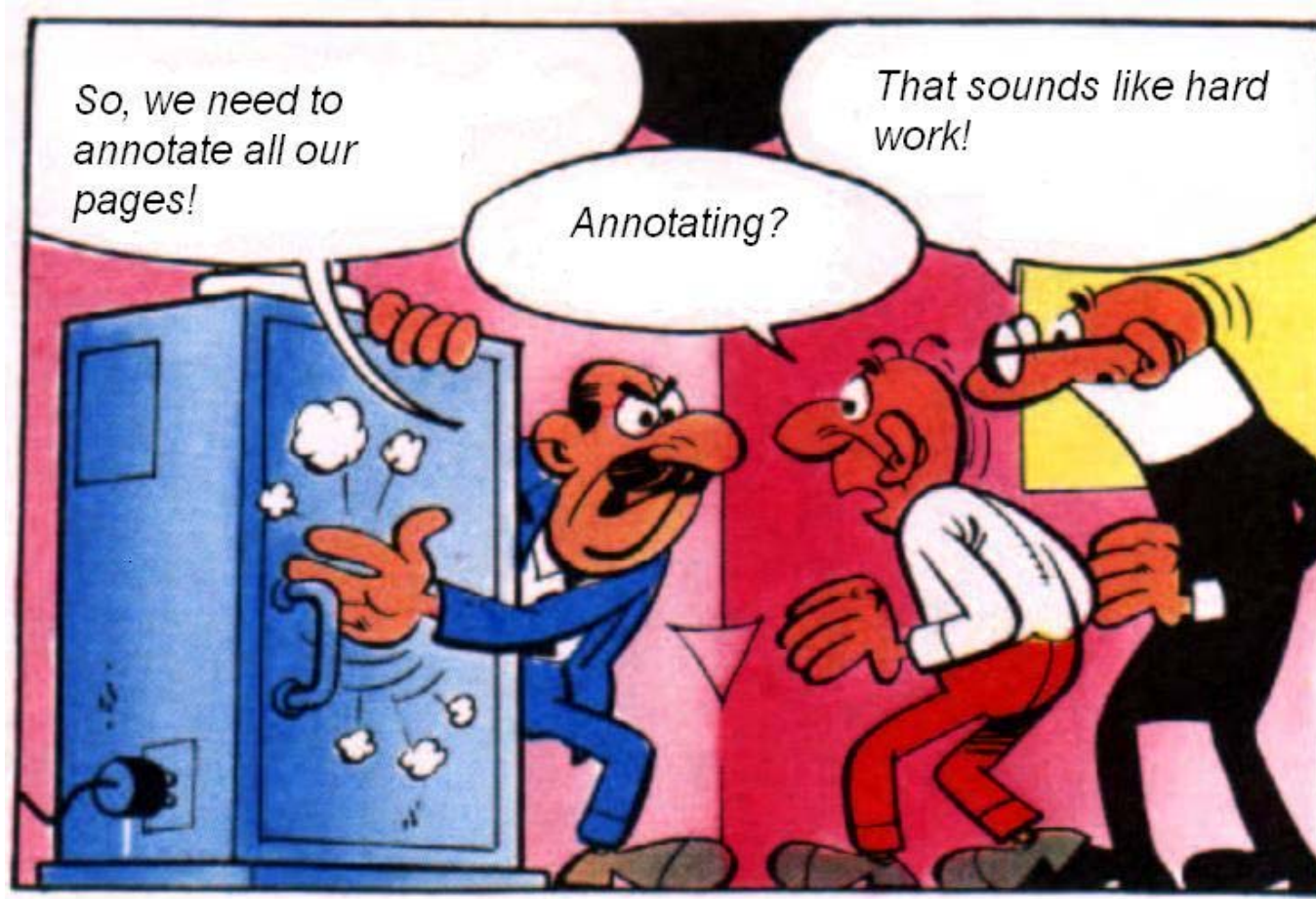
Semantic Web
2009-07-10

- “Unstructured” text data is the **primary** form of human-generated information
 - ◆ Blogs, web pages, news, scientific literature, online reviews, ...
 - ◆ Semi-structured data (database generated): see Prof. Bing Liu’s KDD webinar: <http://www.cs.uic.edu/~liub/WCM-Refs.html>
 - ◆ The techniques discussed here are complimentary to structured object extraction methods
- Need to extract **structured** information to effectively manage, search, and mine the data



The annotation problem from a scientific point of view <is web>







- Information Extraction: mature, but **active research area**
 - ◆ Intersection of Computational Linguistics, Machine Learning, Data mining, Databases, and Information Retrieval
 - ◆ Traditional focus on **accuracy** of extraction

For years, [Microsoft Corporation](#) [CEO Bill Gates](#) was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

Select Name
From PEOPLE
Where Organization = 'Microsoft'

PEOPLE

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

Bill Gates
Bill Veghte

(from William Cohen's IE tutorial, 2003)

- **Information Extraction Tasks**

- ◆ Entity tagging
- ◆ Relation extraction
- ◆ Event extraction

- **Scaling up Information Extraction**

- ◆ Focus on scaling up to large collections (where data mining can be most beneficial)
- ◆ Other dimensions of scalability

- Extracting entities and relations: this talk
 - ♦ **Entities**: named (e.g., Person) and generic (e.g., disease name)
 - ♦ **Relations**: entities related in a predefined way (e.g., Location of a Disease outbreak, or a CEO of a Company)
 - ♦ **Events**: can be composed from multiple relation tuples

- Common extraction subtasks:
 - ♦ **Preprocess**: sentence chunking, syntactic parsing, morphological analysis
 - ♦ Create **rules** or **extraction patterns**: hand-coded, machine learning, and hybrid
 - ♦ **Apply** extraction patterns or rules to **extract** new information
 - ♦ **Postprocess** and **integrate** information
 - Co-reference resolution, deduplication, disambiguation

- Identifying mentions of entities (e.g., person names, locations, companies) in text
 - ◆ MUC (1997): Person, Location, Organization, Date/Time/Currency
 - ◆ ACE (2005): more than 100 more specific types
- Hand-coded vs. Machine Learning approaches
- Best approach depends on entity type and domain:
 - ◆ **Closed class** (e.g., geographical locations, disease names, gene & protein names): hand coded + dictionaries
 - ◆ **Syntactic** (e.g., phone numbers, zip codes): regular expressions
 - ◆ **Semantic** (e.g., person and company names): mixture of context, syntactic features, dictionaries, heuristics, etc.
 - ◆ “Almost solved” for common/typical entity types

- Useful for data warehousing, data cleaning, web data integration

Address	House number	Building	Road	City	State	Zip
	4089	Whispering Pines	Nobel Drive	San Diego	CA	92122

Citation Ronald Fagin, *Combining Fuzzy Information from Multiple Systems*, Proc. of ACM SIGMOD, 2002

Segment(s _i)	Sequence	Label(s _i)
S ₁	Ronald Fagin	Author
S ₂	Combining Fuzzy Information from Multiple Systems	Title
S ₃	Proc. of ACM SIGMOD	Conference
S ₄	2002	Year

- Easy to construct in some cases
 - ◆ e.g., to recognize prices, phone numbers, zip codes, conference names, etc.
- Intuitive to debug and maintain
 - ◆ Especially if written in a “high-level” language:

```
ContactPattern ← RegularExpression(Email.body, "can be reached at")  
[IBM Avatar]
```

- ◆ Can incorporate domain knowledge
- Scalability issues:
 - ◆ Labor-intensive to create
 - ◆ Highly domain-specific
 - ◆ Often corpus-specific
 - ◆ Rule-matches can be expensive

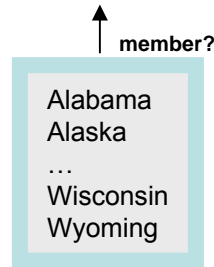
- Can work well when lots of training data easy to construct
- Can capture complex patterns that are hard to encode with hand-crafted rules
 - ◆ e.g., determine whether a review is positive or negative
 - ◆ extract long complex gene names
 - ◆ Non-local dependencies

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

[From AliBaba]

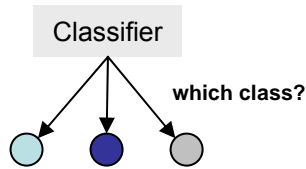
Lexicons

Abraham Lincoln was born in Kentucky.



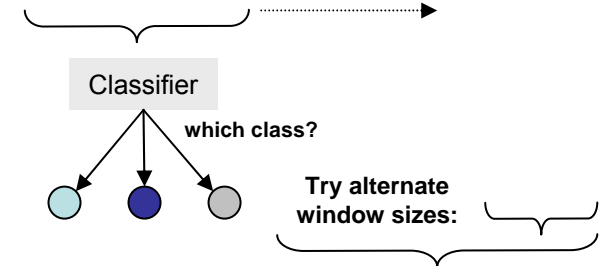
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



Sliding Window

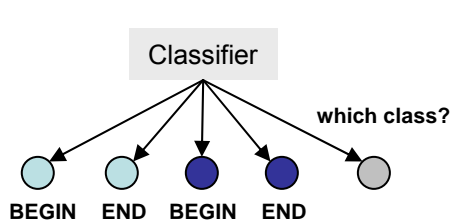
Abraham Lincoln was born in Kentucky.



Boundary Models

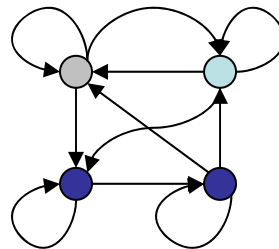
Abraham Lincoln was born in Kentucky.

BEGIN



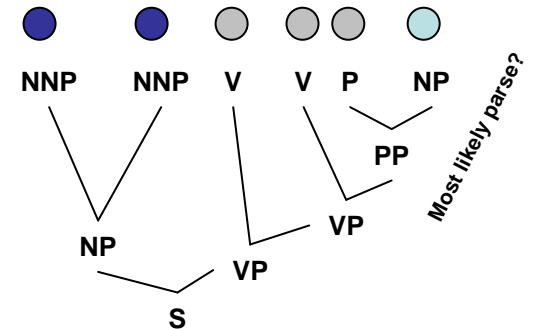
Finite State Machines

Abraham Lincoln was born in Kentucky.



Context Free Grammars

Abraham Lincoln was born in Kentucky.



...and beyond

Any of these models can be used to capture words, formatting or both.

For details: [Feldman, 2006 and Cohen, 2004]

- Naive Bayes
- SRV [Freitag 1998], Inductive Logic Programming
- Rapier [Califf and Mooney 1997]
- Hidden Markov Models [Leek 1997]
- Maximum Entropy Markov Models [McCallum et al. 2000]
- Conditional Random Fields [Lafferty et al. 2001]

- Scalability
 - ◆ Can be labor intensive to construct training data
 - ◆ At run time, complex features can be expensive to construct or process (batch algorithms can help: [Chandel et al. 2006])

- **ABNER:**
 - ♦ <http://www.cs.wisc.edu/~bsettles/abner/>
 - ♦ Linear-chain conditional random fields (CRFs) with orthographic and contextual features.

- **Alias-I LingPipe**
 - ♦ <http://www.alias-i.com/lingpipe/>

- **MALLET:**
 - ♦ http://mallet.cs.umass.edu/index.php/Main_Page
 - ♦ Collection of NLP and ML tools, can be trained for name entity tagging

- **MinorThird:**
 - ♦ <http://minorthird.sourceforge.net/>
 - ♦ Tools for learning to extract entities, categorization, and some visualization

- **Stanford Named Entity Recognizer:**
 - ♦ <http://nlp.stanford.edu/software/CRF-NER.shtml>
 - ♦ CRF-based entity tagger with non-local features

- Statistical named entity tagger
 - ◆ Generative statistical model
 - Find most likely tags given lexical and linguistic features
 - Accuracy at (or near) state of the art on benchmark tasks

- Explicitly targets scalability:
 - ◆ ~100K tokens/second runtime on single PC
 - ◆ Pipelined extraction of entities
 - ◆ User-defined mentions, pronouns and stop list
 - Specified in a dictionary, left-to-right, longest match
 - ◆ Can be trained/bootstrapped on annotated corpora

- Overview of Information Extraction
 - ◆ Entity tagging
 - ◆ **Relation extraction**
 - ◆ Event extraction

- Scaling up Information Extraction
 - ◆ Focus on scaling up to large collections (where data mining and ML techniques shine)
 - ◆ Other dimensions of scalability

- Extract tuples of entities that are **related** in predefined way

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly **Ebola** epidemic in **Zaire**, is finding itself hard pressed to cope with the crisis...

Relation Extraction

Disease Outbreaks relation

Date	Disease Name	Location
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Knowledge engineering

- Experts develop rules, patterns:
 - ♦ Can be defined over lexical items: “<company> located in <location>”
 - ♦ Over syntactic structures: “((Obj <company>) (Verb located) (*) (Subj <location>))”
- Sophisticated development/debugging environments:
 - ♦ Proteus, **GATE**

Machine learning

- **Supervised**: Train system over **manually labeled** data
 - ♦ Soderland et al. 1997, Muslea et al. 2000, Riloff et al. 1996, Roth et al 2005, Cardie et al 2006, Mooney et al. 2005, ...
- **Partially-supervised**: train system by **bootstrapping** from “seed” examples:
 - ♦ Agichtein & Gravano 2000, Etzioni et al., 2004, Yangarber & Grishman 2001, ...
 - ♦ “Open” (**no seeds**): Sekine et al. 2006, Cafarella et al. 2007, Banko et al. 2007
- **Hybrid** or **interactive** systems:
 - ♦ Experts interact with machine learning algorithms (e.g., active learning family) to iteratively refine/extend rules and patterns
 - ♦ Interactions can involve annotating examples, modifying rules, or any combination

Table 2: The contrast between traditional and open IE.

	Traditional IE	Open IE
Input	Corpus + Labeled Data	Corpus + Domain-Independent Methods
Relations	Specified In Advance	Discovered Automatically
Complexity	$O(D \cdot R)$ <i>D</i> documents, <i>R</i> relations	$O(D)$ <i>D</i> documents

- learn a general model of *how* relations are expressed (in a particular language), based on unlexicalized features such as part-of-speech tags. (Identify a verb)
- Learn domain-independent regular expressions. (Punctuations, Commas).

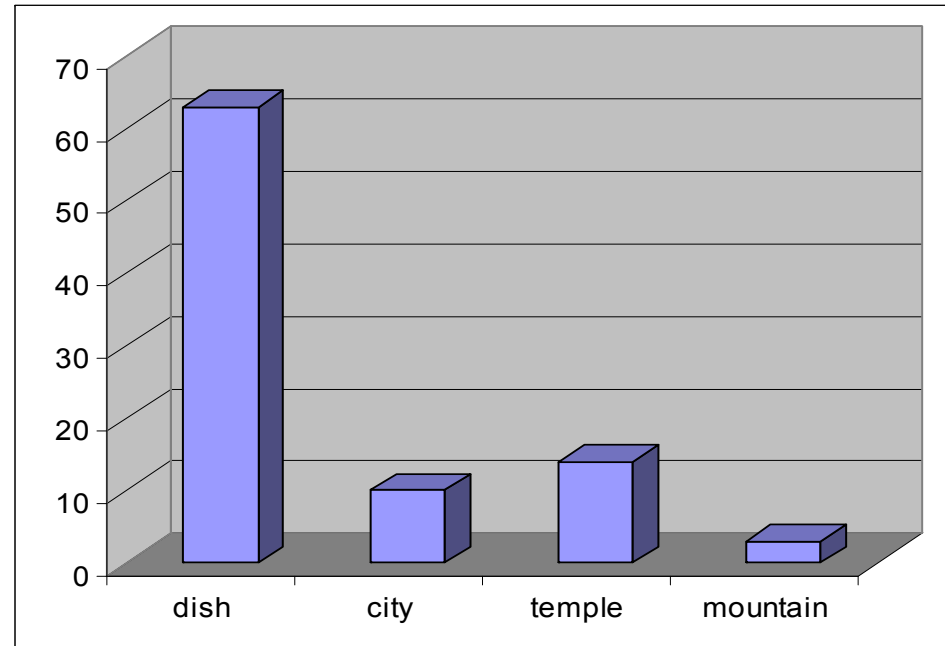
Table 1: Taxonomy of binary relationships. Nearly 95% of 500 randomly selected sentences belong to one of the eight categories noted here.

Relative Frequency	Category	Simplified Lexico-Syntactic Pattern
37.8	Verb	E_1 Verb E_2 <i>X established Y</i>
22.8	Noun + Prep	E_1 NP Prep E_2 <i>X settlement with Y</i>
16.0	Verb + Prep	E_1 Verb Prep E_2 <i>X moved to Y</i>
9.4	Infinitive	E_1 to Verb E_2 <i>X plans to acquire Y</i>
5.2	Modifier	E_1 Verb E_2 Noun <i>X is Y winner</i>
1.8	Coordinate _n	E_1 (and , - :) E_2 NP <i>X-Y deal</i>
1.0	Coordinate _v	E_1 (and ,) E_2 Verb <i>X, Y merge</i>
0.8	Appositive	E_1 NP (: ,)? E_2 <i>X hometown : Y</i>

- There is a huge amount of implicit knowledge in the Web
- Make use of this implicit knowledge together with statistical information to propose formal annotations and overcome the vicious cycle:

semantics \approx syntax + statistics?

- Annotation by maximal statistical evidence



What is Laksa?

A: dish B: city

C: temple D: mountain

- „cities such as Laksa“ 0 hits
- „dishes such as Laksa“ 10 hits
- „mountains such as Laksa“ 0 hits
- „temples such as Laksa“ 0 hits

⇒ Google knows more than all of you together!

⇒ Example of using syntactic information + statistics to derive semantic information

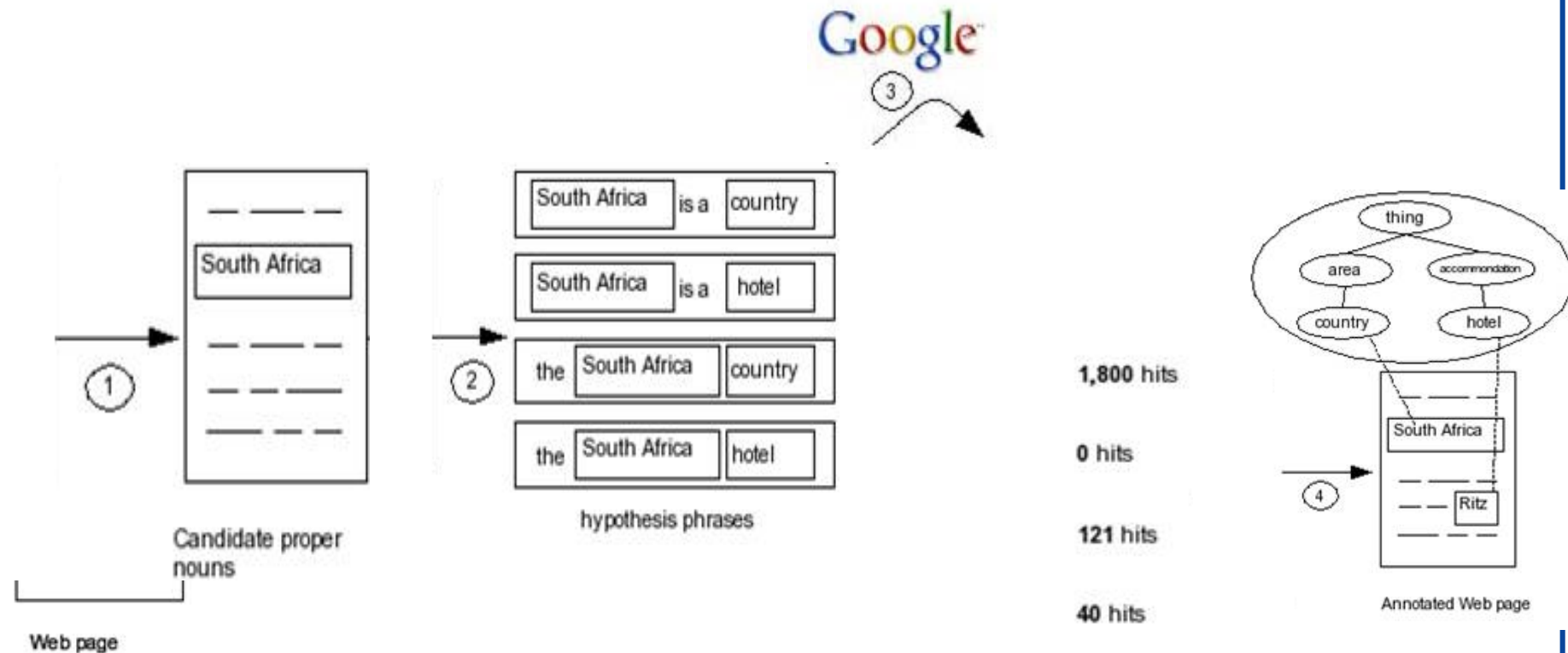
- HEARST1: <CONCEPT>s such as <INSTANCE>
- HEARST2: such <CONCEPT>s as <INSTANCE>
- HEARST3: <CONCEPT>s, (especially/including) <INSTANCE>
- HEARST4: <INSTANCE> (and/or) other <CONCEPT>s

- Examples:
 - ◆ dishes such as Laksa
 - ◆ such dishes as Laksa
 - ◆ dishes, especially Laksa
 - ◆ dishes, including Laksa
 - ◆ Laksa and other dishes
 - ◆ Laksa or other dishes

- DEFINITE1: the <INSTANCE> <CONCEPT>
- DEFINITE2: the <CONCEPT> <INSTANCE>

- APPOSITION:<INSTANCE>, a <CONCEPT>
- COPULA: <INSTANCE> is a <CONCEPT>

- Examples:
 - the Laksa dish
 - the dish Laksa
 - Laksa, a dish
 - Laksa is a dish



- Instance $i \in I$, concept $c \in C$, pattern $p \in \{\text{Hearst1}, \dots, \text{Copula}\}$ **$count(i, c, p)$** returns the number of Google hits of instantiated pattern

$$count(i, c) := \sum_p count(i, c, p)$$

- E.g. $count(\text{Laksa}, \text{dish}) := count(\text{Laksa}, \text{dish}, \text{def1}) + \dots$
- Restrict to the best ones beyond threshold θ

$$R_\theta := \left\{ (i, c_i) \mid i \in I, c_i := \arg \max_{c \in C} (count(i, c)) \wedge count(i, c) \geq \theta \right\}$$

Atlantic city 1520837

Bahamas island 649166

USA country 582275

Connecticut state 302814

Caribbean sea 227279

Mediterranean sea 212284

Canada country 176783

Guatemala city 174439

Africa region 131063

Australia country 128607

France country 125863

Germany country 124421

Easter island 96585

St Lawrence river 65095

Commonwealth state 49692

New Zealand island 40711

Adriatic sea 39726

Netherlands country 37926

St John church 34021

Belgium country 33847

San Juan island 31994

Mayotte island 31540

EU country 28035

UNESCO organization 27739

Austria group 24266

Greece island 23021

Malawi lake 21081

Israel country 19732

Perth street 17880

Luxembourg city 16393

Nigeria state 15650

St Croix river 14952

Nakuru lake 14840

Kenya country 14382

Benin city 14126

Cape Town city 13768

- Corpus: 45 texts from <http://www.lonelyplanet.com/destinations>
- Ontology: tourism ontology from GETESS project
 - ◆ #concepts: original – 1043; pruned – 682
- Manual Annotation by two subjects:
 - ◆ A: 436 instance/concept assignments
 - ◆ B: 392 instance/concept assignments
 - ◆ Overlap: 277 instances (Gold Standard)
 - ◆ A and B used 59 different concepts
 - ◆ Categorical (Kappa) agreement on 277 instances:
 $\kappa = 63.5\%$

- **Cohens Kappa** for interrater reliability

$$\kappa = \frac{p_{agree} - p_{random}}{1 - p_{random}}$$

- p_{agree} : which fraction of times to raters agree
- p_{random} : which fraction of times would they agree by chance

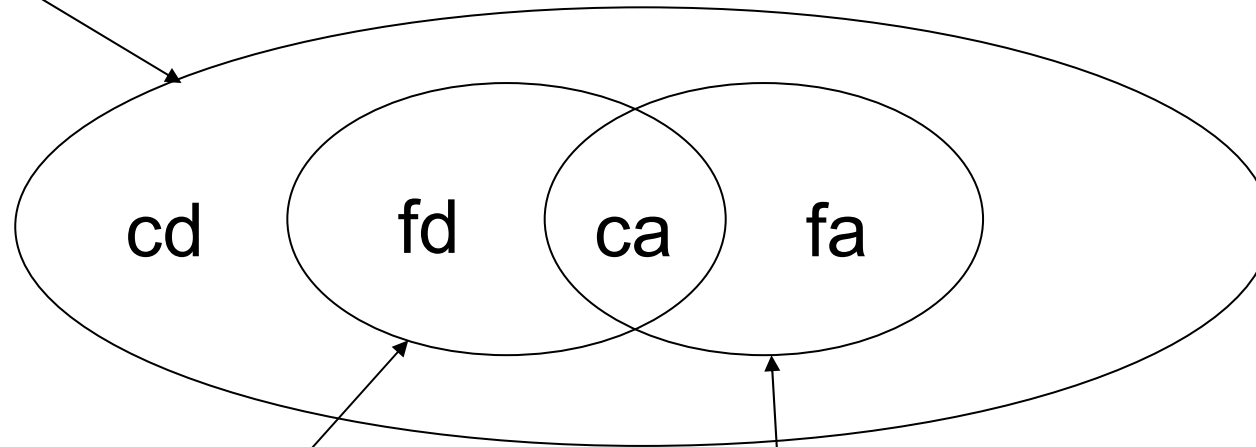
- $\kappa = 0 \Rightarrow$ no agreement
- $\kappa = 1 \Rightarrow$ perfect agreement
- $\kappa < 0 \Rightarrow$ tending towards disagreement

- Correct decisions:
 1. **correct alarms (ca)**
 2. **correct dismissals (cd)**

- Erroneous decisions:
 1. **false alarms (fa)**
 2. **false dismissals (fd)**

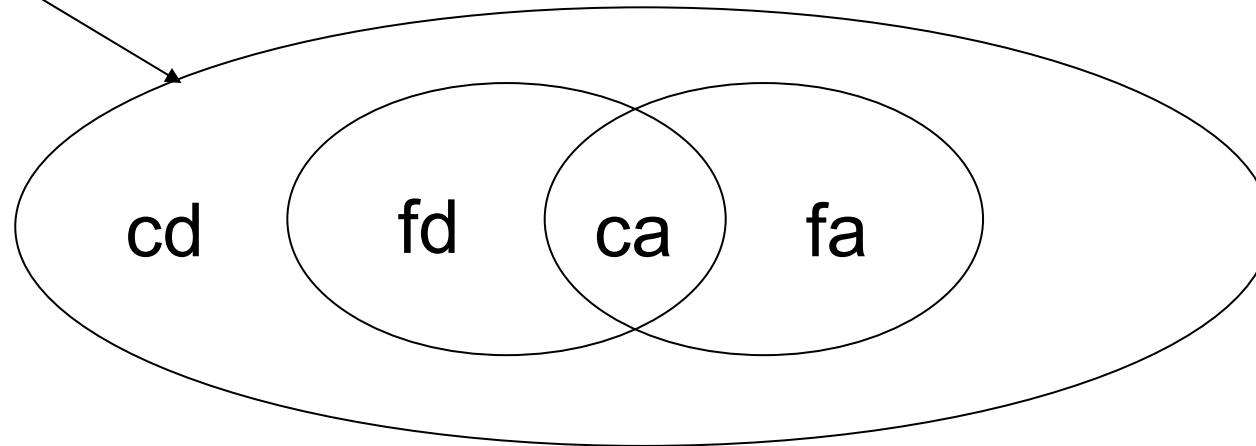
- **fa, fd, ca, cd** represent the numbers of times that documents from a result set fall into any of the four possible categories

Item set



User judgment	System judgment	
	relevant	irrelevant
relevant	ca	fd
irrelevant	fa	cd

Item set



$$Prec = \frac{ca}{ca + fa} \in [0, 1]$$

$$Rec = \frac{ca}{ca + fd} \in [0, 1]$$

	User 1	User 2	System
1	Niger: country	Niger: river	Niger: country
2	Germany: country	Germany: country	Germany: country
3	Main: river	Main: river	Main: street
4	India: country	India: region	India: country
	$\kappa = (\frac{1}{2} - \frac{1}{3}) / 1 - (\frac{1}{3})$ $= \frac{1}{4}$		

Assume Gold Standard: User 1

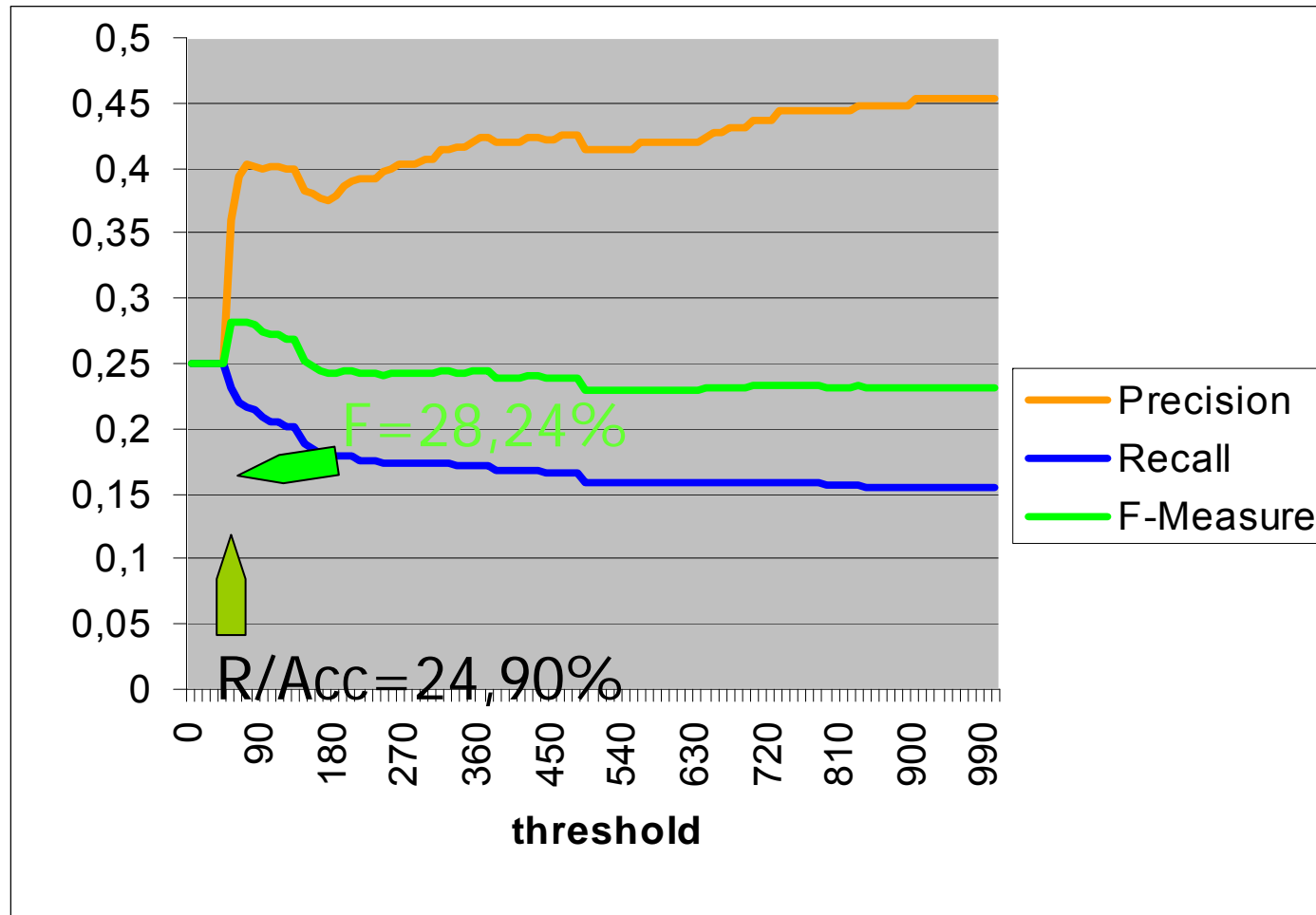
Then: $ca = 3$
 $fd = 1$
 $fa = 1$

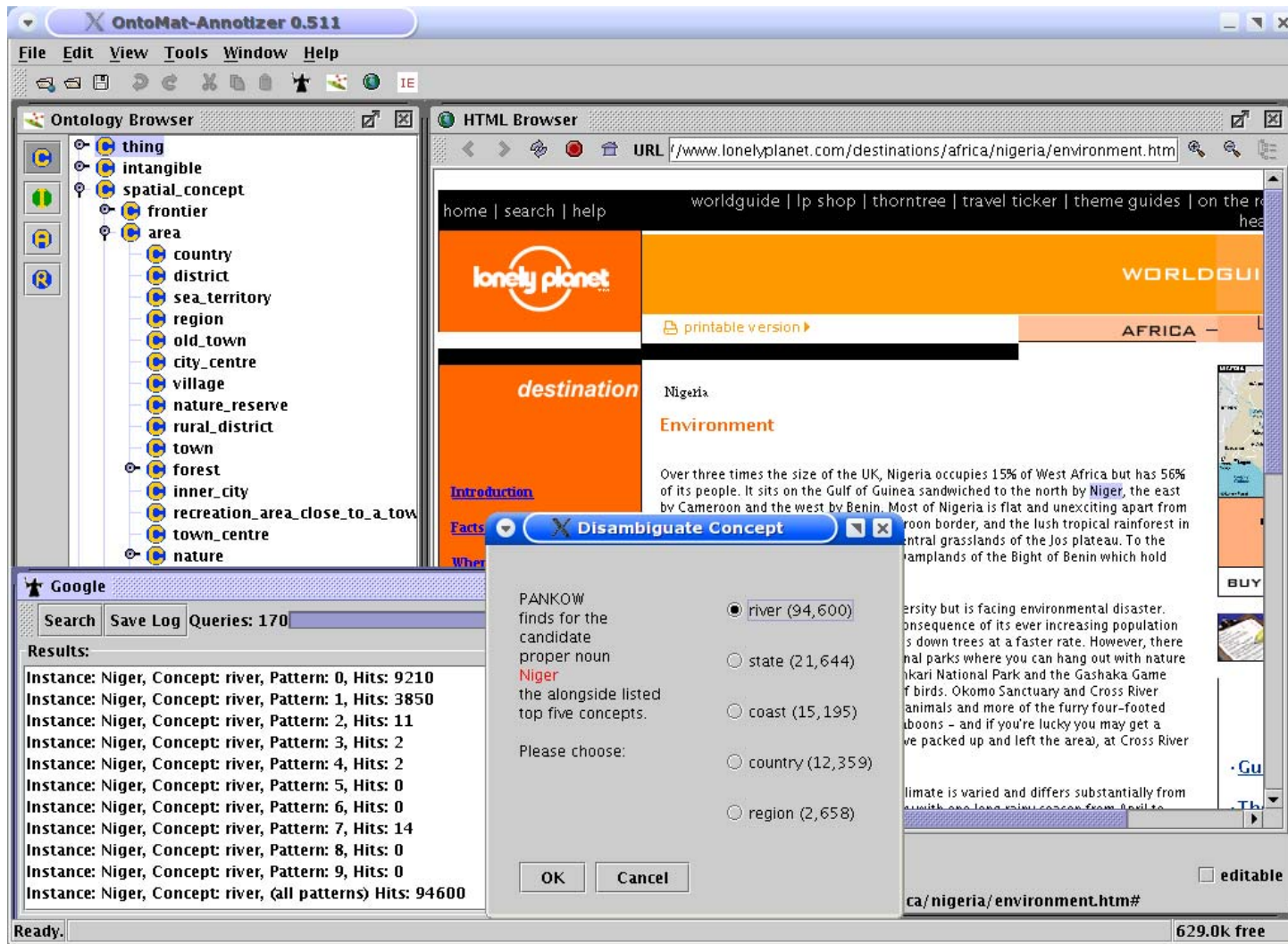
$Prec = 3 / (3+1) = 3/4$
 $Rec = 3 / (3+1) = 3/4$

Assume Gold Standard: User 2

Then: $ca = 1$
 $fd = 3$
 $fa = 3$

Open world: often $Prec@10$ and Recall unknown





Information Extraction – Extracting Knowledge from Wikipedia

Steffen Staab
Maciej Janik

Semantic Web
2009-07-10

- Sören Auer, Jens Lehmann. „What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content.” ESWC 2007
- The amount of knowledge in Wikipedia
 - ◆ English ~ 3M articles
 - ◆ German ~1M articles
 - ◆ Polish ~620K articles
 - ◆ Over 250 languages
 - ◆ Created and edited by users, for other users to read
 - ◆ Arranged into categories, included some structure
 - ◆ ... but generally not machine understandable

- DBpedia is a community effort to
 - ◆ extract structured (“infobox”) information from Wikipedia
 - ◆ provide a query endpoint to the dataset
 - ◆ interlink the DBpedia dataset with other datasets on the Web
- Why it works?
 - ◆ Data: Wikipedia full dumps freely available
 - ◆ Use of open source software



- Structured and “standardized” knowledge in Wikipedia
- Reused for multiple object that share similar semantics

```
1  {{Infobox Town AT |
2  name = Innsbruck |
3  image_coa = InnsbruckWappen.png |
4  image_map = Karte-tirol-I.png |
5  state = [[Tyrol]] |
6  regbzk = [[Statutory city]] |
7  population = 117,342 |
8  population_as_of = 2006 |
9  pop_dens = 1,119 |
10 area = 104.91 |
11 elevation = 574 |
12 lat_deg = 47 |
13 lat_min = 16 |
14 lat_hem = N |
15 lon_deg = 11 |
16 lon_min = 23 |
17 lon_hem = E |
18 postal_code = 6010-6080 |
19 area_code = 0512 |
20 licence = I |
21 mayor = Hilde Zach |
22 website = [http://innsbruck.at] |
23 }}
```

Innsbruck	
	
Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km ²
Population density	1,119 /km ²
Elevation	574 m
Coordinates	47°16′ N 11°23′ E 
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at 

dbpprop:area	▪ 104.91 (xsd:double)
dbpprop:areaCode	▪ 512 (xsd:integer)
dbpprop:district	▪ Statutory city (en)
dbpprop:elevation	▪ 574 (xsd:integer)
dbpprop:hasPhotoCollection	▪ http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Innsbruck
dbpprop:imageCaption	▪ Innsbruck (en)
dbpprop:imageCoa	▪ http://upload.wikimedia.org/wikipedia/en/a/aa/InnsbruckWappen.png
dbpprop:imagePhoto	▪ http://upload.wikimedia.org/wikipedia/commons/0/07/IMG_9039-Innsbruck.JPG
dbpprop:imagesize	▪ 300px (en)
dbpprop:latDeg	▪ 47 (xsd:integer)
dbpprop:latHem	▪ N (en)
dbpprop:latMin	▪ 16 (xsd:integer)
dbpprop:licence	▪ I (en)
dbpprop:lonDeg	▪ 11 (xsd:integer)
dbpprop:lonHem	▪ E (en)
dbpprop:lonMin	▪ 23 (xsd:integer)
dbpprop:mayor	▪ dbpedia:Hilde_Zach
dbpprop:name	▪ Innsbruck (en)
dbpprop:popDens	▪ 1119 (xsd:integer)
dbpprop:population	▪ 117916 (xsd:integer)
dbpprop:populationAsOf	▪ 01.01.2007 (en)
dbpprop:postalCode	▪ 6010-6080 (en)

- Identify pages with templates
- Choose well-populated / used templates
 - ◆ Unused templates can only bring errors
- Parse template (XMLized Wikipedia format)
 - ◆ Get attributes → relations
 - ◆ Get values → objects
- Create relevant triples from extracted information
 - ◆ URI-fy references
 - ◆ Add data types (if known from template context)

- Not all pages include templates
 - ◆ Even within the same topic not all pages may have it

- Template definition flaws
 - ◆ Not well-formed – include presentation properties
 - ◆ Multiple values as objects for more intuitive presentation
 - [[Innsbruck]], [[Austria]]
 - ◆ Complex and redundant attribute values
 - height=5'11" (180cm)
 - ◆ One subject can have multiple templates defined
 - Infobox_Film, Infobox Film, Infobox_film, ...
 - ◆ Multiple attribute names to define the same relationship

- Exploiting rich Wikipedia linking
 - ◆ Unnamed links to other entities in Wikipedia
 - HREFs (for now ...)
 - ◆ Categorization
 - Wikipedia categories
 - Not strict hierarchy, rather thesaurus
 - ◆ Multiple languages
 - Variety of languages for the same entity
 - ◆ Links to open data and other web resources
 - e.g. geo locations
 - ◆ Links to other Wiki projects
 - Wiktionary
 - Wikimedia Commons

Amsterdam



The Keizersgracht at dusk

Location of Amsterdam

Coordinates:  52°22'23"N 4°53'32"E

Country	Netherlands
Province	North Holland
Government	
- Type	Municipality
- Mayor	Job Cohen ^[1] (PvdA)
- Aldermen	Lodewijk Asscher Carolien Gehrels Tjeerd Herrema Maarten van Poelgeest Marijke Vos
- Secretary	Erik Gerritsen
Area ^{[2][3]}	
- City	219 km² (84.6 sq mi)
- Land	166 km² (64.1 sq mi)
- Water	53 km² (20.5 sq mi)
- Urban	1,003 km² (387.3 sq mi)
- Metro	1,815 km² (700.8 sq mi)
Elevation ^[4]	2 m (7 ft)
Population (1 October 2008) ^{[5][6]}	
- City	755,269
- Density	4,459/km² (11,548.8/sq mi)
- Urban	1,364,422
- Metro	2,168,372
- Demonym	Amsterdammer
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postcodes	1011 – 1109
Area code(s)	020

Website: www.amsterdam.nl

```
@prefix dbpedia <http://dbpedia.org/resource/>.
@prefix dbterm <http://dbpedia.org/property/>.

dbpedia:Amsterdam
  dbterm:officialName "Amsterdam" ;
  dbterm:longd "4" ;
  dbterm:longm "53" ;
  dbterm:longs "32" ;
  ...
  dbterm:leaderTitle "Mayor" ;
  dbterm:leaderName dbpedia:Job_Cohen ;
  ...
  dbterm:areaTotalKm "219" ;
  ...
dbpedia:ABN_AMRO
  dbterm:location dbpedia:Amsterdam ;
  ...
```

```
<http://dbpedia.org/resource/Amsterdam>  
owl:sameAs <http://rdf.freebase.com/ns/...> ;  
owl:sameAs <http://sws.geonames.org/2759793> ;  
...
```

```
<http://sws.geonames.org/2759793>  
owl:sameAs <http://dbpedia.org/resource/Amsterdam>  
wgs84_pos:lat "52.3666667" ;  
wgs84_pos:long "4.8833333" ;  
geo:inCountry <http://www.geonames.org/countries/#NL> ;  
...
```

Processors can switch automatically from one to the other...

Extracted Wikipedia

→ DBpedia.org

Linked Open Data

SPARQL Query Interface



navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

search

interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox


- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages

- Bosanski
- Brezhoneg
- Български
- Català
- Česky

August Strindberg

From Wikipedia, the free encyclopedia

 This article **needs additional citations for verification**. Please help improve this article by adding **reliable references**. Unsourced material may be **challenged** and removed. *(December 2006)*

"Strindberg" redirects here. For other uses, see Strindberg (disambiguation).

Johan August Strindberg (ⓘ pronounced ⓘ) (22 January 1849 – 14 May 1912) was a Swedish playwright and writer. He is arguably the most influential and most important of all Swedish authors, and one of the most influential Scandinavian authors, along with Knut Hamsun, with whom he fraternized while in Paris in the mid 1890s, Henrik Ibsen, Søren Kierkegaard and Hans Christian Andersen. Strindberg is known as one of the fathers of modern theatre. His work falls into two major literary movements, Naturalism and Expressionism.^[1]

Contents [hide]

- 1 Biography
 - 1.1 Early years
 - 1.2 Career
 - 1.3 Politics
 - 1.4 Writing
 - 1.5 Other interests
 - 1.6 Personal life
- 2 Bibliography
 - 2.1 Drama
 - 2.2 Poetry, fiction, and autobiography
- 3 In popular culture
- 4 Gallery
- 5 References
- 6 Sources
- 7 External links

Biography [edit]

Early years [edit]

Strindberg was the third son of Carl Oscar Strindberg, a shipping agent, and Ulrika Eleonora (Nora) Norling. Ulrika was twelve years Carl's junior and of humble origin, called a "domestic servant woman" by Strindberg. He used this

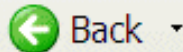
August Strindberg



Born	Johan August Strindberg 21 January 1849 Stockholm, Sweden
Died	14 May 1912 (aged 63) Stockholm, Sweden
Occupation	Playwright, author
Literary movement	Naturalism Expressionism
Signature	



Address http://dbpedia.org/page/August_Strindberg



Search



Favorites



About: [August Strindberg](#)

An Entity in Data Space: dbpedia.org

Johan August Strindberg was a Swedish writer, playwright, and painter. Along with Henrik Ibsen, Søren Kierkegaard and Hans Christian Andersen he is arguably the most influential and most important of the fathers of modern theatre. His work falls into two major literary movements, Naturalism and Expressionism.

Property	Value
dbpedia-owl:birthdate	1849-01-21 (xsd:date)
dbpedia-owl:birthplace	dbpedia:Stockholm dbpedia:Sweden
dbpedia-owl:deathdate	1912-05-14 (xsd:date)
dbpedia-owl:deathplace	dbpedia:Stockholm dbpedia:Sweden
dbpedia-owl:movement	dbpedia:Expressionism dbpedia:Naturalism_%28literature%29
dbpedia-owl:occupation	dbpedia:Painting dbpedia:Playwright dbpedia:Writer

[p:abstract](#)

Johan August Strindberg was a Swedish writer, playwright, and painter. He fraternized while in Paris in the mid-1860s with the Symbolists. August Strindberg, de son nom complet August Fredrik Strindberg, né le 22 août 1849 à Stockholm, était un écrivain suédois. ヨハン・アウグスト・ストリンドベリ(12月13日)は、スウェーデンの作家。(ja) Johan August Strindberg (Stockholm, 22 august 1849 - 14 maj 1912) toneelschrijver naam gemaakt, maar zijn werk omvat vrijwel elk literair genre. August Strindberg (ur. 22 stycznia 1849 w Sztokholmie, zm. 14 maja 1912) powieści, esejów i utworów poetyckich, malarz i fotograf. Uznawany za

1. Use URIs as names for things
2. Use http URIs

3. When someone looks up a name, provide useful information

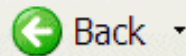
4. Include links to other URIs so that they can discover more things

properties also get URIs

rdfs:label	August Strindberg (en) August Strindberg (fr) ヨハン・アウグスト・ストリンドベリ August Strindberg (nl) August Strindberg (pl) Стриндберг, Юхан Август (ru) August Strindberg (sv) August Strindberg (es) August Strindberg (fi) August Strindberg (it) August Strindberg (pt) 奥古斯特·斯特林堡 (zh) August Strindberg (da) August Strindberg (de) August Strindberg (no)
owl:sameAs	http://www4.wiwiss.fu-berlin.de/gutendata/resource/people/Strindberg_August_1849-1 http://rdf.freebase.com/ns/guid.9202a8c04000641f80000000000617ed
skos:subject	dbpedia:Category:1912_deaths dbpedia:Category:1849_births dbpedia:Category:Alchemists dbpedia:Category:Expressionist_dramatists_and_playwrights dbpedia:Category:French-language_writers dbpedia:Category:Modernist_drama%2C_theatre_and_performance dbpedia:Category:People_from_Stockholm dbpedia:Category:Swedish-language_writers dbpedia:Category:Swedish_dramatists_and_playwrights dbpedia:Category:Swedish_socialists dbpedia:Category:Swedish_writers dbpedia:Category:Swedish_novelists dbpedia:Category:Uppsala_University_alumni dbpedia:Category:Naturalist_dramatists_and_playwrights
foaf:depiction	http://upload.wikimedia.org/wikipedia/commons/thumb/4/49/August_Strindberg.jpg/200px-August_Strindberg.jpg
foaf:img	http://upload.wikimedia.org/wikipedia/commons/4/49/August_Strindberg.jpg
foaf:name	August Strindberg
foaf:page	http://en.wikipedia.org/wiki/August_Strindberg
is dbpedia-owl:author of	dbpedia:Inferno_%28Strindberg%29
is dbpedia-owl:influenced of	dbpedia:Emanuel_Swedenborg dbpedia:Otto_Weininger dbpedia:S%C3%B8ren_Kierkegaard
is dbpedia-owl:influences of	dbpedia:Tennessee_Williams
is dbpedia-owl:writer of	dbpedia:List_of_Swedish_novels
is p:author of	dbpedia:List_of_Swedish_novels

Include links to other URIs so that they can discover more things

properties get URIs



Search



Favorites



About: August Strindberg

An Entity in Data Space: dbpedia.org

Johan August Strindberg was a Swedish writer, playwright, and painter. Along with Knut Hamsun, with whom he fraternized while in Paris in the mid 1890s, with whom he is arguably the most influential and most important of one of the fathers of modern theatre. His work falls into two major literary movements, Naturalism and Expressionism.

Property	Value
dbpedia-owl:birthdate	1849-01-21 (xsd:date)
dbpedia-owl:birthplace	dbpedia:Stockholm dbpedia:Sweden
dbpedia-owl:deathdate	1912-05-14 (xsd:date)
dbpedia-owl:deathplace	dbpedia:Stockholm dbpedia:Sweden
dbpedia-owl:movement	dbpedia:Expressionism dbpedia:Naturalism_%28literature%29
dbpedia-owl:occupation	dbpedia:Painting dbpedia:Playwright dbpedia:Writer
p:abstract	Johan August Strindberg was a Swedish writer, playwright, and painter. fraternized while in Paris in the mid 1890s, ... »more« (en) August Strindberg, de son nom complet Johan August Strindberg, né le 22 janvier 1849 à Stockholm, était un écrivain, dramaturge ... »more« (fr) ヨハン・アウグスト・ストリンドベリ (12px Johan August Strindberg (ja) Johan August Strindberg (Stockholm, 22 januari 1849 – aldaar, 14 mei 1912) was een Zweeds toneelschrijver naam gemaakt, maar zijn werk omvat vrijwel elk literair genre. August Strindberg (ur. 22 stycznia 1849 w Sztokholmie, zm. 14 maja 1912) – polski powieściopisarz, esejista i utworów poetyckich, malarz i fotograf. Uznawany za

- Templates cover only very small fraction of knowledge
- Named links in Wikipedia do not provide all required relationships
- Large knowledge resides within “plain” HREFs
... but this knowledge is hidden in text created for humans

Need to parse free text to get meaningful relationships

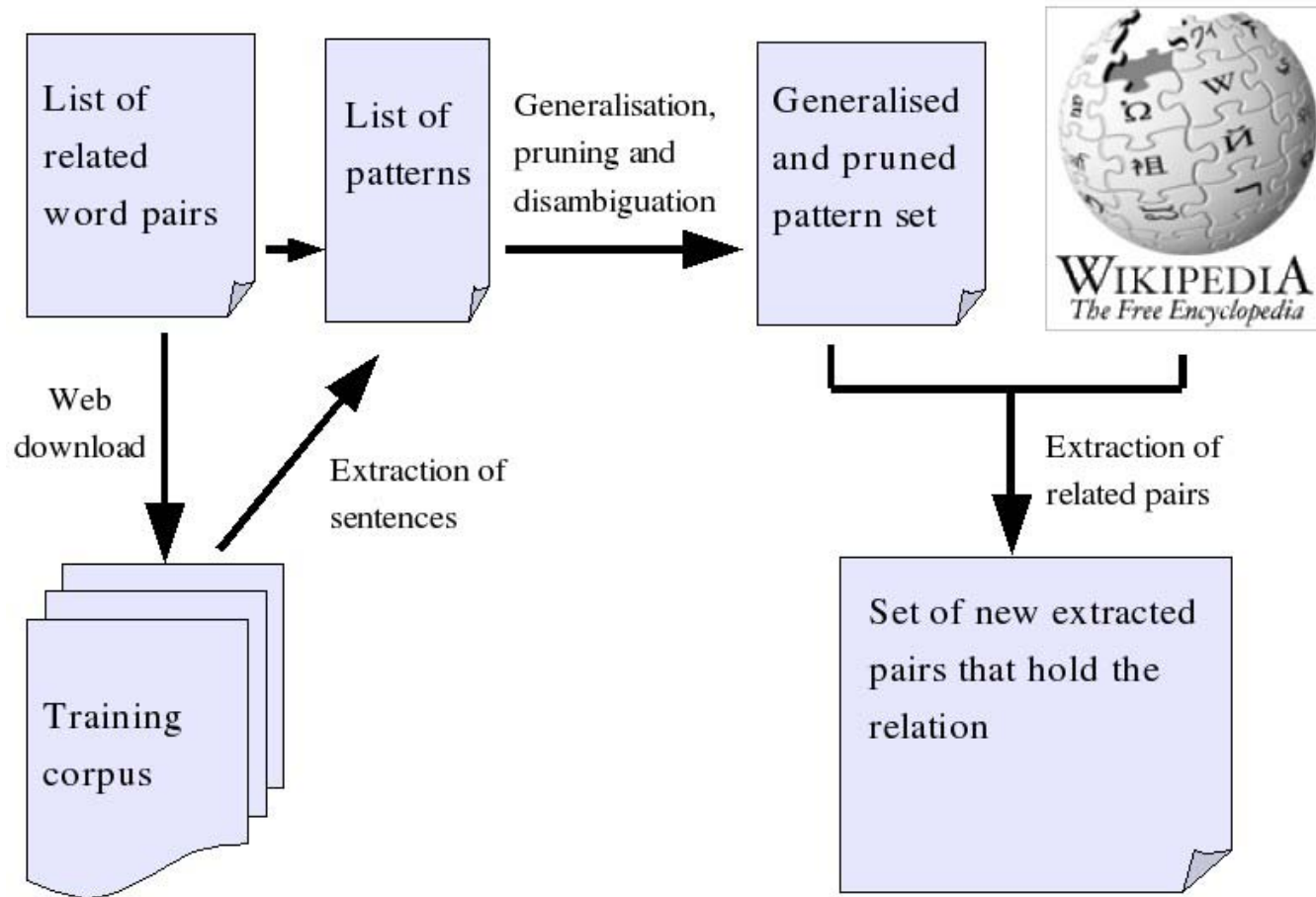
More Patterns....

- Rule and Heuristic based method
 - ◆ Pattern-based approach
 - ◆ Uses WordNet

 - ◆ YAGO [Suchanek et. al, 2007]
 - ◆ [Ruiz-Casado et. al. 2006]
 - ◆ [Weld et al]
 - ◆ [Suchanek et al 2009]

- Learning common patterns for interesting relationships
 - ◆ is-author-of, is-the-capital-of, is-employee-of, ...

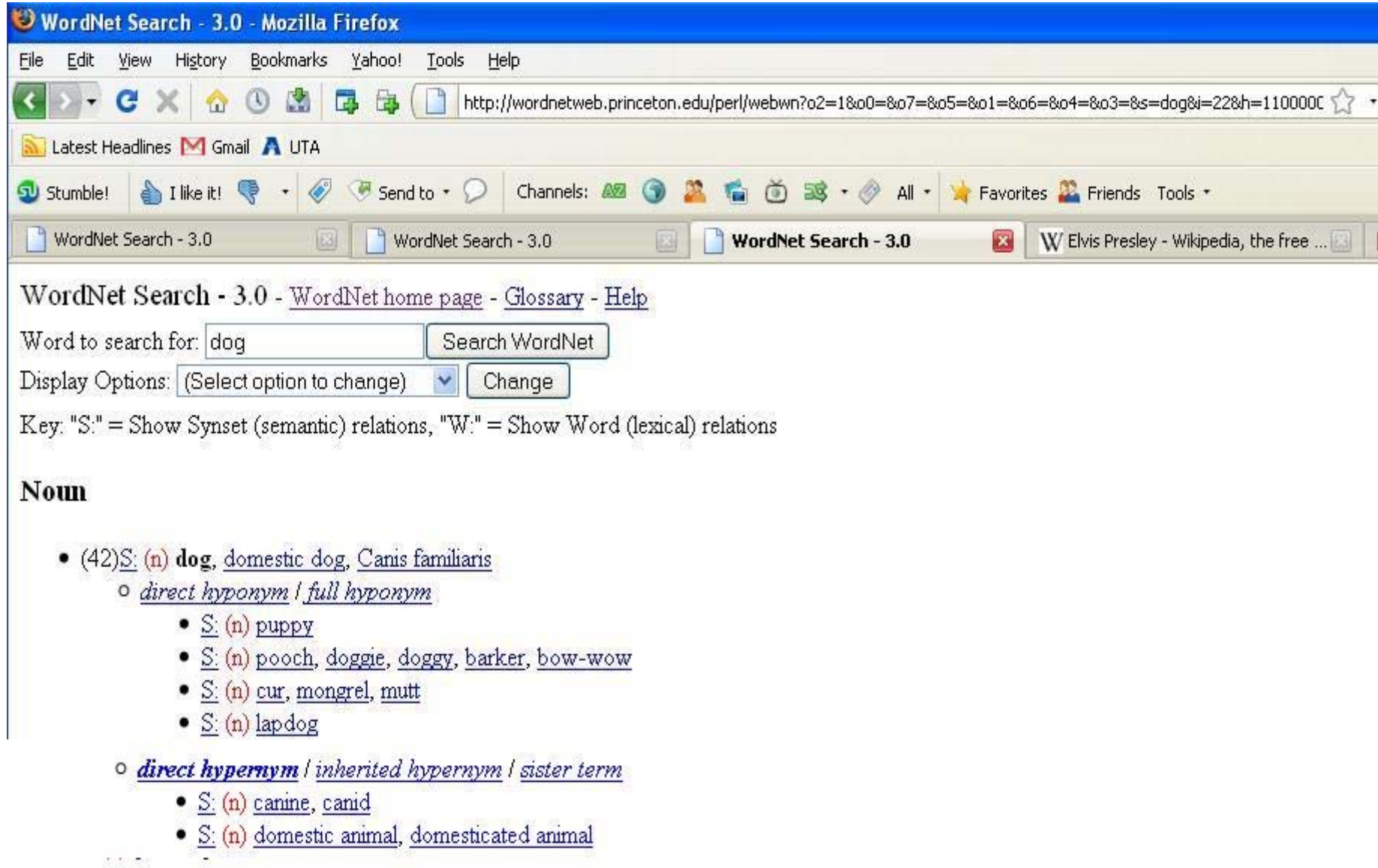
 - ◆ Seed initial patterns and look for them in training corpus
 - ◆ Extract common patterns between linked entries
 - Search for support in Wikipedia
 - Search for pattern support using search engines
 - ◆ Generalize it
 - ◆ Apply it to extract more information



- Example (known entities are underlined):
 - Alfred Hitchcock directed the famous film Psycho
 - Alfred Hitchcock directed the well known film Psycho
 - ?x directed the famous film ?y
 - ?x directed the well known film ?y
 - ?x directed the famous | well known film ?y
 - ?x directed the * famous | known film ?y
- Apply pattern
 - Alfred Hitchcock directed the famous film The Birds
 - Bernardo Bertolucci directed the well known film The Last Emperor
 - Woody Allen directed the amusing and famous film Annie Hall

- Good patterns
 - ◆ Support in training data
 - ◆ Found in Wikipedia documents
 - ◆ Unambiguous – matching same types / topics
 - Not the best one: **?x's ?y**
 - Einstein's Theory of General Relativity
 - Bosco's The Garden of Delights
 - Tolkien's Lord of the Rings
 - Need pruning
 - ◆ Support in free text search
 - Check how often the pairs you found are matched with this pattern in web documents

- Lexical database for the English language
- Created at the Cognitive Science Laboratory of Princeton University
- Groups English words into sets of synonyms called *synsets*
- Provides short, general definitions
- Provides hypernym/hyponym relations
 - ♦ e.g. canine is hypernym, dog is hyponym



WordNet Search - 3.0 - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

http://wordnetweb.princeton.edu/perl/webwn?o2=1&o0=&o7=&o5=&o1=&o6=&o4=&o3=&s=dog&j=22&h=110000C

Latest Headlines Gmail UTA

Stumble! I like it! Send to Channels: Favorites Friends Tools

WordNet Search - 3.0 WordNet Search - 3.0 WordNet Search - 3.0 Elvis Presley - Wikipedia, the free ...

WordNet Search - 3.0 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- (42)S: (n) **dog**, [domestic dog](#), [Canis familiaris](#)
 - [direct hyponym / full hyponym](#)
 - S: (n) [puppy](#)
 - S: (n) [pooch](#), [doggie](#), [doggy](#), [barker](#), [bow-wow](#)
 - S: (n) [cur](#), [mongrel](#), [mutt](#)
 - S: (n) [lapdog](#)
 - [direct hypernym / inherited hypernym / sister term](#)
 - S: (n) [canine](#), [canid](#)
 - S: (n) [domestic animal](#), [domesticated animal](#)

- Goal: create class hierarchy
 - ◆ e.g. singer subclassOf performer
performer subclassOf artist
- hyponymy relation from WordNet
- Wikipedia class 'American people in Japan' is subclass of WordNet class 'person'