Semantics I

IN4325 – Information Retrieval

Today

- We finally finish the evaluation part
- Knowledge-based semantics
  - applications
- Statistics-based "semantics"
  - applications

Evaluation: state-of-the-art

Turning away from binary qrels

1960s

Cranfield

what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft.

Cranfield

what agreement is found between theoretically predicted instability times and experimentally measured collapse times for compressed columns in creep.

today

ClueWeb09

cheap internet

ClueWeb09

dinosaurs

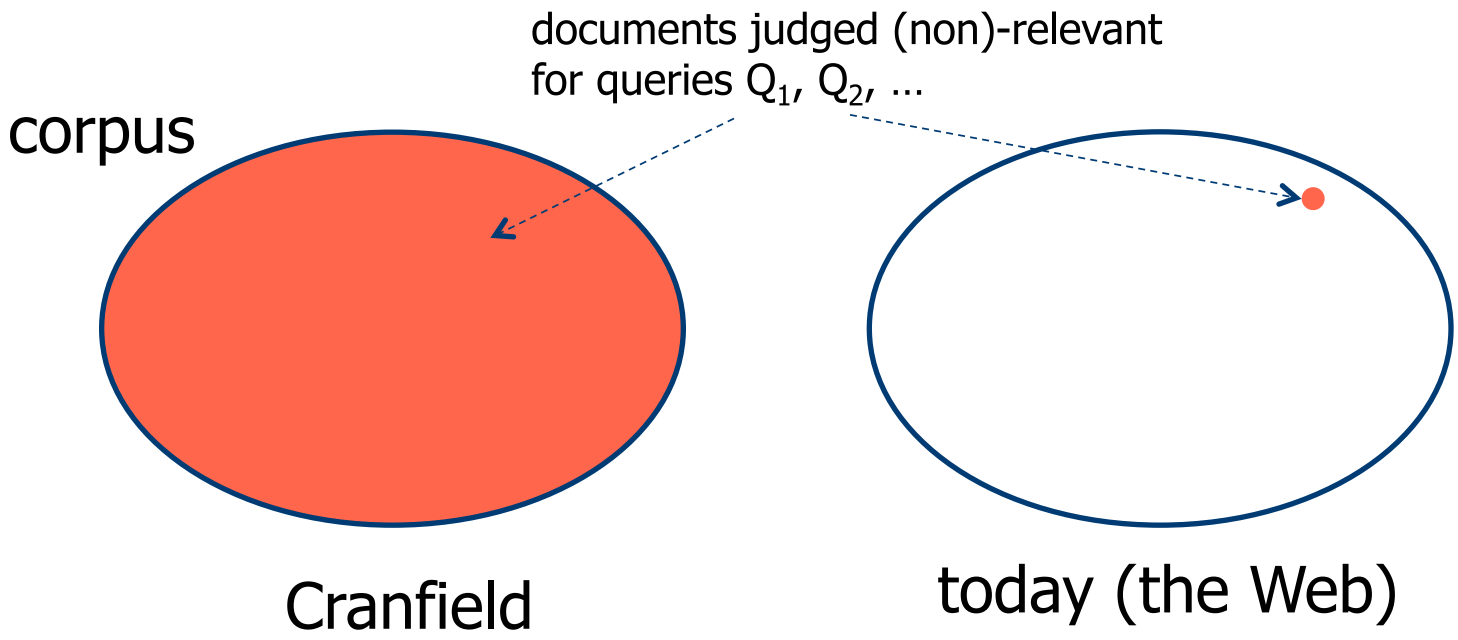
ClueWeb09

yahoo

ClueWeb09

solar panels

Cranfield paradigm vs. today



Normalized Discounted Cumulative Gain

Järvelin & Kekäläinen, 2002 [5]

- Graded relevance scales (e.g. 0-3)
- Used in Web search evaluations
- NDCG measures the „gain" of documents
- Assumptions
  - highly relevant documents are more valuable than marginally relevant documents
  - The greater the ranked position of a relevant document, the less valuable it is for the user (less likely that the user will examine it)
    - User has limited time, may have seen the information already, cognitive load, etc.

Normalized Discounted Cumulative Gain

Järvelin & Kekäläinen, 2002 [5]

- Direct cumulative gain can be defined iteratively

$$G' = < 3,2,3,0,0,1,2,2,3,0,\ldots >$$

$$CG_i = \begin{cases} G_1, if\ i = 1 \\ CG_{i-1} + G_i, otherwise \end{cases}$$

$$CG' = < 3,5,8,8,8,9,11,13,16,16,\ldots >$$

CG at each rank can be read directly

Normalized Discounted Cumulative Gain

Järvelin & Kekäläinen, 2002 [5]

- **Discounted cumulative gain**: reduce the document score as ist rank increases (but not too steeply)
  - Divide the document score by the log of ist rank
  - Base of the logarithm determines discount factor

$$\log_2 1 = 0$$

$$\log_2 2 = 1$$

$$\log_2 1024 = 10$$

$$DCG_i = \begin{cases} CG_i, \mathit{if}\ \ i < b \\ CG_{i-1} + G_i / \log_b i, \mathit{if}\ \ i \geq b \end{cases}$$

$$assume\ b = 2$$

$$CG' = < 3,5,8,8,8,9,11,13,16,16,... >$$

$$DCG = < 3,5,6.89,6.89,7.28,7.99,8.66,9.61,9.61... >$$

Normalized Discounted Cumulative Gain

Järvelin & Kekäläinen, 2002 [5]

- Normalized DCG: compare DCG to the theoretically best possible
- Ideal vectors constructed as follows:

$$BV_i = \begin{cases} 3, if\ \ i \leq m, \\ 2, if\ \ m < i \leq m+l, \\ 1, if\ \ m+l < i \leq m+l+k, \\ 0, otherwise \end{cases}$$

k, l, m are the relevant documents at relevance levels 1, 2 and 3 for *Q*

based on the search topic, not retrieval results

$$I' =< 3,3,3,2,2,2,1,1,1,1,0,0,0,... >$$

$$CG_{I'} =< 3,6,9,11,13,15,16,17,18,19,19,19,19,... >$$

$$DCG_{I'} =< 3,6,7.89,8.89,9.75,10.52,10.88,11.21,11.53,11.83,... >$$

Normalized Discounted Cumulative Gain

Järvelin & Kekäläinen, 2002 [5]

- Normalized DCG: compare DCG to the theoretically best possible
  - The DCG vectors are divided component-wise by the corresponding ideal DCG vectors
    - 1: ideal performance
    - [0,1): share of ideal performance

- In practice NDCG for query Q at rank k:

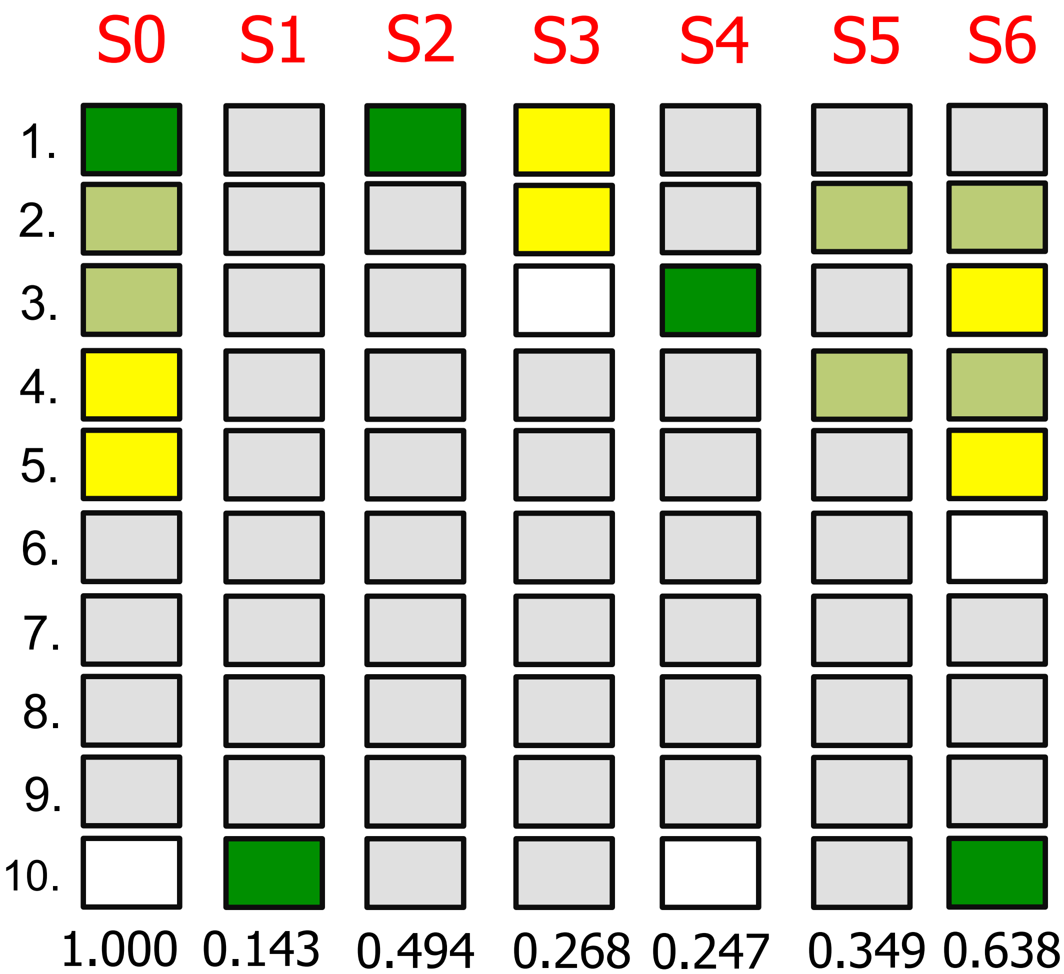
rel. score assessors gave to D at query j

$$NDCG(Q,k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)}$$

normalization factor so that a perfect ranking's NDCG at k for query j is 1

NDCG examples

One query, seven systems.



relevance scores 0-3

0: not relevant

1: marginally relevant

2: somewhat relevant

3: very relevant

In total:

NDCG@10

(trec_eval scores)

Latent spaces & concepts

Document: The apple is red.
Red is a color. Eat an apple.

Query: red

Document: 1 2 3 4.
4 3 5 6. 7 8 2.

Query: 4

What is wrong with keyword-based IR?

- Relies on syntax alone (string matching)
- Shortcomings: multi-word expressions, synonymy, polysemy, generalizations (vocabulary problem)

<DOC>**Space station** makes progress with computers</DOC>
Query: “ISS”

<DOC>**Apple PCs gaining gaming credibility**</DOC>
<DOC>**Canada imposes apple moth restrictons**</DOC>
Query: “apple”

<DOC>**Germany’s Merkel meets treaty opponent Poland**</DOC>
Query: “German chancellor”

What is wrong with keyword-based IR?

- Relies on
- Shortcom
  generaliz

<DOC>**Spa**
Query: “

<DOC>**App**
<DOC>**Can**
Query: “apple”

<DOC>**Germany’s Merkel meets treaty opponent Poland**</DOC>
Query: “German chancellor”

$$D_1 = \begin{pmatrix} Netherlands, 1 \\ Holland, 25 \\ country, 5 \\ .... \end{pmatrix}, D_2 = \begin{pmatrix} Netherlands, 1 \\ Holland, 0 \\ country, 5 \\ ... \end{pmatrix}, Q = \{the\ Netherlands\}$$

separate dimensions in VSM – no connections

The advantages of concepts/semantics

- If the machine "knows" the meaning of terms and phrases, the retrieval effectiveness improves
  - The **vocabulary gap** between users and text authors vanishes

> Our world is being revolutionized by data-driven methods: access to large amounts of data has generated new insights and opened exciting new opportunities in commerce, science, and computing applications. Processing the enormous quantities of data necessary for these advances requires large clusters, making distributed computing paradigms more crucial than ever. **MapReduce** is a programming model for expressing distributed computations on massive datasets and an execution framework for large-scale data processing on clusters of commodity servers. The programming model provides an easy-to-understand abstraction for designing scalable algorithms, while the execution framework transparently handles many system-level details, ranging from scheduling to synchronization to fault tolerance. This book focuses on **MapReduce** algorithm design, with an emphasis on text processing algorithms common in natural language processing, information retrieval, and machine learning. We introduce the notion of **MapReduce** design patterns, which represent general reusable solutions to commonly occurring problems across a variety of problem domains. This book not only intends to help the reader "think in **MapReduce**", but also discusses limitations of the programming model as well.

query: hadoop

- Goal: retrieval based on the semantics (the topics) of documents and queries

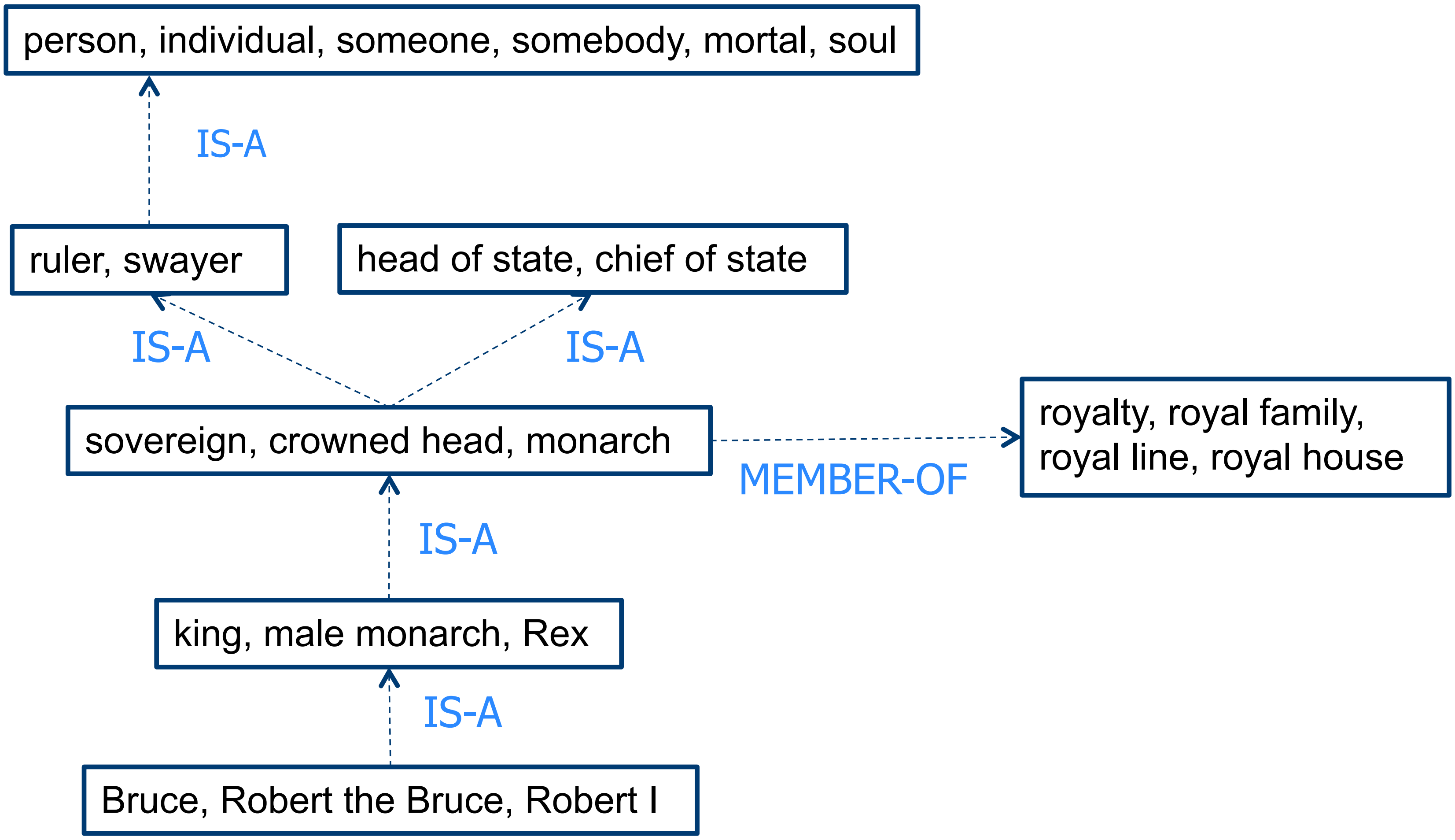
Concept extraction from documents

- Source of extra-document knowledge: knowledge bases vs. statistics

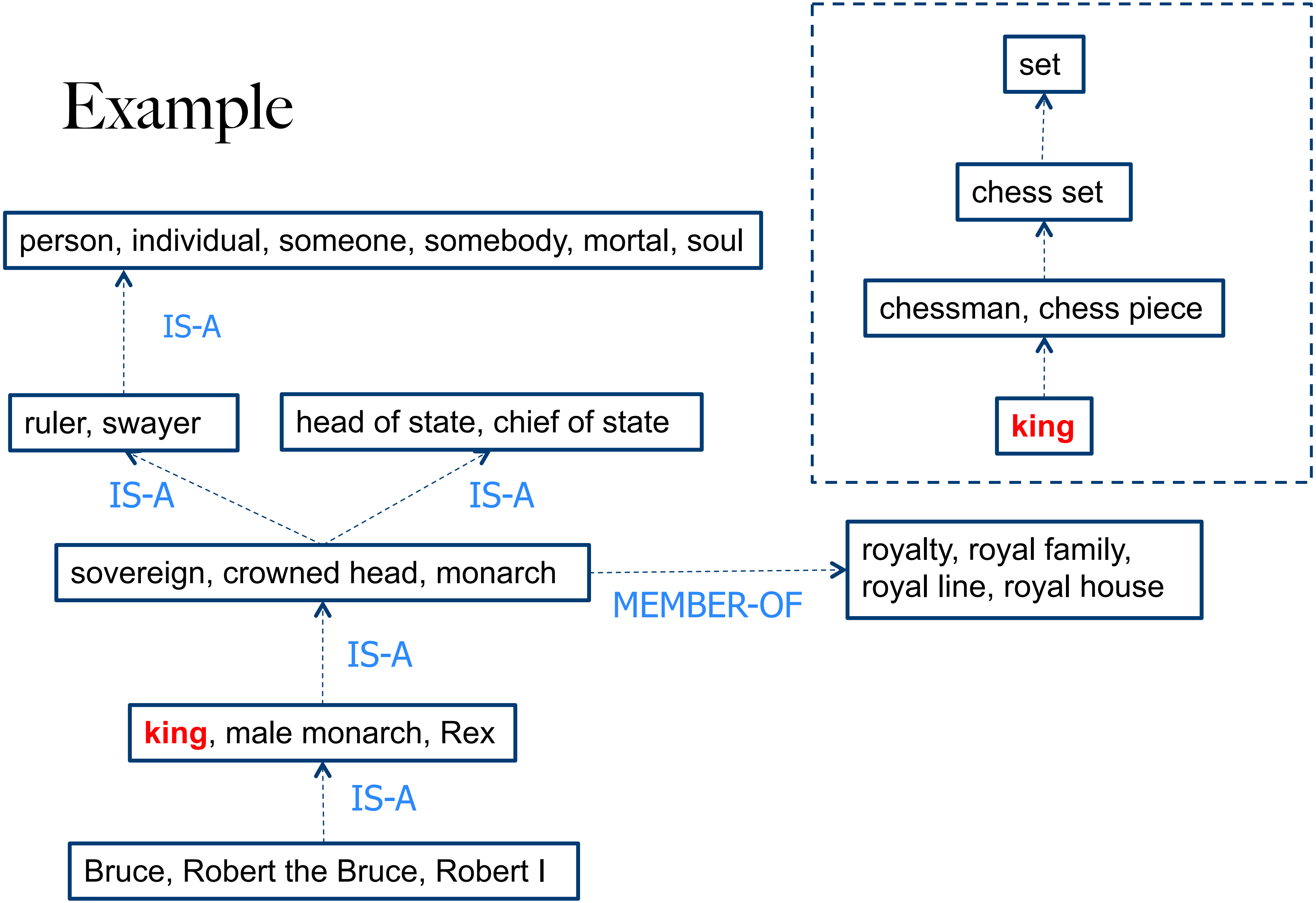
- Knowledge bases
  - Examples: WordNet, UMLS
  - Advantages: precise, ~faultless
  - Disadvantages: maintainance, up-to-datedness, corpus-dependency

- Statistics (LSA, PLSA, LDA, HAL ….)
  - Derived from the corpus itself or external corpora (the Web)
  - Advantages: automatic generation, corpus-dependency
  - Disadvantages: noisy, error-prone

Example


person, individual, someone, somebody, mortal, soul
IS-A
ruler, swayer
head of state, chief of state
IS-A
IS-A
sovereign, crowned head, monarch
MEMBER-OF
royalty, royal family, royal line, royal house
IS-A
king, male monarch, Rex
IS-A
Bruce, Robert the Bruce, Robert I

Example


person, individual, someone, somebody, mortal, soul
IS-A
ruler, swayer
head of state, chief of state
IS-A
IS-A
sovereign, crowned head, monarch
MEMBER-OF
royalty, royal family, royal line, royal house
IS-A
king, male monarch, Rex
IS-A
Bruce, Robert the Bruce, Robert I
set
chess set
chessman, chess piece
king

WordNet

http://wordnet.princeton.edu/

- Lexical database of the English language
  - Models common-sense world knowledge

- Inspired by psycholinguistic theories of human lexical memory

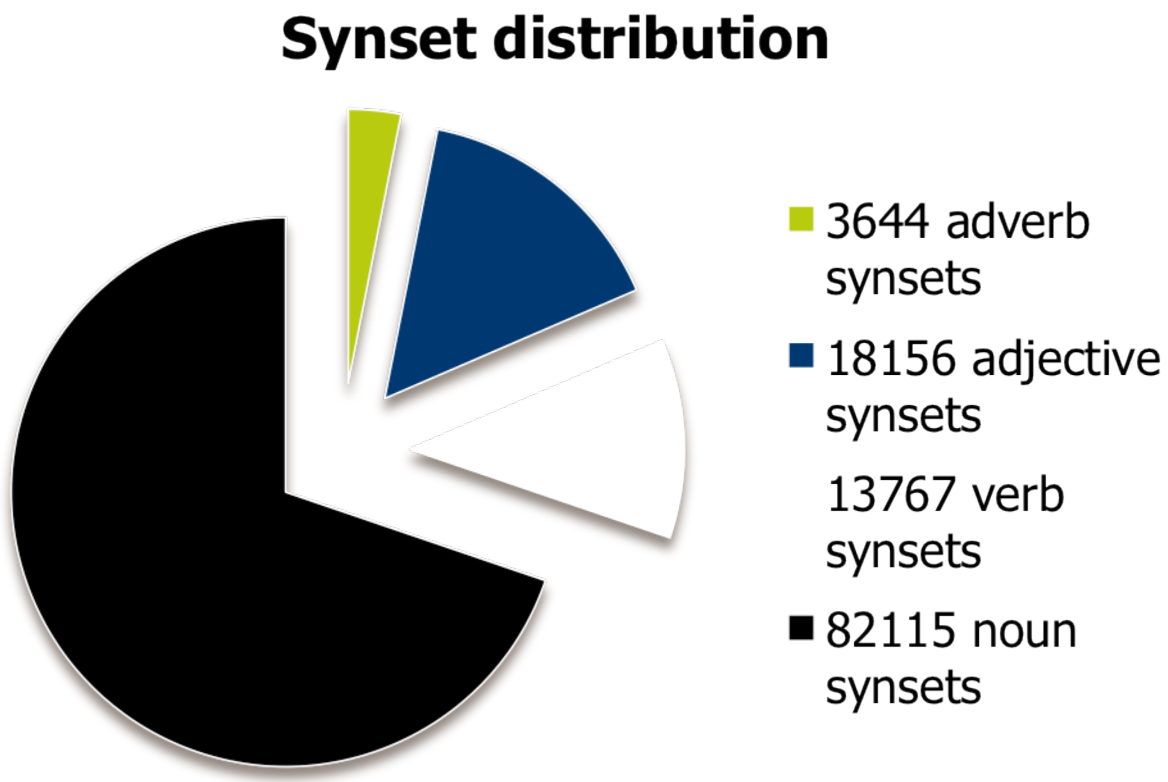
- Manual effort led by George A. Miller (Princeton University)

- Nouns, adjectives, adverbs and verbs are organized by lexical concepts (sets of synonyms -> synsets)

- The syntactic categories form largely separate networks

- Lexical concepts linked through relations

WordNet 3.0 statistics



Most forms have few senses, a few have many senses (Zipf distribution)

Example: instrument

WordNet's sense entries consist of a set of synonyms, dictionary style definitions (gloss) and example uses

Noun

- S: (n) **instrument** (a device that requires skill for proper use)
- S: (n) **instrument**, tool (the means whereby some act is accomplished) *"my greed was the instrument of my destruction"; "science has given us new tools to fight disease"*
- S: (n) **instrument**, pawn, cat's-paw (a person used by another to gain an end)
- S: (n) legal document, legal instrument, official document, **instrument** ((law) a document that states some contractual relationship or grants some right)
- S: (n) instrumental role, **instrument** (the semantic role of the entity (usually inanimate) that the agent uses to perform an action or start a process)
- S: (n) musical instrument, **instrument** (any of various devices or contrivances that can be used to produce musical tones or sour

Verb

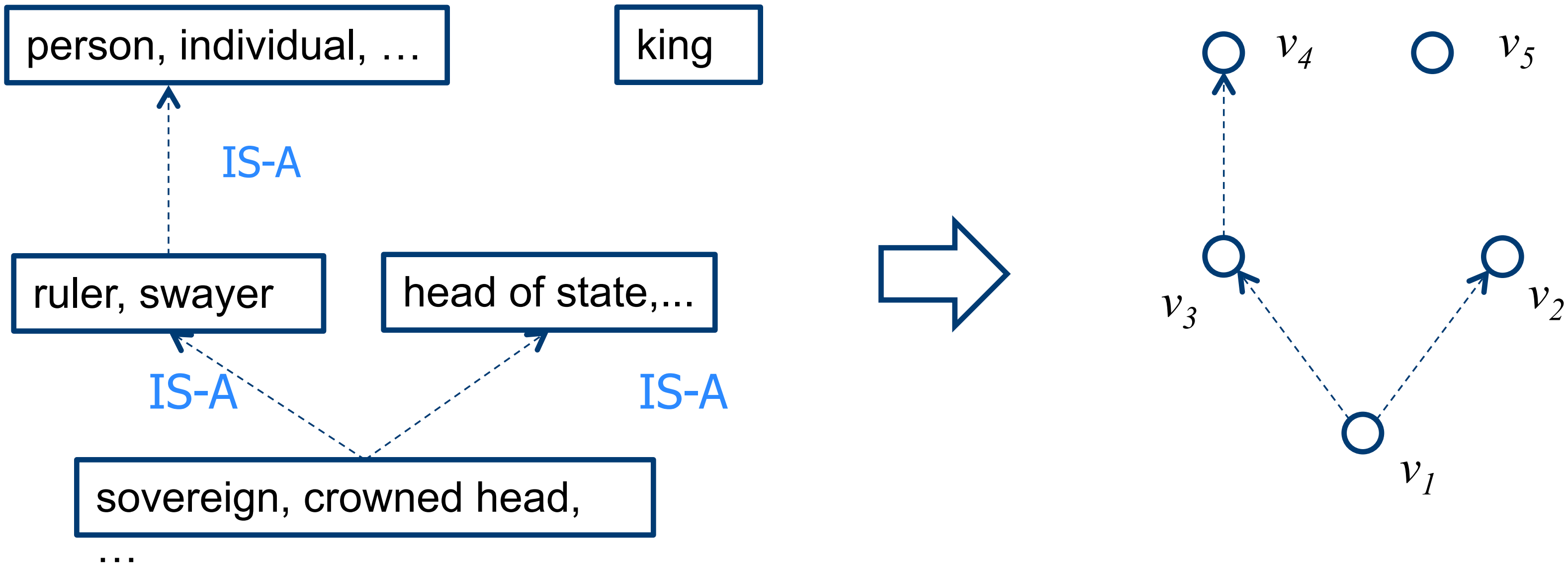
- S: (v) **instrument** (equip with instruments for mea controlling)
- S: (v) **instrument**, instrumentate (write an instrum
- S: (v) **instrument** (address a legal document to)

- A word sense is made up of synonyms (words with similar/identical meaning)
- Each word sense is mapped to a synset -> a synset is a set of words that are interchangeable in some context

Important WordNet relations

- Noun relations
  - Hyponymy / hypernymy (“is-a”)
    - *red is a hyponym of color*; color is a hypernym of red
  - Holonymy / meronymy (“part-of”)
    - a toe is a meronym of leg, leg is a holonym of toe
  - Antonymy
    - king is an antonym of queen, queen is an antonym of king
- Verb relations
  - Hypernymy, antonymy
  - Entailment
    - to snore lexically entails to sleep (you have to sleep to be able to snore)
  - Troponymy
    - to limp is a troponym of to walk (a more specific description of walking)
- Adverb/adjective relations
  - Antonymy

WordNet as a graph



- In-degree d_{in} of a node: number of incoming edges $d_{in}(v_1)=0,\ d_{in}(v_2)=1$
- Out-degree d_{out} of a node: number of outgoing edges $d_{out}(v_1)=2,\ d_{out}(v_2)=0$
- Path P: sequence of nodes with an edge between neighbouring nodes

  $P=\{v_1,v_3,v_4\}$
- Distance $d(v_i,v_j)$ between nodes v_i and v_j: length of the shortest path between them $d(v_1,v_4)=2,\ d(v_1,v_2)=1,\ d(v_1,v_1)=0,\ d(v_1,v_5)=inf.$

Relatedness measures on WordNet

- Calculate the distance/relatedness between words or synsets within the same syntactic category
- Proposed measures rely on different resources and apply to different categories

- Semantic *relatedness*: relationship between concepts is determiend by any kind of relation
- Semantic *similarity*: considers the IS-A relation only

- Two types
  - Based on senses
  - Based on words (usually involves all senses and min/max/av. scores)

Leacock-Chodorow similarity [5]

- Measurement *based on the IS-A shortest path* $len(c_1,c_2)$ *between two synsets* c_1, c_2
- Path length scaled by the overall depth D (~18 for nouns) of the taxonomy

$$sim(c_1,c_2) = -\log \frac{len(c_1,c_2)}{2D}$$

- Examples (maximum score across senses)
  - `sim(car#n,vehicle#n)=2.59`
  - `sim(car#n,boat#n)=2.08`
  - `sim(car#n,birthday_cake#n)=1.05`

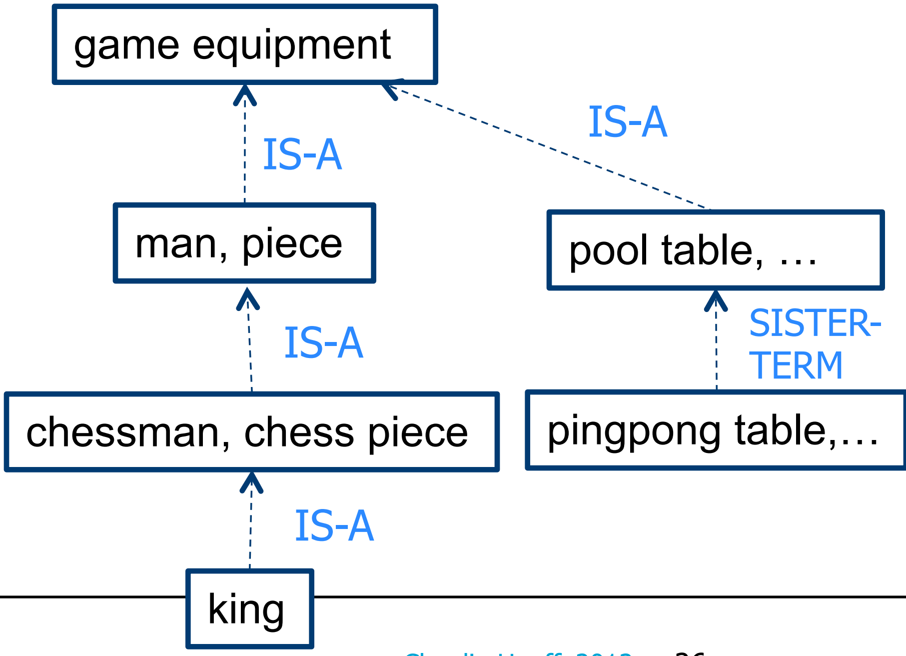
Hirst-St-Onge relatedness [5]

- Two concepts c_1 and c_2 are semantically close, if their synsets are connected by a path that is not too long and does not change direction too often
- Approach utilizes all relations within WordNet

$$rel(c_1,c_2) = C - pathLength - k \times d$$

$$d: number\ of\ changes\ in\ direction$$

$$C, k: parameters$$




```
rel(car#n,vehicle#n)=6.0
rel(car#n,boat#n)=3.0
rel(car#n,birthday_cake#n)=0.0
```

Resnik similarity [5]

```
sim(car#n,vehicle#n)=5.53
sim(car#n,boath#n)=5.53
sim(car#n,birthday_cake#n)=0.61
```

- Information content IC is a measure of specificity of a concept
  - *High IC indicates specificity*
  - *Low IC indicates generality*

- IC estimated by counting concept frequencies in a corpus
  - Sense-tagged corpus: use counts directly
  - Untagged corpus: assign counts to all synsets that contain the word
  - Counts trickle up the IS-A hierarchy (giant counts towards grownup)

- Similarity between two concepts is the IC of their lowest common superordinate

$$IC(c_i) = -\log(P_{ml}(c_i)) \qquad sim(c_1, c_2) = IC(lso(c_1, c_2))$$

Resnik similarity [5]

sim(car#n,vehicle#n)=5.53
sim(car#n,boath#n)=5.53
)=0.61

- Information co[r
  - *High IC indic[a*
  - *Low IC indica*
- IC estimated by
  - Sense-tagged
  - Untagged cor ... d
  - Counts trickle ... b)

This huge expansion will turn the whole of Britain into a <tag "530184">giant</> museum-cum-theme-park sometime next century.
Most Irishmen are very proud of what they have here, of the <tag "530189-p">Giant's</> Causeway and all along the Irish coastline.
Ford, the US car <tag "530183">giant</>, has already declared its interest in moving in on Jaguar.
The readiness of <tag "530184">giant</> Japanese and German companies to invest in new equipment and new lines of business contributes directly to industrial performance.

- Similarity between two concepts is the IC of their lowest common superordinate

$$IC(c_i) = -\log(P_{ml}(c_i)) \qquad sim(c_1,c_2) = IC(lso(c_1,c_2))$$

Which measure performs best? [5]

- Set of ground truth judgments
  - Rubenstein & Goodenough (1965) created 65 pairs of words; 51 human assessors rated their similarity of meaning (scale 0.0-4.0)

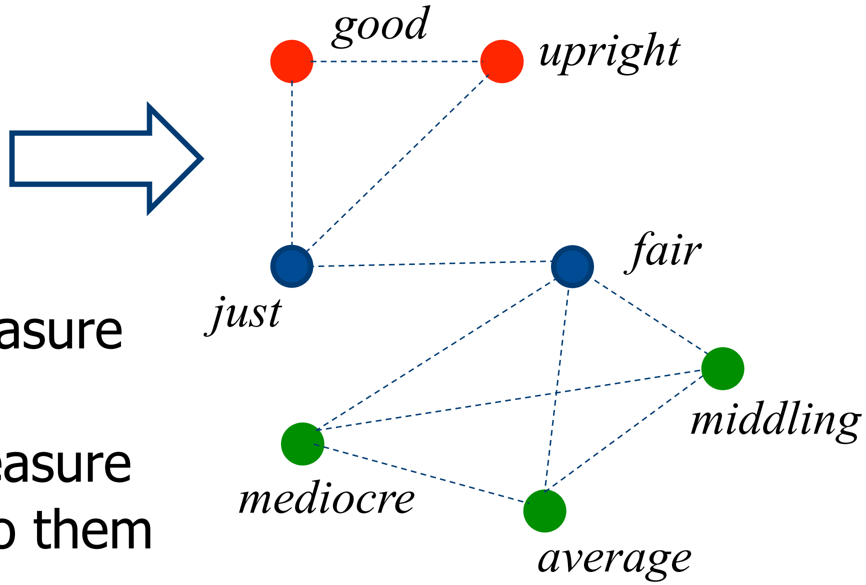
- Evaluation in terms of the correlation between measure based similarity and user based similarity judgments

Leacock-Chodorow *	**0.84**
Hirst-St-Onge	0.79
Resnik	0.78

Word-based distances

- Graph with **words** as nodes; edges between those words that appear in one synset

```
{good,just,upright}
{fair,just}
{average,fair,mediocre,middling}
```

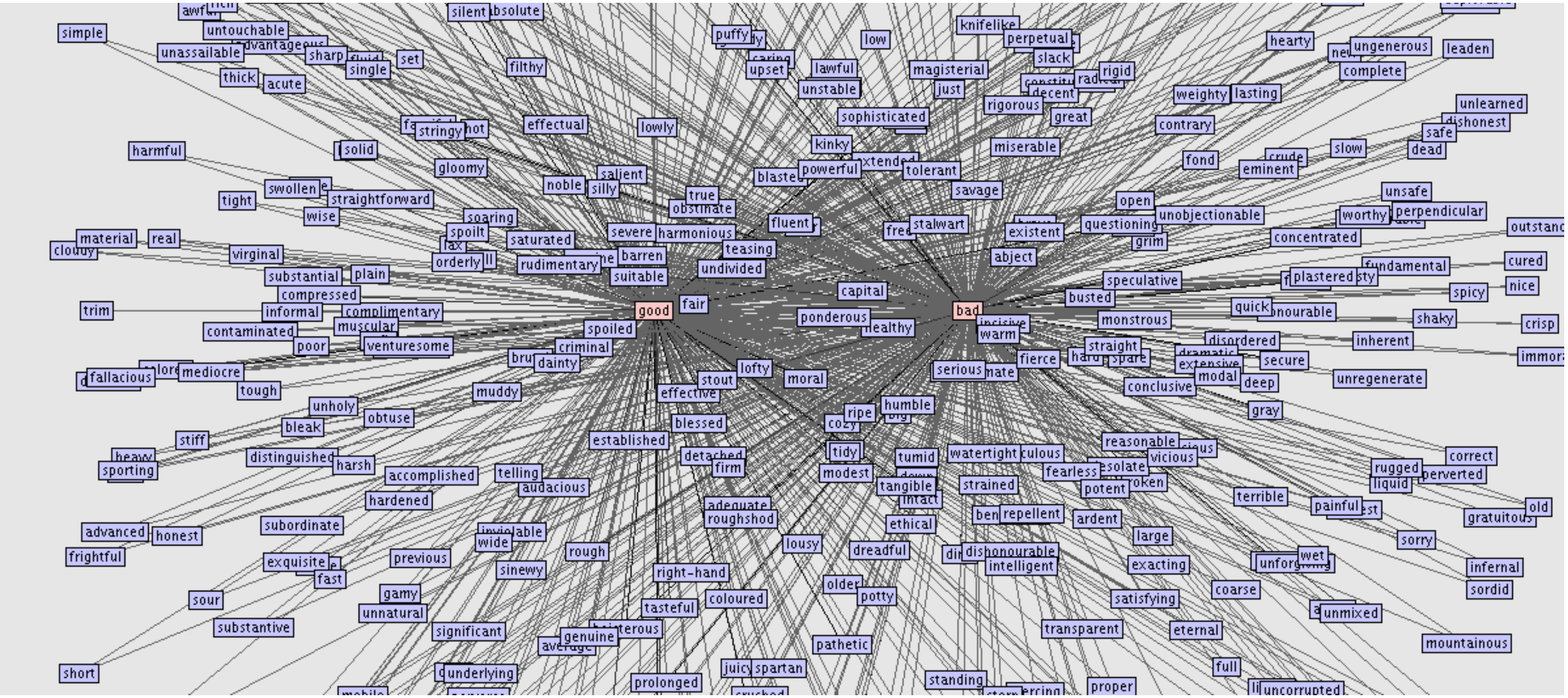


- Distance between 2 words as a measure of similarity, but $d(good,bad)=4$
- Use reference words w_1, w_2 and measure the relative distance of a word w_x to them

$$relDist(w_x) = \frac{d(w_x,w_1) - d(w_x,w_2)}{d(w_1,w_2)}, \; relDist \in [-1,1]$$

Word-based distances

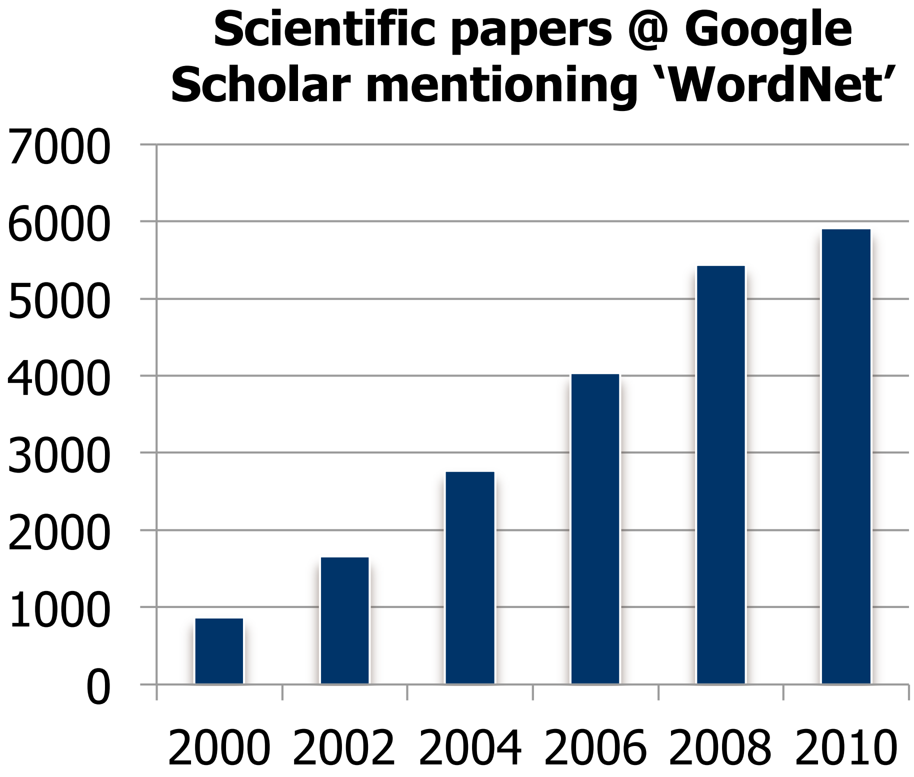
Kamps et al, 2004



Source: http://staff.science.uva.nl/~kamps/wordnet/

WordNet applications

- Word sense disambiguation
- Refinement of search queries
- Semantic orientation of texts
- Detection and correction of spelling errors
- Summarization
- …



Search query refinement

Nemrava, 2006

P_0 *such as* P_1, P_2, …, P_{n-1} *(and|or)* P_n

- Motivation: keyword searches are often ambiguous
  - "Java", "Python", "Pluto"
- Idea: give the user a choice between possible hypernyms

  *if you do not have a query log

1. Look up WordNet synsets and glosses for the query
2. Process the glosses (POS tagger, stopword removal) and keep the nouns as potential hypernyms
3. Apply Hearst patterns on the Web & retrieve #result pages for each candidate
4. Consider the candidate with the highest score as hypernym

Search query refinement

1. WordNet glosses for query term "Pluto"

- **SYN1** a small planet and the farthes known planet from the sun; has the most elliptical orbit of all the planets
- **SYN2** (Greek mythology) the god of the underworld in ancient mythology; brother of Zeus and husband of Persephone
- **SYN3** a cartoon character created by Walt Disney

2. Candidate nouns

- **SYN1** planet, sun, orbit, planets
- **SYN2** Greek, god, underworld, mythology, brother, Zeus, husband, Persephone
- **SYN3** cartoon, character, Walt, Disney

3. Hearst patterns and page counts (shown for SYN1 only)

- "Pluto is a planet" (1550), "Pluto is planet" (145) "Pluto is a sun" (2), "Pluto is sun" (0)
- "Pluto is an orbit" (1), "Pluto is orbit" (1) "Pluto is a planets" (0), "Pluto is planets" (0)

4. Refinement offers: "Pluto planet", "Pluto god", "Pluto cartoon"

Detection & correction of spelling errors

- Intuition: in coherent texts, many instances of related pairs of words can be found
- Words are disambiguated where possible by accepting senses that are semantically related to possible senses of other nearby words

I walk along the river ➔ I talk along the river
His dog barks ➔ His dog parks
The sun shines ➔ The son shines

Detection & correction of spelling errors

Error detection approach: Hirst & Budanitsky, 2005 [7]

1. Consider word w that appears only once in the text and is semantically unrelated to nearby words

2. Create spelling variations w_1, w_2, w_3, … of w by a single insertion, deletion or transposition

walk → talk
barks → _arks
probalbe→ probable

3. Check w_1, w_2, w_3, … for their semantic relatedness to nearby words

4. Suggest the most closely related found variation to the user

Drawbacks of WordNet

- Coverage and named entities (extra lecture about NEs)
  - George W. Bush, Barack Obama, Queen Beatrix
  - York ("House of York")
- The structure is somewhat arbitrary
  - {water, body of water}
    - {thing}
      - {physical entity}
- Synsets can be difficult to distinguish, e.g. `fear#v`

```
• SYN1 be afraid or feel anxious or apprehensive about a
  possible or probable situation or event
• SYN2 be afraid or scared of; be frightened of
• SYN3 be uneasy or apprehensive about
```

Query translation for cross-lingual IR

Nguyen et al., 2009

- First of all: Wikipedia != WordNet (EuroWordNet)
- But: when relying on Wikipedia concepts and their hyperlinks, the graph structures appear similar
  - Concept: Wikipedia article
  - Edge between concepts: hyperlinks between articles
- Given a query in language A, it is translated to language B using **only** Wikipedia
- Idea: use cross-lingual links for translation

cross-lingual links: same concept in different languages

Query translation for cross-lingual IR

Nguyen et al., 2009

- Why Wikipedia
  - Better coverage of named entities, many domain-specific concepts
  - Community keeps it up-to-date
  - Wikipedia articles provide more context (than glosses)
  - Synonyms are provided "for free" (redirect pages)
  - Disadvantage: coverage of common terms (e.g. "house" has many unusual senses)

- Approach:

$\text{query}_{\text{English}} \Rightarrow \text{WP}_{\text{English}} \Rightarrow \text{WP}_{\text{German}} \Rightarrow \text{query}_{\text{German}}$

Query translation for cross-lingual IR

Nguyen et al., 2009

1. Determine the most relevant concepts to the original query after a keyword-based search with the whole query
2. Search on every single query term using
   - The **internal links** from the concepts retrieved in 1.
   - The **text** and title of the articles retrieveed in 1.
3. Add articles that redirect to the found concepts (synonyms, spelling variants)
4. Create new translated query

not available in dictionaries

$\text{query}_{\text{English}} \Rightarrow \text{WP}_{\text{English}} \Rightarrow \text{WP}_{\text{German}} \Rightarrow \text{query}_{\text{German}}$

Query translation for cross-lingual IR

Nguyen et al., 2009

CLEF 2004

```
<title>Atlantis-Mir Koppeling</title>
<desc>Vind documenten over de eerste space shuttle
aankoppeling tussen de Amerikaanse shuttle Atlantis
en het Mir ruimte station</desc>
```

⇩ search WP with full query

```
WP:space_shuttle_atlantis, WP:mir_(ruimtestation)
```

⇩ search WP with query terms on links and content

```
amerika, atlantis_(disambiguation), koppeling, mir
mir_(disambiguation), roskosmos,atlantis_(ruimteveer),…
```

Query translation for cross-lingual IR

Nguyen et al., 2009

CLEF 2004

translate the concepts by following their cross-lingual links to English

```
America, coupling, russian_federal_space_agency,
shuttle, space_shuttle_atlantis, station, …
```

exploit redirects

```
space_shuttle_atlantis: atlantis_(space_shuttle),
ov-104, ss_atlantis, …
```

weigh concepts found early on higher,
use keyword and phrases in the final query

```
"station"^1.0 station^1.0 "russian federal space agency"^1.0
……space^3.0 shuttle^3.0 atlantis^3.0…"mir"^3.0 mir^3.0
```

Query translation for cross-lingual IR

Nguyen et al., 2009

- Evaluation: retrieval of English documents using Dutch, French and Spanish queries
  - Optimum: same effectiveness as the monolingual run

Language	MAP
English (monolingual)	0.3407
French	0.2278 (66.9%)
Spanish	0.2181 (64.0%)
Dutch	0.2038 (59.8%)

Query translation for cross-lingual IR

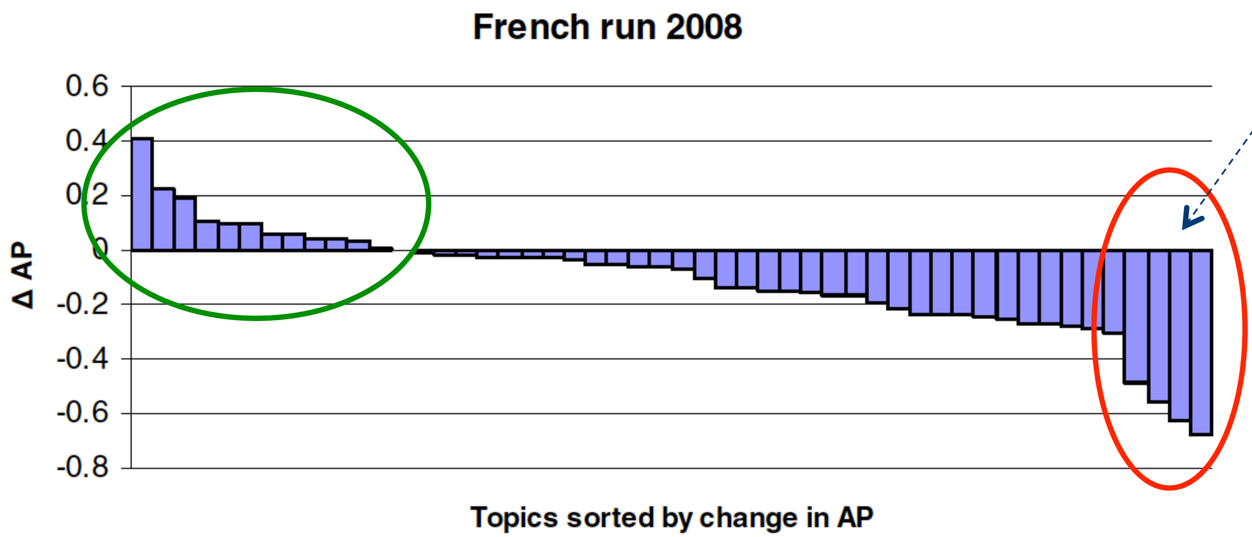
Nguyen et al., 2009

- Evaluation: retrieval of English documents using Dutch, French and Spanish queries

When does it fail?

"fictives" (fr.) means "fictional" (engl.)

- Planets_in_science_fiction
- Fictional_brands



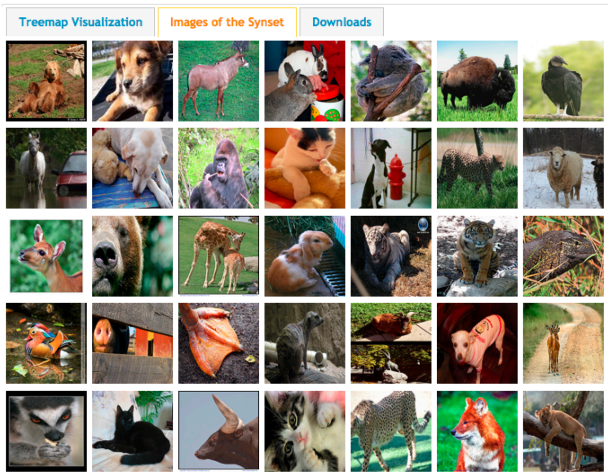
Alternatives to WordNet or Wikipedia

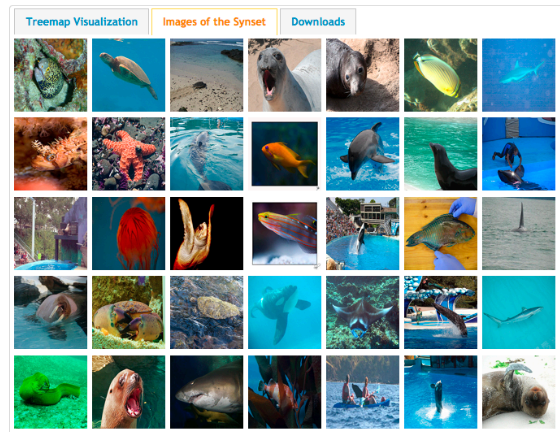
- Yago (Suchanek et al., 2007)
  - Combines WordNet and Wikipedia
  - Contains 2 million entities, “knows” 20 million facts
  - Manually confirmed accuracy of 95%
  - Queries have the form <entity><relation><entity>
- WordNet Affect
  - Manual labeling of 3000 synsets with affective concepts
  - 11 affective labels, e.g.
    - Emotion (`anger#n#1, fear#v#1`)
    - Cognitive state (`confusion#n#2, dazed#adj#2`)
    - Emotional response (`cold_sweat#n#1, tremble#v#1`)

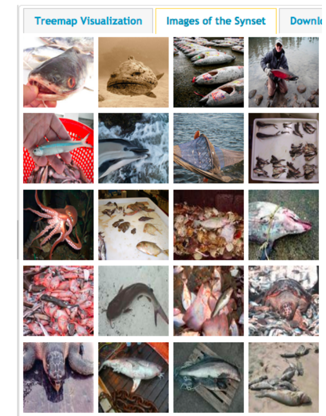
ImageNet

Deng et al., 2009

- Image retrieval (visual codewords)
- Text-to-image synthesis







animal, animate being, beast, brute, creature, fauna

Marine animal, marine creature, sea animal, sea creature

by-catch, bycatch

IS-A

IS-A

- cross-modal semantic relationships
- decrease semantic gap (MM lecture)

http://www.image-net.org/

Latent Semantic Analysis

Deerwester & Dumais, 1990

- Find the documents' hidden meaning (obscured by noise)
- Distributional hypothesis: terms with similar/related meanings tend to co-occur **with the same other words**

The **Netherlands** is a constituent country of the Kingdom of the Netherlands, located mainly in North-West Europe and with several islands in the Caribbean. Mainland Netherlands borders the North Sea to the north and west, Belgium to the south, and Germany to the east, and shares maritime borders with Belgium, Germany and the United Kingdom. It is a parliamentary democracy organised as a unitary state. The country capital is Amsterdam and the seat of government is The Hague.

The main cities in **Holland** are Amsterdam, Rotterdam and The Hague. Amsterdam is formally the capital of the Netherlands and its largest city. The Port of Rotterdam is Europe's largest and most important harbour and port. The Hague is the seat of government of the Netherlands. These cities, combined with Utrecht and other smaller municipalities, effectively form a single city—a conurbation called Randstad.

Latent Semantic Analysis

Deerwester & Dumais, 1990

- Idea: transformation of the term space into a concept space that captures most of the variance in the collection
  - Documents and terms now indirectly related through concepts
  - The originally sparse space is now dense (dimensionality reduction)

terms $\Leftrightarrow$ concepts $\Leftrightarrow$ documents

  - Documents sharing often co-occurring terms are close to each other in concept space

$$D_1 = \{Holland\},\ D_2 = \{Netherlands\}$$

    - Distant in term space, close in concept space

LSA & Singular value decomposition

Deerwester & Dumais, 1990

- Factorization of term document matrix leads to concept space
- SVD to find correlations among the rows and columns
    - Every real matrix has a SVD decomposition

$$T = U \Sigma V^T$$

$$\begin{array}{cc} & \text{document} \\ \text{term} & \begin{pmatrix} tf_{1,1} & tf_{1,2} & \dots & tf_{1,n} \\ tf_{2,1} & tf_{2,2} & \dots & tf_{2,n} \\ \dots & \dots & \dots & \dots \\ tf_{m,1} & tf_{m,2} & \dots & tf_{m,n} \end{pmatrix}_{m \times n} \end{array} = \left(\begin{bmatrix} \\ u_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ u_m \\ \\ \end{bmatrix} \right)_{m \times m} \times \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m \end{pmatrix}_{m \times n} \times \begin{pmatrix} [\quad v_1 \quad] \\ \vdots \\ [\quad v_n \quad] \end{pmatrix}_{n \times n}$$

singular values

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$$

LSA & Singular value decomposition

Deerwester & Dumais, 1990

- Perfect reconstruction of the term-document matrix (sparse) is possible

$$U \Sigma V^T = T$$

- Denser and smaller matrix if only the first k singular values are used (dimensionality reduction)

$$U \Sigma_k V^T \approx T \, , with \; k \ll m$$

- Benefits of SVD
  - Transformation of correlated variables into uncorrelated ones
  - Ordering of dimensions according to the data's greatest variability
  - Best approximation of original data with fewer dimensions

LSA & Singular value decomposition

Deerwester & Dumais, 1990

- Keep the k most important concepts (often [100,300])
  - Best approximation of the term-document matrix in the least square sense
  - Each document is explained by concept vectors
- For retrieval purposes, the (keyword) query needs to be translated into the concept space: $q' = q^T U_k$
  - Retrieval as in the vector space model
- Good retrieval performance for small corpora
- Concept vectors not always easily explainable

Latent Semantic Analysis

Example concepts

celtic	game	serbs
club	players	serb
race	season	un
handicap	cup	bosnia
rangers	club	goal
football	match	try
trainer	team	penalty
clubs	league	theatre

highest scoring terms of three LSI concept vectors; corpus: GH95 (GeoCLEF)

PLSA: probabilistic LSA

Hofman, 1999 [4]

- Co-occurrence statistics modelled through latent class variables ("topics") in a statistical model
- Each word is generated from a single topic; different words in a document may be generated from different topics

PLSA: probabilistic LSA

Hofman, 1999 [4]

- Co-occurrence statistics modelled through latent class variables ("topics") in a statistical model
- Each word is ge[r
a document ma[y

Brazil has become the sixth-biggest economy in the world, overtaking the UK, the country's finance minister says.
Carmaker Nissan is to build a new model at its Sunderland factory, with investment of £125m, which it says will create 2,000 jobs.
Arsene Wenger's Arsenal will give it "a real go" as they try to overturn a 4-0 Champions League deficit against AC Milan.

Sports **Economy** **Cars**

PLSA: probabilistic LSA

Hofman, 1999 [4]

- Co-occurrence statistics modelled through latent class variables ("topics") in a statistical model
- Each word is generated from a single topic; different words in a document may be generated from different topics

- Generative process of observing word w in document D
  1. Select document D with probability $P(D)$
  2. Pick latent class z from $\{z_1, \dots z_k\}$ with probability $P(z|D)$
  3. Generate a word w with "factor" probability $P(w|z)$

$$P(w,D) = P(D) \underbrace{\sum_{z \in Z} P(w \mid z) P(z \mid D)}_{P(w|D)}$$

PLSA: probabilistic LSA

Factor examples [4]

4 factors from a 128 factor decomposition (TDT-1 corpus)
8 most likely terms per factor (i.e. *P(w|z)*)

plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	austronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
"plane"	"space shuttle"	"family"	"hollywood"

flight

love

Summary

- Semantics
  - Important
  - And cool ☺
- Two camps
  - Knowledge-based
  - Statistics-based
- More on Monday!

Sources

① Introduction to Information Retrieval. Manning et al. 2008
② Information retrieval. Keith van Rijsbergen. 1979
③ Managing gigabytes, Witten et al. 1999
④ Unsupervised learning by probabilistic latent semantic analysis. Hofmann. 2001
⑤ Cumulated gain-based evaluation of IR techniques. Järvelin & Kekäläinen. 2002
⑥ Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. Budanitsky & Hirst. 2001
⑦ Correcting real-world spelling errors by restoring lexical cohesion. Hirst & Budanitsky. 2005