Mining Meaning From Wikipedia

PD Dr. Günter Neumann

LT-lab, DFKI, Saarbrücken

Outline

1. Introduction
2. Wikipedia
3. Solving NLP tasks
4. Namend Entity Disambiguation
5. Information Extraction
6. Ontology Building and the Semantic Web

1. Introduction

- Meaning:
  - Concepts, topics, fact descriptions, semantic relations, ways of organizing information
- Mining
  - Gathering meaning into machine-readable structures (e.g., ontologies)
  - Using meaning in areas like IR and NLP
- Wikipedia:
  - The largest and most widely-used encyclopedia in existence
  - Partially validated, trusted, multilingual, multimedia text data

Traditional approaches to Mining Meaning

- Carefully hand-crafted rules
  - High quality, but restricted in size and coverage
  - Needs input of experts, however very expensive to keep with developments
  - e.g., Cyc ontology
    - Hundreds of conbtributors and 20 years of development
    - Still limited size and patchy coverage

Traditional approaches to Mining Meaning

- Statistical inference
  - Scarifice quality and go for quantity by performing large-scale analysis of unstructured text
  - Might be applicable for specific domain and text data/corpora
  - Problems in generalization or moving into new domains and tasks

2. Wikipedia: a middle ground

- Combines quality and quantity through mix of scale and structure
  - 2 millions of articles and 1000 of contributors
  - 18 GB of text
  - extensive network of links, categories, infoboxes provide explicitly defined (shallow) semantics
- Note:
  - Restricted trust & credibility compared to traditional rule-based approaches, because contributors are largely unknown and un-experts
  - Only represents a small snapshot of human language use in the web!

Wikipedia: A resource for mining meaning

- Wikipedia offers a unique, entirely open, collaborative editing process
  - Approx. 250 languages are covered
  - „Emerging semantics" through collaborative „use of language" (cf. Wittgenstein)
- Self-organizing system, but controlled
  - To avoid „edit wars", sophisticated Wikipedia policies (must be followed) and guidelines (should be followed) are established

Wikipedia: A resource for mining meaning

- Implications for mining
  - Constantly growing and changing data
  - How to evaluate systems that use Wikipedia ? How to determine „ground truth"?
- Most researchers use Wikipedia as a „product"
  - Data basis for extracting information/meaning
- In principle also possible: consider Wikipedia as a „process"
  - Infrastructure allows „reasoning" about „how something has been written", e.g., mining of versions/authors, discussions etc.
  - Cross-lingual analysis for cultural/socio data mining ?

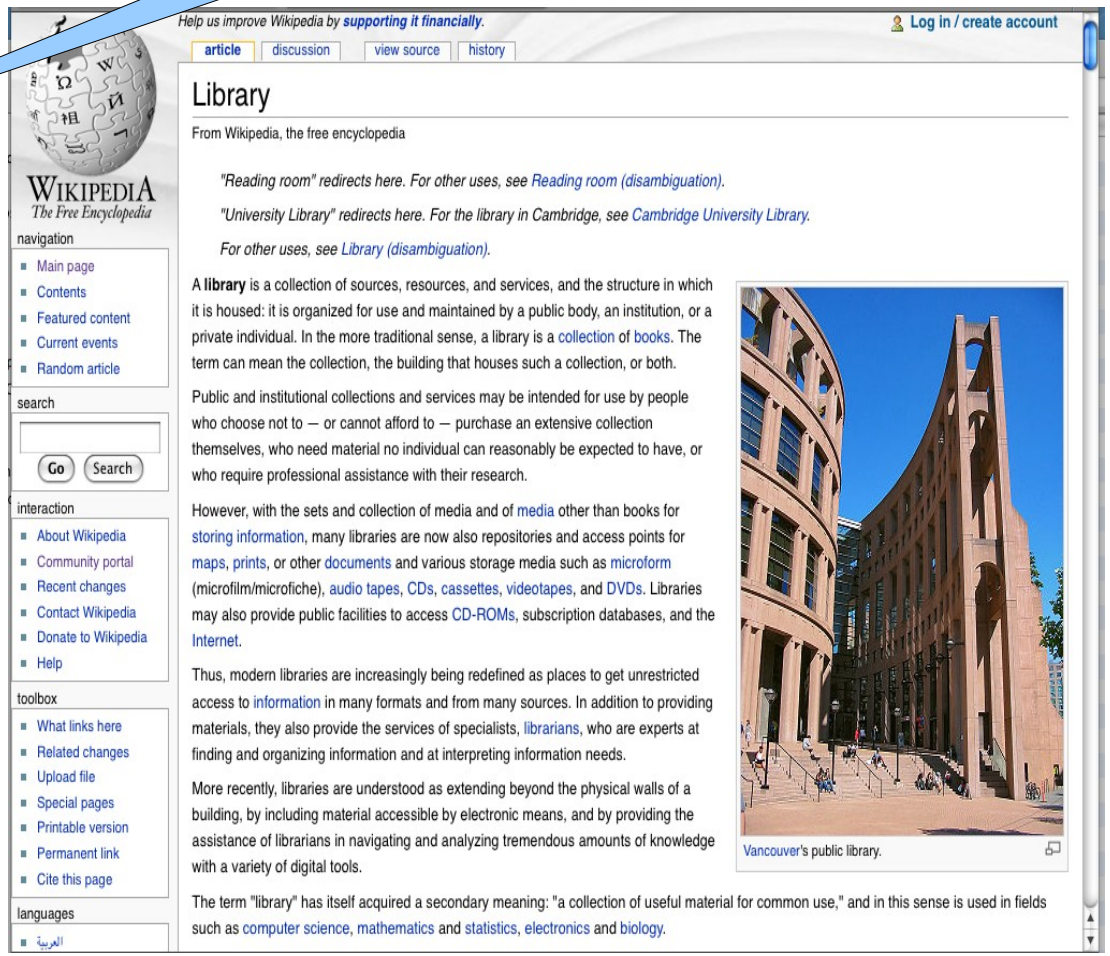
Wikipedia's structure

- Articles
- Redirects
- Disambiguation pages
- Hyperlinks
- Category structure
- Templates/Infoboxes
- Discussion pages
- Edit histories

Wikipedia article

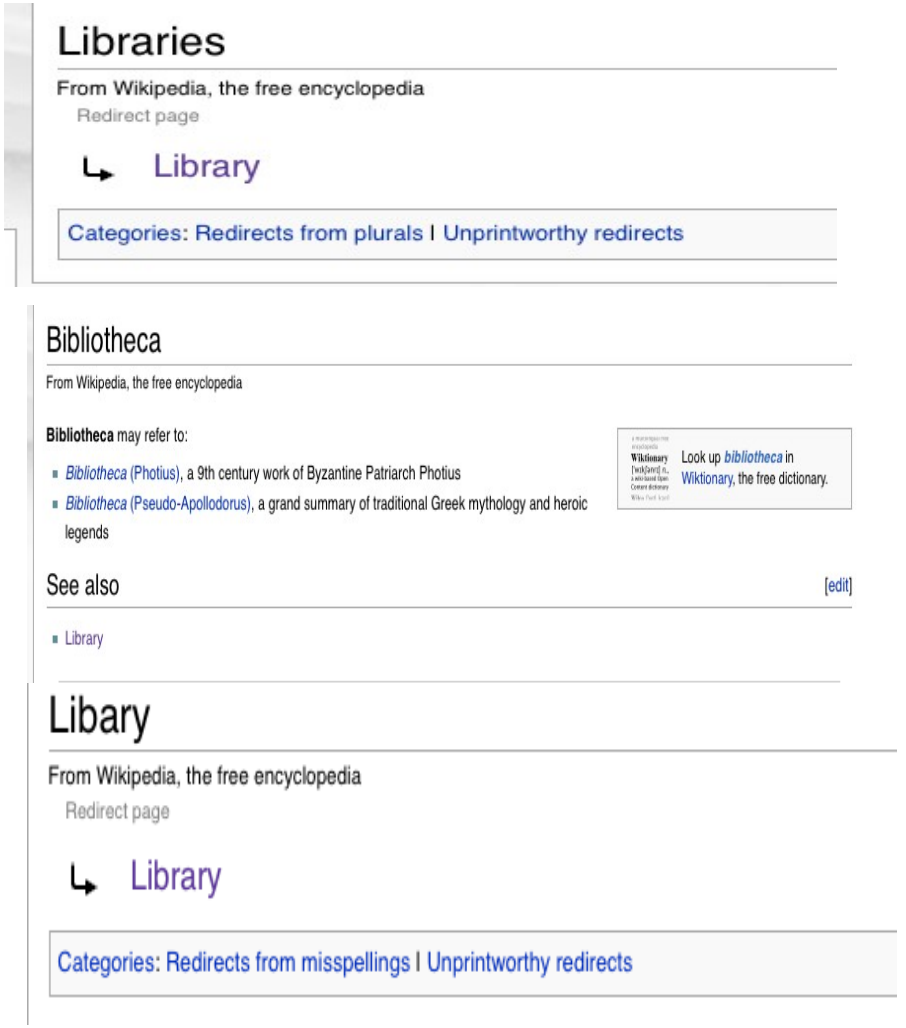
Optic nerve (the nerve)
vs.
Optic Nerve (the comic book)

- Article = Concept
- Title resembles term in thesaurus (capitalization might be important)
- Articles begin with a brief overview of the topic
- First sentence defines the entity and its type
- Scale:
  - ~10M articles in 250 languages
  - e.g., 2M English, 0.8M German



Wikipedia redirects

- A page with just text in form of a directive
- Goal:
  - Have a single article for equivalent terms
- ~3M in English Wikipedia
- Usable for resolving synonyms, since an external thesaurus is not necessary



Libraries

From Wikipedia, the free encyclopedia

Redirect page

↳ Library

Categories: Redirects from plurals | Unprintworthy redirects

Bibliotheca

From Wikipedia, the free encyclopedia

Bibliotheca may refer to:

- *Bibliotheca* (Photius), a 9th century work of Byzantine Patriarch Photius
- *Bibliotheca* (Pseudo-Apollodorus), a grand summary of traditional Greek mythology and heroic legends

Look up *bibliotheca* in Wiktionary, the free dictionary.

See also [edit]

- Library

Libary

From Wikipedia, the free encyclopedia

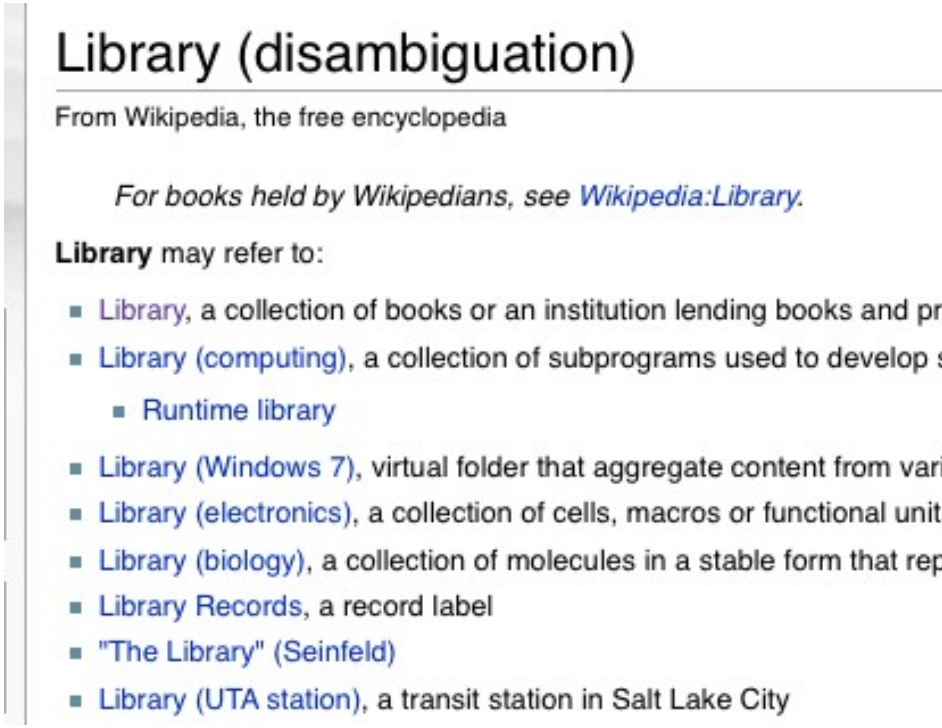
Redirect page

↳ Library

Categories: Redirects from misspellings | Unprintworthy redirects

Wikipedia disambiguation page

- A page with possible meanings (i.e., articles) of a term
- Snippets as brief descriptions of a term (article)
- English Wiki as 0.1M disamig. Pages
- Usable for processing homonyms



Library (disambiguation)

From Wikipedia, the free encyclopedia

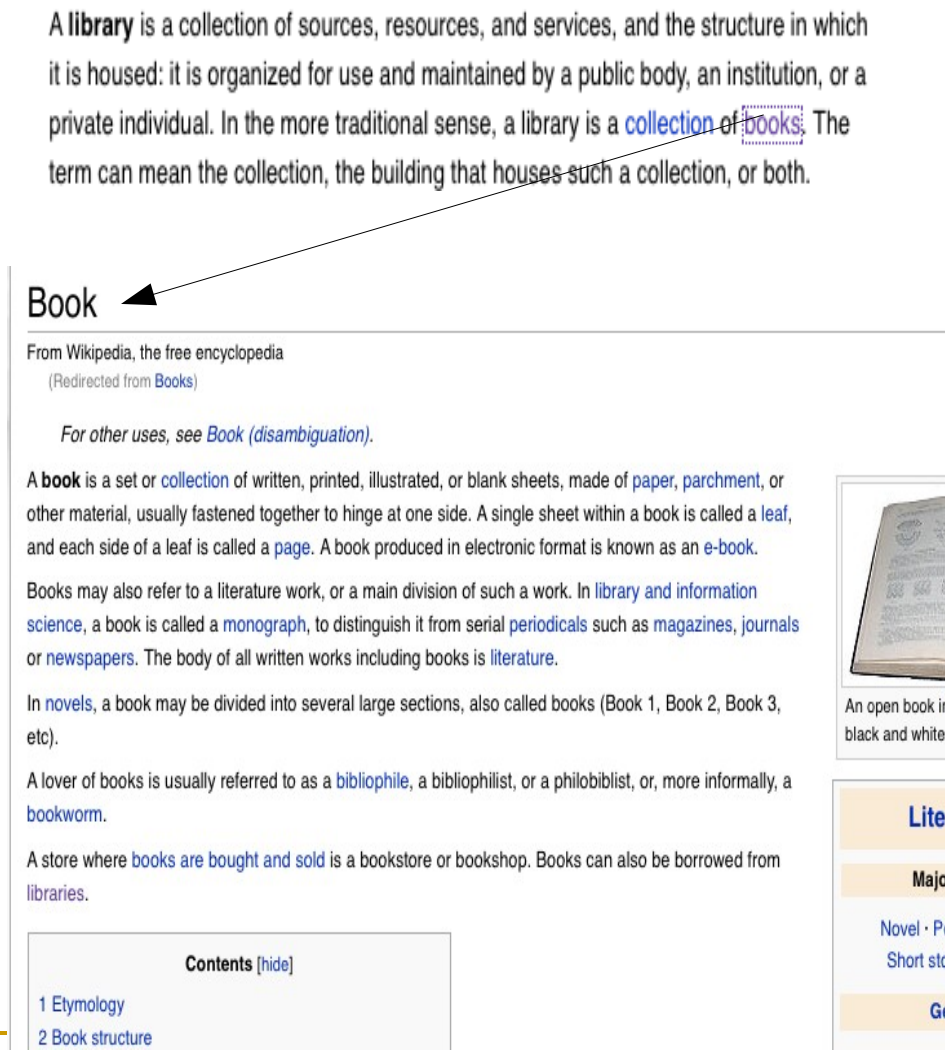
For books held by Wikipedians, see Wikipedia:Library.

Library may refer to:

- Library, a collection of books or an institution lending books and pr
- Library (computing), a collection of subprograms used to develop
  - Runtime library
- Library (Windows 7), virtual folder that aggregate content from vari
- Library (electronics), a collection of cells, macros or functional unit
- Library (biology), a collection of molecules in a stable form that rep
- Library Records, a record label
- "The Library" (Seinfeld)
- Library (UTA station), a transit station in Salt Lake City

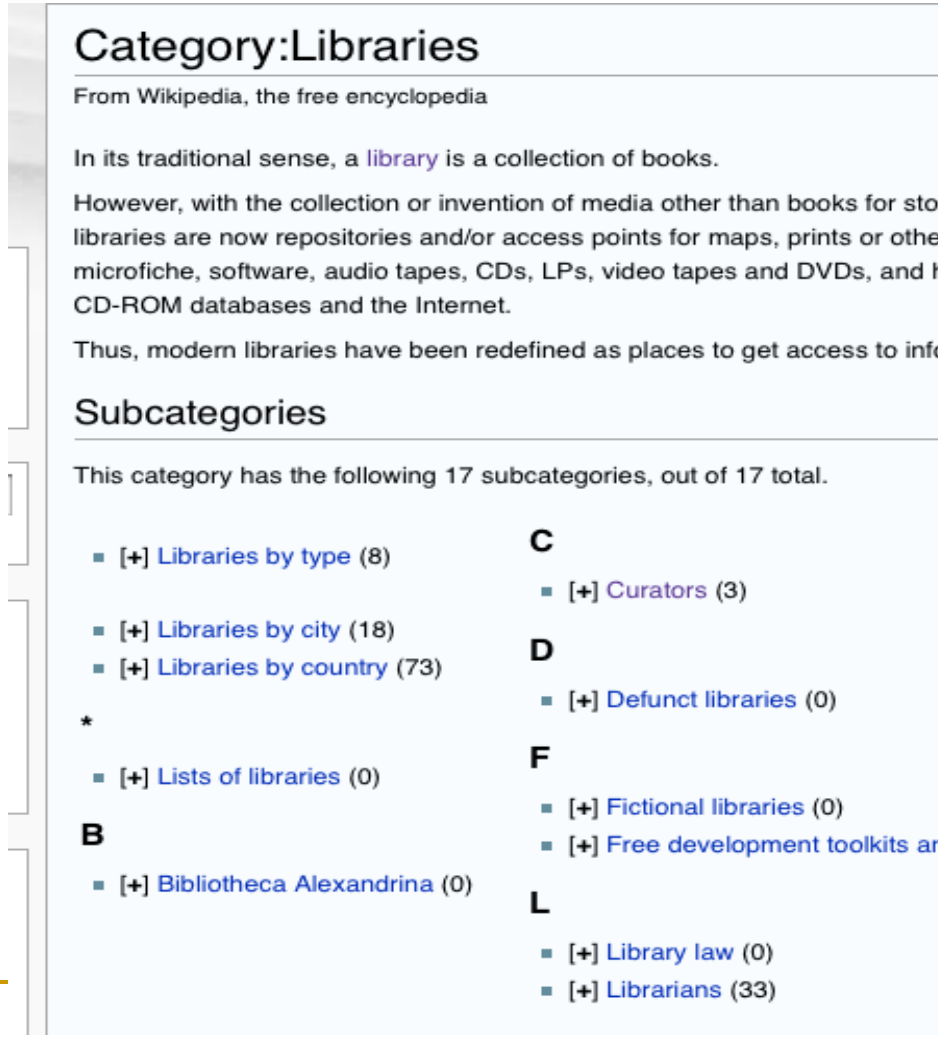
Wikipedia hyperlinks

- Hyperlink are links from articles to other articles
- ~60M links in English Wikipedia
- Usable for
  - Lexical semantics
  - Associative relationship
  - Density/Ranking



Wikipedia categories

- Merely nodes for organizing articles with minimum of explanatory text
- Goal:
  - Represent information hierarchy
  - Overall structure is a DAG
- Status
  - Still in development, no clean definition,
  - Most links are ISA, others represent more different types, e.g., meta categories for editorial purposes



Category:Libraries

From Wikipedia, the free encyclopedia

In its traditional sense, a library is a collection of books.

However, with the collection or invention of media other than books for sto
libraries are now repositories and/or access points for maps, prints or othe
microfiche, software, audio tapes, CDs, LPs, video tapes and DVDs, and h
CD-ROM databases and the Internet.

Thus, modern libraries have been redefined as places to get access to info

Subcategories

This category has the following 17 subcategories, out of 17 total.

- [+] Libraries by type (8)
- [+] Libraries by city (18)
- [+] Libraries by country (73)

*

- [+] Lists of libraries (0)

B

- [+] Bibliotheca Alexandrina (0)

C

- [+] Curators (3)

D

- [+] Defunct libraries (0)

F

- [+] Fictional libraries (0)
- [+] Free development toolkits ar

L

- [+] Library law (0)
- [+] Librarians (33)

Wikipedia templates

- Templates often look like text boxes with a different background color from that of normal text.
- They are in the template namespace, i.e. they are defined in pages with "Template:" in front of the name.
- They are like text patterns to add information


Library

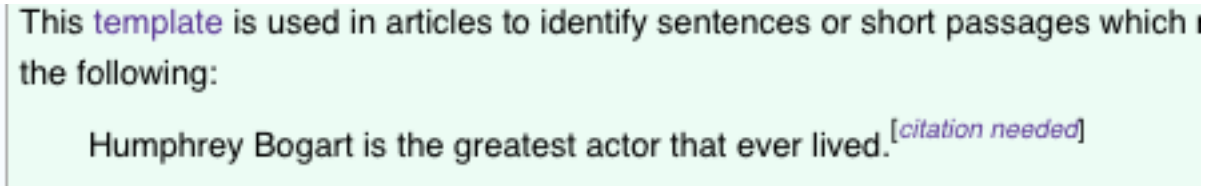
From Wikipedia, the free encyclopedia

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

"University Library" redirects here. For the library in Cambridge, see Cambridge University Library.

For other uses, see Library (disambiguation).

A **library** is a collection of sources, resources, and services, and the structure in which it is housed: it is organized for use and maintained by a public body, an institution, or a


This template is used in articles to identify sentences or short passages which
the following:

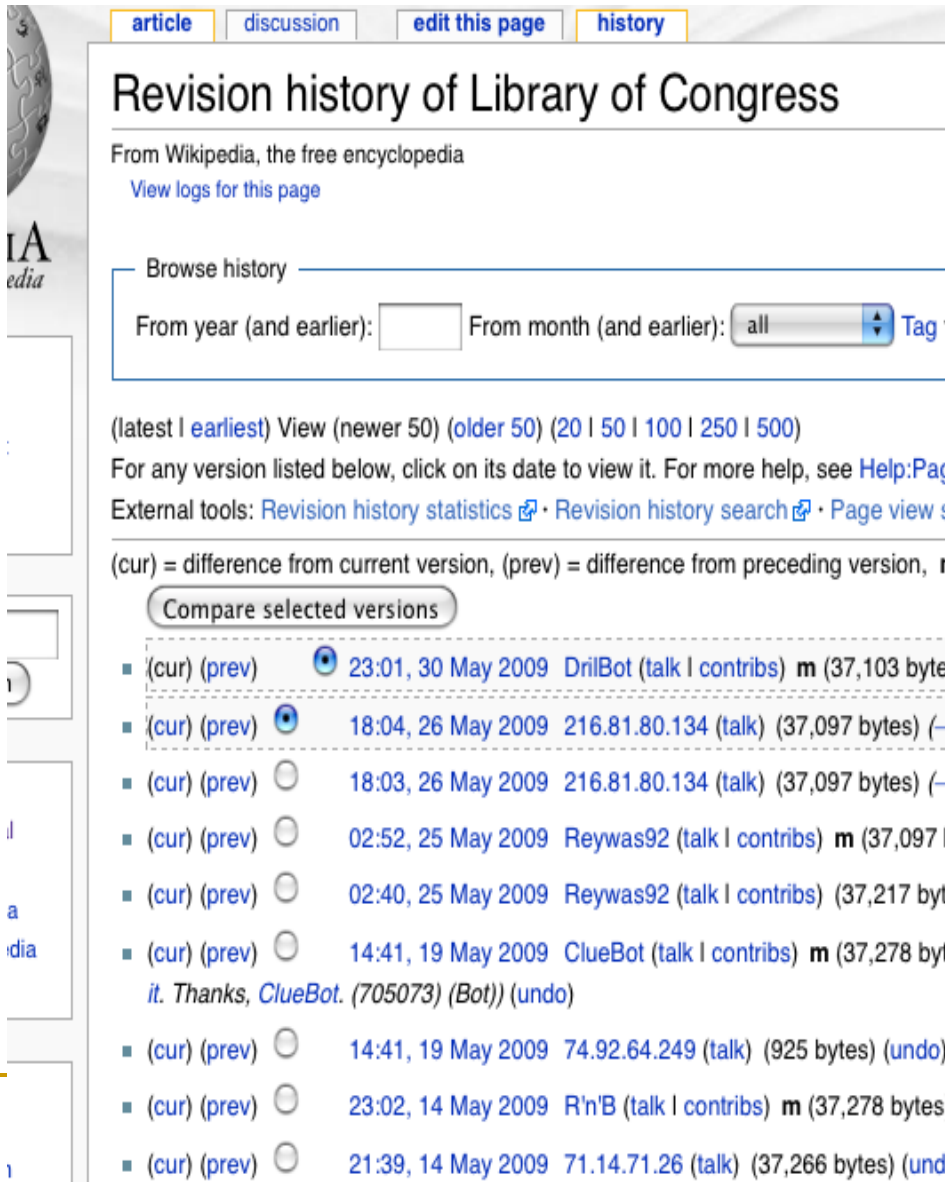
Humphrey Bogart is the greatest actor that ever lived.[*citation needed*]

Wikipedia infoboxes

- An infobox is a special type of template that displays factual information in a structured uniform way.
- ~8000 different infobox templates
- Still not standardized, e.g., names/values of attributes.
- Ako semi-structured IE templates

Wikipedia discussion & edit histories

- Each article has an associated **talk page** representing a forum for discussion as to how it might be critized, improved or extended
- Contains edit development & corresponding author (alias)
- Both Wikipedia structures are not much used in data mining so far.



Perspectives on Wikipedia

- Wikipedia as an encyclopedia
- Wikipedia as a large corpus
  - Large text sources, well-written, well-formulated
  - Partially annotated through tags
  - Partial multilingual alignment
- Wikipedia as a thesaurus
  - Compare and augment with traditional thesauri
  - extract/compute crosslingual thesauri

Perspectives on Wikipedia

- Wikipedia as a database
  - Massive amount of highly structured information
  - Several projects try to make it available, e.g. DBPedia
- Wikipedia as an ontology
  - Articles can be considered as conceptual elements
  - explicit/implicit lexical semantics relationships
- Wikipedia as a network structure
  - The hyperlinked structures make Wikipedia a microcosmos of the Web
  - Development of new ranking algorithm, e.g., to find related articles or cluster articles under different criteria
  - Apply WordNet similarity measures to Wikipedia's category graph

3. Solving NLP tasks

- Two major groups
  - symbolic methods, where system utilizes a manually encoded repository of human language
    - Low coverage, e.g., WordNet
  - Statistical methods, which infer properties of language by processing large text corpora
    - Upper performance bounds probably only can improve when symbolic knowledge is integrated (hybrid approaches)

Four NLP problems in which Wikipedia has been used

- Semantic relatedness
- Word sense disambiguation
- Co-reference resolution
- Multilingual alignment

Four NLP problems in which Wikipedia has been used

- Semantic relatedness
- Word sense disambiguation
- Co-reference resolution
- Multilingual alignment

Semantic Relatedness

- Semantic relatedness determines how much two concepts (e.g., doctor & hospital) are related by using all relations between them, e.g., is-a, has-part, is-made-of, …
  - Only if is-a then we call it semantic similarity
- Usually, relatedness is computed using
  - predefined taxonomies (e.g., is-a) and other relations, e.g., has-part, is-made-of
  - Statistical methods to analyze term co-occurrence in large corpora

Evaluation

- Standard corpora
  - M&C: a list of 30 noun pairs, cf. Miller & Charles, 1991
  - R&G: 65 synonymous word pairs, cf. Rubenstein & Goodenough, 1965
  - WS-353: a list of 353 word pairs, cf. Finkelstein et al. 2002
    - http://alfonseca.org/eng/research/wordsim353.html
- Best pre-Wikipedia result
  - 0.86 correlation for M&C by Jiang & Conrath, 1997
    - based on human similarity judgment
    - A mixed statistical approach + WordNet
  - 0.56 for WS-353 by Finkelstein using LSA

Wikipedia based Semantic Relatedness

- Strube & Ponzetto, AAAI-2006
  - WikiRelate!
- Gabrilovic & Markovitch, IJCAI-2007
  - Explicit Semantic Analysis (ESA)
- Milne, 2007
  - Use of internal linkstructure of Wikipedia articles

Approach 1: WikiRelate!

- Re-calculation of different measures developed for WordNet using Wikipedia's category structure
- Best performing measure: normalized path measure, cf. Leacock & Chodorow, 1998:
  - $lch(c_1,c_2) = -log(length(c_1,c_2)/2D))$
  - $length(c_1,c_2)$: shortest path, D: max. depth of taxonomy
- Result:
  - WordNet-based measures still better on M&C and R&G
  - Wikipedia-based measures are better on WS-353 (0.62)
    - Why ? WordNet is too fine-grained and sometimes do not match the user's intuition (cf. Jaguar vs Stock)

Approach 2: Explicit Semantic Analysis

- Idea: use centroid-based classifier to map input text to a vector of weighted Wikipedia articles
  - Bank of Amazon → vector(Amazon River, Amazon Basin, Amazon Rainforest, Amazon.com, Rainforest, Atlantic Ocean, Brazil, ...)
- Relatedness(c_1, c_2)
  - cosinus(a_1, a_2), where a_i is article of concept c_i
- Result:
  - WS-353: ESA=0.75, LSA=0.56
  - Open-Directory-Project = 0.65 → Wikipedia'quality is greater

ESA: More details

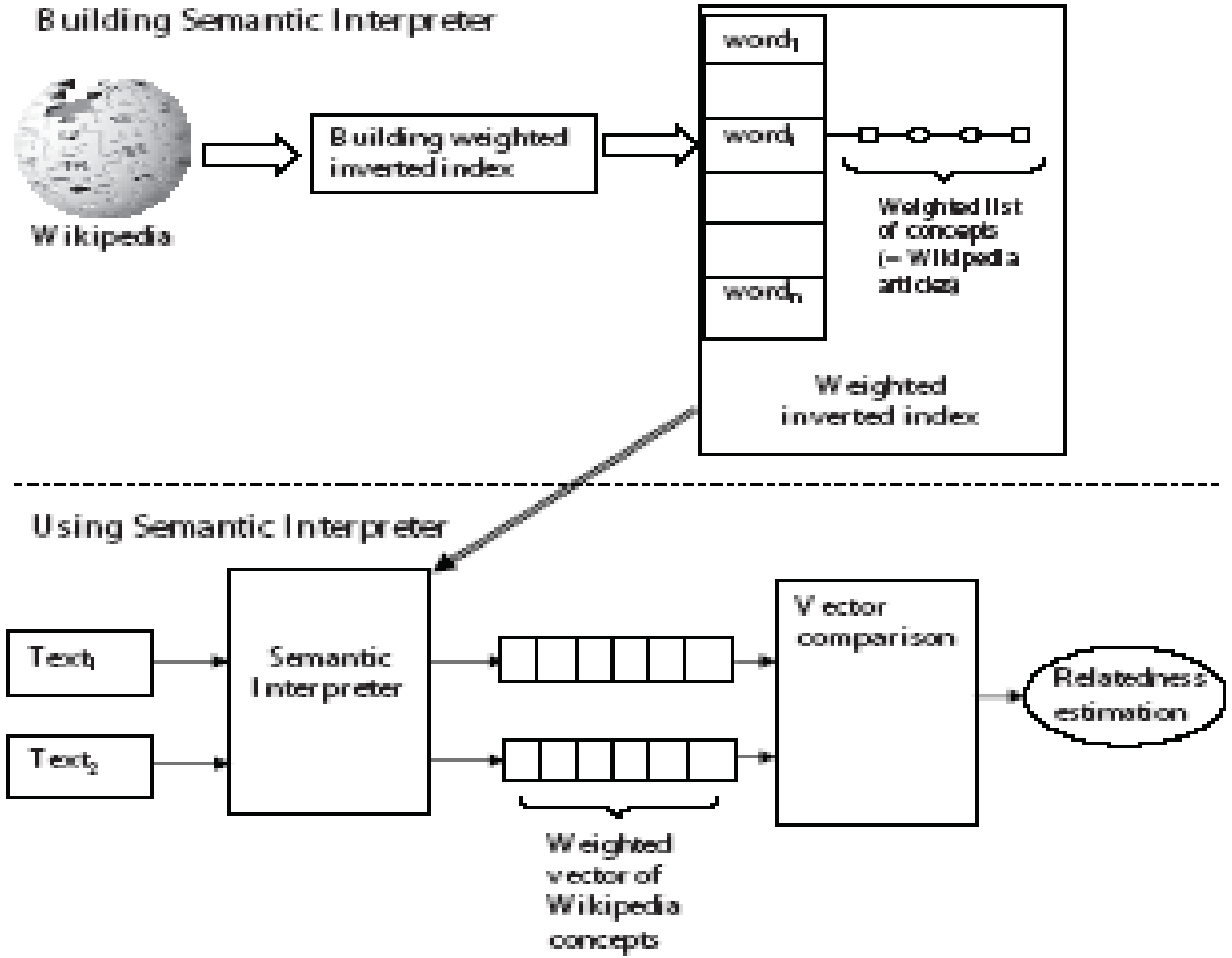
- T = {w1…wn} be input text
- <vi> be T's TFIDF vector
  - vi is the weight of word wi
- Wikipedia concept cj , {cj ∈ c1, . . . , cN}
  - N = total number of Wikipedia concepts
- Let <kj> be an inverted index entry for word wi
  - where kj quantifies the strength of association of word wi with Wikipedia concept cj

Explicit Semantic Analysis

- the semantic interpretation vector V for text T is a vector of length N, in which the weight of each concept cj is defined as

$$\sum_{w_i \in T} v_i \cdot k_j$$

- To compute semantic relatedness of a pair of text fragments we compare their vectors using the cosine metric


Building Semantic Interpreter
Wikipedia
Building weighted inverted index
word1
wordi
wordn
Weighted list of concepts (= Wikipedia articles)
Weighted inverted index
Using Semantic Interpreter
Text1
Text2
Semantic Interpreter
Weighted vector of Wikipedia concepts
Vector comparison
Relatedness estimation

Example: small text input

#	Input: *"equipment"*	Input: *"investor"*
1	Tool	Investment
2	Digital Equipment Corporation	Angel investor
3	Military technology and equipment	Stock trader
4	Camping	Mutual fund
5	Engineering vehicle	Margin (finance)
6	Weapon	Modern portfolio theory
7	Original equipment manufacturer	Equity investment
8	French Army	Exchange-traded fund
9	Electronic test equipment	Hedge fund
10	Distance Measuring Equipment	Ponzi scheme

First ten concepts in sample interpretation vectors

Example: large text input

#	Input: *"U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials."*	Input: *"The development of T-cell leukaemia following the otherwise successful treatment of three patients with X-linked severe combined immune deficiency (X-SCID) in gene-therapy trials using haematopoietic stem cells has led to a re-evaluation of this approach. Using a mouse model for gene therapy of X-SCID, we find that the corrective therapeutic gene IL2RG itself can act as a contributor to the genesis of T-cell lymphomas, with one-third of animals being affected. Gene-therapy trials for X-SCID, which have been based on the assumption that IL2RG is minimally oncogenic, may therefore pose some risk to patients."*
1	Iraq disarmament crisis	Leukemia
2	Yellowcake forgery	Severe combined immunodeficiency
3	Senate Report of Pre-war Intelligence on Iraq	Cancer
4	Iraq and weapons of mass destruction	Non-Hodgkin lymphoma
5	Iraq Survey Group	AIDS
6	September Dossier	ICD-10 Chapter II: Neoplasms; Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism
7	Iraq War	Bone marrow transplant
8	Scott Ritter	Immunosuppressive drug
9	Iraq War- Rationale	Acute lymphoblastic leukemia
10	Operation Desert Fox	Multiple sclerosis

First ten concepts in sample interpretation vectors

Example (texts with ambiguous words)

#	Ambiguous word: "Bank"		Ambiguous word: "Jaguar"	
	"Bank of America"	*"Bank of Amazon"*	*"Jaguar car models"*	*"Jaguar (Panthera onca)"*
1	Bank	Amazon River	Jaguar (car)	Jaguar
2	Bank of America	Amazon Basin	Jaguar S-Type	Felidae
3	Bank of America Plaza (Atlanta)	Amazon Rainforest	Jaguar X-type	Black panther
4	Bank of America Plaza (Dallas)	Amazon.com	Jaguar E-Type	Leopard
5	MBNA	Rainforest	Jaguar XJ	Puma
6	VISA (credit card)	Atlantic Ocean	Daimler	Tiger
7	Bank of America Tower, New York City	Brazil	British Leyland Motor Corporation	Panthera hybrid
8	NASDAQ	Loreto Region	Luxury vehicles	Cave lion
9	MasterCard	River	V8 engine	American lion
10	Bank of America Corporate Center	Economy of Brazil	Jaguar Racing	Kinkajou

First ten concepts in sample interpretation vectors

Empirical Evaluation

- Wikipedia
  - parsing the Wikipedia XML dump, we obtained 2.9 Gb of text in 1,187,839 articles
  - removing small and overly specific concepts (those having fewer than 100 words and fewer than 5 incoming or outgoing links), 241393 articles were left
  - 389,202 distinct terms

Empirical Evaluation

- Open Directory Project
  - hierarchy of over 400,000 concepts and 2,800,000 URLs.
  - crawling all of its URLs, and taking the first 10 pages encountered at each site
  - 70 Gb textual data. After removing stop words and rare words, we obtained 20,700,000 distinct terms

Datasets and Evaluation Procedure

- The WordSimilarity-353 (WS-353) collection
  - contains 353 word pairs. Each pair has 13–16 human judgements
  - Spearman rank-order correlation coefficient was used to compare computed relatedness scores with human judgements
  - Spearman rank-order correlation (http://webclass.ncu.edu.tw/~tang0/Chap8/sas8.htm)

Datasets and Evaluation Procedure

- 50 documents from the Australian Broadcasting Corporation's (ABC) news mail service [Lee et al., 2005]
  - These documents were paired in all possible ways, and each of the 1,225 pairs has 8–12 human judgements
  - When human judgements have been averaged for each pair, the collection of 1,225 relatedness scores have only 67 distinct values.
  - Spearman correlation is not appropriate in this case, and therefore we used Pearson's linear correlation coefficient
  - http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Results for ESA

- word relatedness (WS-353)

Algorithm	Correlation with humans
WordNet [Jarmasz, 2003]	0.33–0.35
Roget's Thesaurus [Jarmasz, 2003]	0.55
LSA [Finkelstein *et al.*, 2002]	0.56
WikiRelate! [Strube and Ponzetto, 2006]	0.19 – 0.48
ESA-Wikipedia	0.75
ESA-ODP	0.65

- text relatedness (ABC)

Algorithm	Correlation with humans
Bag of words [Lee *et al.*, 2005]	0.1–0.5
LSA [Lee *et al.*, 2005]	0.60
ESA-Wikipedia	0.72
ESA-ODP	0.69

Approach 3: Wikipedia hyperlinks

- Milne, 2007, only uses articles' internal links structure
- Relatedness of two terms:
  - Determine articles
  - Create vector from the links inside the articles that point to other articles
  - Each link is weighted by the inverse number of times it is linked from other Wikipedia articles
    - The less common the link, the higher its weight.
- Example:
  - Bank of America is the largest commercial <bank> in the <United States> by both <deposits> and <market capitalization>
  - 4 links
  - <market capitalization> gets higher weight than <United States>, and hence has semantic relatedness with <Bank of America>

Results for Wikipedia link structure

- Results on WS-353:
  - Manual disambiguation: 0.72
  - Automatic disambiguation (max. similarity): 0.45
- Milne & Witten (2008) improved disambiguation:
  - Conditional probability of the sense given the term
    - „Leopard" most often links to animal article than to Mac OS article
    - Normalized Google distance of term, cf. Cilibrasi & Vitanys's 2002 instead of cosinus-measure
    - Degree of collocation of two terms in Wikipedia
  - Summing over these 3 parameters, they obtain 0.69 on WS-353
    - But approach is less complex than approach of Gabrilovich & Markovitch

Summary of Results

Method	M&C	R&G	WS-353
WordNet [Strube and Ponzetto, 2006]	0.82	0.86	full: 0.36 test: 0.38
WikiRelate! [Ponzetto and Strube, 2007]	0.49	0.55	full: 0.49 test: 0.62
ESA [Gabrilovich and Markovitch, 2007]	0.73	0.82	0.75
WLVM [Milne, 2007]	n/a	n/a	man: 0.72 auto: 0.45
WLM [Milne and Witten, 2008]	0.70	0.64	0.69

Table 2. Overview of semantic relatedness methods.

Four NLP problems in which Wikipedia has been used

- Semantic relateness
- Word sense disambiguation
- Co-reference resolution
- Multilingual alignment

Word Sense Disambiguation

- Goal: resolving polysemy
  - A polyseme is a word or phrase with multiple, related meanings.
  - A word is judged to be polysemous if it has two senses of the word whose meanings are related.
- Standard technology
  - Dictionary or thesaurus that defines the inventory of possible senses
- Wikipedia as an alternative resource
  - Each article describes a concept, i.e., a possible sense for words and phrases that denote it

Example: Wood

- A piece of a tree or a geographical area with many trees

He could see wood around the house.

Figure 3. What is the meaning of *wood* in both examples?

Main Idea behind Word Sense Disambiguation

- Identify the context and analyze which of the possible senses fit it best.
- The following cases will be considered
  - Disambiguating phrases in running text
  - Disambiguating named entities
  - Disambiguating thesaurus & ontology terms

Disambiguating phrases in running text

- Goal: discover the intended senses of words and phrases
- WordNet: a popular resource, but
  - Linguistic (disambiguation) techniques must be essentially perfect to help
  - WordNet defines word senses very fine-grained making it difficult to differentiate them
- Wikipedia:
  - Defines only those senses on which its contributors reach consensus
  - Include an extensive description of each rather than WordNet's brief gloss.

Wikification, Mihalcea & Csomai, 2007

- Use Wikipedia's content as a sense inventory in its own.
  - Ako Wikipedia-based Text Understanding
- Find significant topics in a text and link them to Wikipedia articles.
- Simulates, how Wikipedia authors manually insert hyperlinks.

Wikification: Find significant topics and link them to Wiki documents.

Figure 1: A news story that has been automatically augmented with links to relevant Wikipedia articles

Step 1: Extraction

- Identify important terms to be highlighted as links in a text
- Consider only terms appearing > 5 times in Wikipedia
- Imporant terms:
  - measure relationship of a term occuring as anchor text in articles & total number of articles it appears in
- Use a predefined threshold for those terms which should be highlighted as links
  - F-measure of 55% obtained on a set of manually annotated Wikipedia articles

Step 2: Disambiguation

- The highlighted terms are disambiguated to Wikipedia articles that capture the indented sense.
  - Jenga is a popular beer in the bars of Thailand.
  - bar → bar (establishment) article
- Given a term, those articles are candidates which contain the term has anchor text.

Machine Learning approach for step 2.

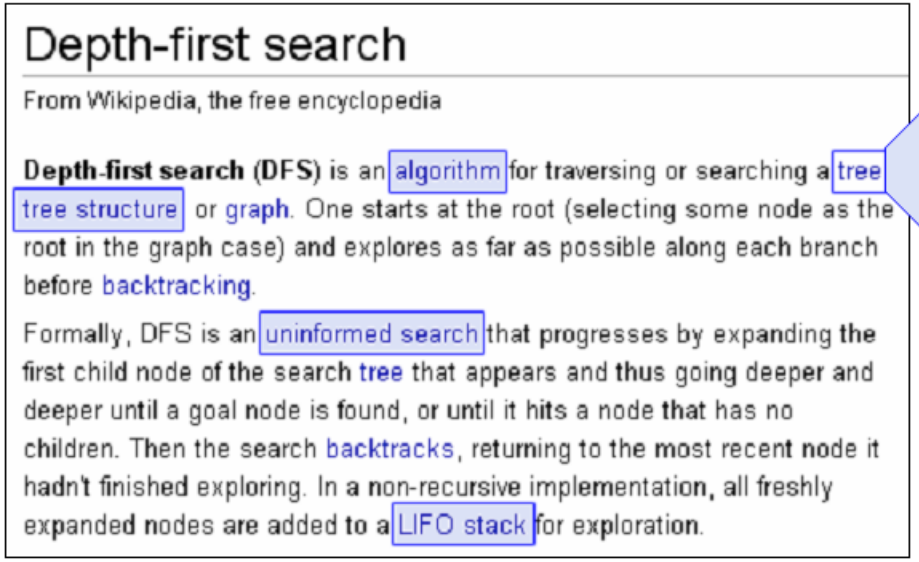
- Supervised: already annotated Wikipedia articles serve as training data
- Features:
  - POS, -3/+3-window+ POS
  - Computed for each ambiguous term that appeas as anchor text of a hyperlink
- Learner: Naive Bayes classifier
- Result: F = 87,7% on 6500 examples

Learning to link in Wikipedia

- Milne & Witten, 2008
- Two important concepts
  - Commonness
  - relatedness

Learning to disambiguate links - commonness

- balancing the commonness of a sense with its relatedness to the surrounding context
- *commonness* (*prior probability*): the number of times a wiki document is used as a destination in Wikipedia



Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (DFS) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

Figure 2: Disambiguating tree *using surrounding unambiguous links as context.*

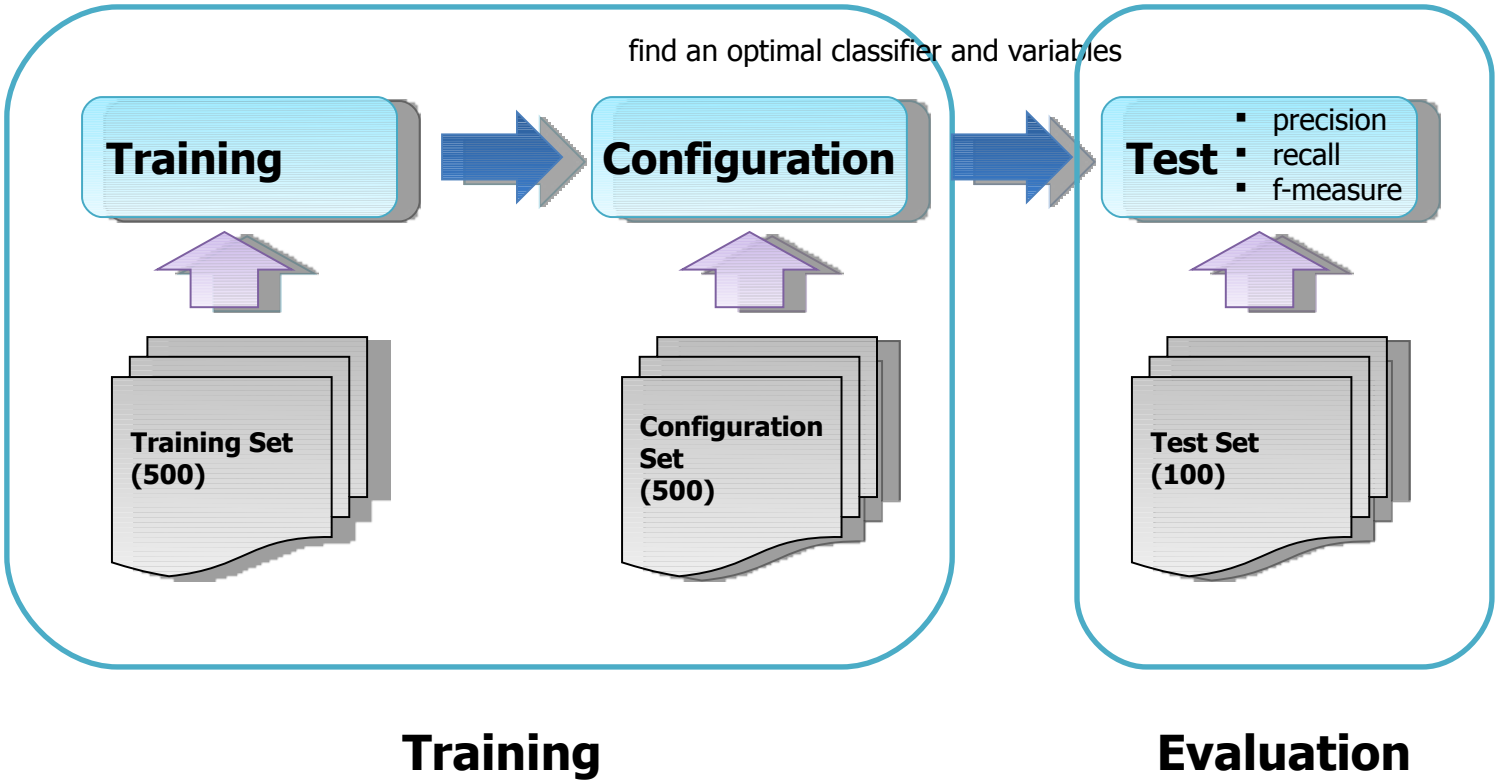
Learning to disambiguate links - relatedness

- Comparing each possible sense with its surrounding context
  - Words consisting context also may be ambiguous
  - Use un ambiguous words that has only one sense
    - ex) algorithm, uniformed search, LIFO stack
  - Reduced to selecting the sense article that has most in common with all of the context articles

$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

  - a,b: articles of interest
  - A, B: sets of all articles that link to a and b
  - W: a set containing all articles in Wikipedia
- some context terms are better than others

Training – Configuration – Test



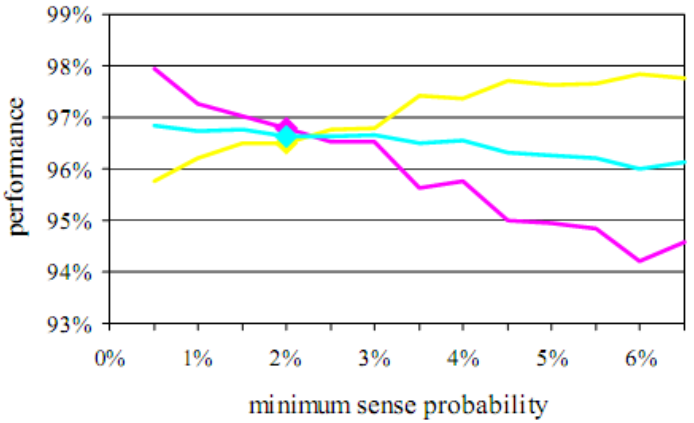
Learning to disambiguate links

– configuration and attribute selection

- identifying the most suitable classification algorithm

	recall	precision	f-measure
Naïve Bayes	96.6	95.0	95.8
C4.5	96.8	**96.5**	96.6
Support Vector Machines	96.5	96.0	96.3
Feature selected C4.5	96.8	**96.5**	96.6
Bagged C4.5	**97.3**	**96.5**	**96.9**

- setting minimum probability of s considered by the algorithm
  - reduce the required time to compare relatedness between context and candidate senses



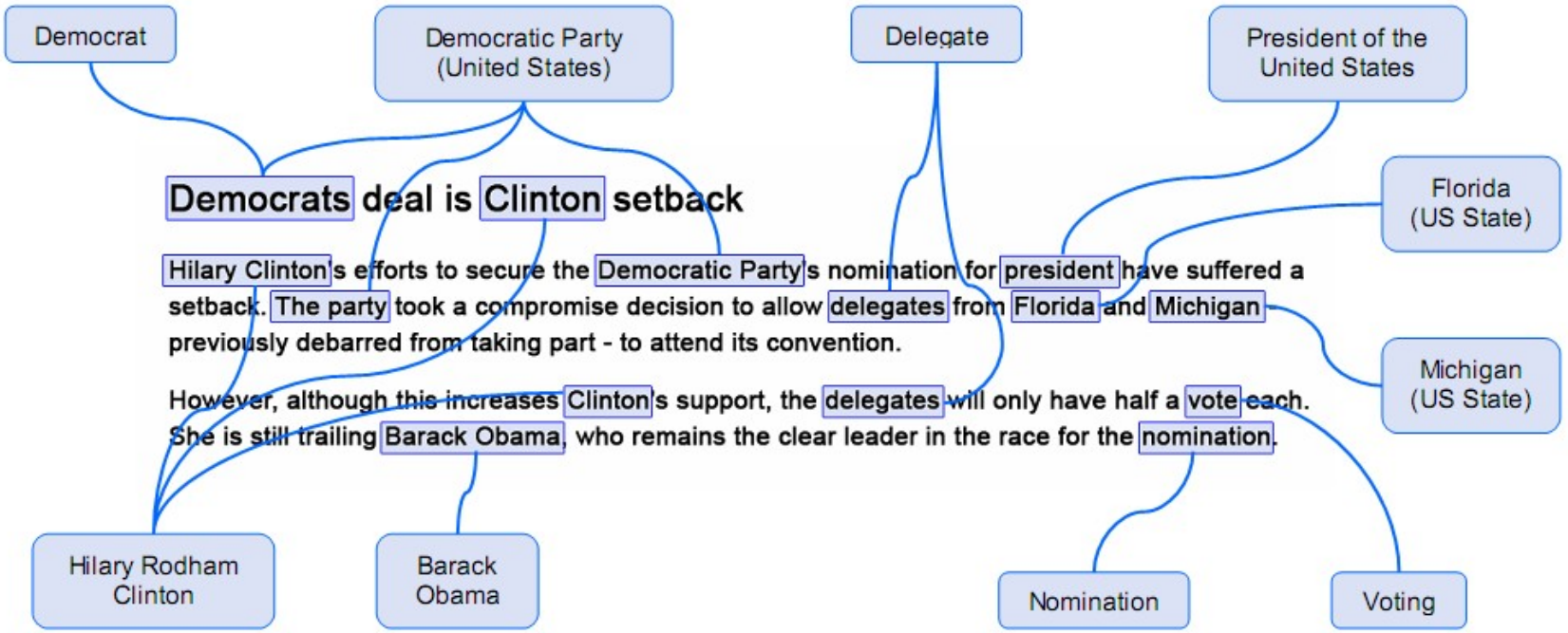
Learning to disambiguate links - evaluation

	recall	precision	f-measure
Random sense	56.4	50.2	53.1
Most common sense	92.2	89.3	90.7
Medelyan *et al.* (2008)	92.3	93.3	92.9
Most valid sense	95.7	**98.4**	**97.1**
All valid senses	**96.6**	97.0	96.8

Learning to detection links

- Naïve approach (Mihalcea and Csomai 2008)
  - If probability that a word or phrase had been linked to an article exceeds a certain threshold, a link is attached to it
- Presented approach
  - Machine learning link detector that uses various features
    - Link probability
    - Relatedness
    - Disambiguation confidence
    - Generality: the minimum depth at which it is located in Wikipedia's category tree
    - Location and Spread
      - first occurrence, last occurrence, spread (distance between them)

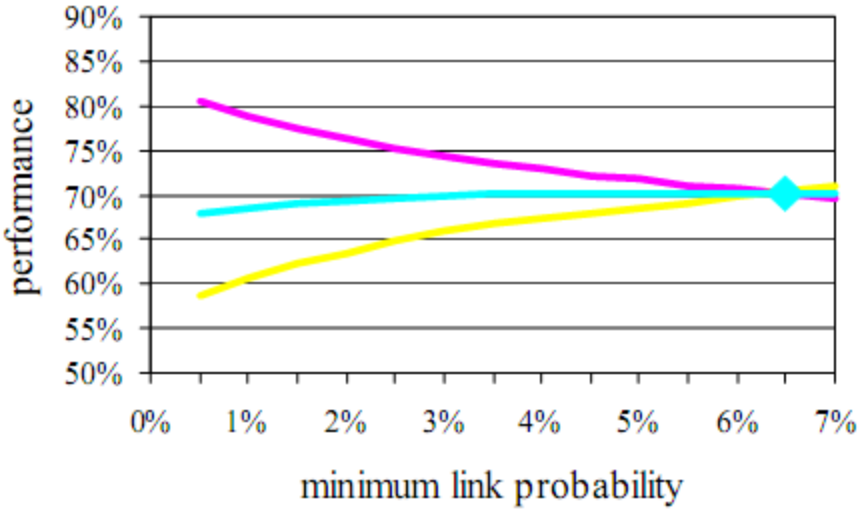
Learning to detection links (cont'd)



Learning to detection links

- training and configuration, and evaluation

	recall	precision	f-measure
Naïve Bayes	70.2	70.3	70.2
C4.5	**77.6**	72.2	74.8
Support Vector Machines	72.5	**75.0**	73.7
Bagged C4.5	77.3	72.9	**75.0**



	recall	precision	f-measure
Wikify (estimate)	46.5	49.6	48.0
Wikify (upper bound)	53.4	55.9	54.6
New link detector	**73.8**	**74.4**	**74.1**