Mining Meaning From Wikipedia

PD Dr. Günter Neumann

LT-lab, DFKI, Saarbrücken
Outline

1. Introduction
2. Wikipedia
3. Solving NLP tasks
4. Named 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 contributors 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 unexperts
  - 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

- 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

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

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

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 private individual. In the more traditional sense, a library is a collection of books. The term can mean the collection, the building that houses such a collection, or both.
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 storing libraries are now repositories and access points for maps, prints or other microfiche, software, audio tapes, CDs, LPs, video tapes and DVDs, and digital CD-ROM databases and the Internet.

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

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)
- Bibliotheca Alexandrina (0)
- Curators (3)
- Defunct libraries (0)
- Fictional libraries (0)
- Free development toolkits and libraries (0)
- 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.

This template is used in articles to identify sentences or short passages which need 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](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:
  - \( lch(c_1,c_2) = -\log(\text{length}(c_1,c_2)/2D)) \)
  - \( \text{length}(c_1,c_2) \): 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**($c_1, c_2$)
  - $\cos(a_1, a_2)$, where $a_i$ is article of concept $c_i$

- **Result:**
  - WS-353: ESA=0.75, LSA=0.56
  - Open-Directory-Project = 0.65 → Wikipedia's quality is greater
ESA: More details

- $T = \{w_1...w_n\}$ be input text
- $<v_i>$ be $T$'s TFIDF vector
  - $v_i$ is the weight of word $w_i$
- Wikipedia concept $c_j$, $\{c_j \in c_1, ... , c_N\}$
  - $N =$ total number of Wikipedia concepts
- Let $<k_j>$ be an inverted index entry for word $w_i$
  - where $k_j$ quantifies the strength of association of word $w_i$ with Wikipedia concept $c_j$
Explicit Semantic Analysis

- The semantic interpretation vector $V$ for text $T$ is a vector of length $N$, in which the weight of each concept $c_j$ 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

Weighted list of concepts

(= Wikipedia articles)

Using Semantic Interpreter

Text₁

Semantic Interpreter

Text₂

Vector comparison

Relatedness estimation

Weighted vector of Wikipedia concepts
**Example: small text input**

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

First ten concepts in sample interpretation vectors
Example: large text input

<table>
<thead>
<tr>
<th>#</th>
<th>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.”</th>
<th>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.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iraq disarmament crisis</td>
<td>Leukemia</td>
</tr>
<tr>
<td>2</td>
<td>Yellowcake forgery</td>
<td>Severe combined immunodeficiency</td>
</tr>
<tr>
<td>3</td>
<td>Senate Report of Pre-war Intelligence on Iraq</td>
<td>Cancer</td>
</tr>
<tr>
<td>4</td>
<td>Iraq and weapons of mass destruction</td>
<td>Non-Hodgkin lymphoma</td>
</tr>
<tr>
<td>5</td>
<td>Iraq Survey Group</td>
<td>AIDS</td>
</tr>
<tr>
<td>6</td>
<td>September Dossier</td>
<td>ICD-10 Chapter II: Neoplasms; Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism</td>
</tr>
<tr>
<td>7</td>
<td>Iraq War</td>
<td>Bone marrow transplant</td>
</tr>
<tr>
<td>8</td>
<td>Scott Ritter</td>
<td>Immunosuppressive drug</td>
</tr>
<tr>
<td>9</td>
<td>Iraq War- Rationale</td>
<td>Acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>10</td>
<td>Operation Desert Fox</td>
<td>Multiple sclerosis</td>
</tr>
</tbody>
</table>

First ten concepts in sample interpretation vectors
Example (texts with ambiguous words)

<table>
<thead>
<tr>
<th>#</th>
<th>Ambiguous word: “Bank”</th>
<th>Ambiguous word: “Jaguar”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“Bank of America”</td>
<td>“Jaguar car models”</td>
</tr>
<tr>
<td>1</td>
<td>Bank</td>
<td>Jaguar (car)</td>
</tr>
<tr>
<td>2</td>
<td>Bank of America</td>
<td>Jaguar S-Type</td>
</tr>
<tr>
<td>3</td>
<td>Bank of America Plaza (Atlanta)</td>
<td>Jaguar X-type</td>
</tr>
<tr>
<td>4</td>
<td>Bank of America Plaza (Dallas)</td>
<td>Jaguar E-Type</td>
</tr>
<tr>
<td>5</td>
<td>MBNA</td>
<td>Jaguar XJ</td>
</tr>
<tr>
<td>6</td>
<td>VISA (credit card)</td>
<td>Daimler</td>
</tr>
<tr>
<td>7</td>
<td>Bank of America Tower, New York City</td>
<td>British Leyland Motor Corporation</td>
</tr>
<tr>
<td>8</td>
<td>NASDAQ</td>
<td>Loreto Region</td>
</tr>
<tr>
<td>9</td>
<td>MasterCard</td>
<td>Loreto Region</td>
</tr>
<tr>
<td>10</td>
<td>Bank of America Corporate Center</td>
<td>Economy of Brazil</td>
</tr>
</tbody>
</table>

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
Results for ESA

- **word relatedness (WS-353)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation with humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet [Jarmasz, 2003]</td>
<td>0.33–0.35</td>
</tr>
<tr>
<td>Roget’s Thesaurus [Jarmasz, 2003]</td>
<td>0.55</td>
</tr>
<tr>
<td>LSA [Finkelstein et al., 2002]</td>
<td>0.56</td>
</tr>
<tr>
<td>WikiRelate! [Strube and Ponzetto, 2006]</td>
<td>0.19 – 0.48</td>
</tr>
<tr>
<td>ESA-Wikipedia</td>
<td>0.75</td>
</tr>
<tr>
<td>ESA-ODP</td>
<td>0.65</td>
</tr>
</tbody>
</table>

- **text relatedness (ABC)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation with humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of words [Lee et al., 2005]</td>
<td>0.1–0.5</td>
</tr>
<tr>
<td>LSA [Lee et al., 2005]</td>
<td>0.60</td>
</tr>
<tr>
<td>ESA-Wikipedia</td>
<td>0.72</td>
</tr>
<tr>
<td>ESA-ODP</td>
<td>0.69</td>
</tr>
</tbody>
</table>
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
    - 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

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

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.

Iranian POW negotiator holds talks with Iraqi ministers

The head of Iran's prisoner of war commission met with two Iraqi Cabinet ministers Saturday in a bid to glean information about thousands of Iranian PCWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister Mohammed Saeed al-Sanhaf told Abdullah al-Najafi that the two states needed to "speed up the closure of what remains from the POW and Missing-In-Action file," INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the 1980-88 Iran-Iraq War. The countries accuse each other of hiding POWs and preventing visits by the International Committee of the Red Cross to prisoner camps.

The ICRC representative in Baghdad, Manuel Bessler, told The Associated Press that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 since 1990.

More than 1 million people were identified as civil law detainees in the largest exchange.

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
- Important terms:
  - measure relationship of a term occurring 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 appears 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

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 unambiguous 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

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

- \(a, b\): 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

Training Set (500) → Configuration Set (500) → Test Set (100)

Training

Configuration

Test

find an optimal classifier and variables

- precision
- recall
- f-measure

Training Evaluation
Learning to disambiguate links – configuration and attribute selection

- identifying the most suitable classification algorithm

<table>
<thead>
<tr>
<th></th>
<th>recall</th>
<th>precision</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>96.6</td>
<td>95.0</td>
<td>95.8</td>
</tr>
<tr>
<td>C4.5</td>
<td>96.8</td>
<td>96.5</td>
<td>96.6</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>96.5</td>
<td>96.0</td>
<td>96.3</td>
</tr>
<tr>
<td>Feature selected C4.5</td>
<td>96.8</td>
<td>96.5</td>
<td>96.6</td>
</tr>
<tr>
<td>Bagged C4.5</td>
<td>97.3</td>
<td>96.5</td>
<td>96.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>recall</th>
<th>precision</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sense</td>
<td>56.4</td>
<td>50.2</td>
<td>53.1</td>
</tr>
<tr>
<td>Most common sense</td>
<td>92.2</td>
<td>89.3</td>
<td>90.7</td>
</tr>
<tr>
<td>Medelyan et al. (2008)</td>
<td>92.3</td>
<td>93.3</td>
<td>92.9</td>
</tr>
<tr>
<td>Most valid sense</td>
<td>95.7</td>
<td>98.4</td>
<td>97.1</td>
</tr>
<tr>
<td>All valid senses</td>
<td>96.6</td>
<td>97.0</td>
<td>96.8</td>
</tr>
</tbody>
</table>
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)
Democrats deal is Clinton setback

Hilary Clinton’s efforts to secure the Democratic Party’s nomination for president have suffered a setback. The party took a compromise decision to allow delegates from Florida and Michigan, previously debarred from taking part - to attend its convention.

However, although this increases Clinton’s support, the delegates will only have half a vote each. She is still trailing Barack Obama, who remains the clear leader in the race for the nomination.
Learning to detection links

- training and configuration, and evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>70.2</td>
<td>70.3</td>
<td>70.2</td>
</tr>
<tr>
<td>C4.5</td>
<td>77.6</td>
<td>72.2</td>
<td>74.8</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>72.5</td>
<td>75.0</td>
<td>73.7</td>
</tr>
<tr>
<td>Bagged C4.5</td>
<td>77.3</td>
<td>72.9</td>
<td>75.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikify (estimate)</td>
<td>46.5</td>
<td>49.6</td>
<td>48.0</td>
</tr>
<tr>
<td>Wikify (upper bound)</td>
<td>53.4</td>
<td>55.9</td>
<td>54.6</td>
</tr>
<tr>
<td>New link detector</td>
<td>73.8</td>
<td>74.4</td>
<td>74.1</td>
</tr>
</tbody>
</table>