NLP & IR – Whither We Goest¹?

Elizabeth D. Liddy
Dean, iSchool
Syracuse University
Syracuse, New York, USA

¹Archaic second person singular present of GO
Overview Questions

- What was early information retrieval like? (before it was called search!)
- How was NLP first applied to the task?
- Which levels of language analysis were utilized?
  - Which were successful? Which were not?
- Why were other levels not incorporated?
- Do we now see that these levels can and need to be investigated?
- If they are, how might they change how we do IR, as well as what tasks we use it for?
Overview Questions

• What was early information retrieval like? (before it was called search!)
• How was NLP first applied to the task?
• Which levels of language analysis were utilized?
  • Which were successful? Which were not?
• Why were other levels not incorporated?
• Do we now see that these levels can and need to be investigated?
• If they are, how might they change how we do IR, as well as what tasks we do it for?
Overview Questions – 1

- Calvin Mooers
  - American computer scientist at MIT
  - Coined term "information retrieval“ in his 1948 Master's thesis
- Also authored Mooers’ Law (not to be confused with Moore’s Law) and its corollary in 1959:
  1. An information retrieval system will tend not to be used whenever it is more painful & troublesome for a customer to have information than for him not to have it.
  2. Where an information retrieval system tends not to be used, a more capable information retrieval system may tend to be used even less.
Looking Backward – 1

- Dialog
  - World's 1st online IR system used globally with significant databases
  - Operational long before the Internet
  - Initially completed in 1966 by Roger Summit
  - New 3rd generation computers could store massive databases centrally and offer worldwide access
  - Revolutionary to provide information services to a global market from a centralized computer facility
  - Real beginning of the commercial field of IR
    - Some problems same today / many problems have been resolved / new problems developed
Dialog was used by librarians on behalf of end users.

There was a single database, that of NASA, and the system allowed only a single person searching at a time.

Reduced the turn-around time for searching the NASA STAR database from 14 hours (plus mail and handling) when done on NASA headquarters IBM 1410 computer to a few minutes at the remote site.

And, the search could be modified during the process without having to reformulate the entire search.

Next, developed similar search engine for Department of Education, known as ERIC
## Everything has Changed!

<table>
<thead>
<tr>
<th></th>
<th>Early IR</th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users</strong></td>
<td>Professional Searchers</td>
<td>Everyone!</td>
</tr>
<tr>
<td><strong>Sources</strong></td>
<td>Bibliographic Databases</td>
<td>Web; Enterprise Systems</td>
</tr>
<tr>
<td><strong>Genres</strong></td>
<td>Scientific, Legal, Medical Reports</td>
<td>News, Blogs, Retail, Email</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>Dedicated Terminals</td>
<td>Laptops &amp; Mobile</td>
</tr>
<tr>
<td><strong>Queries</strong></td>
<td>Boolean</td>
<td>2.5 words</td>
</tr>
</tbody>
</table>
In early systems, Boolean queries were needed
- AND, OR, NOT, (later added ADJ, NEAR)
- But deemed too difficult for non-trained searchers

Even though natural language provides same logic
- Modifiers – adjectives & adverbs
- Conjunctions
- Prepositional phrases
- Negations & other adverbial modifiers

System developers turned away from both Boolean queries & full NL queries

New statistical matching alone on high-volume databases could provide relevant documents

Has led to 2.5 word Google queries of today
- Exceptions – TREC, CLEF
Overview Questions – 2

• What was early information retrieval like? (before it was called search!)

• How was NLP first applied to the task?

• Which levels of language analysis were utilized?
  • Which were successful? Which were not?

• Why were other levels not incorporated?

• Do we now see that these levels can and need to be investigated?

• If they are, how might they change how we do IR, as well as what tasks we do it for?
Levels of Language Understanding

- Synchronic Model of Language
  - Morphological
  - Lexical
  - Syntactic
  - Semantic
  - Discourse
  - Pragmatic
Early Minimal Utilization of NLP

- Inflectional stemming alone
- Deletion of function words
- Parsing was tried and dropped
  - Too early in the development of parsers
    - Slow and produced too many parses
    - Didn’t experiment with partial parsing
  - Matching algorithms not sophisticated enough to use the resulting parses
- Eventually, semantic resources were introduced
  - WordNet
  - Ontologies
  - Word sense disambiguation algorithms
  - Named entity recognition
Does Search Today Need NLP?

• The pure size of the collection of resources that engines such as Google search, typically results in finding something you want, but there are 2 facts to keep in mind:

1. The raw size of the web, makes redundancy a key factor
   • You may **not** find the document / web site that is the **best** hit because it uses a synonymous phrasing of your search
   • But it is highly likely that at least someone used exactly the phrasing you searched on, and that’s what you’ll find.

2. However, there are precious collections which doctors, lawyers & scientists search on, where redundancy won’t help
   • There can very well be just **1** document that answers their question & the author of that document may not have used the phrase searched on.
Overview Questions

• What was early information retrieval like? (before it was called search!)
• How was NLP first applied to the task?
• Which levels of language analysis were utilized?
  • Which were successful? Which were not?
• Why were other levels not incorporated?
• Do we now see that these levels can and need to be investigated?
• If they are, how might they change how we do IR, as well as what tasks we do it for?
Alternative Search Needs Today

- Given short queries, real-time, social, & exploratory search, what could NLP be doing for IR?
  - Just lexical semantic expansion on query?
  - What could higher levels of NLP enable us to accomplish?

- More than surface level representation & matching
  - Typing in a phrase whose site you want to go to – e.g. Facebook, American Airlines, Starbucks

- Seeing an increasing recognition of applications where today’s simple, known-item search does not accomplish the goals
Exploratory search¹

- An information-seeking, problem-solving context that is open-ended
- Searches are typically opportunistic, iterative, and multi-tactical
- When the need is not well-defined in the user’s mind
  - Broad curiosity
  - Learning in unfamiliar areas
  - Scientific discovery
  - Decision making
  - Problem solving
- Differs from what is referred to by librarians as ‘known item search’
  - The look-up based model of IR
  - The pragmatic level of language understanding is useful

Large volumes of electronic data are searched with the intent of finding information to use as evidence in a civil or criminal legal case.

Includes email, IM, tweets, blog postings, text documents, databases, web sites, and any other electronically-stored information which could be relevant evidence.

Current experiments search large heterogeneous digital collections using topical representations that approximate how real lawyers would go about effective discovery.

Results are compared to those of legal teams.
• **Topic 102.** Documents referring to marketing or advertising restrictions proposed for inclusion in, or actually included in, the Master Settlement Agreement (MSA), including, but not limited to, restrictions on advertising on billboards, stadiums, arenas, shopping malls, buses, taxis, or any other outdoor advertising.

• **Topic 52.** Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.

• **Topic 13.** All documents to or from employees of a tobacco company or tobacco organization referring to the marketing, placement, or sale of chocolate candies in the form of cigarettes.
eDiscovery – 2

- Typically many thousands of requests in a single case
  - 1,726 in the Philip Morris tobacco case
  - Searched against 32 million records / documents

- Requests are broader and vaguer than web queries
  - Many can only be satisfied across multiple documents

- High recall is important
  - Jury is instructed to construe missing information as contrary to interests of party that failed to produce it
  - High precision improves efficiency & reduces cost, but there are less onerous sanctions for lack of precision
eDiscovery – 3

- Compared 2 commercial systems on task of categorizing documents as responsive or not, to manual attorney review, to see if IR could perform as “reasonably” as humans
  - Judged 1,600,047 documents for “responsiveness”
  - Humans, systems, and TREC performed equally

- TREC legal track assesses the ability of IR systems to meet eDiscovery needs
  - 55% Recall & 21% Precision – best overall performer
  - 62% Recall & 81% Precision – best interactive Topic

- Level of agreement, both human & system, also similar to TREC
  - Only 71% agreement amongst TREC human assessors
  - Concluded that low human agreement negates notion of ground truth on which precision and recall can be computed¹

Recognizing Connotative Meaning

- Evaluating feasibility of a system being able to detect what a text suggests beyond what is explicitly stated
  - A *sincere* apology
  - An *urgent* request for help
- Preliminary experiments demonstrate agreement among humans on when connotative meaning is present
- Using these cumulative results to train an ML system to recognize the potential for connotative meaning
- Seeking author-independent features for distinguishing implicit connotation across sets of texts
- One IR application is to detect “disgruntled members”
- Pragmatic level of language understanding
Pragmatics

- Functional perspective - study of language in use
- Aspects of language which require context to be understood
  - How is the situational context grammaticalized?
- Goal is to recognize the extra meaning that humans read into utterance without actually being encoded in them
- Relative emphasis:
  - Important in representing / searching informal texts
  - Focus on dialogue, interaction, real-time, social, and exploratory search
- Key to understanding user’s intent / plan in their query
Plan Recognition

- Propositional content does not always fully communicate the speaker’s intent
- How do we understand an utterance which, on the surface means one thing, but clearly means another?

  - We recognize their plan!

  1a. User: *Do you know when the train leaves for Boston?*
  1b. System: Yes.

  2a. User: *Does the train for Washington leave at 4:00?*
  2b. System: No.

- System relies on surface-level syntax & semantics to process user’s questions, but not pragmatic knowledge – user’s plan
Plan Recognition

Utterance / Request

A states

B hears
Plan Recognition

Utterance / Request

Goal / Plan

A states has

B hears infers
Plan Recognition

Utterance / Request

Goal / Plan

states

A

B

has

indirect

unstated

question

answers

infers

hears

implicit
Polarity Recognition Task

• Polarity: mutual opposition; a relation between two opposite attributes or tendencies

• Measuring & utilizing polarity of text
  • Negative or positive attitude of a news reporter
  • Favorable or unfavorable review of a product
  • Right or left political leaning of speaker
  • Certainty or uncertainty about what’s reported
  • Positive or negative scholarly reference citations
  • Denotative or connotative meaning conveyed

• Large amounts of text would benefit from this
  • Web sites
  • Blogs
  • Product sites
  • Reviews
How is Polarity Recognition Done Now?

- 1 to 20 NLP features of text can be used
  - Target verb, syntactic phrase type, voice, ‘affect’ word count, association w/ known set of words, etc
  - Main reliance on lexical level, but taste of pragmatics included
- Machine Learning based on humanly annotated data
  - Typically, 50 to 100 documents
- Machine Learning methods
  - N-gram
  - Support Vector Machine
  - Naïve Bayes
- Semantic Orientation: PointWise Mutual Information (SO-PMI)

- Algorithm that estimates the semantic orientation (+ or - valence) of each word in a text
  - Evaluative character and degree of a word
- Base SO of a word on its statistical association with a set of positive and negative paradigm terms
- Computes a measure of association using:
  - Pointwise Mutual Information (PMI)
  - Latent Semantic Analysis (LSA)
  - Semantic distance in WordNet

Paradigm Terms

- Studies use wide range of # of paradigm words
  - 1,300 for Hatzivassiloglou & McKeown, 1997
  - 14 for Turney & Littman, 2003
    - + good, nice, excellent, positive, fortunate, correct, superior
    - - bad, nasty, poor, negative, unfortunate, wrong, inferior
- Applications may retrain with specialized words for their domain
- Humans - high agreement on SO of words
  - 97% agreement amongst 4 subjects (Hatzivassiloglou & McKeown, 1997)

Computing SO-PMI

- Word co-occurrence counts for PMI formula are determined using Information Retrieval
  - Each content word in the ‘text’ being processed is used as a query to a search engine
  - Search a corpus using NEAR operation for each paradigm term to determine count
    - Words occurring within 10 words
    - Possible with 14 queries / runs for each term in text (if using Turney & Littman)

- Compute + or - valences of words from text to determine text’s overall orientation
Does this Matter in Users’ Searches?

- **Product Reviews**
  - 81% of Internet users search for product reviews
  - 20% do so on each typical day
  - 73% - 87% say reviews have a significant influence on their purchase
  - 32% have provided a rating on a product, service, or person via an online ratings system

- **Political Opinions**
  - In recent US elections, 34% Americans searched the web to find perspectives that differed from their own
Application 1 - Classifying Reviews

• Task – processing new reviews
  • Part-Of-Speech tagging
  • Recognition of phrases with ADJs or ADVs
  • Calculate SO-PMI of each phrase
  • Compute average SO of extracted phrases
  • Assign + or – to the review

• Challenge
  • Positive reviews may have words with negative SO, although subjective evaluation is positive

• Typical performance is in range of 80 – 95%
Application 2 – Shopping Engines

- Shopping.com
- Epinions
- Dealtime
- Provide product reviews
  - More than simple thumbs up or thumbs down
  - Summarized form
  - Detailed reviews
- Data is mined from full reviews
  - Pros
  - Cons
- Next Goal
  - More specialized focus of SO
Sample Reviews of iPod

Pros:
- Wi-Fi and Safari, you'll never be lost again
- Beautiful graphics and screen
- Very slim and lightweight
- Cover Flow is a beautiful feature
- Can see photos on portable device with amazing graphics
- It's set up so you can surf the web and still listen to your music at the same time

Cons:
- Wi-Fi can be temperamental
- Safari keyboard is missing some needed symbols
- No hard drive capability
- Awful headphones
SentiWordNet

- A newer lexical resource in which each Synset of WORDNET is associated to 3 numerical scores
  1. Is it subjective?
  2. If so, is it’s subjectivity Positive or Negative?
  3. If Positive or Negative, to what degree are terms contained in the Synset subjective?
    - E.g. Weakly Positive, Mildly Positive, Strongly Positive

- As a graded lexical resource, it appears to provide enough information to capture nuances

- Freely available for research purposes, with a Web-based graphical user interface

SentiWordNet Evaluation

- Produced a human labelling of a subset of 1,000 WORDNET synsets to use as a “gold standard”
  - Synsets were tagged by five different evaluators
  - For each synset, each evaluator attributes a score for each of the three labels of interest such that the three scores sum up to 1.0 through a graphical interface

- Evaluated these scores as compared to scores assigned automatically to the same synsets in SentiWordNet
  - Results showed that use of SentiWordNet produces a significant improvement over baseline system & not using any specialized lexical resource
  - Significant improvement with respect to the use of other opinion-related lexical resources.
SentiWordNet Synset Evaluation

PN polarity

Positive  Subjective  Negative

SO polarity

Objective

Term Sense Position
SentiWordNet Evaluation

- Produced a human labelling of a subset of 1,000 WORDNET synsets to use as a “gold standard”
  - Synsets were tagged by five different evaluators
  - For each synset, each evaluator attributes a score for each of the three labels of interest such that the three scores sum up to 1.0 through a graphical interface

- Evaluated human scores as compared to scores assigned automatically to the same synsets in SentiWordNet
  - Results showed that use of SentiWordNet produces a significant improvement over baseline system & not using any specialized lexical resource
  - Significant improvement with respect to the use of other opinion-related lexical resources.
What Makes a Review Helpful?

- UCD researchers developed a system to rate helpfulness of online customer reviews
- Applied ML to determine factors that contribute to a helpful review
  1. **Reputation** – analyzes all previous reviews written by same author to determine if author is known for providing useful comments.
  2. **Social** – tabulates how often reviewers respond to other reviews, improving the quality of their own statements about products or services.
  3. **Sentiment** – acknowledges that users respond more to positive reviews & takes the score attributed to a product into account.
  4. **Content** – looks at how well a review is written by form & length as poorly written reviews are considered less helpful by users.
Application 3 - “Critical” Nature of Citations

- Scientific texts also contain *subjective* content
- Indicated by ‘critical’ citing relations to documents
  - Similar to Shepardizing in the legal domain which reveals whether a court ruling has been upheld or reversed
- **Opportunity:** *CiteSeer* has 10,000+ hits a day
  - Shows snippet of text around citation, but may not be the referring text (69% found in prior sentences)
  - Could increase usefulness with “critical” relations
- Citations can be enriched by indicating relation
  - *Criticizes* (theory / method / results / conclusions)
  - *Supports* ..............
  - *Based on* ..............
- Enables more precise retrieval
  - E. g. “Find papers which criticize Liddy, 2007.”

Teufel, S. Argumentative Zoning for Improved Citation Indexing. AAAI EAAT Symposium, 2004
Recognizing “Critical” Relations

- Verb clusters identified for change, failure, contrast, etc
  - Change – adapt, augment, combine, extend, modify, refine
  - Failure – problematic, contradict, overgeneralize, resort to
  - Contrast – be distinct from, conflict, differ from, differentiate

- Produces enriched citation indexing
  - Not just citation of document being referred to
  - Also, indication of the nature of the critical citing relationship
  - Text to which citation is referring

- Trained on 80 papers on which nature of the citing relationships had been manually annotated

- Evaluation showed significantly improved retrieval results when looking for critical relations between papers
Certainty

- The quality / state of being free from doubt, especially on the basis of evidence.

Earlier work:
- Types of subjectivity (Liddy et al. 1993; Wiebe 1994, 2000; Wiebe et al. 2001),
- Adverbs and modality (Hoye, 1997),
- Hedging in different kinds of discourse
- Expressions of (un)certainty in English (from applied linguistics)

Goal – utilize degree of ‘conviction’ of textual statements for improved retrieval
Four-Dimensional Relational Model for Certainty Categorization

**D1: PERSPECTIVE**
- **Writer’s Point of View**
  - Directly involved 3rd parties (e.g. witnesses, victims)
  - Indirectly involved 3rd parties (e.g. experts, authorities)

**D2: FOCUS**
- **Abstract Information** (e.g. opinions, judgments, attitudes, beliefs, emotions, assessments, predictions)
- **Factual Information** (e.g. concrete facts, events, states)

**D3: TIMELINE**
- **Past Time** (i.e. completed, recent in the past)
- **Present Time** (i.e. immediate, current, incomplete, habitual)
- **Future Time** (i.e. predicted, scheduled)

**D4: LEVEL**
- Absolute
- High
- Moderate
- Low

Pilot Study Data

- 32 articles published by same newspaper
  - 685 sentences, excluding headlines
  - Editorials & news reports

- Frequency of occurrence of explicit certainty markers
  - In 32 articles, an average of **53%** sentences were identified as having recognizable certainty markers

- Certainty marker examples:
  - *it was not even clear that*...
  - *remains to be seen*...
  - *don’t believe they will*...
  - *not necessarily*...
  - *Estimated*...
  - *seems exaggerated*...
  - *would probably have to*...
  - *will almost certainly have to*...
Experimental Performance Results

- Test set of 113 instances served as gold standard for testing system performance
  - Mainly lexical clues
  - Some phrasal
  - Some tense

- 82 (72.6%) were identified by the system correctly

- Ongoing development
  - Discourse level features
  - Larger corpora
  - Different genre
Potential Applications

- **Alerting** analysts to above or below normal level *certainty* and associating it with its source

- **Searching** by level and point of view
  - *Which aspects of foreign policy does Pres. Obama sound most certain about in his speeches?*

- **Rank ordering** retrieved documents by *certainty* of author or author’s report of *certainty* of others
  - Decreases amount of uncertain information in retrieved reports
  - Prioritizes sources that provide *highly certain* information

- **Summarizing** per certainty level, per topic, per document, and across documents

- **Inferring** true state of affairs from high level *certainty* statements from multiple sources
5 – Using Credibility of Blog Posts for IR

- Goal - Improve rankings in response to query
- Developed a framework for assessing blog credibility
  1. Blogger’s expertise & offline identity disclosure
  2. Blogger’s trustworthiness & value system
  3. Information quality
  4. Appeals & triggers of a personal nature
- A subset of factors tested on TREC ’06 data
  - Those that can be estimated automatically from test collections, and don’t use external information

Assessing Credibility of Blog Posts

- Work by Weerkamp & de Rijke, 2008
- 2 types of indicators
  - Post level & blog level
    - Topic-dependent
      - Timeliness with triggering event
      - Semantic similarity – language use compared to news on topic
    - Topic-independent
      - Capitalization, shouting, spelling, length, spam, comments, regularity, consistency
- Combination of all indicators produced MAP score higher than best performer in TREC ’06
- Top 3 performance on p@10

Let’s Pause, and ask….

• What is it we want to accomplish in IR now?
  • Is it just more relevant sites for the same 2.5 word queries?
  • What about the needs of financial, patent, and legal professionals, or more sophisticated enterprise search?

• Will the trend towards more sophisticated usage of prior search results and users’ search behavior be able to improve retrieval?

• How can NLP contribute to accomplishing this goal?
  • Can the higher levels of language analysis be made computably tractable enough to use?
  • Can we operationalize what we know about pragmatics for real-world, real-time search improvements?
  • Will genres such as tweets, blog posts, Facebook pages provide more context, thereby enabling Pragmatics to contribute?
Today's Challenges / Opportunities – 1

- Real Time Search
  - Emphasis on currency / ‘as it’s happening’ / recency
    - E.g. Up-to-the-minute info on traffic delays
    - Should this be a pull technology or a push technology against standing queries?
  - More about conversations, recent news, trendy topics
  - Twitter, blogs, Facebook
  - Are these ‘shallow’, ‘superficial’ questions?
    - Should search be done differently?
    - Isn’t this the perfect opportunity to use pragmatics?
  - How are parameters for currency integrated with relevancy in displaying results?
    - Now add in credibility and certainty
Today’s Challenges / Opportunities – 2

- Social / Collaborative search
  - Live updates from different blogs / Twitter accounts
  - Is agreement important? Or uniqueness?
  - Weighting by links, retweets, # of followers?
  - Again, credibility & certainty!

- Mobile search
  - How to combine / weight geographic proximity, relevance, & semantic orientation?
  - Should the fact that it’s mobile, fundamentally change the way search is done?
Today’s Challenges / Opportunities – 3

- Question-Answering in medical insurance domain
- Situation:
  - If patient chooses a physician who is not part of their insurance plan, costs are more than if you stayed in network
  - Insurance company reimburses a “usual & customary rate”
  - Large government investigation has found that insurance companies underpay patients for ‘out of network’ charges
- $100 million US government suit of companies
- Syracuse University & Fair Health have been awarded the contract
  - Determine new, fairer rates based on all medical insurance claims data across US
  - Build a consumer web site, which interacts with consumers on very complex questions
    - Ultimate challenge in use of NLP for IR
Again, Let’s pause, and ask....

- What is it we want to accomplish in IR now?
  - Is it just more relevant sites for the same 2.5 word queries?
  - What about the needs of financial, patent, and legal professionals, or more sophisticated enterprise search?

- Will the trend towards more sophisticated usage of prior search results and users’ search behavior be able to improve retrieval?

- How can NLP contribute to accomplishing this goal?
  - Can the higher levels of language analysis be made computably tractable enough to use?
  - Can we operationalize what we know about pragmatics for real-world, real-time search improvements?
  - Will genres such as tweets, blog posts, Facebook pages provide more context, thereby enabling Pragmatics to contribute?