# The State of the Art in Computational Linguistics: How to get at information encoded in natural language

Amazon Developer Conference
January 17th, 2007

Emily M. Bender
Assistant Professor & Faculty Director
Professional Master's in Computational Linguistics
University of Washington
ebender@u.washington.edu

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

# Goals of this talk

- Overview of the field of computational linguistics
- Pointers to resources that are currently available
- Give a sense of the state of the art ...
- ... by means of a handful of relevant examples

# Overview

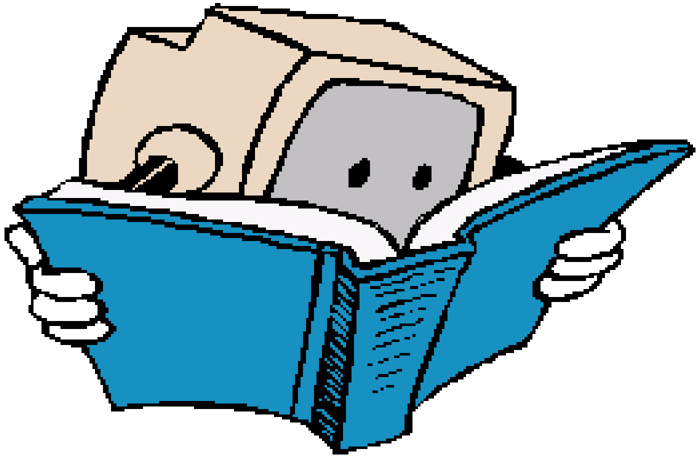
- What is NLP and what it is good for?
- Resources
- Stepping stones
- Wrap up

# Overview

- What is NLP and what it is good for?
- Resources
- Stepping stones
- Wrap up

# What is NLP?

- NLP: The processing of natural language text by computers
    - for practical applications
    - ... or linguistic research
- NLU: NLP with the goal of extracting meaning from the text for further machine processing



# NLP/NLU for Amazon.com

- Classification of reviews
- Classification of books (similar topics, similar styles)
- Smarter searching
- Searching across languages (e.g., book reivews)
- Semi-automated customer service
- Other?

# Human Language Understanding

- Relies on a wealth of intricate grammatical knowledge
- Is supported by an even greater wealth of world knowledge
- This means that information stored in natural language text requires a complex set of keys

# Levels of linguistic structure

- Phonetics: Speech sounds, how we make them, how we perceive them
- Phonology: The grammatical structure of sounds and sound systems
- Morphology: How meaningful sub-word units combine to make words
- Syntax: How words combine to make sentences
- Semantics (lexical, propositional): What words mean and how those meanings combine to make sentence meanings
- Pragmatics: How sentence meanings are used to convey communicative intent
- ...

# Pervasive ambiguity

- Phonetic: *It's hard to wreck a nice beach.*
- Morphological: *This choice is undoable.*
- Syntactic: *Time flies like an arrow.*
- Semantic: *Every person read some book.*
- Pragmatic: *You should take those penguins to the zoo!*

# And that’s only the tip of the iceberg!

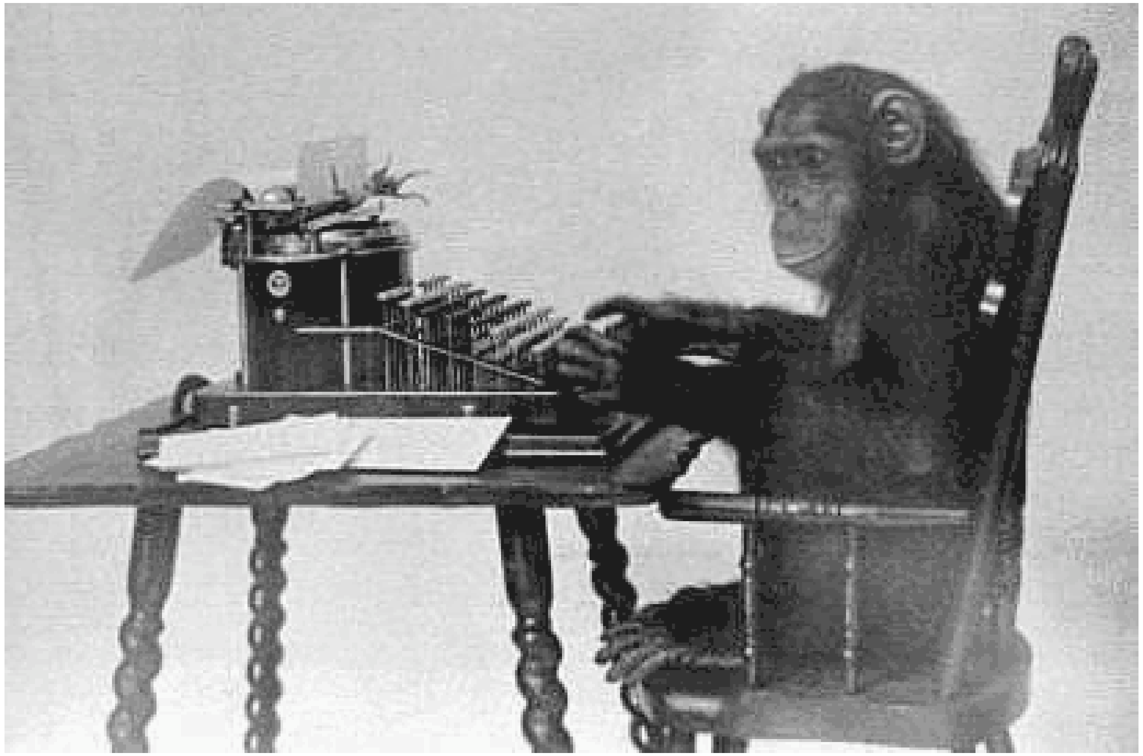
- Ambiguities are typically independent, leading to combinatorial explosions.
- *Have that report on my desk by Friday* (32-ways ambiguous)
- Humans are generally bad at detecting ambiguity, a consequence of being so good at *resolving* it.
- In NLP, stochastic models usually stand in for the common sense knowledge people use.

# NLP: Subtasks

- Tokenization: sentence segmentation, word segmentation
- Morphological analysis: POS tagging, stemming, full morphological analysis
- Named-entity recognition
- Word-sense disambiguation
- Parsing (to syntactic or semantic structure)
- Reference resolution
- Dialogue management
- Generation
- Document classification
- ...

# NLP: Spectrum of approaches

- Knowledge engineering
- Stochastic models
  - Supervised v. unsupervised training
  - Incorporation of hand-made resources
- Hybrid approaches



# Overview

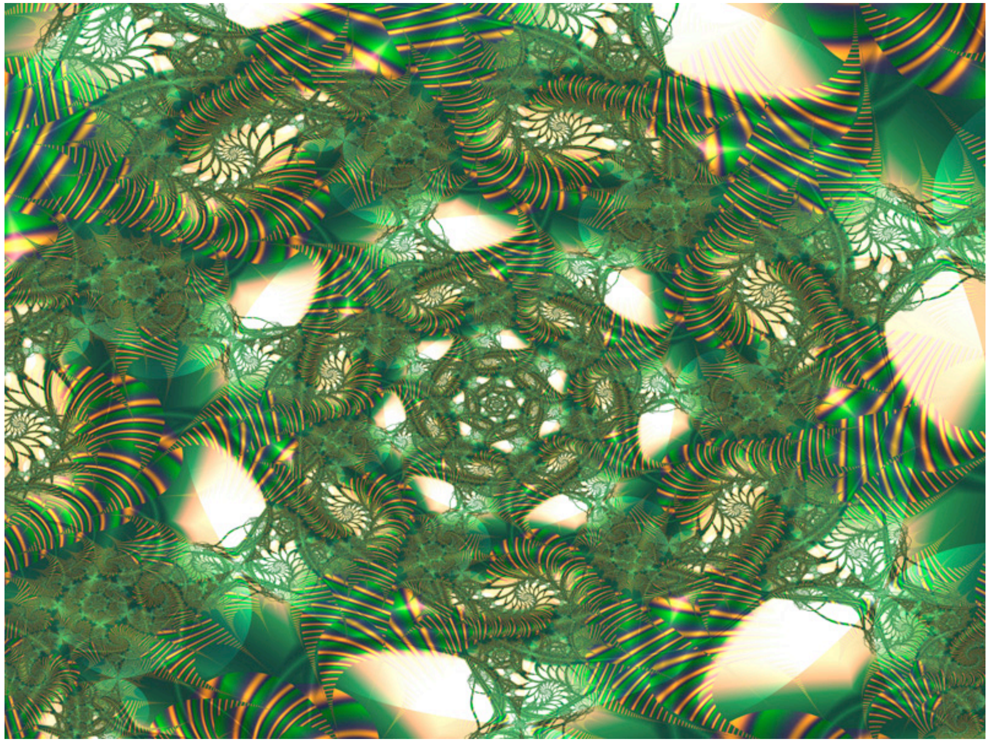
- What is NLP and what it is good for?
- Resources
- Stepping stones
- Wrap up

# Resources: Treebanks

- Collections of text with syntactic structures for each sentence
- Most famous: Penn Treebank (PTB)
- Innovations: Propbanks, dynamic precision-grammar based treebanks (e.g., Redwoods: http://www.delph-in.net/redwoods)

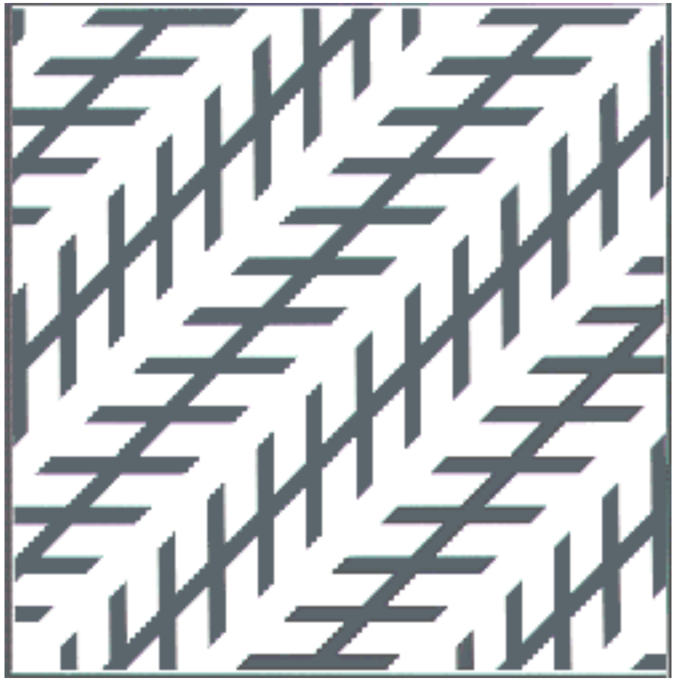
# Resources: WordNets

- Representation of word senses through “synsets”: Collections of words which are near synonyms (on one of their senses)
- (Proto-)Ontology: hyponymy and hypernymy relations between synsets
- Available now for many languages: http://www.globalwordnet.org
- FrameNet: Includes annotation of semantic arguments http://framenet.icsi.berkeley.edu



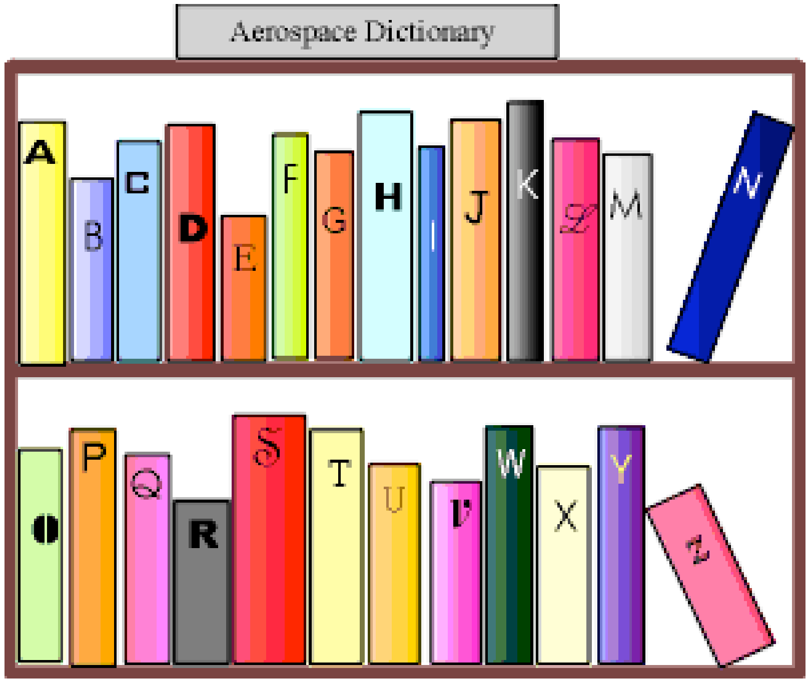
# Resources: Parallel text

- Translations of a text into one or more languages
- Sometimes with sentence-level alignment
- More rarely with word-level alignment
- Most famous: Canadian Hansards, Europarl



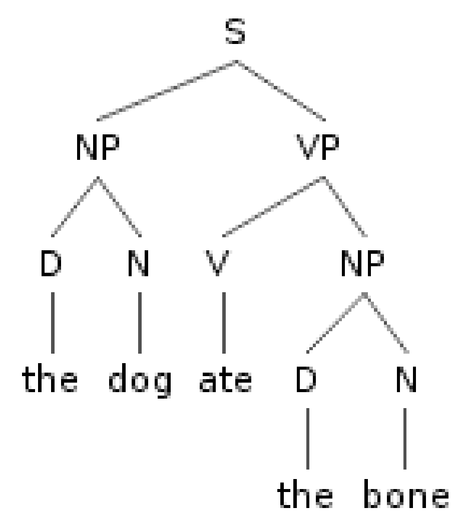
# Resources: Machine-readable dictionaries

- Monolingual, associate word forms with:
    - prose definitions
    - pronunciations
    - part-of-speech annotation
- Bilingual
    - word-sense aligned
    - ... or not (more common)



# Resources: Morphosyntactic analysis

- Part of speech tagging (e.g., Brill's tagger)
- Morphological analysis/stemmers (e.g., Porter Stemmer)
- Statistical parsing (e.g., Collins parser, RASP)



# Resources: Summary

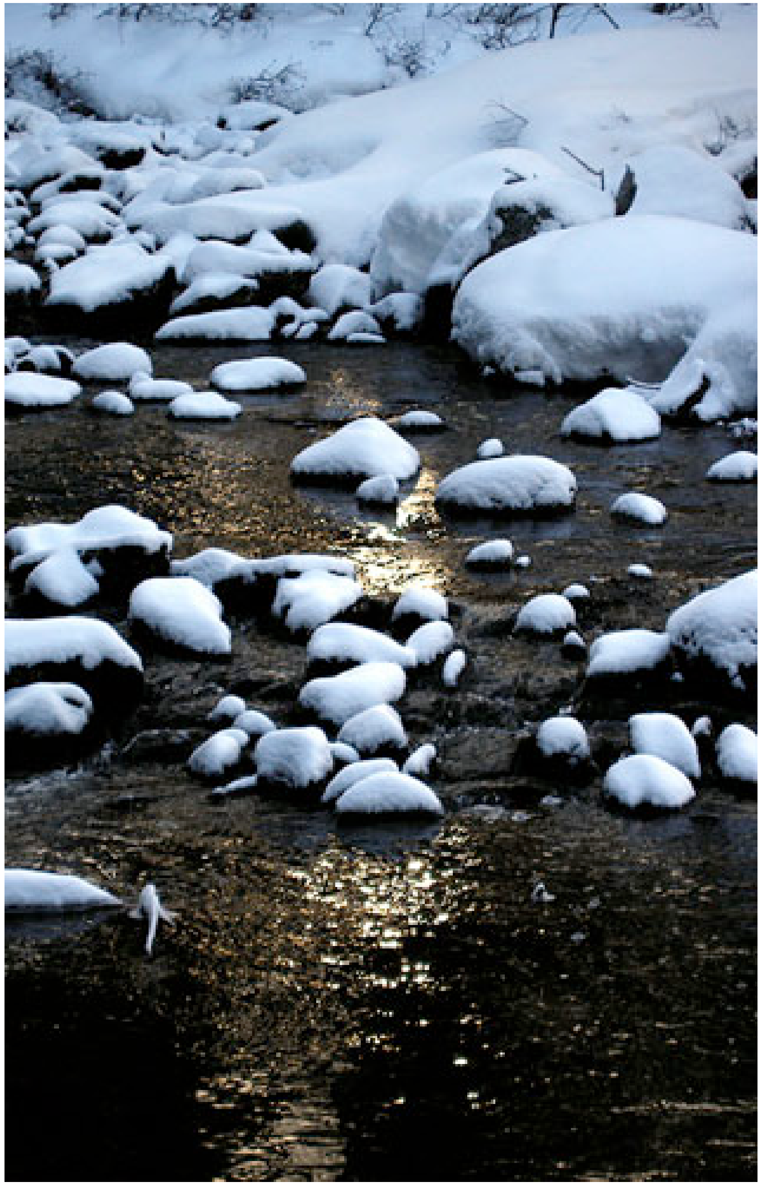
- Availability varies widely by language
- Reusability is typically an important goal
- Primary outlet: Linguistic Data Consortium (http://ldc.upenn.edu)
- Other sources: B2B companies, academic websites

# Overview

- What is NLP and what it is good for?
- Resources
- Stepping stones
- Wrap up

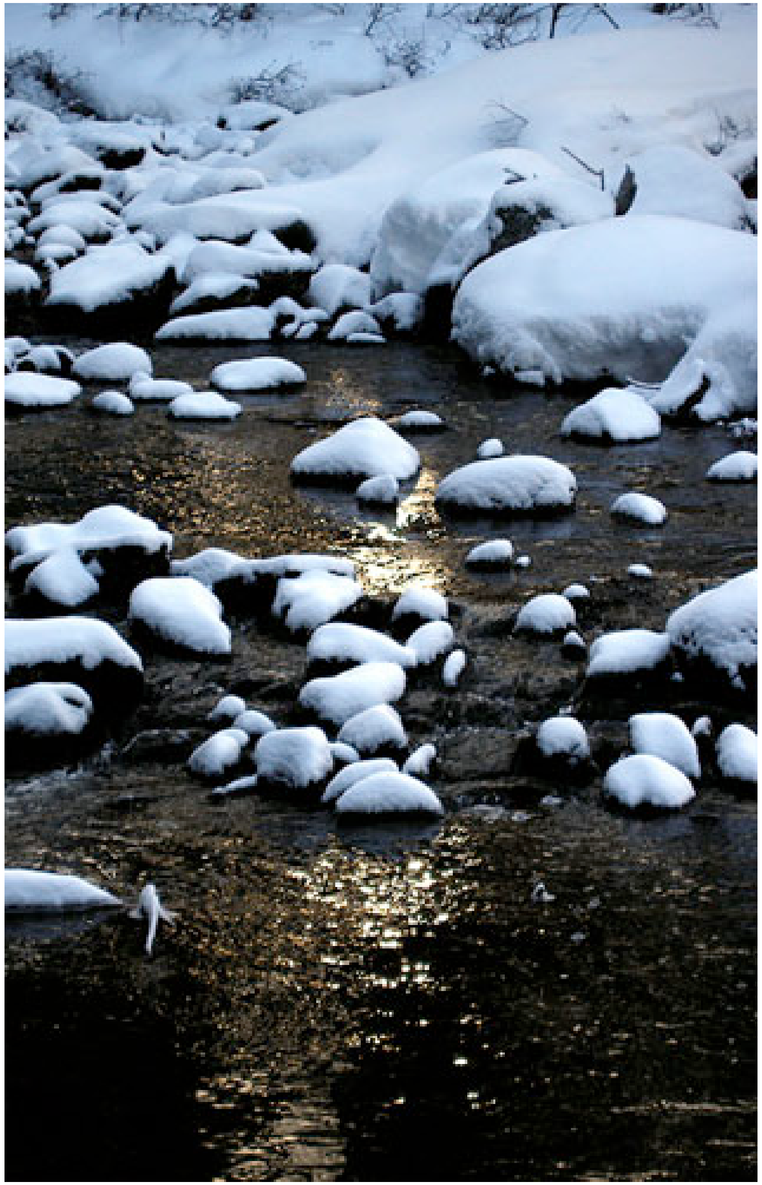
# Stepping-stones: Overview

- Gliozzo & Strapparava (2006): Cross-language text categorization
- Lv et al (2006): Personalized search
- Miyao et al (2006): Retrieval of relational concepts from massive text databases



# Stepping-stones: Overview

- Presentation of problem
- Resources
- Methodology
- Evaluation



# Gliozzo & Strapparava: Problem

- Classification of texts in one language when training data is only available in another language
- Useful in cross-language question answering, categorization of documents in community based trade, construction of comparable corpora for MT
- Text classification exploits the tendency for documents about the same topic to use the same words
- Documents in different languages don't use the same words!
- ... mostly: The exceptions are (some) cognates and (some) proper names in languages with sufficiently similar orthographies
- Can such overlaps in vocabulary be used to 'seed' cross-language text classification?
- How much do bilingual dictionaries help?

# Gliozzo & Strapparava: Resources

- *AdnKronos* corpus of Italian and English news
  - 32,354 Italian articles, 27,821 English articles
  - Classified into four categories: Quality of life, Made in Italy, Tourism, Culture and School
- MultiWordNet, for Italian (linked to English): 58,000 Italian word senses, 41,500 lemmas, and 32,700 synsets
- Collins machine-readable Italian-English dictionary: 37,727 English head words and 32,602 Italian head words
- svmlight (Joachims 2002) for implementing Support Vector Machines

# Gliozzo & Strapparava: Methodology

- Monolingual text classification: Create a vector for each document in *k*-dimensional space, where *k* is the size of the vocabulary of the whole corpus of texts.
- Estimate similarity between two texts by taking the cosine of their vectors in this Vector Space Model (VSM)
- Multilingual text classification: Translate classical VSM into a multilingual domain VSM, using Latent Semantic Analysis (LSA)
- The LSA (performed through Singular Value Decomposition) has the effect of “pulling along” additional words into domains established by the shared vocabulary.
- The greater the shared vocabulary, the more precise the method.
- Boosting with word-sense aligned dictionaries allows even more words to be pulled in.
- Without word-sense alignment, too much noise in second-degree terms.

# Gliozzo & Strapparava: Evaluation (1/2)

- Train on 75% of Italian data, test on 25% of English data, and vice versa
- Multilingual domain model trained on training portion of both corpora (without classifications annotated)
- Baseline for comparison is the Bag of Words model: only actual overlapping vocabulary is considered.
- Results reported as F-measure (harmonic mean of precision and recall)

# Gliozzo & Strapparava: Evaluation (2/2)

| | Train: English Test: Italian | Train: Italian Test: English |
|---|---|---|
| No dictionaries | .65 (.45) | .55 (.41) |
| Word-sense aligned dict. | .72 (.59) | .69 (.61) |
| Collins dictionary | .89 (.93) | .89 (.92) |

# Lv et al: Problem


- Search engines aren't responsive to user context.
- Can implicit feedback (recent queries + documents the user chooses to view) be used to improve search?

# Lv et al: Methodology (1/2)


- Create a graph such that:
    - terms and documents are both represented as nodes
    - edges represent term occurrence in documents
    - edge weights represent term frequency in documents
- Assign “authority” scores to pages and “hub” scores to terms
- Iteratively update “authority” and “hub” scores based on the “mutual reinforcement principle”


# Lv et al: Methodology (2/2)

- Extract query terms from recent query logs
  - Use VSM to calculate similarity to current query
  - For queries over threshold, select top 30% (by tf*idf) from clicked documents in stored query
- Extract query terms from immediately viewed documents
  - By term frequency in viewed documents
  - Inverse document frequency in entire search result (snippets only)
- Expand and rerank list of query results

# Lv et al: Evaluation (1/2)


- 9 student participants tried two variants of the system plus two comparison systems on queries from TREC and HTRDP (English and Chinese IR competitions)
- Participants could expand queries at will and click on documents
- Results from all systems combined and evaluated for relevance
- Resulting precision in Top 5, 10, 20, 30 results calculated

# Lv et al: Evaluation (2/2)

| Top 5 precision | English | Chinese |
|---|---|---|
| Google | .558 | .464 |
| UCAIR | .655 | -- |
| PAIR, no QE | .681 | .521 |
| PAIR | .706 | .585 |

# Miyao et al: Problem

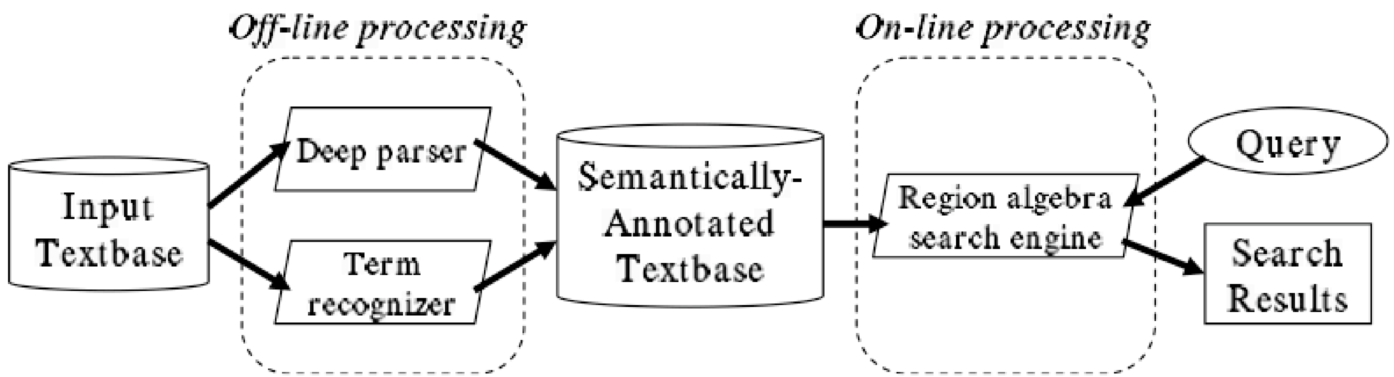
- Biomedical results are reported in natural language text.
- MEDLINE indexes 4500 journals (14,785,094 articles as of 2006).
- Researchers want answers to queries like: “What triggers diabetes?”, “What inhibits ERK2?”
- State-of-the-art: Keyword based searches.
- Can semantic search (using ontologies and parsing for predicate argument structure) do better?
- Big problem: Lots of text, a broad range of concepts
- Also narrow: Queries target simple relations between two entities

# Miyao et al: Resources

- Ontologies: GENA (metadatabase of genes and gene products; Koike & Takagi 2004); UMLS (other biomedical and health concepts; Lindberg et al 1993)
    - Map textual expressions to real-world entities
- Term recognizer: map expressions in the text to ontology entries (Tsuruoka and Tsujii 2004)
- Parsing technology: A probabilistic HPSG parser (Miyao & Tsujii 2005), which extracts predicate argument structure. (97.6% coverage on MEDLINE corpus)
    - *exclude* (ARG1: *CRP*, ARG2: *thrombosis*)
- Treebank: GENIA Treebank (Tateisi et al 2005), contains biomedical domain text

# Miyao et al: Methodology

- Parse corpus offline, store predicate-argument structures in a structured database.
- Run term recognizer to annotate sentences with links to ontology




- Convert queries to extended region algebra
- Match queries to semantic annotations to return relevant passages

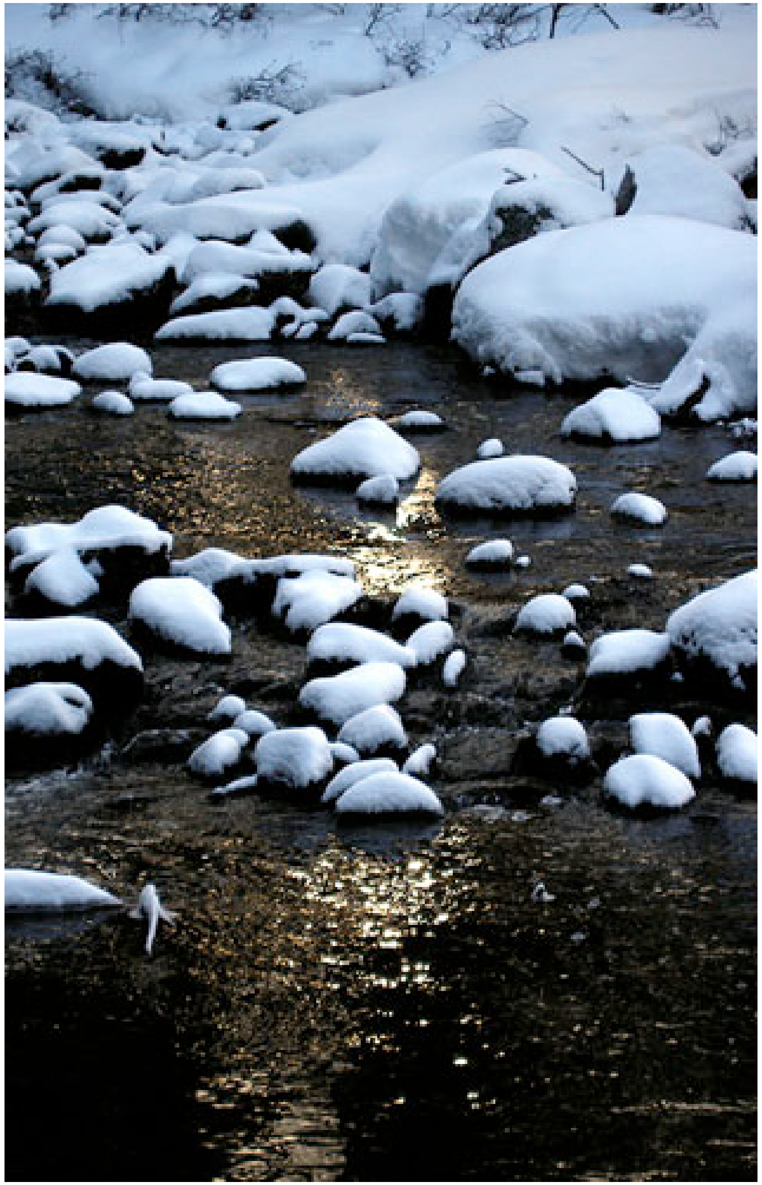
# Miyao et al: Evaluation

- 8 sample queries
- Max 100 results per query
- With and without ontological mapping; keyword or semantic matching
- Present results from different conditions in random order to biologist for evaluation
- Keyword precision ranges from 0-74%
- Semantic search precision 60-97%
- Effect of ontological mapping also clear
- This task values precision over recall

Demo

# Stepping-stones: Overview

- Gliozzo & Strapparava (2006): Cross-language text categorization
- Lv et al (2006): Personalized search
- Miyao et al (2006): Retrieval of relational concepts from massive text databases



# Overview

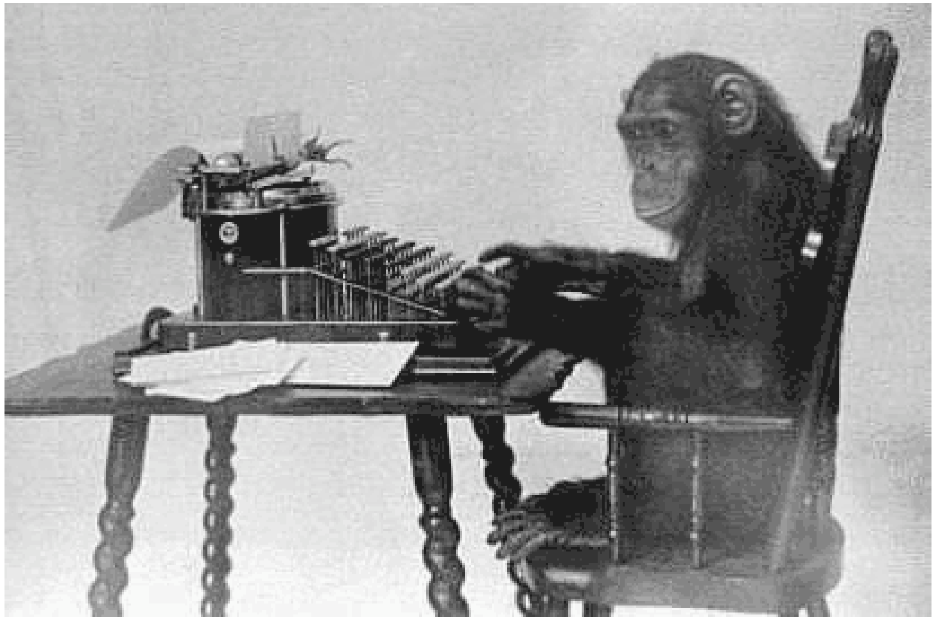
- What is NLP and what it is good for?
- Resources
- Stepping stones
- Wrap up

# Evaluation

- For many NLP tasks, appropriate evaluation metrics remain elusive
    - What is the “right” answer for machine translation, text summarization?
    - In other cases, annotated corpora can provide a gold-standard for comparison
- A lot of NLP work is driven by competitions or shared tasks, which provide standardized, competitive evaluation

# Knowledge engineering and machine learning

- After swinging hard towards machine learning, the pendulum is returning to hybrid approaches
- Knowledge engineering contributes precision, depth of analysis
- Machine learning contributes robustness and scalability



# The promise of NLP

- The amount of information stored in digitized text is increasing every day
- NLP provides improved access to information:
    - Machine translation and other multilingual NLP
    - Automated question answering based on web content
    - NLP for business intelligence
    - ...

# To learn more...

- The ACL recently launched a wiki:

http://aclweb.org/aclwiki

- Papers from top conferences back to 1965 are available online:

http://acl.ldc.upenn.edu

- Computational linguistics at the University of Washington

http://www.compling.washington.edu

Thank you!