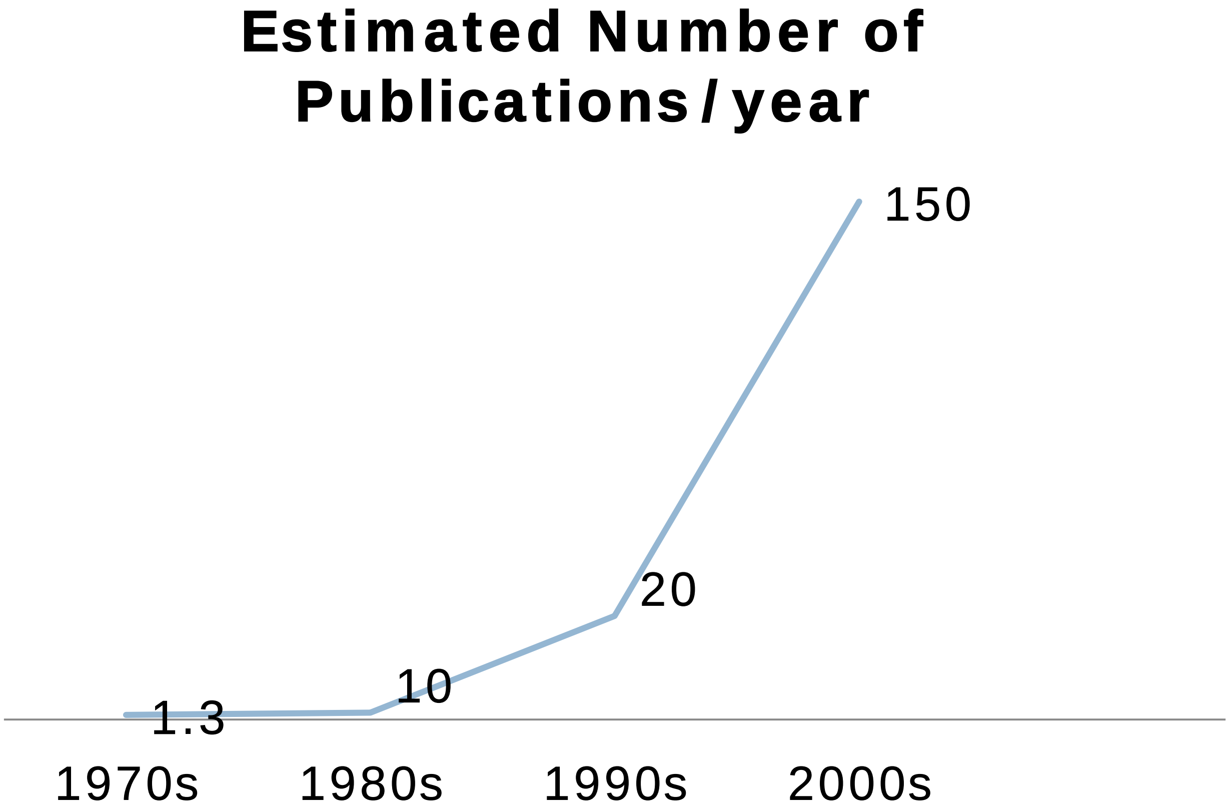
WORKSHOP ON NATURAL LANGUAGE PROCESSING: STATE OF THE ART, FUTURE DIRECTIONS AND APPLICATIONS FOR ENHANCING CLINICAL DECISION MAKING

Carol Friedman

Department of Biomedical Informatics, Columbia University

NLP in the Biomedical Domain

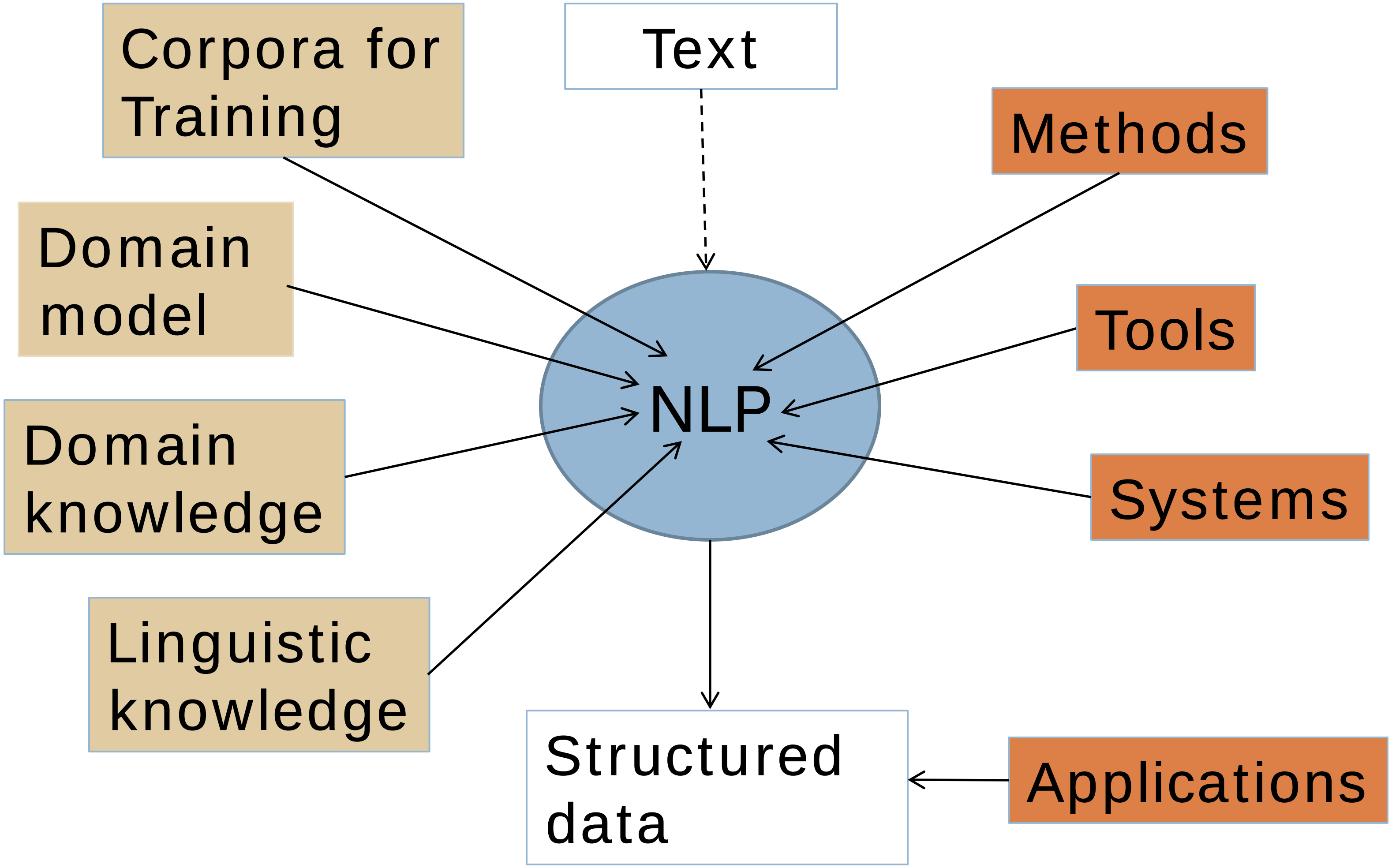


Goal of NLP Workshop

Identify

- Achievements
- Critical challenges
- Recommend future directions

Aspects of NLP



Corpora for Training

Text

Methods

Domain model

Tools

NLP

Domain knowledge

Systems

Linguistic knowledge

Structured data

Applications

Applications: clinical

- Patient care
  - Decision support, quality measures, coding, reduce errors, improve documentation, health information exchange
- Secondary data use
  - Clinical trial recruitment
  - Identify phenotypes
  - Knowledge acquisition and discovery
- Summarization
- Translation
- Tailoring information for consumers
- Computer-generated explanations

Applications: Biomedical

- Improve access to information in text, on Web
- Facilitate curation
- Knowledge acquisition
- Integration of knowledge from multiple sources and disciplines
- Question answering
- Summarization

BioNLP Milestones

- 1960s-70s: Start of clinical NLP
- 1970s, 1980s: Feasibility of structuring clinical information
  - Sager – comprehensive NLP system
- Early 1990s: Demonstration that NLP could be used to improve care
  - Haug (*Symtext*: rule-based syntactic, statistical semantics)
  - Friedman & Hripcsak (*MedLEE*: rule-based semantic/syntactic)

BioNLP: important clinical NLP

- Early-mid 1990s
  - Chute, Elkin: compositionality, terminology, ontology, & NLP
  - Baud, Scherrer, & Rassinoux: ontology-driven semantics, multi-lingual NLP
  - Hahn: Discourse analysis, ontology-based NLP
  - Zweigenbaum: Ontology-driven, semantic analysis of terms

BioNLP Milestones

- Côté RA, Rothwell DJ: SNOMED-standardizing structure of medical language (1980s)
- NLM
  - Lindberg DA, Humphreys BL: UMLS, a critical knowledge source for medical informatics and NLP (late 1980s)
  - McCray: Specialist system: NLP system(early 1990s)
    - McCray, Browne - comprehensive medical lexicon
  - PubMed: Abstracts and MeSH annotations

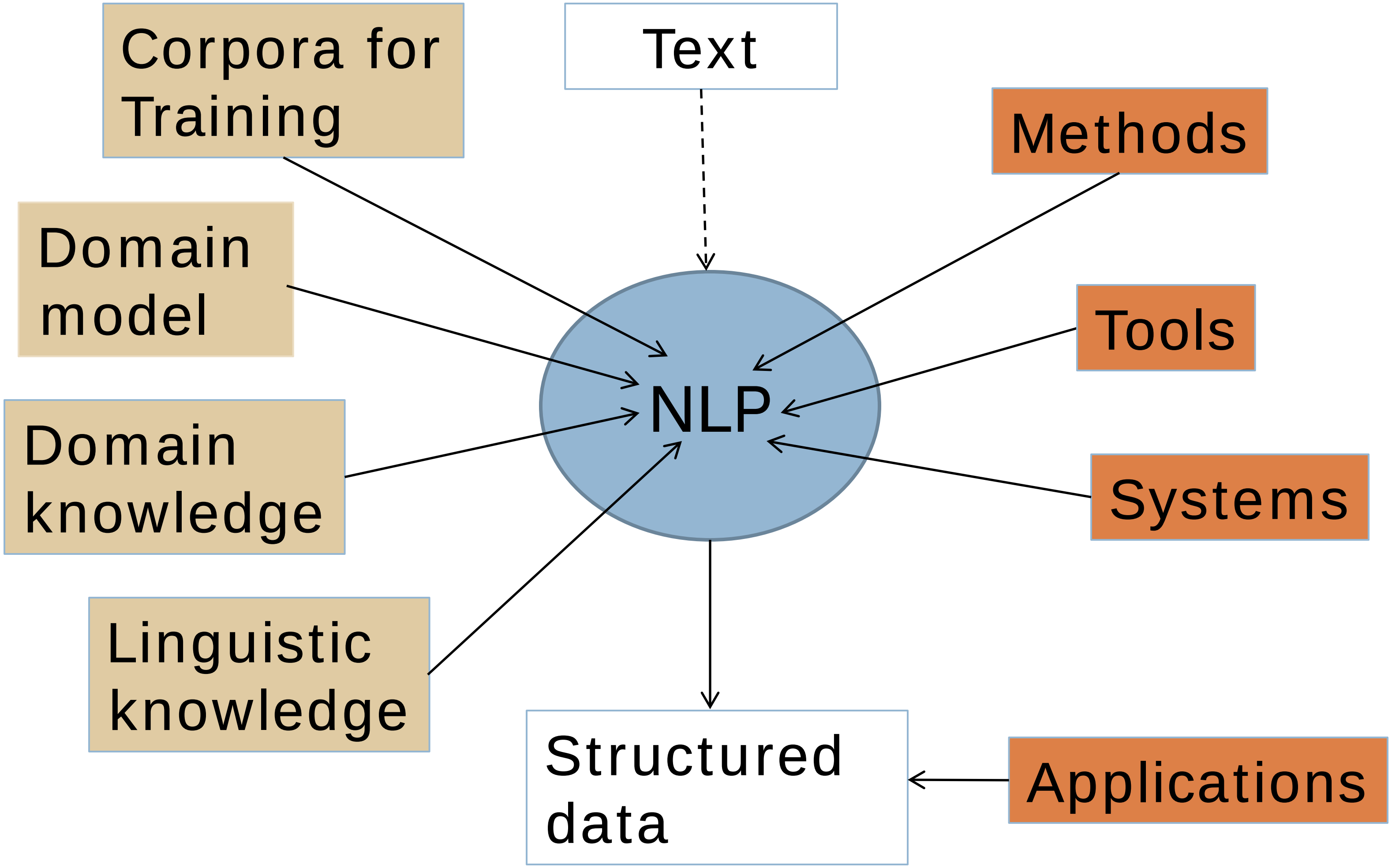
BioNLP Milestones: genomics literature

- NLP in biomolecular domain: named entity recognition, molecular relations, connecting information
  - Late 1990s: Tsujii, Park, Rindflesch, Aronson, Hunter
  - Early 2000s: Rzhetsky, Wong, Raychaudhuri
- Corpora/challenges
  - GENIA corpus: Tsujii
  - BioCreative challenges: Hirschman, Valencia
  - TREC Genomics Track: Hersh
  - BioNLP workshops & challenges

BioNLP Milestones - tools

- MetaMap (Aronson): text to UMLS concepts
- SemRep (Rindflesch): extraction of predications
- Open Source NLP clinical systems
  - NegEx & ConTEXT (Chapman): negation detection expanded to detection of temporality, experiencer
  - caTIES (Crowley): pathology diagnoses
  - cTAKES (Savova, Chute): general information extraction of clinical notes
  - Orbit Project: biomedical informatics tools
    - orbit.nlm.nih.gov

Aspects of NLP



Corpora for Training

Text

Methods

Domain model

Tools

NLP

Domain knowledge

Systems

Linguistic knowledge

Structured data

Applications

General Language Linguistic Knowledge/Tools/Corpora

- Natural Language Tool Kit (NLTK)
  - www.nltk.org
- LingPipe
  - www.alias-i.com/lingpipe
- OpenNLP
  - incubator.apache.org/opennlp
- UIMA
  - uima.apache.org
- Chris Manning's list of resources
  - www-nlp.stanford.edu/links/statnlp.html

Domain Linguistic Knowledge: Lexical

- NLM Resources
  - UMLS Metathesaurus: domain terms
  - UMLS Semantic Network: semantic categories
  - UMLS Specialist NLP tools
  - NCBI resources: biomolecular, species, ...

- OBO (Open Biological and Biomedical Ontologies)

Domain Models

- Critical for interoperability, sharing, and health information exchange
- Models for concepts
- Models for relations

Domain Concept Models

Many domain ontologies/terminologies

- UMLS containing >160 sources
  - MeSH
  - SNOMED
  - RXNORM
  - ICD-9
  - LOINC
- Open Biological and Biomedical Ontologies (gene ontology, cell ontology, chemical, phenotype, disease, ...)

Domain Models of Relations

Clinical domain: represent concepts and their modifiers/qualifiers

- Canon effort
- Galen effort
- Clinical Element Model (Sharp, I2B2, QueryHealth,…)
- http://wiki.siframework.org/

Domain Models of Relations

Biomedical Domain: predicate-argument (PAS) representational models

- Predicates and Arguments with semantic roles
- Models for specific verbs (PASBio, BioProp)
- SemRep predications
  - Based on 26 UMLS relations (causes, disrupts, treats, ...)

Domain Specific Purpose Models

- Representing specific types
  - Guidelines/Clinical Trials
    - EON, GLIF, Arden
- Representing Temporal Data
  - TimeML
  - Temporal constraint structure

Annotated Domain Corpora: Biomedical Literature

- PubMed – MeSH
- GENIA – semantic, syntactic, entities, relations
- BioCreAtIvE: annotated for realistic tasks
  - gene, protein mentions/ normalization/molecular interactions/cross-species
- PASBio,BioProp: predicate-arguments for specific verbs
- BioScope, BioInfer: negation, uncertainty & scope (some clinical)
- WSD, MSH WSD test collections: annotations of 50 & 203 ambiguous terms

Domain Corpora: Raw Clinical Documents

- Cincinnati Children’s Hospital
  - De-identified pediatric corpus
- Pittsburgh
  - De-identified reports from multiple hospitals
- MIMIC
  - Longitudinal de-identified reports
    - 26,000 patients in ICU setting
    - > 1 million notes
    - Discharge summaries, ECG/echo/radiology reports, and doctor and nursing notes
    - ICD-9 codes

Domain Corpora: Annotated Clinical Documents

- Cincinnati’s Children Hospital
  - Radiology reports: ICD-9 coding annotations
- I2B2 Challenges (2007-2012)
  - De-identified discharge summaries: annotated for various challenges
- TREC Medical Records Track

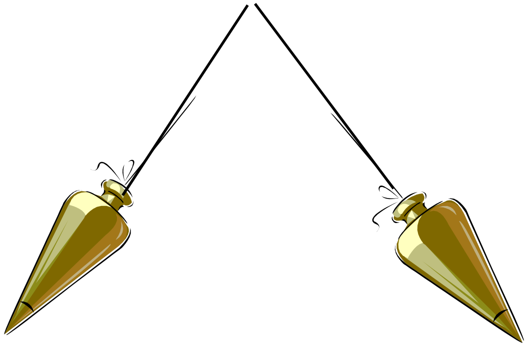
Challenges & Future Directions

Issues/Future Directions

- Access to more clinical notes & larger variety
- New methods vs. incremental methods
- More varied applications
- Evaluation
  - Important to learn from results
  - Some tasks more difficult than others: Why?
    - General vs. specific task
    - NLP issues vs. other reason
    - Domain reasoning

Issues/Future Directions: Linguistic Trends

Empirical corpus-based (before late 1950s)



Manual rule-based, linguistic-expertise (late 1950-late 1980s)

Statistical corpus-based (late 1980s–present)

Issues/Future Directions: Development of hybrid methods

Advantages of statistical methods

- Automated detection of textual patterns possible
- Many machine learning (ML) tools available
- Annotation & tools enable
  - Rapid implementation
  - Implementation without linguistic expertise
- Easy to experiment with different features, ML methods

Issues/Future Directions: Development of hybrid methods

Some disadvantages also

- Annotation is costly
- Performance depends on having similar corpora
- Statistical patterns are not intuitive
- Error analysis difficult to perform
- Errors cannot be rapidly fixed
  - Requires more annotated text or
  - Changes in method

Issues/Future Directions: Development of hybrid methods

Need synergistic models

- Methods that integrate
  - Expert rules
  - Domain knowledge
  - Machine learning
- Methods that allow experts to overrule
- More linguistically intuitive

Issues/Future Directions: Lexical knowledge in clinical domain

Identifying senses of abbreviations clinicians use

- Not defined in reports, often contain 2-3 letters
- Typical
  - Ca (*cancer*, *calcium* as measurement, *calcium* as medication)
  - PD (*Parkinson disease*, *primary care physician, peritoneal dialysis, pancreatic duct*)
- Atypical
  - HF
  - RH
  - b4

Issues/Future Directions: Word sense disambiguation

- Critical and difficult problem
- Large number of ambiguous words
- Performance varies for individual ambiguous words
  - Local vs. global vs. contextual vs. knowledge-based features

Issues/Future Directions: Domain Models

- Continue representational modeling work
  - Include rich features that affect meaning/use
  - Expand predicate-argument relations in clinical domain
  - Evaluate models for accuracy & coverage based on real text

Future Directions: Balance & Broaden NLP research portfolio

- Improve data entry
  - Reduce use of abbreviations
  - Reduce cut/paste
  - Improve template creation and use
- Improve EHR documentation
- Develop cutting-edge applications
- Summarization
- Question-answering
- Improve access to information for consumers
- Knowledge acquisition, integration, and discovery

Issues/Future Direction

Keep up the momentum!