Information Extraction from Dirty Notes for Clinical Decision Support

Clare T Grasso, UMBC
Anupam Joshi, PhD (Advisor)
Eliot Siegel, MD
Requirements

- Information Retrieval & Extraction
  - High precision
    - Top 1-2 Ranking
  - High recall

- Clinical Decision Support
  - Near real-time performance constraints
    - 3 seconds or less
    - Over thousands of pages of text
Semantic Search
Use Case: Pain

Attributes:
• Location
• Severity
• Quality/Type

• Onset
• Duration

Patient came in complaining of abdominal pain. Symptoms started 2 weeks ago, sudden, usually lasts intermittently. He rates the pain as 8/10 with zero being no pain and 10 being worst possible pain. Pain is described as aching, shooting, squeezing and throbbing.

Dirty Notes

Characterized by nonstandard use of:
Abbreviations
Punctuation
Capitalization
White space
Etc.

1. CHIEF COMPLAINT:
ABD PAIN @ OLD SURGICAL SITE, ONSET THIS AM, DENIES SOB, DENIES C/P, PT. THOUGHT PROCESS OFF/CONFUSING THIS THOUGHTS, DR. XXXXXX MADE AWARE PT IN ACUTE AREA BED #5
Data Properties

- Training Data
  - 1500 pages @ 10 point font
  - 1-2% of the lines contain data relating to patient pain
Using IR with Machine Learning

- Vectorization
  - TF.IDF
  - Cosine normalization
- Query Terms
  - “learned” by ML Algorithms
- Performance Efficiency
  - 0.6 seconds
Machine Learning Results

- Naive Bayes
- SVM Linear
- SVM RBF
- SVM Poly
- LogReg
- SGD

Bar chart showing precision, recall, and F-score for different machine learning models.
Current Research Approach

• Proximity-Weighted Scoring
  • Score relevancy of each document by how many query terms appear within a set window size of $\omega$.

• Ontology Driven Retrieval
  • Ontology specifies concepts and how they relate to each other (eg, is-a, has-a).
  • Concepts are expressed in words and can be used as query terms
  • Relationships can be used to generalize / specialize the query terms and results
Anchor-Proximity IR

• Develop ontology for pain
• Define anchor terms = [‘pain’, ’painful’]
• Retrieval Algorithm
  • Use a sliding window $\omega = 10$ terms
  • Relevance Score is:
    • number of ontology terms in the document
    • that are within $\omega$ terms of an anchor term
Anchor-Proximityity Results

Proximity vs. Anchor-Proximity IR

$\omega = 10$
Questions?

THE LD$_{50}$ OF TOXICITY DATA IS 2 KILOGRAMS PER KILOGRAM.