

# Beyond Dictionary Match for Disease Named Entity Recognition

Robert Leaman

Department of Computer Science

Chris Miller & Graciela Gonzalez

Department of Biomedical Informatics

Arizona State University, Tempe AZ



# Motivation

- Named entity recognition (NER)
  - Find all entities of a specified type (diseases, genes, drugs, etc) in a given text
  - Not trivial: many terms ambiguous
    - Neurofibromatosis 2 (Nf 2); AD
    - Dictionary insufficient
- Building block task
  - e.g. gene/disease relationship extraction
  - High performance required
- State of the art technique: sequence-based machine learning e.g. conditional random fields



# Related Work & Corpus

- Disease corpus by Jimeno et. al. (EBI)
  - Evaluation centered on dictionary matching
  - Did not contain mention locations - *required for applying sequence-based machine learning*

Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics. 2008 9(Suppl 3):S3.

- Our effort: update corpus
  - Re-annotate for mention location
  - Annotate corpus sentences not prev. annotated
  - Currently 2427 sentences, 3367 disease annotations

# Methods

- BANNER - biomedical NER system
  - Portable, trainable (CRF), configurable
  - Open source: <http://banner.sourceforge.net>
- Disease dictionary
  - Extracted from UMLS MetaThesaurus
  - 167,944 entries
  - Added to BANNER as an input feature
- Evaluation by 5x2 cross-validation

# Evaluation Results

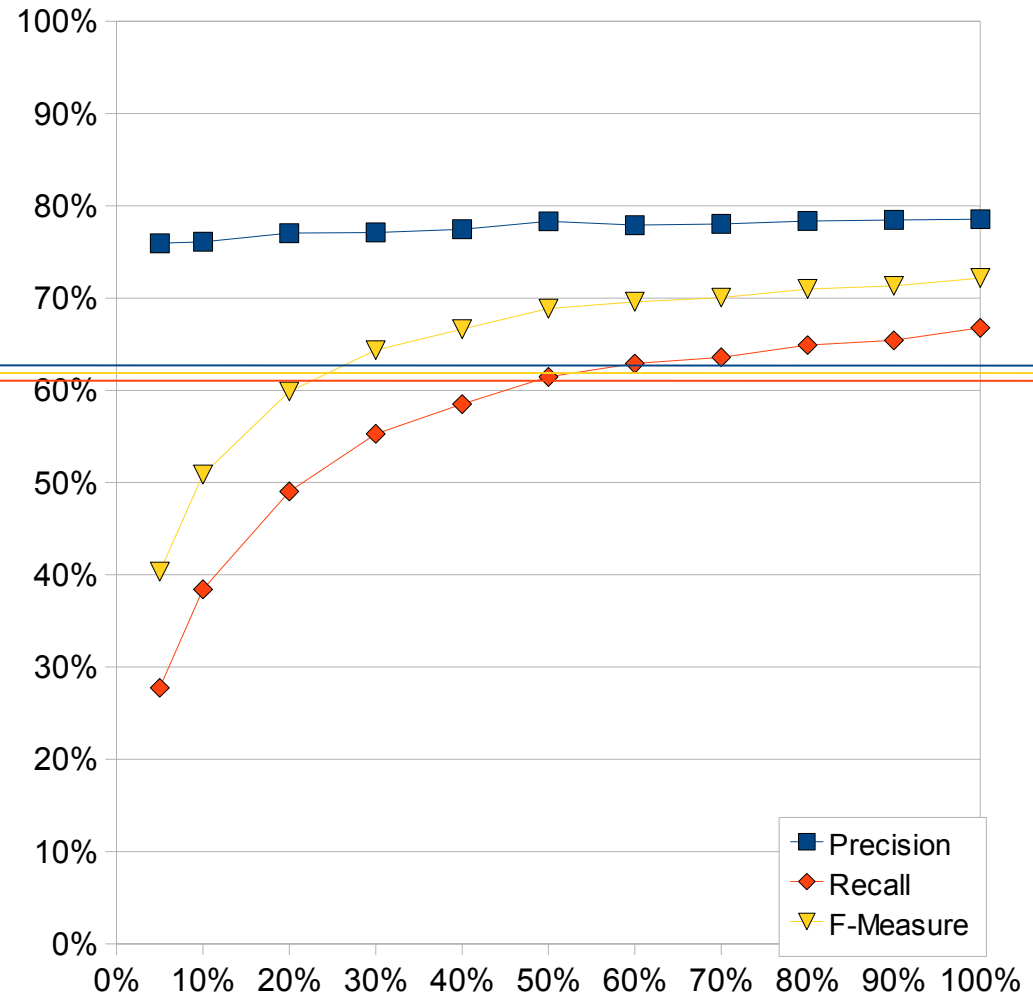
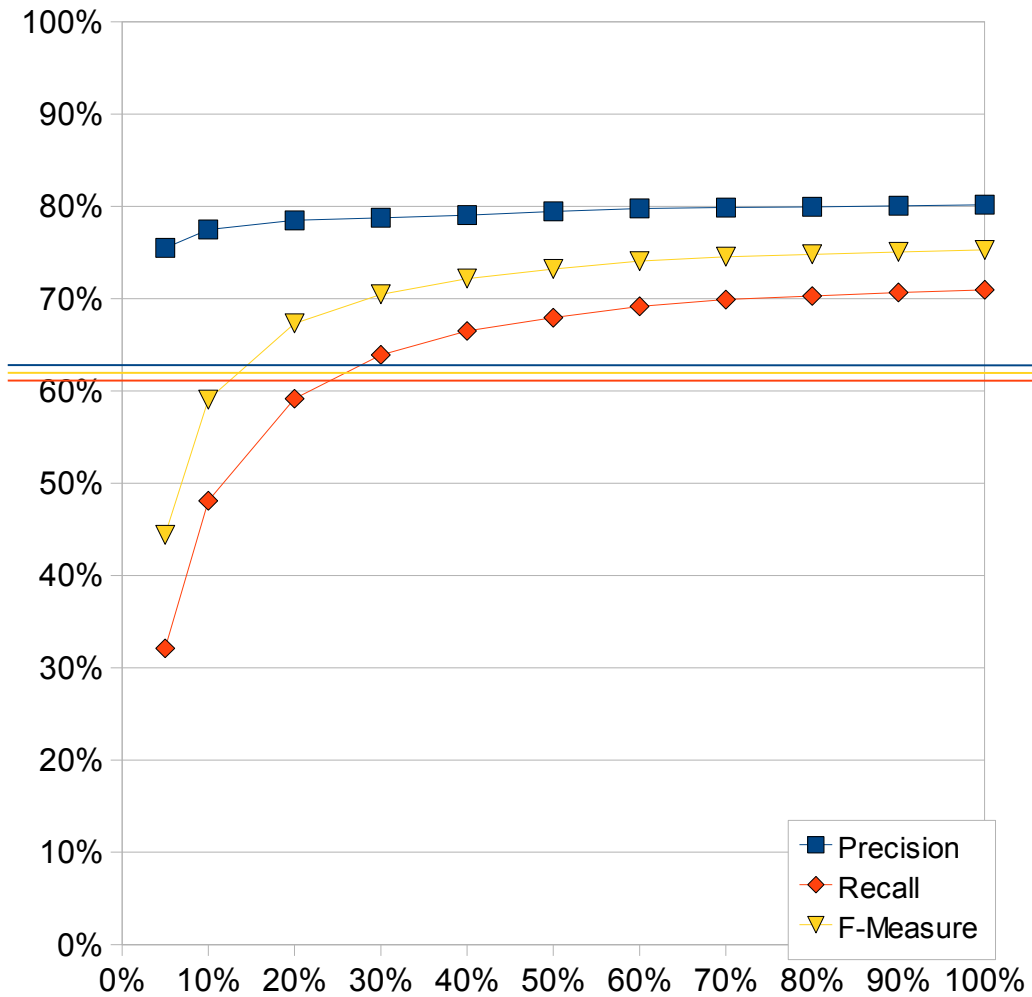
<b>System</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
BANNER with UMLS-based dictionary	80.78	70.66	75.36
- with CRF order = 1	80.17	70.95	75.27
- without UMLS-based dictionary	78.55	66.77	72.16
UMLS-based disease dictionary	62.72	61.78	62.24
MetaMap (Jimeno et.al.)	31.27	30.07	30.66
Statistical (Jimeno et.al.)	26.07	30.18	27.97

# Ablation Study

How many sentences do we need?

## BANNER, with dictionary

## BANNER, without dictionary



Percentage of training data used