CS54701: Information Retrieval

Natural Language Processing in IR

7 April 2016

Prof. Chris Clifton

Indiana Center for Database Systems

1

Need for NLP

- Vector space model limitations
  - Words in combination carry more/different meaning than isolation
  - President flew
    - to Washington
    - from the Revolution
- Words can mean different things
- Relative importance of different words
- Words vs. Concepts

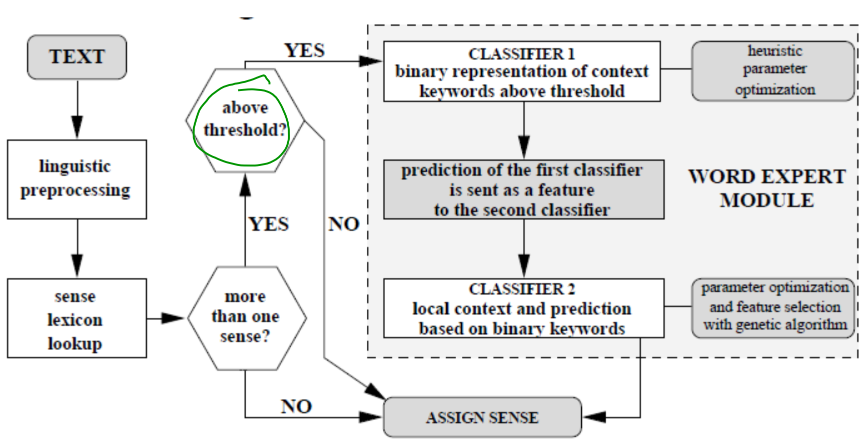
2

Different meanings

- NLP Task: *Word Sense Disambiguation*
  - Given word, dictionary of multiple meanings
  - Determine from context which meaning applies
- Hard problem
  - SensEval 3 (2004): 65% accuracy

3

"Winner": GAMBL
Decadt, Hoste, Daelemans, Bosch



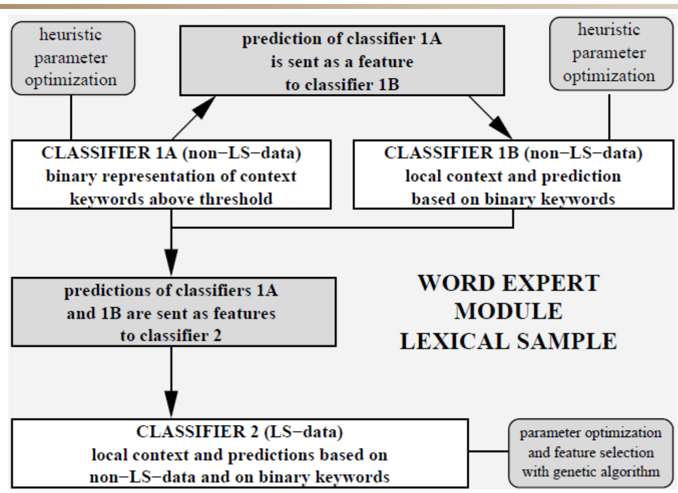
4

"Winner": GAMBL
Decadt, Hoste, Daelemans, Bosch

- Initial phase: Linguistic analysis
  - Tokenize
  - Part-of-speech
  - Grammatical relations
- Training data
  - Senseval-3 task (7860 words)
  - SemCor (WordNet), previous SenseEval (555,269 words)

5

"Winner": GAMBL
Decadt, Hoste, Daelemans, Bosch



6

“Winner”: GAMBL
Decadt, Hoste, Daelemans, Bosch

- Cascaded Classifiers
  - First stage: Broad context
    - Three sentences
    - Instance-based learning
  - Second stage: Narrow context
    - Seven words
    - Result of 1st classifier
    - Genetic algorithm

7

Different meanings

- NLP Task: *Word Sense Disambiguation*
  - Given word, dictionary of multiple meanings
  - Determine from context which meaning applies
- Hard problem
  - SensEval 3 (2004): 65% accuracy
    - “just choose most frequent sense” 60%
    - Inter-annotator agreement 72.5%

8

Words vs. Concepts

- Named Entity Recognition
  - People
  - Places
  - Organizations
  - Dates
  - …

  Success story – effective, learn new types of NER
- Coreference Resolution
  - Different names for same entity in same document

9

NER – CoNLL 2003 Winner
Florian, Ittycheriah, Jing, Zhang

- Label each word
  - Start, continue, or end a named entity
- Key: good features
  - Words and part of speech, 5 word window
  - Prefix, suffixes of surrounding words
  - Word "flags" such as *firstCap*, *2digit, allCaps*
  - Gazetteer – 130k known names
  - Output of existing NER systems trained for different output categories

10

Winner: Ensemble
Florian, Ittycheriah, Jing, Zhang

- Multiple classifiers
  - Robust risk minimization
  - Maximum entropy
  - Transformation-based learning
  - Hidden Markov model
- Weighted voting
- Results: 89% accuracy
  - Baseline 60%

11

Template Analysis

Named Entity Recognition on Steroids

- Given a "template" of desired structured information
  - Fill in fields of template from analysis of document
- Fields:
  - Entities (named entities)
  - Relationships
  - Time/date/order

12

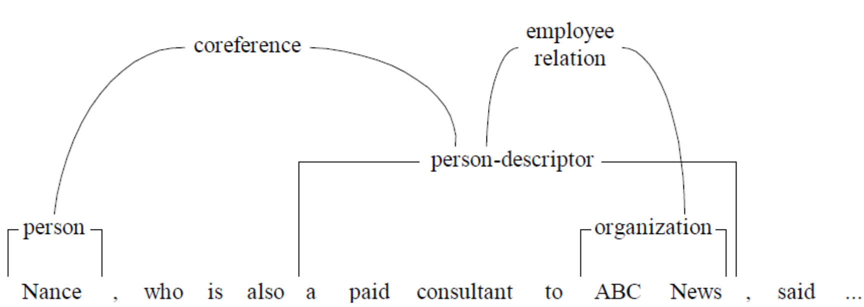
Template Analysis: Example

NAME:	Fletcher Maddox Maddox
DESCRIPTOR:	former Dean of the UCSD Business School his father the firm's CEO
CATEGORY:	PERSON
NAME:	Oliver
DESCRIPTOR:	His son Chief Scientist
CATEGORY:	PERSON
NAME:	Ambrose
DESCRIPTOR:	Oliver's brother the CFO of L.J.G.
CATEGORY:	PERSON
NAME:	UCSD Business School
DESCRIPTOR:	

PERSON	Employee_of	ORGANIZATION
Fletcher Maddox Fletcher Maddox Oliver Ambrose	Employee_of Employee_of Employee_of Employee_of	UCSD Business School La Jolla Genomatics La Jolla Genomatics La Jolla Genomatics
ARTIFACT	Product_of	ORGANIZATION
Geninfo	Product_of	La Jolla Genomatics
LOCATION	Location_of	ORGANIZATION
La Jolla	Location_of	La Jolla Genomatics
CA	Location_of	La Jolla Genomatics

13

Message Understanding Conferences



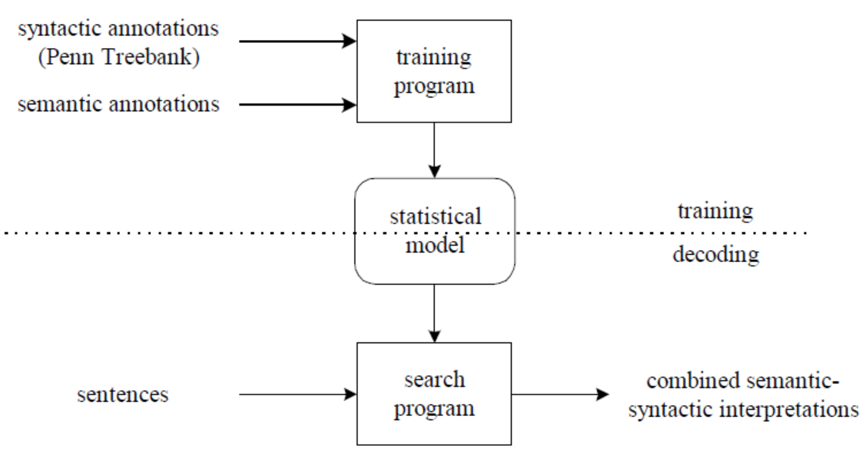
14

SIFT: *Miller, Crystal, Fox, Ramshaw, Schwartz, Stone, Weischedel*

- Language model approach
  - Uses Hidden Markov Models
- Sentence-level model
  - Part of speech
  - Named Entity
  - Parse (grammatical)
  - Relationships
- Uses "outside" training data
  - Penn Treebank, additional domain-specific text

15

SIFT: Sentence Model



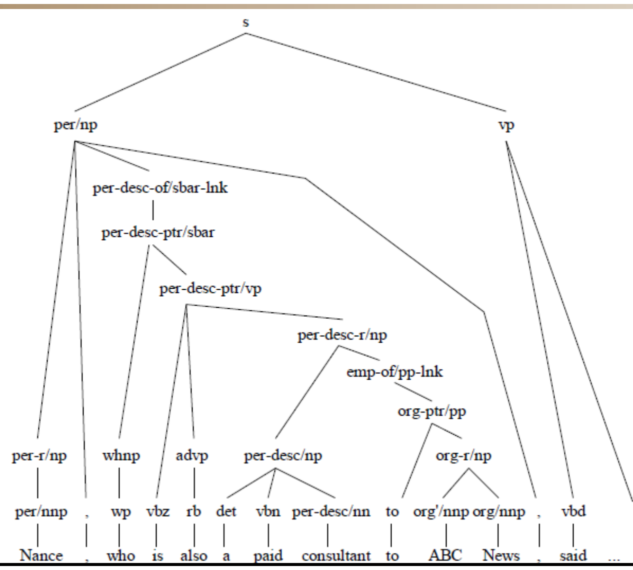
16

SIFT: Additional semantics

- Further breakdown (e.g., distinguish title from name in Named Entity)
- Semantic labeling
- Co-reference
- *Probability labels for all of these*

17

SIFT: Sentence-level output



18

Cross-Sentence Model

- Similar approach
- Uses sentence parse/labeling as input

19

Basic Tools

- Part of Speech tagging
- Sentence diagramming

20