Lister Hill National Center
for Biomedical Communicatio0ns

Development of a Semantic Type Based WSD Tool

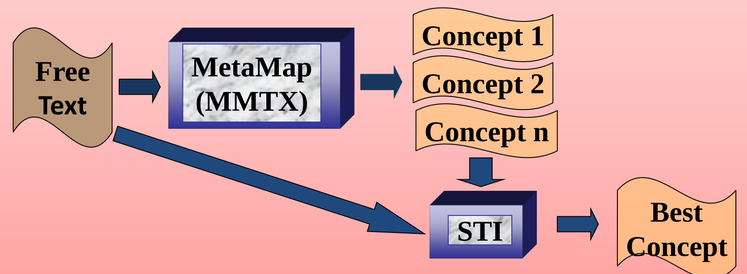
Chris J. Lu, Ph.D., Susanne M. Humphrey, Willie J. Rogers, Allen C. Browne

Text Categorization
http://specialist.nlm.nih.gov/tc

Objective:

Develop an ST-based WSD tool, ST-WSD, for distribution in the open source Text Categorization (TC) package

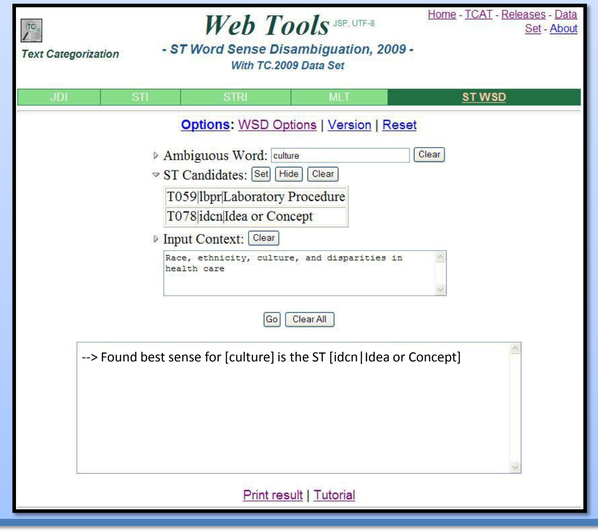
ST Word Sense Disambiguation:



Example:

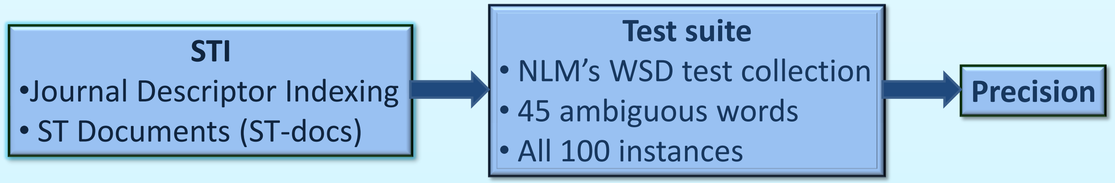
- **Free Text:**
  Race, ethnicity, culture, and disparities in health care
- **Ambiguous word:** culture
  - **Anthropological Culture - Idea or Concept**
  - Laboratory Culture - Laboratory Procedure

ST-WSD:



Approach:

I. Test suite: Find best ST Documents (precision on NLM's WSD test collection) through test suite



II. ST-docs Enhancement:

- Weighted Frequency (WF): use occurrence data in the form of weighted frequency instead of counting a word only once in ST-Docs
- 1 SG: words associated with only one Semantic Type Group (SG)
- Refined by STRI to choose good representative words:
  - Top n rule: must be ranked in the top n (15)
  - ST score must be within one standard deviation (StdDev) of the top score
- Add words associated with Multiple SGs (MSG) if the ST is in the top 3

III. ST-WSD Tool Enhancement:

- Combined scores (CS): select the ST with highest relative combined score (absolute difference between WC and DC based score)
- Ambiguous sentences filter option
- Force ambiguous word and its morphological variants to be legal words

IV. Results:

ID	ST Documents	Precision
A	Baseline - 2008	73.81%
B	Weighted Frequency (WF)	76.29%
C	WF – 1SG	76.85%
D	WF – 1SG: StdDev & Top 15	78.07%
E	WF 1SG: StdDev & Top 15; MSG:Top 3	78.71%
F	ST Document E witch CS	**79.05%**