INFERRING META-DATA

based on:

Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources

Tandeep Sidhu, Judith Klavans, and Jimmy Lin
College of Information Studies
University of Maryland
College Park, MD 20742

OUTLINE

1. Introduction
2. Sidhu et. al
   a) Resources
   b) Methods
   c) Results
   d) Analysis
3. Conclusion

1. INTRODUCTION

- Meta-Data = data about data
  - Type of Meta-Data
    - Technical..
    - Structural..
    - …
    - Guide/ Descriptive!
      is used to help humans find specific items and is usually expressed as a set of keywords in a natural language
- Challenge: Extract meta-data automatically
  - Problem: Ambiguity

    one word - at least two different senses

1. INTRODUCTION
AMBIGUITY - EXAMPLE

- Wings
  - **Sense#1**: Used for accessories that project outward from the shoulder of a garment and are made of cloth or metal.
  - **Sense#2**: Lateral parts or appendages of a work of art, such as those found on a triptych.
  - **Sense#3**: The areas offstage and to the side of the acting area.
  - **Sense#4**: The two forward extensions to the sides of the back on an easy chair.
  - **Sense#5**: Subsidiary parts of buildings extending out from the main portion.

2. SIDHU ET. AL

- extracting meta-data from descriptions
  - Domain: art and architecture
    - Special vocabulary!
  - using existing resources
    - Lexicon, data set, additional algorithms

developing a disambiguation algorithm

2. SIDHU ET. AL - RESOURCES

- domain-specific Lexicon
  - Art and Architecture Thesaurus (AAT)
- Data Set
  - National Gallery of Art (NGA) online archive
- Disambiguation Tool
  - SenseRelate AllWords with WordNet

2. SIDHU ET. AL - RESOURCES
ART AND ARCHITECTURE THESAURUS (AAT)

- widely-used multi-faceted thesaurus of terms
  for the cataloging and indexing of art, architecture, artifactual, and archival materials
- 31,000 structured records (= sense)

  unique ID, preferred name, record description, variant names, broader, narrower, and related terms
- Problem with ambiguity

# of Senses	# of Homonyms	Example
2	1097	bells
3	215	painting
4	50	alabaster
5	39	wings
6	9	boards

2. SIDHU ET. AL - RESOURCES
NATIONAL GALLERY OF ART ONLINE ARCHIVE

- collection of paintings, sculpture, decorative arts, and works from the Middle Ages to the present
- randomly selected 20 images with corresponding text from this collection

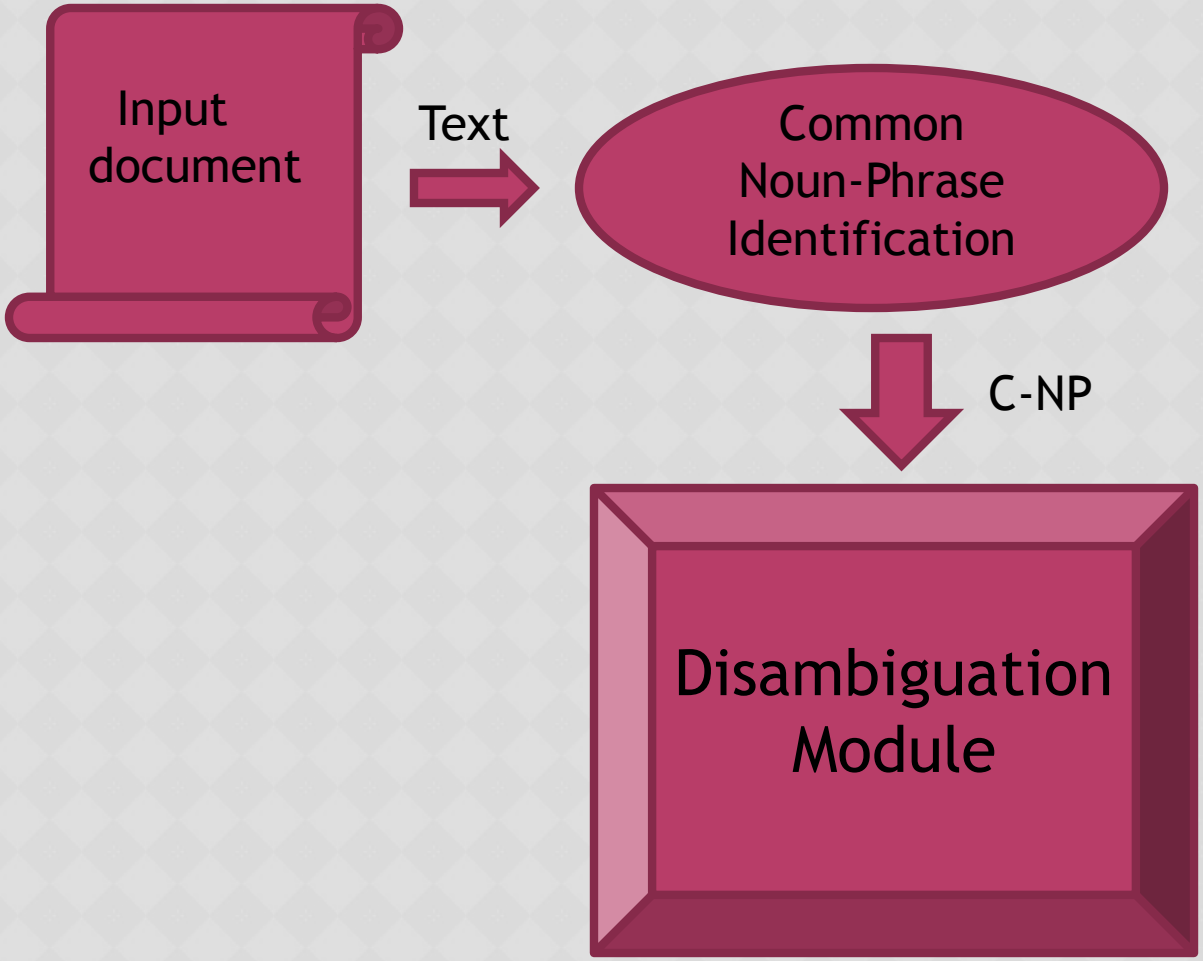
Term extraction

Training Data Set (128 ambiguous of 326 terms)	Test Data Set (96 ambiguous of 275 terms)

2. SIDHU ET. AL - RESOURCES
SENSERELATE ALLWORDS AND WORDNET

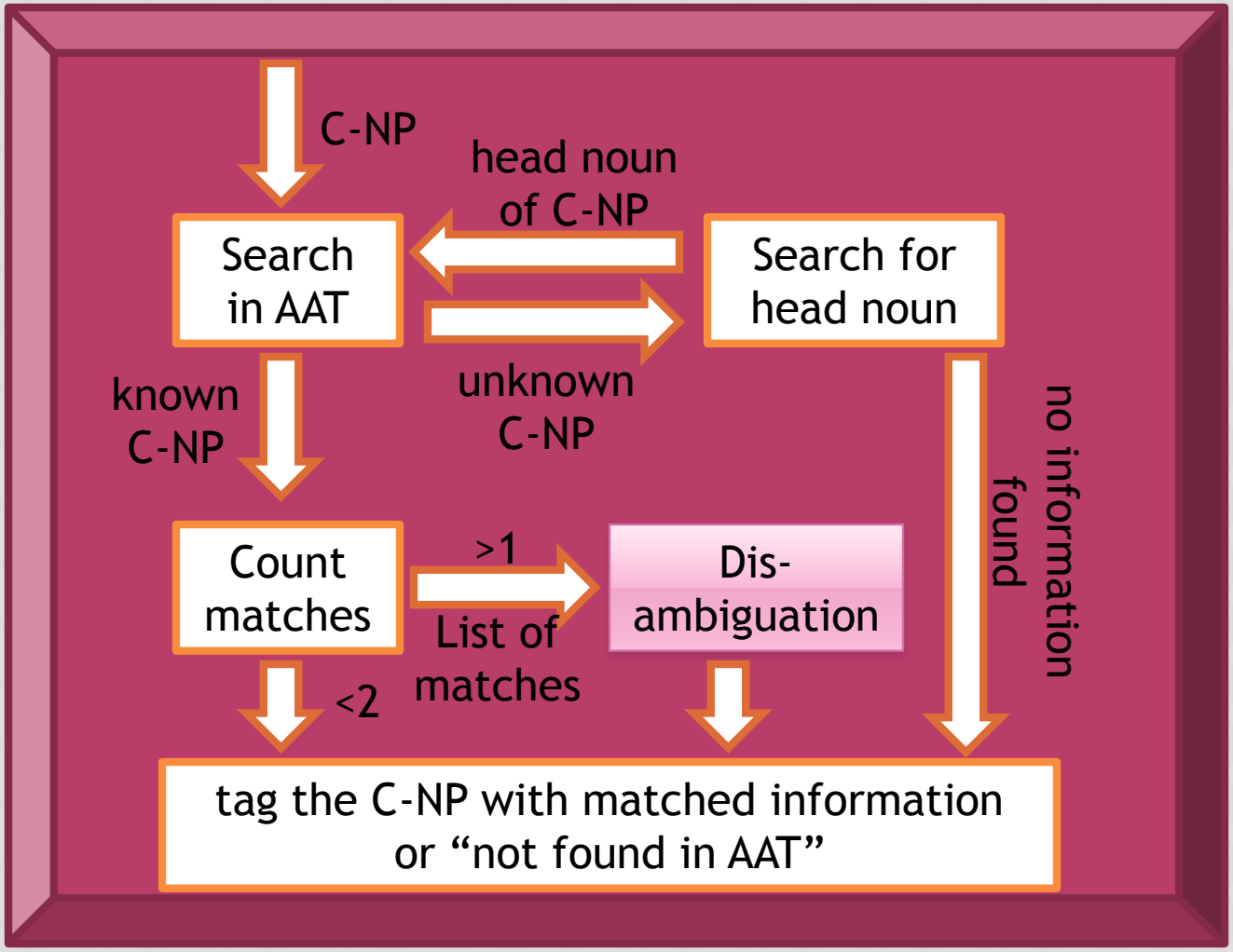
- SenseRelate AllWords
  - Disambiguates all the words in that sentence using word sense definitions from WordNet
- WordNet
  - a large lexical database of English nouns, verbs, adjectives, and adverbs

2. SIDHU ET. AL - METHODS
DISAMBIGUATION ALGORITHM



2. SIDHU ET. AL - METHODS

DISAMBIGUATION ALGORITHM



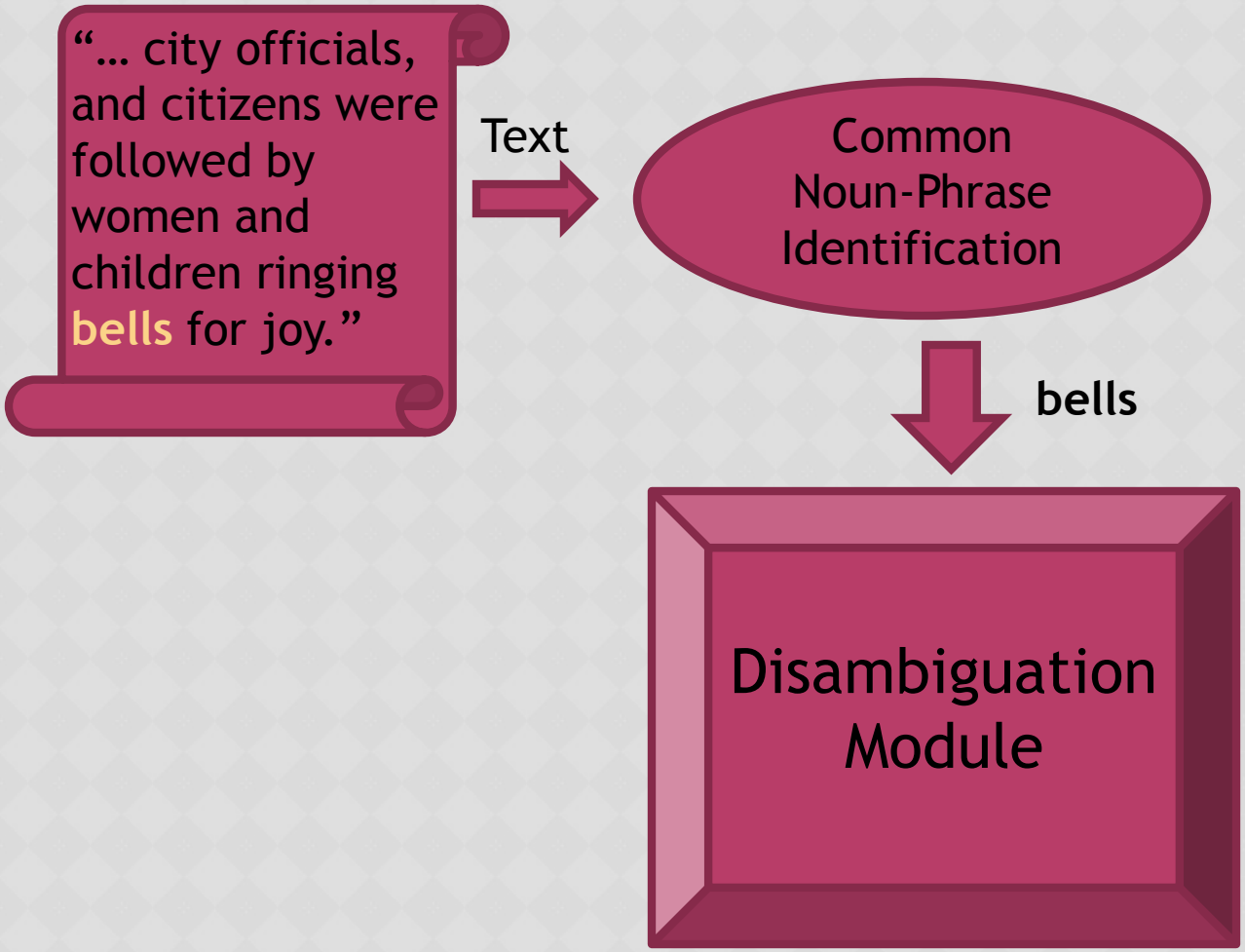
2. SIDHU ET. AL - METHODS
DISAMBIGUATION ALGORITHM

Techniques for Disambiguation

1. Lookup Modifiers
2. SenseRelate AllWords
3. Select Best Record Match
4. Use Most Common Sense from WordNet

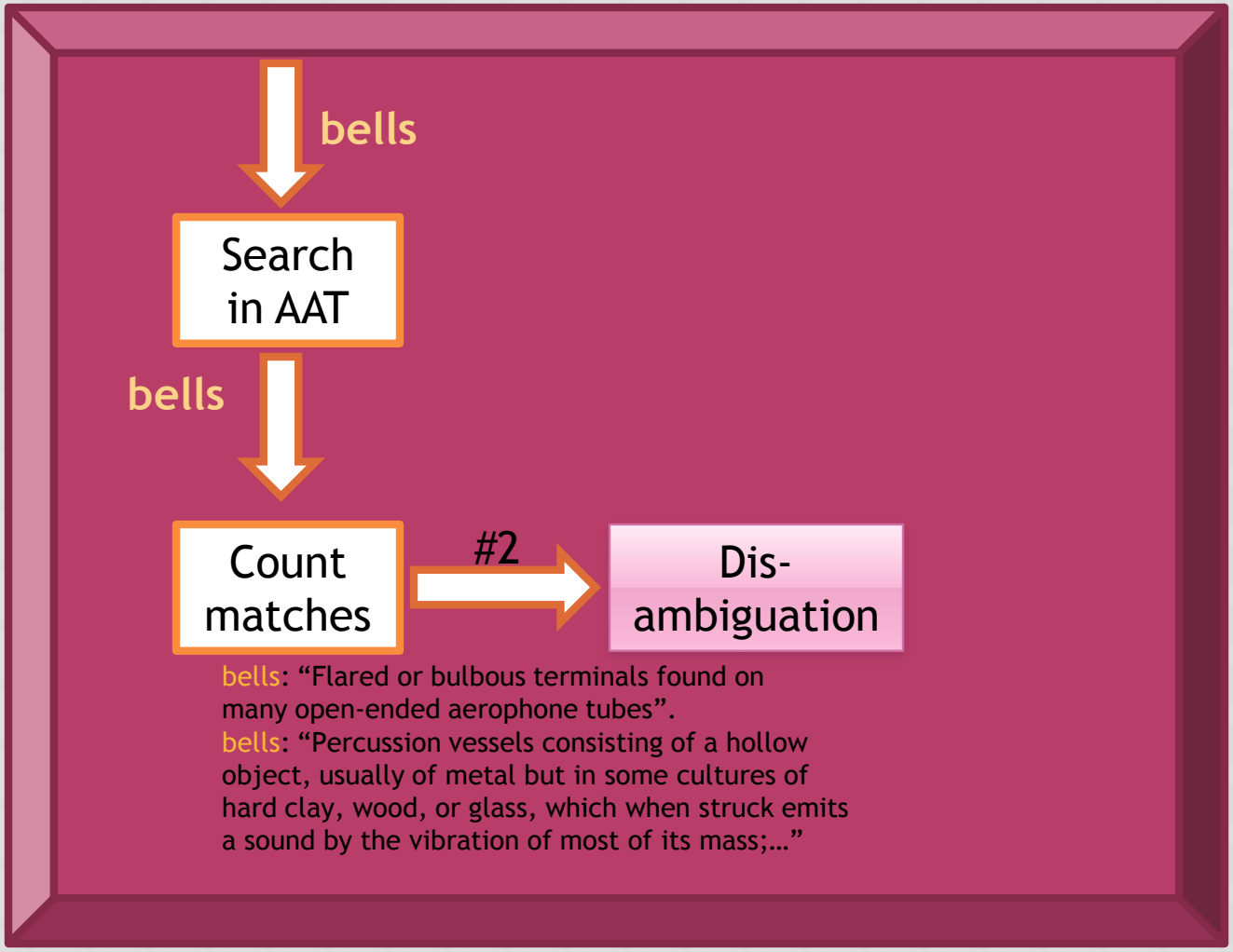
2. SIDHU ET. AL - METHODS

DISAMBIGUATION ALGORITHM - EXAMPLE



2. SIDHU ET. AL - METHODS

DISAMBIGUATION ALGORITHM - EXAMPLE



2. SIDHU ET. AL - METHODS
DISAMBIGUATION ALGORITHM

Disambiguation with SenseRelate:

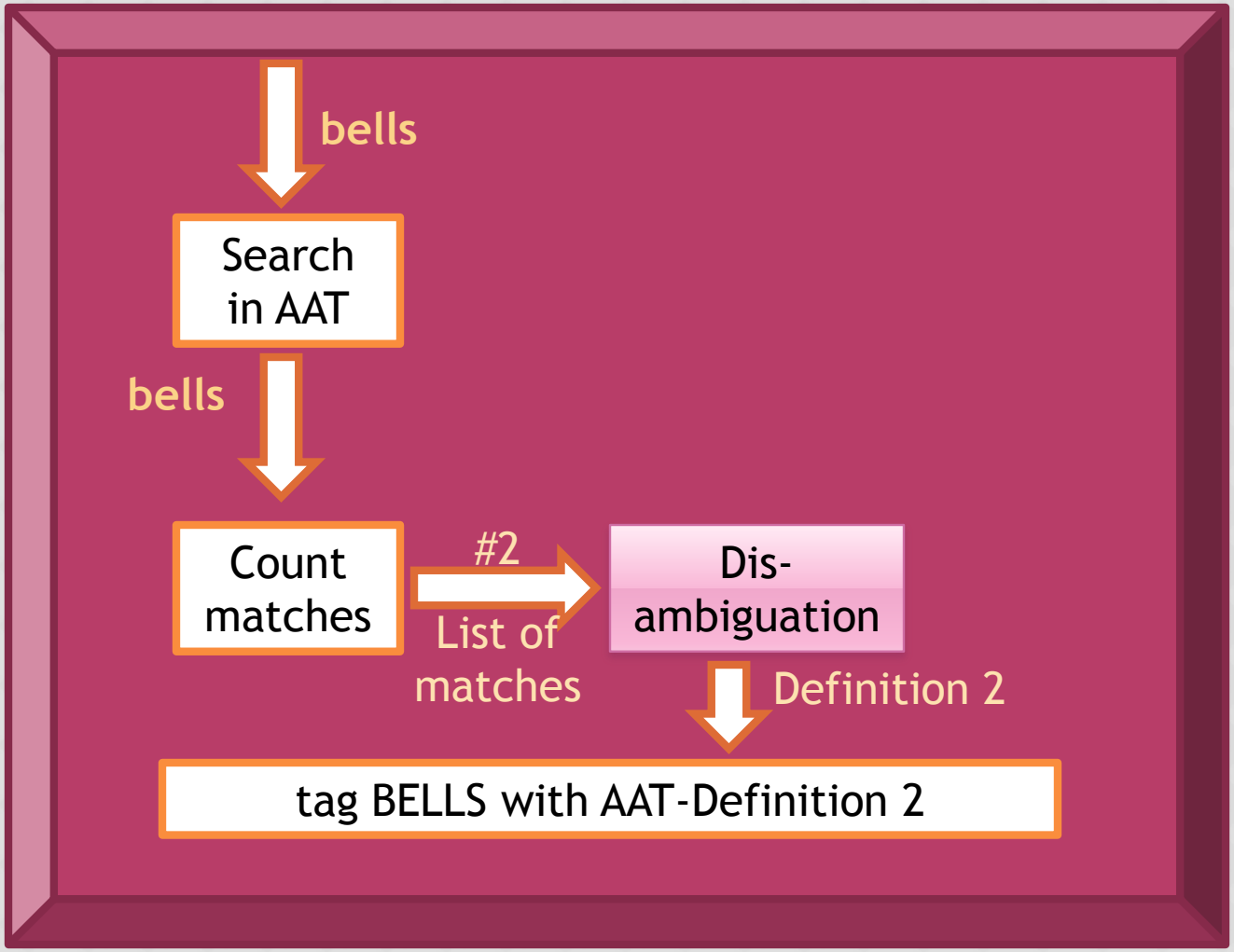
Output:
city#n#1 official#n#1 and citizen#n#1 be#v#1 follow#v#20 by#r#1 woman#n#1 and child#n#1 ringing#a#1 **bell#n#1** for joy#n#1

bells is WordNet-Sense1, "a hollow device made of metal that makes a ringing sound when struck"

Comparison	Score	Word Overlap
AAT – Definition 1 and WordNet Sense1	0	none
AAT – Definition 2 and WordNet Sense1	4	hollow, metal, sound, struck

2. SIDHU ET. AL - METHODS

DISAMBIGUATION ALGORITHM - EXAMPLE



2. SIDHU ET. AL - RESULTS

- Methodologies
- Overall Results
- results for Ambiguous terms

2. SIDHU ET. AL - RESULTS
METHODOLOGIES

- Evaluation
  - Correct AAT record is matched
  - Correct AAT record is in the top3 of the algorithm
  - Correct AAT record is in the top5 of the algorithm
  - Baseline - Without other disambiguation algorithm
  - Data set manually labeled by two different people
    = **ground truth**

2. SIDHU ET. AL - RESULTS

OVERALL RESULTS

Evaluation	Labeler 1		Labeler 2	
	training	test	training	test
Algorithm Accuracy	76%	74%	68%	73%
Baseline Accuracy	69%	72%	62%	69%
Top3	84%	79%	78%	79%
Top5	88%	81%	79%	80%

2. SIDHU ET. AL - RESULTS

RESULTS FOR AMBIGUOUS TERMS

Evaluation	Labeler 1		Labeler 2	
	training	test	training	test
Algorithm Accuracy	55%	50%	48%	53%
Baseline Accuracy	35%	42%	32%	39%
Top3	71%	63%	71%	68%
Top5	82%	68%	75%	71%

2. SIDHU ET. AL - ANALYSIS

USED TECHNIQUES

Row	Technique	Training	Test
One	Lookup M.	1	3
Two	SenseRelate	108	63
Three	Best Record	14	12
Four	Most Common	5	18

2. SIDHU ET. AL - ANALYSIS
OCCURRED ERRORS

Technique	Reason for Error	Error Count
SenseRelate	SenseRelate picked wrong WordNet sense	16
	WordNet does not have the sense	8
	Definitions did not overlap	11
	Other reasons	10
Best Record Match		10
Lookup Modifier		0
Most Common Sense		3

3. CONCLUSION

- Extracting automatically

  ➡ done, but not very efficient

- Exchange of variable components make sense
  - better disambiguation tool
  - better ground truth to compare
  - bigger data set

  ➡ better results possible

THANK YOU
FOR YOUR ATTENTION

REFERENCES

- Tandeep Sidhu; Judith Klavans; Jimmy Lin. Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: *Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH-07)*, 2007.