Projektseminar: Text Mining for Historical Documents
(WS 2010/11)

Inferring Meta-Data

Patricia Helmich

Basiert auf dem Paper: 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

Problem: mining text for image metadata

- Computational Linguistics for Metadata Building (CliMB) project:
  → improve image access by automatically extracting metadata from text associated with images (subject term acces)
- Part of this main problem: word sense disambiguation
  → avoid leading the image searcher to a wrong image as a result of ambigous metadata
  → subject of this presentation
- Domain: art and architecture domain (highly specialized technical vocabulary)
- Disambiguation algorithm: tries to choose the correct sense of nouns in textual descriptions of art object (with respect to a domain-specific thesaurus: the Art and Architecture Thesaurus (AAT))

Word Sense Disambiguation

- Basic challenge in computational linguistics
- Task: mining scholarly text for metadata terms
  → Word Sense Disambiguation: clarify ambigous terms
- Development of an algorithm that takes noun phrases and assigns a sense to the head noun or phrase
- Hypothesis: Accurate assignment of senses to metadata index terms will result in higher precision for searchers
- Finding subject terms and mapping them to a thesaurus:
  → time-intensive task for catalogers
  → automate this task
- Manual disambiguation would be slow, tedious and unrealistic

Resources

- The Art and Architecture Thesaurus (AAT)
  - a widely-used multi-faceted thesaurus of terms for the domain of art, architecture, artifactual and archival materials
  - each concept is described through a record with a unique ID, the preferred name, the record description, variant names, broader, narrower, and related names
  - 31,000 records in total, and 1,400 homonyms (records with same preferred name)
  - In this context: record ≈ sense
  - Two tasks addressed with the algorithm:
    - primary focus on: mapping a term to the correct sense in the AAT
    - The task of selecting amongst closely related terms in the AAT is handled with a simply ranking approach

Resources

- The Test Collection
  - The data set used for the evaluation of the algorithm
    - → from the National Gallery of Art (NGA) online archive
    - → covers paintings, sculpture, decorative arts, works from the Middle Ages to the present
  - 20 images randomly selected with corresponding text
    - → extracted noun phrases form the data set
    - → data set divided in two parts:
      - Trainings set: 326 terms (train the algorithm)
      - Test set: 275 terms (evaluate the algorithm)
  - A groundtruth for the data set is created manually by two labelers
    - → assign an AAT-ID to each term
    - → terms not appearing in the AAT were given an AAT record value of zero
  - Interannotator agreement was pretty high (85%)

Resources

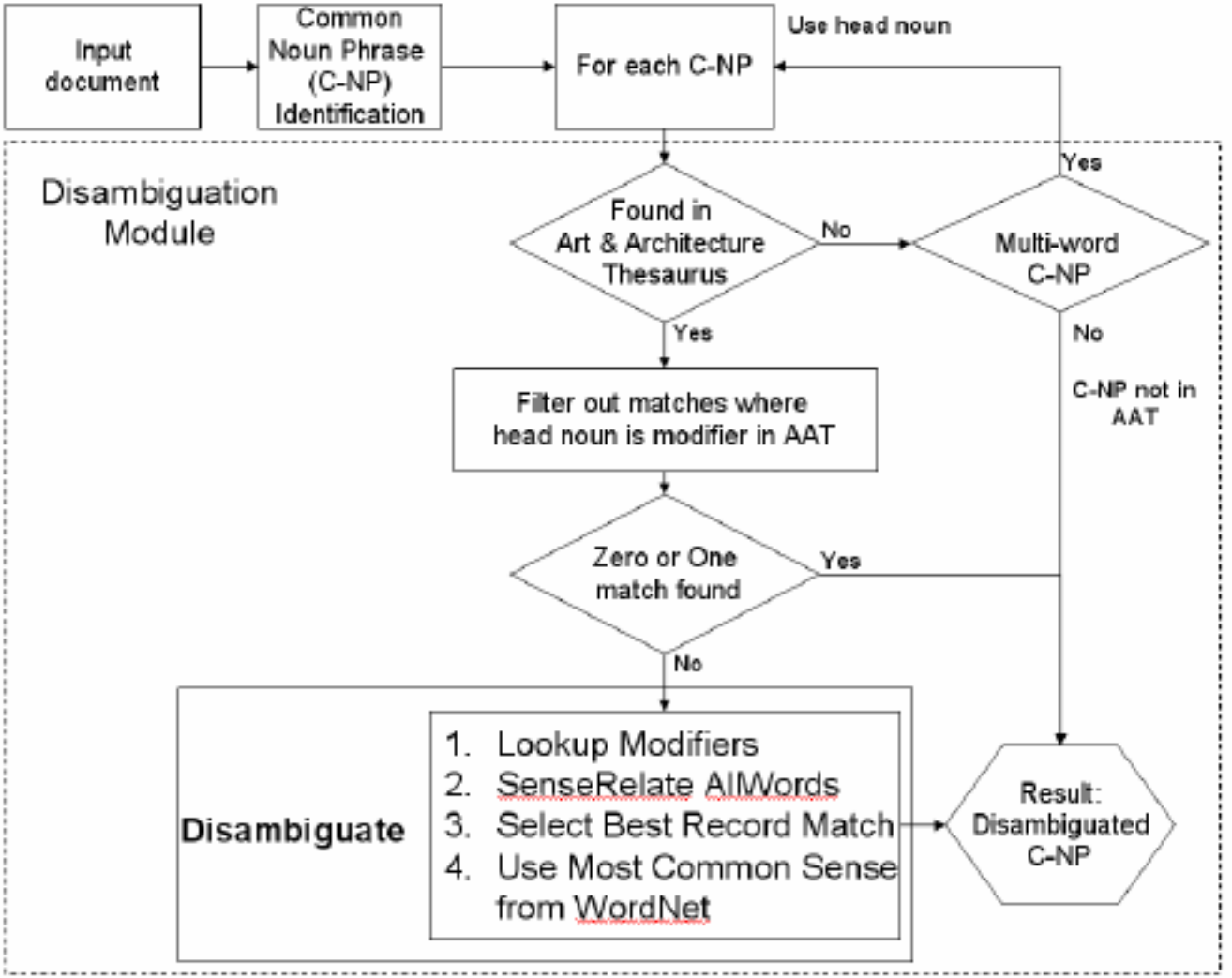
- SenseRelate AllWords and WordNet
  - SenseRelate AllWords

    → Perl program

    → performs basic disambiguation of words with the help of WordNet

    → adapted for the AAT senses

Disambiguation Algorithm



Techniques for Disambiguation

1. Use all modifiers that are in the noun phrase to find the correct AAT record
2. Use SenseRelate AllWords and WordNet
   - → result: WordNet sense of the noun phrase / its head noun
   - → examine which of the AAT senses best matches with the WordNet sense definition (word overlapping technique)
3. Use the AAT record names (preferred and variant) to find the one correct match, the one that matches best is chosen as the correct record
4. If none of these three techniques achieves success
   - → use the most common sense definition for a term (from WordNet) in conjunction with the AAT results and word overlapping

➡ if all the techniques fail, the first AAT record is selected as the correct one

Results

- 3 methods to evaluate the performance of the algorithm

  (1) Computes whether the algorithm picked the correct AAT record

  (2) Computes whether the correct record is among the top three top three records picked picked by the algorithm

  (3) Computes whether the correct record is among the Top5

- The AAT records were ranked according to their preferred name for the baseline

  → AAT records that match the term in question appear on top, followed by records that partially matched the term

Results

- Overall results
  - Results for the trainings set (n = 326 terms)

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	76%	68%
Baseline Accuracy	69%	62%
Top3	84%	78%
Top5	88%	79%

  - Results for the test set (n = 275 terms)

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	74%	73%
Baseline Accuracy	72%	69%
Top3	79%	79%
Top5	81%	80%

Results

- Results for ambigous terms
  - Results for the trainings set (n = 128 terms)

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	55%	48%
Baseline Accuracy	35%	32%
Top3	71%	71%
Top5	82%	75%

  - Results for the test set (n = 96 terms)

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	50%	53%
Baseline Accuracy	42%	39%
Top3	63%	68%
Top5	68%	71%

Analysis of the methods

- Breakdown of AAT mappings by the disambiguation techniques

Row	Technique	Training Set(n=128)	Test Set (n=96)
One	Lookup Modifier	1	3
Two	SenseRelate	108	63
Three	Best Record Match	14	12
Four	Most Common Sense	5	18

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 Com-mon Sense		3

- Breakdown of the errors in the algorithm under training set (55 total errors)

Conclusion

- Possible to create an automated system for word sense disambiguation in a domain with specialized vocabulary
- Great potential in rapid development of metadata for digital collections
- In order to integrate the program in the CliMB Toolkit, still much work has to be done:
  - Improve the algorithm's accuracy (currently 48-55%)
    → e.g. reimplement concepts behind SenseRelate (currently the work depends on the external program SenseRelate → causes errors)
  - Better and more groundtruth necessary
    → noun phrases like favour, kind, certain aspects, etc. have to be eliminated from the dataset
    → image catalogers instead of project members as labelers
  - Test the program on more collections