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
tsidhu@umiacs.umd.edu, {jklavans, jimmylin}@umd.edu

Abstract

We address the problem of mining text for relevant image metadata. Our work is situated in the art and architecture domain, where highly specialized technical vocabulary presents challenges for NLP techniques. To extract high quality metadata, the problem of word sense disambiguation must be addressed in order to avoid leading the searcher to the wrong image as a result of ambiguous — and thus faulty — metadata. In this paper, we present a disambiguation algorithm that attempts to select the correct sense of nouns in textual descriptions of art objects, with respect to a rich domain-specific thesaurus, the Art and Architecture Thesaurus (AAT). We performed a series of intrinsic evaluations using a data set of 600 subject terms extracted from an online National Gallery of Art (NGA) collection of images and text. Our results showed that the use of external knowledge sources shows an improvement over a baseline.

1. Introduction

We describe an algorithm that takes noun phrases and assigns a sense to the head noun or phrase, given a large domain-specific thesaurus, the Art and Architecture Thesaurus[1] (published by the Getty Research Institute). This research is part of the Computational Linguistics for Metadata Building (CLiMB) project (Klavans 2006, Klavans *in preparation*), which aims to improve image access by automatically extracting metadata from text associated with images. We present here a component of an overall architecture that automatically mines scholarly text for metadata terms. In order to filter and associate a term with a related concept, ambiguous terms must be clarified. The disambiguation of terms is a basic challenge in computational linguistics (Ide and Veronis 1990, Agirre and Edmonds 2006).

As more non-specialists in digital libraries search for images, the need for subject term access has increased. Subject terms enrich catalog records with valuable broad-reaching metadata and help improve image access (Layne 1994). Image seekers will receive more relevant results if image records contain terms that reflect conceptual, semantic, and ontological relationships. Furthermore, subject terms associated with hierarchical and faceted thesaural senses promise to further improve precision in image access. Such terms map to standardized thesaurus records that include the term's preferred, variant, and related names, including both broader and specific concepts, and other related concepts. This information can then be filtered, linked, and subsequently tested for usefulness in performing richer image access. As with other research on disambiguation, our hypothesis is that accurate assignment of senses to metadata index terms will results in higher precision for searchers. This hypothesis will be fully tested as we incorporate the disambiguation module in our end-to-end CLiMB Toolkit, and as we perform user studies.

Finding subject terms and mapping them to a thesaurus is a time-intensive task for catalogers

[1] http://www.getty.edu/research/conducting_research/vocabularies/aat/

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 25–32,
Prague, 28 June 2007.

(Rasmussen 1997, Ferguson and Intner 1998). Doing so typically involves reading image-related text or other sources to find subject terms. Even so, the lack of standard vocabulary in extensive subject indexing means that the enriched number of subject terms could be inadvertently offset by the vocabulary naming problem (Baca 2002).

This paper reports on our results using the subject terms in the AAT; the CLiMB project is also using the Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN). Since the focus of this paper is on disambiguation of common nouns rather than proper nouns, the AAT is our primary resource.

2. Resources

2.1 Art and Architecture Thesaurus (AAT)

The AAT is a widely-used multi-faceted thesaurus of terms for the cataloging and indexing of art, architecture, artifactual, and archival materials. Since the AAT offers a controlled vocabulary for recording and retrieval of data in object, bibliographic, and visual databases, it is of interest to a wide community.

In the AAT, each concept is described through a record which has a unique ID, preferred name, record description, variant names, broader, narrower, and related terms. In total, AAT has 31,000 such records. For the purpose of this article, a record can be viewed as synonymous with sense. Within the AAT, there are 1,400 homonyms, *i.e.,* records with same preferred name. For example, the term *wings* has five senses in the AAT (see Figure 1 below).

Wings (5 senses):

- 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.

Figure 1: Selection of AAT records for term "wings"

Table 1 shows the breakdown of the AAT vocabulary by number of senses with a sample lexical item for each frequency.

# of Senses	# of Homonyms	Example
2	1097	bells
3	215	painting
4	50	alabaster
5	39	wings
6	9	boards
7	5	amber
8	2	emerald
9	1	plum
10	1	emerald green
11	1	magenta
12	1	ocher
13	1	carmine
14	2	slate

Table 1: Scope of the disambiguation problem in AAT

Note that there are potentially three tasks that could be addressed with our algorithm: (i) mapping a term to the correct sense in the AAT, (ii) selecting amongst closely related terms in the AAT, and (iii) mapping synonyms onto a single AAT entry. In this paper, our primary focus is on task (i); we handle task (ii) with a simple ranking approach; we do not address task (iii).

Table 1 shows that multiple senses per term makes mapping subject terms to AAT very challenging. Manual disambiguation would be slow, tedious, and unrealistic. Thus we explore automatic methods since, in order to identify the correct sense of a term in running text, each of these senses needs to be viewed in context.

2.2 The Test Collection

The data set of terms that we use for evaluation comes from the National Gallery of Art (NGA) online archive[2]. This collection covers paintings, sculpture, decorative arts, and works from the Middle Ages to the present. We randomly selected 20 images with corresponding text from this collection and extracted noun phrases to form the data set. The data set was divided into two categories: the training set and the test set. The training set consisted of 326 terms and was used

[2] http://www.nga.gov/home.htm

to develop the algorithm. The test set consisted of 275 terms and was used to evaluate.

Following standard procedure in word sense disambiguation tasks (Palmer et al. 2006), groundtruth for the data set was created manually by two labelers (referred to as Labeler 1 and Labeler 2 in Section 4 below). These labelers were part of the larger CLiMB project but they were not involved in the development of the disambiguation algorithm. The process of creating the groundtruth involved picking the correct AAT record for each of the terms in the data set. Terms not appearing in the AAT (as determined by the labelers) were given an AAT record value of zero. Each labeler worked independently on this task and had access to the online version of the AAT and the text where each term appeared. Interannotator agreement for the task was encouragingly high, at 85% providing a notional upper bound for automatic system performance (Gale et al. 1992).

Not all terms in this dataset required disambiguation; 128 terms (out of 326) under the training set and 96 terms (out of 275) under the test set required disambiguation, since they matched more than one AAT record. The dataset we selected was adequate to test our different approaches and to refine our techniques. We intend to run over more data as we collect and annotate more resources for evaluation.

2.3 SenseRelate AllWords[3] and WordNet[4]

SenseRelate AllWords (Banerjee and Pederson 2003, Patwardhan et al. 2003) is a Perl program that our algorithm employs to perform basic disambiguation of words. We have adapted SenseRelate for the purpose of disambiguating AAT senses.

Given a sentence, SenseRelate AllWords disambiguates all the words in that sentence. It uses word sense definitions from WordNet (in this case WordNet 2.1), a large lexical database of English nouns, verbs, adjectives, and adverbs. As an example, consider the text below:

> With **more** than **fifty** individual scenes, the altarpiece was about fourteen feet wide.

The SenseRelate result is:

> With **more#a#2** than **fifty#n#1** individual#n#1 scene#n#10 the altarpiece#n#1 be#v#1 about#r#1 fourteen#n#1 foot#n#2 wide#a#1

In the above example, more#a#2 means SenseRelate labeled *more* as an adjective and mapped it to second meaning of *more* (found in WordNet). fifty#n#1 means SenseRelate labeled *fifty* as a noun and mapped it to first meaning of *fifty* (found in WordNet). Note, that fifty#n#1 maps to a sense in WordNet, whereas in our algorithm it needs to map to an AAT sense. In Section 3, we show how we translate a WordNet sense to an AAT sense for use in our algorithm.

To perform disambiguation, SenseRelate requires that certain parameters be set: (1) the number of words around the target word (also known as the context window), and (2) the similarity measure. We used a value of 20 for the context window, which means that SenseRelate will use 10 words to the left and 10 words to the right of the target word to determine the correct sense. We used *lesk* as the similarity measure in our algorithm which is based on Lesk (1986). This decision was based on several experiments we did with various context window sizes and various similarity measures on a data set of 60 terms.

---

[3] http://sourceforge.net/projects/senserelate

[4] http://wordnet.princeton.edu/

3. Methodology

3.1 Disambiguation Algorithm

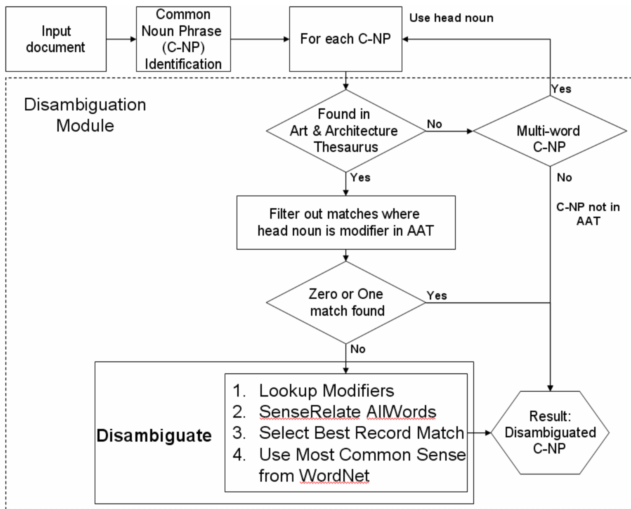

Figure 2: Disambiguation Algorithm

Figure 2 above shows that first we identify the noun phrases from the input document. Then we disambiguate each noun phrase independently by first looking it up in the AAT. If a record is found, we move on to the next step; otherwise we look up the head noun (as the noun phrase) in the AAT.

Second, we filter out any AAT records where the noun phrase (or the head noun) is used as an adjective (for a term like *painting* this would be *painting techniques*, *painting knives*, *painting equipment*, etc). Third, if zero records are found in the AAT, we label the term as "not found in AAT." If only one matching record is found, we label the term with the ID of this record. Fourth, if more than one record is found, we use the disambiguation techniques outlined in the next section to find the correct record.

3.2 Techniques for Disambiguation

For each of the terms, the following techniques were applied in the order they are given in this section. If a technique failed to disambiguate a term, we applied the next technique. If none of these techniques was able to disambiguate, we selected the first AAT record as the correct record. Findings for each technique are provided in the Results section below.

First, we used all modifiers that are in the noun phrase to find the correct AAT record. We searched for the modifiers in the record description, variant names, and the parent hierarchy names of all the matching AAT senses. If this technique narrowed down the option set to one record, then we found our correct record. For example, consider the term *ceiling coffers.* For this term we found two records: coffers (coffered ceiling components) and coffers (chests). The first record has the modifier *ceiling* in its record description, so we were able to determine that this was the correct record.

Second, we used SenseRelate AllWords and WordNet. This gave us the WordNet sense of our noun phrase (or its head noun). Using that sense definition from WordNet, we next examined which of the AAT senses best matches with the WordNet sense definition. For this, we used the word overlapping technique where we awarded a score of N to an AAT record where N words overlap with the sense that SenseRelate picked. The AAT record with the highest score was selected as the correct record. If none of the AAT records received any positive score (above a certain threshold), then it was decided that this technique could not find the one correct match.

As an example, consider finding the correct sense for the single word noun *bells* using SenseRelate:

1. Given the input sentence:

   "… city officials, and citizens were followed by women and children ringing **bells** for joy."

2. Search for AAT records. There are two records for the *bells* in AAT:

   a. **bells:** "Flared or bulbous terminals found on many open-ended aerophone tubes".
   b. **bells**: "Percussion vessels consisting of a hollow object, usually of metal but in some cultures of hard clay, wood, or glass, which when struck emits a sound by the vibration of most of its mass;…"

3. Submit the input sentence to SenseRelate, which provides a best guess for the corresponding WordNet senses for each word.

4. Get SenseRelate output, which indicates that the WordNet definition for *bells* is WordNet-Sense1, *i.e.,* "a hollow device made of metal that makes a ringing sound when struck"

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

5. Find the correct AAT match using word overlap of the WordNet definition and the two AAT definitions for *bells*:

   WordNet: "a hollow device made of metal that makes a ringing sound when struck"

   compared with:

   AAT: "Flared or bulbous terminals found on many open-ended aerophone tubes"

   and *compared with*:

   AAT: "Percussion vessels consisting of a hollow object, usually of metal but in some cultures of hard clay, wood, or glass, which when struck emits a sound by the vibration of most of its mass;…"

6. The second AAT sense is the correct sense according to the word overlap (see Table 2 below):

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

Table 2: Word Overlap to Select AAT Definition

Notice that we only used the AAT record description for performing the word overlap. We experimented by including other information present in the AAT record (like variant names, parent AAT record names) also, but simply using the record description yielded the best results.

Third, we used AAT record names (preferred and variant) to find the one correct match. If one of the record names matched better than the other record names to the noun phrase name, that record was deemed to be the correct record. For example, the term *altar* more appropriately matches *altars* (religious building fixtures) than *altarpieces* (religious visual works). Another example is *children,* which better matches *children* (youth) than *offspring* (people by family relationship).

Fourth, if none of the above techniques succeeded in selecting one record, we used the most common sense definition for a term (taken from WordNet) in conjunction with the AAT results and word overlapping mentioned above to find the one correct record.

4. Results and Evaluation

4.1 Methodologies

We used three different evaluation methods to assess the performance of our algorithm. The first evaluation method computes whether our algorithm picked the correct AAT record (*i.e.,* the AAT sense picked is in agreement with the groundtruth). The second method computes whether the correct record is among the top three records picked by our algorithm. In Table 3 below, this is referred to as *Top3*. The third evaluation method computes whether the correct record is in top five records picked by our algorithm, *Top5*. The last two evaluations helped us determine the usability of our algorithm in situations where it does not pick the correct record but it still narrows down to top three or top five results.

We ranked the AAT records according to their preferred name for the baseline, given the absence of any other disambiguation algorithm. Thus, AAT records that exactly matched the term in question appear on top, followed by records that partially matched the term. For example, for term *feet*, the top three records were *feet* (terminal elements of objects), *French feet* (bracket feet), and *Spanish feet* (furniture components). For the noun *wings*, the top three records were *wings* (shoulder accessories), *wings* (visual works components), and *wings* (backstage spaces).

4.2 Overall Results

In this section, we present evaluation results for all the terms. In the next section, we present results for only those terms that required disambiguation.

Overall results for the training set (326 terms) are shown in Table 3. This table shows that overall accuracy of our algorithm is 76% and 68% for Labeler 1 and Labeler 2, respectively. The baseline accuracy is 69% for Labeler 1 and 62% for Labeler 2. The other two evaluations show much better results. The Top 3 and Top5 evaluations have accuracy of 84% and 88% for Labeler 1 and accuracy of 78% and 79% for Labeler 2. This argues for bringing in additional techniques to enhance the SenseRelate approach in order to select from *Top3* or *Top5*.

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

Table 3: Results for Training Set (n=326 terms)

In contrast to Table 3 for the training set, Table 4 shows results for the test set. Labeler 1 shows an accuracy of 74% on the algorithm and 72% on the baseline; Labeler 2 has an accuracy of 73% on the algorithm and 69% on the baseline.

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

Table 4: Results for Test Set (n=275 terms)

4.3 Results for Ambiguous Terms

This section shows the results for the terms from the training set and the test set that required disambiguation. Table 5 below shows that our algorithm's accuracy for Labeler 1 is 55% compared to the baseline accuracy of 35%. For Labeler 2, the algorithm accuracy is 48% compared to baseline accuracy of 32%. This is significantly less than the overall accuracy of our algorithm. Top3 and Top5 evaluations have accuracy of 71% and 82% for Labeler 1 and 71% and 75% for Labeler 2.

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

Table 5: Ambiguous Terms for Training (n=128 terms)

Similar results can be seen for the test set (96 terms) in Table 6 below. Labeler 1 shows an accuracy of 50% on the algorithm and 42% on the baseline; Labeler 2 has an accuracy of 53% on the algorithm and 39% on the baseline.

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

Table 6: Results for Ambiguous Terms under the Test Set (n=96 terms)

4.4 Analysis

Table 7 shows that SenseRelate is used for most of the AAT mappings, and provides a breakdown based upon the disambiguation technique used. Row One in Table 7 shows how few terms were disambiguated using the lookup modifier technique, just 1 in the training set and 3 in the test set.

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

Table 7: Breakdown of AAT mappings by Disambiguation Technique

Rows Two and Three show that most of the terms were disambiguated using the SenseRelate technique followed by the Best Record Match technique. The Most Common Sense technique (Row Four) accounted for the rest of the labelings.

Table 8 gives insight into the errors of our algorithm for the training set terms:

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

Table 8: Breakdown of the errors in our algorithm under training set (58 total errors)

Table 8 shows the following:

(1) Out of the total of 58 errors, 16 errors were caused because SenseRelate picked the wrong WordNet sense.

(2) 8 errors were caused because WordNet did not contain the sense of the word in which it was being used. For example, consider the term *workshop*. WordNet has two definitions of *workshop*:

i. "small workplace where handcrafts or manufacturing are done" and

ii. "a brief intensive course for a small group; emphasizes problem solving"

but AAT has an additional definition that was referred by term *workshop* in the NGA text:

> "In the context of visual and decorative arts, refers to groups of artists or craftsmen collaborating to produce works, usually under a master's name"

(3) 11 errors occurred because the AAT record definition and the WordNet sense definition did not overlap. Consider the term *figures* in the sentence, "As with The Holy Family, the style of the figures offers no clear distinguishing characteristic." Then examine the AAT and WordNet sense definitions below for *figures*:

AAT sense: "Representations of humans or animals"

WordNet sense: "a model of a bodily form (especially of a person)"

These definitions do not have any words in common, but they discuss the same concept.

(4) 10 errors occurred in the Best Record Match technique, 0 errors occurred under the Lookup Modifier Technique, and 3 errors occurred under the Most Common Sense technique.

5. Conclusion

We have shown that it is possible to create an automated program to perform word sense disambiguation in a field with specialized vocabulary. Such an application could have great potential in rapid development of metadata for digital collections. Still, much work must be done in order to integrate our disambiguation program into the CLiMB Toolkit, including the following:

(1) Our algorithm's disambiguation accuracy is between 48-55% (Table 5 and Table 6), and so there is room for improvement in the algorithm. Currently we depend on an external program (SenseRelate) to perform much of the disambiguation (Table 7). Furthermore, SenseRelate maps terms to WordNet and we then map the WordNet sense to an AAT sense. This extra step is overhead, and it causes errors in our algorithm. We can either explore the option of re-implementing concepts behind SenseRelate to directly map terms to the AAT, or we may need to find additional approaches to employ hybrid techniques (including machine learning) for disambiguation. At the same time, we may benefit from the fact that WordNet, as a general resource, is domain independent and thus offers wider coverage. We will need to explore the trade-off in precision between different configurations using these different resources.

(2) We need more and better groundtruth. Our current data set of noun phrases includes term like *favor*, *kind*, and *certain aspects*. These terms are unlikely to be used as meaningful subject terms by a cataloger and will never be mapped to AAT. Thus, we need to develop reliable heuristics to determine which noun phrases are potentially high value subject index terms. A simple frequency count does not achieve this purpose.

Currently we are evaluating based on groundtruth that our project members created. Instead, we would like to extend the study to a wider set of image catalogers as labelers, since they will be the primary users of the CLiMB tool. Image catalogers have experience in finding subject terms and mapping subject terms to the AAT. They can also help determine which terms are high quality subject terms.

In contrast to working with the highly experienced image cataloger, we also want to extend the study to include various groups with different user needs. For example, journalists have ongoing needs for images, and they tend to search by subject. Using participants like these for markup and evaluation promises to provide comparative results, ones which will enable us to effectively reach a broad audience.

We also would like to test our algorithm on more collections. This will help us ascertain what kind of improvements or additions would make CLiMB a more general tool.

6. Acknowledgements

We thank Rachel Wadsworth and Carolyn Sheffield. We also acknowledge Philip Resnik for valuable discussion.

7. References

Baca, Murtha, ed. 2002. Introduction to art image access: issues, tools, standards, strategies. Getty Research Institute.

Banerjee, S., and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on ArtificialIntelligence*, 805–810.

Ferguson, Bobby and Sheila Intner. 1998. Subject Analysis: Blitz Cataloging Workbook. Westport, CT:Libraries Unlimited Inc.

Gale, W. A., K. W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 101-112, Montreal, Canada.

Ide, Nancy M. and Jean Veronis. 1990. Mapping Dictionaries: A Spreading Activation Approach. In *Proceedings of the 6th Annual Conference of the UW Centre for the New OED and Text Research*, 52-64 Waterloo, Ontario.

Lesk, Michael. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of ACM SIGDOC Conference*, 24-26, Toronto, Canada.

Klavans, Judith L. 2006. Computational Linguistics for Metadata Building (CLiMB). In *Procedings of the OntoImage Workshop,* G. Greffenstette, ed. *Language Resources and Evaluation Conference* (LREC), Genova, Italy.

Klavans, Judith L. (in preparation). Using Computational Linguistic Techniques and Thesauri for Enhancing Metadata Records in Image Search: The CLiMB Project.

Layne, Sara Shatford. 1994. Some issues in the indexing of images. *Journal of the American Society for Information Science*, 583-588.

Palmer, Martha, Hwee Tou Ng, & Hoa Trang Dang. 2006. Evaluation of WSD Systems. *Word Sense Disambiguation: Algorithms and Applications*. Eneko Agirre and Philip Edmonds, ed. 75-106. Dordrecht, The Netherlands:Springer.

Patwardhan, S., S. Banerjee, S. and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 241–257.

Rasmussen, Edie. M. 1997. Indexing images. *Annual Review of Information Science and Technology* (ARIST), 32, 169-196.