Methods for Precise Named Entity Matching in Digital Collections

Peter T. Davis
Columbia University
450 C.S. Building
New York, NY 10027
ptd7@cs.columbia.edu

David K. Elson
Columbia University
450 C.S. Building
New York, NY 10027
delson@cs.columbia.edu

Judith L. Klavans
Columbia University
508 Butler Library
New York, NY 10027
klavans@cs.columbia.edu

Abstract

In this paper, we describe an interactive system, built within the context of CLiMB project, which permits a user to locate the occurrences of named entities within a given text. The named entity tool was developed to identify references to a single art object (e.g. a particular building) with high precision in text related to images of that object in a digital collection. We start with an authoritative list of art objects, and seek to match variants of these named entities in related text. Our approach is to "decay" entities into progressively more general variants while retaining high precision. As variants become more general, and thus more ambiguous, we propose methods to disambiguate intermediate results. Our results will be used to select records into which automatically generated metadata will be loaded.

1. Computational Linguistics and Metadata

CLiMB (Computational Linguistics for Metadata Building,[1] funded by the Mellon Foundation) is an interdisciplinary project that aims to improve access to scholarly digital image collections by extracting descriptive metadata about images from related texts. With the large volume of image collections now being scanned, it is prohibitively expensive for specialized image catalogers to manually assign robust metadata to every image. By analyzing scholarly texts and associating their contents to related images, CLiMB tools will explore the potential to identify descriptive metadata which can be used to enrich catalog records. To test the process, these records will be mounted in a standard retrieval platform, where users will search for images related to particular keywords generated by CLiMB.

Although the current CLiMB project is aimed at text associated with the information in image collections, our techniques and tools are applicable to texts of many types, not just those associated with images. The application to images is one of the ways to experiment with using computational linguistic techniques to enrich catalog records. Since we are testing with text associated with images, we have designed a narrow testbed with which to measure success. If our techniques prove useful, then they should be applicable to a wide range of text types, and to languages other than English.

[1] http://www.columbia.edu/cu/cria/climb

2. The Role of Named Entities

In traditional image search platforms, the name of an art object generally serves as the key to a record. These object names tend to be complex, with a series of variants, all of them listed in catalog records to improve user search. The process of automatically associating blocks of prose from scholarly publications with image catalog records requires the identification of these art objects. Because images are discretely grouped into records about specific art objects (the domain of CLiMB's focus), we must be able to confidently identify which sections of a text are "about" which art objects. Each collection might choose a different type of entity as the art object. For example, of the three collections selected for the CLiMB project, each has a different type of object to be identified: for a collection of architectural drawings, the object is a project name for the architects; for a collection of images on South Asian temples, the object is a geographic location of the temple site; for a collection of images of paper gods from China, the art object is the name of the god or gods depicted[2].

A given art object is identified with a set of related named entities, which we call Art Object Identifiers (AO-ID's). An AO-ID can be given *a priori* to CLiMB tools by authority lists such as those from which image catalogers draw. Typically, AO-IDs are complex, with variations, which are often not very obvious. This results in their being difficult to find automatically in a text (see, for example, Table 1, row

[2] http://www.columbia.edu/cu/cria/climb/collections.html

1). Similarly, for one of the paper gods in the Chinese paper gods collection, there are over 25 variants for most god names, many of which are translations, different transliterations, or different dialects.

One of the collections with which we are developing CLiMB tools is from the American architects Charles and Henry Greene[3]. In this case, the authority list of AO-IDs is a complete list of projects built by the Greene brothers. The image collection consists of architectural drawings and photographs of the finished projects, many of which are private houses. Each image is cataloged in a record that is assigned to a particular AO-ID from the project list as the key to that record. Before passages in text can be mined for metadata, they must be classified as being "about" the record's images. A high precision for this step is crucial, since all further processing depends on the association of metadata with the AO-ID that it describes. Thus, in the precision/recall trade-off, we have opted for precision. (See [10] for a user evaluation of index terms, comparing the value of precision vs. recall.)

The most obvious way to identify a passage as relating to a particular AO-ID is simply to look for the frequency of occurrences of that AO-ID. However, this method by itself is problematic: it is highly unlikely that an AO-ID supplied either by a user or by an authority list will appear verbatim in the text. Table 1, row 1 shows the official AO-ID for the William R. Thorsen House project which does not appear at all in the major scholarly text on the Greene brothers [2]. Thus, a method for finding AO-ID variants is needed.

3. Case Study: Named Entities in Text

Early research in computing and the humanities showed that the frequency of mention of a given term in a section of a text is a straighforward method of identifying segments of text about that term. This applies not only to the use of named entities in identifying meaningful segments, but also to common nouns. Indeed, this principle underlies the term frequency and inverse document frequency relationship (tf*idf) measure popularized by [9].

Within the CLiMB project, we have explored in depth the use of named entities within [2], where different Greene & Greene projects are discussed in succession. We have started with this particular author as representative of the texts that will be processed within the project. A paragraph with a high frequency of mentions of a particular project tends to signify the beginning of a discussion of that project. The challenge of identifying such matches to the authority term is well known in the library[1] and computer science[8] communities, since a project can be known by many different variants. The house that the Greenes designed for William Thorsen, for example, is sometimes referred to as *the Thorsen house* and *the William R. Thorsen house.* Most commonly, a referential term, e.g *the house*, is used in context. By contrast, the AO-ID provided by catalogers is given as a complete, unambiguous *William R. Thorsen House (Berkeley, Calif)* which does not appear at all in the text itself. Thus, the identification of both variants and referents is required to automatically link full terms to the AO-ID to which they refer. We are exploring two techniques in the identification of AO-IDs. One is the use of a named entity finder operating directly on texts. The other, reported in this paper, is on the use of authoritative lists, either via user input or from a given list of precompiled AO-IDs.

[3]http://gamblehouse.usc.edu/architects/index.html

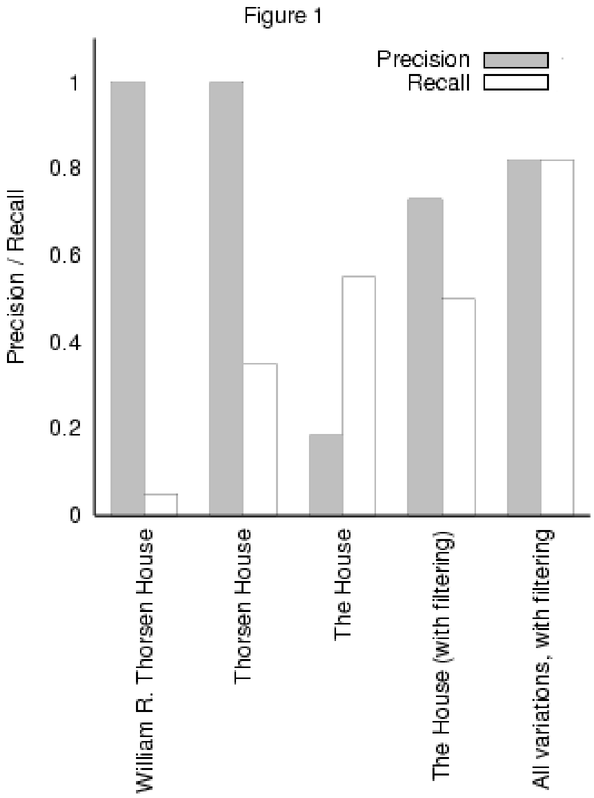
Table 1. Frequency of AO-ID variants in chapter 5 of *Greene & Greene* **(approximately 11,500 words)**

Project name variant	Occurrences
William R. Thorsen House (Berkeley, Calif)	0
William R. Thorsen House (Calif)	0
William R. Thorsen House	1
William Thorsen House	0
Thorsen House	7
The House	65

4. Using Decay to Locate Variants and Referents

We have built a tool which takes as input a user provided art object name, and a given text over which to search. Building on the role of heads and modifiers [10], we repeatedly "decay" the AO-ID by sequentially removing modifiers, causing it to become more general. We then locate occurrences of the term's variants among the noun phrases in the a text (using tools such as LTChunk from the University of Edinburgh[5]). We observe that the frequency of a variant in chapter 5 of *Greene & Greene* generally increases with the state of its decay (see Table 1). Once we remove all the modifiers, we add the determiner *the* to the head to ensure that all the occurrences are mentions of specific houses.

Our initial results showed that in chapter 5 of [2], 22 noun phrases directly refer to the Thorsen house. Of those, 19 are decayed variants of the original AO-ID. The remaining three, *the Thorsens' Berkeley residence, the project*, and *the Berkeley residence*, may be obtainable through further semantic manipulation of the original AO-ID. Precision, however, is crucial; though none of the other variants of the Thorsen AO-ID match to an unrelated noun phrase, *the house* appears 65 times, of which only 12 are in reference to



the Thorsen house (see Fig. 1). Because the 7 occurrences of *Thorsen House* are accurate, it is reasonable to hypothesize that appearances of *the house* near those occurrences are more likely to be coreferents to the AO-ID than those that do not. To test this hypothesis, we modified our algorithm to use high-precision, low-recall matches as seeds for correctly matching more ambiguous terms nearby[6]. Extrapolating the decay technique to all 253 AO-IDs in the project list, there are 27 such seeds, including the 7 occurrences of *Thorsen House* and analogous ones for 8 other projects. An occurrence of *the house*, then, is assigned to refer to the project whose seed occurs most recently.

Of the 15 occurrences of *the house* linked to the Thorsen project by this technique, 11 are linked accurately. When these results are combined with those of the seed matches, both a precision and recall of .82 are achieved for identifying references to the Thorsen house in the chapter. Despite being below 90%, this precision is high considering the difficulty of the problem. For example, in the MUC-7 coreference task, the the highest average F_1 was 61.8%, which is below our F_1 of 82% [7]. In future work, we aim for above 90% precision, which is a benchmark for precision proposed by [4].

Figure 1 shows the precision and recall of variants and referents in correctly identifying mentions of the Thorsen house in Chapter 5 of [2]. The more specific variants do not occur at all, while *the house* is too ambiguous. Results are best when combining the accurate forms with seed-based filtering on the ambiguous form.

While our results are encouraging for identifying references to the Thorsen house, there remain challenges in developing this technique into an automatic tool for heterogeneous texts. For example, it took manual intervention to know which AO-ID variant was sufficiently precise but maximally frequent to use as a seed for disambiguating the more general variant.

In future work, we will test these techniques over additional texts and explore ways to incorporate additional authority lists. Our goal is to identify and label named entities, using as much *a priori* information as possible, with the ability to fall back to sensible guessing when no authoritative information is available.

References

[1] M. Baca (ed). Introduction to art image access: issues, tools standards, strategies. Los Angeles: Getty Research Institute, 2002.

[2] E. Bosley, Greene & Greene, Phaidon Press, 2000.

[3] M. Collins, A New Statistical Parser Based on Bigram Lexical Dependencies, Proceedings of the 34th Meeting of the ACL, Santa Cruz, 1996.

[4] J. Cowie, W. Lehnert, Information Extraction, Communications of the ACM, 39 (1), 1996.

[5] S. Finch, A. Mikheev: A Workbench for Finding Structure in Texts, Proceedings of the Fifth Conference of Applied Natural Language Processing (ANLP), Washington D.C., 1997.

[6] S. Lappin and H. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535-562.

[7] MUC-7 - Proceedings of the 7th Message Understanding Conference. http://www.muc.saic.com, 1998.

[8] V. Ng, C. Cardie, Improving Machine Learning Approaches to Coreference Resolution, Proceedings of the 40th Meeting of the ACL, Philadelphia, PA, 2002.

[9] G. Salton, ed. 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall: Englewood Cliffs, NJ.

[10] N. Wacholder, D. K. Evans, and J. L. Klavans, Automatic identification and organization of index terms for interactive browsing. ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2001.