Adaptive Linking between Text and Photos Using Common Sense Reasoning

Henry Lieberman and Hugo Liu

MIT Media Laboratory
20 Ames St., E15-320G
Cambridge, MA 02139, USA
{lieber, hugo}@media.mit.edu

Abstract. In a hypermedia authoring task, an author often wants to set up meaningful connections between different media, such as text and photographs. To facilitate this task, it is helpful to have a software agent dynamically adapt the presentation of a media database to the user's authoring activities, and look for opportunities for annotation and retrieval. However, potential connections are often missed because of differences in vocabulary or semantic connections that are "obvious" to people but that might not be explicit.

ARIA (Annotation and Retrieval Integration Agent) is a software agent that acts an assistant to a user writing e-mail or Web pages. As the user types a story, it does continuous retrieval and ranking on a photo database. It can use descriptions in the story to semi-automatically annotate pictures. To improve the associations beyond simple keyword matching, we use natural language parsing techniques to extract important roles played by text, such as "who, what, where, when". Since many of the photos depict common everyday situations such as weddings or recitals, we use a common sense knowledge base, Open Mind, to fill in semantic gaps that might otherwise prevent successful associations.

1 Introduction

As digital photography becomes more popular, consumers will need better ways to organize and search their large collections of images, perhaps collected over a lifetime. Just as people compile ordinary photos into albums and scrapbooks in order to share stories with friends and family, people will want to share stories online. It is popular for users to engage in the hypermedia authoring task of sharing stories both by email and through a web page. However, there are few tools available which assist the user in their task of selecting the pictures to use to tell stories with.

ARIA [6], the software agent presented in this paper, aims to facilitate the user's storytelling task by observing the user as she tells a story, and opportunistically suggesting photos which may be relevant to what the user is typing. When a user incorporates one of the system's photo suggestions by dragging the photo into the story, our system will automatically associate with the photo any relevant keywords and phrases from the story context.

1.1 ARIA

The ARIA Photo Agent combines an email client or web page editor with a database of the user's photos, as shown in Fig. 1.

Fig. 1. A screenshot of ARIA which combines an email panel (left) with a photo database (right) that dynamically reorders itself in real-time, as the user types

Photos are automatically imported into ARIA when a digital camera flash card is inserted into the computer. Rather than requiring the user to organize photos into a directory or album structure, our system tags photos with text annotations, organized into "who, what, where, when" for each picture. The user composes an email or web page in the text client on the left. Whenever a photo is dragged from the photo pane into the text pane, new annotations are automatically associated with the photo. These annotations consist of people, places, things, and events, and are extracted from the story text adjacent to the photo in the text. Users can also edit and add to the annotations associated with a photo by double-clicking on that photo.

As the user goes about his authoring task, the photo agent monitors what he types, and in real-time, the agent reorders the annotated photos in the photo pane to suggest photos that might be relevant to the user. A photo is deemed relevant if any of its annotations can be potentially linked with the current focus of the text, either through explicit keywords, or through a variety of semantic connections.

1.2 Our Approach

ARIA goes beyond the naïve approach of suggesting photos by a simple match between keywords in a photo's annotations with keywords in the story. Such an approach often misses potential connections between keywords with different vocabu-

lary, or keywords that exhibit *implicit semantic connectedness*. By this, we mean that it may be obvious to a person that two different keywords are conceptually related, such as "bride" and "wedding;" however, computer programs cannot usually understand such connections. Our approach remedies the problems associated with naïve keyword matching by applying natural language parsing techniques to the annotation process, and commonsense reasoning to the retrieval of pictures.

To address the issue of different vocabulary, we apply natural language techniques to the annotation process such that we extract concepts rather than keywords from the text. Unlike keywords, concepts are not sensitive to morphological variation, such as "bridesmaids" versus "bridesmaid," or abbreviations or near synonyms, such as "LA" versus "Los Angeles." We map keywords into concepts using a morphological tool and abbreviation and near synonym dictionary.

In cases where potential connections are missed due to keywords that are related semantically rather than explicitly, "common sense" can help. Consider a photo annotated with "bride." Knowing some everyday knowledge about the world, ARIA can infer concepts closely related to "bride" such as "groom," "wedding," "flower girl," and "wife." These concepts are related socially, spatially, and temporally to the original concept. Expanding the original annotation with semantically related concepts gives the software agent more opportunities to recommend relevant photos to the user.

1.3 Shaping Authoring Behavior

ARIA assists the user in her authoring task by annotating photos with descriptions from the story, and dynamically adapting the presentation of the photos while the user writes the story.

Our preliminary experience with ARIA shows that the opportunistic presentation of photos can even shape the user's authoring behavior. While a user may begin the authoring task with a predisposition to tell the story a certain way, she may change her mind if the agent suggests an interesting photo at an unexpected moment. This might cause her to recall a memory, think differently, and tell the story differently. As the story unfolds, the presentation of photos will adapt accordingly, and if the agent suggests another interesting photo, the user may again revise her authoring behavior, and so on. The interaction between ARIA and the user may be able to stimulate interesting changes in the user's authoring behavior.

This paper is structured as follows: First, we discuss the source and nature of the corpus of common sense knowledge used by ARIA. Second, we present how natural language processing can incorporate commonsense knowledge in the automated annotation mechanism. Third, we discuss how commonsense can be used with shallow inference in the adaptive retrieval mechanism. Fourth, we compare our approach to related work. The paper concludes with a discussion of the benefits and limitations of our approach, and the application of commonsense to other domains.

2 Open Mind: A Source of Common Sense

The source of the commonsense knowledge used by ARIA is the Open Mind Commonsense Knowledge Base (OMCS) [11] – an endeavor at the MIT Media Laboratory that allows a web-community of teachers to collaboratively contribute to a knowledge base of "common sense." OMCS contains over 400,000 semi-structured English sentences, each of which represents a simple fact about the everyday world. Some examples of entries in the knowledge base are as follows:

1. Something you may find in a restaurant is a waiter.
2. Something that might come after a wedding is a wedding reception.
3. People get married at weddings.

OMCS is often compared with its more famous counterpart, the CYC Knowledge Base [5]. CYC contains over 1,000,000 hand-entered rules of common sense. The difference between OMCS and CYC is that CYC is meant for a more formal type of reasoning using logic, while OMCS's English sentence representation may not be constrained enough for formal logic. Even though OMCS is noisier than CYC and inherits the ambiguities associated with its natural language representation, it is still suitable to our task because we only need binary semantic relations to make adaptive linking work. This can be achieved through shallow techniques.

3 Common Sense for Parsing

When a user drags and drops a photo into the story, the description of the photo given in the story is used to automatically annotate the photo. The annotations extracted from the text are the semantically important concepts of person, place, thing, and event, which can be used to answer the "who, what, and where" questions about a photo. For the natural language parser to correctly identify these semantic types, it needs dictionaries of concepts falling under each type.

To recognize people's names, we obtain a dictionary of first names from the Web, and combine that with regular expressions to recognize full names. Geographical places are also mined from databases on the Web and added to the parser's semantic lexicon. As for everyday places, thing, and events, we extract dictionaries from Open Mind. The extraction is fairly straightforward, because many of the sentence patterns, or ontological relations, found in Open Mind sufficiently constrain the semantic types of the slots.

The result is a lexicon of words and phrases with their associated semantic type. The natural language parser uses this to enhance a syntactic parse tree with semantic and thematic phrasal tags. The resulting tree represents an event structure using an ontology based on the work of Jackendoff [3]. Below is an example of a sentence and its parse.

Sentence:

Last weekend, I went to Ken and Mary's wedding in San Francisco, and I took gorgeous pictures of the Golden Gate Bridge.

Event Structure Parse:

```
(ROOT (ASSERTION (TIME ARIA_DATESPAN{03m09d2002y-
03m10d2002y} ) , (ASSERTION (PERSON I ) (ACTION
went (PROPERTY to (EVENT (THING (PERSON Ken and )
(PERSON Mary 's ) ) wedding ) ) (PROPERTY in (PLACE
San Francisco ) ) ) , and (ASSERTION (PERSON I )
(ACTION took (THING (THING gorgeous pictures )
(PROPERTY of (PLACE the Golden Gate Bridge ) ) ) )
) ) . ) )
```

As shown in this example, knowledge mined from Open Mind and the Web allows a semantically meaningful parse to be produced. ARIA uses heuristics to decide which people, places, things, and events are relevant to the photo and should be used to annotate the photo placed adjacent to this sentence in the story.

4 Commonsense Inference for Adaptive Retrieval

ARIA uses commonsense relations mined out of Open Mind to expand annotations with semantically connected concepts that make adaptive retrieval of photos possible. To do this, a resource was automatically constructed from Open Mind by applying sentence patterns to the corpus, and extracting simple predicate argument structures (usually a binary relation). Arguments are normalized into syntactically neat concepts, and these concepts, together with the predicate relations, are used to construct a spreading activation network of nodes and directed edges. The edges between the concept nodes in the network represent the notion of semantic connectedness. The task of expanding an annotation with its related concepts is modeled as spreading activation over the network. Another way to think about spreading activation is as inference directed by the strength of relations between concepts (edge weight).

In this section, we describe how a subset of the knowledge in OMCS is extracted and structured to be useful to annotation expansion, and how spreading activation can return semantically connected concepts. Examples of actual runs of the concept expansion are given.

4.1 Extracting Concepts and Relations from OMCS

The first step of extracting predicate argument structures from OMCS is to apply a fixed set of mapping rules to the sentences in Open Mind. Each mapping rule captures a different commonsense relation that may be valuable to facilitating the retrieval task in our problem domain. The relations of interest fall under the following general categories of knowledge:

1. Classification: A dog is a pet
2. Spatial: San Francisco is part of California

3. Scene: Things often found together are: restaurant, food, waiters, tables, seats
4. Purpose: A vacation is for relaxation; Pets are for companionship
5. Causality: After the wedding ceremony comes the wedding reception.
6. Emotion: A pet makes you feel happy; Rollercoasters make you feel excited and scared.

In our extraction system, mapping rules can be found under all of these categories. To explain mapping rules, we give an example of knowledge from the aforementioned Scene category:

```
somewhere THING1 can be is PLACE1
somewherecanbe
THING1, PLACE1
0.5, 0.1
```

This rule contains a sentence pattern with the variables THING1 and PLACE1 binding to some text blob, and the name of the predicate that this relation should map to. Text blobs are normalized into concepts using a sieve-like grammar. The pair of numbers on the last line represents the confidence weights given to forward relation (left to right), and backward relation (right to left), respectively, for this predicate relation. This also corresponds to the weights associated with the directed edges between the nodes, THING1 and PLACE1 in the spreading activation network representation.

It is important to distinguish the value of the forward relation on a particular rule, as compared to a backward relation. For example, let us consider the fact, "somewhere a bride can be is at a wedding." Given the annotation "bride," it may be very useful to return "wedding." However, given the annotation "wedding," it is arguably not as useful to return all the things found at a wedding such as "bride," "groom," "wedding cake," "priest," etc. For our problem domain, we will generally penalize the direction in a relation that returns hyponymic (taxonomic child) concepts as opposed to hypernymic ones (taxonomic parent).

Approximately 20 mapping rules are applied to all the sentences (400,000+) in the OMCS corpus, and a set of 50,000 predicate argument relations is extracted. These structures are compiled into a spreading activation network consisting of 30,000 concept nodes and 160,000 direct edges. The average branching factor is 5.

4.2 Expansion as Spreading Activation

In spreading activation, the origin node is the annotation or concept we wish to expand and it is the first node to be activated. Next, the nodes one hop away from the origin node are activated, then two levels away, and so on. Nodes will continue to be activated so long as their activation score meets the activation threshold, which is a number between 0 and 1.0. Given nodes A and B, where A has one edge pointing to B, the activation score (AS) of B can be constructed:

$$AS(B) = AS(A) * weight(edge(A, B))$$

When no more nodes are activated, we have found all the relevant concepts that expand the input concept. One problem that can arise with spreading activation is that nodes that are activated two or more hops away from the origin node may quickly lose relevance, causing the search to lose focus. This can be due to noise. Because concept nodes do not make distinctions between different word senses, it is possible that a node represents many different word senses. Therefore, activating more than one hop away risks exposure to noise. Although associating weights with the edges provides some measure of relevance, these weights form a homogenous class for all edges of a common predicate (recall that the weights came from mapping rules).

We identify two opportunities to re-weight the graph to improve relevance: reinforcement and popularity. Both are relatively common techniques associated with spreading activation, but we motivate their explanations in the context of common sense.

Reinforcement

We make the observation that if the concept "bride" is connected to "groom," both directly, and through "wedding," then "groom" is more semantically relevant to "bride" because two paths connect them. This is the idea of *reinforcement*. Looking at this another way, if three or more concepts are mutually connected, as all the concepts about a wedding might be, they form a cluster, and any two concepts in the cluster have enhanced relevance because the other concepts provide additional paths for reinforcement. Applying this, we re-weight the graph by detecting clusters and increasing the weight on edges within the cluster.

Popularity

The second observation we make is that if an origin node A has a path through node B, and node B has 100 children, then each of node B's children are less likely to be relevant to node A than if node B had had 10 children. This is a common notion used in spreading activation, often referred to as "fan-out" [10].

We refer to nodes with a large branching factor as being popular. It so happens that popular nodes in our graph tend to be very common concepts in commonsense, or tend to have many different word senses, or word contexts. This causes its children to be in general, less relevant.

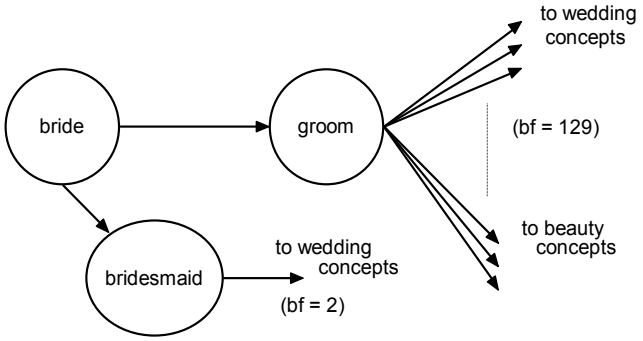

Fig. 2. Illustrating the negative effects of popularity

As illustrated in Figure 2, the concept *bride* may lead to *bridesmaid* and *groom*. Whereas *bridesmaid* is a more specific concept, not appearing in many contexts, *groom* is a less specific concept. In fact, different senses and contexts of the word can mean "the groom at a wedding," or "grooming a horse," or "he is well-groomed." This causes *groom* to have a much larger branching factor.

Despite being a knowledge base of common sense, there seems to be more value associated with more specific concepts than general ones. To apply this principle, we visit each node and discount the weights on each of its edges based on the following heuristic (α and β are constants):

$$newWeight = oldWeight * discount$$

$$discount = \frac{1}{\log(\alpha * branchingFactor + \beta)}$$

4.3 Example

Below is actual output of the concept expansion program using an activation threshold of 0.1.

```
>>> expand("bride")
('wedding', '0.3662') ('woman', '0.2023')
('ball', '0.1517') ('tree', '0.1517')
('snow covered mountain', '0.1517')
('flower', '0.1517') ('lake', '0.1517')
('cake decoration', '0.1517') ('grass', '0.1517')
('groom', '0.1517') ('tender moment', '0.1517')
('veil', '0.1517') ('tuxedo', '0.1517')
('wedding dress', '0.1517') ('sky', '0.1517')
('hair', '0.1517') ('wedding boquet', '0.1517')
```

5 Related Work

The state-of-the-art in image annotation for consumer photography is probably best represented by Kuchinsky et. al. [4]. Kuchinsky does not observationally learn annotations from text descriptions, but it does use some image analysis to propose annotations. Budzik and Hammond's Watson [1] is an agent that observes user actions and automates retrieval, but does not consider annotation. Neither of the aforementioned programs provides real-time recommendations of images or adaptively links text with images through semantic connectedness.

The concept expansion mechanism proposed here is not necessarily a new approach, but performing concept expansion with commonsense relations *is* new. In the

past, other dictionary-like resources such as lexical semantic relations [12], and keyword co-occurrence statistics [9] have been used. The limitations of these resources have been that for the most part, they operate on a word, rather that concept level. In addition, the size and variety of their relational ontologies have been a limiting factor. For example, OMCS gives us numerous relations including temporal, social, and emotion but a resource like WordNet [2] can only give us a small set of nymic relations. Represented as semi-structured English sentences, it is also relatively easy to augment the relational ontology, and easy to update.

6 Conclusion

In this paper, we presented ARIA, a software agent that facilitates a hypermedia authoring task. While the user tells a story in an email client, the agent observes the text pane and continuously presents suggestions of photos that may be relevant to the context of the developing story. By using a semantically enriched parsing technique on description text, the agent is able to automatically annotate photos used in the story with semantically important concepts like the "who, what, and where" of the photo. Then using concepts and relations extracted from Open Mind, the photo recommendation mechanism is able to adaptively present not only photos whose annotations explicitly match the text, but also photos whose annotations exhibit implicit semantic connectedness to the text.

In user testing [6], we saw not only that ARIA adapts to the user, but that the user adapts to ARIA. Often a user's typing will bring up some photos relevant to the user's current text, but that also trigger the user's memory, encouraging him or her to explain related pictures in subsequent text, triggering new picture retrieval. This *mutual adaptation* is an important characteristic of adaptive systems, and our users particularly liked the continual interplay between their story and ARIA's suggestions.

Another example of a system that successfully integrates common sense knowledge into an interactive application is Erik Mueller's Common Sense Calendar [8]. It makes "sanity checks" such as helping you avoid situations like inviting a vegetarian friend to a steak house for dinner. We think applications like this, and ARIA, show that it is not necessary to find complete solutions to the common sense reasoning problem in order to make common sense knowledge useful in an interactive application. All you have to do is use a little common sense.

References

1. Budzik, J. and Hammond, K. J.: User Interactions with Everyday Applications as Context for Just-in-Time Information Access, ACM Conf. Intelligent User Interfaces (IUI 2000), ACM Press, New York, (Jan. 2000), pp.44-51.
2. Fellbaum, C. (Ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA. (1998).
3. Jackendoff, R.: Semantic structures. Cambridge, MA: MIT Press, (1990).

4. Kuchinsky, A., Pering, C., Creech, M. L., Freeze, D., Serra, B., and Gwizdka, J.: FotoFile: a consumer multimedia organization and retrieval system, ACM Conference on Human-Computer Interface, (CHI-99) Pages 496 – 503, Pittsburgh, (May 1999).
5. Lenat, D.: The dimensions of context-space, Cycorp technical report, (1998), www.cyc.com.
6. Lieberman, H., Rosenzweig, E., and Singh, P.: Aria: An Agent For Annotating And Retrieving Images. IEEE Computer, (July 2001), pp. 57-61.
7. Minsky, M.: Commonsense-Based Interfaces. Communications of the ACM. Vol. 43, No. 8 (August, 2000), Pages 66-73
8. Mueller, E. T.: A calendar with common sense. In Proceedings of the 2000 International Conference on Intelligent User Interfaces, 198-201. New York: Association for Computing Machinery. (2000).
9. Peat, H. J. and Willett, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the ASIS, 42(5), (1991), 378--383.
10. Salton G. and Buckley C.: On the Use of Spreading Activation Methods in Automatic Information Retrieval, In Proc. 11th Ann. Int. ACM SIGIR Conf. on R&D in Information Retrieval (ACM), (1988), 147-160.
11. Singh, P.: The Public Acquisition of Commonsense Knowledge. AAAI Spring Symposium, Stanford University, Palo Alto, CA, (2002).
12. Voorhees, E.: Query expansion using lexical-semantic relations. In Proceedings of ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval. (1994) 61-69.