Adaptive Linking between Text and Photos Using Common Sense Reasoning

Henry Lieberman and Hugo Liu

MIT Media Laboratory, Software Agents Group
20 Ames St., E15-320G
Cambridge, MA 02139, USA
lieber@media.mit.edu, hugoliu@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.

Introduction

The task described in this paper is the robust retrieval of annotated photos by a keyword query. By "annotated photos", we mean a photo accompanied by some metadata about the photo, such as keywords and phrases describing people, things, places, and activities depicted in the photo. By "robust retrieval", we mean that photos should be retrievable not just by the explicit keywords in the annotation, but also by other related keywords conceptually related to the event in the photo.

In the retrieval sense, annotated photos behave similarly to documents because both contain text, which can be exploited by conventional IR techniques. In fact, the common query enrichment techniques such as thesaurus-based keyword expansion developed for document retrieval may very well work for the photo retrieval domain.

However, keyword expansion using thesauri is limited in its usefulness because keywords expanded by their synonyms can still only retrieve documents directly related to the original keyword. Furthermore, naïve synonym expansion may actually contribute more noise to the query and negate what little benefit keyword expansion may add to the query, namely, if keywords cannot have their word sense disambiguated, then synonyms for all the word senses of a particular word may be used in the expansion, and this has the potential to retrieve many irrelevant documents.

Relevant Work

Attempting to overcome the limited usefulness of keyword expansion by synonyms, various researchers have tried to use slightly more sophisticated resources for query expansion. These include dictionary-like resources such as lexical semantic relations (Voorhees, 1994), and keyword co-occurrence statistics (Peat, 1991), as well as more resources generated dynamically through relevance feedback, like global document analysis (Xu, 1996), and collaborative concept-based expansion (Klink, 2001)

Although some of these approaches are promising, they share some of the same problems as naïve synonym expansion. Dictionary-like resources such as WordNet (Fellbaum, 1998) and co-occurrence frequencies, although more sophisticated that just synonyms, still operate mostly on the word-level and suggest expansions that are lexically motivated rather than conceptually motivated. Relevance feedback, though somewhat more successful than dictionary approaches, requires additional user action and we cannot consider it fully automated retrieval, which makes it inappropriate for our task.

Photos vs. Documents

With regard to our domain of photo retrieval, we make a key observation about the difference between photos and documents. We argue that a photo has more structure and is more predictable than a document, even though that structure may not be immediately evident. The contents of

a typical document such as a web page are hard to predict, because there are too many types and genres of web pages and the content doesn't usually follow a stereotyped structure. However, with typical photos, such as one found in your photo album, there is more predictable structure. Photos capture people and things in certain places, participating in certain events and activities. Many of these situations depicted, such as a wedding, a hike, or a recital are common to human experience, and therefore have a high level of predictability.

Take for example, a picture annotated with the keyword *"bride"*. Even without looking at the photo, a person may be able to successfully guess who else is in the photo, and what situation is being depicted. Common sense says that brides are usually found at weddings, that people found around her are usually the groom, the father of the bride, bridesmaids, that weddings may take place in a chapel or church, that there may be a wedding cake, walking down the aisle, and a wedding reception.

World Semantics

Knowledge about the spatial, temporal, and social relations of the everyday world, as exemplified above, is part of commonsense knowledge. We also call this *world semantics*, referring to the meaning of concepts and how concepts relate to each other in the world.

The mechanism we propose for robust photo retrieval uses a world semantic resource in order to expand concepts in existing photo annotations with concepts that are, *inter alia,* spatially, temporally, and socially related. More specifically, we automatically constructed our resource from a corpus of English sentences about commonsense by first extracting predicate argument structures, and then compiling those structures into a Concept Node Graph, where the nodes are commonsense concepts, and the weighted edges represent commonsense relations. The graph is structured like MindNet (Dolan, 1998). Performing concept expansion using the graph is modeled as spreading activation. Relevance of a concept is measured as world semantic proximity between nodes on the graph.

This paper is structured as follows: First, we discuss the source and nature of the corpus of commonsense knowledge used by our mechanism. Second, a discussion follows regarding how our world semantic resource was automatically constructed from the corpus. Third, we show the spreading activation strategy for robust photo retrieval. The paper concludes with a discussion of the larger system to which this mechanism belongs, potential application of this type of resource in other domains, and plans for future work.

Open Mind: A Corpus of Commonsense Knowledge

The source of the world semantic knowledge used by our mechanism is the Open Mind Commonsense Knowledge Base (OMCS) (Singh, 2002) - an endeavor at the MIT Media Laboratory that aims to allow a web-community of teachers to collaboratively build a database of "common sense" knowledge. OMCS contains over 400,000 semi-structured English sentences about commonsense, organized into an ontology of commonsense relations such as the following:

- A is a B
- You are likely to find A in/at B
- A is used for B

By semi-structured English, we mean that many of the sentences loosely follow one of 20 or so sentence patterns in the ontology. However, the words and phrases represented by A and B (see above) are not restricted. Some examples of sentences in the knowledge base are:

- Something you find in (a restaurant) is (a waiter)
- The last thing you do when (getting ready for bed) is (turning off the lights)
- While (acting in a play) you might (forget your lines)

The parentheses above denote the part of the sentence pattern that is unrestricted. The major limitations of OMCS are two-fold. First, there is much ambiguity, such as the lack of disambiguated word senses, and the lack of ability to fully parse the more complex sentences. Second, some of the knowledge is noisy in that many sentences don't truly reflect common sense, and there are no claims made about the coverage of the commonsense in the knowledge base.

The Open Mind Commonsense Knowledge Base is often compared to its more famous counterpart, the CYC Knowledge Base (Lenat, 1998). CYC contains over 1,000,000 hand-entered rules that constitute "common sense". Unlike OMCS, CYC represents knowledge using formal logic, and ambiguity is minimized. Unfortunately, the CYC corpus is not publicly available at this time.

Even though OMCS is a slightly noisy and ambiguous corpus, it is still suitable to our task because we only need to mine conceptual relations out of it using shallow techniques.

Constructing a World Semantic Resource

In this section, we describe how a subset of the knowledge in OMCS is extracted and structured to be useful to the photo retrieval task. First, we apply sentence pattern rules to the raw OMCS corpus and extract crude predicate argument structures, where predicates represent

commonsense relations and arguments represent commonsense concepts. Second, concepts are normalized using natural language techniques. Third, the predicate argument structures are read into a Concept Node Graph, where nodes represent concepts, and edges represent predicate relationships. Edges are weighted to indicate the strength of the relationship.

Extracting Predicate Argument Structures

The first step of extracting predicate argument structures is to apply a fixed number of mapping rules to the sentences in OMCS. Each mapping rule captures a different commonsense relation. Commonsense relations, insofar as what interests us for constructing our world semantic resource for photos, fall under the following general categories of knowledge:

- Classification: A dog is a pet
- Spatial: San Francisco is part of California
- Scene: Things often found together are: restaurant, food, waiters, tables, seats
- Purpose: A vacation is for relaxation; Pets are for companionship
- Causality: After the wedding ceremony comes the wedding reception.
- 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
```

The first line in a mapping rule is a sentence pattern. THING1 and PLACE1 approximately bind to a word or phrase, which is later mapped to a set of canonical commonsense concepts. Line 2 specifies the name of this predicate relation. Line 3 specifies the arguments to the predicate, and corresponds to the variable names in line 1. 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 graph 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 commonsense 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 seems to be less useful to return *"bride," "groom," "wedding cake," "priest,"* and all the other things found in a wedding. For our problem domain, we will generally penalize the direction in a relation that returns hyponymic concepts as opposed to hypernymic ones.

Approximately 20 mapping rules are applied to all the sentences (400,000+) in the OMCS corpus. From this, a crude set of predicate argument relations are extracted. At this time, the text blob bound to each of the arguments needs to be normalized into concepts.

Normalizing Concepts

Because any arbitrary text blob can bind to a variable in a mapping rule, these blobs need to be normalized into concepts before they can be useful. There are two categories of concepts that can accommodate the majority of the commonsense knowledge in OMCS: Noun Phrases (things, places, people), and Activity Phrases (e.g.: *"walk the dog," "buy groceries."*).

To normalize a text blob into a Noun Phrase or Activity Phrase, we part-of-speech tag the blob, and use these tags filter the blob through a miniature noun phrase and activity phrase grammar. If the blob does not fit the grammar, it is massaged until it does or it is rejected altogether. The final step involves normalizing the verb tenses and the number of the nouns. Only after this is done can our predicate argument structure be added to our repository. A brief description of each of the aforementioned steps is given below.

The aforementioned noun phrase and activity phrase grammar is as follows:

```
NOUN PHRASE:
(PREP) (DET|POSS-PRON) NOUN
(PREP) (DET|POSS-PRON) NOUN NOUN
(PREP) NOUN POSS-MARKER (ADJ) NOUN
(PREP) (DET|POSS-PRON) NOUN NOUN NOUN
(PREP) (DET|POSS-PRON) (ADJ) NOUN PREP NOUN
ACTIVITY PHRASE:
(PREP) (ADV) VERB (ADV)
(PREP) (ADV) VERB (ADV) (DET|POSS-PRON) (ADJ) NOUN
(PREP) (ADV) VERB (ADV) (DET|POSS-PRON) (ADJ) NOUN NOUN
```

The application of the grammar behaves like a filter. If the input to a grammar rule matches any optional tokens, which are in parentheses, then this is still considered a match, but the output will filter out any optional fields. For example, the phrase, *"in your playground"* will match the first rule, but after the grammar is applied, the phrase will become just *"playground"*.

Concept Node Graph

To model concept expansion as a spreading activation task, we convert the predicate argument structures gathered previously into a Concept Node Graph shown in Figure 1.

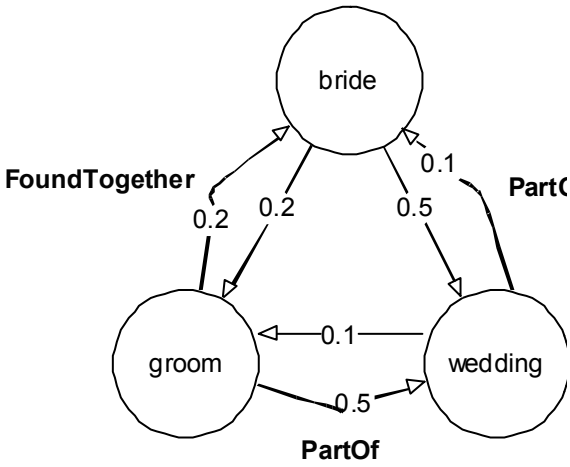

Figure 1. A portion of the Concept Node Graph. Nodes are concepts, and edges correspond to predicate relations.

The following statistics were compiled on the automatically constructed resource:

- 400,000+ sentences in OMCS corpus
- 50,000 predicate argument structures extracted
- 20 predicates in mapping rules
- 30,000 concept nodes
- 160,000 edges
- average branching factor of 5

Concept Expansion By Spreading Activation

In this section, we explain how concept expansion is modeled as spreading activation. We propose two heuristics for re-weighting the graph to improve relevance. Examples of the spreading activation are then given.

In spreading activation, the origin node is the concept we wish to expand and it is the first node to be activated. Next, all 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 1 edge pointing to B, the activation score (AS) of B is:

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

When no more nodes are activated, we have found all the concepts that expand the input concept.

Heuristics to Improve Relevance

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. One reason for this is 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.

Reinforcement

Figure 2. An example of reinforcement

As illustrated in figure 2, we make the observation that if node C is two hops away from the origin node A through path P, and also one hop away from node A through path Q, then node C is more likely to be relevant than if path Q had not existed. This is the idea of *reinforcement*. We define reinforcement as two or more corroborating pieces of evidence, represented by paths, that two nodes are related. The stronger the reinforcement, the higher the potential relevance. Looking at this another way, if three or more nodes are mutually connected, they form a cluster, and any two nodes in the cluster have enhanced relevance because the other nodes 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

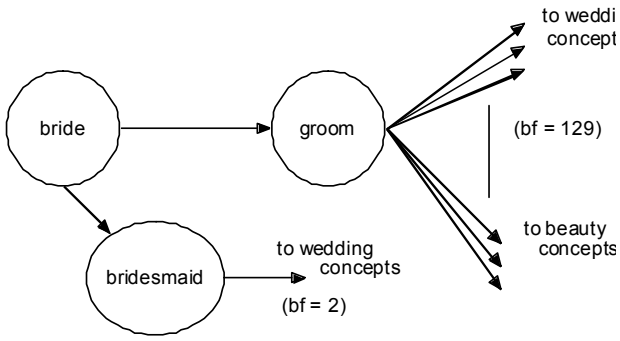

Figure 2. Illustrating the negative effects of popularity

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 based on empirical evidence with commonsense knowledge, though it has also a common technique used in spreading activation (Salton, 1988).

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.

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.

It seems that even though our knowledge is commonsense, there is 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 rule (_ and _ are constants):

$$newWeight = oldWeight * discount$$

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

Examples

Below are actual runs 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')

>>> expand('london')
('england', '0.9618') ('ontario', '0.6108')
('europe', '0.4799') ('california', '0.3622')
('united kingdom', '0.2644') ('forest', '0.2644')
('earth', '0.1244')

>>> expand("symphony")
('concert', '0.5') ('piece music', '0.4')
('kind concert', '0.4') ('theatre', '0.2469')
('screw', '0.2244') ('concert hall', '0.2244')
('xylophone', '0.1') ('harp', '0.1')
('viola', '0.1') ('cello', '0.1')
('wind instrument', '0.1') ('bassoon', '0.1')
('bass fiddle', '0.1')

>>> expand("listen to music")
('relax', '0.4816') ('be entertained', '0.4816')
('have fun', '0.4')
("understand musician's feelings", '0.4')
('feel emotionally moved', '0.4')
('hear music', '0.4') ('end silence', '0.4')
('understand', '0.4') ('mother', '0.2')
('feel emotion brings', '0.136')
('get away', '0.136') ('listen', '0.136')
('change psyche', '0.136') ('show', '0.1354')
('dance club', '0.1295') ('frisbee', '0.1295')
('scenery', '0.124') ('garden', '0.124')
('spa', '0.124') ('bean bag chair', '0.124')
```

Conclusion

In this paper, we presented a mechanism for robust photo retrieval through performing concept expansion on the photo's annotations using a world semantic resource. The resource was automatically constructed from the publicly available Open Mind Commonsense corpus. Sentence patterns were applied to the corpus, and simple predicate argument structures were extracted. After normalizing arguments into syntactically neat concepts, a weighted concept node graph was constructed. Concept expansion is modeled as spreading activation over the graph. To improve relevance in spreading activation, the graph was re-weighted according to arguments for reinforcement and against popularity.

This work has not yet been formally evaluated. Any evaluation will likely take place in the context of the larger system that this mechanism is used in, called (A)nnotation and (R)etrieval (I)ntegration (A)gent (Lieberman, 2001) ARIA is an assistive software agent which automatically learns annotations for photos by observing how users place photos in emails and web pages. It also monitors the user as she types an email and finds opportunities to suggest relevant photos. The idea of using world semantics to make the retrieval robust comes from recognizing that given one annotation about the situation depicted in a photo, it becomes possible to predict other annotations given some knowledge about spatial, temporal, and social relations about the everyday world. While the knowledge extracted from OMCS does not give very complete coverage of many different concepts, we believe that what concept expansions are done have added to the value of the retrieval process. Sometimes the concept expansions are irrelevant, but because ARIA engages in opportunistic retrieval, the user does not suffer as a result. We sometimes refer to ARIA as being "fail-soft" because the user does not rely on the software to suggest photos. But when the software does suggest a relevant photo, the user feels that ARIA has been helpful.

Robust photo retrieval is not the only IR task in which semantic resources extracted from OMCS have been successfully applied. (Liu, 2001a) used OMCS to generate effective search queries by analyzing the user's search goals. (Liu, 2001b) describes how story scripts can be generated given a seed story sentence.

In general, OMCS can benefit any program that deals with concepts of the everyday world, or assists the user in performing an everyday task. However, because of such limitations as noise and ambiguity associated with this corpus, it is likely to be only useful at a very shallow level, such as in various IR tasks.

Future work is planned to improve the performance of the mechanism presented in this paper. One major limitation that we have encountered is noise, stemming from ambiguous word senses and contexts. To overcome this, we hope to apply word sense disambiguation techniques to the concepts and the query, using word sense co-occurrence statistics, WordNet, or LDOCE. A similar approach could be taken to disambiguate meaning contexts, but it is less clear how to proceed. We also hope to migrate from sentence patterns to a broad coverage parser so that we can extract more kinds of commonsense relations from the corpus that cannot be ordinarily parsed by a sentence pattern.

Acknowledgements

We thank our colleagues Push Singh, and Kim Waters at the MIT Media Lab, Tim Chklovski at the MIT AI Lab, and Erik Mueller at IBM who are also working on the problem of Commonsense, for their contributions to our collective understanding of the issues. We would especially like to thank Push for leading the Open Mind effort and for his advocacy of commonsense reasoning.

References

Borchardt, G. C. (1992). Understanding Causal Descriptions of Physical Systems. *Proc. AAAI Tenth National Conference on Artificial Intelligence,* San Jose, CA, 2-8.

Davis, E. (1990) *Representations of commonsense knowledge*. San Mateo, Calif.: Morgan Kaufmann.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Ide, N. and Véronis, J. (Eds.) 1998. Special Issue on Word Sense Disambiguation. *Computational Linguistics*, 24(1).

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,* pp. 501-661. Prentice Hall.

Klink, S. (2001) Query reformulation with collaborative concept-based expansion. Proceedings of the First International Workshop on Web Document Analysis, Seattle, WA.

Lenat, D. (1998) *The dimensions of context-space*, Cycorp technical report, www.cyc.com.

Lieberman, H., & Selker, T. (2000). Out of context: Computer systems that adapt to, and learn from, context. *IBM Systems Journal,* 39(3,4):617-632.

Lin, D. (1998). Using collocation statistics in information extraction. In *Proc. of the Seventh Message Understanding Conference (MUC-7).*

Liu, H., Lieberman, H., Selker, T. (2002a). GOOSE: A Goal-Oriented Search Engine With Commonsense. *In Proceedings of the 2002 International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Malaga, Spain

Liu, H., Singh, P. (2002b). MAKEBELIEVE: Using Commonsense to Generate Stories. *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-02)* -- Student Abstract. Seattle, WA.

Minsky, M. Commonsense-Based Interfaces. Communications of the ACM. Vol. 43, No. 8 (August 2000), Pages 66-73

Mueller, E. T. (2000). A calendar with common sense. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, 198-201. New York: Association for Computing Machinery.

Peat, H. J. and Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the ASIS, 42(5), 378--383.

Salton G. and Buckley C. (1988). 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),* 147-160.

Shneiderman, B., Byrd, D., and Croft, B. (1998) Sorting out searching: A user-interface framework for text searches, *Communications of the ACM 41*, 4 (April 1998), 95-98.

Singh, P. (2002). The public acquisition of commonsense knowledge. *In Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.* Palo Alto, CA, AAAI.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. *In Proc. of ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval*, pages 61—69.

Xu, J. and Croft, W.B. (1996). Query Expansion Using Local and Global Document Analysis. *In Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4—11.