Towards Automatic Image Region Annotation - Image Region Textual Coreference Resolution

Emilia Apostolova
College of Computing and Digital Media
DePaul University
Chicago, IL 60604, USA
emilia.aposto@gmail.com

Dina Demner-Fushman
Communications Engineering Branch
National Library of Medicine
Bethesda, MD 20894, USA
ddemner@mail.nih.gov

Abstract

Detailed image annotation necessary for reliable image retrieval involves not only annotating the image as a single artifact, but also annotating specific objects or regions within the image. Such detailed annotation is a costly endeavor and the available annotated image data are quite limited. This paper explores the feasibility of using image captions from scientific journals for the purpose of automatically annotating image regions. Salient image clues, such as an object location within the image or an object color, together with the associated explicit object mention, are extracted and classified using rule-based and SVM learners.

1 Introduction

The profusion of digitally available images has naturally led to an interest in the field of automatic image annotation and retrieval. A number of studies attempt to associate image regions with the corresponding concepts. In (Duygulu et al., 2002), for example, the problem of annotation is treated as a translation from a set of image segments (or blobs) to a set of words. Modeling the association between blobs and words for the purpose of automated annotation has also been proposed by (Barnard et al., 2003; Jeon et al., 2003).

A recurring hindrance that appears in studies aiming at automatic image region annotation is the lack of an appropriate dataset. All of the above studies use the Corel image dataset that consists of 60,000 images annotated with 3 to 5 keywords. The need for an image dataset with annotated image regions has been recognized by many researchers. For example, Russell et al (2008) have developed a tool and a general purpose image database designed to delineate and annotate objects within image scenes.

The need for an image dataset with annotated object boundaries appears to be especially pertinent in the biomedical field. Organizing and using for research the available medical imaging data proved to be a challenge and a goal of the ongoing research. Rubin et al (2008), for example, propose an ontology and annotation tool for semantic annotation of image regions in radiology.

However, creating a dataset of image regions manually annotated and delineated by domain experts, is a costly enterprise. Any attempts to automate or semi-automate the process would be of a substantial value.

This work proposes an approach towards automatic annotation of regions of interest in images used in scientific publications. Publications abundant in image data are an untapped source of annotated image data. Due to publication standards, meaningful image captions are almost always provided within scientific articles. In addition, image Regions of Interest (ROIs) are commonly referred to within the image caption. Such ROIs are also commonly delineated with some kind of an overlay that helps locating the ROI. This is especially true for hard to interpret scientific images such as radiology images. ROIs are also described in terms of location within the image, or by the presence of a particular color. Identifying ROI mentions within image captions and visual clues pinpointing the ROI within the image would be the first step in building an object

Proceedings of NAACL HLT 2009: Short Papers, pages 41–44,
Boulder, Colorado, June 2009.

1. Object Location - explicit ROI location, e.g. front row, background, top, bottom, left, right. *Shells of planktonic animals called formainifera record climatic conditions as they are formed. This one, Globigerinoides ruber, lives year-round at the surface of the Sargasso Sea. The form of the live animal is shown **at right**, and its shell, which is actually about the size of a fine grain of sand, **at left**.*
2. Object Color - presence of a distinct color that identifies a ROI. *Anterior SSD image shows an elongated splenorenal varix (**blue area**). The varix travels from the splenic hilar region inferiorly along the left flank, down into the pelvis, and eventually back up to the left renal vein via the left gonadal vein. The kidney is encoded **yellow**, the portal system is encoded **magenta**, and the spleen is encoded **tan**.*
3. Overlay Marker - an overlay marker used to pinpoint the location of the ROI, e.g. arrows, asterisks, bounding boxes, or circles. *Transverse sonograms obtained with a 7.5-MHz linear transducer in the subareolar region. **The straight arrows** show a dilated tubular structure. **The curved arrow** indicates an intraluminal solid mass.*
4. Overlay Label - an overlay label used to pinpoint the location of the ROI, e.g. numbers, letters, words, abbreviations. *Location of the calf veins. Transverse US image just above ankle demonstrates the paired posterior tibial veins (**V**) and posterior tibial artery (**A**) imaged from a posteromedial approach. Note there is inadequate venous flow velocity to visualize with color Doppler without flow augmentation.*

Table 1: Image Markers divided into four categories, followed by a sample image caption[1] in which Image Markers are marked in bold, Image Marker Referents are underlined.

delineated and annotated image dataset.

2 Problem Definition

The goal of this research is to locate visually salient image region characteristics in the text surrounding scientific images that could be used to facilitate the delineation of the image object boundaries. This task could be broken down into two related subtasks - 1) locating and classifying textual clues for visually salient ROI features (Image Markers), and 2) locating the corresponding ROI text mentions (Image Marker Referents). Table 1 gives a classification of Image Markers including examples of Image Markers and Image Marker Referents. Figure 1 shows the frequency of Image Marker occurrences.

[1]The captions were extracted from Radiology and Radiographics © Radiological Society of North America and Oceanus © Woods Hole Oceanographic Institution.

3 Related Work

Cohen et al (2003) attempt to identify what they refer to as "image pointers" within captions in biomedical publications. The image pointers of interest are, for example, image panel labels, or letters and abbreviations used as an overlay within the image, similar to the Overlay Labels described in Table 1. They developed a set of hand-crafted rules, and a learning method involving Boosted Wrapper Induction on a dataset consisting of biomedical articles related to fluorescence microscope images.

Deschacht and Moens (2007) analyze text surrounding images in news articles trying to identify persons and objects in the text that appear in the corresponding image. They start by extracting persons' names and visual objects using Named Entity Recognition (NER) tools. Next, they measure the "salience" of the extracted named entities within the text with the assumption that more salient named entities in the text will also be present in the accompanying image.

Davis et al (2003) develop a NER tool to identify references to a single art object (for example a specific building within an image) in text related to art images for the purpose of automatic cataloging of images. They take a semi-supervised approach to locating the named entities of interest by first providing an authoritative list of art objects of interest and then seeking to match variants of the seed named entities in related text.

4 Experimental Methods and Results

4.1 Dataset

The chosen dateset contains more than 60,000 images together with their associated captions from three online life and earth sciences journals[1]. 400 randomly selected image captions were manually annotated by a single annotator with their Image Markers and Image Marker Referents and used for testing and for cross-validation respectively

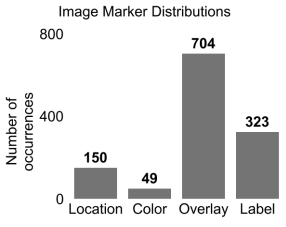

Figure 1: Distribution of Image Marker types across 400 annotated image captions.

in the two methods described below.

4.2 Rule Based Approach

First, we developed a two-stage rule-based, bootstrapping algorithm for locating the image markers and their coreferents from unannotated data. The algorithm is based on the observation that textual image markers commonly appear in parentheses and are usually closely related semantic concepts. Thus the seed for the algorithm consists of:

1. The predominant syntactic pattern - parentheses, as in *'hooking of the soft palate (arrow)'*. This pattern could easily be captured by a regular expression and doesn't require sentence parsing.

2. A dozen seed phrases (e.g *'left'*, *'circle'*, *'asterisk'*, *'blue'*) identified by initially annotating a small subset of the data (20 captions). Wordnet was used to look up and prepare a list of their corresponding inherited hypernyms. This hypernym list contains concepts such as *'a spatially limited location'*, *'a two-dimensional shape'*, *'a written or printed symbol'*, *'a visual attribute of things that results from the light they emit or transmit or reflect'*. Best results were achieved when inherited hypernyms up to the third parent were used.

In the first stage of the algorithm, all image captions were searched for parenthesized expressions that share the seed hypernyms. This step of the algorithm will result in high precision, but a low recall since image markers do not necessarily appear in parentheses. To increase recall, in stage 2 a full text search was performed for the stemmed versions of the expressions identified in stage 1.

A baseline measure was also computed for the identification of the Image Marker Referents using a simple heuristic - the coreferent of the Image Marker is usually the closest Noun Phrase (NP). In the case of parenthesized image markers, it is the closest NP to the left of the image marker; in the case of non-parenthesized image markers, the referent is usually the complement of the verb; and in the case of passive voice, the NP preceding the verb phrase. The Stanford parser was used to parse the sentences.

Table 2 summarizes the results validated against the annotated dataset (excluding the 20 captions used to identify the seed phrases). It appears that the relatively low accuracy for Image Marker Referent identification was mostly due to parsing errors since the syntactic structure of the image caption texts is quite distinct from the Penn Treebank dataset used for training the Stanford parser.

	Precision	Recall	F1-score
Image Marker	87.70	68.10	76.66
Image Marker Referent	Accuracy	59.10	

Table 2: Rule-based approach results for Image Marker and Image Marker Referent identification. Image Marker Referent results are reported as accuracy because the algorithm involves locating an Image Marker Referent for each Image Marker. Referent identification accuracy was computed for all annotated Image Markers.

Kind	k_{-5}	...	k_0	...	k_{+5}
Orth	o_{-5}	...	o_0	...	o_{+5}
Stem	s_{-5}	...	s_0	...	s_{+5}
Hypernym	h_{-5}	...	h_0	...	h_{+5}
Dep Path	d_{-5}	...	d_0	...	d_{+5}
Category			$\mathbf{[c_0]}$		

Table 3: Features from a surrounding token window are used to classify the current token into category $\mathbf{[c_0]}$. Best results were achieved with a five-token window.

4.3 Support Vector Machines

Next we explored the possibility of improving the rule-based method results by applying a machine learning technique on the set of annotated data. Support Vector Machines (SVM) (Vapnik, 2000) was the approach taken because it is a state-of-the-art classification approach proven to perform well on many NLP tasks.

In our approach, each sentence was tokenized, and tokens were classified as Beginning, Inside, or Outside an Image Marker type or Image Marker Referent. Image Marker Referents are not related to Image Markers and creating a classifier trained on this task is planned as future work. SVM classifiers were trained for each of these categories, and combined via 'one-vs-all' classification (the category of the classifier with the largest output was selected). Features of the surrounding context are used as shown in Table 3 and Table 4.

Table 5 summarizes the results of a 10-fold cross-validation. SVM performed well overall for identifying Image Markers, Location being the hardest because of higher variability of expressing ROI position. Image Marker Referents are harder to classify,

Token Kind	The general type of the sentence token (Word, Number, Symbol, Punctuation, White space).
Orthography	Orthographic categorization of the token (Upper initial, All capitals, Lower case, Mixed case).
Stem	The stem of the token, extracted with the Porter stemmer.
Wordnet Superclass	Wordnet hypernyms (nouns, verbs); the hypernym of the derivationally related form (adjectives); the superclass of the pertanym (adverbs).
POS Category	POS categories extracted using Brill's tagger.
Dependency Path*	The smallest sentence parse subtree including both the current token and the annotated image marker(s), encoded as an undirected path across POS categories.

Table 4: Orthographic, semantic, and grammatical classification features computed for each token (*Dependency Path is used only for classifying Image Marker Referents).

as deeper syntactic knowledge is necessary. Idiosyncratic syntactic structures in image captions pose a problem for the general-purpose trained Stanford parser and performance is hindered by the accuracy of computing Dependency Path feature.

5 Conclusion and Future Work

We explored the feasibility of determining the content of ROIs in images from scientific publications using image captions. We developed a two-stage rule-based approach that utilizes WordNet to find ROI pointers (Image Markers) and their referents. We also explored a supervised machine learning approach. Both approaches are promising. The rule-based approach seeded with a small manually annotated set resulted in 78.7% precision and 68.1% recall for Image Markers recognition. The SVM approach (which requires a greater annotation effort) outperformed the rule based approach (p=93.6%, r=87.7%). Future plans include training SVMs on the results of the rule-based annotation. Further work is also needed in improving Image Marker Referent identification and co-reference resolution. We also plan to involve two annotators in order to collect a more robust dataset based on inter-annotator agreement.

	Precision	Recall	F1-score
Location	60.93	45.15	51.86
Color	100.00	51.32	67.82
Overlay Marker	97.43	95.39	96.39
Overlay Label	85.74	87.69	86.70
Overall	93.64	87.69	90.56
Image Marker Referent	Accuracy	61.15	

Table 5: SVM classification results for the four types of Image Markers, and for Image Marker Referents. LibSVM software was used (3-degree polynomial kernel, cost parameter = 1, τ = 0.6 empirically determined).

References

K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei, and M.I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

W.W. Cohen, R. Wang, and R.F. Murphy. 2003. Understanding captions in biomedical publications. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 499–504. ACM New York, NY, USA.

P.T. Davis, D.K. Elson, and J.L. Klavans. 2003. Methods for precise named entity matching in digital collections. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 125–127. IEEE Computer Society Washington, DC, USA.

K. Deschacht and M. Moens. 2007. Text analysis for automatic image annotation. In *Proceedings of the 45th Annual ACL Meeting*, pages 1000–1007. ACL.

P. Duygulu, K. Barnard, JFG de Freitas, and D.A. Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *LECTURE NOTES IN COMPUTER SCIENCE*, pages 97–112.

J. Jeon, V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM New York, NY, USA.

D. Rubin, P. Mongkolwat, V. Kleper, K. Supekar, and D. Channin. 2008. Medical imaging on the Semantic Web: Annotation and image markup. In *AAAI Spring Symposium Series, Semantic Scientific Knowledge Integration*.

B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1):157–173.

V.N. Vapnik. 2000. *The Nature of Statistical Learning Theory*. Springer.