Explorations in Automatic Image Annotation using Textual Features

Chee Wee Leong
Computer Science & Engineering
University of North Texas
cheeweeleong@my.unt.edu

Rada Mihalcea
Computer Science & Engineering
University of North Texas
rada@cs.unt.edu

Abstract

In this paper, we report our work on automatic image annotation by combining several textual features drawn from the text surrounding the image. Evaluation of our system is performed on a dataset of images and texts collected from the web. We report our findings through comparative evaluation with two gold standard collections of manual annotations on the same dataset.

1 Introduction

Despite the usefulness of images in expressing ideas, machine understanding of the meaning of an image remains a daunting task for computers, as the interplay between the different visual components of an image does not conform to any fixed pattern that allows for formal reasoning of its semantics. Often, the machine interpretation of the concepts present in an image, known as *automatic image annotation*, can only be inferred by its accompanying text or co-occurrence information drawn from a large corpus of texts and images (Li and Wang, 2008; Barnard and Forsyth, 2001). Not surprisingly, humans have the innate ability to perform this task reliably, but given a large database of images, manual annotation is both labor-intensive and time-consuming.

Our work centers around the question : Provided an image with its associated text, can we use the text to reliably extract keywords that relevantly describe the image ? Note that we are not concerned with the generation of keywords for an image, but rather their *extraction* from the related text. Our goal eventually is to automate this task by leveraging on texts which are naturally occurring with images. In all our experiments, we only consider the use of nouns as annotation keywords.

2 Related Work

Although automatic image annotation is a popular task in computer vision and image processing, there are only a few efforts that leverage on the multitude of resources available for natural language processing to derive robust linguistic based image annotation models. Most of the work has posed the annotation task as a classification problem, such as (Li and Wang, 2008), where images are annotated using semantic labels associated to a semantic class.

The most recent work on image annotation using linguistic features (Feng and Lapata, 2008) involves implementing an extended version of the continuous relevance model that is proposed in (Jeon et al., 2003). The basic idea underlying their work is to perform annotation of a test image by using keywords shared by similar training images. Evaluation of their system performance is based on a dataset collected from the news domain (BBC). Unlike them, in this paper, we attempt to perform image annotation on datasets from unrestricted domains. We are also interested in extending the work pursued in (Deschacht and Moens, 2007), where *visualness* and *salience* are proposed as important textual features for discovering named entities present in an image, by extracting other textual features that can further improve existing image annotation models.

3 Data Sets

We use 180 images collected from the Web, from pages that have a single image within a specified size range (width and height of 275 to 1000 pixels). 110 images are used for development, while the remaining 70 are used for test. We create two different gold standards. The first, termed as *Intuitive annotation standard* ($GS_{intuition}$), presents a user with the image in the absence of its associated text, and asks the user for the 5 most relevant annotations. The second, called *Contextual annotation standard* ($GS_{context}$), provides the user with a list of candidates[1] for annotation, with the user free to choose any of the candidates deemed relevant to describe the image. The user, however, is not con-

[1] Union of candidates proposed by all systems participating in the evaluation, including the baseline system

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 56–59,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

strained to choose any candidate word, nor is she obligated to choose a specified number of candidates. For each image I in the evaluation set, we invited five users to perform the annotation task per gold standard. The agreement is 7.78% for $GS_{intuition}$ and 22.27% for $GS_{context}$, where we consider an annotation that is proposed by three or more users as one that is being agreed upon. The union of their inputs forms the set $GS_{intuition}(I)$ and $GS_{context}(I)$ respectively. We do not consider image captions for use as a gold standard here due to their absence in many of the images – a random sampling of 15 images reveals that 7 of them lack captions. Contrary to their use as a proxy for annotation keywords in (Feng and Lapata, 2008; Deschacht and Moens, 2007), where evaluation is performed on datasets gleaned from authoritative news websites, most captions in our dataset are not guaranteed to be noise free. However, they are used as part of the text for generating annotations where they exist.

4 Automatic Image Annotation

We approach the task of automatic image annotation using four methods. Due to the orthogonal nature in their search for keywords, the output for each method is generated separately and later combined in an unsupervised setting. However, all four methods perform their discrimination of words by drawing information exclusively from the text associated to the image, using no image visual features in the process.

4.1 Semantic Cloud (Sem)

Every text describes at least one topic that can be semantically represented by a collection of words. Intuitively, there exists several "clouds" of semantically similar words that form several, possibly overlapping, sets of topics. Our task is to select the dominant topic put forward in the text, with the assumption that such a topic is being represented by the largest set of words. We use an adapted version of the K-means clustering approach, which attempts to find natural "clusters" of words in the text by grouping words with a common centroid. Each centroid is the semantic center of the group of words and the distance between each centroid and the words are approximated by ESA (Gabrilovich and Markovitch, 2007). Further, we perform our experiments with the following assumptions : (1) To maximize recall, we assume that there are only two topics in every text. (2) Every word or collocation in the text must be classified under one of these two topics, but not both. In cases, where there is a tie, the classification is chosen randomly. For each dominant cluster extracted, we rank the words in decreasing order of their ESA distance to the centroid. Together, they represent the gist of the topic and are used as a set of candidates for labeling the image.

4.2 Lexical Distance (Lex)

Words that are lexically close to the picture in the document are generally well-suited for annotating the image. The assumption is drawn from the observation that the caption of an image is usually located close to the image itself. For images without captions, we consider words surrounding the image as possible candidates for annotation. Whenever a word appears multiple times within the text, its occurrence closest to the image is used to calculate the lexical distance. To discriminate against general words, we weigh the Lexical Distance Score (LDS) for each word by its *tf * idf* score as in the equation shown below :

$$LDS(W_i) = \textit{tf} * \textit{idf}(W_i)/LS(W_i) \qquad (1)$$

where LS(W_i) is the minimum lexical distance of W_i to the image, and *idf* is calculated using counts from the British National Corpus.

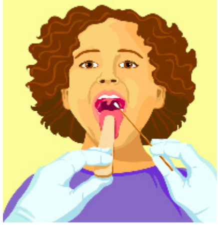
4.3 Saliency (Sal)

To our knowledge, all word similarity metrics provide a symmetric score between a pair of words w_1 and w_2 to indicate their semantic similarity. Intuitively, this is not always the case. In psycholinguistics terms, uttering w_1 may bring into mind w_2, while the appearance of w_2 without any contextual clues may not associate with w_1 at all. Thus, the degree of similarity of w_1 with respect to w_2 should be separated from that of w_2 with respect to w_1. We use a directional measure of similarity:

$$DSim(w_i, w_j) = \frac{C_{ij}}{C_i} * Sim(w_i, w_j) \qquad (2)$$

where C_{ij} is the count of articles in Wikipedia containing words w_i and w_j, C_i is the count of articles containing words w_i, and Sim(w_i, w_j) is the cosine similarity of the ESA vectors representing the two words. The *directional weight* (C_{ij}/C_i) amounts to the degree of association of w_i with respect to w_j. Using the directional inferential similarity scores as directed edges and distinct words as vertices, we obtain a graph for each text. The directed edges denotes the idea of "recommendation" where we say w_1 recommends w_2 if and only if there is a directed edge from w_1 to w_2, with the weight of the recommendation being the directional similarity score. By employing the graph iteration algorithm proposed in (Mihalcea and Tarau, 2004), we can compute the rank of a vertex in the entire graph. The output generated is a sorted list of words in decreasing order of their ranks, which serves as a list of candidates for annotating the image. Note that the top-ranked word must infer some or all of the words in the text.

Table 1: An image annotation example

	
Sem	**symptoms**, **treatment**, **medical treatment**, **medical care**, **sore throat**, fluids, **cough**, **tonsils**, **strep throat**, **swab**
Lex	**strep throat**, **cotton swab**, lymph nodes, rheumatic fever, **swab**, **strep**, fever, **sore throat**, **lab**, **scarlet fever**
Sal	**strep**, **swab**, nemours, **teens**, ginger ale, grapefruit juice, **sore**, antibiotics, **kids**, fever
Pic	**throat**, runny nose, **strep throat**, **sore throat**, hand washing, orange juice, 24 hours, **medical care**, beverages, lymph nodes
Combined	**treatment**, **cough**, **tonsils**, **swab**, fluids, **strep throat**
Doc Title	**strep throat**
*tf * idf*	**strep**, **throat**, antibiotics, **symptoms**, **child**, **swab**, fever, **treatment**, **teens**, nemours
GS$_{context}$	medical care, medical treatment, doctor, cotton swab, treatment, tonsils, sore throat, swab, throat, sore, sample, symptoms, throat, cough, medication, bacteria, lab, scarlet fever, strep throat, teens, culture, kids, child, streptococcus, doctor, strep
GS$_{intuition}$	tongue, depressor, exam, eyes, cartoon, doctor, health, child, tonsils, fingers, hair, mouth, dentist, sample, cloth, curly, tip, examine

4.4 Picturable Cues (Pic)

Some words are more *picturable* than others. For instance, it is easy to find a picture that describes the word *banana* than another word *paradigm*. Clearly, picturable words in the associated text of an image are natural candidates for labeling it. Unlike the work in (Deschacht and Moens, 2007), we employ a corpus-based approach to compute word to word similarity. We collect a list of 200 manually-annotated words[2] that are deemed to be picturable by humans. We use this list of words as our set of seed words, S_{seed}. We then iterate a bootstrapping process where each word in the text is compared to every word in the set of seed words, and any word having a maximum ESA score of greater than 0.95 is added to S_{seed}. Similarly, the maximum ESA score of each word over all S_{seed} words is recorded. This is the picturability score of the word.

5 Experiments and Evaluations

We investigate the performance of each of the four annotation methods individually, followed by a combined approach using all of them. In the individual setting, we simply obtain the set of candidates proposed by each method as possible annotation keywords for the image. In the unsupervised combined setting, only the labels proposed by all individual methods are selected, and listed in reverse order of their combined rankings.

We allow each system to produce a re-ranked list of top k words to be the final annotations for a given image. A system can discretionary generate less (but not more) than k words that is appropriate to its confidence level. Similar to (Feng and Lapata, 2008), we evaluate our systems using precision, recall and F-measure for k=10, k=15 and k=20 words.

For comparison, we also implemented two baselines systems: *tf * idf* and *Doc Title*, which simply takes all the words in the title of the web page and uses them as annotation labels for the image. In the absence of a document title, we use the first sentence in the document. The results for $GS_{intuition}$ and $GS_{context}$ are tabulated in Tables 2 and 3 respectively. We further illustrate our results with an annotation example (an image taken from a webpage discussing strep throat among teens) in Table 1. Words in bold matches $GS_{context}$ while those underlined matches $GS_{intuition}$.

6 Discussion

As observed, the system implementing the Semantic Cloud method significantly outperforms the rest of the systems in terms of recall and F-measure using the gold standard $GS_{intuition}$. The unsupervised combined system yields the highest precision at 16.26% (at k=10,15,20) but at a low recall of 1.52%. Surprisingly, the baseline system using *tf * idf* performs relatively well across all the experiments using the gold standard $GS_{intuition}$, outperforming two of our proposed methods Salience (Sal) and Picturability Cues (Pic) consistently for all k values. The other baseline, *Doc Title*, records the highest precision at 16.33% at k=10 with a low recall of 3.81%. For k=15 and k=20, the F-measure scored 6.31 and 6.29 respectively, both lower than that scored by *tf * idf*.

[2] http://simple.wikipedia.org/wiki/Wikipedia:Basic_English_picture_wordlist

Table 2: Results for Automatic Image Annotation for $GS_{intuition}$. In both Tables 2 and 3, statistically significant results are marked with *(measured against Doc Title, p<0.05, paired t-test), $^{\times}$(measured against tf*idf, p<0.1, paired t-test), †(measured against tf*idf, p<0.05, paired t-test).

	$GS_{intuition}$								
	k=10			k=15			k=20		
	P	R	F	P	R	F	P	R	F
Sem	11.71	**6.25***	**8.15**	11.31	**8.91**$^{*\times}$	**9.97***†	10.36	**9.45**$^{*\times}$	**9.88***†
Lex	9.00	4.80	6.26	7.33	5.86	6.51	7.14	7.62	7.37
Sal	4.57	2.43	3.17	6.28	5.03	5.59	6.38	6.78	6.57
Pic	7.14	3.81	4.97	6.09	4.87	5.41	5.64	6.02	5.82
Combined	16.26	1.52	2.78	**16.26**†	1.52	2.78	**16.26**†	1.52	2.78
Doc Title	**16.33**	3.81	6.18	15.56	3.96	6.31	15.33	3.96	6.29
*tf * idf*	9.71	5.18	6.76	8.28	6.63	7.36	7.14	7.62	7.37

Table 3: Results for Automatic Image Annotation for $GS_{context}$

	$GS_{context}$								
	k=10			k=15			k=20		
	P	R	F	P	R	F	P	R	F
Sem	71.57	**26.20***†	**38.36***†	68.00	**37.34***†	**48.21***†	64.56	**47.17***†	**54.51***†
Lex	61.00	22.23	32.59	58.95	32.37	41.79	56.92	41.68	48.12
Sal	46.42	16.99	24.88	51.14	28.08	36.25	54.59	39.80	46.04
Pic	51.71	21.12	29.99	56.85	31.22	40.31	56.35	41.26	47.64
Combined	**75.60***†	4.86	9.13	**75.60***†	4.86	9.13	**75.60***†	4.86	9.13
Doc Title	32.67	5.23	9.02	32.33	5.64	9.60	32.15	5.70	9.68
*tf * idf*	55.85	20.44	29.93	54.19	29.75	38.41	49.07	35.93	41.48

When performing evaluations using the gold standard $GS_{context}$, significantly higher precision, recall and F-measure values are scored by all the systems, including both baselines. This is perhaps due to the availability of candidates that suggests a form of cued recall, rather than free recall, as is the case with $GS_{intuitive}$. The user is able to annotate an image with higher accuracy e.g. labelling a Chihuahua as a *Chihuahua* instead of a *dog*. Again, the Semantic Cloud method continues to outperform all the other systems in terms of recall and F-measure consistently for k=10, k=15 and k=20 words. A similar trend as observed using the gold standard of $GS_{intuition}$ is seen here, where again our combined system favors precision over recall at all values of k.

A possible explanation for the poor performance of the Saliency method is perhaps due to over-specific words that infer all other words in the text, yet unknown to the knowledge of most human annotators. For instance, the word *Mussolini*, referring to the dictator *Benito Mussolini*, was not selected as an annotation for an image showing scenes of World War II depicting the Axis troops, though it suggests the concepts of *war*, *World War II* and so on. The Pic method is also not performing as well as expected under the two gold annotation standards, mainly due to the fact that it focuses on selecting picturable nouns but not necessarily those that are semantically linked to the image itself.

7 Future Work

The use of the semantic cloud method to generate automatic annotations is promising. Future work will consider using additional semantic resources such as ontological information and encyclopaedic knowledge to enhance existing models. We are also interested to pursue human knowledge modeling to account for the differences in annotators in order create a more objective gold standard.

References

Kobus Barnard and David Forsyth. 2001. Learning the semantics of words and pictures. In *Proceedings of International Conference on Computer Vision*.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *Proceedings of the Association for Computational Linguisticd*.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of the Association for Computational Linguistics*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *International Joint Conferences on Artificial Intelligence*.

J Jeon, V Lavrenko, and R Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jia Li and James Wang. 2008. Real-time computerized annotation of pictures. In *Proceedings of International Conference on Computer Vision*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *in Proceedings of Empirical Methods in Natural Language Processing*.