Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Automatic Caption Generation for News Images using Phrase Based Model

[1]Priyanka M. Kadhav, [2]Pritam Nikam
[1]Dept. of Wireless Communication and Computing, ABHA Gaikwad-Patil College of Engineering, Nagpur, India
[2]Chief Engineer, Samsung R &D, Research Institute, India

***Abstract— This paper deal with the confront the trouble of automatic generation of caption for news images which are collocated with thematically related documents as well as development of efficient tools that generate description for images automatically which is more advantageous to image search engines that get welfare from image description in supporting more accurate and targeted queries for end users. Searching of an image is a technique to find the required image from the collection of images. Here captions get generated automatically from the images with the help of associated document by considering that image and documents shares the common topics (variables). Further it captures the image's contents and consists of two factors that are content selection and surface realization. Content selection shows relationship between appearance of certain features in a documents with the appearance of corresponding features in a given images whereas surface realization arbitrate verbalization of the chosen contents. Annotation process applies over the images and documents collection available in the database and generating the keywords for the image. By using those keywords caption is automatically generated with the help of extractive and abstractive caption generation models. In the paper we are going to used phrase based model for caption generation as a consequence with phrase tree construction. After making comparisons between abstractive and extractive method it is examine that output of abstractive model is better than extractive method.***

***Keywords— Caption generation, Parse tree construction, stemming, stop word removal, surface realization.***

## I. INTRODUCTION

Recently immense growth have witnessed in the amount of digital information available on the internet including billions of images, documents, articles, books, sound, video, and social networks . Many online sites publish images with their stories and even provide photo feeds related to current events. Browsing, searching and retrieving pictures from such a prodigious collection encompass much interest towards information and image retrieval. Automatic generation of caption is mostly used in our real world life because more and more news televisions can show the news image with the help of caption generation. According to that everybody can see the news images with their correct information. A good caption must be succinct and informative that correctly identifies the subject of a picture. It should established picture relevance to articles and create news worthy text that draws reader into article and need not describe image content in detail [1].

Caption is a few line of text that appears below an image which is used to explain or elaborate that image. Along with the titles and section heading, caption are the most commonly read words in an article. Automatic caption generation is also helpful for searching or image retrieval. During searching process any image can be retrieve on the basis of the collocated textual information e.g. include image file name or format name or the text surrounding that image but those images that are not associated with particular text cannot retrieved. Therefore proposed system generates description for image automatically without using the dictionaries that performs mapping between image region and words by using associated document with the image. Our task is automatically generating caption for news images, so our database consist of image and its associated document is as follows and we have to generate caption:

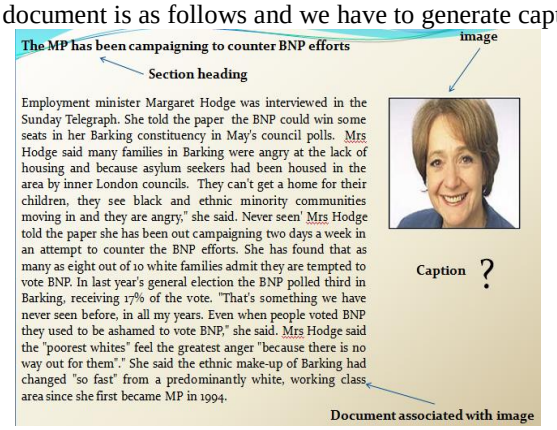


Fig 1: Example of Dataset consist of news image and document

 

## II. LITERATURE SURVEY

An image comprehension is considered as most popular factor in computer vision. An image is an artifact that depicts or records visual perception, for example some pictures has selfsame appearance as that of a many subjects including any person or objects. As mentioned earlier that image description should generate from the images. To generate description from the images two steps must be followed. The first one is to analyze the image with the help of image processing techniques and extracts foremost factors from the images by means of some extraction methods which is then transcribe into natural language text by taking into account text generation engine. In addition to above mentioned technique, image description can be generating by other fashion also. For example, Hiroshi Miki, Atsuhiro Kojima, and Koichi Kise [4], evaluate various usages of objects from the images and also identify different functions of an object, and in turn classify them accordingly by means of probability networks. This paper represents the method of creation of model of object recognition that can be done by examine relationship between human actions and function of object. Although to generate such a model various methods are available. M. Higuchi, S. Aoki, A. Kojima, and K. Fukunaga describe relationship between human actions and objects in terms of recognized the scenes [5], such relationship propose hierarchical model. However it is difficult to create suitable hierarchical model so that M. Mitani, M. Takaya, A. Kojima, and K. Fukunaga [6], develop a concept in which objects are recognize on the basis of human actions. One of the most effective mechanisms like probabilistic mechanisms by M. Saitou, A. Kojima, T. Kitahashi, and K. Fukunaga [7], describe development of relationship between human action and objects automatically in a probabilistic way rather than creating it manually. In object recognition technique extract feature values of human actions from environmental map. These feature values of human actions deliver not only human pose and actions but also select target regions. In this way object recognition can be done by getting target objects.

Other work describe by Patrick Hede, Pierre-Alain Moellic, Joel Bourgeoys, Magali Joint, Corinne Thomas [8], has described usually to represent images of objects in some natural language or in a human readable form image annotation system is utilize in image base management. Automatic image annotation also known as automatic image tagging or linguistic indexing is the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. This application is used in image retrieval systems to organize and locate images of interest from a database. Their system describe that manual database creations were required on account of images attributes i.e. color and texture to generate a caption in a natural language. Several steps must be followed in order to initiate such a description of Images. The first step is to perform segmentation technique over the images with respect to available objects in the image. Next step is to retrieve the attributes from the obtainable database, and finally description is generated using templates. But manual database creation is very monotonous and time consuming job and also it is relatively unrealistic.

B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu [9], signify most effectual methods which are image parsing and text generation. In order to generate image text description in detail a parsing technique is used. Parsing technique shows correspondence between different sharing visual patterns within a image and cut up image into various parts namely scenes-object-parts. Specifically inputted images get decompose by means of image parsing engine. For example, assume an image of scenery in which a man can standing near the river and carry a bag. This image is considered as a scene which is going to decompose into various objects like sky, tree, water, and person, ground which are subsequently converted into semantic form. Finally the task of text description is to generate meaningful and human readable text. This paper included image parsing engine that parse the input image into various parse graph and An And-or Graph (AoG) that shows syntactic and semantic relation between visual features of images scenes, objects, parts. AoG not only parse the image into top down approach but also provides mechanism that transfigure parse graph into semantic representation. Semantic web furnish interconnection between various semantic elements that are capture from previous method.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Juli Hockenmaier, David Forsyth [10], demonstrate the formation of sentences from the required images by simply compare given images and sentences of documents and obtained a score in the paper related to every picture tells a story. This score in used to solely attach given sentence to the images. This method simply retrieves a given sentence rather than composing a new one. Here also an images split into three parts scenes, objects, and action applied over the scenes and objects. It contains two main factors Illustration and annotation. From the collection of pictures find picture suggested by keywords by an illustration method. Using an image annotation model, the picture is described with keywords, which are eventually giving a detailed explanation into a human readable sentence. Headline generation bears some resemblance to the caption generation task, where the target is to create a very short summary for a document. Therefore create a caption that not only summarizes the document but is also faithful to the image's content.

In the paper published by V.Ordonez, G. Kulkarni, and T.L. Berg [11], Advances in Neural Information Processing Systems, describe the caption generation by using large collection of images based on word based model by taking into account about 1 millions photographs which is really a huge and enormous database. But in word based model some drawbacks arises. As the image annotation model does not take function words into account. Image annotation mechanism also auspiciously used relevance models, mainly implemented for information retrieval. In image annotation model keywords are used to convert image into human readable form. In this paper Extractive caption generation technique were used. Therefore explore extractive and abstractive summarization models that rely on visual information to drive the generation process. There are various caveats with extracts. There is often no single sentence in the document that uniquely describes the image's content. Most of the time keywords are available into the document but interspersed across multiple sentences. Captions (sometimes longer than the average document sentence), which are not giving a lots of information in a few words and overall not as strong as human-written

 

captions. For these reasons, turn to abstractive caption generation and present models based on single words but also phrases.

M. Banko, V. Mittal, and M. Witbrock [12][13], can help to generate headline and image description by means of statistical machine translation. Summarization technique is used to comprehend the given documents that give rise to suitable summary. From a source documents this technique should produce summary in a succinct manner. Previously extractive summarization methods were used for creation of summary. From a source documents an entire sentence or a complete paragraph merely extracts by extractive summarization technique. Such an extracted data get organize in some order to form summary. But this method does not provide required results because extracted information is not in a concise manner. And we need a sentence that should be not only short but also explain the entire document. This paper explain a approach that can help to form coherent summary which is very short and expressive by virtue of two mechanisms selection and ordering along with generalization. Selection method includes selection of necessary data from the documents and arranges those data by means of ordering.

As discuss earlier database of images and its related documents is necessary for development of all the obligatory steps to perform annotation of images. But formation and collection of such database is challenging and time consuming. To overcome such a drawback Y. Feng and M. Lapata [14] create a database of images that are naturally associated with the news articles without overhead of manual annotation because documents or news articles associated with images can heighten the image annotation process. Such document contains foremost information that is used to create image description.

## III. SYSTEM DESIGN

A great deal of work has focused on development of one method that generates image description automatically. Although various methods are available. One of the methods to get more targeted image is to create image description with the help of dictionaries that perform mapping between visual region and words manually. These approaches are very difficult and time consuming. This paper uses a mechanism works on short summary of the image's content to produce descriptions from the images automatically and therefore providing longer and more targeted queries.

**Dataset:** The dataset consist of number of news article which contains news images with their associated documents. The dataset covers a wide range of topics including national and international politics, technology, sports, education, and so on. News articles normally use color images which are around 200 pixels wide and 150 pixels high. The average sentence length is 20.5 words and the average document length 421.5 words. The document vocabulary is 26,795 words.

**Image Annotation**: Annotation process consists of two main tasks. First is feature extraction from the images and convert it into visual words with help of SIFT algorithm. Second is creating keywords for the image with the help of associated document with image using LDA algorithm. Automatic annotation detects and labels semantic content of images with a set of keywords automatically.

In annotation process first, we use SIFT algorithm and LDA. Visual features receive a discrete representation and each image is treated as a bag of visual words. In order to does this use the Scale Invariant Feature Transform (SIFT) algorithm [2]. SIFT is an algorithm in computer vision to detect and describe local features in images. An application includes finger print recognition system, face recognition, content based image retrieval system, satellite mechanism, image processing, etc. To describe features of an object it is necessary to take out required points from the object available in an image. Such a description which is extracted with respect to the features of an object is used to get needed image among the collection of object in an image. As image may contain noise therefore it should necessary to take into account one important point that even if after arising noise, features extracted from the image's object must remains noticeable. One more prime attribute of these features is that from image to image its relative position should not vary. Take one example if only corners of a car is consider as features of an object, therefore less importance given to the its position means whether car is moving or in a stationary form, but if another sides are consider excepting the corners, the recognition would fail if the car is moving or stationary. In a similar way if any changes are made in a flexible region of an object then the features located in such a region get change automatically. However, SIFT is a novel algorithm which perceive various features from the images even if some disparity caused in feature of images.

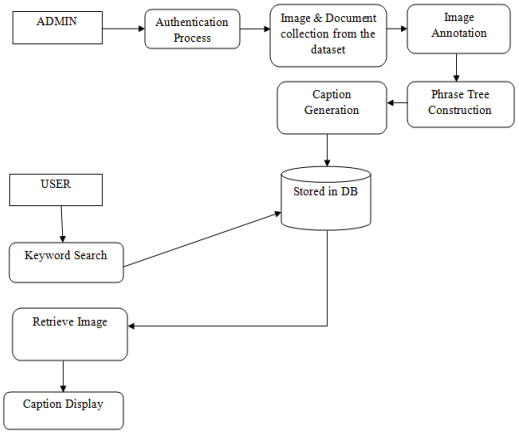


Fig 2. Block Diagram of automatic caption generation

 

In SIFT algorithm key points of objects are first extracted from a set of reference images and stored in a database. To recognize object in a targeted images comparison can be done with reference to features between database image ant targeted images using Euclidian distance of their feature vectors. Output of above steps identified collection of attributes that matches object's location and various factors related to it to filtrate nearest match. Generalized transformation technique applied by making use of hash table mechanism to determine cluster. Each cluster of three or more features that agree on an object and its pose is then subject to further detailed model verification and subsequently outliers are discarded. Ultimately the discovery of whether the object in an image is present or not is done by the probability of group of features. In addition to collection of good matches, set of false matches also design. Any object that can go through all this above steps without any fail identified as a veracious and accurate. SIFT features have been shown to be superior to other descriptors and are considered state of the art in object recognition. Typically, the SIFT descriptor is used to convert extracted image patches into visual words called as visual modality.

Next task is to convert documents into bag of words to make image and document into same format. In our paper bag of phrases are created to simplify the caption generation task. For that first we check the wordings in those documents with the help of part-of-speech. After that we perform stop-word removal and stemming processes. Stop word process is used to split the sentence into individual words. Stop words are common words that carry less important meaning than keywords. Usually search engines remove stop words from a keyword phrase to return the most relevant result. i.e. stop words drive much less traffic than keywords. Words like the, is, of, to, etc called function words are stop words this words are store and further used to create parse tree. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, For example, words such as running, runs are stemmed to run. Here we are getting bag of words but we need phrases so create phrases with the help of Parse tree construction. Parse tree constructed by using Standford parser. Parse tree contains root, sibling, leaf node and it also find the dependency between the phrases and create bag of phrases. To create parse tree we need function words that is obtained from stop words removal steps. Now image and document are of same format.

**Keyword generation**: To create keywords for the image Latent Dirichlet Allocation (LDA) is used with the help of associated document with the image by assuming that image and document are sharing some common topics. In the proposed method, visual terms and textual terms are treated as same in the document represent as mixed modality so this LDA is represented as MixLDA. MixLDA represents documents as mixtures of topics that spit out words with certain probabilities.

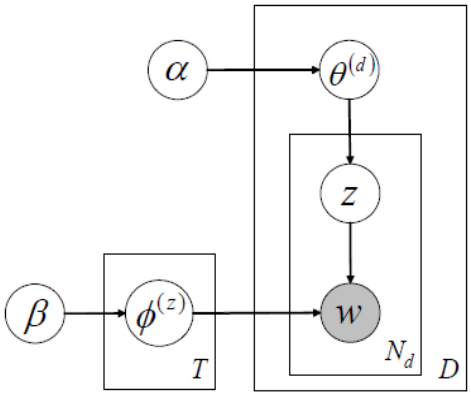


Fig. 3 Graphical model of LDA

Where D is the collection of documents, d is specific document from collection D, N is total number of words ,α is per-document topic distribution,β is per-topic word distribution ,$\theta^{(d)}$ is the topic distribution over document d ,$\varphi^{(j)}$ is word distribution over topic j, Z is the topic indicator, w is the specific word.

Probability of word can be calculated is as follows:

$$P(word) = \sum_{k=1}^{K} P(word|Topics\ k)P(\text{Topics}\ k) \quad \text{..... (1)}$$

$$P(wi) = \sum_{j=1}^{T} P(wi|zi = j)P(zi = j) \quad \text{..... (2)}$$

Where $P(w_i|z_i=j) = \varphi^{(j)}$

$P(z_i=j)= \theta^{(d)}$

Equation 2 is represented as follows

$P(wi)= \varphi^{(j)} . \theta^{(d)}$ …… (3)

After generating keywords using LDA algorithm, then find the maximal probability keywords and is given by the following equation

**W1=arg max Π P(Wt|I,D)**

Where

W is the set of textual words

 

I, D are represented jointly as the concatenation of textual and visual terms and it can be represented as jointly by dmix. It can be given by the following equation

**W1=arg max Π P(Wt|dmix)**

In this way maximal keywords are generated. This keywords are the keywords for image.

**Caption generation:** Now we have description keywords and nag of phrases **.**With the help of phrase dependency predicted by the Standford parser, abstractive caption generation generated phrases for the keywords. A phrase is simply a head and its dependent. Finally around those phrases extract dependency from the document with the help of Standford parser and glue them together (using n-gram model).

$$P(\rho 1, \rho 2, \ldots, \rho m) \approx \prod_{j=1}^{m} P(\rho j \in C | \rho j \in D)$$

$$. \prod_{j=2}^{m} P(\rho j \mid \rho j - 1)$$

$$. P(len(C) = \prod_{j=1}^{m} len(\rho j))$$

$$. \prod_{i=3}^{\sum_{j=1}^{m} len(\rho j)} Padp(wi \mid wi - 1, wi - 2)$$

**Image retrieval:** After generating caption automatically, the captions and the particular image are stored in the database. The final process is to search the particular images from the search engine. Here, the user can enter the keyword of a particular image into the search engine; it can retrieve the correct image from the database. The image can be displayed with their caption.

**Example**

Document associated with image**:** Employment minister Margaret Hodge was interviewed in the Sunday Telegraph. She told the paper the BNP could win some seats in her Barking constituency in May's council polls. Mrs Hodge said many families in Barking were angry at the lack of housing and because asylum seekers had been housed in the area by inner London councils. They can't get a home for their children, they see black and ethnic minority communities moving in and they are angry," she said. Never seen' Mrs Hodge told the paper she has been out campaigning two days a week in an attempt to counter the BNP efforts. She has found that as many as eight out of 10 white families admit they are tempted to vote BNP. In last year's general election the BNP polled third in Barking, receiving 17% of the vote. "That's something we have never seen before, in all my years. Even when people voted BNP they used to be ashamed to vote BNP," she said. Mrs Hodge said the "poorest whites" feel the greatest anger "because there is no way out for them"." She said the ethnic make-up of Barking had changed "so fast" from a predominantly white, working class area since she first became MP in 1994.

Image:

Generated Caption: "That is the key thing that has created the environment the BNP has sought to exploit," she said.

## IV. CONCLUSIONS

In this paper we studied the efficient method of automatic text generation from the images and searching methods that search the images with respect to content of images, without giving more importance to surrounding text. So that we get more targeted images. Also shows relationship between content selection and surface realization to achieve caption generation using phrase based model for image annotation method. Here comparative analysis has been studied different mechanisms with reference to various methods that are used to generate caption and the text from the images. From the comparison it is found that number of factors was found that affect the performance of text generation

## REFERENCES

[1] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," IEEE Trans. Image Processing, vol. 10, no. 1, pp. 117-130, 2001.

[2] D. Lowe, "Object Recognition from Local Scale-Invariant Features," Proc. IEEE Int'l Conf. Computer Vision, pp. 1150-1157, 1999.

 

[3] A. Ahmed, E.P. Xing, W.W. Cohen, and R.F. Murphy, "Structured Correspondence Topic Models for Mining Captioned Figures in Biological Literature," Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining, pp. 39-48, 2009.

[4] A. Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga, "Recognition and Textual Description of Human Activities by Mobile Robot," Proc. Third Int'l Conf. Innovative Computing Information and Control, pp. 53-56, 2008.

[5] M. Higuchi, S. Aoki, A. Kojima, and K. Fukunaga. Scene recognition based on relationship between human actions and objects. In *17th International Conference on Pattern Recognition*, volume 3, pages 73–78, Aug. 2004.

[6] M. Mitani, M. Takaya, A. Kojima, and K. Fukunaga. Environment recognition based on analysis of human actions for mobile robot. In *the 18th International Conference on Pattern Recognition*, volume 4, pages 782–786, Aug. 2006.

[7] M. Saitou, A. Kojima, T. Kitahashi, and K. Fukunaga. Dynamic recognition of human actions and objects using dual hierarchical models. In *the First International Conference on Innovative Computing*, pages 306–309, Aug. 2006.

[8] P. He´de, P.A. Moe¨llic, J. Bourgeoys, M. Joint, and C. Thomas, "Automatic Generation of Natural Language Descriptions for Images," Proc. Recherche d'Information Assiste´e par Ordinateur, 2004.

[9] B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu, "I2T: Image Parsing to Text Description," Proc. IEEE, vol. 98, no. 8, pp. 1485- 1508, 2009.

[10] A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every Picture Tells a Story: Generating Sentences from Images," Proc. 11th European Conf. Computer Vision, pp. 15-29, 2010.

[11] V. Ordonez, G. Kulkarni, and T.L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," Advances in Neural Information Processing Systems, vol. 24, pp. 1143-1151, 2011.

[12] M. Banko, V. Mittal, and M. Witbrock, "Headline Generation Based on Statistical Translation," Proc. 38th Ann. Meeting Assoc. for Computational Linguistics, pp. 318-325, 2000.

[13] M. Witbrock and V. Mittal, "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 315-316, 1999.

[14] Y. Feng and M. Lapata, "Automatic Image Annotation Using Auxiliary Text Information," Proc. 46th Ann. Meeting Assoc. of Computational Linguistics: Human Language Technologies, pp. 272- 280, 2008.

 