ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 2, Ver. 1 (April - June 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

A Systematic Approach for News Caption Generation

[I]Vini Varghese, [II]J Saravanan
[I,II]Dept. of CSE, M.E In CSE (4TH SEM), Mahendra Institute Of Engg. and Technology, Mahendrapuri, Mallasamudram (W), Thiruchengode, Namakkal, India

Abstract
Automatic image caption generation is of great interest to many image related applications.Now a day's, whenever retrieving images from the search Engines that retrieves images without analyzing their content, simply by matching user queries against the image's file name and format, user-annotated tags, captions, and, generally, text surrounding the image. Also the retrieved image does not contain any textual data along with the images.
We introduced the task of automatic caption generation for news images. The task fuses insights from computer vision and natural language processing and holds promise for various multimedia applications, such as image retrieval, development of tools supporting news media management, and for individuals with visual impairment. It is possible to learn a caption generation model from weakly labeled data without costly manual involvement. Instead of manually creating annotations, image captions are treated as labels for the image. Although the caption words are admittedly noisy compared to traditional human-created keywords, we show that they can be used to learn the correspondences between visual and textual modalities, and also serve as a gold standard for the caption generation task. We have presented extractive and abstractive caption generation models. A key aspect of our approach is to allow both the visual and textual modalities to influence the generation task.

Keywords
Caption generation, image annotation, summarization, topic models.

I. Introduction

Automatic image caption generation is of great interest to many image related applications. Examples include image search engines and tools for helping people with visual impairment to access multimedia information in the same way as sighted people. However, relatively little work has focused on the interplay between visual and linguistic information in literature. Existing efforts often follow a two-step natural language generation framework consisting of content selection and surface realization.
This paper is concerned with the task of automatically generating captions for images, which is important for many imagerelated applications. Examples include video and image retrieval as well as the development of tools that aid visually impaired individuals to access pictorial information. Our approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned and colocated with thematically related documents. Our model learns to create captions from a database of news articles, the pictures embedded in them, and their captions, and consists of two stages. Content selection identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content. We approximate content selection with a probabilistic image annotation model that suggests keywords for an image. The model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics) and is trained on a weakly labeled dataset (which treats the captions and associated news articles as image labels). Inspired by recent work in summarization, they propose extractive and abstractive surface realization models. Experimental results show that it is viable to generate captions that are pertinent to the specific content of an image and its associated article, while permitting creativity in the description. Indeed, the output of our abstractive model compares favorably to handwritten captions and is often superior to extractive method.
In this paper, we explore the feasibility of creating captions, using annotation keywords, for images associated with news documents. The availability of the accompanying news documents in our dataset enables us to formulate the generation module so that it resembles text summarization. We then propose both extractive and abstractive caption generation models. The backbone for both approaches is the probabilistic
image annotation model that suggests content for an image given this image and its associated document. We can then simply identify the sentences in the document that share these keywords or create a new caption that is potentially more concise but also informative and fluent. Specifically, for extractive
models, we examine how to establish criterions for selecting sentences that are similar to the image content. We also present abstractive caption generation models that operate over image description keywords and document phrases. Their combination gives rise to many caption realizations which we select probabilistically by taking intoaccount dependency and word order constraints. Experimental results show that bothapproaches generate readable captions with littlehuman involvement. Our abstractivemodel defined over phrases yields more grammatical output than word-based methods.

II. Proposed System

In this paper, we tackle the related problem of generating captions for news images. Our approach leverages the vast resource of pictures available on the web and the fact that many of them naturally co-occur with topically related documents and are captioned. We focus on captioned images embedded in news articles, and learn both models of content selection and surface realization from data without requiring expensive manual annotation. At training time, our models learn from images, their captions, and associated documents, while at test time they are given an image and the document it is embedded in and generate a caption. Compared to most work on image description generation, our approach is shallower, it does not rely on dictionaries specifying image-to-text correspondences, nor does it use a human-authored grammar for the caption creation task. It uses the document co-located with the image as a proxy for linguistic, visual, and world-knowledge.

Advantages:

- Content selection and surface realization from data without requiring expensive manual annotation.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

- It does not rely on dictionaries specifying image-to-text correspondences, nor does it use a human-authored grammar for the caption creation task.
- It reduces the need for human supervision.

III. System Details

We implement a system architecture for solving this problem.

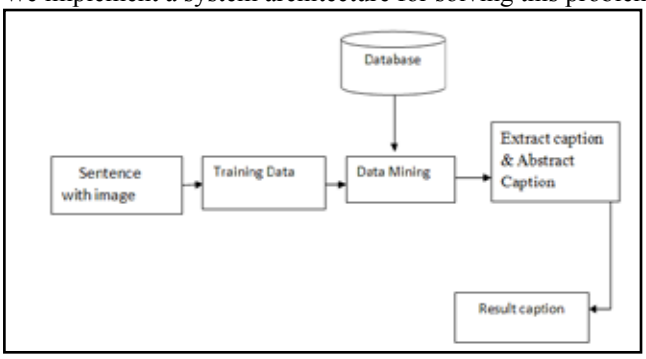
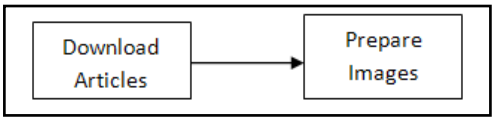


Fig. 1: System architecture

A. Data Collection

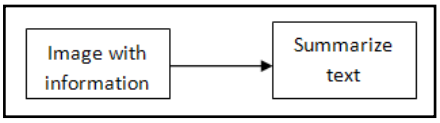


Data collection usually takes place early on in an improvement project, and is often formalised through a data collection plan which often contains the following activity.

1. Pre collection activity — agree on goals, target data, definitions, methods
2. Collection — data collections
3. Present Findings — usually involves some form of sorting analysis and/or presentation.

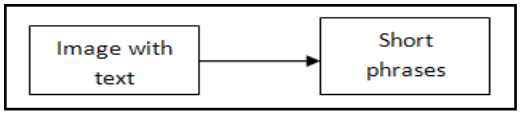
We created our own dataset by downloading articles from the News websites. The dataset covers a wide range of topics including national and international politics, technology, sports, education, and so on. News articles normally use color images which are around 200 pixels wide and 150 pixels high. The captions tend to use half as many words as the document sentences and more than 50 percent of the time contain words that are not attested in the document.

B. Input preparation



The document should contain the necessary background information which the image describes or supplements. And also we can exploit the rich linguistic information inherent in the text and address caption generation with methods relative to text summarization without extensive knowledge engineering. The caption generation task is not constrained in any way, words and syntactic structures are chosen with the aim of creating a good caption rather than rendering the task acceptable to current vision and language generation techniques.

C. Abstractive caption



We turn to abstractive caption generation and present models based on single words but also phrases. Content selection is modeled as the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent of other words in the headline. They also take the distribution of the length of the headlines into account in an attempt to relative to the model toward generating output of reasonable length.

D. Extractive caption

A phrase may refer to any group of words In linguistics, a phrase is a group of words (or sometimes a single word) that form a constituent and so function as a single unit in the syntax of a sentence. A phrase is lower on the grammatical hierarchy than a clause.

This Extractive caption mostly focuses on sentence extraction. The idea is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document. Without a great deal of linguistic analysis, it is possible to create summaries for a wide range of documents, independently of style, text type, and subject matter. For our caption generation task, we need only extract a single sentence. And our guiding hypothesis is that this sentence must be maximally similar to the description keywords generated by the annotation model.

Explanation:

Search engines deployed on the web retrieve images without analyzing their content, simply by matching user queries against collocated textual information. Examples include metadata (e.g., the image's file name and format), user-annotated tags, captions, and, generally, text surrounding the image. As this limits the applicability of search engines (images that do not coincide with textual data cannot be retrieved), a great deal of work has focused on the development of methods that generate description words for a picture automatically.

This approach can create sentences of high quality that are both meaningful and fluent; however, the reliance on manually created resources largely limits the deployment of existing methods to real-world applications. Developing dictionaries that specify exhaustively image-to-text correspondences is a difficult and time-consuming task that must be repeated for new domains and languages. The same is also true for templates and rule-based generation methods; the former are typically specific to the domain in question and not portable to new tasks, whereas the latter can be more general and linguistically sophisticated, albeit with extensive knowledge engineering.

IV. Caption Generation

A caption, also known as a cutline, is text that appears below an image. Most captions draw attention to something in the image that is not obvious, such as its relevance to the text. Captions can consist of a few words of description, or several sentences. Writing good captions is difficult, and the examples below may

be helpful. Along with the title, the lead, and section headings, captions are the most commonly read words in an article, so they should be succinct and informative.

In this project we are using stemming algorithm:-

Stemming Algorithm:

In information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form, for example,

connection
connections
connective ---> connect
connected
connecting

Algorithm techniques

Lookup algorithms

A simple stemmer looks up the inflected form in a lookup table. The advantages of this approach is that it is simple, fast, and easily handles exceptions. The disadvantages are that all inflected forms must be explicitly listed in the table: new or unfamiliar words are not handled, even if they are perfectly regular (e.g. iPads ~ iPad), and the table may be large. For languages with simple morphology, like English, table sizes are modest, but highly inflected languages like Turkish may have hundreds of potential inflected forms for each root.

The Production Technique

The lookup table used by a stemmer is generally produced semi-automatically. For example, if the word is "run", then the inverted algorithm might automatically generate the forms "running", "runs", "runned", and "runly". The last two forms are valid constructions, but they are unlikely to appear in a normal English-language text.

Suffix-stripping algorithms

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Suffix stripping algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations (like 'ran' and 'run'). The solutions produced by suffix stripping algorithms are limited to those lexical categories which have well known suffixes with few exceptions. This, however, is a problem, as not all parts of speech have such a well formulated set of rules. Lemmatisation attempts to improve upon this challenge.

Example

A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument".

Criteria for Good Caption

There are several criteria for a good caption. A good caption

1. clearly identifies the subject of the picture, without detailing the obvious.
2. is succinct.
3. establishes the picture's relevance to the article.
4. provides context for the picture.
5. draws the reader into the article.

Different people read articles in different ways. Some people start at the top and read each word until the end. Others read the first paragraph and scan through for other interesting information, looking especially at pictures and captions. Those readers, even if the information is adjacent in the text, will not find it unless it is in the caption. However, it is best not to tell the whole story in the caption, but use the caption to make the reader curious about the subject.

Clear identification of the subject

One of a caption's primary purposes is to identify the subject of the picture. Make sure your caption does that, without leaving readers to wonder what the subject of the picture might be. Be as unambiguous as practical in identifying the subject. What the picture is is important, too. If the illustration is a painting, the painter's Wikilinked name, the title, and a date give context. The present location may be added in parentheses: (Louvre).

Technical images

Technical images like charts and diagrams may have captions that are much longer than other images. Prose should still be succinct, but the significance of the image should be fully explained. Any elements not included in a legend or clearly labelled should be defined in the caption. A substantial, full discussion of a technical image may be confined to the caption if it improves the structure of the prose in the main article.

For maps and other images with a legend, the {{legend}} template can be used in the caption instead of (or in addition to) including the legend explaining the color used in the image. This makes the legend more readable, and allows for easy translation into other languages.

V. Image Annotation

Automatic image annotation (also known as automatic image tagging or linguistic indexing) is the process by which a computer system automatically assigns metadata in a digital image. This application of computer vision techniques is used in image retrieval systems to the for of captioning or keywords to organize and locate images of interest from a database.

This method can be regarded as a type of multi-class image classification with a very large number of classes as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

used by machine learning techniques to attempt to automatically apply annotations to new images. The first methods learned the correlations between image features and training annotations, then techniques were developed using machine translation to try to translate the textual vocabulary with the 'visual vocabulary', or clustered regions known as blobs. Work following these efforts have included classification approaches, relevance models and so on.

The advantages of automatic image annotation versus content-based image retrieval are that queries can be more naturally specified by the user . CBIR generally (at present) requires users to search by image concepts such as color and texture, or finding example queries. Certain image features in example images may override the concept that the user is really focusing on. The traditional methods of image retrieval such as those used by libraries have relied on manually annotated images, which is expensive and time-consuming, especially given the large and constantly growing image databases in existence.

Feature extraction and image representation

In image classification and retrieval, images are represented using low level features. Because an image is an unstructured array of pixels, the first step in semantic understanding is to extract efficient and effective visual features from these pixels. Appropriate feature representation significantly improves the performance of the semantic learning techniques. While both global and region based image representations are used in the existing image retrieval techniques, the trend is towards using region based features. Region based feature extraction needs prior image segmentation while global features are directly calculated from the whole image. In the following, we first briefly review common image segmentation algorithms used in AIA techniques. Then, various feature extraction techniques will be reviewed in detail.

A. Image segmentation

Image segmentation is usually the first step to extract region based image representation. The segmentation algorithm divides images into different components based on feature homogeneity. A number of segmentation approaches exist in the literature, such as grid based, clustering based, contour based, model based, graph based, and region growing based method. This section provides a brief review of segmentation methods commonly used in AIA. For a comprehensive segmentation review, readers are referred to.

Because automatic image segmentation is a difficult task, many techniques simplify this task using grid based approach to roughly segment images into blocks. Visual features are then extracted from these blocks. Block based approach takes little computations; however, this simple technique does not describe the semantic components in images well. A single block often consists of parts of visually different objects. Furthermore, it is difficult to determine the size of blocks for image representation. Therefore, region features are usually not accurate. If appropriately applied, it can be used in domain specific applications, e.g., medical image classification.

The main idea of contour based segmentation is to evolve a curve around an object. The evolution stops when the curve coincides with the boundary of an object. Unlike the cluster based segmentation algorithm, contour based segmentation algorithms do not need the prior assumption of the number of clusters. The underlying problem in this approach is the dependency on accurate edge detection which is subject to noise. Therefore, it often needs human to define rough boundary outline which makes the approach to be applicable only to specific domain, e.g., image processing tools. Segmentation algorithms based on statistical models have also been proposed. Among them, Blobworld is widely used. In Blobworld, each pixel is represented by an 8 dimensional feature vector of colour, texture and position. An image pixel is then modelled as a random variable with Gaussian mixture distributions. The number of regions and Gaussian parameters are calculated using *expectation maximisation* (EM) algorithm. Once the model parameters are found, the pixel–region relationship is calculated using the posterior probabilities. The pixel–region relationship is used to determine the image segmentation. One of the major issues with this approach is that the computation is very expensive because the EM is an optimisation algorithm.

The widely used JSEG algorithm is a region growing approach. It groups pixels or smaller regions into larger regions. At first, pixel colours of the image are quantised into a number of classes and pixels in the image are replaced with the colour class labels. A class map is formed and region growing is followed on the class map. Pixels with more homogeneous neighbours are assumed to be interior pixels of possible regions. These pixels are selected as candidate seed points and regions are grown around these seed areas. As this method looks for both colour and texture homogeneity, the segmented regions have highly homogeneous characteristics. It has been widely used in image retrieval.

Image segmentation is a complex issue and a large research topic. Segmentation performance usually depends on applications. For image retrieval purpose, the region boundary does not have to be accurate as long as the region is homogenous. However, regions from segmentation are usually contaminated with segments from neighbouring regions. This problem can be overcome by a clean-up post-processing.

B. Colour features

Colour is one of the most important features of images. Colour features are defined subject to a particular colour space or model. A number of colour spaces have been used in literature such as, RGB, LUV, HSV, HMMD.

Once the colour space is specified, colour feature is extracted from images or regions. A number of important colour features have been proposed in the literatures, including colour histogram, colour moments (CM), *colour coherence vector* (CCV), colour correlogram, etc. MPEG-7 also standardizes a number of colour features including *dominant colour descriptor* (DCD), *colour. layoutdescriptor* (CLD), *colour structure descriptor* (CSD), and *scalable colour descriptor* (SCD).Colour moments are one of the simplest features. They are used in many retrieval systems. The common moments are mean, standard deviation and skewness. Usually they are calculated for each colour channels (components) separately. Therefore, nine features form the feature vector. These features are useful when they are calculated for region or object. However, the moments are not enough to represent all the colour information of an image.

C. Texture features

Texture is another important image feature. While colour is usually a pixel property, texture can only be measured from a group of pixels. Due to its strong discriminative capability, texture feature is widely used in image retrieval and semantic learning techniques. Texture has been well studied in image processing and

computer vision area. A number of techniques have been proposed to extract texture features. Based on the domain from which the texture feature is extracted, they can be broadly classified into spatial texture feature extraction methods and spectral texture feature extraction methods. In the following, we describe these techniques.

1. Spatial texture feature extraction methods

In spatial approach, texture features are extracted by computing the pixel statistics or finding the local pixel structures in original image domain. The spatial texture feature extraction techniques can be further classified as structural, statistical and model based.

Structural techniques describe textures using a set of texture primitives (texon or texture elements) and their placement rules. Textons are organised into a string descriptor, and syntactical pattern recognition techniques are used to find similarity of two descriptors.

Statistical texture feature characterises texture as a measure of low level statistics of grey level images. The common spatial domain statistical features are moments, Tamura texture features and features derived from grey level co-occurrence matrix (GLCM). Statistical features are compact and robust because they are derived from large support.In model based techniques, texture is interpreted using stochastic (random) or generative models. Model parameters characterize the underlying texture property of the image. Popular texture models are Markov random field (MRF), simultaneous auto-regressive (SAR) model, fractal dimension (FD), etc. As these models involve optimisation, they are usually computationally expensive.

Spatial texture methods are easy to understand and many of them even have semantics. They do not require regular region shape and can be applied to irregular regions straightforwardly. However, these features are usually sensitive to noise and distortions. Furthermore, many of these methods involve complex search and optimisation processes which have no general solutions.

2. Spectral texture feature extraction techniques

In spectral texture feature extraction techniques, an image is transformed into frequency domain and then feature is calculated from the transformed image. The common spectral techniques include Fourier transform (FT) , discrete cosine transform(DCT) wavelet, and Gabor filters. FT and DCT are very fast to compute but are not scale and rotation invariant. Wavelet is both efficient and robust, but it only captures horizontal and vertical features. Among them, Gabor features are most robust because it captures image features in multi-orientations and multi-scales. Recently, researches on multi-resolution analysis have shown that curvelet features have significant advantages over Gabor features and wavelet features, because curvelet features are more effective in capturing curvilinear properties, like lines and edges.

The problem with those spectral methods is that they can only be applied to square regions due to the use of FFT. Most of the existing region based techniques define a region as a set of small blocks of size 4×4 pixels and apply spectral transform on those blocks , because small blocks are likely homogenous. Features of a region are then calculated as the average features of those blocks.

VI. Summarization

Background

Text summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. In principle, TS is possible because of the naturally occurring redundancy in text and because important (salient) information is spread unevenly in textual documents. Identifying the redundancy is a challenge that hasn't been fully resolved yet.

There is no single definition for salience and redundancy given that different users of summaries may have different backgrounds, tasks, and preferences. Salience also depends on the structure of the source documents. Since information that the user already knows should not be included in a summary and at the same time information that is salient for one user may not be for another, it is very difficult to achieve consistent judgments about summary quality from human judges. This fact has made it difficult to evaluate automatic summarization.

Taxonomically one can distinguish among the following types of summaries: extractive/non-extractive, generic/query-based, single-document/ multi-document, and monolingual/multilingual/ crosslingual. Most existing summarizers work in an extractive fashion, selecting portions of the input documents (e.g., sentences) that are believed to be more salient. Non-extractive summarization includes dynamic reformulation of the extracted content, involving a deeper understanding of the input text, and is therefore limited to small domains. Query-based summaries are produced in reference to a user query (e.g., summarize a document about an international summit focusing only on the issues related to the environment) while generic summaries attempt to identify salient information in text without the context of a query. The difference between single- and multi-document summarization (SDS and MDS) is quite obvious, however some of the types of problems that occur in MDS are qualitatively different from the ones observed in SDS: e.g., addressing redundancy across information sources and dealing with contradictory and complementary information.

A number of evaluation techniques for summarization have been developed. They are typically classified into two categories. Intrinsic measures attempt to quantify the similarity of a summary with one or more model summaries produced by humans. Intrinsic measures include Precision, Recall, Sentence Overlap, Kappa, and Relative Utility. All of these metrics assume that summaries have been produced in an extractive fashion. Extrinsic measures include using the summaries for a task, e.g., document retrieval, question answering, or text classification.

A. Keyphrase Extraction

Task description and example

The task is the following. You are given a piece of text, such as a journal article, and you must produce a list of keywords or keyphrases that capture the primary topics discussed in the text. In the case of research articles, many authors provide manually assigned keywords, but most text lacks pre-existing keyphrases. For example, news articles rarely have keyphrases attached, but it would be useful to be able to automatically do so for a number of applications discussed below. Consider the example text from a recent news article:

"The Army Corps of Engineers, rushing to meet President Bush's promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Press".
An extractive keyphrase extractor might select "Army Corps of Engineers", "President Bush", "New Orleans", and "defective flood-control pumps" as keyphrases. These are pulled directly from the text. In contrast, an abstractive keyphrase system would somehow internalize the content and generate keyphrases that might be more descriptive and more like what a human would produce, such as "political negligence" or "inadequate protection from floods". Note that these terms do not appear in the text and require a deep understanding, which makes it difficult for a computer to produce such keyphrases. Keyphrases have many applications, such as to improve document browsing by providing a short summary. Also, keyphrases can improve information retrieval — if documents have keyphrases assigned, a user could search by keyphrase to produce more reliable hits than a full-text search. Also, automatic keyphrase extraction can be useful in generating index entries for a large text corpus.

Keyphrase extraction as supervised learning

Beginning with the Turney paper, many researchers have approached keyphrase extraction as a supervised machine learning problem. Given a document, we construct an example for each unigram, bigram, and trigram found in the text. We then compute various features describing each example (e.g., does the phrase begin with an upper-case letter?). We assume there are known keyphrases available for a set of training documents. Using the known keyphrases, we can assign positive or negative labels to the examples. Then we learn a classifier that can discriminate between positive and negative examples as a function of the features. Some classifiers make a binary classification for a test example, while others assign a probability of being a keyphrase. For instance, in the above text, we might learn a rule that says phrases with initial capital letters are likely to be keyphrases. After training a learner, we can select keyphrases for test documents in the following manner. We apply the same example-generation strategy to the test documents, then run each example through the learner. We can determine the keyphrases by looking at binary classification decisions or probabilities returned from our learned model. If probabilities are given, a threshold is used to select the keyphrases. Keyphrase extractors are generally evaluated using precision and recall. Precision measures how many of the proposed keyphrases are actually correct. Recall measures how many of the true keyphrases your system proposed. The two measures can be combined in an F-score, which is the harmonic mean of the two ($F = 2PR/(P+R)$). Matches between the proposed keyphrases and the known keyphrases can be checked after stemming or applying some other text normalization..

Document summarization

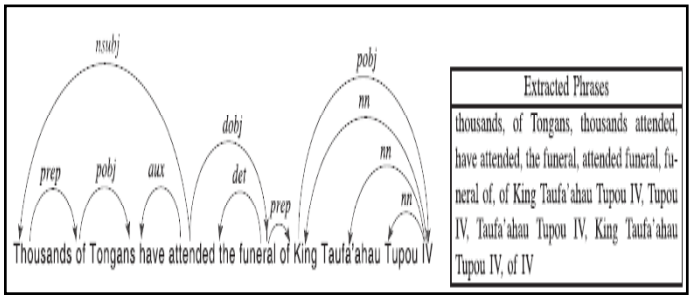
Like keyphrase extraction, document summarization hopes to identify the essence of a text. The only real difference is that now we are dealing with larger text units—whole sentences instead of words and phrases.
Before getting into the details of some summarization methods, we will mention how summarization systems are typically evaluated. The most common way is using the so-called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure. This is a recall-based measure that determines how well a system-generated summary covers the content present in one or more human-generated model summaries known as references. It is recall-based to encourage systems to include all the important topics in the text. Recall can be computed with respect to unigram, bigram, trigram, or 4-gram matching, though ROUGE-1 (unigram matching) has been shown to correlate best with human assessments of system-generated summaries (i.e., the summaries with highest ROUGE-1 values correlate with the summaries humans deemed the best). ROUGE-1 is computed as division of count of unigrams in reference that appear in system and count of unigrams in reference summary.

Algorithm

Extractive and Abstractive Caption Generation

Extractive and Abstractive Caption:

Extractive caption generator draws inspiration from previous work on automatic summarization, most of which focuses on sentence extraction . The idea is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document. Without a great deal of linguistic analysis, it is possible to create summaries for a wide range of documents, independently of style, text type, and subject matter. For our caption generation task, we need only extract a single sentence. And our guiding hypothesis is that this sentence must be maximally similar to the description keywords generated by the annotation model.
The 'phased' nature of extractive activities means that many, or all, phases of an extractives operation –exploration, evaluation, development, production and decommissioning - can occur at the same time. Production will routinely commence while development continues. Exploration and evaluation designed to identify and expand reserves (or to 'prove up' reserves) will continue. Major expansionary expenditure is often deferred to maximise cash flows (e.g., extending a decline or shaft, up-scaling plant capacity, expanding well heads), only being completed as required throughout the life of the operation.



The design of the mining or oil and gas operation seeks to maximise returns from the operation overall and does not explicitly consider individual 'expansions' as production continues, e.g., in a mining operation, ore may be uneconomic on its own but 'blended' with other ore to achieve optimal ore grades for processing, or otherwise uneconomic ore in an underground mine, may be mined in gaining access to high grade ore; and the initial infrastructure is constructed in view of the final scale of the operation, and therefore it may not be supported by cash flows expected from accessible reserves which can currently be extracted.
We define a bag-of-phrases model for caption generation by modifying the content selection and caption length components in (18) as follows:

$$P(\rho_1, \rho_2, \ldots, \rho_m) \approx \prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in \mathcal{D}) \cdot P\left(len(C) = \sum_{j=1}^{m} len(\rho_j)\right) \cdot \prod_{i=3}^{L} P_{adap}(w_i | w_{i-1}, w_{i-2}),$$

The term P(ᵖj ϵ C|ᵖj ϵ D) models the probability of phrase ᵖj appearing in the caption given that it also appears in the document and is estimated as

$$P(\rho_j \in C | \rho_j \in \mathcal{D}) = \prod_{w_j \in \rho_j} P(w_j \in C | w_j \in \mathcal{D}),$$

where wj is a word in the phrase pj
We therefore attempt to take phrase adjacency constraints into account by estimating the probability of phrase pj attaching to the right of phrase pi as

$$P(\rho_j|\rho_i) = \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} p(w_j|w_i) = \frac{1}{2} \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} \left\{ \frac{f(w_i, w_j)}{f(w_i, -)} + \frac{f(w_i, w_j)}{f(-, w_j)} \right\},$$

After integrating the adjacency probabilities into , the caption generation model becomes

$$P(\rho_1, \rho_2, \ldots, \rho_m) \approx \prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in \mathcal{D}) \cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1}) \cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j)) \cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}).$$

VII. Evaluation

We evaluated the performance of our models automatically, and also by eliciting human judgments. Our automatic evaluation was based on Translation Edit Rate (TER, Snover et al. 2006), a measure commonly used to evaluate the quality of machine translation output. We chose to use TER over other metrics with similar properties such as BLEU (Papineni et al., 2002) since it can account for word reordering and be applied to individual sentences without any adjustments. TER is defined as the minimum number of edits a human would have to perform to change the system output so that it exactly matches a reference translation. In our case, the original captions written by the BBC journalists were used as reference:

$$\text{TER}(E, E_r) = \frac{\text{Ins} + \text{Del} + \text{Sub} + \text{Shft}}{N_r}$$

where E is the hypothetical system output, Er the reference caption, and Nr the reference length. The number of possible edits include insertions (Ins), deletions (Del) substitutions (Sub) and shifts (Shft). TER is similar to word error rate, the only difference being that it allows shifts. A shift moves a contiguous sequence to a different location within the the same system output and is counted as a single edit. The perfect TER score is 0, however note that it can be higher than 1 due to insertions. The minimum translation edit alignment is usually found through beam search. We used TER to compare the output of our extractive and abstractive models with the original captions and also for parameter tuning.
In our human evaluation study, participants were presented with a document, an associated image, and its caption, and asked to rate the latter on two dimensions :grammaticality (is the sentence fluent or word salad?) and relevance (does it describe succinctly the content of the image and document?). We used a 1–7 rating scale, participants were encouraged to give high ratings to captions that were grammatical and appropriate descriptions of the image given the accompanying document. We randomly selected 12 document-image pairs from the test set and generated captions for them using the best extractive system (KL divergence based), and two abstractive systems(word-based and phrase-based). We also included the original human-authored caption as an upper bound.

VIII. Conclusion

In this paper we have presented a novel task, automatic caption generation for news images, and proposed extractive and abstractive models. A key aspect of our approach is to allow both the visual and textual modalities to influence the generation task. This is achieved through an image annotation model that characterizes pictures in terms of description keywords that are subsequently used to guide the caption generation process. Our results show that the visual information plays an important role in content selection. Simply extracting a sentence from the document often yields an inferior caption. We focus on captioned images embedded in news articles, and learn both models of content selection and surface realization from data without requiring expensive manual annotation. At training time, our models learn from images, their captions, and associated documents, while at test time they are given an image and the document it is embedded in and generate a caption. Compared to most work on image description generation, our approach is shallower, it does not rely on dictionaries specifying image-to-text correspondences, nor does it use a human-authored grammar for the caption creation task. It uses the document co-located with the image as a proxy for linguistic, visual, and world-knowledge. Our innovation is to exploit this implicit information and treat the surrounding document and caption words as labels for the image, thus reducing the need for human supervision.

References

[1] *Vailaya, M. Figueiredo, A. Jain, and H. Zhang,"Image Classification for Content-Based Indexing," IEEE Trans. ImageProcessing, vol. 10, no. 1, pp. 117-130, 2001.*
[2] *A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based ImageRetrieval at the End of theEarly Years, "IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12,pp. 1349-1380, Dec. 2000.*
[3] *P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," Proc. Seventh European Conf. Computer Vision, pp. 97-112, 2002.*
[4] *D. Blei, "Probabilistic Models of Text and Images," PhD dissertation, Univ. of Massachusetts, Amherst, Sept. 2004.*
[5] *K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei,*

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

and M.Jordan, "Matching Words and Pictures," J. Machine Learning Research, vol. 3, pp. 1107-1135, 2002.

[6] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," Proc.IEEE Conf. Computer Vision and Pattern Recognition, pp. 1903-1910, 2009.

[7] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," Proc. 16th Conf. Advances in Neural Information Processing Systems, 2003.

[8] S. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli Relevance Models for Image and Video Annotation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.

[9] A. Kojima, T. Tamura, and K. Fukunaga, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions," Int'l J. Computer Vision, vol. 50,no. 2, pp. 171-184, 2002.

[10] 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.

[11] 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.

[12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L.Berg, "Baby Talk: Understanding and Generating Image Descriptions,"Proc. IEEE Conf. Computer Vision and Pattern Recognition,pp. 1601-1608, 2011.

[13] C.-F. Chai and C. Hung, "Automatically Annotating Images with Keywords: A Review of Image Annotation Systems," Recent Patents on Computer Science, vol. 1, pp. 55-68, 2008.

[14] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, vol. 41, no. 2, pp. 177-196,2001.

[15] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," IEEE Trans. Pattern Analysis andMachine Intelligence, vol. 29, no. 10, pp. 1802-1817, Oct. 2007.

[16] J.-Y. Pan, H.-J. Yang, and C. Faloutsos, "MMSS:Multi-Modal Story-Oriented Video Summarization," Proc. Fourth IEEE Conf.Data Mining, pp. 491-494, 2004.

[17] H. Schmid, "Probabilistic Part-of-Speech Tagging Using DecisionTrees," Proc. Int'l Conf. New Methods in Language Processing, 1994.

[18] 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.

[19] C. Buckley and E.M. Voorhees, "Retrieval System Evaluation,"TREC: Experiment and Evaluation in Information Retrieval,E.M. Voorhees and D.K. Harman, eds., pp. 53-78, MIT Press, 2005.

[20] E.W. Noreen, Computer-Intensive Methods for Testing Hypotheses: An Introduction. John Wiley & Sons, Inc., 1989.