ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 7, Issue, 7, pp. 12497-12499, July, 2016

International Journal of Recent Scientific Research

Research Article

AUTOMATIC CAPTION GENERATION

Nivetha G., GeethaPriya K and Vanitha M

Department of Computer Science (Final year), Vel Tech High Tech Dr.Rangaranjan Dr.Sakunthala Engineering College, Avadi

ARTICLE INFO

Article History:

Received 20th April, 2016
Received in revised form 29th May, 2016
Accepted 30th June, 2016
Published online 28th July, 2016

Key Words:

Qatari Fish, Marine Parasites, Fish diseases, Aquaculture, Fisheries, Arabian Gulf.

ABSTRACT

The novel task of automatically generating caption which fuses insights from computer vision and natural language processing and holds future for different multimedia applications, such as image retrieval, development of tools supporting various fields of media management. It is possible to learn a caption generation model from weakly labeled data without costly human involvement. Instead of manually creating annotations, captions are treated as labels for the image. Although the captions generated are noisy compared to human-created keywords, we show that they can be used to learn the relation between visual and textual modalities, and also serve as a optimum for the caption generation task. We have presented caption generation models using the content selection and surface generation. A key aspect of our approach is to allow both the graphic and textual modalities to influence the caption generation task.

INTRODUCTION

Our approach leverages the vast resource of images available on the web and the fact that many of them are captioned and collocated 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, we 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 methods. Content selection suggests keywords for the image using image annotation model.

Content selection it decides what information to be included in the text. It includes document structuring showing how to organize text. Lexicalisation is the process of choosing particular words or phrases. Then Referring expression generation takes place deciding what properties should be used in referring to an entity. It practically constructs a set of messages from the underlying data (entities, concepts and relations).

Surface realization is the process of mapping underlying content of text to a grammatically correct sentence that expresses the desired meaning. Realization is also a subtask of natural language generation, which involves creating an actual text in a human language (English, French, etc.) from a syntactic representation. There are a number of software packages available for realization Tasks of surface realization includes Sentence planning (micro-planning) and Surface realizer (proper). Sentence planning includes word and syntax selection. Surface realizer is the task of creating linear text from structured input

System Analysis

Many of the 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.

Corresponding author: **Nivetha G*
Department of Computer Science (Final year), Vel Tech High Tech Dr.Rangaranjan Dr.Sakunthala Engineering College, Avadi

The web retrieve images without analyzing their content, simply by matching user queries against collocated textual information.

Images that do not coincide with textual data cannot be retrieved

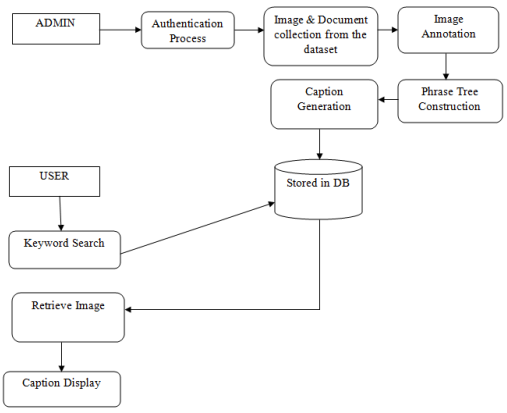
Human Authored Grammar

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. 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.

Advantages

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

Architecture



Implementation Methodology

1. Data Collection
2. Input preparation
3. Abstractive caption
4. Extractive caption

Data Collection

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

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

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.

Extractive Caption

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.

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 (and hence, improve) 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. No true multilingual summarization systems exist yet, however, cross-lingual approaches have been applied successfully.

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.

Traditionally, summarization has been mostly applied to two genres of text: scientific papers and news stories. These genres are distinguished by a high level of stereotypical structure. In both these domains, simply choosing the first few sentences of a text or texts provides a baseline that few systems can better and none can better by much. Attempts to summarize other texts, e.g., fiction or email, have been somewhat less successful.

Recently, summarization researchers have also investigated methods of text simplification (or compression). Typically, these methods apply to a single sentence at a time. Simple methods include dropping unimportant words (determiners, adverbs). Complex methods involve reorganizing the syntactic parse tree of the sentence to remove sections or to rephrase units in shorter form. Language modeling approaches in TS have mostly focused on this method.

CONCLUSION

The task fuses insights from computer vision and natural language processing and holds promise for various multimedia applications, such as image and video retrieval, development of tools supporting news media management, and for individuals with visual impairment. As a departure from previous work, we have approached this task in a knowledge-lean fashion by leveraging the vast resource of images available on the Internet and exploiting the fact that many of these co-occur with textual information (i.e., captions and associated documents). Our results show that it is possible to learn a caption generation model from weakly labeled data without costly manual involvement The dataset we employed contains real-world images and exhibits a large vocabulary including both concrete object names and abstract keywords; 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. Moreover, this news dataset contains a unique component, the news document, which provides both information regarding to the image's content and rich linguistic information required for the generation procedure.

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. A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early 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, and M.Jordan, "Matching Words and Pictures," *J. Machine Learning Research*, vol. 3, pp. 1107-1135, 2002.

****

How to cite this article:

Nivetha G., GeethaPriya K and Vanitha M.2016, Automatic Caption Generation. *Int J Recent Sci Res.* 7(7), pp. 12497-12499.