AutoCAP: An Automatic Caption Generation System based on the Text Knowledge Power Series Representation Model

Krishnapriya P S
M.Tech Dept of CSE
NSS College of Engineering
Palakkad, Kerala

Usha K
Associate Prof. Dept of CSE
NSS College of Engineering
Palakkad, Kerala

ABSTRACT
This paper describes Automatic Caption generation for news Articles, it is an experimental intelligent system that generates presentations in text based on the text knowledge power series representation model. Captions or titles are useful for users who only need information on the main topics of an article. Using current extractive summarization techniques, it is not able to generate a coherent document summary shorter than a single opinion, or to produce a brief that conforms to specific linguistic constraints. The power series representation (PSR) model, which has a low complex computation in text knowledge constructing process. This model can provide rich knowledge and automatic construction. Our model experience to design captions from a database of news articles, and from the associated images in them, and their captions, and consists of two stages. Text assertions is that the keywords/terms that stand for the common knowledge and Text association rules, are referred to these term's relations that are causations and reflect the semantic relationships in a text. It is viable to generate captions from the news by mapping the PSR and the image associated with the news article.

General Terms
Power series representation model.

Keywords
Natural language generation, Computer vision, human concept learning, knowledge representation.

1. INTRODUCTION
We introduce a caption generation system that selects caption words throughout the entire text and the associated labeled image. Natural language, whether spoken, written, or typed, makes up much of human communication. A significant amount of this language describes the visual world either directly around us or in images and video. Connecting visual imagery with visually descriptive language is a challenge for computer vision that is becoming more relevant as recognition and detection methods are beginning to work. Humans can prepare a concise description in the form of a sentence relatively comfortably. Such characteristics might associate the most interesting objects, what they are doing, and where this is happening. These representations are rich, because they are in sentence form. They are accurate, with good agreement between annotators. They are succinct much is neglected, because humans offer not to mention objects or events that they judge to be less significant. Finally, they are consistent.

The power series representation model [1] is a standard approach to text description generation. The standard approach to text description generation adopts a two-stage framework consisting of text assertion and text association rules . The former stage analyzes the terms, whose term frequencies are very high, will be seen as text assertions. whereas the second stage determines the rules, that is the text association rules are referred to these terms' relations that are causations and reflect the semantic relationships in a text. Both stages are usually manually developed.

This approach can create sentences of high quality that are both meaningful and fluent. In this paper, we tackle the related problem of generating captions automatically for news articles. 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 automatic captioning without requiring expensive manual annotation. At training time, our models learn from images and privileged documents, while at trial time they are given an image and the document it is surrounded in and procreate the caption. Our innovation is to exploit this entangled information and treat the surrounding document and caption words as signs for the image, thus reducing the need for human supervision. However, we argue that the superfluity inherent in such a multimodal dataset allows the development of a fully unsupervised caption generation system.

2. RELATED WORK
Most previous work on summarization focused on extractive methods, investigating issues such as cue phrases (Luhn, 1958), positional indicators (Edmundson, 1964), lexical occurrence statistics (Mathis et al., 1973), probabilistic measures for token salience (Salton et al., 1997), and the use of implicit discourse structure (Marcu, 1997). Work on combining an information extraction phase followed by generation has also been reported: for instance, the FRUMP system (DeJong, 1982) used templates for both in.

He´de et al. [2] generate descriptions for images of objects shot in unvarying background. Their system depends on a manually created database of objects indexed by an image signature (e.g., color and texture) and two keywords (the object's name and collection of objects). Images are first segmented into objects, their approval is regained from the database, and portraiture is generated using patterns. Other work (e.g., [4]) generates representations for human activities in office scenes. The idea is to prepare features of human motion from video key frames and interleave them with a concept hierarchy of actions to create a case frame from which a natural language sentence is generated. Yao et al. [3] present a general framework for generating text descriptions of image and video content based on image parsing. A multi sentence description is generated using a document planner and a surface realizer. The task of learning visual models from

images and videos with accompanying captions can be naturally casted into this framework. with tags [5], news photos with captions [6], and movies with scripts [7]. Among them, images with captions are especially interesting, as they contain richer information. Berg et al. [8] and Guillaumin et al. [6] show that names extracted from news captions using natural language processing can be used to cluster the faces appearing in news images.

Early work on connecting words and pictures focused on associating individual words with image regions [9], [10] for tasks such as clustering, auto-annotation or auto-illustration. Other work has made use of text as a source of noisy labels for predicting the content of an image. This works especially well in constrained recognition scenarios for recognizing particular classes of objects such as for labeling faces in news photographs with associated captions [11]. The keywords of images and the frequent words in sentences are the important parts.

3. POWER SERIES REPRESENTATION MODEL

3.1 The model

Power series representation conceptually modeled as consisting of two major sub-tasks: text assertion and text association rules. Parameters for statistical models of both of these tasks were estimated from a training corpus of approximately 25,000 Reuters news-wire articles on politics, technology, health, sports and business. The target documents that the system needed to learn the translation mapping to, were the headlines accompanying the news articles. The documents were preprocessed before training: formatting and mark-up information, such as font changes and SGML/HTML tags, was removed; punctuation, except apostrophes, was also removed.

3.2 Human concept learning and concept algebra

Based on Feldman's linearity hypothesis [14], [15] take an example the "world" can be represented as W. The objects represented as {σ1 , σ2 , . . . , σd} corresponds to a concept of a human concept learning process. This called as a

"language" to express the structure of this "world" W. Also there is α, β, ω notations, where α is for the set of constant properties, i.e., the affirmations in "world" W; ω for the set of paired implications; and β for the rest of "world" W, the unconstrained properties. Every possible "world," which can be produced by these types of concepts alone, can be expressed as the Cartesian product of the lattice for ω with the lattice for β, conjoined with the properties α. Equation illustrates that human concept learning is a linear process.

$$\alpha.[\omega \times \beta], W \Big|_{\Sigma} \subset \alpha.[\omega \times \beta]$$

3.3 Text representation

In consideration of object attribute relation model [16][17], a text can be regarded as a concept, whereas sentences that are contained in the text can be regarded as objects belonging to the text and the terms in sentences are the attributes of these objects. In this paper, we regard the terms as attributes of describing text knowledge, a sentence or paragraph as an object, therefore a text which consists of many sentences and paragraphs can be regarded as a concept. Then, we represent the text knowledge according to the linearity hypothesis of human concept learning. Thus, based on the similarity between the concept and the text, an example about text fragments is given as follows: we find a property language text , which includes all the properties from the text fragment including **S1** and **S2** .

S1 : That *girl* stands on the *right*, whose *skirt* is *green*.

S2 : Two *boys* stand on the *left*, whose *t-shirts* are also *green*.

3.4 Text assertion

Text assertion is the keywords/terms that stand for the common knowledge in a text are referred as text assertions. Text assertions have the simplest and the most easily understood information in a text. Moreover, the system learn a model of the relationship between the appearance of some features in a document and the appearance of corresponding features in summary. In this paper, text assertions are also defined as the keywords/terms that often appear in a text. In another word, the terms, whose term frequencies are very high, will be seen as text assertion.

3.5 Text association rules

Text association rules are referred to these terms' relations that are causations and reflect the semantic relationships in a text. Generally, text association rules are mined from those terms that are adjacent to each other and own a high co-occurrence appearance in a text.

1) For every text, calculate each term's frequency, by which the terms are ranked in descending order.

2) According to Salton *et al.* [18], the terms with the highest term frequencies are always the top 4% in the ranked list.

4. AUTOMATIC CAPTION GENERATION

The power series representation (PSR) model, which has a low complex computation in text knowledge building process, is intended to influence the contradiction between carrying rich knowledge and automatic construction. The model acquires to create captions from a database of news articles, the images securely surrounded in them, and their captions, and consists of two stages. Text assertions is that the keywords/terms that stand for the common knowledge and Text association rules, are referred to these term's relations that are causations and reflect the semantic relationships in a text. It is viable to generate captions from the news by mapping the PSR and the image associated with the news article.

We have to compose the words into readable captions. News article captions mostly use words from the beginning of the text, also stated in (Zajic et al., 2002). The top-scoring words over the whole story are selected and highlighted. From the given news the frequent words are selected and it will further checked with the keywords of images. Then the system determines which is needed for the generation task.

"We are delighted to bring to Tamil Nadu and Chennai, the M-Pesa, a mobile banking service, in partnership with ICICI Bank. M-Pesa, is a safe, secure and convenient service to transfer money and make payments beyond the reach of traditional banking channels," Mr. Mehrotra said.Observing that less than five per cent of villages in the country have access to banking services, Vodafone India, Business Head (M-Pesa), Suresh Sethi said: "Financial inclusion is a national priority and we believe that with M-Pesa, we now have the ideal offering to facilitate the same across the country in compliance with applicable regulations." The service would be available across 5,103 authorised agents, including 840 Vodafone exclusive stores, in Tamil Nadu, including Chennai. Vodafone has launched and M-Pesa services in various telecom circles, including New Delhi, Mumbai, Kolkata, West Bengal, Punjab, Uttar Pradesh East and Western regions, Bihar.	 Vodafone India, 840 Vodafone exclusive Chennai. Vodafone has launched and banking services and M-Pesa services in telecom circles.

Fig 1: Automatic caption generation

5. CONCLUSION

We presented a new large-scale database of images and captions, for the automatic caption generation task. It is designed to be representative of the challenges in learning automatically from realistic image and caption pairs mined freely from the Internet. Using the power series representation model the complications are removed. The models such as CTM, LDA has complicated computation and even lack the ability of flexible knowledge-based reasoning. Using VSM lose many text knowledge and OWL cannot be constructed automatically. Extensive experiments on the whole and various subsets of the database show the merits of our database. The task fuses 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.

6. REFERENCES

[1] 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.

[2] 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'InformationAssiste´e par Ordinateur, 2004.

[3] 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.

[4] 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.

[5] Y. Wang and G. Mori. A discriminative latent model of image region and object tag correspondence. In NIPS, 2010.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In CVPR, 2008.

[7] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In CVPR, 2009.

[8] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who's in the picture. In NIPS, 2004.

[9] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan, "Matching Words and Pictures," J. Machine Learning Research, vol. 3, pp. 1107-1135, 2003.

[10] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object Recognition as Machine Translation," Proc. European Conf. Computer Vision, 2002.

[11] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, E. Learned- Miller, Y.-W. Teh, and D.A. Forsyth, "Names and Faces," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[12] K. Saenko and T. Darrell, "Unsupervised Learning of Visual Sense Models for Polysemous Words," Proc. Neural Information Processing Systems, 2008