Improving Automatic Image Captioning Using Text Summarization Techniques

Laura Plaza[†], Elena Lloret[‡], and Ahmet Aker[*]

[†]Universidad Complutense de Madrid, C/Prof. José García Santesmases, s/n 28040 Madrid, Spain
lplazam@fdi.ucm.es
[‡]University of Alicante, Apdo. de correos, 99, E-03080 Alicante, Spain
elloret@dlsi.ua.es
[*]University of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, UK
a.aker@dcs.shef.ac.uk

Abstract. This paper presents two different approaches to automatic captioning of geo-tagged images by summarizing multiple web-documents that contain information related to an image's location: a graph-based and a statistical-based approach. The graph-based method uses text cohesion techniques to identify information relevant to a location. The statistical-based technique relies on different word or noun phrases frequency counting for identifying pieces of information relevant to a location. Our results show that summaries generated using these two approaches lead indeed to higher ROUGE scores than n-gram language models reported in previous work.

1 Introduction

The number of images with location information is growing exponentially with the rapid development of online photo sharing services and increasing prevalence of camera phones with embedded GPS and compass. Additionally, many legacy photographs and other images are stored or tagged by place names or contain minimal captions that include geographical information. In all these cases the small or non-existent amount of textual information associated with the image is of limited usefulness for image indexing, organization, and/or search. What would be useful is a means to automatically generate or augment captions for images from their geo-referencing information.

Aside from application to image indexing, organization and search, the capability to automatically caption geo-referenced images has further potential applications. It could, for instance, help users gain quick access to the information they seek about a place of interest just by taking its picture. Such textual information could also be used by a journalist who is planning to write an article about a building, or by a tourist who seeks further places to visit nearby.

Attempts towards automatic generation of image captions have been previously reported. Deschacht & Moens [1] and Mori et al. [2] generate image captions automatically by analyzing image-related text from the immediate context of the image, e.g. the surrounding text in HTML documents. The authors identify named entities and other noun phrases in the image-related text and assign these to the image as captions. Other attempts towards automatic generation of image captions generate captions based on the immediate textual context of the image with or without consideration of image related features such as colour, shape or texture [1–9]. However, Marsch & White [10] argue that the content of an image and its immediate text have little semantic agreement and this can, according to Purves et al. [11], be misleading to image retrieval. Furthermore, these approaches assume that the image has been obtained from a document. In cases where there is no document associated with the image, which is the scenario we are principally concerned with, these techniques are not applicable.

Following the general approach proposed by Aker and Gaizauskas [12], in this paper we describe different methods for automatic image captioning starting with only a set of place names pertaining to an image – for example ⟨ {St. Paul's, London}⟩. Place names can be obtained automatically given GPS coordinates and compass information using techniques such as those described in Xin et al. [13] – that task is not the focus of this paper.

Aker and Gaizauskas [12] have argued that humans appear to have a conceptual model of what is salient regarding a certain object type (e.g. church, bridge, etc.) and that this model informs their choice of what to say when describing an instance of this type. They also experimented with representing such conceptual models using n-gram language models derived from corpora consisting of collections of descriptions of instances of specific object types (e.g. a corpus of descriptions of churches, a corpus of bridge descriptions, and so on) and reported results showing that incorporating such n-gram language models as a feature in a feature-based extractive summarizer improves the quality of automatically generated summaries. However, the authors report that the quality of language model biased summaries was still not at satisfactory level.

In this paper we experiment with two different approaches to generate summaries related to images: a graph-based and a statistical-based approach. The graph-based method enriches the words in the documents with further concepts and relations from WordNet to better capture syntactic different but semantically similar feature descriptions used to described places. The statistical-based approach concentrates on long descriptive noun phrases frequency count to identify salient feature descriptions. Our results show that our methods indeed score better than the ones reported in Aker and Gaizauskas [12].

In the following we first describe the set of images, their model summaries and the retrieval of related web-documents (section 2). In section 3 we present the summarizers used to caption images. We discuss the results of evaluating automatic summaries against the human created captions in section 4, and draw conclusions and future lines of work in section 5.

2 Corpus

For evaluation we use the image collection described in Aker and Gaizauskas [14]. The image collection contains 308 different images with manually assigned place names. For each image there are up to four short descriptions or model summaries. The model summaries were created manually based on image descriptions taken from *VirtualTourist* and contain a minimum of 190 and a maximum of 210 words.

To generate automatic captions for the images, Aker and Gaizauskas [12] automatically retrieved the top ten related web-documents for each image using the Yahoo! search engine and the place name associated with the image as a query. The text from these documents was extracted using an HTML parser and passed to their summarizer. We also used these documents to generate image captions.

3 Summarizers

3.1 A semantic-graph based summarizer

The summarizer has been already presented in previous work and evaluated in two different domains: news items and biomedical papers [15]. In this paper, we focus on image caption generation. First, it should be noted that the system was not originally designed to deal with multi-document summarization. To overcome this shortcoming, we simply merge all documents about the same topic into a single document, and run the summarizer over it. After producing the summary, we apply a textual entailment module to detect and remove redundancy [16]. The summarizer first applies a shallow preprocessing over the document, including sentence detection, POS tagging and removing stopwords and high frequency terms. It next translates the text in the document to WordNet concepts, using the *lesk* algorithm (as implemented in the WordNet Sense Relate package [17]) to disambiguate the meaning of each term in the document according to its context. After that, the resulting WordNet concepts are extended with their hypernyms, building a graph representation for each sentence in the document, where the vertices represent distinct concepts in the sentence and the edges represent *is-a* relations. The system then merges all the sentence graphs into a single document graph, which is extended with a further semantic relation, so that every pair of leaf vertices whose similarity (calculated in terms of WordNet concepts gloss overlaps, using the *WordNet Similarity* package [18]) exceeds a certain threshold. Each edge in the document graph is assigned a weight that is directly proportional to the depth in the hierarchy of the nodes that it links (that is, the more specific are the concepts connected by a link, the more weight is assigned to it).

Once the graph is built, the vertices are ranked according to their *salience* or prestige. The salience of a vertex is calculated as the sum of the weight of the edges connected to it. The top n vertices are grouped into *Hub Vertices Sets (HVS)*, which represent sets of concepts strongly related in meaning. These will constitute the centroids of the clusters. A *degree-based clustering* method [19] is then executed over the graph and, as a result, a variable number of clusters or subgraphs are obtained. The working hypothesis is that each of these clusters represents a different *subtheme* or topic within the document, and that the most central concepts in a cluster (the so called HVS) give the necessary and sufficient information related to its topic. The process continues by calculating the similarity between all the sentence graphs and each cluster. To this aim, a non-democratic vote mechanism [20] is used, so that each vertex (v_k) of a sentence (S_j) gives to each cluster (C_i) a different number of votes (w_i, j) depending on whether v_k belongs or not to the HVS of that cluster. The similarity is computed as the sum of the votes given by all vertices in the sentence to each cluster. Finally, under the hypothesis that the cluster with more concepts represents the main theme in the document, and hence the only one that should contribute to the summary, the N sentences with greater similarity to this cluster are selected.

3.2 A Statistical-based Summarizer

In previous work [21], different techniques were shown to be appropriate for the text summarization task. More specifically, three features were analyzed (term-frequency, textual entailment, and the code quantity principle), and the performance of several approaches employing such features on their own, as well as in combination, was investigated. After the research, it was concluded that the combination of all the techniques within the same approach led to the best results, outperforming by approximately 10% the best system in DUC 2002. Therefore, in this paper, we follow the same idea and we take an improved version of this approach as the basis for generating text summaries. However, it should be noted that this approach has been only evaluated over newswire, whereas in this research we focus on a completely different type of documents, *image captions*, which can be considered as one of the many new textual genres born with the Web 2.0. Besides the general assessment of the whole statistical-based summarization approach, the use of such corpus will also allow us to analyze whether the suggested techniques could be domain-independent or not.

Next, each of the features are briefly described. First, a textual entailment tool [16] is used to detect redundant information in the document and, as a consequence, the repeated information is removed. The other features, term-frequency and the code quantity principle[1] are used to measure the importance of each sentence in the document, thus assigning a score to each one based on the frequency of the words that a noun phrase contains and normalizing this value with respect to the number of total noun phrases. A detailed description of these features can be found in [21]. It is worth stressing upon the fact that the summarization approach was originally working only for single-document

[1] The code quantity principle is a linguistic theory that proves the existence of a proportional relation between the importance of a piece of information and the number of text units it contains [22]. In our work, noun phrases are the text units taken into account.

summarization. In order to allow it to deal with several documents, we adapted it in the same way we did in the semantic-graph based approach previously explained.

The summarization process starts with an initial stage where some basic preprocessing is carried out, which includes sentence segmentation, tokenization, part-of-speech tagging, and *stopwords* removal. At this stage, the frequency of each remaining word is counted and stored. Then, the textual entailment module is run over the documents in order to detect potential sentences with repeated information. Further on, once we have the text without redundant information, a relevance sentence detection stage computes a score for each sentence, based on the frequency of the words that appear in a noun phrase. Therefore, sentences not only with longer noun phrases but also with the most frequent words within these noun phrases are considered more important, and consequently, are assigned a higher score. This score is normalized by the number of noun phrases a sentence has. Finally, the summarization selects the highest ranked sentences, and presents them in the same order as they appeared in the original documents, in a preliminary attempt of maintaining the coherence of the text. The final summary is made up from these selected sentences.

4 Results

4.1 Experimental framework

To evaluate both approaches, we use the image caption collection described in Section 2. We generate 200-words long summaries for the images from this collection, each of one is described by ten different documents, and compare the automatic summaries against the model summaries written by humans.

Following the Document Understanding Conferences [23], the ROUGE evaluation metric [24] is used for assessing the summarizers. ROUGE compares automatically generated summaries (called *peers*) against human-created summaries (called *models*), and computes a set of different measures to estimate content coverage in an automatically generated summary. In particular, we compute ROUGE-2 and ROUGE-SU4 recall scores. In short, ROUGE-2 evaluates bi-gram co-occurrence between the peer and model summaries, while ROUGE-SU4 allows bi-grams to have intervening word gaps no larger than four words.

As baseline we generate summaries using the Wikipedia article describing each image, from which we select the first 200 words. We look at these summaries as a difficult goal to achieve: first, it must be taken into account that these articles have been created by humans; second, the first paragraph in a Wikipedia article is usually just a summary of the entire document content; and third, Wikipedia articles almost exclusively contain salient information to the subject matter, and so do not present other information somehow related to the topic but not important (e.g. nearby hotels, restaurants, transport services, or even advertising).

4.2 Results and Discussion

Table 1 shows the ROUGE-2 and ROUGE-SU4 recall values for the Wikipedia baseline summaries as well as for the two suggested approaches: *semantic-graphs* and *statistical-based*. It also includes the best results of the n-gram *language models*, as reported in [12].

Table 1. ROGUE results for the different summarization approaches

	Wikipedia	**Semantic-graphs**	**Statistical-based**	**Language models**
Rouge-2	0.096***	0.089*	0.086	0.071
Rouge-SU4	0.142***	0.142***	0.134	0.119

It can be observed that both systems (semantic-graphs and statistical-based) achieve better results than language models in both ROUGE metrics. Besides, as expected, Wikipedia summaries significantly outperform the other summarizers (Wilcoxon Signed Ranks test[2]). However, the difference between Wikipedia and our two summarizers is less than we would have anticipated, specially in the ROUGE-SU4 score, which seems to indicate that both approaches provide a good approximation to the problem of summarizing information related to tourist images. Regarding the comparison between the two approaches presented here, the semantic-graph based method obtains significantly better for ROUGE-2 and ROUGE-SU4 score, for different confidence intervals.

In the remaining of the section, we will try to elucidate the reasons for the unfavorable differences between the summaries generated by our systems and Wikipedia summaries. Regarding the graph-based approach, the main problem is directly related to the type of the documents to summarize: in most of these documents, the salient information is concerned with proper nouns describing monuments, cities, beaches, etc., that are not likely to be found in WordNet (e.g. *Sacre Coeur*, *Santorini* or *Ipanema*). If no concept is found in the ontology for these terms, the document graph will be inevitably losing essential information to identify the topics covered in the document.

As far as the statistical-based approach is concerned, the main problem lies also in the nature of the corpus. Most documents in the corpus contain sentences with a high number of noun phrases, but which are unrelated to the topic (e.g ‘*‘Mahogany, Maple, crown mouldings, multiple Viking ovens, Sub-Zero refrigerators, antique...* ”). According to the code quantity principle feature, these types of sentences are scored higher, thus being considered relevant to incorporate them to the summary. In these cases, the quality of the generated summaries is directly affected by these sentences.

[2] We use the following conventions for indicating significance level in the tables: *** = p < .0001, ** = p < .001, * = p < .05 and no star indicates non-significance.

5 Conclusion

In this paper we presented two different approaches – a semantic graph-based and a statistical-based approach – for automatically generating image captions from several documents retrieved from the Internet. The former takes into consideration the salience of the WordNet concepts in the text to identify important contents. The latter relies on long descriptive noun phrases together with the frequency of terms to identify relevant information. The results of both systems are highly satisfactory. They compare positively with previous approaches and their ROUGE scores are not far from those of the Wikipedia summaries.

The type of documents in hand, most of them extracted from tourist information websites, makes automatic summarization even more challenging than in other domains. In most of these documents, only a few information is relevant to the image, while the rest can be considered as noisy information (e.g. nearby hotels and other tourist services, advertisements from the website that hosts the information...). Besides, these documents are highly redundant.

However, the results reported show that there is room for improvement. In the future, we plan to overcome the limitations of our approaches that have been identified after the analysis of the results obtained. Incorporating some module able to identify the noisy information in the documents and filter it out would undoubtedly beneficial for both systems. In the case of the statistical-based approach, a query-focused summarization approach would be necessary to identify only sentences talking about the topic (i.e. the place).

Acknowledgments

This research is funded by the Spanish Government through the FPU program and the projects TIN2009-14659-C03-01, TSI 020312-2009-44, and TIN2009-13391-C04-01; by Conselleria d'Educació - Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/288); and the FPI program (BES-2007-16268) from the Spanish Ministry of Science and Innovation (project TEXT-MESS (TIN2006-15265-C06-01)). It is also funded by the TRIPOD project supported by the European Commission (Contract No. 045335).

References

1. Deschacht, K., Moens, M.: Text Analysis for Automatic Image Annotation. Proc. of the 45 th ACL (2007)
2. Mori, Y., Takahashi, H., Oka, R.: Automatic word assignment to images based on image division and vector quantization. In: Proc. of RIAO 2000: Content-Based Multimedia Information Access. (2000)
3. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: International Conference on Computer Vision. Volume 2. (2001) 408–415
4. Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In Seventh European Conference on Computer Vision (ECCV) **4** (2002) 97–112

5. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. The Journal of Machine Learning Research **3** (2003) 1107–1135
6. Pan, J., Yang, H., Duygulu, P., Faloutsos, C.: Automatic image captioning. In: IEEE International Conference on Multimedia and Expo. Volume 3. (2004)
7. Feng, Y., Lapata, M.: Automatic Image Annotation Using Auxiliary Text Information. Proc. of ACL 2008, Columbus, Ohio, USA (2008)
8. Satoh, S., Nakamura, Y., Kanade, T.: Name-It: naming and detecting faces in news videos. Multimedia, IEEE **6**(1) (1999) 22–35
9. Berg, T., Berg, A., Edwards, J., Forsyth, D.: Whos in the Picture? In: Proc. Of Advances in Neural Information Processing Systems Conference. (2005)
10. Marsh, E., White, M.: A taxonomy of relationships between images and text. Journal of Documentation **59** (2003) 647–672
11. Purves, R., Edwardes, A., Sanderson, M.: Describing the where–improving image annotation and search through geography. 1st Intl. Workshop on Metadata Mining for Image Understanding (2008)
12. Aker, A., Gaizauskas, R.: Summary Generation for Toponym-Referenced Images using Object Type Language Models. International Conference on Recent Advances in Natural Language Processing (RANLP) (2009)
13. Fan, X., Aker, A., Tomko, M., Smart, P., Sanderson, M., Gaizauskas, R.: Automatic Image Captioning From the Web For GPS Photographs. In: Proc. of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval. (2010)
14. Aker, A., Gaizauskas, R.: Model summaries for location-related images. In: Proc. of the 7th conference on International Language Resources and Evaluation. (2010)
15. Plaza, L., Díaz, A., Gervás, P.: Concept-graph based biomedical automatic summarization using ontologies. In: Coling 2008: Proc. of the 3rd Textgraphs workshop on Graph-based Algorithms for NLP. (2008) 53–56
16. Ferrández, O., Micol, D., Muñoz, R., Palomar, M.: A perspective-based approach for solving textual entailment recognition. In: Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. (2007) 66–71
17. Patwardhan, S., Banerjee, S., Pedersen, T.: Senserelate::targetword: a generalized framework for word sense disambiguation. In: Proc. of the ACL 2005 on Interactive poster and demonstration sessions, Morristown, NJ, USA (2005) 73–76
18. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity - measuring the relatedness of concepts. In: Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04). (2004) 1024–1025
19. Erkan, G., Radev, D.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research **22** (2004) 457–479
20. Yoo, I., Hu, X., Song, I.Y.: A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinformatics **8**(9) (2007)
21. Lloret, E., Palomar, M.: A gradual combination of features for building automatic summarisation systems. In: Proc. of the 12th International Conference on Text, Speech and Dialogue. (2009) 16–23
22. Givón, T. In: Syntax: A functional-typological introduction, II. John Benjamins (1990)
23. Dang, H.: Overview of DUC 2005. DUC 05 Workshop at HLT/EMNLP (2005)
24. Lin, C.: ROUGE: A Package for Automatic Evaluation of Summaries. Proc. of the Workshop on Text Summarization Branches Out (WAS 2004) (2004) 25–26