Generating Image Captions using Topic Focused Multi-document Summarization

Robert Gaizauskas
Natural Language Processing Group
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
R.Gaizauskas@sheffield.ac.uk

In the near future digital cameras will come standardly equipped with GPS and compass and will automatically add global position and direction information to the metadata of every picture taken. Can we use this information, together with information from geographical information systems and the Web more generally, to caption images automatically?

This challenge is being pursued in the TRIPOD project (http://tripod.shef.ac.uk/) and in this talk I will address one of the subchallenges this topic raises: given a set of toponyms automatically generated from geo-data associated with an image, can we use these toponyms to retrieve documents from the Web and to generate an appropriate caption for the image?

We begin assuming the toponyms name the principal objects or scene contents in the image. Using web resources (e.g. Wikipedia) we attempt to determine the types of these things – is this a picture of church? a mountain? a city? We have constructed a taxonomy of such image content types using on-line collections of captioned images and for each type in the taxonomy we have constructed several collections of texts describing that type. For example, we have a collection of captions describing churches and a collection of Wiki pages describing churches. The intuition here is that these collections are examples of, e.g. the sorts of things people say in captions or in descriptions of churches. These collections can then be used to derive models of objects or scene types which in turn can be used to bias or focus multi-document summaries of new images of things of the same type.

In the talk I report results of work we have carried out to explore the hypothesis underlying this approach, namely that brief multi-document summaries generated as image captions by using models of object/scene types to bias or focus content selection will be superior to generic multi-document summaries generated for this purpose. I describe how we have constructed an image content taxonomy, how we have derived text collections for object/scene types, how we have derived object/scene type models from these collections and how these have been used in multi-document summarization. I also discuss the issue of how to evaluate the resulting captions and present preliminary results from one sort of evaluation.

Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, page 1
Manchester, August 2008