Model Summaries for Location-related Images

Ahmet Aker, Robert Gaizauskas

Department of Computer Science, University of Sheffield
Sheffield, UK, S1 4DP
Email: a.aker@dcs.shef.ac.uk, r.gaizauskas@dcs.shef.ac.uk

Abstract

At present there is no publicly available data set to evaluate the performance of different summarization systems on the task of generating location-related extended image captions. In this paper we describe a corpus of human generated model captions in English and German. We have collected 932 model summaries in English from existing image descriptions and machine translated these summaries into German. We also performed post-editing on the translated German summaries to ensure high quality. Both English and German summaries are evaluated using a readability assessment as in DUC and TAC to assess their quality. Our model summaries performed similar to the ones reported in Dang (2005) and thus are suitable for evaluating automatic summarization systems on the task of generating image descriptions for location related images. In addition, we also investigated whether post-editing of machine-translated model summaries is necessary for automated ROUGE evaluations. We found a high correlation in ROUGE scores between post-edited and non-post-edited model summaries which indicates that the expensive process of post-editing is not necessary.

1. Introduction

In recent years the number of images on the Web has grown immensely, facilitated by the development of affordable digital hardware and the availability of online image sharing social sites. To support indexing and retrieval of these images different approaches to automatic image captioning have been proposed (Deschacht and Moens, 2007; Mori et al., 2000; Barnard and Forsyth, 2001; Duygulu et al., 2002; Barnard et al., 2003; Pan et al., 2004; Feng and Lapata, 2008; Satoh et al., 1999; Berg et al., 2005). We have experimented with multi-document summarization techniques to automatically generate extended image captions from web-documents related to images (Aker and Gaizauskas, 2008; Aker and Gaizauskas, 2009). Unlike other works, we have focussed our attention on images of static features of the built or natural landscape (e.g. buildings, mountains, etc.) that do not contain objects which move about in such landscapes (e.g. people, cars, etc.).

Multi-document summarization is a well established research area. It has been also a focus of the Document Understanding Conference (Dang, 2005; Dang, 2006) (DUC) and the Text Analysis Conference (TAC)[1] where participating systems are evaluated on a summarization task within a specific domain. The evaluation is performed both manually and automatically. For manual evaluation the output summaries are rated by humans along various dimensions using a five point scale. Automatic evaluation is performed using ROUGE metrics (Lin, 2004) that express the degree of n-gram overlap between automatic and human generated summaries (model summaries). Model summaries have been provided by DUC and TAC for diverse domains (e.g. news, biography, etc.). However, there is no publicly available data set to evaluate the performance of different summarization systems on the task of generating location-related extended image captions. In this paper we describe a corpus of human generated model captions in English and German. The corpus is free for download[2].

The data we provide is a set of model captions/summaries for location related images along with their images. The model summaries have been constructed using descriptions of locations taken from *VirtualTourist*[3]. VirtualTourist is a social web site where visitors upload their pictures and descriptions of places they have visited. Using VirtualTourist descriptions or captions we collected a corpus of 932 model summaries for 307 different locations along with their images. The model summaries are written in English and have been manually assembled by eleven humans. We translated these summaries into German using the Langrid machine translator (Ishida, 2006) and post-edited them. The original model summaries as well as the translations and their post-edited versions are evaluated based on readability criteria similar to those used in DUC and TAC. The results of our evaluation are comparable with those reported

[1] http://www.nist.gov/tac/

[2] See www.dcs.shef.ac.uk/ ahmet for details.

[3] www.virtualTourist.com

by DUC for the readability of model summaries.

In this paper we first describe the collection of model summaries. In section 3 we discuss the multi-lingual caption generation. Finally we report the results of manual readability evaluation of model summaries for both languages and conclude the paper in section 5.

2. Model Summary Collection

One approach to gathering model captions/summaries is to begin by retrieving documents from the web using as a query the name of the image subject (e.g. "Westminster Abbey"). These are the same documents from which an automated caption generator will generate its summary. The model caption/summary is then generated by humans who read the documents and write a model summary. This is the approach to model summary creation followed in DUC and TAC. In a variant of this approach, humans gather "information nuggets" – atomic facts – from the retrieved documents and use these to assess automatic summaries by checking the extent to which the automatically generated summary contains the nuggets (Voorhees, 2003).

In both variants of this approach the generation of model summaries requires that humans read all the documents to be summarized and select the content to go into a summary. However, going through all the documents and reading them is a labor-intensive task (in time and money).

To reduce the burden of model summary generation we have used *VirtualTourist* as a resource. VirtualTourist is one of the largest online travel communities in the world, where over six million travelers around the world share information in form of image descriptions or captions. Collecting model summaries from these existing descriptions has the following advantages:

Descriptions for locations are "natural" model summaries in that they are written to concisely convey essential information about the subject of the image. Furthermore since they are captions spontaneously written and associated by humans with images there is an argument for preferring them as model captions to summaries artificially created from documents mentioning the location.

Descriptions are shorter (average length is 87 words) than documents which need to be summarized if following a DUC-like approach. This reduces the time required for reading to the time which needs to be spent reading only a few short image descriptions.

Descriptions are more focused on the location The descriptions are focused on the location in the image. If the description, for instance, is about a church, then it usually contains information about when the church was built, the name of the architect or designer, where the church is located, how to reach the church by public transport, etc.. This contrasts with documents retrieved using the location name as a query, which may have a different focus and either not contain the relevant information or contain it in a non-obvious place in the document, i.e. the content selection has already been done in the image descriptions, again reducing the time and effort needed to create model summaries.

Given the advantages of reduced time/effort in creating a model summary resource we decided to use VirtualTourist image descriptions as model summaries. Note that this choice distinguishes the summary/caption evaluation from other summary evaluations in that the reference or model summaries are not derived from the documents from which the automated summaries are themselves generated. A likely consequence of this is that the automated summary scores will be lower than when the reference and peer summaries are generated from the same source; however, given the redundancy of information on the web, this effect should not be too high for well-known locations.

2.1. Collection Procedure

VirtualTourist organizes images by location. We selected popular cites including *London, Edinburgh, New York, Venice, Florence, Rome*, assigned different sets of cities to eleven different human subjects and asked them to collect up to four model summaries for each place from existing captions with lengths ranging from 190 to 210 words (Figure 1).

During the collection it was ensured that the summaries did not contain personal information and that they did genuinely describe a place, e.g. *Westminster Abbey*. If the captions did contain personal information this was removed. Where a caption was deemed too short, i.e. the number of words was less than 190, more than one caption was used to build a model summary. We also ensured that summaries did not contain redundant information. If the caption contained more than 210 words the less important information was deleted – this was a subjective decision made by the person collecting the captions.

We collected model summaries for 307 different places from various cities around the world. The number of places with four model summaries is 170, with

Table 1: Model summary about the Eiffel Tower in English and its post-edited German machine translation.

The Eiffel Tower is the most famous place in Paris. It is made of 15,000 pieces fitted together by 2,500,000 rivets. It's of 324 m (1070 ft) high structure and weighs about 7,000 tones. This world famous landmark was built in 1889 and was named after its designer, engineer Gustave Alexandre Eiffel. It is now one of the world's biggest tourist places which is visited by around 6,5 million people yearly. There are three levels to visit: Stages 1 and 2 which can be reached by either taking the steps (680 stairs) or the lift, which also has a restaurant "Altitude 95" and a Souvenir shop on the first floor. The second floor also has a restaurant "Jules Verne". Stage 3, which is at the top of the tower can only be reached by using the lift. But there were times in the history when Tour Eiffel was not at all popular, when the Parisians thought it looked ugly and wanted to pull it down. The Eiffel Tower can be reached by using the Mtro through Trocadro, Ecole Militaire, or Bir-Hakeim stops. The address is: Champ de Mars-Tour Eiffel.	Der Eiffelturm ist der bekannteste Platz in Paris. Er ist aus 15.000 zusammengesetzten Stcken mit 2.500.000 Nieten gemacht. Er ist ein 324 m (1070 ft) hoches Bauwerk und wiegt rund 7.000 Tonnen. Dieses weltberhmte Wahrzeichen wurde im Jahr 1889 gebaut und nach seinem Konstrukteur, Ingenieur Gustave Alexandre Eiffel benannt. Er ist heute einer der weltweit grten touristischen Orte, der jhrlich von rund 6,5 Millionen Menschen besucht wird. Es gibt drei Ebenen, die man besuchen kann: Die Stufen 1 und 2, die entweder ber die die Treppen (680 Stufen) oder den Aufzug erreicht werden knnen und die auch ein Restaurant "Altitude 95" und ein Souvenir-Geschft im ersten Stock haben. Der zweite Stock hat auch ein Restaurant "Jules Verne". Stufe 3, die an der Spitze des Turms ist, kann nur mit dem Lift erreicht werden. Aber es gab Zeiten in der Geschichte, als der Eiffelturm berhaupt nicht bei allen beliebt war und als die Pariser dachten, er sei hsslich und ihn abreien wollten. Der Eiffelturm kann erreicht werden, indem Sie die Mtro durch Trocadro, Ecole Militaire oder Bir-Hakeim Haltestellen benutzten. Die Adresse ist: Champ de Mars-Tour Eiffel.

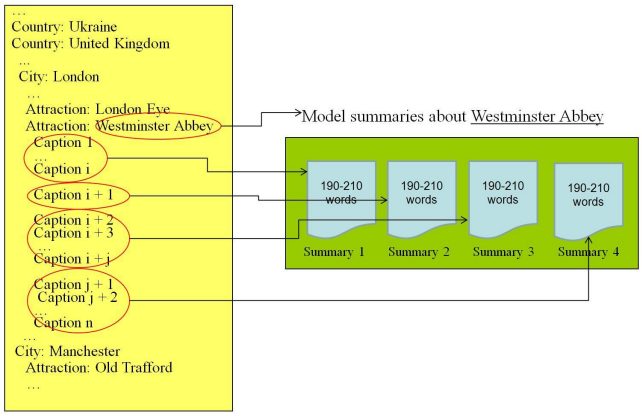

Figure 1: Model Summary Collection.

three 41, with two 33 and 63 had only one model summary. This leads to 932 model summaries in total. On average 7.5 minutes was spent to generate one single model summary with 116 hours being spent to gather the entire collection. An example model summary about the *Eiffel Tower* is shown in Table 1.

We also manually categorized each of the 307 places by scene type (Aker and Gaizauskas, 2009; Gornostay and Aker, 2009). Table 2 shows 60 different scene types (ranging from rural types such as mountains, valley, etc. to urban types such as churches, museums, etc.) covered by our 307 places. Table also gives details about the number of places and number of model summaries in each scene type.

3. Multi-Lingual Model Summaries

We used the Langrid (Ishida, 2006) system to translate the English model summaries into German[4]. Langrid is a lightweight infrastructure which enables users to (1) add language resources such as dictionaries, thesauri and corpora, morphological analyzers, translation and paraphrasing systems and (2) combine existing language resources to compose services to perform machine translation tasks.

After machine translation the German summaries were post-edited. We randomly selected 200 machine translated model summaries and asked two German native speakers to post-edit them. The two speakers were given the English model summary as well as the corresponding German machine translated one and were asked to correct all mistranslations and grammatical errors. Each participant corrected 100 model summaries. The participants spent on average 12 minutes to edit a single summary.

4. Evaluation

Our model summaries were evaluated by a readability assessment in the same way as in DUC and TAC. DUC and TAC use a manual assessment scheme to measure the quality of the automatically generated summaries

[4]We used machine translated summaries because there exists no web resource in German that is comparable in size to VirtualTourist

Table 2: Scene types, number of different places (plcs) and the number of model summaries (sums) within a scene type. These scene types are covered by our 307 places.

Scene Type	plcs	sums	Scene Type	plcs	sums
mountain	7	18	cemetery	1	4
street	6	13	college	3	5
beach	7	18	house	5	13
cave	1	1	village	5	8
zoo	4	10	abbey	1	4
hill	5	16	church	11	32
lake	3	6	museum	17	55
pub	2	2	basilica	2	8
gate	1	4	glacier	1	1
temple	8	29	parliament	3	12
statue	2	8	market	2	8
railway	2	3	ski resort	1	1
avenue	2	7	stadium	2	5
theatre	2	8	aquarium	2	5
cathedral	11	35	bridge	9	31
opera house	4	16	palace	14	52
railway station	1	4	mosque	4	13
waterfall	3	4	road	1	1
valley	1	1	island	7	14
area	5	15	volcano	2	4
skyscraper	2	5	monument	10	31
district	3	11	boulevard	1	2
university	6	14	building	9	23
park	14	45	gallery	2	7
venue	1	1	canal	1	6
observation wheel	1	4	tower	8	31
prison	2	7	residence	2	6
castle	14	51	square	18	63
hotel	4	7	garden	4	14
river	8	26	chapel	1	4

as well as the model summaries. In these exercises the human subjects are presented with the summaries and asked to assess each summary based on the following criteria (each criterion has a five point scale with high scores indicating a better result in relation to that criterion) (Dang, 2005; Dang, 2006):

- **Grammaticality**: The caption does not have formatting or capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
- **Redundancy**: There is no unnecessary repetition in the caption. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Westminster Abbey") when a pronoun ("it") would suffice.
- **Clarity**: It is easy to identify who or what the pronouns and noun phrases in the caption are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.
- **Focus**: The caption has a focus; sentences should only contain information that is related to the rest of the caption.
- **Structure**: The caption is well-structured and well-organized. The caption should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Table 3: Readability five point scale evaluation results for 489 English model summaries.

Feature	5	4	3	2	1
grammaticality	79.8%	15.2%	2.8%	1.8%	0.4%
redundancy	94.9%	3.1%	1.8%	0.2%	0%
clarity	93.9%	4.3%	0.81%	0.4%	0.6%
focus	90.6%	7%	2%	0.4%	0%
structure	90.8%	5.3%	2.9%	0.81%	0.2%

4.1. Evaluation of English model summaries

To assess all English model summaries is labour-intensive work. Thus, we randomly selected half (in total 489 out of 932) of the model summaries and distributed them among three different humans who were not the summary collector. The humans were assigned different summary sets and were asked to assess the summaries according to the criteria described above. The results in terms of percentage of summaries with same scores at each level are shown in Table 3.

Most of the model summaries (94% or more) obtained scores at level 4 or above. These results show that our summaries are high quality model summaries and are appropriate for automatic evaluation of systems producing image captions. The results are also comparable with those reported at DUC concerning the readability assessment of model summaries. Dang (2005; 2006) reported 95% of model summaries received scores at level 4 or above in each of the criteria.

Table 4: Readability five point scale evaluation results for 100 machine translated German model summaries with (PostEdit) and without (NonPostEdit) post-editing.

Feature	5	4	3	2	1
grammaticalityNonPostEdit	0%	3%	15%	56%	26%
redundancyNonPostEdit	84%	13%	3%	0%	0%
clarityNonPostEdit	70%	15%	15%	0%	0%
focusNonPostEdit	80%	18%	2%	0%	0%
structureNonPostEdit	2%	81%	17%	0%	0%
grammaticalityPostEdit	78%	16%	6%	0%	0%
redundancyPostEdit	96%	2%	2%	0%	0%
clarityPostEdit	94%	2%	2%	2%	0%
focusPostEdit	86%	10%	4%	0%	0%
structurePostEdit	90%	4%	6%	0%	0%

4.2. Evaluation of German model summaries

For the evaluation of machine translated model summaries we randomly selected 100 summaries from the 200 model summaries set we used for post-editing (see Section 3). We first evaluated the 100 model summaries without having post-edited and later their post-edited versions. We asked two German native speakers who were not the post-editors to do the assessment. The first person evaluated the summaries without post-editing and the second person evaluated their post-edited versions. The results of this evaluation are shown in Table 4.

Apart from the grammaticality measure the non post-edited machine translated summaries perform almost as well as their post-edited versions. The rather unsatisfactory results in the grammaticality measure are mainly due to the use of slang in the English model summaries which gets incorrectly translated. Furthermore, some English sentences contain several main clauses separated with "and". Such constructions are not grammatical in German and have to be transformed into sub clauses in German, which is not done by the machine translation system. In addition, articles are sometimes incorretly translated or verbs are used in the wrong tense or discarded completely. Finally, in some translated sentences the word order is wrong. However, the results for the post-edited versions show that these errors in machine translated summaries can be corrected with little effort and corrected summaries can be used as model summaries. This is also supported by significantly high correlation results obtained after comparing different automated summaries to German machine translated summaries and their post-edited versions using ROUGE.

We generated two different types of German summaries for 28 different images: baseline and automated summaries. The baseline summaries are generated from the top-ranked documents retrieved in the Yahoo! search results for each image's place name. From these documents we created a summary by selecting sentences from the beginning until the summary reaches a length of 200 words. The automated summaries are generated using the summarization system described in (Aker and Gaizauskas, 2009). We compared both baseline and automated summaries against the machine translated model summaries as well as against their post-edited versions. Following the Document Understanding Conference (DUC) (Dang, 2005; Dang, 2006) evaluation standards we used for the comparison ROUGE 2 and ROUGE SU4 as evaluation metrics. ROUGE 2 gives recall scores for bi-gram overlap between the automatically generated summaries and the model ones. ROUGE SU4 allows bi-grams to be composed of non-contiguous words, with a maximum of four words between the bi-grams.

After obtaining the ROUGE scores we computed two-tailed Pearson correlation coefficients between them. We observed that the comparison between the baseline and machine translated summaries and baseline and post-edited machine translated summaries correlate with 94% in ROUGE 2 ($p<.01$) and 95% in ROUGE SU4 ($p<.01$). For the automated summaries (comparison between the automated and machine translated summaries and automated and post-edited machine translated summaries) we obtained a similar observation: 89% correlation in ROUGE 2 ($p<.01$) and 93% correlation in ROUGE SU4 ($p<.01$).

5. Conclusion

In this paper we described a set of 932 model summaries for 307 different locations. The summaries are collected from existing image captions and are written in English. We used machine translation to translate the original model summaries into German. The original as well as the machine translated summaries are assessed by humans based on readability criteria similar to those used by DUC and TAC. The evaluations show that both English and German machine translated summaries are at high quality. In future we plan to enrich the data set with Italian and Latvian summaries. The model summaries are free for download.

6. Acknowledgment

The research reported was funded by the TRIPOD project supported by the European Commission under the contract No. 045335. We would like to thank Virtualtourist for allowing us to use the captions and

the images. We also would like to thank Emina Kurtic, Edina Kurtic, Lejla Kurtic, Olga Nesic, Mesude Bicak, Asmaa Elhannani, Murad Abouammoh, Azza Al-Maskari, Mariam Kiran, Rafet Ongun and Dilan Paranavithana who took part in the experiments.

7. References

A. Aker and R. Gaizauskas. 2008. Evaluating automatically generated user-focused multi-document summaries for geo-referenced images. *Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008) Workshop on Multi-source, Multilingual Information Extraction and Summarization (MMIES2), 2008.*

A. Aker and R. Gaizauskas. 2009. Summary Generation for Toponym-Referenced Images using Object Type Language Models. *International Conference on Recent Advances in Natural Language Processing (RANLP) September 14-16, 2009, Borovets, Bulgaria*.

K. Barnard and D. Forsyth. 2001. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, volume 2, pages 408–415. Vancouver: IEEE.

K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei, and M.I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

T.L. Berg, A.C. Berg, J. Edwards, and DA Forsyth. 2005. Whos in the Picture? In *Advances in Neural Information Processing Systems 17: Proc. Of The 2004 Conference*. MIT Press.

H.T. Dang. 2005. Overview of DUC 2005. *DUC 05 Workshop at HLT/EMNLP*.

H.T. Dang. 2006. Overview of DUC 2006. *National Institute of Standards and Technology*.

K. Deschacht and M.F. Moens. 2007. Text Analysis for Automatic Image Annotation. *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics. East Stroudsburg: ACL.*

P. Duygulu, K. Barnard, JFG de Freitas, and D.A. Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *In Seventh European Conference on Computer Vision (ECCV)*, 4:97–112.

Y. Feng and M. Lapata. 2008. Automatic Image Annotation Using Auxiliary Text Information. *Proc. of Association for Computational Linguistics (ACL) 2008, Columbus, Ohio, USA*.

T. Gornostay and A. Aker. 2009. Development and Implementation of Multilingual Object Type Toponym-Referenced Text Corpora for Optimizing Automatic Image Description. *Proc. of the 15th Annual International Conference on Computational Linguistics and Intellectual Technologies Dialogue 2009, Bekasovo, Russia.*

T. Ishida. 2006. Language grid: An infrastructure for intercultural collaboration. In *Applications and the Internet, 2006. SAINT 2006*, page 5.

C.Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proc. of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

Y. Mori, H. Takahashi, and R. Oka. 2000. Automatic word assignment to images based on image division and vector quantization. In *Proc. of RIAO 2000: Content-Based Multimedia Information Access*.

J.Y. Pan, H.J. Yang, P. Duygulu, and C. Faloutsos. 2004. Automatic image captioning. In *Multimedia and Expo, 2004. ICME'04. IEEE International Conference on*, volume 3.

S. Satoh, Y. Nakamura, and T. Kanade. 1999. Name-It: naming and detecting faces in news videos. *Multimedia, IEEE*, 6(1):22–35.

E.M. Voorhees. 2003. Overview of the TREC 2003 Question Answering Track. *Proc. of the Twelfth Text REtrieval Conference (TREC 2003)*, 142.