Developing a Dataset for Evaluating Approaches for Document Expansion with Images

Debasis Ganguly, Iacer Calixto, Gareth J.F. Jones
ADAPT Centre, School of Computing, Dublin City University, Ireland
{icalixto, dganguly, gjones}@computing.dcu.ie

Abstract

Motivated by the adage that a "picture is worth a thousand words" it can be reasoned that automatically enriching the textual content of a document with relevant images can increase the readability of a document. Moreover, features extracted from the additional image data inserted into the textual content of a document may, in principle, be also be used by a retrieval engine to better match the topic of a document with that of a given query. In this paper, we describe our approach of building a ground truth dataset to enable further research into automatic addition of relevant images to text documents. The dataset is comprised of the official ImageCLEF 2010 collection (a collection of images with textual metadata) to serve as the images available for automatic enrichment of text, a set of 25 benchmark documents that are to be enriched, which in this case are children's short stories, and a set of manually judged relevant images for each query story obtained by the standard procedure of depth pooling. We use this benchmark dataset to evaluate the effectiveness of standard information retrieval methods as simple baselines for this task. The results indicate that using the whole story as a weighted query, where the weight of each query term is its tf-idf value, achieves an precision of 0.1714 within the top 5 retrieved images on an average.

Keywords: Image Retrieval, Document Augmentation with Images

1. Introduction

Document expansion, in addition to inserting text and hyperlinks, can also involve adding non textual content such as images that are topically related to document text, in order to enhance the readability (or in fact indexability) of the text. For example, in (Hall et al., 2012), Wikipedia articles are augmented with images retrieved from the *Kirklees* image archive, where automatically extracted key concepts from the Wiki text passages were used to formulate the queries for retrieving the images.

The aim of our work, reported in this paper, is to build up a dataset for evaluation of the effectiveness of automated approaches to document expansion with images. In particular, the problem that we address in this paper is that of augmenting the text of children's short stories (e.g. fairy tales and fables) with images in order to help improve the readability of the stories for small children according to the adage that "a picture is worth a thousand words"[1]. The "document expansion with images" methodologies, developed and evaluated on this dataset, can also be applied to augment other types of text documents, such as news articles, blogs etc.

The illustration of children's stories is a particular instance of the general problem of automatic text illustration, an inherently multimodal problem that involves image processing and natural language processing. A related problem to automatic text illustration is that of automatic textual generation of image description. This problem is in fact under active research and has drawn significant research interests in recent years (Feng and Lapata, 2010; Vinyals et al., 2014; Karpathy and Fei-Fei, 2014; Xu et al., 2015).

2. Outline of our work

In order to share a dataset for text augmentation with images among researchers, and to encourage them to use this dataset for research purposes, we organized a shared task, named "Automated Story Illustration"[2], as a part of the Forum of Information Retrieval and Evaluation (FIRE) 2015 workshop[3]. The goal of this task was to automatically illustrate children's short stories by retrieving a set of images that can be considered relevant to illustrate the concepts (agents, events and actions) of a given story. The data resource comprised of the text of children's short stories and a set of relevant images associated to each story. The data resource described in this paper is the outcome of this shared task.

In contrast to the standard keyword-based ad-hoc search for images (Caputo et al., 2014), there are no currently available explicitly user formulated keyword-based queries for the task of automated story illustration. Instead, each text passage acts as an implicit query for which images need to retrieved to augment it. To illustrate the task output with an example, let us consider the story "The Ant and the Grasshopper" shown in Figure 1. In the text we underline the key concepts that are likely to be used to formulate queries for illustrating this story. Additionally, we show a set of manually collected images from the results of Google image search[4] executed with each of these underlined phrases as queries. It can be seen that the story with these sample images is likely to be more appealing to a child rather than the plain raw text. This is because, with the accompanying images, children can potentially relate to the concepts described in the text, e.g. the top left image shows a child how does a "summer day's field" look like.

3. Dataset Description

It is worth mentioning that we use Google image search in our example of Figure 1 for illustrative purposes only. However, in order to achieve a fair comparison between au-

[1] http://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words

[2] http://srv-cngl.computing.dcu.ie/StoryIllustrationFireTask/

[3] http://fire.irsi.res.in/fire/

[4] https://images.google.com/

IN a field one summer's day a Grasshopper was hopping about, chirping and singing to its heart's content. An Ant passed by, bearing along with great toil an ear of corn he was taking to the nest. "Why not come and chat with me, said the Grasshopper, "instead of toiling and moiling in that way?" "I am helping to lay up food for the winter," said the Ant, "and recommend you to do the same." "Why bother about winter?" said the Grasshopper; "we have got plenty of food at present." But the Ant went on its way and continued its toil. When the winter came the Grasshopper had no food, and found itself dying of hunger, while it saw the ants distributing every day corn and grain from the stores they had collected in the summer. Then the Grasshopper knew: "IT IS BEST TO PREPARE FOR THE DAYS OF NECESSITY."

Figure 1: The story of "The Ant and the Grasshopper" with a sample annotation of images from the web. Images were manually retrieved using Google image search. The key terms used as queries in the Google image search are underlined in the text.

tomated approaches to the story illustration task, it is imperative to construct a dataset comprised of a static document collection, a set of test queries (text from stories), and the relevance assessments for each story.

The static image collection that we use for this image search task is the ImageCLEF 2010 Wikipedia image collection released (Popescu et al., 2010). For the queries, we used popular children's fairy tales since most of them are available in the public domain and are freely distributable. In particular, we make use of 22 short stories collected from "Aesop's Fables"[5].

3.1. Manual Annotation of the Short Story Text

The first research challenge for an automated story illustration approach is to extract the key concepts from the text passages in order to formulate suitable queries for retrieving relevant images, e.g. an automated approach should extract "summer day field" as a meaningful unit for illustration. The second research challenge is to make use of these extracted concepts or phrases to construct queries and perform retrieval from the collection of images, which in this case is the ImageCLEF collection.

In order to facilitate participants to concentrate on retrieval only, we manually annotated the short stories with concepts that are likely to require illustration. Participants volunteering for the annotation task, were instructed to highlight parts of the stories that they feel would better be understood by children with the help of illustrative images. In total, five participants annotated 22 stories, three participants annotating 4 each and the remainder two annotating 5 each.

For other participants who wanted to automatically extract the concepts from a story for the purpose of illustration, we encouraged them to develop automated approaches and then compare their results with the manually annotated ones. Some possible alternatives for a participating system could be the use of a shallow natural language processing (NLP) technique, such as named entity recognition and chunking, to first identify individual query concepts and then retrieve candidate images for each of these. Another approach could be using the entire text as a query and to cluster the result-list of documents to identify the individual query components.

3.2. Pooled Relevance Assessments

An important component in an information retrieval (IR) dataset is the set of relevance assessments for a query. To obtain the set of relevant images for each story, we undertake a IR procedure called *pooling*, where a pool of documents, i.e. the set of top ranked documents from retrieval systems with different settings, is assessed manually for relevance (Voorhees and Harman, 1999). The relevance judgements for our dataset are obtained as follows.

Firstly, in order to be able to search for images with ad-hoc keywords, we indexed the ImageCLEF collection. The extracted text from the caption of each image in the ImageCLEF collection was indexed as a retrievable document. The ImageCLEF collection was indexed with Lucene[6], an open source IR system in Java.

Secondly, we made use of each manually annotated concept as an individual query that was executed on the document collection of the ImageCLEF. For example, for the story shown in Figure 1, the queries that were executed on the index of the image metadata were "field one summer's day", "grasshopper" and so on. Each query was executed with three different retrieval model settings, namely BM25, LM and tf-idf with default parameter settings in Lucene.

To construct the pool of candidate relevant images for each story, we fused the ranked lists obtained with each individual query, with three different retrieval settings for each query using the standard COMBSUM merging technique (Shaw et al., 1994). Finally, the top 20 documents from this fused ranked list were assessed for relevance. The relevance assessment for each manually annotated concept for each story was conducted by the same participant who created the annotation in the first place. This ensured that the participants had a clear understanding of the relevance criteria. Participants were asked to assign relevance on a five point scale ranging from absolutely non-relevant to highly relevant.

[5] https://en.wikipedia.org/wiki/Aesop

[6] https://lucene.apache.org/

Approach	MAP	P@5	P@10
Unweighted qry terms	0.0275	0.1048	0.0905
tf-idf weighted qry terms	**0.0529**	**0.1714**	**0.1238**

Table 1: Retrieval effectiveness of simple baseline approaches averaged over 22 stories.

4. Experimental Results

In this section, we describe initial experiments that we conducted on our dataset, which are intended to act as baselines for future work on this dataset. As our first baseline, we simply use all the words in a story to create a query. We then use this query to retrieve a list of images from the indexed collection of image captions. The retrieval model that we use is language modeling (LM) with Jelinek Mercer smoothing (Ponte and Croft, 1998). The purpose of this baseline is to see whether it is effective to use the whole text from a story to formulate a query.

Our second baseline also uses the entire text of a story as a query. However, instead of using unweighted query terms as in baseline 1, in this approach we assign weights to each query term. The weight that we use for each query term is the tf-idf value of that term in the collection. The objective in this experimental setting is to see whether the tf-idf weights of the terms can reasonably approximate the importance of each term in the stories and thus in turn may help in effective query formulation. For example, if the terms corresponding to the important concepts in a story get assigned higher tf-idf values, a matching of these terms in the image captions will help retrieve captions with these words to be retrieved in the top ranks.

Although it is possible to prune query terms whose weights (tf-idf values) fall below a certain threshold, to keep our baseline simple, we did not prune any query terms. It is worth mentioning here that the two baselines that we use are quite simple because our intention is to see how simple methods perform, before attempting to apply more involved approaches for this task.

The evaluation measures that we use for measuring the effectiveness of retrieving relevant images for enriching text documents, are the mean average precision (MAP) and precision at top cut-off ranks of 5 and 10, denoted by $P@5$ and $P@10$ respectively, in Table 1. This retrieval task is more of a precision oriented task than recall oriented one, as a result of which it is more important to retrieve relevant images corresponding to the topics (entities or events) in a story at the top ranks than to achieve high recall values by retrieving all relevant images corresponding to the topics of a story from the collection. This is the reason why we use MAP and $P@k$ as the evaluation measure for this task, since both are precision oriented metrics.

In Table 1, we observe that simply using all terms of a story as a query to retrieve a ranked list of images does not produce satisfactory results, as can be seen from the low MAP and $P@k$ values. In contrast, even a very simple approach of weighting the terms in the text of the story by their tf-idf weights can produce a significant improvement in the results. Motivated by this result, we believe that shallow NLP techniques to extract useful concepts can further improve the results.

5. Conclusions and Future work

In this paper, we described the construction of a dataset for the purpose of evaluating automated approaches for document augmentation with images. In particular, we addressed the problem of automatically illustrating children stories. Our constructed dataset comprises of 22 children stories as the set of queries and uses the ImageCLEF document collection as the set of retrievable images. The dataset also comprises manually annotated concepts within each story that can potentially be used as queries to retrieve a collection of relevant images for each story. In fact, the retrieval results obtained with the manual annotations can act as strong baselines to compare against approaches that automatically extract the concepts from a story. For evaluation purposes, the dataset contains the relevance assessments, i.e. a set of images relevant to the core topics (entities or events) of each story. Our initial experiments suggest that the dataset can be used to compare and evaluate various approaches to automated augmentation of documents with images. We demonstrate that a tf-idf based term weighting for the query terms can prove useful in improving retrieval effectiveness, thus leaving open the future directions of research for effective query representation for this task.

Acknowledgements

This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (Grant No: 13/RC/2106).

6. Bibliographical References

Caputo, B., Müller, H., Martínez-Gómez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N. B., Üsküdarli, S., Paredes, R., Cazorla, M., García-Varea, I., and Morell, V. (2014). ImageCLEF 2014: Overview and Analysis of the Results. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, pages 192–211.

Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hall, M. M., Clough, P. D., de Lacalle, O. L., Soroa, A., and Agirre, E. (2012). Enabling the Discovery of Digital Cultural Heritage Objects Through Wikipedia. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '12, pages 94–100, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karpathy, A. and Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281. ACM.

Popescu, A., Tsikrika, T., and Kludas, J. (2010). Overview of the wikipedia retrieval task at imageclef 2010. In Martin Braschler, et al., editors, *CLEF (Notebook Papers/LABs/Workshops)*.

Shaw, J. A., Fox, E. A., Shaw, J. A., and Fox, E. A. (1994). Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Voorhees, E. M. and Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.