Overview of the Automated Story Illustration Task at FIRE 2015

Debasis Ganguly
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
dganguly@computing.dcu.ie

Iacer Calixto
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
icalixto@computing.dcu.ie

Gareth Jones
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
gjones@computing.dcu.ie

ABSTRACT

In this paper, we describe an overview of the shared task (track) carried out as part of the Forum of Information Retrieval and Evaluation (FIRE) 2015 workshop. The objective in this task is to illustrate a passage of text automatically by retrieving a set of images and then inserting them at appropriate places in the text. In particular, for this track, the text to be illustrated is a set of short stories (fables) for children. Some of the research challenges for the participants in developing an automated story illustrating system involve developing techniques to automatically extract out the concepts to be illustrated from a full story text, explore how to use these extracted concepts for query representation in order to retrieve a ranked list of images per query, and finally investigating how merge the ranked lists obtained from each individual concept to present a single ranked list of candidate relevant images per story. In addition to reporting an overview of the approaches undertaken by two participating groups who submitted runs for this task, we also report two of our own baseline approaches for tackling the problem of automated story illustration.

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 of the text. For example, in [3], 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. This automatic augmentation of documents can be useful for various purposes, such as enhancing the readability of text for children enabling them to learn and engage with the content more, for making it easier for medical students to learn more about a disease or its syndromes by looking at related images, etc.

The aim of our work, reported in this paper, is to build up a dataset for evaluating the effectiveness of automated approaches for document expansion with images. In particular, the problem that we address in the 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 [2, 7, 4, 8].

The rest of the paper is organized as follows. In Section 2, we present a brief overview of the task objectives. In Section 3, we describe how the dataset (queries and relevance judgments) is constructed. Section 4 describes our own initial experiments so as to obtain our own baselines on the dataset constructed. Section 5 provides a brief overview of the approaches undertaken by the participating groups and presents the official results. Finally Section 6 concludes the paper with directions for future work.

2. TASK DESCRIPTION

In order to share among researchers a dataset for text augmentation with images, and to encourage them to use this dataset for research purposes, we are organizing 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 is 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.

In contrast to the standard keyword-based ad-hoc search for images [1], there exists no explicitly user formulated keyword based queries in this task. 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 the story. Additionally, we show a set of manually collected

[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/`

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."

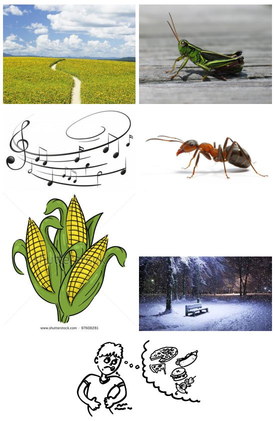

Figure 1: The story of "The Ant and the Grasshopper" with a sample annotation of images from the web. Images were manually retrieved with Google image search. The key terms used as queries in Google image search are underlined in the text.

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, the 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 purpose only. However, in order to achieve a fair comparison between automated approaches to the story illustration task, it is imperative to build up 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 task is the ImageCLEF 2010 Wikipedia image collection released [6]. For the queries, we used popular children's fairy tales since most of them are available in the public domain and freely distributable. In particular, we make use of 22 short stories collected from "Aesop's Fables"[5].

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 cases 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, we got five participants annotating 22 stories, three annotating 4 each and the rest two annotating 5 each. Each story was annotated by a single participant only.

For other participants who want 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. A participating system may use shallow natural language processing (NLP) techniques, such as named entity recognition and chunking, to first identify individual query concepts and then to retrieve candidate images for each of these. Another approach may be to use the entire text as query and then to cluster the result-list of documents to identify the individual query components.

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 the standard *pooling* procedure of IR, where a pool of documents, i.e. the set of top ranked documents from retrieval systems with different settings, is assessed manually for relevance. The relevance judgements for our dataset are obtained as follows.

Firstly, in order to be able to search for images with adhoc keywords, we indexed the ImageCLEF collection. In particular, 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 make use of the manually annotated concepts as an individual query that is executed on the document collection of the ImageCLEF. To construct the pool, we obtain runs with different retrieval models, such as BM25, LM and tf-idf with default parameter settings in Lucene and finally fuse the ranked lists with the standard COMBSUM merging technique.

Finally, top 20 documents from this fused ranked list were then 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. The participants were asked to assign relevance on a five point scale ranging from absolutely non-relevant to highly relevant.

4. OUR BASELINES

In this section, we describe some initial experiments that we conducted on our dataset, meant 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 by making use of the similarity of the query with the caption texts of the images in the index. The retrieval model that we use is the LM with Jelinek Mercer smoothing [5]. As a second baseline, we still use all the words in the story but this time weight each

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

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

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

query term by its tf-idf score. It is worth mentioning here that the two baselines that we use are quite simple because our intention is to see how simple methods can perform, before attempting to apply more involved approaches for this task.

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.

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. 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. We believe that shallow NLP techniques to extract useful concepts can further improve the results.

5. SUBMITTED RUNS

Two participating groups submitted runs for this task. The details about each group is shown in Table 2. The first group (Group 1) employed a word embedding based approach to expand the annotated concepts of each story to formulate a query and retrieve a ranked list of images. Only the text of the image captions was used for computing similarities with the queries. The similarity function employed was tf-idf. The second group (Group 2) used Terrier for indexing the Image CLEF 2010 collection. For retrieval, they applied the Divergence from Randomness (DFR) model similarity function of Terrier.

Table 3 shows the official results evaluated on the submitted runs by the two participating groups. Each participating group were allowed to submit three runs. While group 1 submitted only one run, the second group submitted three. It can be seen that the run submitted by Group 1 is comprised of a higher number of retrieved documents (6405) than the submitted runs of group 2 (about 100). Due to a higher value of the average number of retrieved images per story by group 1 (6405/22 ≈ 291) in comparison to group 2 (100/22 ≈ 4.5), group 1 achieves a higher recall and MAP (compare the #relret and MAP values in Table 3). However, the submitted runs from group 2 were scored high on precision, e.g. compare the MRR and the P@5 values between the runs of the two groups.

A comparison of the official results and our own baselines (see Tables 3 and 1 shows that none of the submitted runs were able to outperform the simple baseline approaches that we had experimented with. More investigation is required to comment on this observation which we leave for future work.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we describe the construction of a dataset for the purpose of evaluating automated approaches for document augmentation with images. In particular, we address 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 in 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 out the concepts from a story. The dataset contains the relevance assessments for each story obtained with pooling to a depth of 20.

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 some of the future directions of research for effective query representation for this task.

Grp	Affiliation	#members
1	Amrita Vishwa Vidyapeetham, Coimbatore, India	3
2	i) Charotar University of Science and Technology, Anand, India; ii) L.D.R.P. College, Gandhinagar, India; iii) Gujarat University, Ahmedabad, India.	4

Table 2: Participating groups for FIRE Automated Story Illustration task 2015.

Grp	Run	Evaluation Metrics					
Id	Id	#ret	#relret	MAP	MRR	B-pref	P@5
1	1	6405	**255**	**0.0107**	**0.1245**	0.1241	0.0636
2	1	92	16	0.0047	0.3708	0.0074	0.1273
2	2	95	20	0.0053	0.2997	0.0095	**0.1545**
2	3	100	13	0.0030	0.2504	0.0065	0.0909

Table 3: Official results of the FIRE Automated Story Illustration task 2015. The evaluation measures are averaged over the set of 22 stories (#rel: 2068).

References

[1] B. Caputo, H. Müller, J. Martínez-Gómez, M. Villegas, B. Acar, N. Patricia, N. B. Marvasti, S. Üsküdarli, R. Paredes, M. Cazorla, I. García-Varea, and V. Morell. 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, 2014.

[2] Y. Feng and M. Lapata. 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, 2010. Association for Computational Linguistics.

[3] M. M. Hall, P. D. Clough, O. L. de Lacalle, A. Soroa, and E. Agirre. 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, 2012. Association for Computational Linguistics.

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

[5] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.

[6] A. Popescu, T. Tsikrika, and J. Kludas. Overview of the wikipedia retrieval task at imageclef 2010. In M. Braschler, D. Harman, and E. Pianta, editors, *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

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

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