INAOE's participation at ImageCLEF 2016: Text Illustration Task

Luis Pellegrin, A. Pastor López-Monroy,
Hugo Jair Escalante, and Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
{pellegrin,pastor,hugojair,mmontesg}@inaoep.mx

Abstract. In this paper we describe the participation of the Language Technologies Lab of INAOE at ImageCLEF 2016 teaser 1: Text Illustration (TI). The goal of the TI task consists in finding the best image that describes a given document query. For evaluating this task, there is a dataset containing web pages having text and images. We address the TI as a purely Information Retrieval (IR) task, for a given document query we search for the most similar web pages and use the associated images to them as illustrations. In this way, queries are used to retrieve related images from web pages, but the retrieval result are only the associated images. For this, we represent the web pages and queries using state-of-the-art text representations. Those representations incorporate information that allows us to exploit textual or semantic aspects. According to ImageCLEF 2016 evaluation, the proposed approach holds the best performance for the TI task.

Keywords: text illustration, image retrieval, document representation.

1 Introduction

Since 2010, ImageCLEF promotes research into annotation of images using noisy web-page data. Following the same path, for the 2016 edition [1] two new tasks were introduced as teasers: Text illustration and Geolocation, this paper focuses in the former. The goal of the Text Illustration task consists in finding the best illustration, from a set of reference images, for a given text-document. Unlike the problem of illustrating a sentence formed by few words, the TI is a much more challenging task. The reason of this is that we want to illustrate a whole document (i.e. web page) including a number of different topics. In this regard, the used dataset consists in images embedded in web pages.

We address the TI problem as an Information Retrieval (IR) task. The hypothesis is that related web pages have related images. Thus, the document queries to be illustrated are a target set of web pages, which we illustrate using the embedded images of the retrieved web pages. For this, we bring two popular representations from the IR field, that do not take into account visual characteristics of images. On the one hand, the *bag-of-words* representation defines each document as histograms of word occurrences. On the other hand, the Word2vec representation incorporates distributional semantics to text documents with learned word vectors [2]. Finally, as work in progress we experiment with a third novel multimodal representation, where the textual and visual information are used to produce a multimodal representation of the queries. Such representation, allows us to directly retrieve images from a reference image dataset. The official results in the evaluation are encouraging and lays the background for future avenues of inquiry.

The remainder of the paper is organized as follows. Section 2 describes our method; Section 3 shows the obtained experimental results; finally, in Section 4 some conclusions of this work are presented.

2 Text illustration using an IR-based approach

To approach the TI task we consider the following elements in our strategy. Let $\mathcal{Q} = \{q_1, \ldots, q_m\}$ be the set of document queries to be illustrated. Also, let $\mathcal{D} = \{(d_1, I_1), \ldots, (d_n, I_n)\}$ be the set of web pages of d_n documents and I_n images pairs in the collection. Finally, let $\mathcal{V} = \{w_1, \ldots, w_r\}$ be the textual features extracted from documents in the reference collection $\mathcal{D}$. The general process of the proposed approach has two stages. The first consist in representing each query $\mathbf{q}_j$ and each document $\mathbf{d}_i$ into the same space $\mathbb{R}^r$. In the second stage, each query $q_j \in \mathcal{Q}$ is used to retrieve the k most similar web pages $\{(d_h, I_h) : (d_h, I_h) \in \mathcal{D}\}$ to q_j. The final result only considers the I_h elements as the resultant illustration set. The rest of this section explains the stages in detail.

2.1 Representing documents

The first stage in our strategy requires computing query vectors $\mathbf{q}_j = \langle w_1, \ldots, w_r \rangle$ and document vectors $\mathbf{d}_i = \langle w_1, \ldots, w_r \rangle$ in a space $\mathbb{R}^r$. For this, we relied in two different textual representations exploiting word occurrences (i.e., in BoW) and co-occurrences (i.e., in Word2vec) in documents, as described below. Note that $|r|$ is defined according to each representation. For the case of BoW, $|r| = |\mathcal{V}|$. In the case of Word2vec, $|r|$ is number of hidden neurons used to represent the learned word vectors.

Bag-of-Words (BoW) Under BoW each document is represented by taking each word in the vocabulary as an attribute to build document vectors $\mathbf{d}_i = \langle w_1, \ldots, w_r \rangle$. Intuitively, the BoW is an histogram representing word frequencies in each document. BoW representation was built filtering terms with high frequency and using TF-IDF (Term Frequency Inverse Document Frequency) weighting scheme [3].

Word2vec Adaptation The purpose of Word2vec is to build accurate representations of words in a space $\mathbb{R}^r$. The main goal is that semantically related words should have similar word vectors in $\mathbb{R}^r$ [2]. For instance, *Paris* vector are close to *Berlin* vector, since both are capitals. Surprisingly, Mikolov et. a.l (2013) also showed other generalizations using specific lineal operations. For example, *France-Paris+Berlin* result in a very close vector to *Germany.* In this paper, we exploit the use of learned word vectors from Wikipedia using Word2vec [2]. For our experiments, the learned word vectors from each document are used to compute the average document vector as in [4]. The idea is that the average of those word representations, should capture rich notions of semantic relatedness and compositionality of the whole document.

2.2 Retrieval stage

In this stage, a document query q_j under a specific representation is used to retrieve a set of relevant items $\{(d_h, I_h) : (d_h, I_h) \in \mathcal{D}\}$. Note that only the textual information from web pages and textual queries are used in the retrieval stage, but the reported results correspond to the immersed images in the retrieved items. For the retrieval stage we used the cosine similarity measure, which is defined in Equation 1.

$$similarity(q_j, d_i) = cosine(\mathbf{q}_j, \mathbf{d}_i) = \frac{\mathbf{q}_j * \mathbf{d}_i}{||\mathbf{q}_j||||\mathbf{d}_i||} \tag{1}$$

where $\mathbf{q}_j, \mathbf{d}_i$ are the representations of the document query q_j, and the i^{th} document d_i from the collection, respectively. This equation iterates over all documents from $\mathcal{D}$, then the images associated to the k most similar documents to q_j are used to illustrate it.

3 Experimental Results

In this section we present qualitative and quantitative results of the proposed approach in the TI task.

3.1 Quantitative results

In Table 1, it can be seen the performance of the proposed representations for TI. The table reports scores from the metric proposed in [5], where basically the recall is evaluated at the k-th rank position (R@K) of the ground truth images. Several values of k are reported, in Table 1 we can see the scores that correspond to the test set.

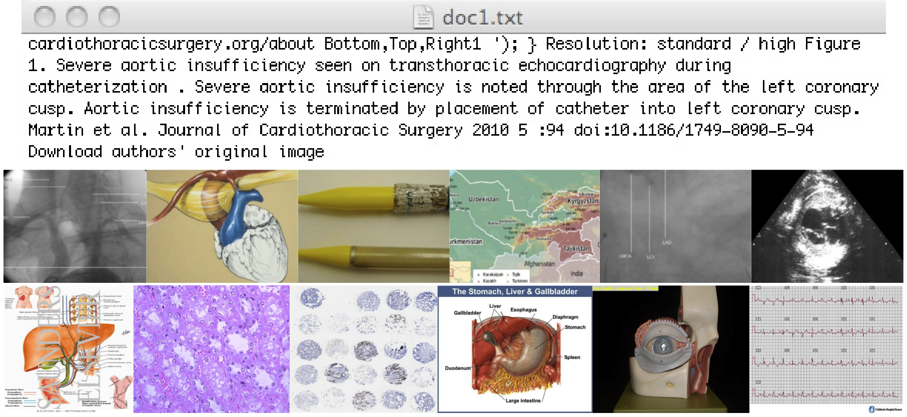
Our best score is reported by run1, which uses the BoW under a TF-IDF weighting scheme filtering 5% of the highest frequent terms. The results obtained by run1 validate our hypothesis that related images appear in related web pages. On the other hand, the run2 and run3 report scores obtained by Word2vec representation (denoted as d2v), both runs use also a filtering of 5%, but with and without TF-IDF weighting respectively. In these latter results, we consider that the representation is affected by noise when increasing the number of used word to build it. Although, a Word2vec representation helps to retrieve similar documents (as is showed in Figure 1), we have found that this representation is more confident with short documents or in specific domains. However, in documents with diversity of topics the performance decrease (see Figure 2) because of the great variety of different words involved.

Table 1: Recall@K for full 180K test set.

Team	RUN	Recall (%)						
		R@1	R@5	R@10	R@25	R@50	R@75	R@100
Baseline	chance	0.00	0.00	0.01	0.01	0.03	0.04	0.05
CEA	cbs.flickrgroup.FS.valid	0.02	0.10	0.22	0.48	0.84	1.16	1.44
	cbs.wordnet.FS.valid	0.03	0.12	0.23	0.53	0.97	1.38	1.74
	cbs.mergeA.valid	0.14	0.56	0.97	1.90	2.98	3.82	4.47
	cbs.mergeB.valid	0.18	0.63	1.05	1.97	3.00	3.87	4.51
	cbs.mergeC.valid	0.18	0.62	1.04	1.95	2.99	3.85	4.50
	wam5.kcca1.idsQueries.all.valid	0.11	0.36	0.62	1.11	1.68	2.11	2.47
	warm7.idsQueries.10BWS.all.valid	0.18	0.63	1.07	1.93	2.93	3.69	4.33
INAOE	**run1.bow+tfidf.thr5p**	**28.75**	**63.50**	**75.48**	**84.39**	**86.79**	**87.36**	**87.59**
	run2.d2v.thr5p	2.57	5.65	7.71	11.76	16.69	20.34	23.40
	run3.d2v+tfidf.thr5p	3.68	7.73	10.46	15.62	21.36	25.48	28.78

3.2 Qualitative results

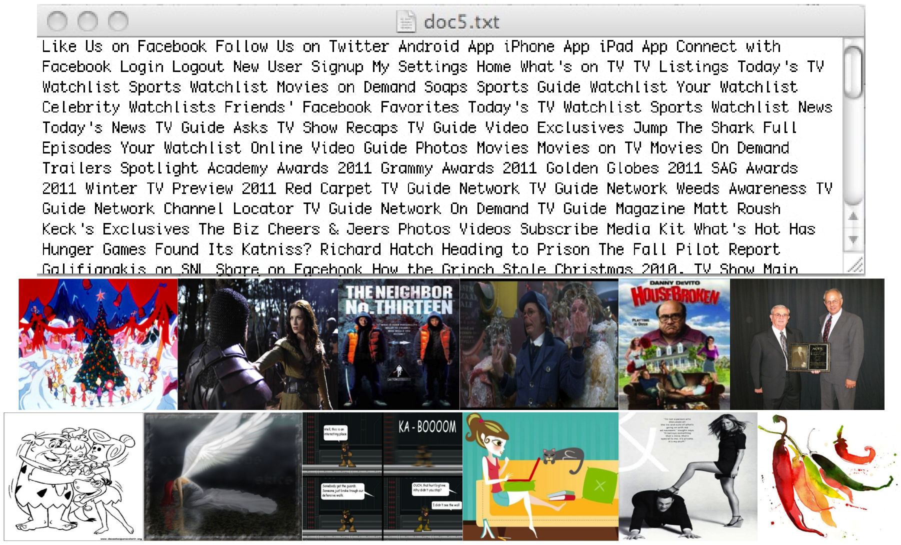
In this subsection we compare the proposed representation. In Figure 1, we show top retrieved images that illustrate the document query under two representations. In this case, the document query consists in a short text, we can see that both representations show relevant images to illustrate the text. An interesting output is obtained by run3 that shows diversity on retrieved images.



```
cardiothoracicsurgery.org/about Bottom,Top,Right1 '); } Resolution: standard / high Figure
1. Severe aortic insufficiency seen on transthoracic echocardiography during
catheterization . Severe aortic insufficiency is noted through the area of the left coronary
cusp. Aortic insufficiency is terminated by placement of catheter into left coronary cusp.
Martin et al. Journal of Cardiothoracic Surgery 2010 5 :94 doi:10.1186/1749-8090-5-94
Download authors' original image
```

Fig. 1: Given the text document (top), the top of retrieved images. First row, output images from run1. Second row, output images from run3.

On the other hand, the Figure 2 shows a long document used as query. Again, the outputs of run1 and run3 are compared. Despite that in the document are included great quantity of topics, the image retrieval of run1 is effective, but the image retrieval of run3 includes few relevant images. Taking as examples the Figures 1 and 2, we can see that the quantity of terms and rich vocabularies contained in the documents is an important factor for selecting the representation. While Word2vec representation seems to be robust in short documents or documents in a specific domain, the BoW representation plus weighting TF-IDF shows to be a better option for the case of long documents.



```
Like Us on Facebook Follow Us on Twitter Android App iPhone App iPad App Connect with
Facebook Login Logout New User Signup My Settings Home What's on TV TV Listings Today's TV
Watchlist Sports Watchlist Movies on Demand Soaps Sports Guide Watchlist Your Watchlist
Celebrity Watchlists Friends' Facebook Favorites Today's TV Watchlist Sports Watchlist News
Today's News TV Guide Asks TV Show Recaps TV Guide Video Exclusives Jump The Shark Full
Episodes Your Watchlist Online Video Guide Photos Movies Movies on TV Movies On Demand
Trailers Spotlight Academy Awards 2011 Grammy Awards 2011 Golden Globes 2011 SAG Awards
2011 Winter TV Preview 2011 Red Carpet TV Guide Network TV Guide Network Weeds Awareness TV
Guide Network Channel Locator TV Guide Network On Demand TV Guide Magazine Matt Roush
Keck's Exclusives The Biz Cheers & Jeers Photos Videos Subscribe Media Kit What's Hot Has
Hunger Games Found Its Katniss? Richard Hatch Heading to Prison The Fall Pilot Report
Galifianakis on SNL Share on Facebook How the Grinch Stole Christmas 2010. TV Show Main
```

Fig. 2: Given the text document (top), the top of retrieved images. First row, output images from run1. Second row, output images from run3.

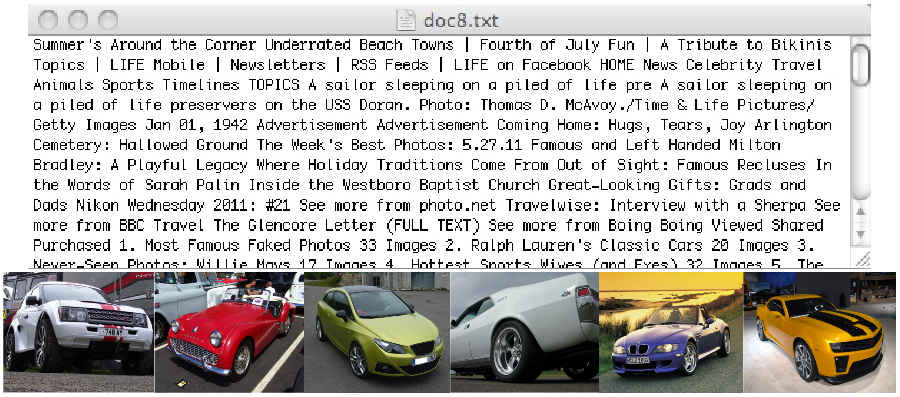
3.3 Work in progress: representing documents in a visual space

We have worked with a visual representation but it is not reported in Table 1. Unfortunately, we were not able to submit a run because of the tight time for the deadline. Nevertheless, we also present an in-house evaluation showing qualitative results.

For representing documents in a visual space, we used a multimodal representation $\mathcal{M}$ composed by visual prototypes. The construction of $\mathcal{M}$ is performed in an unsupervised way by using images immersed in web pages. The idea is that images can be represented by two different modalities: a visual representation extracted from the image $\mathcal{I}$, and a textual representation extracted from the web pages $\mathcal{D}$. In $\mathcal{M}$ for every word in $\mathcal{D}$ a visual prototype is formed, where each prototype is a distribution over visual representation (more detail of this approach in [6]). We used a reference image dataset (training set of [1]) for construction of $\mathcal{M}$.

The aim of this representation is to include the visual information in the text illustration. Under this representation, the words from a given document query are seen in function of its visual representation. First, using visual prototypes of words extracted from a query, then an average visual prototypes is formed. Second, using average visual prototype as query, then we retrieve some related images. In other words, the document query is translated to a visual document and used it to retrieve images, as a CBIR (Content-Based Image Retrieval) task.

In Figure 3, we show one favorable case for the visual representation. However, the average visual prototype in this case is formed by three words of the document query with the highest weight. For this kind of representation, we have observed that the more terms in the document query, the more noisy the visual representation is. As conclusion, a visual document representation is formed only by few words, so it is necessary a keyword extraction process on document query.



```
Summer's Around the Corner Underrated Beach Towns | Fourth of July Fun | A Tribute to Bikinis
Topics | LIFE Mobile | Newsletters | RSS Feeds | LIFE on Facebook HOME News Celebrity Travel
Animals Sports Timelines TOPICS A sailor sleeping on a piled of life pre A sailor sleeping on
a piled of life preservers on the USS Doran. Photo: Thomas D. McAvoy./Time & Life Pictures/
Getty Images Jan 01, 1942 Advertisement Advertisement Coming Home: Hugs, Tears, Joy Arlington
Cemetery: Hallowed Ground The Week's Best Photos: 5.27.11 Famous and Left Handed Milton
Bradley: A Playful Legacy Where Holiday Traditions Come From Out of Sight: Famous Recluses In
the Words of Sarah Palin Inside the Westboro Baptist Church Great-Looking Gifts: Grads and
Dads Nikon Wednesday 2011: #21 See more from photo.net Travelwise: Interview with a Sherpa See
more from BBC Travel The Glencore Letter (FULL TEXT) See more from Boing Boing Viewed Shared
Purchased 1. Most Famous Faked Photos 33 Images 2. Ralph Lauren's Classic Cars 20 Images 3.
Never-Seen Photos: Willie Mays 17 Images 4. Hottest Sports Wives (and Exes) 32 Images 5. The
```

Fig. 3: Given the text document (top), the top of retrieved images. Output images using visual representation.

4 Conclusions

In this paper we presented an IR approach to address the Text Illustration task. The documents are defined as textual, semantic or visual representations. The performed experiments under these different representations give an initial point of comparison for future approaches. According to the performed evaluation we conclude that, related web pages have related images, then it is possible to retrieve highly relevant elements using IR techniques. On the one hand, the BoW obtained outstanding performances because of the filtering of high frequent terms and the discriminative information captured by TF-IDF weighting scheme. On the other hand, Word2vec representation did not obtain reliable representations because of the great diversity of words involved in web pages. Such diversity makes difficult to build accurate document representations using the simple average of words. Our perspectives for future work include exploring relationships between representation to incorporate mix information (textual-visual) and adding a keyword extraction for the document query.

Acknowledgments. This work was supported by CONACYT under project grant CB-2014-241306 (Clasificación y recuperación de imágenes mediante técnicas de minería de textos). The first author was supported by CONACyT under scholarship No. 214764, and López-Monroy thanks for doctoral scholarship CONACyT-México 243957.

References

1. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, Évora, Portugal (2016)
2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
3. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28** (1972) 11–21
4. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR **abs/1405.4053** (2014)
5. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Int. Res. **47** (2013) 853–899
6. Pellegrin, L., Vanegas, J.A., Arevalo, J., Beltrán, V., Escalante, H.J., Montes-Y-Gómez, M., González, F.: INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, Toulouse, France (2015)