The BreakingNews Dataset

Arnau Ramisa*
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
aramisa@iri.upc.edu

Fei Yan*
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
f.yan@surrey.ac.uk

Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
fmoreno@iri.upc.edu

Krystian Mikolajczyk
Department of Electrical and Electronic Engineering, Imperial College London, UK
k.mikolajczyk@imperial.ac.uk

Abstract

We present BreakingNews, a novel dataset with approximately 100K news articles including images, text and captions, and enriched with heterogeneous meta-data (e.g. GPS coordinates and popularity metrics). The tenuous connection between the images and text in news data is appropriate to take work at the intersection of Computer Vision and Natural Language Processing to the next step, hence we hope this dataset will help spur progress in the field.

1 Introduction

Current successes in the crossroads between NLP and computer vision indicate that the techniques are mature for more challenging objectives than those posed by existing datasets. The NLP community has been addressing tasks such as sentiment analysis, popularity prediction, summarization, source identification or geolocation to name a few, that have been relatively little explored in computer vision. *BreakingNews* is a large-scale dataset[1] of news articles with rich meta-data and, we believe, an excellent benchmark for taking joint vision and language developments a step further. In contrast to existing datasets, the link between images and text in BreakingNews is not as direct, i.e., the objects, actions and attributes of the images may not explicitly appear as words in the text (see example in Fig. 1). The visual-language connections are more subtle and learning them will require the development of new inference tools able to reason at a higher and more abstract level. Furthermore, besides tackling article illustration or image captioning tasks, the proposed dataset is intended to address new challenges, such as source/media agency detection, estimation of GPS coordinates, or popularity prediction (which we annotate based on the reader comments and number of re-tweets).

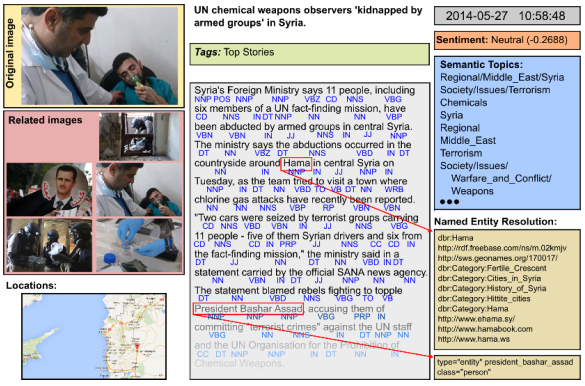

Figure 1: Example article with annotations from the BreakingNews dataset.

In (Ramisa et al., 2016) we present several baseline results for different tasks using this dataset.

2 Description of the Dataset

The *BreakingNews* dataset consists of approximately 100,000 articles published between the 1st of January and the 31th of December of 2014. All articles include at least one image, and cover a wide variety of topics, including sports, politics, arts, healthcare or local news.

The main text of the articles was downloaded using the IJS newsfeed (Trampuš and Novak, 2012), which provides a clean stream of semantically enriched news articles in multiple languages from a pool of *rss* feeds.

We restricted the articles to those that were written in English, contained at least one image, and originated from a shortlist of highly-ranked news media agencies (see Table 1) to ensure a degree of

* denotes equal contribution

[1] http://www.iri.upc.edu/people/aramisa/BreakingNews/index.html

Proceedings of the 6th Workshop on Vision and Language, pages 38–39,
Valencia, Spain, April 4, 2017.

Source	num. articles	avg. len. article	avg. num. images	avg. len. caption	avg. num. comments	avg. len. comment	avg. num. shares	% geo-located
Yahoo News	10,834	521 ± 338	1.00 ± 0.00	40 ± 33	126 ± 658	39 ± 71	n/a	65.2%
BBC News	17,959	380 ± 240	1.54 ± 0.82	14 ± 4	7 ± 78	48 ± 21	n/a	48.7%
The Irish Independent	4,073	555 ± 396	1.00 ± 0.00	14 ± 14	1 ± 6	17 ± 5	4 ± 20	52.3%
Sydney Morning Herald	6,025	684 ± 395	1.38 ± 0.71	14 ± 10	6 ± 37	58 ± 55	718 ± 4976	60.4%
The Telegraph	29,757	700 ± 449	1.01 ± 0.12	16 ± 8	59 ± 251	45 ± 65	355 ± 2867	59.3%
The Guardian	20,141	786 ± 527	1.18 ± 0.59	20 ± 8	180 ± 359	53 ± 64	1509 ± 7555	61.5%
The Washington Post	9,839	777 ± 477	1.10 ± 0.43	25 ± 17	98 ± 342	43 ± 50	n/a	61.3%

Table 1: Dataset statistics. Mean and standard deviation, usually rounded to the nearest integer.

consistency and quality. Given the geographic distribution of the news agencies, most of the dataset is made of news stories in English-speaking countries in general, and the UK in particular. For each article we downloaded the images, image captions and user comments from the original article webpage. News article images are quite different from those in existing captioned images datasets like Flickr8K (Hodosh et al., 2013) or MS-COCO (Lin et al., 2014): often include close-up views of a person (46% of the pictures in BreakingNews contain faces) or complex scenes. Furthermore, news image captions use a much richer vocabulary than in existing datasets (e.g. Flickr8K has a total of 8,918 unique tokens, while eight thousand random captions from BreakingNews already have 28,028), and they rarely describe the exact contents of the picture.

We complemented the original article images with additional pictures downloaded from Google Images, using the full title of the article as search query. The five top ranked images of sufficient size in each search were downloaded as potentially related images (in fact, the original article image usually appears among them).

Regarding measures of article popularity, we downloaded all comments in the article page and the number of shares on different social networks (e.g. Twitter, Facebook, LinkedIn) if this information was available. Whenever possible, in addition to the full text of the comments, we recovered the thread structure, as well as the author, publication date, likes (and dislikes) and number of replies. Since there were no share or comments information available for "The Irish Independent", we searched Twitter using the full title and collected the tweets that mentioned a name associated with the newspaper (e.g. @Independent_ie, Irish Independent, @IndoBusiness) or with links to the original article in place of comments. We considered the collective number of re-tweets as shares of the article. The IJS Newsfeed annotates

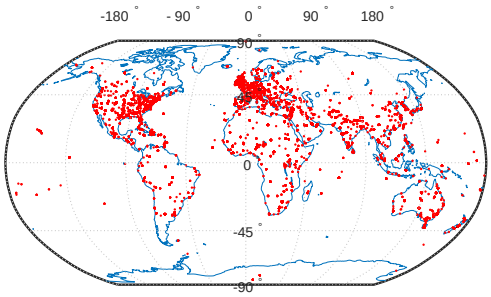

Figure 2: Ground truth geolocations of articles.

the articles with geolocation information both for the news agency and for the article content. This information is primarily taken from the provided RSS summary, but sometimes it is not available and then it is inferred from the article using heuristics such as the location of the publisher, TLD country, or the story text. Fig. 2 shows a distribution of news story geolocation.

Finally, the dataset is annotated for convenience with shallow and deep linguistic features (e.g. part of speech tags, inferred semantic topics, named entity detection and resolution, sentiment analysis) with *XLike* [2] and *Enrycher* [3] NLP pipelines.

References

M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk. 2016. Breakingnews: Article annotation by image and text processing. *CoRR*, abs/1603.07141.

M. Trampuš and B. Novak. 2012. Internals of an aggregated web news feed. In *International Information Science Conference IS*, pages 431–434.

[2] http://www.xlike.org/language-processing-pipeline/

[3] http://ailab.ijs.si/tools/enrycher/