# A Corpus of Images and Text in Online News

**Laura Hollink*, Adriatik Bedjeti*, Martin van Harmelen* and Desmond Elliott*†**
*Centrum Wiskunde & Informatica (CWI)
Amsterdam, The Netherlands
{l.hollink, a.bedjeti}@cwi.nl; martin@vanharmelen.com

†ILLC, University of Amsterdam
Amsterdam, The Netherlands
d.elliott@uva.nl

**Abstract**

In recent years, several datasets have been released that include images and text, giving impulse to new methods that combine natural language processing and computer vision. However, there is a need for datasets of images in their natural textual context. The ION corpus contains 300K news articles published between August 2014 - 2015 in five online newspapers from two countries. The 1-year coverage over multiple publishers ensures a broad scope in terms of topics, image quality and editorial viewpoints. The corpus consists of JSON-LD files with the following data about each article: the original URL of the article on the news publisher's website, the date of publication, the headline of the article, the URL of the image displayed with the article (if any), and the caption of that image. Neither the article text nor the images themselves are included in the corpus. Instead, the images are distributed as high-dimensional feature vectors extracted from a Convolutional Neural Network, anticipating their use in computer vision tasks. The article text is represented as a list of automatically generated entity and topic annotations in the form of Wikipedia/DBpedia pages. This facilitates the selection of subsets of the corpus for separate analysis or evaluation.

**Keywords:** online news; image features; topic extraction

## 1. Introduction

In recent years, several datasets have been released that include images and text (Ferraro et al., 2015; Bernardi et al., 2016), giving impulse to new methods that combine natural language processing and computer vision, such as automatic image description (Fang et al., 2014) and image–sentence matching (Hodosh et al., 2013). Current datasets typically consist of user-captioned images from Flickr[1] (Ordonez et al., 2011; Chen et al., 2015) or images with descriptions produced by crowd workers (Rashtchian et al., 2010; Elliott and Keller, 2013; Zitnick and Parikh, 2013; Lin et al., 2014; Young et al., 2014). These datasets can be used to train systems to produce Flickr-like captions or crowdsourced descriptions, but we argue there is a need for more datasets for generating *captions in context*. This argument has previously been made by Feng and Lapata (2008), who released a dataset of 3,361 news articles from BBC News, including images and real-world captions; Tirilly et al. (2010) collected the texts, images and captions of 27,000 French newspaper articles, and used the real-world captions as a ground truth to evaluate an image annotation algorithm. As a continuation of these efforts, we present the *Images in Online News* (ION) corpus.

The ION corpus contains news articles published between August 2014 - 2015 in five online newspapers from two different countries. It includes more than 323,707 articles, making it larger than the existing datasets of news images and text. The 1-year coverage over multiple publishers ensures a broad scope in terms of topics, image quality, and editorial / political viewpoints. We hope the dataset is valuable not only for language and vision researchers, but also for communication scientists studying images in the media. The current working practice of most communication scientists is to manually analyze source materials, which is time consuming and leads to studies on small datasets. For example, in Greenwood and Jenkins (2015), 192 news photos were coded to determine the visual framing of conflict in magazines; in Esser (2008), a team of coders watched 45.3 hours of televised news to compare news culture in four countries. We hypothesize these types of studies will eventually be accompanied by automatic analysis of larger quantities of visual material. By publishing this corpus we aim to contribute to this research direction.

The ION corpus consists of JSON-LD files with the following data about each article: the original URL of the article on the news publisher's website, the date of publication, the headline of the article, the URL of the image displayed with the article (if any), and the caption of that image (if present). Neither the article text nor the images themselves are included in the corpus due to varying copyright restrictions, unlike those in the BBC News Dataset (Feng and Lapata, 2008). In the ION corpus the images are distributed as high-dimensional feature vectors extracted from a Convolutional Neural Network, anticipating their use in computer vision tasks. The article text is represented as a list of automatically generated entity and topic annotations in the form of Wikipedia/DBpedia pages. This facilitates, for example, the selection of subsets of the corpus for separate analysis or evaluation. The use of DBpedia allows us to take advantage of the rich and diverse knowledge in the Linked Open Data Cloud.

This paper details how the corpus is created and how it can be used. The JSON-LD dataset files are accompanied by the source code used to extract metadata from the websites, and to obtain and process the images and texts.

[1] https://www.flickr.com/

## 2. Dataset Creation

Newspaper websites are selected for inclusion in the ION corpus based on estimates of their number of page views and unique visitors, as provided by the web analytics company Alexa. The dataset includes five of the most visited websites in the category 'newspapers'[2]: the US-based *The New York Times*, *The Washington Post* and *The Huffington Post*, and the UK-based *The Daily Mail* and *The Independent*. ION includes all articles between August 13, 2014 and August 13, 2015. Figure 1 illustrates the steps in the dataset creation process. Each step will be briefly discussed below.

### 2.1. Retrieving Articles and Metadata Extraction

For each newspaper, we retrieve the URLs of all news articles published in the past year. The URLs are obtained using the search engines provided on the newspaper websites or their online archives. Given that each search engine was implemented differently – for example, with respect to handling of stop words, wild cards and article metadata – each website requires a different query to retrieve all articles. *The New York Times* and *The Daily Mail* returned the most results with a '*' query; for *The Washington Post* we used 'the', whereas for *The Independent* 'has OR have OR had OR independent'. For *The Huffington Post* the URLs were taken from their sorted online archive. The result lists are filtered to include only regular news articles and no images galleries, video clips without text, recipes, or employee profile pages. If no such filtering was possible, as was in the case of *The Huffington Post*, a filtering was performed based on the URL-patterns.

Second, we download the HTML stored at each URL and extract the publication date, headline, article text, image, image caption, and image URL based on tailor-made XPath queries. Figure 2 shows an overview of the components of a news article that are being used in the creation of the ION corpus. The article text and image files are stored locally as `.txt` and `.jpg` files, respectively. They are not published as part of the corpus due to copyright reasons, but are processed locally to extract topics and image feature vectors, as described below. News articles occasionally contain multiple images, but we only provide the URL, caption and feature vectors of the first appearing image.

### 2.2. Topic Annotation

The article texts are processed using the TextRazor API. TextRazor[3] is a commercial tool that provides several NLP modules. We use its entity linking service, which scored best in terms of precision (but not recall) in a recent comparison to other entity linkers (Derczynski et al., 2015). This service annotates a given input text with Wikipedia page URLs. We use it to obtain two types of annotations: (1) high-level categories (called "Coarse Topics" in TextRazor), which are selected from a limited number of Wikipedia Category pages, and (2) more specific entities and topics, which can be any type of Wikipedia entry.

[2] `http://www.alexa.com/topsites/category/Top/News/Newspapers`

[3] `https://www.textrazor.com`

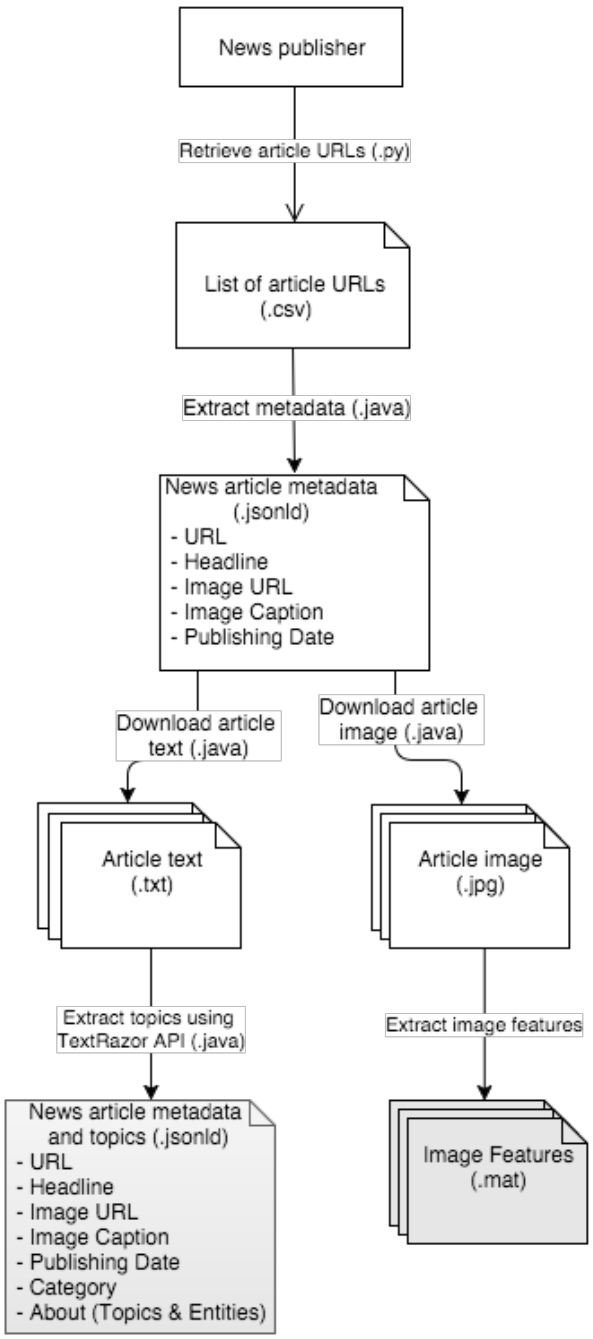


Figure 1: The ION dataset construction pipeline. The highlighted entities (`.jsonld` and `.mat`) represent the published dataset.

Given an input text, TextRazor returns a ranked list of annotations. We retain only those with a relevance score of 1.0. For the current dataset this results in an average of 5 high-level categories and 25 topics / entities per article. Figure 3 gives an overview of the contents of the dataset by showing the distribution of high-level categories for each news publisher.

### 2.3. Visual Feature Extraction

The images are available as high-dimensional feature vectors extracted from a Convolutional Neural Network object recognition model (Simonyan and Zisserman, 2015). We use CNN-based image features because they have proven

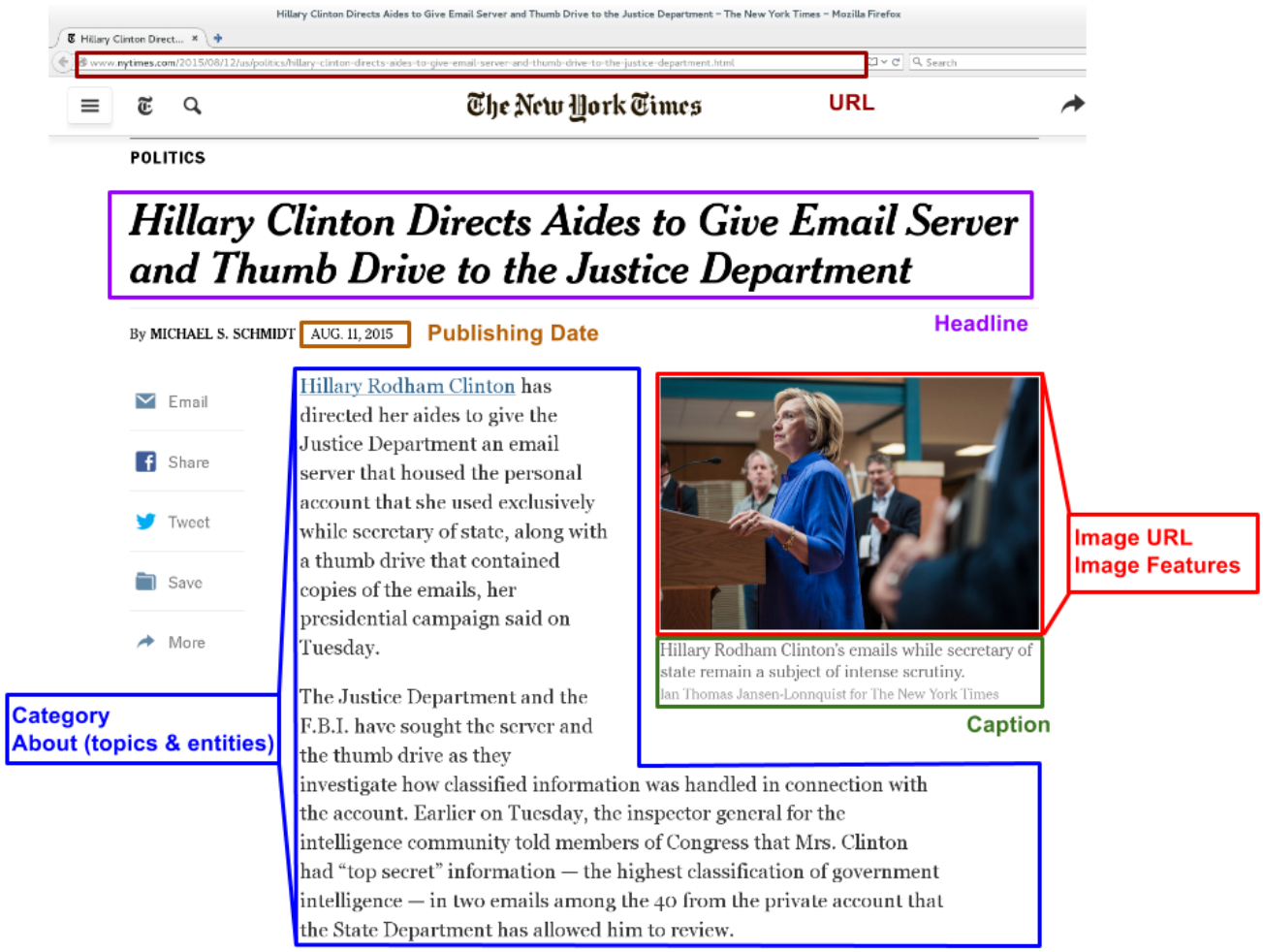


Figure 2: The metadata extracted from articles on *The New York Times*

useful for a wide-range of computer vision tasks. The feature vectors were extracted from the final fully-connected layer of the VGG-19 model (specifically, the layer labelled 'relu7'), originally trained for 1000-class object recognition (Russakovsky et al., 2015). The resulting 4096-dimension feature vector of all images of a news publisher are stored as a separate MATLAB[4] (.mat) file, which together with the JSON-LD file forms the ION corpus.

## 3. Format, availability and reuse

The ION corpus is published as JSON-LD, a specification for representing Linked Data in the popular JSON format (Sporny et al., 2014), and accompanied by MATLAB files for the image feature vectors. Schema.org[5] was chosen as the vocabulary for the Linked Data properties for its wide-spread use in various communities in academia and industry. An excerpt of an article from *The New York Times* is presented in Appendix A.

The JSON-LD and MATLAB files are available for download from a persistent URL: `http://persistent-identifier.org/?identifier=urn:nbn:nl:ui:18-24394`. The data are stored as one JSON-LD file per news website.

All code is available on *GitHub*[6] under an MIT license[7]. The repository includes the URL lists and the code to extract metadata, to download the article text and images, and to extract topics using the TextRazor API. Publishing the source code helps users judge the quality, coverage, and reliability of the corpus (Traub and van Ossenbruggen, 2015). It will also alleviate the efforts needed to create similar corpus, or over different periods of time. Finally, it facilitates downloading the article texts and images from the article URLs for further processing.

## 4. Expected Use

We hope the ION corpus will be valuable for many tasks in computer vision, natural language processing, semantic web, and media studies. Below we give three examples of tasks that we are currently exploring.

**News Image Captioning** This is the task of generating a caption of a news article image (Feng and Lapata, 2008). The main difference between news image captioning and automatic image description is news images are rarely captioned with literal texts (Hodosh et al., 2013). We aim to stimulate research into this task by providing real-world

[4] `http://mathworks.com/products/matlab/`
[5] `http://schema.org/`
[6] `https://github.com/abedjeti/ADS-NewsArticles`
[7] `https://opensource.org/licenses/MIT`

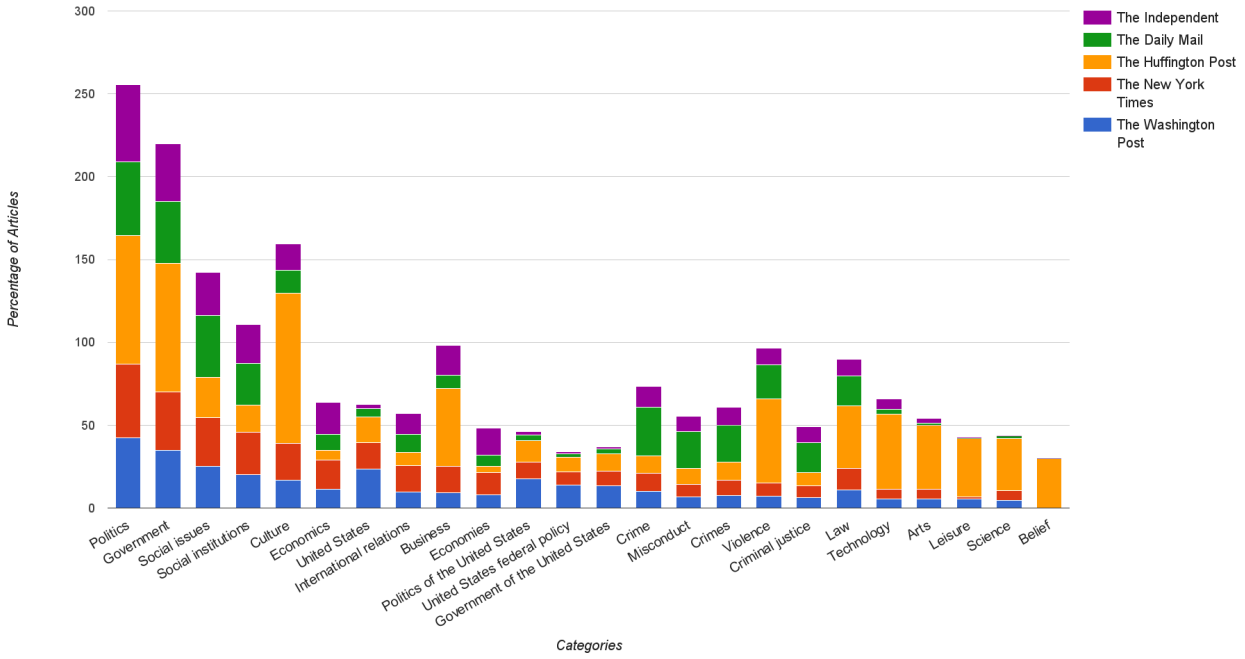


Figure 3: The distribution of high-level categories ("Coarse topics") from TextRazor for all five news publishers, based on the percentage of articles about a particular category

captions, in contrast to those in existing datasets, which are typically collected by crowdsourcing.

In this task, the surrounding article context is expected to improve the performance of caption generation models. The Extended Relevance Annotation Model was significantly improved with additional context from the surrounding article (Feng and Lapata, 2008). The ION corpus provides some of the original context (headlines, topics, publisher, date), aiming to progress research into how each part of the article context contributes to the model. The eventual inclusion of articles from many news providers will allow researchers to study the effect of context across publishers.

**Training visual classifiers using captions** Learning visual classifiers is expensive because of the need for large amounts of labeled images. Hence, researchers have sought alternatives to manual annotation. For example, Guillaumin et al. (2010) use Flickr image tags to learn classifiers. Text associated with images are another alternative to manually created training sets, assuming the text mentions what is visible in the images. Tirilly et al. (2010) use captions as a ground truth to evaluate their annotation algorithm. To support this kind of research on how to use captioned images for training visual classifiers, Ozcan et al. (2011) released the FAN-Large database of 125,479 captioned images of celebrities downloaded from the Internet. The ION corpus is suitable for similar purposes, such as training visual classifiers for individual political leaders. Note, however, that while the FAN-Large corpus includes crowd-sourced annotations of the faces in the images based on the captions, the ION corpus only includes the raw captions.

**Large scale media analysis** Images are a powerful tool in the framing of a story in a news article (Rodriguez and Dimitrova, 2011). Several studies have analyzed the content of news images to identify how a particular event or topic is framed (Fahmy and Kim, 2008; Parry, 2010). We are currently exploring to what extent the ION corpus can be used to identify patterns in the (editors') choice of what to display with news articles about a particular event or topic. As an example, we look at what is depicted with articles about male and female political leaders. For this task, we need both the visual feature vectors and the extracted Wikipedia/DBpedia annotations. The first will be used for automatic image annotation. The latter will be used to select articles about male and female leaders. The background knowledge in DBpedia is crucial here, both for identifying whether a detected entity is a political leader and for information about their gender. Whether this kind of study will prove to be feasible depends heavily on the quality of the visual classifiers and topic/entity linkers (in this case TextRazor). For future work, we plan an evaluation of a sample of the annotations.

## 5. Conclusion and Future work

We introduced the *Images in Online News* (ION) corpus of images and text in online news articles. The main advantages of collecting image–text datasets from online news articles are: the more natural relationship between the images and text, compared to crowdsourced datasets. And in comparison to social media datasets, the images in online news appear in a broader article-wide context.

The corpus contains more than 300K news articles from five news publishers. The source code for collecting ION is freely available to help users judge the quality, coverage, and reliability of the corpus. We hope the source code will help with future efforts to create similar corpora or to extend the ION corpus to different periods of time.

For future work, we will evaluate of the accuracy of the TextRazor topic annotations over a sample of the corpus. It would be useful to know how well TextRazor performs on such a broad corpus. We would like to experiment with generating image captions in the context of entire newspaper articles. Finally, we plan on extending ION to also cover non-English news publishers, especially because TextRazor can extract topic information in multiple languages.

## Acknowledgements

Adriatik Bedjeti was supported by Amsterdam Data Science. Desmond Elliott was supported by ERCIM ABCDE Fellowship 2014-23.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *JAIR*, pages 409–442.

Chen, J., Kuznetsova, P., Warren, D. S., and Choi, Y. (2015). Deja image-captions: A corpus of expressive descriptions in repetition. In *NAACL*.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32 – 49.

Elliott, D. and Keller, F. (2013). Image Description using Visual Dependency Representations. In *EMNLP*.

Esser, F. (2008). Dimensions of political news cultures: Sound bite and image bite news in france, germany, great britain, and the united states. *The International Journal of Press/Politics*, 13(4):401–428.

Fahmy, S. and Kim, D. (2008). Picturing the iraq war: Constructing the image of war in the british and us press. *International Communication Gazette*, 70(6):443–462.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al. (2014). From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *ACL*, pages 272–280.

Ferraro, F., Mostafazadeh, N., Huang, T.-H. K., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (2015). A survey of current datasets for vision and language research. In *EMNLP*, pages 207–213.

Greenwood, K. and Jenkins, J. (2015). Visual framing of the syrian conflict in news and public affairs magazines. *Journalism Studies*, 16(2):207–227.

Guillaumin, M., Verbeek, J., and Schmid, C. (2010). Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 902–909. IEEE.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Lin, T., Maire, M., Belongie, S., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Ozcan, M., Luo, J., Ferrari, V., and Caputo, B. (2011). A large-scale database of images and captions for automatic face naming. Technical report, Idiap.

Parry, K. (2010). A visual framing analysis of british press photography during the 2006 israel-lebanon conflict. *Media, War & Conflict*, 3(1):67–85.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Rodriguez, L. and Dimitrova, D. V. (2011). The levels of visual framing. *Journal of Visual Literacy*, 30(1):48–65.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2014). Json-ld 1.0. *W3C Recommendation (January 16, 2014)*.

Tirilly, P., Claveau, V., Gros, P., et al. (2010). News image annotation on a large parallel text-image corpus.

Traub, M. C. and van Ossenbruggen, J. (2015). Workshop on tool criticism in the digital humanities. In *Workshop on Tool Criticism in the Digital Humanities*, volume 2, page 7.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Zitnick, C. L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *CVPR*, pages 3009–3016.

# A JSON-LD format of the ION corpus

Listing 1: An excerpt from the ION corpus showing the JSON-LD context and one article from the New York Times.


```
{
"@context": {
        "@vocab": "http://schema.org/",
        "datePublished": {
                "@type": "http://www.w3.org/2001/XMLSchema#dateTime"
        }
},
"@id": "http://www.nytimes.com/",
"@type": "Newspaper",
"name": "New York Times",
"@reverse": {
   "publisher": [
   {
        "@id": "1cf5e45097bb2e1f036824790e955cd52e84751d",
        "datePublished": "2015-08-12",
        "headline": "Hillary Clinton Directs Aides to Give Email Server and Thumb Drive to the
            Justice Department",
        "url": "http://www.nytimes.com/2015/08/12/us/politics/hillary-clinton-directs-aides-to-give-
            email-server-and-thumb-drive-to-the-justice-department.html",
        "image": {
                "@id": "1cf5e45097bb2e1f036824790e955cd52e84751d.jpg",
                "caption": "Hillary Rodham Clinton's emails while secretary of state remain a subject
                     of intense scrutiny",
                "url": "http://static01.nyt.com/images/2015/08/12/us/12EMAILS/12EMAILS-master675.jpg"
        },
        "category":[
                "http://en.wikipedia.org/wiki/Category:Technology",
                "http://en.wikipedia.org/wiki/Category:Politics",
                "http://en.wikipedia.org/wiki/Category:International_relations",
                "http://en.wikipedia.org/wiki/Category:Security",
                "http://en.wikipedia.org/wiki/Category:Government_information",
                "http://en.wikipedia.org/wiki/Category:United_States_federal_policy",
                "http://en.wikipedia.org/wiki/Category:Espionage",
                "http://en.wikipedia.org/wiki/Category:Privacy",
                "http://en.wikipedia.org/wiki/Category:Intelligence_(information_gathering)",
                "http://en.wikipedia.org/wiki/Category:Secrecy",
                "http://en.wikipedia.org/wiki/Category:Information_sensitivity",
                "http://en.wikipedia.org/wiki/Category:Politics_of_the_United_States",
                "http://en.wikipedia.org/wiki/Category:National_security",
                "http://en.wikipedia.org/wiki/Category:American_politicians"
        ],
        "about": [
                "http://en.wikipedia.org/wiki/United_States_Department_of_State",
                "http://en.wikipedia.org/wiki/Hillary_Clinton",
                "http://en.wikipedia.org/wiki/Classified_information_in_the_United_States",
                "http://en.wikipedia.org/wiki/Email",
                "http://en.wikipedia.org/wiki/Federal_Bureau_of_Investigation",
                "http://en.wikipedia.org/wiki/Server_(computing)",
                "http://en.wikipedia.org/wiki/Classified_information"
        ]
      }
    ]
  }
}
```