Multimodal Topic Labelling

Ionut Sorodoc[1], **Jey Han Lau**[1,2], **Nikolaos Aletras**[3] **and Timothy Baldwin**[1]

[1]Computing and Information Systems, The University of Melbourne
[2]IBM Research
[3]Amazon.com
{ionutsorodoc,jeyhan.lau,nikos.aletras}@gmail.com
tb@ldwin.net

Abstract

Topics generated by topic models are typically presented as a list of topic terms. Automatic topic labelling is the task of generating a succinct label that summarises the theme or subject of a topic, with the intention of reducing the cognitive load of end-users when interpreting these topics. Traditionally, topic label systems focus on a single label modality, e.g. textual labels. In this work we propose a multimodal approach to topic labelling using a simple feedforward neural network. Given a topic and a candidate image or textual label, our method automatically generates a rating for the label, relative to the topic. Experiments show that this multimodal approach outperforms single-modality topic labelling systems.

1 Introduction

LDA-style topic models (Blei et al., 2003) are a popular approach to document clustering, with the "topics" (in the form of multinominal distributions over words) and topic allocations per document (in the form of a multinomial distribution over the topics) providing a powerful document collection visualisation, gisting and navigational aid (Griffiths et al., 2007; Newman et al., 2010a; Chaney and Blei, 2012; Sievert and Shirley, 2014; Poursabzi-Sangdeh et al., 2016).

Given its internal structure, an obvious way of presenting a topic t is as a ranked list of the highest-probability terms w_i based on $\Pr(w_i|t)$, often simply based on a fixed "cardinality" (i.e. number of topic words) such as 10. However, this has a number of disadvantages: (a) there is a cognitive load in forming an impression of what concept the topic represents from its topic words (Aletras et al., 2014; Aletras et al., 2017); (b) there is a potential bias in presenting the topic based on a fixed cardinality (Lau and Baldwin, 2016); and (c) it can be hard to interpret mixed or incoherent topics (Newman et al., 2010b). Automatic topic labelling methods have been proposed to assist with topic interpretation, e.g. based on text (Lau et al., 2011; Bhatia et al., 2016) or images (Aletras and Stevenson, 2013; Aletras and Mittal, 2017), with recent work showing that the optimal modality (i.e. text or image) for topic labelling varies across topics (Aletras and Mittal, 2017).

The focus of this paper is the automatic rating of a textual or image label for a given topic. Our contributions are as follows:

1. we develop and release a novel topic labelling dataset with manually-scored image and text labels for a diverse set of topics; one particular point of divergence from other text–image datasets is that text and image labels are rated on a common scale, and the optimal modality (text vs. image) for a given topic input must be selected as part of the output; and
2. we propose two deep learning approaches to automatically rate multimodal topic label candidates, which we show to outperform single-modality topic labelling benchmarks.

The code and dataset associated with this paper are available at: `https://github.com/sorodoc/multimodal_topic_label`.

2 Related work

Topic labelling methods usually involve two main steps: (1) the generation of candidate labels (e.g. text or images) for a given topic; and (2) the ranking of candidate labels by relevance to the topic. Textual labels have been sourced from in a number of different ways, including noun chunks from a reference corpus (Mei et al., 2007), Wikipedia ar-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 701–706,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

ticle titles (Lau et al., 2011; Aletras and Stevenson, 2014; Bhatia et al., 2016), or short text summaries (Cano Basave et al., 2014; Wan and Wang, 2016). Images are often selected from Wikipedia or the web based on querying with topic words (Aletras and Stevenson, 2013; Aletras and Mittal, 2017). Recent work on topic labelling has shown that text or image embeddings can improve candidate label generation and ranking (Bhatia et al., 2016; Aletras and Mittal, 2017).

Bhatia et al. (2016) use `word2vec` (Mikolov et al., 2013) and `doc2vec` (Le and Mikolov, 2014) to represent topics and candidate textual labels in the same latent semantic space. The most relevant textual labels for a topic are selected from Wikipedia article titles using the cosine similarity between the topic and article title embeddings. Finally, top labels are re-ranked in a supervised fashion using various features such as the PageRank score of the article in Wikipedia (Brin and Page, 1998), trigram letter ranking (Kou et al., 2015), topic word overlap, and word length of the label.

Aletras and Mittal (2017) use pre-computed dependency-based word embeddings (Levy and Goldberg, 2014) to represent the topics and the caption of the images, as well as image embeddings using the output layer of VGG-net (Simonyan and Zisserman, 2014) pretrained on ImageNet (Deng et al., 2009). A concatenation of these three vectors is the input to a simple deep neural network with four hidden layers and a sigmoid output layer to predict the relevance score.

Textual or visual modalities for labelling topics have been studied extensively, although independently from one another. Our work differs from the single-modality methods described above in that it uses a joint model to predict the continuous-valued rating for both textual and image labels. This is, to the best of our knowledge, the first attempt at joint multimodal topic labelling.

3 Dataset

Several annotated datasets have been developed in previous work for topic labelling, although they have been based on a particular label modality (i.e. text or images). For example, Aletras and Stevenson (2013) used topics generated from New York Times articles and collected image labels with human ratings, while Bhatia et al. (2016) extended the work of Lau et al. (2011) and annotated textual labels for topics generated from four distinct domains. The topics of these different datasets do not overlap, and as such have little utility for our multimodal method. To this end, we develop a new dataset which contains human-assigned ratings for two topic label modalities (textual and image) for the same set of topics.

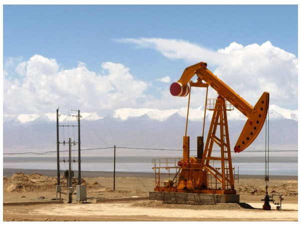
Topic Terms	oil, energy, gas, water, power, fuel, global, price, plant, natural
Image Label	
Mean Rating	2.83
Textual Label	Energy Development
Mean Rating	2.14

Table 1: Example of a topic and its textual and image labels.

We build on the dataset of Bhatia et al. (2016), which has ratings for textual labels. This dataset contains 228 topics generated from 4 different domains: BLOGS, BOOKS, NEWS and PUBMED.[1] Each topic has 19 textual labels which were rated by human judges on a scale of 0–3, where 0 represents a poor label and 3 indicates a perfect label. We chose this dataset due to the diversity of sources represented in the topics.

We use the 228 topics and generate image labels for each topic following the method of Aletras and Stevenson (2013).[2] We follow the annotation approach of Bhatia et al. (2016), collecting ratings based on an ordinal scale of 0–3. We use Amazon Mechanical Turk to crowdsource the ratings, and have each image labelled by 8 workers. To aggregate the ratings for a label, we compute its mean rating.

For quality control, we embedded a bad label into the HIT for each topic by sampling a label candidate for a topic from a different domain, under the assumption that an out-of-domain label is highly unlikely to be appropriate. Workers who rate these control labels greater than 1 are

[1] To clarify, the original topics were generated by Lau et al. (2011); Bhatia et al. (2016) collected human ratings for textual labels on these topics using their methodology.

[2] We use the Bing Search API as our search engine.

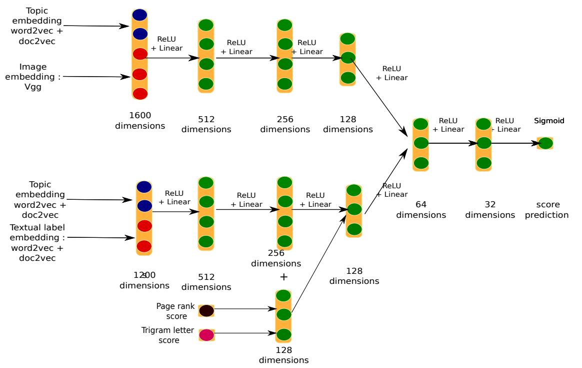


Figure 1: Multimodal model for topic labelling (`joint-NN`)

recorded, and those who fail more than 50% of control labels are filtered out of the dataset.

In total, 353 turkers participated in the image labelling task, at an average error percentage of 16% (based on the control images). A total of 42 turkers were filtered out, on the basis of having an error rate of more than 50%.

An example of a topic and its image and textual labels, and their associated mean ratings, is presented in Table 1. The mean rating for the textual labels is 1.57 with a variance of 0.29, while the mean rating for the image labels is 1.84 with a variance of 0.51. That is, the image labels are, on average, better quality, but there is equally more variability among the image labels.

To summarise, our final dataset consists of 4560 images and 4332 textual labels for 228 topics (20 images and 19 textual labels for each topic). To the best of our knowledge, it is the first dataset which has ratings for two topic label modalities. In addition to benefiting topic labelling research, it has potential applications in other language and vision tasks such as image captioning.

4 Models

Our baseline model (`baseline`) combines the two methodologies of Aletras and Mittal (2017) and Bhatia et al. (2016). That is, we generate and rank textual and image labels based on Bhatia et al. (2016) and Aletras and Mittal (2017) respectively, and then generate a combined ranking based on the predicted ratings.[3] The baseline model views the two modalities (image and textual labelling) as two distinct tasks and does not leverage potential complementarity between them.

Evaluation	`baseline`	`disjoint` **`-NN`**	`joint` **`-NN`**	**Upper** **Bound**
Multimodal	2.07	2.02	**2.08**	2.74
Visual-Only	1.95	1.98	**1.99**	2.67
Textual-Only	**2.01**	1.87	**2.01**	2.48

Table 2: Top-1 average rating performance. Boldface indicates the best performance for each type of evaluation.

We propose a simple feed-forward neural that jointly re-ranks the two topic label modalities (`joint-NN`). In `joint-NN`, we first generate the candidate image labels and textual labels using the methodologies of Aletras and Mittal (2017) and Bhatia et al. (2016), respectively. However, unlike `baseline` where the labels are ranked separately, `joint-NN` feeds both label modalities into a single network to predict their ratings. The network architecture is depicted in Figure 1.

Each input modality is fed into two dense layers that are unconnected. The hidden representation at the 4th layer of the networks is then passed to a joint/shared hidden layer before the final output layer. All connections between layers are dense connections and the final output layer has a sigmoid activation, while all other hidden layers have ReLU activations. The first four layers are kept separate to allow the network to transform the embeddings from the two different modalities to a common hidden representation. The shared layers leverage potential complementarity between the two label modalities to predict the final label rating.

We generate the textual labels following the label generation methodology of Bhatia et al. (2016), as part of which, the labels and topic terms each have representations based on `doc2vec` and `word2vec` embeddings, respectively. We concatenate all four embeddings and use them as the input for the network.[4] Bhatia et al. (2016) found that letter trigram features and PageRank features were strong features when re-ranking the labels. We borrow this idea, and incorporate these two features into the network by mapping the 2-

[3] Bhatia et al. (2016) originally used SVR to rank textual labels. We re-ran their model using the same features and SVR to predict label ratings, allowing us to combine both textual and image labels and rank them using their predicted ratings.

[4] Each type of embedding has 300 dimensions; the concatenated input thus has $4 \times 300 = 1200$ dimensions.

Topic Terms	food, eat, cook, chicken, recipe, cup, cheese, add, taste, tomato	drive, computer, card, laptop, memory, battery, usb, intel, processor, hard
Image Label		
Predicted Rating	2.53	1.87
Textual Label	Cooking	Desktop Computer
Predicted Rating	1.98	2.20

Table 3: Example of two topics and their generated textual and image labels and predicted ratings.

dimensional input (representing the letter trigram and PageRank features) into a 128-dimension vector and concatenating it with the 256-dimension hidden representation at the third layer (thus yielding a 384-dimension vector).[5]

For the visual labels, the topic terms use the same `doc2vec` and `word2vec` embeddings. For the image labels, we use the representation of the last layer of the VGG Neural Network (Simonyan and Zisserman, 2014). As before, the vectors for the topic terms and image labels are concatenated and fed as input to the network.[6]

As a control to test whether the sharing of weights helps with the prediction of label ratings, we experiment with another network (`disjoint-NN`) that has the same architecture as `joint-NN`, except that the final few layers are not shared and the two networks are trained independently.

5 Experiments and results

Following standard practice in topic labelling evaluation (Lau et al., 2011; Aletras and Stevenson, 2013; Bhatia et al., 2016), we use "top-1 average rating" as the evaluation metric. It computes the mean rating of the top-ranked label generated by the system, and provides an assessment of the absolute utility of the labels. For example, if the top-ranked label predicted by the system has an average rating of 3.0, that means the system are generating perfect topic labels.[7]

We present the results of all systems (`baseline`, `joint-NN` and `disjoint-NN`) in Table 2. Each model is trained using 10-fold cross-validation for 10 epochs. Presented results are an average over 20 runs. We display three types of evaluation: (1) "multimodal", where we pool both label modalities together and evaluate jointly; (2) "visual-only", where we evaluate only the visual labels; and (3) "textual-only", where we evaluate only the textual labels. In addition to the 3 systems, for each topic we determine the rating of the best label and compute its mean over all topics, as the upper bound for the task (labelled "upper bound").

Encouragingly, `joint-NN` — which exploits information from both input modalities — achieves the best performance. The improvement compared to `disjoint-NN` is substantial, and much of the improvement is in the textual labels. However, when compared to `baseline`, most of the gain is in the visual labels. These observations seem a little unintuitive; to better understand them we first look at `baseline` and `disjoint-NN`.

In terms of methodology, the difference between `baseline` and `disjoint-NN` is their rerankers. Both the image label re-rankers of `baseline` and `disjoint-NN` are driven by neural networks, but the re-ranker of `disjoint-NN` has an additional layer (5 vs. 4).[8] The improvement of results for the visual labels could thus be attributed to the additional hidden layer.

On the other hand, the performance difference for the textual labels between `baseline`

[5] We explored incorporating the additional features at different layers, but saw little difference in task performance.

[6] VGG vectors have 1000 dimensions and the `doc2vec` and `word2vec` embeddings each have 300 dimensions; the concatenated input is thus a 1600-dimension vector.

[7] Note that, unlike previous work, we don't evaluate based on nDCG as the candidate set and ratings for each of the individual label modalities and the combined labels differ, meaning that the nDCG numbers are not directly comparable.

[8] The re-ranker of `disjoint-NN` also does not use caption embedding, as it proves to be redundant. All other features are the same.

and disjoint-NN is attributed to the classifiers (baseline = SVR; disjoint-NN = neural network), since they both share the same features. These results suggest that SVR is the superior classifier in this case.

However, when we share the latent representations for the last few layers (joint-NN), we see that results improve substantially. In particularly, textual label performance is on par with baseline, suggesting the addition of image label data helps learn the latent representations of textual labels. As a whole, this suggests there is strong complementarity between the two different modalities of labels and highlights the strength of a multimodal network.

Lastly, it is worth mentioning that the multimodal evaluation yields the highest rating across all systems. This suggests that, consistent with the findings of Aletras and Mittal (2017), different topics may have different optimal label representations (image or textual), and that the best performance is achieved when we allow the model to dynamically select between modalities. We present a sample of generated textual and image label for a topic in Table 3.

Looking at the upper bound, we see there is considerable room for further improvement. The models we have experimented with are based on simple feed-forward architectures, and the input representation is pre-computed, and thus not updated in the network. An immediate direction for future work would be designing end-to-end architectures that take the input as raw features (e.g. using the image pixels for the image labels).

6 Conclusions

In this paper, we have proposed a multimodal approach to automatic topic labelling, based on a deep neural network. Compared to benchmark systems, our joint model achieves the best performance, demonstrating the strength of modelling different label modalities jointly.

Another contribution of the paper is the development of a multimodal dataset which we have released publicly. The dataset, which contains annotations for image and textual labels, could have applications for other multimodal NLP tasks.

Acknowledgements

This research was supported in part by the European Union, through the LCT Masters Erasmus Mundus programme (https://lct-master.org/), and in part by the Australian Research Council.

References

Nikolaos Aletras and Arpit Mittal. 2017. Labeling topics with images using neural networks. In *Proceedings of the 39th European Conference on Information Retrieval (ECIR 2017)*, Aberdeen, UK.

Nikolaos Aletras and Mark Stevenson. 2013. Representing topics using images. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 158–167, Atlanta, USA.

Nikolaos Aletras and Mark Stevenson. 2014. Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 631–636, Baltimore, USA.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2014. Representing topics labels for exploring digital libraries. In *Proceedings of Digital Libraries 2014*, London, UK.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2017. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 953–963, Osaka, Japan.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Sergei Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh World Wide Web Conference*, pages 107–117, Brisbane, Australia.

Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. 2014. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, Baltimore, USA.

Allison June-Barlow Chaney and David M. Blei. 2012. Visualizing topic models. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012)*, pages 419–422, Dublin, Ireland.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255, Miami, USA.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Wanqiu Kou, Li Fang, and Timothy Baldwin. 2015. Automatic labelling of topic models using word vectors and letter trigram vectors. In *Proceedings of the 11th Asian Information Retrieval Societies Conference (AIRS 2015)*, pages 229–240, Brisbane, Australia.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2016)*, pages 483–487, San Diego, USA.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545, Portland, USA.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1188–1196, Beijing, China.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, USA.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499, San Jose, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS-13)*, pages 3111–3119, Lake Tahoe, USA.

David Newman, Timothy Baldwin, Lawrence Cavedon, Sarvnaz Karimi, David Martinez, and Justin Zobel. 2010a. Visualizing document collections and search results using topic mapping. *Journal of Web Semantics*, 8(2–3):169–175.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010b. Evaluating topic models for digital libraries. In *Proceedings of the 2010 Joint Conference on Digital Libraries (JCDL)*, pages 215–224, Gold Coast, Australia.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. Alto: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany.

Carson Sievert and Kenneth E. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the ACL 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, USA.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Xiaojun Wan and Tianming Wang. 2016. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305, Berlin, Germany.