Comparative Summarisation of Rich Media Collections

Umanga Bista
The Australian National University
Canberra, Australia
umanga.bista@anu.edu.au

ABSTRACT

The goal of this thesis is to develop techniques for *comparative* summarisation of *multimodal* document collections. Comparative summarisation is extractive summarisation in *comparative settings*, where documents form two or more groups, e.g. articles on the same topic but from different sources. Comparative summarisation involves, not only, selecting representative and diverse samples within groups, but also samples that highlight commonalities and differences between the groups. We posit that comparative summarisation is a fruitful problem for diverse use cases, such as comparing content over time, authors, or distinct view points. We formulate the problem of comparative summarisation by reducing it to binary classification problem and define objectives to incorporate representativeness, diversity and comparativeness. We design new automatic and crowd-sourced evaluation protocols for summarisation evaluation that scales much better than the evaluations requiring manually created ground truth summaries. We show the efficacy of the approach in a newly curated datasets of controversial news topics. We plan to develop new collection comparison methods for multimodal document collections.

KEYWORDS

Comparative Summarisation, Multimodal Data

1 INTRODUCTION

Document summarisation has been one of the core problems for tackling information overload. Summarisation can be either extractive, where only parts of existing documents are used to create summaries that achieve coverage and diversity, or abstractive, where new content is created. In general the former is simpler and tends to cause fewer grammatical and semantic mistakes than abstractive summarisation. A large range of different models have been applied to extractive summarisation, including methods incorporating diversity measures from information-retrieval [4], structured SVM regularized by constraints for diversity, coverage, and balance [11], hierarchical topic models [6] and optimising discrete submodular functions [12, 18].

In this research, we consider *comparative* summarisation: given *groups* of document collections, construct summaries for each group. For instance, given thousands of news articles per month on a certain topic, e.g. climate change, guncontrol or beefban, groups can be formed by publication time, by source, or by political leaning. *Comparative* summaries aim to highlight the similarities and differences *between* groups and can help answer user questions such as: what is new on climate change this week? what is the same? or even what is different between the coverage in NYTimes and BBC? The existing literatures on comparative summarisation focuses on differences between topics and is limited to selecting sentences [17] or aims to compare differences in similar concepts across documents within a topic [8]. Our focus is more general: selecting documents highlighting differences and similarities between groups in multimodal collections where the groups can be topics, time ranges, etc.

In our recent work [3], we formulate comparative summarisation for differences from a perspective of competing binary classifiers and propose a set of objectives in this formulation which can be optimised using discrete and gradient based optimisation. We show the application of our proposed methods in a controversial news dataset that we curated by collecting tweets, news articles and their images around few controversial topics such as Gun Control, Climate Change, Beefban, etc. We show the efficacy of our model using a scalable evaluation that does not require manually created ground truth summaries. We make classification performance as evaluation for summarisation and our evaluation is verified by a human pilot study.

More specifically, the planned main contributions of this thesis are:

(1) **Problem Definition:** We define the new problem of comparative summarisation of multimodal document collections.
(2) **New approaches for Comparative Summarisation:** We formulate objectives for comparative summarisation from a binary classification perspective that incorporate representativeness, diversity, and discrimination criteria.
(3) **Application:** We apply comparative summarisation to the important task of summarising controversial news topics from different aspects such as source, stance, over time and geography.
(4) **New automatic and crowd sourced evaluation methods for summarisation:** We propose a scalable framework for evaluating comparative summarisation of document collections, which doesn't require manually created ground truth summaries and is backed by a human pilot study.

As many real world datasets consist of multiple modalities such as texts and images, we consider multimodal summarisation for future work where we will summarise multiple modalities jointly. Having images in the summaries will make the results visually appealing and more accurately convey information. While there is some literature on multimodal analysis such as cross modal retrieval [14] and multimodal feature learning [1, 7, 10], comparative summarisation for multimodal collections remains unexplored.

We intend to extend our current works to full comparative summarisation including discrimination and similarity criteria. We also intend to extend it to other comparative and data domains including entity-relation graphs and multimodal datasets.

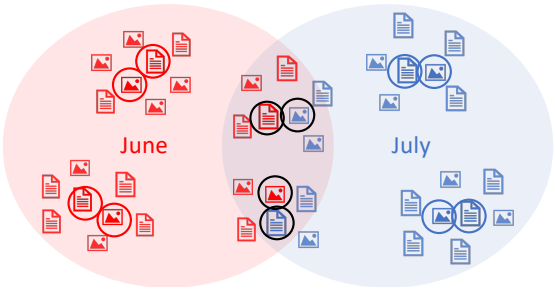

Figure 1: Comparative Summarisation for multimodal data (text + images). Articles and images created in June are within the red shaded region while those created in July are in the blue shaded region. Summary prototypes of each months are circled by respective colors, while summary prototypes from articles and images similar across two months are highlighted by black circles.

2 APPROACH

2.1 Comparative Summarisation

Formally, the comparative summarisation problem is defined on G groups of documents $\{\mathbf{X}_1, \ldots, \mathbf{X}_G\}$, where a group may, for example, correspond to a particular time range. We write the document collection for group g as $\mathbf{X}_g = \{\mathbf{x}_{g,1}, \mathbf{x}_{g,2}, \ldots, \mathbf{x}_{g,N_g}\}$, where N_g is the total number of documents in group g. We represent individual documents as vectors $\mathbf{x}_{g,i} \in \mathbb{R}^d$ (e.g. using averaged word-vectors [13]). Our goal is to summarise each document collection $\mathbf{X}_g$ with a set of *prototypes* $\bar{\mathbf{X}}_g \subset \mathbf{X}_g$ that discriminate from each other, written as $\bar{\mathbf{X}}_g = \{\bar{\mathbf{x}}_{g,1}, \bar{\mathbf{x}}_{g,2}, \ldots, \bar{\mathbf{x}}_{g,M}\}$. For simplicity, we assume the number of prototypes M is the same for each group.

2.1.1 Comparative Summarisation as Binary Classification. For comparative summarisation of the differences, we can think of prototype selection in terms of two competing binary classification objectives: one distinguishing $\bar{\mathbf{X}}_g$ from $\mathbf{X}_g$, and another distinguishing $\bar{\mathbf{X}}_g$ from $\mathbf{X}_{\neg g}$ (documents from groups other than g). In abstract, this suggests a multi-objective optimisation problem of the form,

$$\max_{\bar{\mathbf{X}}_1,\ldots,\bar{\mathbf{X}}_G} \left(\sum_{g=1}^{G} -\mathrm{Acc}(\bar{\mathbf{X}}_g, \mathbf{X}_g), \sum_{g=1}^{G} \mathrm{Acc}(\bar{\mathbf{X}}_g, \mathbf{X}_{\neg g}) \right) \tag{1}$$

where $\mathrm{Acc}(\cdot,\cdot)$ is an accuracy term.

2.1.2 Prototype Selection via Nearest Neighbor. We approximate the accuracy term $\mathrm{Acc}(\cdot,\cdot)$ in Eq 1 with nearest-neighbour classifiers. We adopt a formulation by [18] that maximises the total similarity of every point to its nearest prototype from the same class.

$$\mathcal{U}_{nn}(\bar{\mathbf{X}}) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \max_{m=1:M} k(\bar{\mathbf{x}}_{g,m}, \mathbf{x}_{g,i}) \tag{2}$$

Here k is any similarity function. The nearest neighbour utility function is simple and intuitive, however it only considers representativeness and diversity but misses discriminative criteria.

2.1.3 Prototype selection by Maximum Mean Discrepancy (MMD). *MMD* [5] measures the distance between two distributions $\mathbf{X}$ and $\mathbf{Y}$ and is defined as distance between the expected values of the two distributions in a Reproducing Kernel Hilbert Space (RKHS). Using the kernel trick, we can write $MMD^2(\cdot,\cdot)$ without requiring explicit feature mapping.

$$\mathrm{MMD}^2(\mathbf{X},\mathbf{Y}) = \mathbb{E}_{\mathbf{x},\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')] - 2\cdot\mathbb{E}_{\mathbf{x},\mathbf{y}}[k(\mathbf{x},\mathbf{y})] + \mathbb{E}_{\mathbf{y},\mathbf{y}'}[k(\mathbf{y},\mathbf{y}')] \tag{3}$$

One can think of MMD as implicitly computing a (kernelised) *nearest centroid classifier* to distinguish between $\mathbf{X}$ and $\mathbf{Y}$: MMD is small when this classifier has high expected error. Thus, MMD can be seen as an efficient approximation to classification accuracy $\mathrm{Acc}(\cdot,\cdot)$. This intuition lead to a practical utility function that approximates Equation 1 by taking the difference of two MMD terms:

$$\mathcal{U}_{diff}(\bar{\mathbf{X}}) = \sum_{g} (-\mathrm{MMD}^2(\bar{\mathbf{X}}_g, \mathbf{X}_g) + \lambda \cdot \mathrm{MMD}^2(\bar{\mathbf{X}}_g, \mathbf{X}_{\neg g})) \tag{4}$$

The hyper-parameter λ trades off how well the prototype represents its group, against how well it distinguishes between groups. There are other variants of MMD based comparative summarisation, which are explained in our recent work [3].

2.1.4 Optimising Utility Functions. NN (2) and MMD objectives (4) are discrete optimisation problem, which are NP-Hard. We optimise all objectives greedily and MMD objectives using gradient based optimisation methods. In gradient optimisation, we allow prototypes to come from dataspace rather than restricting to come from dataset. The selected prototypes are then snapped to the nearest document in the group under euclidean distance. The details are explained in our recent work [3].

2.2 Multimodal Summarisation

Given two modalities with some correspondence between the items, e.g. text and images, video and speech, english and German text etc. In multimodal representation learning, the goal is to learn the joint space, in which corresponding data points from different modalities lie close to each other. Approaches include encoder-decoder based model for pairwise ranking [10], Canonical Correlation Analysis(CCA) which is based on normalized correlation [7] and their kernel [7] and deep neural network [1] variants. Once we represent our images and text in a joint space, we could do cross modal retrieval [14] using summaries in either modality or jointly (§2.1) to form coherent summaries across modalities.

3 DATASET AND EVALUATION

3.1 Dataset

We curate an initial list of 10 topics in June 2017 that satisfy the criteria of having non-trivial news coverage and being controversial. The topics are *Beef Ban*, *Capital Punishment*, *Gun Control*, *Climate change*, *Illegal immigration*, *Refugees*, *Gay marriage*, *Animal testing*, *Cyclists on road* and *Marijuana*. We want to focus on controversial topics since they are likely to be discussed in the future as their coverage lasts for a long time. Controversy is an important topic for research in social media and online political discourse, and is also important in real-world applications . To obtain various opinions on contemporary social problems, we choose Twitter as a source with hashtags as query and use hashtag expansion to obtain other similar hashtags [16] . We obtain embedded news articles from Twitter, extract knowledge base entities [2] from the text and also store the their creation timestamp. We also obtain the images from tweets and news articles, store their features and objects obtained

from Convolutional Neural Networks [15] . In our recent paper [3] we used a subset of dataset from 3 topics: Beefban, Capital punishment and Guncontrol with about 15000 articles that span across 14 months altogether.

3.2 Evaluation

3.2.1 Automatic Evaluation. We evaluate the comparative summarisation of differences via classification performance, i.e. we train a classifier such as Support Vector Machine (SVM) or 1-Nearest Neighbor (1NN) from summary prototypes and evaluate the balanced accuracy on a test set. A good summary should be representative of the entire data and hence help the classifier in identifying the group of each test data-point. This framework for evaluating summaries is scalable and does not require the manually created ground truth summaries.

3.2.2 Human Evaluation. We also run a pilot study to demonstrate the efficacy of classification as summarisation evaluation. We present crowd workers with few summary articles from two groups and ask them to classify the test articles into either group. We then evaluate human performance qualitatively and quantitatively, which show that our automatic evaluation framework is a good surrogate for evaluation of comparative summarisation.

4 RESULTS

We now discuss some of the results towards comparative summarisation of differences. We evaluate on a subset of the dataset collected in (§3.1). For this evaluation we use news articles only use news articles which we filtered for spam and deduplicated. We only use the title and first three sentences of news articles and represent them using averaged GLOVE [13].

We compare NN (§2.1.2), MMD (§2.1.3) objectives with baselines including kmeans, mmd-critic [9]. We measure classification performance by SVM and 1NN (§3.2.1). For three news topics (§3.1) with 2 classifier evaluation (SVM and 1NN) and for 4 different settings of the number of prototypes (2,4,8 and 16 per class), there were 24 evaluations settings for automatic evaluation. We found out that gradient based MMD objectives perform best in 15 evaluations, followed by NN in 4, greedy MMD in 3 and k-means in 2 evaluations. In human evaluations, we ask 3 people to classify 6 test articles for each of 21 summaries generated from same dataset using four methods. Human accuracy was 71% on summaries obtained from gradient based MMD, outperforming other methods by at least 7%. The results of human pilot study have moderate inter-annotator agreement and are statistically significant with $p < 0.05$ using paired two tailed t-tests over all 378 judgements. Details of evaluation settings and results are reported in [3].

5 DISCUSSION

In summary, in our recent work[3], we have formulated the problem comparative summarisation for differences of document collections and show an effective large scale evaluation without requiring manually created ground truth summaries. There could be several future work extensions for this work.

One direction for future work lies in extending our comparative summarisation approaches to identify similarities between document collections in addition to identifying differences which the current work focuses on. Another future work lies in formulating and solving the multimodal summarisation. Whether we can use existing approaches in our problem (§2.2) or if they need new formulation is yet to be explored. We also need a scalable evaluation framework for multimodal summarisation that is good surrogate of evaluation in the context of this research problem. Another dimension for future work could be from an application perspective, such as doing comparative summarisation over time, or different users or sources of data from different data domains, or focusing on different input space such as text, images or entities-relation graphs. Out of these vast number of possibilities, this research aims to focus on the problems and approaches that provide maximal benefit to the multimodal collections summarisation.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer.

[3] Umanga Bista, Alex Mathews, Minjeong Shin, Aditya Menon, and Lexing Xie. 2019. Comparative summarisation of Document Collections. In *AAAI*.

[4] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries *(SIGIR)*. ACM.

[5] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. (2012).

[6] Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization *(NAACL)*.

[7] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. (2004).

[8] Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. 2011. Comparative news summarization using linear programming *(HLT)*. ACL.

[9] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability *(NIPS)*.

[10] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. abs/1411.2539 (2014).

[11] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. 2009. Enhancing Diversity, Coverage and Balance for Summarization Through Structure Learning *(WWW)*. ACM.

[12] Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization *(HLT)*. ACL.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation *(EMNLP)*.

[14] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval *(MM)*. ACM.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

[16] John-Paul Verkamp and Minaxi Gupta. 2013. Five Incidents, One Theme: Twitter Spam as a Weapon to Drown Voices of Protest.

[17] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2012. Comparative document summarization via discriminative sentence selection. (2012).

[18] Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning *(ICML)*.