ReadAlong: Reading Articles and Comments Together

Dyut Kumar Sil
Dept CSA, IISc, Bangalore
dyutkumar@csa.iisc.ernet.in

Srinivasan H Sengamedu
Yahoo! Labs, Bangalore
shs@yahoo-inc.com

Chiranjib Bhattacharyya
Dept CSA, IISc, Bangalore
chiru@csa.iisc.ernet.in

ABSTRACT

We propose a new paradigm for displaying comments: showing comments alongside parts of the article they correspond to. We evaluate the effectiveness of various approaches for this task and show that a combination of bag of words and topic models performs the best.

Categories and Subject Descriptors

H.3.m [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL—*Miscellaneous*

General Terms

Algorithms

Keywords

Comments, Comment Alignment, Topic Models

1. INTRODUCTION

Comments are the primary form of user interaction in several sites. Consider, for example, a news site which supports commenting: users can leave comments on various aspects of the news articles. Comments increase user engagement in multiple ways: (1) Commenters can share information/insights with other users. (2) Readers get additional information/perspectives from comments. Users can also rate comments with "Thumbs Up/Down". Comments are usually ranked by time or rating. Popular/Controversial topics attract lots of comments. Hence it is difficult for users to keep track of and assimilate the information in the comments.

Most of the comments are short and are about specific topics discussed in the article. When reading comments, the readers have to mentally map each comment to the parts of the article they discuss. For long articles or for articles with several comments, this requires significant cognitive effort – especially when the reader is not familiar with the topic of the article.

In this paper, we propose a new paradigm for displaying comments: showing comments alongside parts of the article they correspond to. Figure 1(a) shows an example movie review with comments. The comments are on different aspects of the movie. Figure 1(b) shows the review after comments are aligned. The parts of the review which have comments are highlighted with a symbol indicating the presence of comments. Mousing over the symbol shows the associated comments.

To enable this, we need to identify for each comment, those parts of the article which map to the comment. The mapping can be done at the lexical level or at the topic level. For robust topic detection, we augment each article with a few related articles. We show that combining topic representation (topics derived from an enriched corpus) and lexical representation (the classical bag of words representation) provides the best results.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. CHALLENGES

We formally define the problem first.

Problem Statement: Let $D = \{A_i, C_i\}_{i=1}^{M}$ be the dataset of M articles with the associated comments: here A_i is an article and $C_i = \{c_{i1}, c_{i2}, \cdots, c_{in_i}\}$ is the collection of comments for article A_i. For the article A_i, we first identify the set $S_i = \{s_{i1}, s_{i2}, \cdots, s_{im_i}\}$ of topical segments it contains and for each comment $c_{ij} \in C_i$, we then find the article segment $s_{ik} \in S_i$ it corresponds to. □

Comments are usually very brief and heavily leverage the article context. Hence it is very hard to perform the mapping using the words alone because the article and the comment can use different words to discuss the same topic. In Figure 1, the article uses "martial arts master" while the comment makes the clarification "teaches Kung Fu". Because of the vocabulary mismatch, it is ideal to map comments and the article at a topic level. Since the comment and article text are usually short, performing topic modeling on a small corpus does not lead to reliable results. Hence we enhance each article with several related articles before performing topic modeling. This leads to robust topics and hence better alignment.

3. APPROACH

There are three phases to the solution.

Topic segmentation: We topically segment each article and comment. For simplicity, in this paper, we treat each paragraph in the article as well as comment as a single topical segment. In addition, the title of the article as well as the entire article text also constitute article segments. Comments usually have only one topic segment.

Segment representation: The feature description for each segment is discussed in Section 3.1.

Matching: We use cosine similarity between features for matching. For each comment, we calculate the cosine similarity with each segment of the corresponding article. The

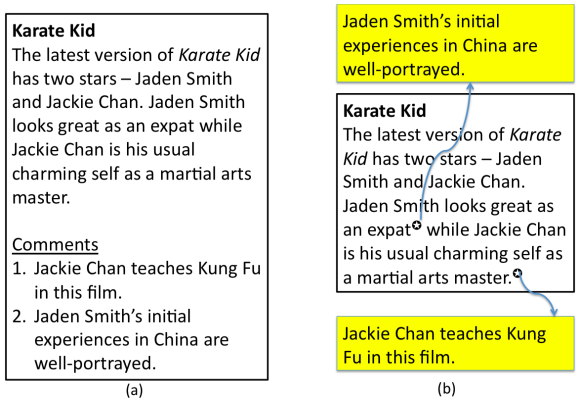

Figure 1: Comment Alignment. (a) Article with comments (b) Article with aligned comments.

segment with the maximum similarity is chosen as the matching segment.

3.1 Representation

As mentioned in the Introduction, the segments are best represented using topics because of the vocabulary mismatch between article and comments. We use Latent Dirichlet Allocation (LDA) [1] to perform topic analysis. In **LDA**, a topic is a probability distribution over words and each document is represented as a mixture θ over topics. When performing LDA, we treat each segment as a document. For an article, topic analysis is performed on the segments of the article and the associated comments. To address the sparseness of the data, we enhance each article with a set of related articles. In this case. LDA is performed on the segments of an article, its comments, and the related articles. The feature vector for each segment is the vector of topic weights, θ.

We also experiment with two baseline feature representations: bag of words (BOW) and semi-supervised PLSA.

In **BOW**, each segment is represented as a vector of IDFs of words it contains – the words being obtained after stemming and stop word removal. Let $\mathbf{v}$ be the feature vector representation for BOW.

SS-PLSA – a semi-supervised generative model based on PLSA – was proposed in [2] for aligning user reviews and blogs with expert reviews. Supervision is provided by steering the learned generative model towards the topics in the expert review through conjugate priors obtained using the expert review. In case of comments, we calculate a model, θ_{ik}, for each article segment $s_{ik} \in S_i$ and compute the probability $P(c_{ij}|\theta_{ik})$ for each comment based on the models. The segment with maximum probability is assigned to the comment. Since the segments of an article are used as "model topics", the enriched corpus does not provide any additional value. Hence we do not use SS-LDA on the enriched corpus.

Finally, we combine both BOW and topic weights. The topic weights are given a global weight α. The feature representation is $[(1-\alpha)\mathbf{v} \;\; \alpha\theta]$. We call this **LDA+BOW**. $\alpha = 0.4$ provided the best results in our experiments.

4. EXPERIMENTAL RESULTS

Dataset: We created a dataset D by collecting 40 news articles along with ≈10 comments for each of the 40 articles from `http://news.yahoo.com`. We also created another dataset $D_{enriched}$ with same articles and comments as of D but each of the articles enriched with additional (4–8) related articles. The related articles for an original article have been found by Google news search (`http://news.google.com`) with title of the original article as search key. Each article had an average of 351.1 words (after stemming and stop word removal), 17.5 segments, and 8.4 comments. The average length of comments is 27.1 words. In $D_{enriched}$, an average of 1316.5 words and 58.3 segments were added per article.

Method	D	$D_{enriched}$
LDA	0.263	0.420
LDA+BOW	0.483	**0.646**
BOW	**0.527**	-
SS-PLSA	0.214	-

Table 1: RI for different methods

We created ground truth for 336 comments of 40 articles. We have experimented with both the datasets D and $D_{enriched}$ with different methods discussed in Section 3.1.

Metric: For an article and comment-set pair (A_i, C_i), let $y_{ij} \subseteq S_i$ be the set of true related article-segments (found by human inspection) for comment c_{ij}.

If $|y_{ij}| > 1$, then c_{ij} has multiple related article-segments or if $|y_{ij}| = 0$, c_{ij} has no related article-segment.

Let r_{ij} be the retrieved result for comment c_{ij}. We consider this to be correct if $r_{ij} \in y_{ij}$.

The *Retrieval Index* is defined as:

$$RI = \frac{|\bigcup_{i=1}^{M}\{c_{ij} \in C_i : r_{ij} \in y_{ij}\}|}{|\bigcup_{i=1}^{M} C_i|}$$

Results: Table 1 shows RI for the different techniques.
(1) It can be seen that BOW technique performs surprisingly well – around 53% of the comments are correctly aligned and is the second best.
(2) Topic models (SS-PLSA and LDA) do not perform very well and LDA performs marginally better compared to SS-PLSA on the original corpus.
(3) Enriched corpus helps LDA improve performance by 60%.
(4) The best performance is achieved by combining LDA with BOW on enriched corpus – 64.6% of the comments are matched correctly. LDA+BOW improves LDA by 54% and BOW by 23%.

5. CONCLUSIONS

In this paper, we have proposed the new problem of aligning comments to relevant parts of the article to reduce the readers' cognitive burden. We argued for matching comments and article segments at the topic level. To overcome the limitations of sparse data, we proposed corpus enrichment. The representation enhancing topics learned on enriched corpus with bag of words performs best on our test dataset. In follow-up work, we plan to use generic knowledge sources like Wikipedia in addition to related news article for enrichment.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *JMLR*, 2003.
[2] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *WWW*, 2008.