Video Episode Boundary Detection with Joint Episode-Topic Model

Shunyao Wang, Ye Tian, Ruidong Wang, Yang Du, Han Yan,

Ruilin Yang and Jian Ma

State Key Lab of Networking and Switching Technology,

Beijing University of Posts and Telecommunications, Beijing, China, 100876

Motivation and Challenges

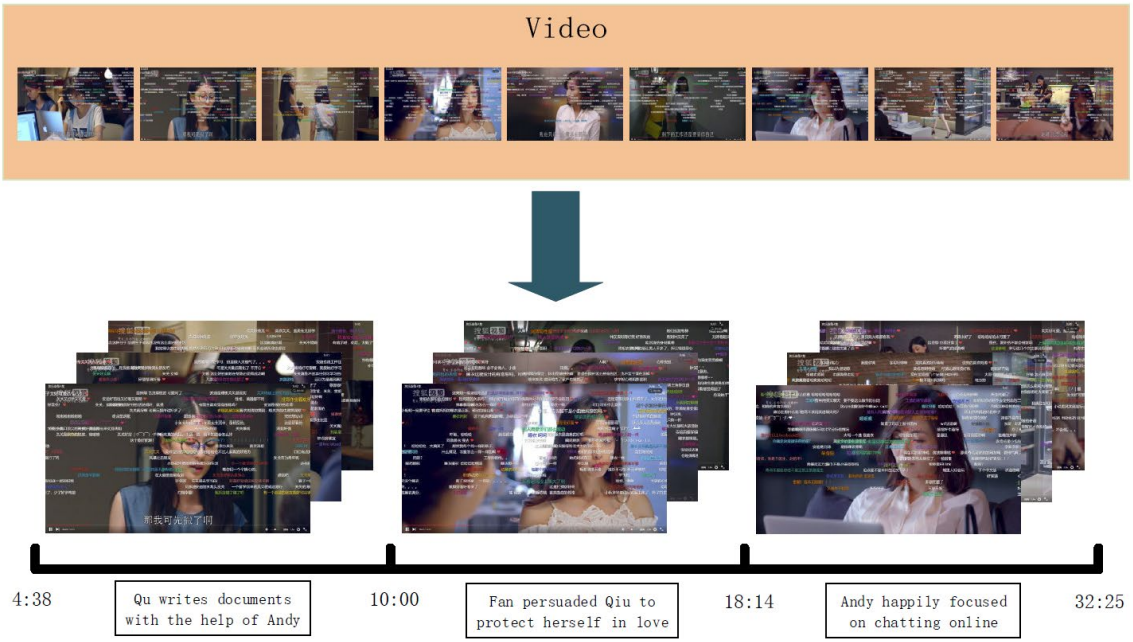
- Most people are used to quickly navigate and locate his concerned video clip according to its corresponding video labels.
- For website operators, manually dividing the episode boundary is a time-consuming and laborious task.
- Traditional scene segmentation algorithms are mostly based on the analysis of frames, which cannot automatically generate labels.
- In recent years, time-synchronized review videos have become increasingly popular, especially in Asia, which provides a new idea for episode boundary detection.

Related Works

- Video or sentence boundary detection
  - Hybrid Keypoint Detection (HKD)
  - Scene Transition Graph (STG)
  - Tilburg Memory Based Learner (TiMBL)
- Video tagging directly by processing time-synchronized reviews
  - Combining user preference with the preceding shots to extract topics
  - Constructing a semantic association graph to filter noisy comments
- Short text modeling
  - Biterm Topic Model (BTM)
  - Common Semantics Topic Model (CSTM)
  - Semantics-assisted Non-negative Matrix Factorization (SeaNMF)

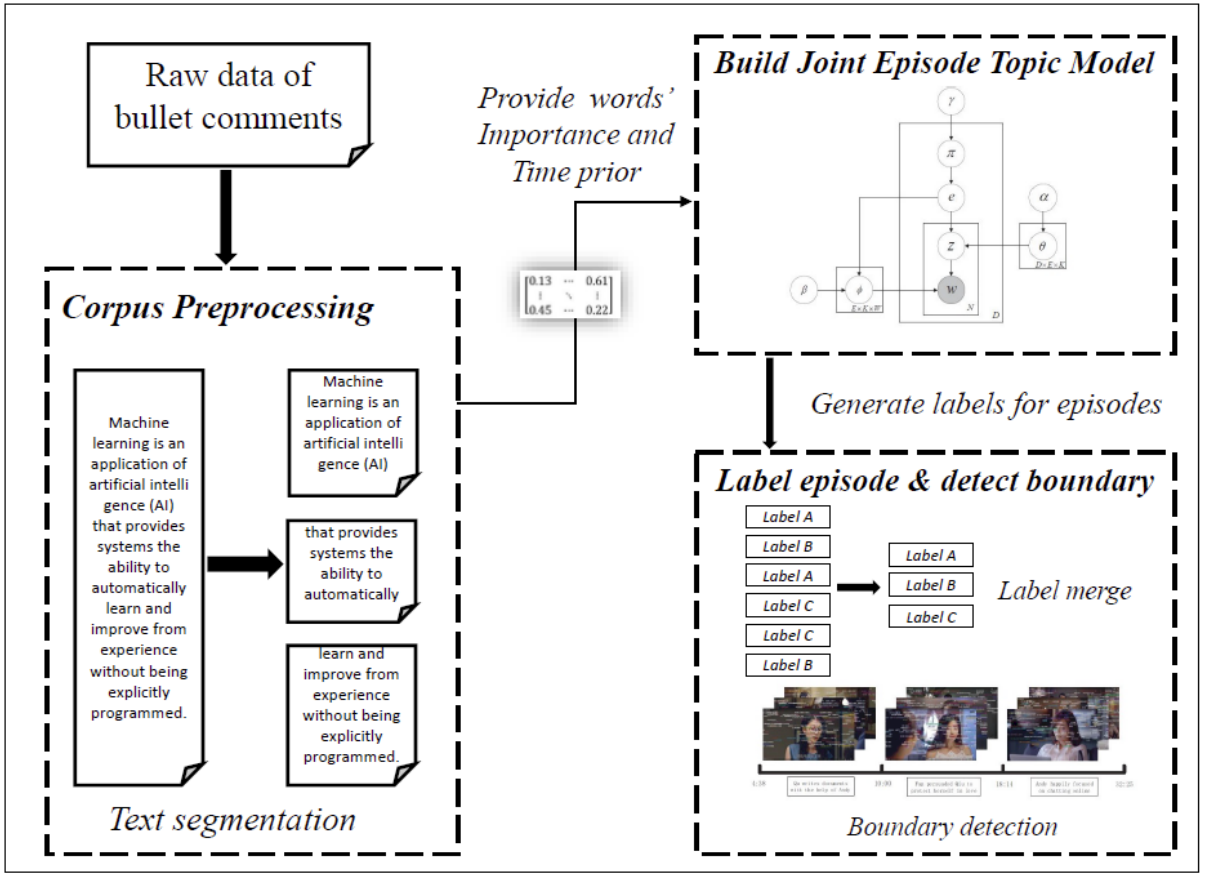
Video Episode Boundary Detection Model （VEBD）

- When the video is playing, the comments are swiped across screen like bullets, so they are also known as bullet screen comment(s) .
- By processing these bullet screen comments filled with screens, VEBD model can divide a complete video into segments based on the correct episodes.



Video Episode Boundary Detection Model （VEBD）

- VEBD model consists of three layers, corpus preprocessing, Gibbs sampling and episode label merging.



Video Episode Boundary Detection Model （VEBD）

- First, this model segments words and removes stop words from the original bullet screen comments file, and divides the file into time slices of equal duration. At the same time, the distributions of time prior and importance prior will be obtained.

- Then, The above data will be entered into the specially designed topic model. After Gibbs sampling, the multinomial distribution of episode, topic and word will be obtained.

- Finally, VEBD model generates episode tags, and merges the similar time slices to get the correct episode boundary division based on temporal similarity and semantic similarity.

Corpus Preprocessing

- The time prior calculation formula of word:

$$\overrightarrow{\lambda_i} = \left\{ e^{\frac{|w_i^1|}{\sum_{n=1} |t_i^n - ep_1|} \over -\xi}, e^{\frac{|w_i^2|}{\sum_{n=1} |t_i^n - ep_2|} \over -\xi}, \cdots, e^{\frac{|w_i^{|E|}|}{\sum_{n=1} |t_i^n - ep_{|E|}|} \over -\xi} \right\}$$

- Importance prior refers to calculating the importance of each word according to TextRank keyword extraction algorithm.

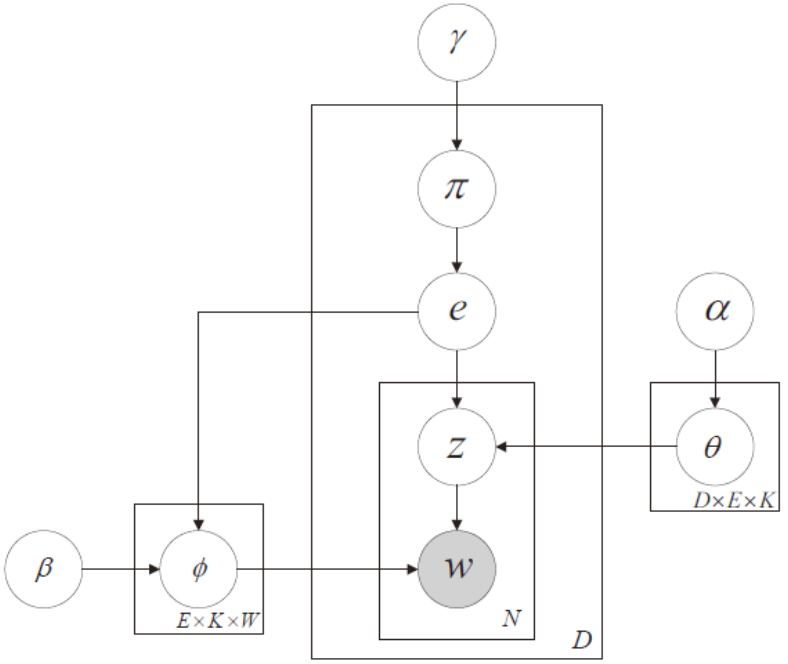
$$r(v_i) = (1-\delta) + \delta \sum_{(j,i) \in edge(v_i)} \frac{w_{ji}}{\sum_{v_k \in edge(v_j)} w_{jk}} r(v_j)$$

$$\sigma_{ij} = \frac{r_{i,j} + \zeta}{\sum_{d \in D} r_{i,d} + |D| \times \zeta}$$

Joint Episode-Topic Model

- JET model is an important part of VEBD model.

- The document layer contains multiple episodes, and topic layer is associated with document and episode layers, and the word layer is associated with episode and topic layers. The episode layer is the connecting bridge and plays an important role in the four layers.



Episode Label Merging

- Select the top k_t topics with the highest probability and select the top k_w words with the highest probability to generate the episode label.

Algorithm Episode label merging

Input: Tags generated after sampling;
Output: Episode label for the video;

$NewTag_1 = RawTag_1$;
for $i = 2$ to E **do**
    if $CurTag_i$ is similar to the current $NewTag_{i-1}$ **then**
        Merge $CurTag_i$ and tag_i to $NewTag_{i-1}$;
    else
        Update $NewTag_i$ with $CurTag_i$;
    end if
end for

Data Description

- Database： 4 famous movies of different genres from the largest Chinese UGC (User Generated Content) video website ‘bilibili’.

- A group of movie fans to watch these movies, and manually record labels, which include timestamps and corresponding episode description text. Then another group of fans voted for the label that they thought best matched the episode

Number	Videos	IMDb Title*	Duration(sec.)	Comments	Genres
Movie_1	*“Memoirs of a Geisha”*	tt0397535	8708	8000	Drama | Romance
Movie_2	*“The Notebook”*	tt0332280	6909	7999	Drama | Romance
Movie_3	*“A Nightmare On Elm Street”*	tt0087800	5460	6681	Horror
Movie_4	*“Halo 4: Forward Unto Dawn”*	tt2262308	5460	7999	Action | Adventure | Science Fiction

Metrics

- Perplexity

$$Perplexity = p(D|M) = \exp^{(-\frac{\sum_{d=1}^{D} \log p(w_d|M)}{\sum_{d=1}^{D} N_d})}$$

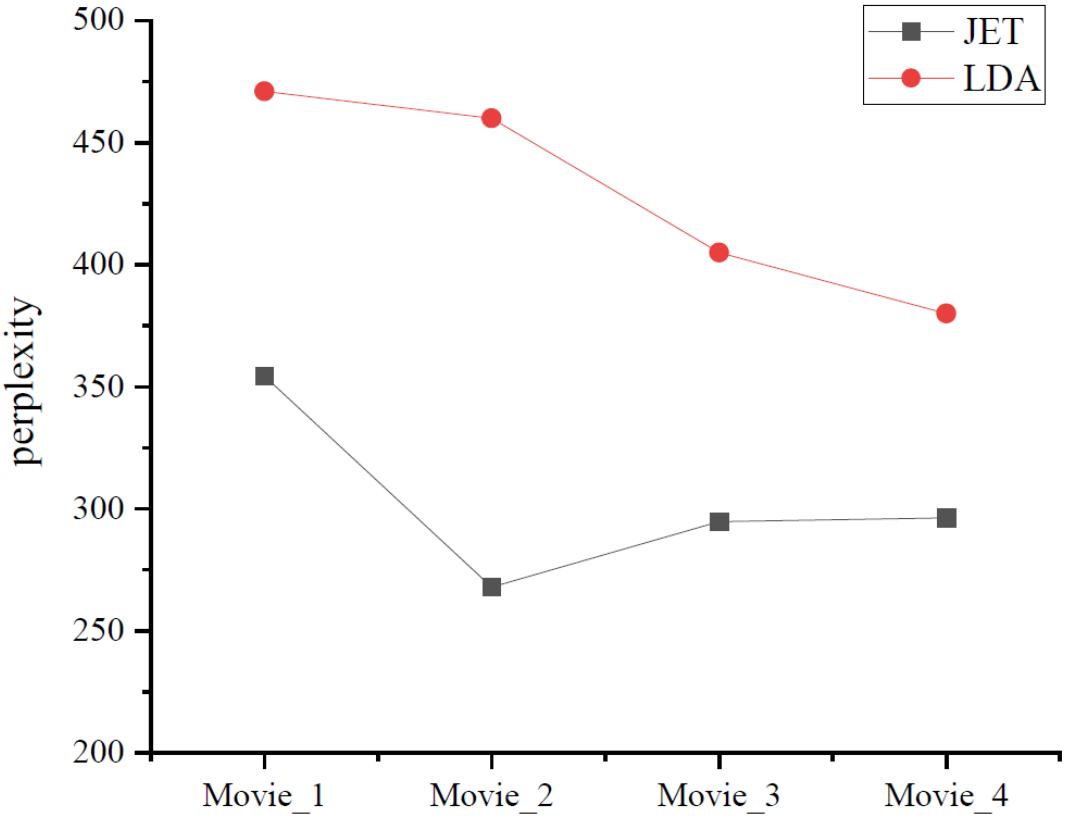
- Precision, Recall and F1-Score

Variable	Meaning
TP(True Positive)	$\|\hat{t} - t\| <= th_t$ and $sen(\hat{s}, s) >= th_s$
FP(False Positive)	$\|\hat{t} - t\| <= th_t$ and $sen(\hat{s}, s) < th_s$
FN(False Negative)	$\|\hat{t} - t\| > th_t$ and $sen(\hat{s}, s) >= th_s$
TN(True Negative)	$\|\hat{t} - t\| > th_t$ and $sen(\hat{s}, s) < th_s$

- Coverage(C), Overflow(O) and F-measure(FCO):

$$F_{CO} = \frac{2*C*(1-O)}{C+(1-O)}$$

Performance Evaluation



Topic Model of VEBD and LDA's perplexity about four movies.

Performance Evaluation

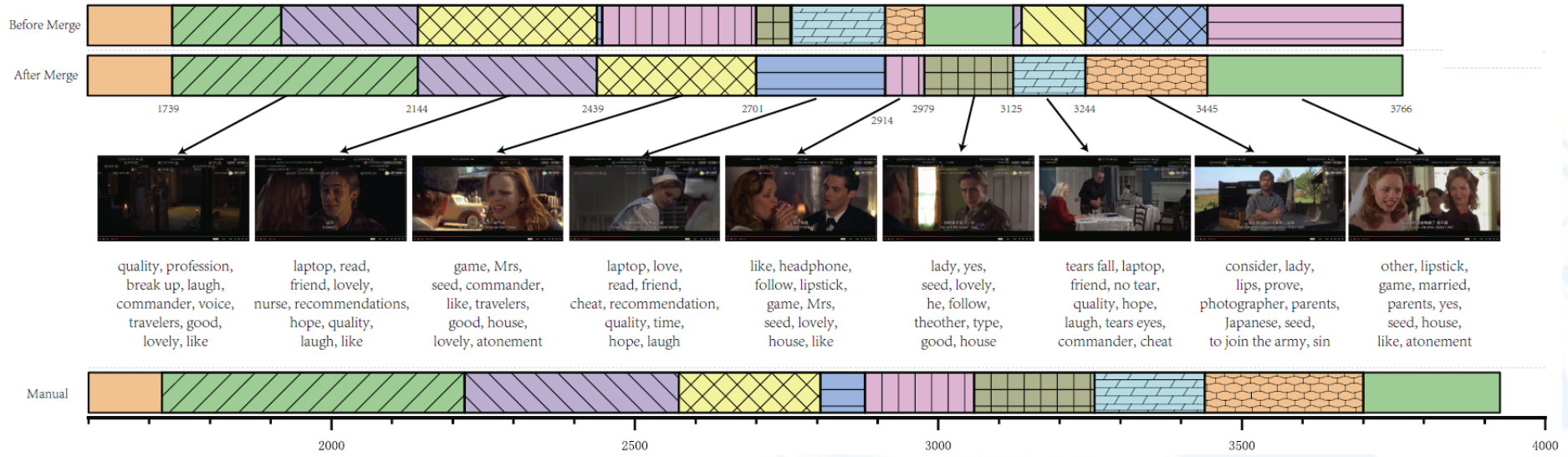
PRECISION, RECALL AND F1-SCORE FOR FOUR MOVIES IN THREE MODELS

	VEBD			TPTM			VEBD-w		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Movie_1	0.2963	0.6154	0.4	0.2174	0.7143	0.3333	0.2273	0.5556	0.3226
Movie_2	0.5556	0.5263	0.5405	0.375	0.3529	0.3636	0.4091	0.5294	0.4615
Movie_3	0.5625	0.5294	0.5455	0.1818	0.6667	0.2857	0.375	0.8571	0.5217
Movie_4	0.5	0.9	0.6429	0.0909	0.5	0.1538	0.4	0.75	0.5217
Average	0.4786	0.642775	0.532225	0.216275	0.558475	0.2841	0.35285	0.673025	0.456875

COVERAGE, OVERFLOW AND F_{CO} FOR FOUR MOVIES IN FOUR MODELS

	VEBD			GSTG			TPTM			VEBD-w		
	C	O	F_{CO}	C	O	F_{CO}	C	O	F_{CO}	C	O	F_{CO}
Movie_1	0.8288	0.8256	0.2882	0.7326	0.8346	0.2699	0.6781	0.7568	0.3580	0.8036	0.8256	0.2866
Movie_2	0.8017	0.8014	0.3183	0.7889	0.9871	0.0254	0.6466	0.7892	0.3179	0.8017	0.8204	0.2935
Movie_3	0.8022	0.7230	0.4118	0.8022	0.8064	0.3119	0.6374	0.7123	0.3964	0.8132	0.7346	0.4002
Movie_4	0.6923	0.3380	0.6768	0.6215	0.6128	0.4771	0.5275	0.4599	0.5337	0.6799	0.3467	0.6663
Average	0.7812	0.6720	0.4238	0.7363	0.8102	0.2711	0.6224	0.6795	0.4015	0.7746	0.6818	0.4116

Case Study



In a part of “The Notebook”, schematic comparison of the labels generated by VEBD model and the ground truth on timeline.

Conclusion

- This paper proposed an unsupervised automatic video episode boundary detection method.
- Compared with the traditional topic model, this article innovatively adds time prior and importance prior.
- It could not only automatically identify each episode boundary, but also detect the topic for video tagging.
- Experiments based on real data show that our model outperforms the existing algorithms in both boundary detection and semantic tagging quality.

Thank you!


BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS
北京邮电大学