Reading the Videos: Temporal Labeling for Crowdsourced Time-Sync Videos based on Semantic Embedding

Laboratory of Semantic Computing and Data Mining

Guangyi Lv

Presented by Lin Wang

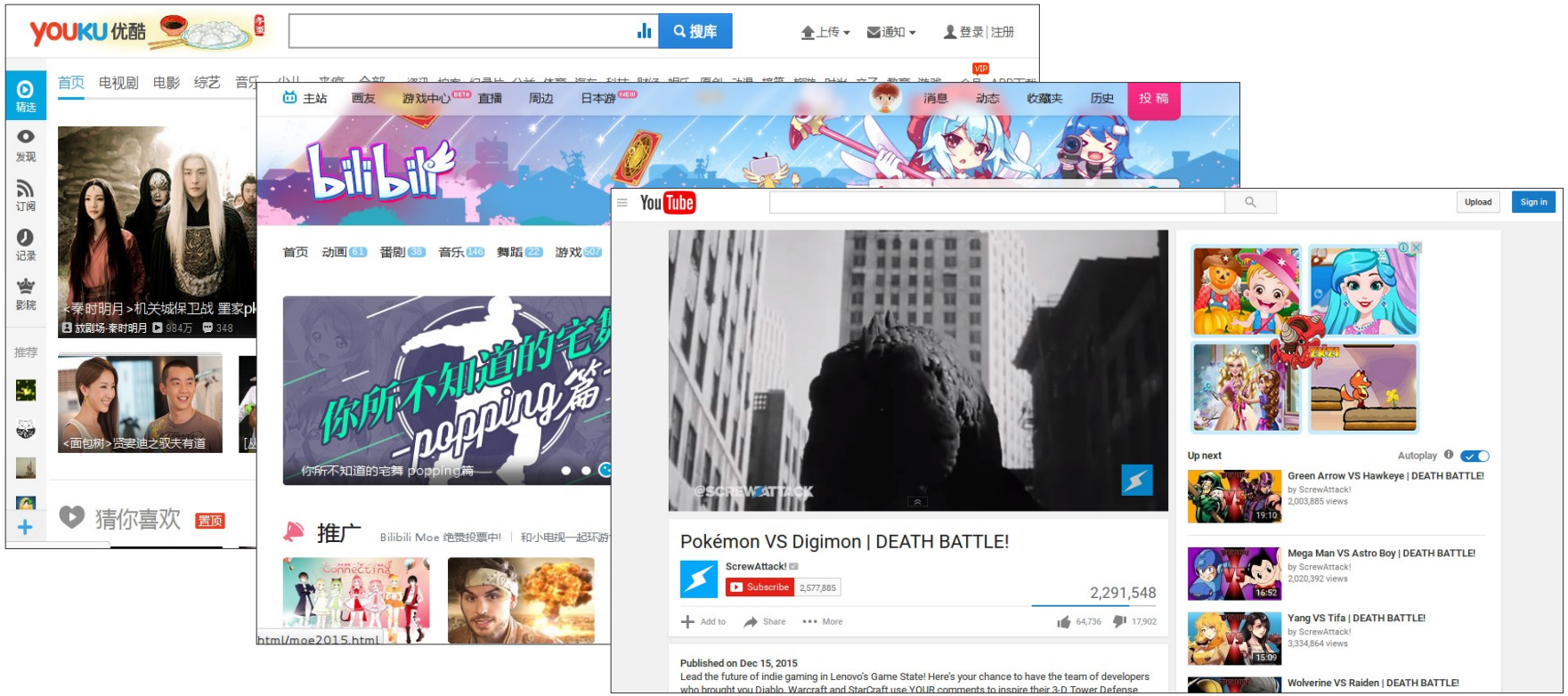
2016-02-14

Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

Background

- The booming of online video-sharing websites raises significant challenges in effective management and retrieval of videos.

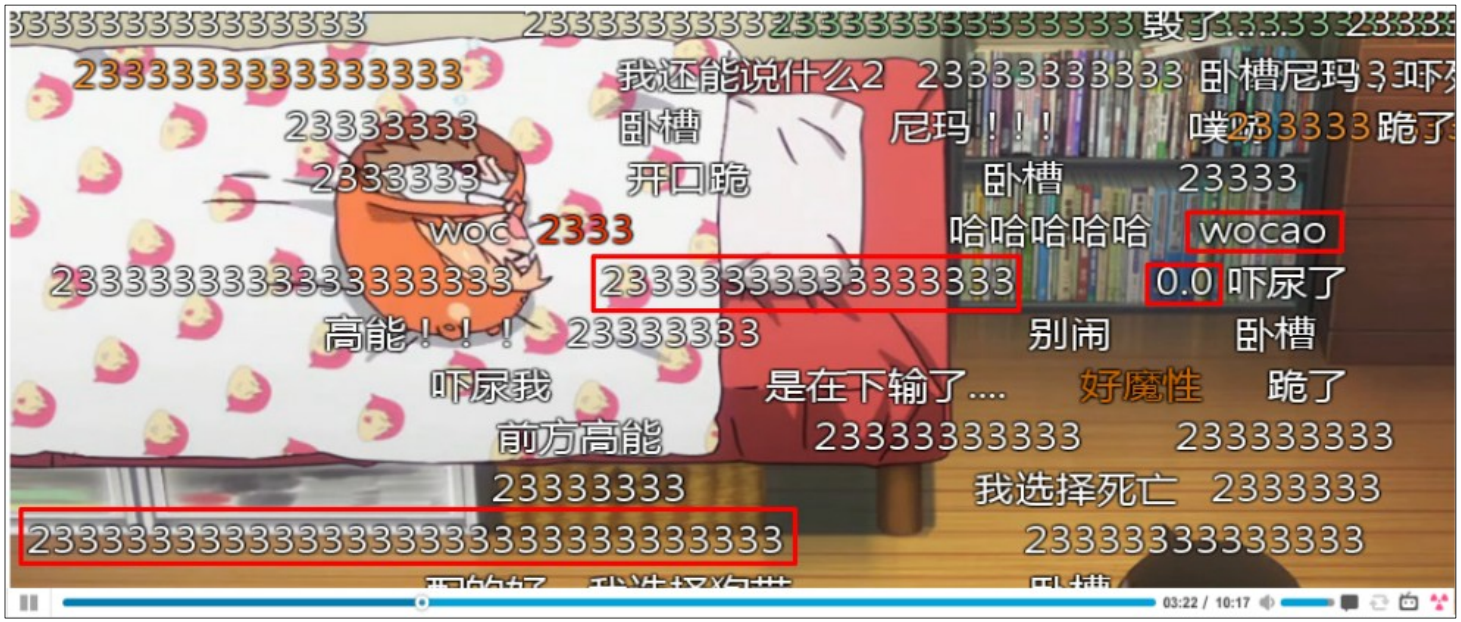


Background

- Precise retrieval on video shots with certain topics has been largely ignored.
- Users tend to view only parts of video shots on certain topics (a certain style, a certain movie star).
- Video labeling with both semantics and timestamps is urgently required.

Bullet-screen Comment

- Bullet-screen comments
  - Real-time comments on video shots are now available.
- New opportunities occur for temporal video labeling.



Bullet-screen Comment - Challenge

- Traditional NLP models may fail due to:
  - Typo errors
    - Especially in Chinese words
  - Informal expressions
    - e.g., "high energy"
    - e.g., "233333"
  - Latent meaning
    - e.g., "u ru sa i !" may relate to "Shana" or ""Kugimiya Rie
    - e.g., "philosopher" may relate to "Billy Herrington"
    - e.g., "⑨" may stand for "Cirno"
- We use deep learning !

Shana

Billy Herrington

Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

Problem Definition

- We target at finding and labeling video "highlights", i.e., video shots focusing on certain topics (labels)

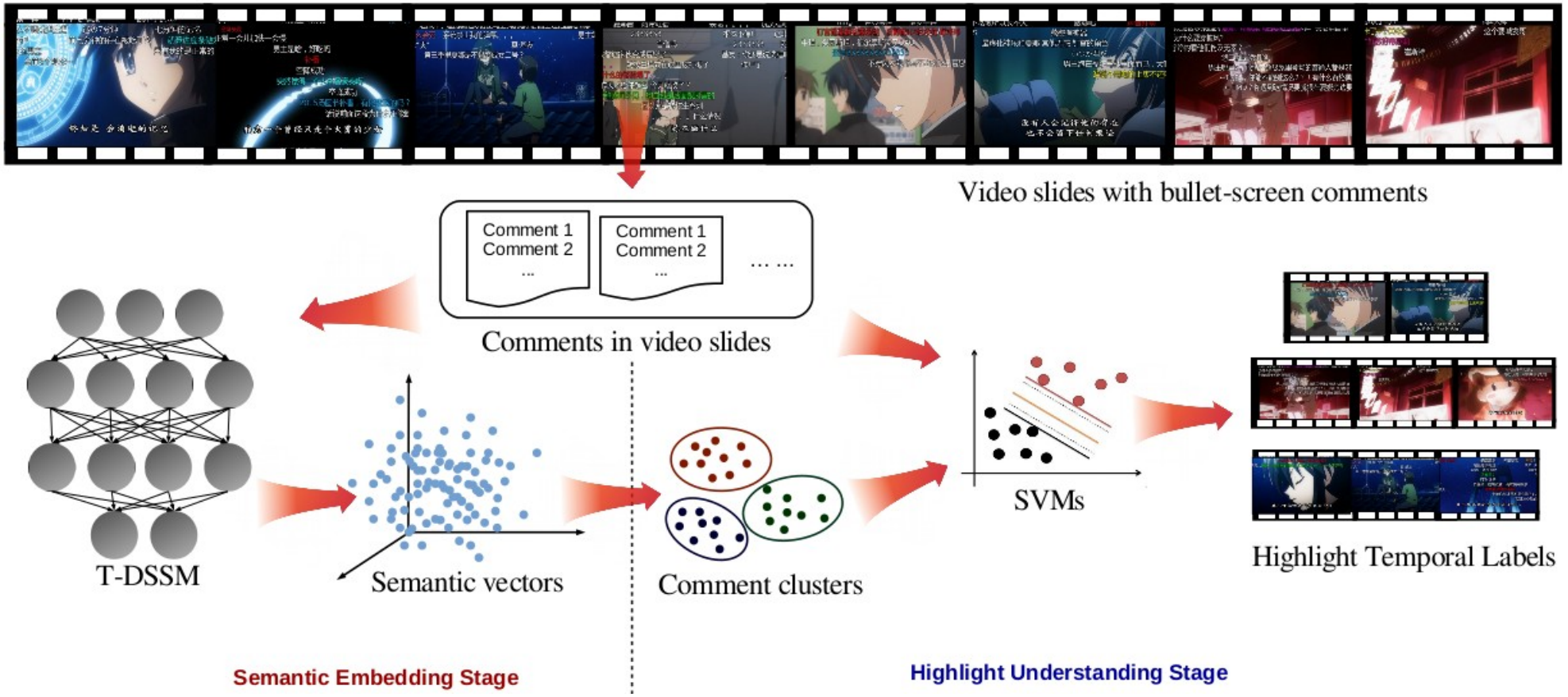
Definition 1 *Given the training set of videos with bullet-screen comments* $\mathbf{C_{train}} = \{< text, time >\}$, *as well as temporal labels* $\mathbf{L_{train}} = \{< t_s, t_e, lt >\}$ *in which* $< t_s, t_e >$ *indicates the timestamps (start and end) and* lt *presents the label type, the target is to precisely assign temporal labels* $\mathbf{L_{predict}} = \{< t_s^{'}, t_e^{'}, lt^{'} >\}$ *to the test set* $\mathbf{C_{test}}$, *where each* $< t_s^{'}, t_e^{'} >$ *indicates a video highlight with corresponding label as* $lt^{'}$.

Two Stage Framework

- Semantic embedding stage
  - Represent bullet-screen comments as corresponding semantic vectors.
- Highlight understanding stage
  - Highlight recognizing and labeling in a supervised way.

The Overall Procedure

- The overall framework:

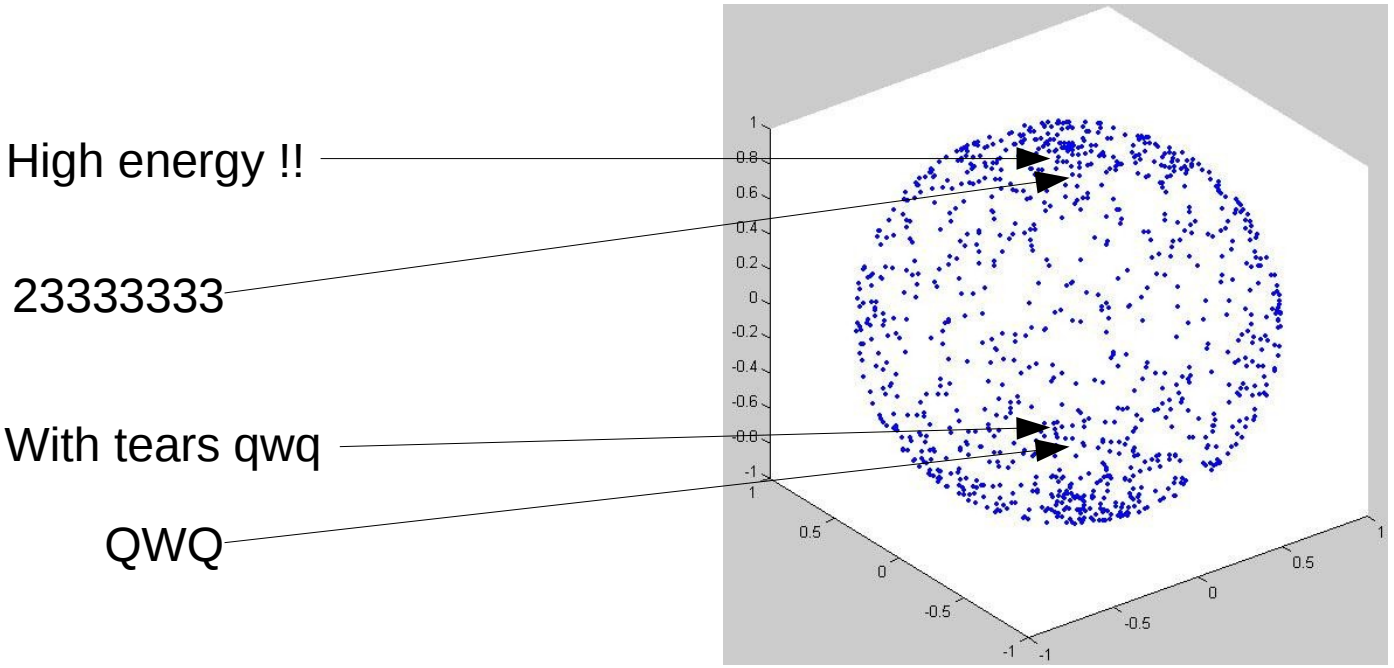


Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

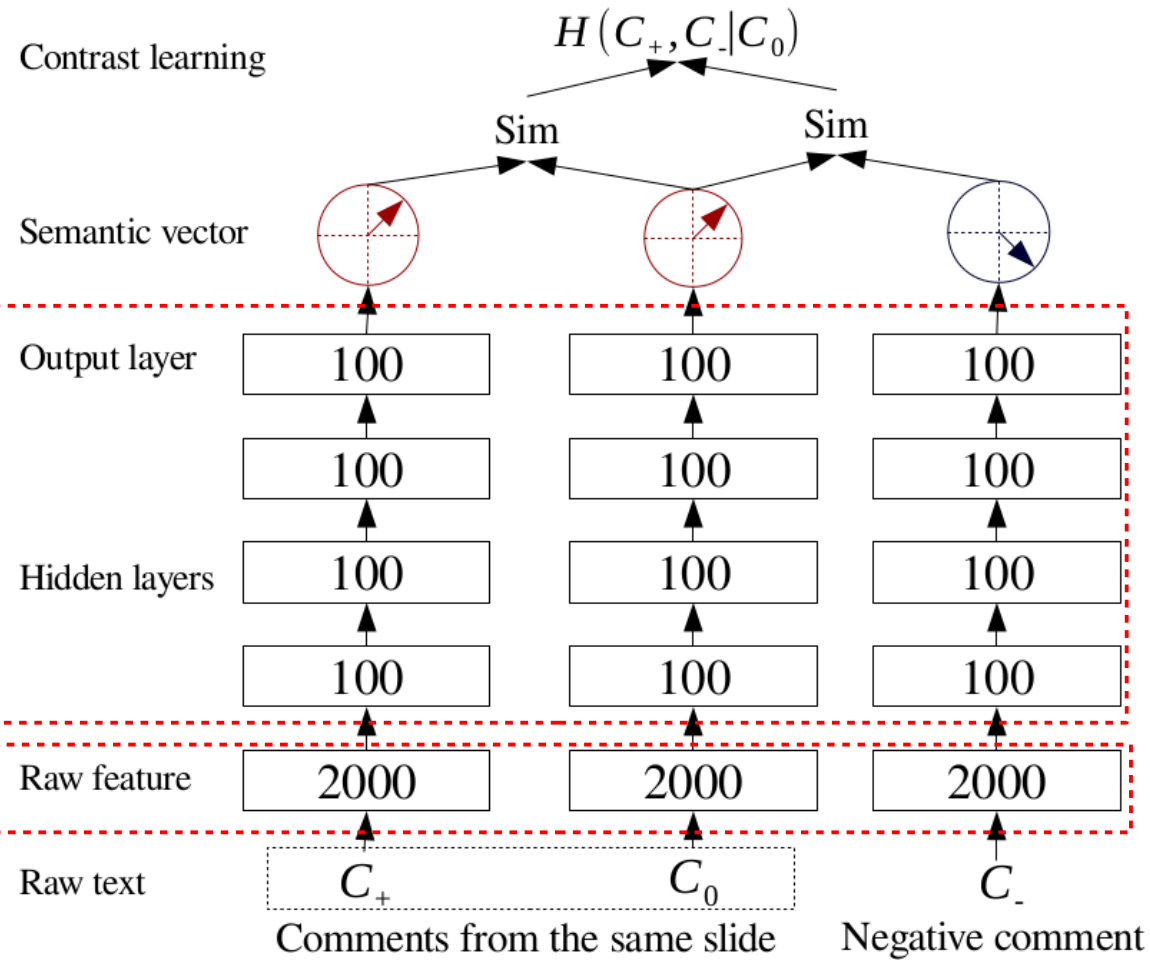
Semantic Embedding

- We design "Temporal Deep Structured Semantic Model" (T-DSSM)
- Represent each bullet-screen comment as corresponding semantic vector.



The Architecture of T-DSSM

- T-DSSM is based on DSSM
- A typical DNN

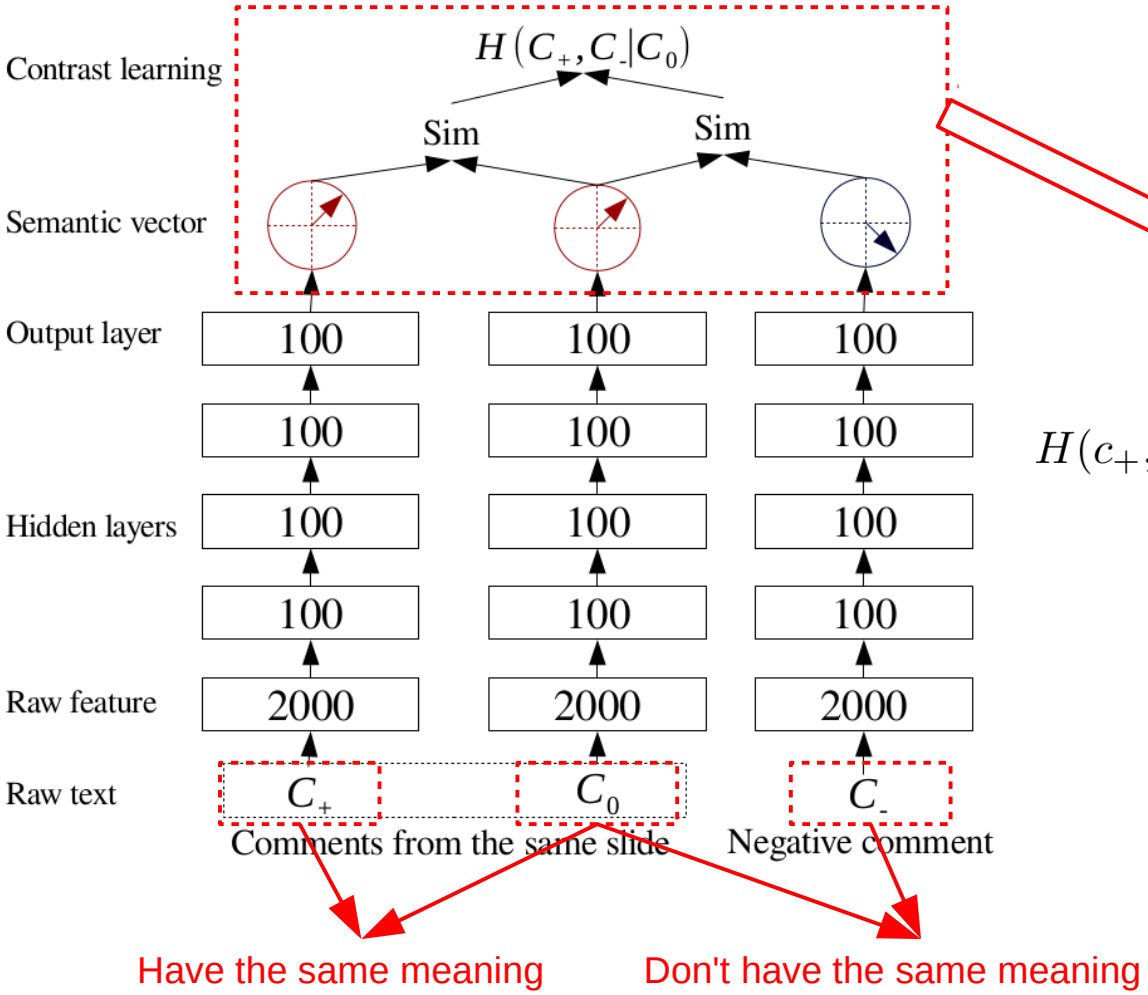


DNN with 3 hidden layers

For simplicity, bag of word feature can be used

Model Learning – Contrast Learning

- Loss function is designed based on contrast learning

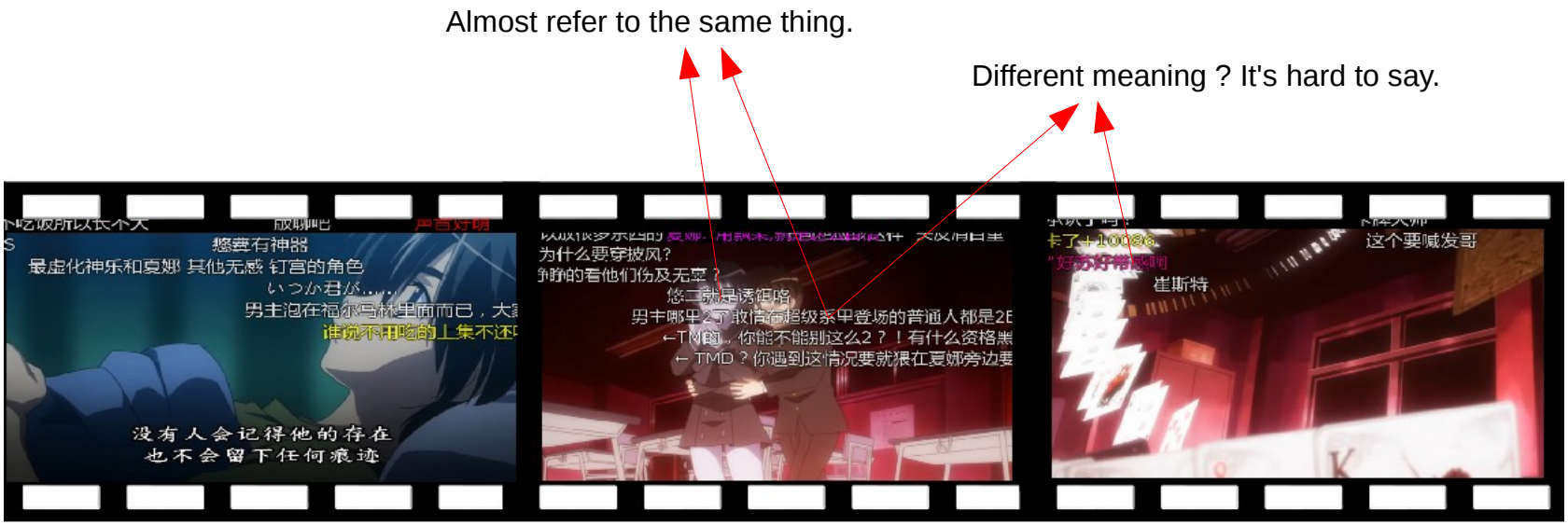


Contrast learning part:

$$H(c_+, c_-|c_0) = \frac{e^{Sim(c_+,c_0)}}{e^{Sim(c_+,c_0)} + e^{Sim(c_-,c_0)}}$$

Model Learning – Temporal Correlation

- Temporal correlation
  - Semantic vectors of adjacent comments could be reasonably similar.
- Hard to find negative comments.
- Simply selecting may result in difficulty in convergence.



Model Learning – EM Algorithm

- We solve this problem iteratively.
- Regard the negative comment as latent variable.
- Maximize the marginal distribution via EM algorithm:

$$L(\theta) = P(c_+|c_0) = \sum_{c_-} P(c_+, c_-|c_0)$$

$$P(c_+, c_-|c_0) = \frac{e^{H(c_+, c_-|c_0)}}{\sum_{c'_+} \sum_{c'_-} e^{H(c'_+, c'_-|c_0)}}$$

Model Learning – EM Algorithm

- Maximize the Q function by "E step" and "M step"
- The posterior probability can be calculated by sampling negative comments

$$Q(\theta|\theta^{(t)}) = E_{c_-|c_+}[logP(c_+, c_-|c_0)]$$

$$P(c_-|c_+) = \frac{e^{Sim(c_+,c_-)}}{\sum_{c'_- \in \mathbf{C}_-} e^{Sim(c_+,c'_-)}}$$

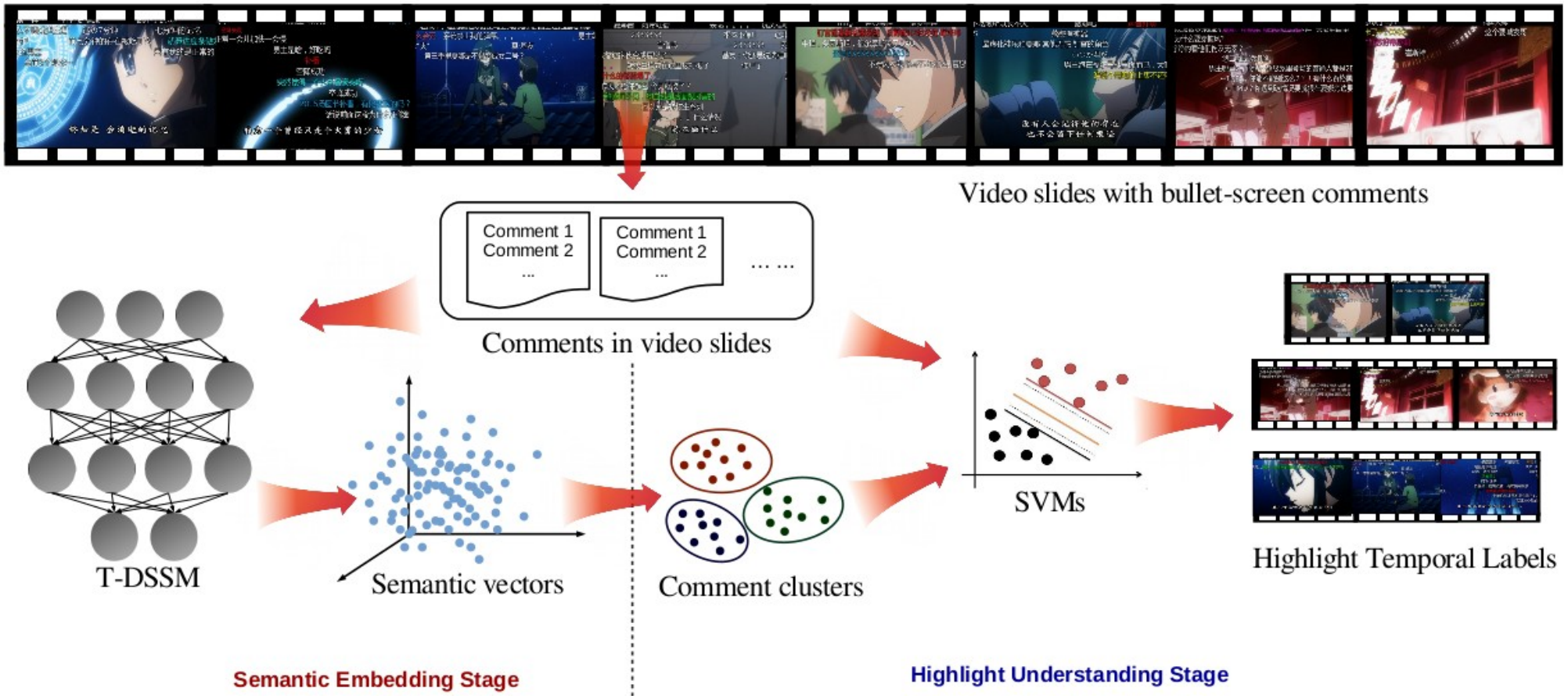
Has the same size with **C+**

Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

Highlight Understanding

- The overall framework:

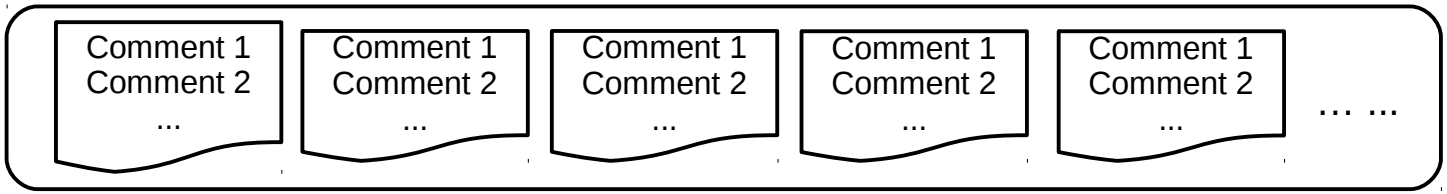


Highlight Understanding

- Temporal label contains time range information
  - Set time-window to split the video stream into slides.
  - Each slide is treat as the basic unit and extract its feature for labeling.
- The feature is presented as latent topics revealed from clustering semantic vectors.

… …

Slide 1 Slide 2 Slide 3 Slide 4 Slide 5

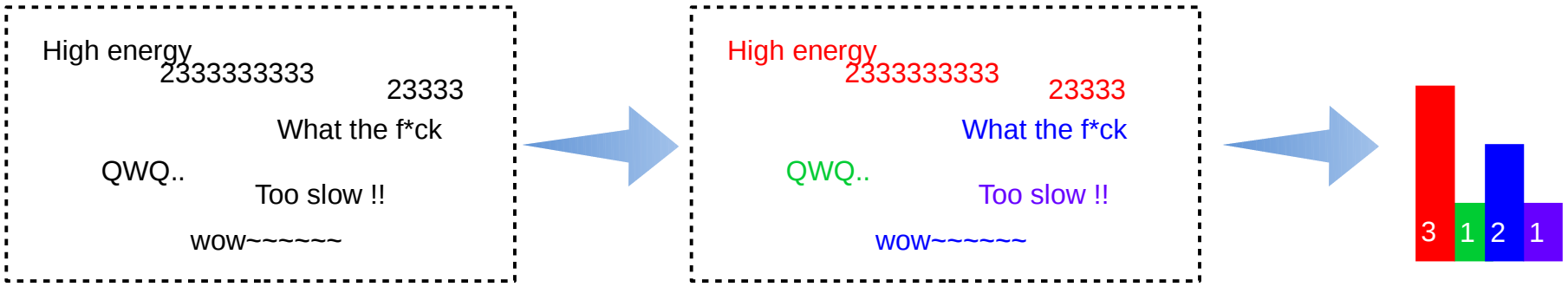


… …

Corresponding comments

Slide Feature

- Label each comment with the corresponding cluster (topic)
- Calculate comment frequency on each topic and denote it as feature "f"



One slide from the video

Label comment based on its cluster

Slide feature “f”

Three Steps

- Recognizing step
  - Semantically concentrating slide
    - Higher variance
    - Lower information entropy
  - Concentrating rating

$$rating = \frac{\frac{\sum_i^k (f_i - \bar{f})^2}{k-1}}{\sum_{\mathbf{p}} -p\log(p)}$$

- Labeling step
  - Train a classifier to map feature → label type
- Merging step
  - Merge the adjacent slides
    - Adjacent in time
    - Have the same label type

Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

Experiments: Data

- Real-world data set extracted from Bilibili
  - http://www.bilibili.com
- 133,250 comments
- 1,600 minutes long videos of different types of animation

Experiments: Baselines

- Word based
  - Generate a distribution of words instead of latent topics for each window-slide.
- LDA based
  - LDA is used to obtain the distribution of topics for each slide.

Experiments: Training Samples

- Three experts in Japanese anime domain
- Label the training samples with 10 types:
  - describe scenes
    - "funny", "moving", "surprising", "sad", "magnificent fighting"
  - describe characters
    - "cool", "lovely girl", "sexy shot"
  - about music
    - "OP", "BGM"

Experiments: Metrics

- Hit Time

$$hit_{time} = \sum_{L_i.lt = L_j^+.lt} L_i \cap L_j^+$$

- Precision, Recall and F1 score
  - Defined based on “hit time”.
- Precision/Recall of labels
  - Measure the ability of discovering different label types.
- Recall of shots

Experimental Results

- The T-DSSM based framework outperforms the other models in all metrics

Model	Word based	LDA	**T-DSSM**
$Precision$	0.3509	0.3695	**0.4575**
$Recall$	0.3885	0.4013	**0.4969**
$F1$	0.3687	0.3847	**0.4764**
$Precision_{label}$	0.4992	0.5139	**0.6103**
$Recall_{label}$	0.4452	0.4547	**0.5738**
$Recall_{shot}$	0.3486	0.3669	**0.4770**

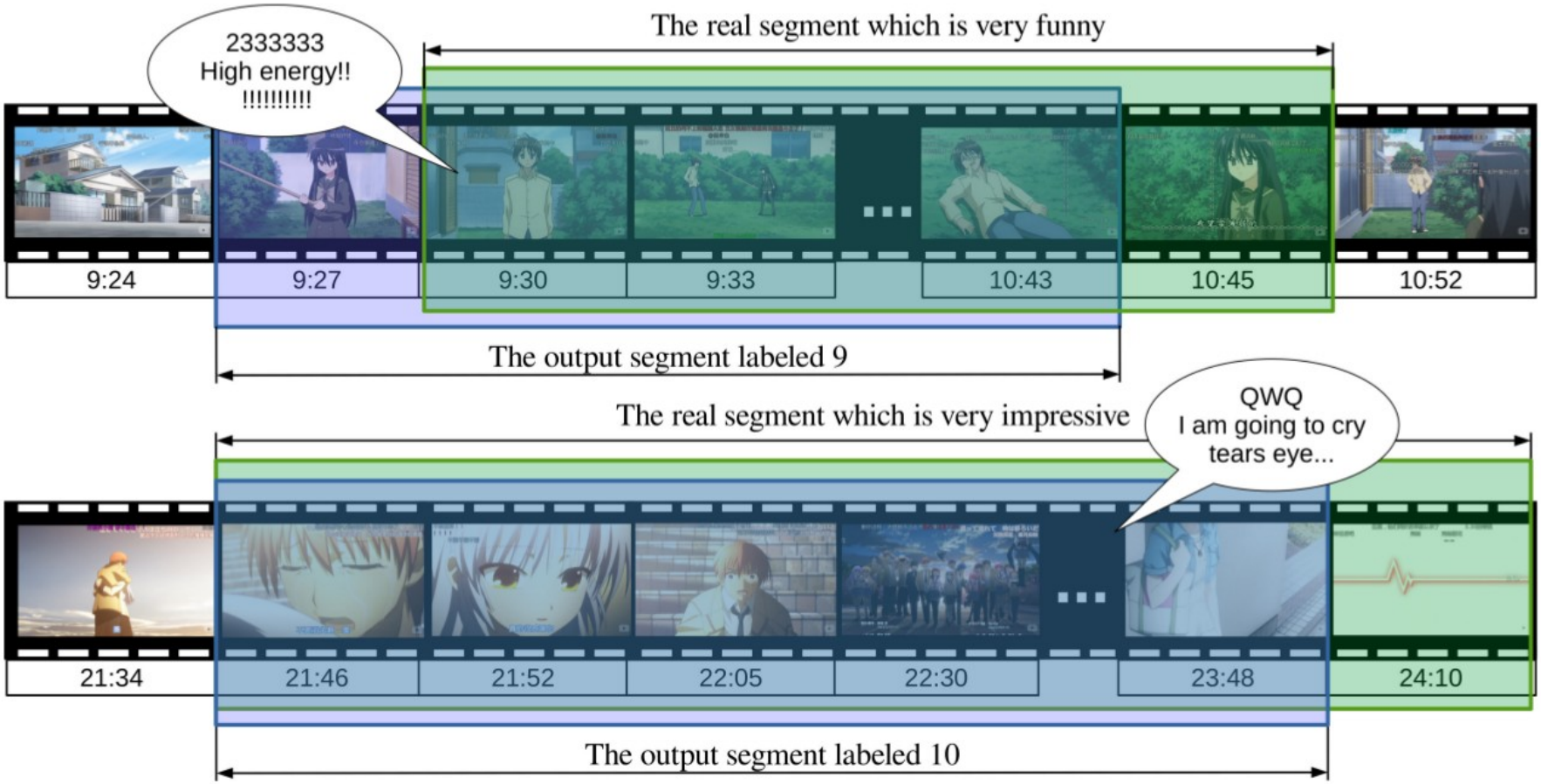
Latent Topics

- Some clustering results.

#	Comment content	Actual topic
Cluster 0	No little TV today, moved 64M no pressure Three years to review ends with flowers I return to the battlefield	These are something appears at the beginning of a video. The user always send comments about greeting to other here, talking about the net speed and other pointless topics
Cluster 1	u ru sa i! u ru sa i! u ru sa i!! We seldom hear the be poker-faced Kugimiya Do you really don't look at Shana Shana my lover Shana quite let us worried about...	Topics associated with Shana who is the actress of a famous Japanese animation called "Burning-Eyed Shana". Note that the word "u ru sa i" is translated directly from Japanese which means "so noisy" in English and it's also known as Shana's pet phrase. Kugimiya is the voice actor of Shana whose full name is Kugimiya Rie.
Cluster 9	WARNING!High energy!High energy! 2333333333 yooooooooooooo press ← will have surprise !!!!!!!!!!!	These comments usually show along with something interesting, exciting, funny or even terrible. This kind of sentences may be the most difficult to understand among comments.
Cluster 10	first animation made me cry== good bye qwq tears eyes ... the BGM is too tear for now QWQ	Bullet-screen comments are also often used to express emotions. Like this cluster, people must be talking about a sad topic along with a moving BGM (Background Music). Note that "qwq" is used as a symbol of cry in which letter "q" looks like a tears eye.

Case Study

- Two typical results.



Outline

- Background
- Problem Definition and Framework
- Semantic Embedding
- Highlight Understanding
- Experiments
- Conclusions

Conclusions

- Proposed a novel video understanding framework to assign temporal labels on highlighted video shots.
- T-DSSM was designed to represent comments into semantic vectors to deal with the informal expression of bullet-screen comments.
- T-DSSM was by trained taking advantage of comments' temporal correlation.
- Video highlight shots were recognized and temporally labeled via mapping semantic vectors in a supervised way.
- Experiments on real-world dataset proved that our framework could label video highlights effectively.

Thank You !

- That's all.
- Thank you !
- Feel free to ask questions ~