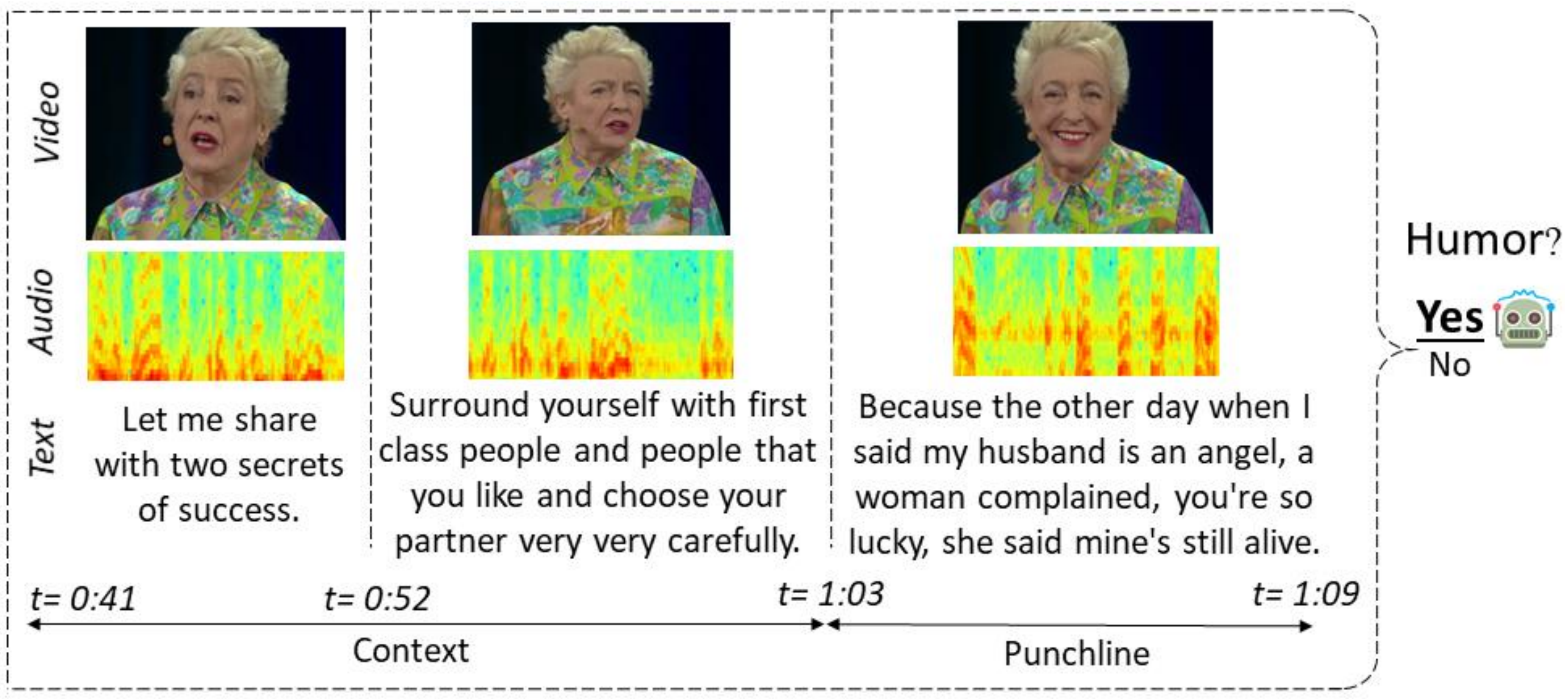
UR-FUNNY : A Multimodal Language Dataset for Understanding Humor

Md Kamrul Hasan* (mhasan8@cs.rochester.edu), Wasifur Rahman*, Amir Zadeh, Jianyuan Zhong,
Md Iftekhar Tanveer, Louis Philippe Morency, Mohammed (Ehsan) Hoque
University of Rochester & Carnegie Mellon University, USA

Motivation

Can computer recognize the punchline of a joke using different modalities (text, audio & video) and background context?


Video
Audio
Text
Let me share with two secrets of success.
Surround yourself with first class people and people that you like and choose your partner very very carefully.
Because the other day when I said my husband is an angel, a woman complained, you're so lucky, she said mine's still alive.
Humor?
Yes
No
t= 0:41
t= 0:52
t= 1:03
t= 1:09
Context
Punchline

Dataset Overview

- **UR-FUNNY: First multimodal** (text, audio & video) dataset for **humor** detection
- **8257** Humor Instances (video) from **TED** Talk
- It has **punchline** & background story **context**
- Average duration of each data = 19.67s ; context = 14.7s & punchline = 4.97s
- **Diverse** in both speakers (1741) and topics (417)
- Total duration is **90.23 hour**

Publicly available to download (data + processed features + code)

Link: https://github.com/ROC-HCI/UR-FUNNY

Dataset Analysis

DATA Acquisition

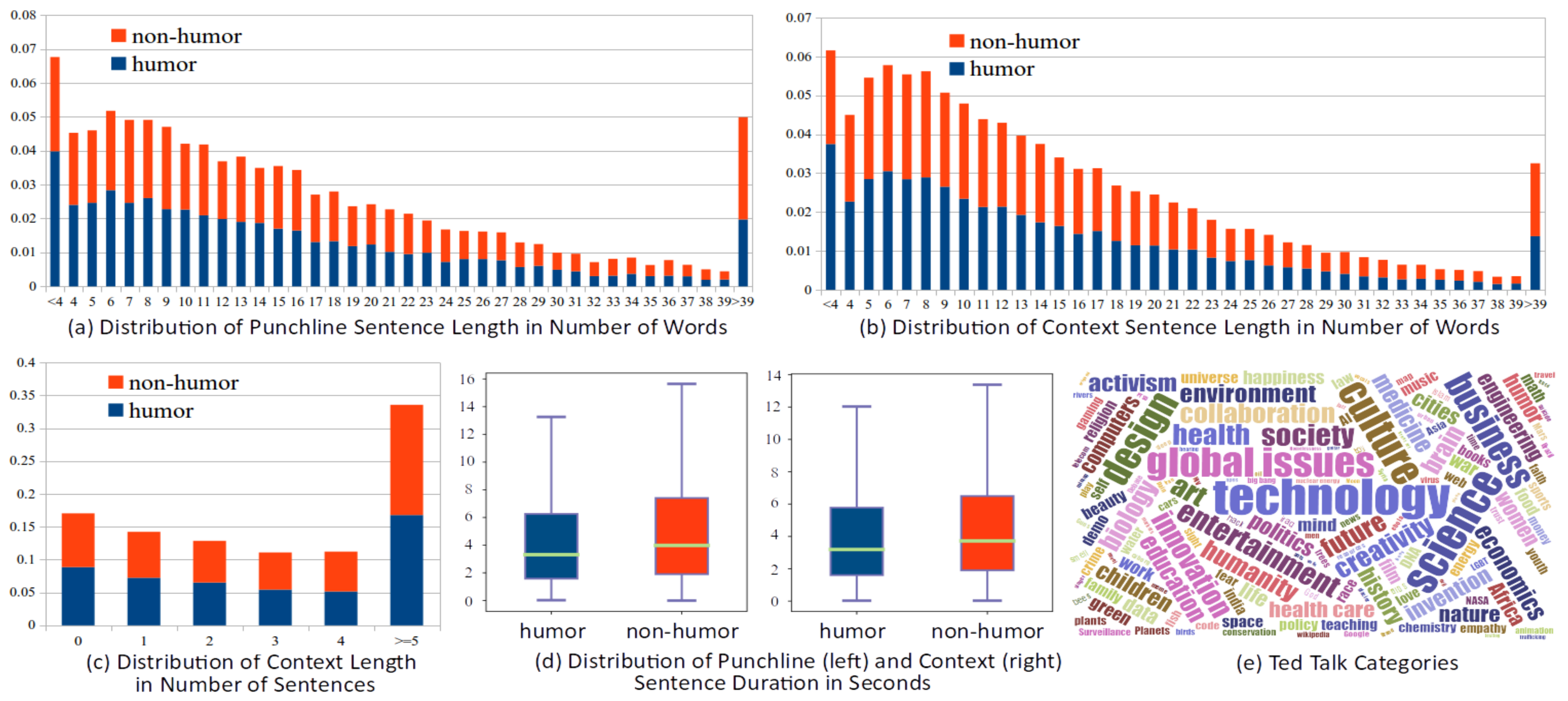
- Collected 1866 **TED talk videos** + transcripts
- **Audience Laughter** markup is used to filter 8257 humorous punchlines from transcript
- **Context** is extracted from the prior sentences to the punchline
- Negative examples are from same videos (**homogenous**)
- Force alignment is used to align text, audio & video
- Preprocessed features: text = glove, audio = COVAREP, video= OpenFace

UR-FUNNY Vs Other Datasets

Dataset	#Pos	#Neg	Modality	Type	#Speaker
16000 One Liner	16000	16000	$\{t\}$	Joke	-
Pun of the Day	2423	2423	$\{t\}$	Pun	-
PTT Jokes	1425	2551	$\{t\}$	Political	-
Ted Laughter	4726	4726	$\{t\}$	Speech	1192
Big Bang Theory	18691	24981	$\{t,a\}$	Tv show	< 50
UR-FUNNY	8257	8257	$\{t,a,v\}$	Speech	1741

$t = text, a = audio, v = video$

Dataset Statistics



(a) Distribution of Punchline Sentence Length in Number of Words

(b) Distribution of Context Sentence Length in Number of Words

(c) Distribution of Context Length in Number of Sentences

(d) Distribution of Punchline (left) and Context (right) Sentence Duration in Seconds

(e) Ted Talk Categories

General		Punchline / Context	
total #video	1866	#sentence in punchline	1
total duration (hour)	90.23	avg #word in punchline	16.14
#humor instances	8257	avg duration of punchline (sec)	4.97
#non-humor instances	8257	avg #sentences in context	2.86
#sentence	63727	avg duration of context (sec)	14.7
avg #word in sentences	15.15	avg #word in context sentence	14.80

Contextual Memory Fusion Network (C-MFN)

Problem Formulation

Set of modalities, $M = \{t, a, v\}$;
$t = text, a = audio, v = vision$

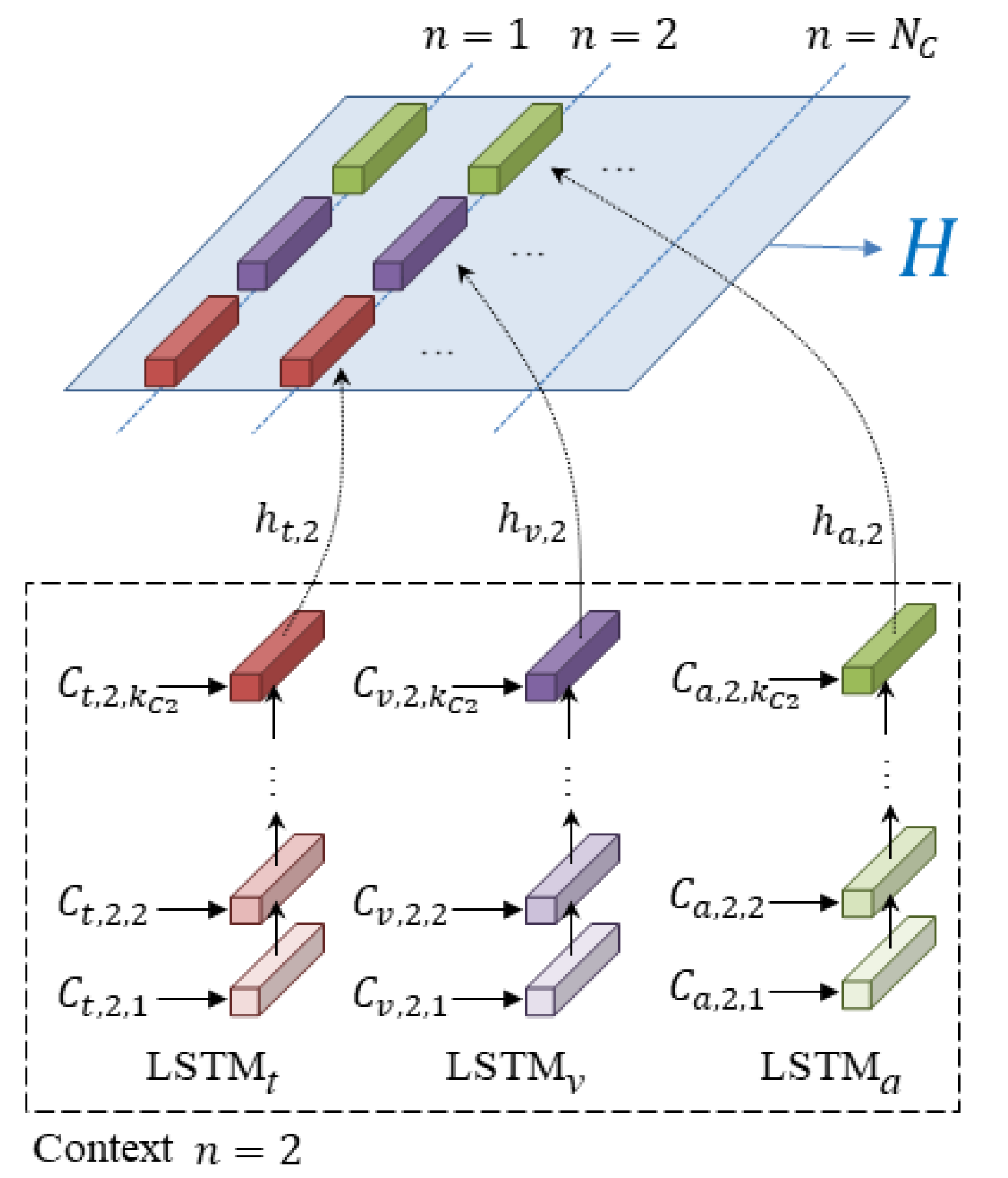
Each instance, $I = \{l, P, C\}$; $l = label, P = punchline, C = Context$

Punchline & context have multiple modalities $P = \{P_m ; m \in M\}$ & $C = \{C_m ; m \in M\}$.

$C_m = [\, C_{m,1}, C_{m,2} , \dots . C_{m,N_c}]$; N_c = number of context sentences

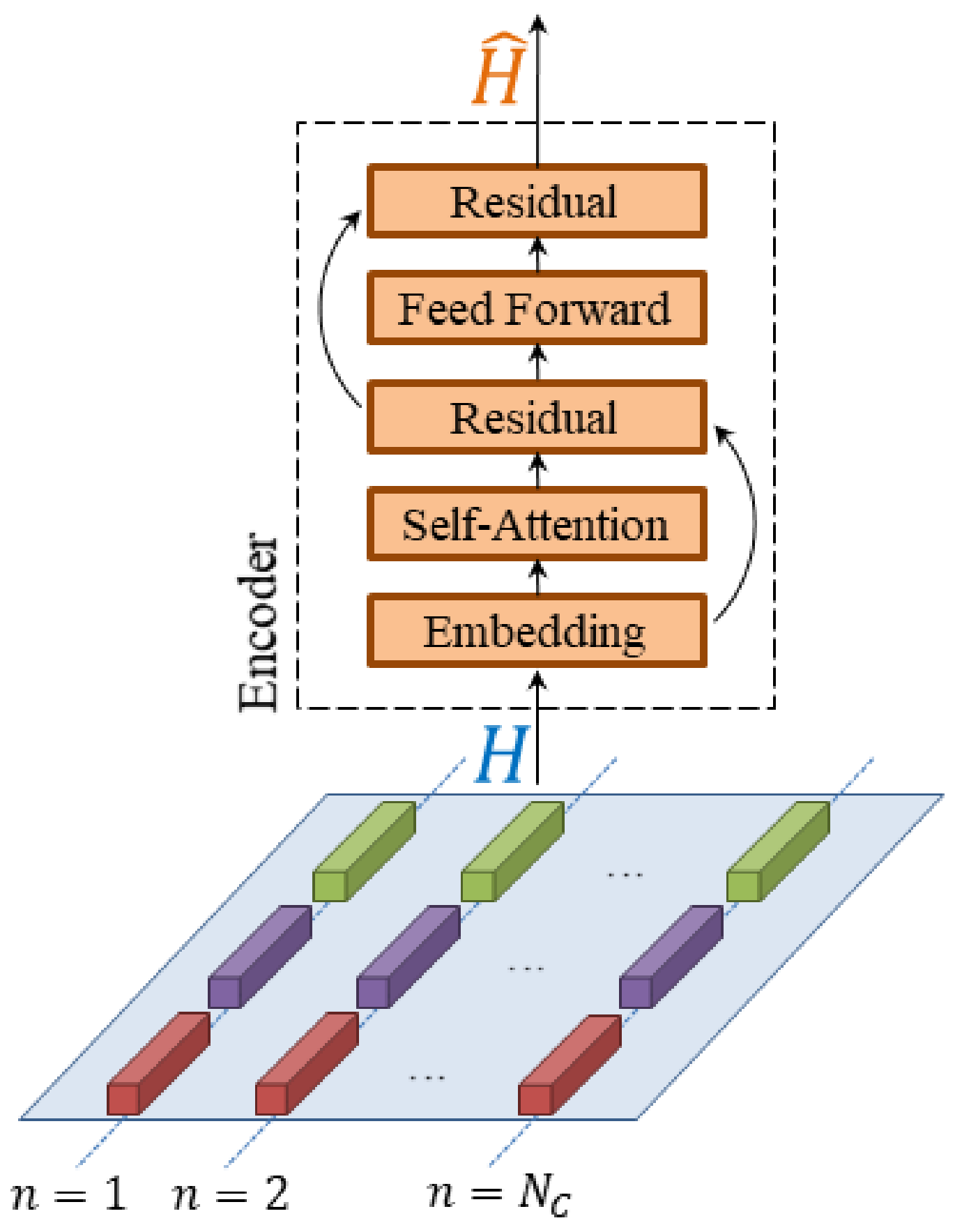
K_p = Number of words in the punchline
K_{C_n} = Number of words in the nth context sentence ; $n \in \{1, N_c\}$

Unimodal Context Network



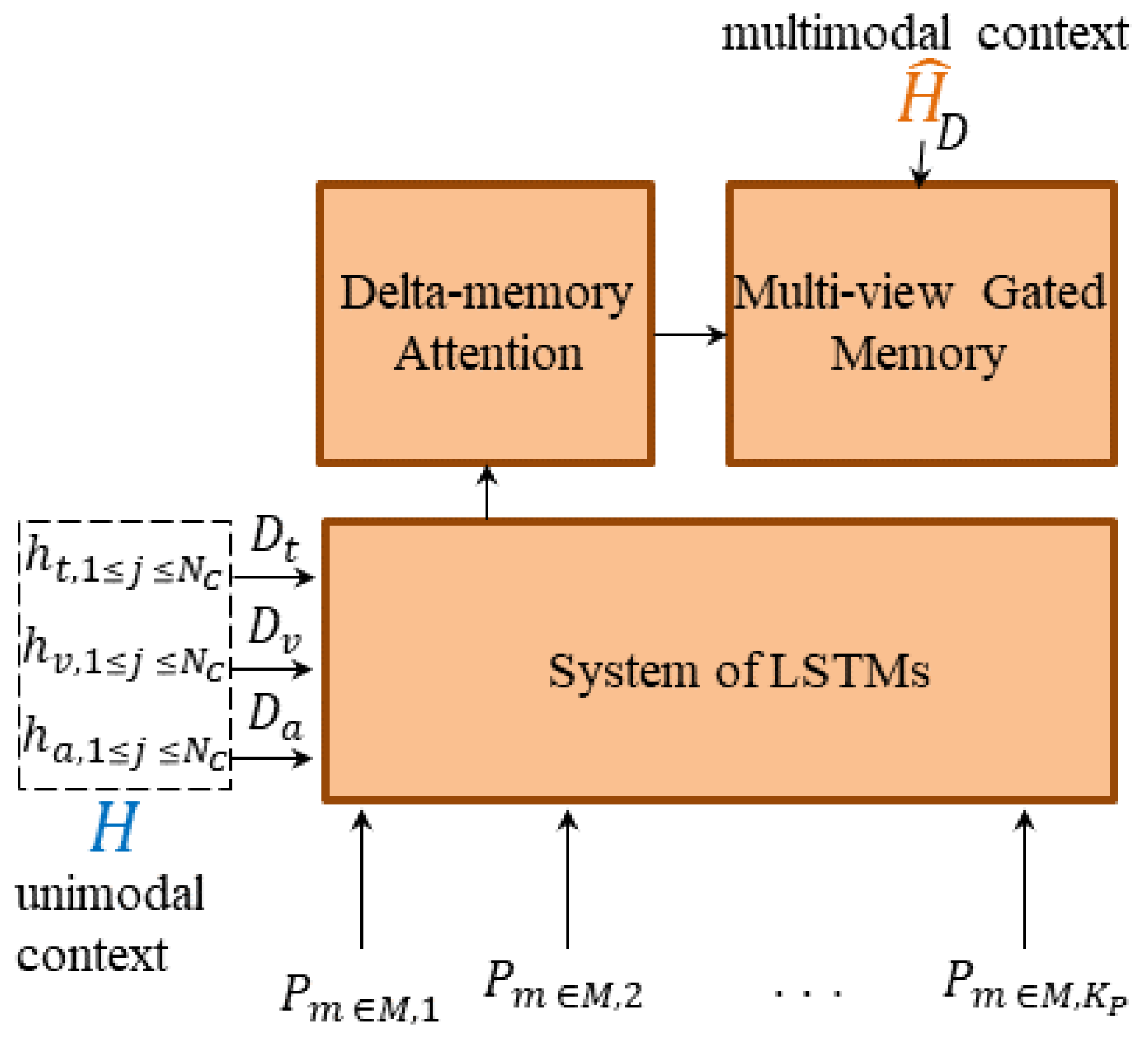
Three LSTM for summarizing three modalities

Multimodal Context Network



Transformer encoder creates multimodal context

Memory Fusion Network (MFN)



LSTM cells & Multi-view memory initialized with unimodal & multimodal context

Ablation Study

Role of context & punchline

C-MFN (P) : This variant uses only punchline; **C-MFN (C)** : This variant uses only context ; **C-MFN:** uses both

Role of different modalities

(T) only text modality is used ; **(A+V)** only vision and acoustic modalities are used; **(T+A+V)** all modalities are used together

Results

Modality	T	A+V	T+A	T+V	T+A+V
C-MFN (P)	62.85	53.3	63.28	63.22	64.47
C-MFN (C)	57.96	50.23	57.78	57.99	58.45
C-MFN	64.44	57.99	64.47	64.22	**65.23**

Performance Metrics: Binary Accuracy

C-MFN that uses both punchline and context along with all three modalities give best performance

Summary

- Humor can be modeled better as multimodal
- Context and punchline are important
- Brings new challenge to Humor understanding by extending the task in multimodal domain