

科研

国际 (index.htm)

科研

(../science/index.htm)

您所在的位置: 首页 - (../index.htm) 科研 - (../index.htm) 国际 (index.htm)

新学期首次! 2名人大高瓴人工智能学子赴美参会

日期: 2023-02-27 访问量: 2708

2023年2月7日至14日, 国际人工智能顶级学术会议 AAI (the 37th AAI Conference on Artificial Intelligence) 在美国华盛顿特区举行。中国人民大学高瓴人工智能学院两名博士研究生王庆梅、吴宜函因文章被接收获邀参会, 与参会学者现场深入交流了研究工作。



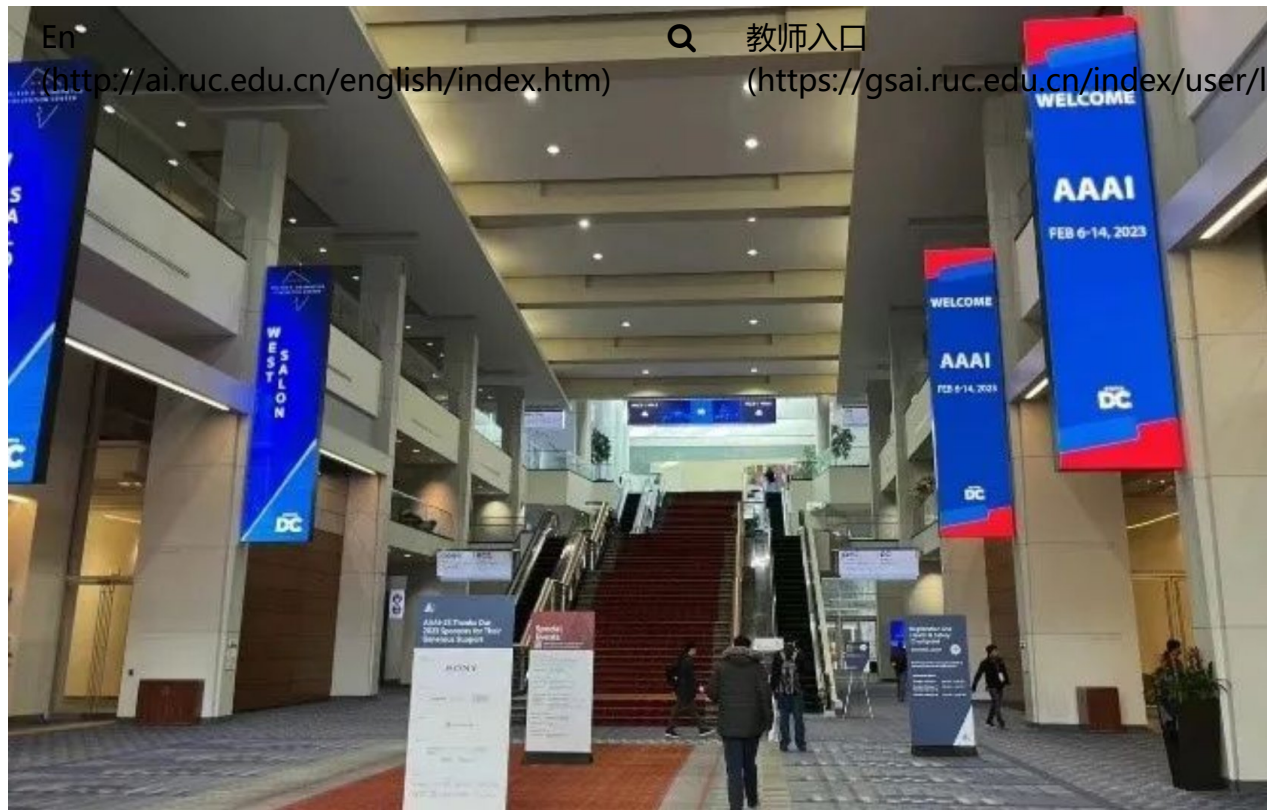
En

(<http://ai.ruc.edu.cn/english/index.htm>)



教师入口

(<https://gsai.ruc.edu.cn/index/user/login.html>)



会议现场



大会开幕式

吴宜函是高瓴人工智能学院直博二年级学生，这也是她第一次出国参加学术会议。她的导师宋睿华长聘副教授积极鼓励她现场参加这次会议。“这次参加AAAI 2023对我来说是一个全新的体验，是一次真正的与我所做的研究领域前沿学者深入交流的好机会。”吴宜函说，“AAAI提供了一个很好的平台，不仅可以和同领域的其他研究人员热烈讨论本领域的问题，交流心得，碰撞思路，还可以广泛了解其他领域的工作，巧借他山之石，激发自己的灵感。”

她关于AI创作的工作“VideoDubber: Machine Translation with Speech-Aware Length Control for Video Dubbing”被AAAI 2023 接收为口头报告（Oral Presentation），在主题为“Speech & Natural Language Processing”的环节中介绍并分享了她们的工作。这篇工作提出了一种针对视频配音任务的机器翻译系统，该系统直接考虑翻译过程中每个词的语音时长并显式控制，以匹配原始语音与翻译后语音的总时长。VideoDubber在四个语言方向(德语→英语、西班牙语→英语、汉语→英语，英语→汉语)上设计实验，结果表明该方法对生成的语音具有较好的长度控制能力，优于基线方法。

“当你的工作被大家不断提问，积极讨论时，你更能感受到自己工作的价值与意义，是十分幸福的。”吴宜函回忆说。此外，在会议上与业内“大佬”的交流，也让她对科研有了更深刻的思考。“他们对工作的热忱与执着，甚至有些执拗的态度深深打动了我，也坚定了我对科研的信心。”



吴宜函同学在主题为
“Speech & Natural Language Processing”
的会议上作报告

En

(http://ai.ruc.edu.cn/english/index.html)

VideoDubber: Machine Translation with Position-aware Length Control in Video Dubbing

(https://gsai.ruc.edu.cn/index/user/login.html)

Yihan Wu¹, Junliang Guo², Xu Tan², Chen Zhang³, Bohan Li³, Ruihua Song¹, Lei He³, Sheng Zhao³, Arul Menezes⁴, Jiang Bian⁴

¹Gaoling School of Artificial Intelligence, Renmin University of China
²Microsoft Research Asia, ³Microsoft Azure Speech, ⁴Microsoft Azure Translation

教师入口


(https://gsai.ruc.edu.cn/index/user/login.html)

ABSTRACT

Video dubbing aims to translate the original speech in a film or television program into the speech in a target language, which can be achieved with a cascaded system consisting of speech recognition, machine translation and speech synthesis. To ensure the translated speech to be well aligned with the corresponding video, the length/duration of the translated speech should be as close as possible to that of the original speech, which requires strict length control. Previous works usually control the number of words or characters generated by the machine translation model to be similar to the source sentence, without considering the asynchronicity of speech as the speech duration of words/characters in different languages varies. In this paper, we propose VideoDubber, a machine translation system tailored for the task of video dubbing, which directly considers the speech duration of each token in translation, to match the length of source and target speech. Specifically, we control the speech length of generated sentence by guiding the prediction of each word with the duration information, including the speech duration of itself as well as how much duration is left for the remaining words. We design experiments on four language directions (German to English, Spanish to English, Chinese to English), and the results show that VideoDubber achieves better length control ability on the generated speech than baseline methods. To make up the lack of real-world datasets, we also construct a real-world test set collected from films to provide comprehensive evaluations on the video dubbing task.

TASK OVERVIEW

Video dubbing aims to translate the original speech in a film or television program into the speech in a target language, and the translated speech should be well aligned with the corresponding video.



METHODS

We propose VideoDubber, a speech-aware length control model to directly control the translated speech duration in video dubbing.

- We make the model aware of how much duration is left at each time-step as well as the speech duration of each word.
- We introduce a special pause word [P] to control the speech length more smoothly by considering the prosody of speech through adjusting the duration of [P].
- Considering the scarcity of real-world video dubbing dataset, we construct a test set to provide comprehensive evaluations of video dubbing systems.

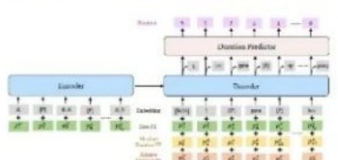


Figure 2: The overall architecture of our machine translation model with speech-aware length control for video dubbing. PE stands for positional embedding and [P] indicates the special pause token. We set $N = 5$ and follow Equation (2) to calculate the relative duration PE.

RESULTS

- VideoDubber consistently outperforms its baselines with a large margin, demonstrating that the proposed speech-aware length control achieves better speech duration synchronization than controlling the number of words/characters.
- More duration information on three kinds of PEs in four language directions, we can find that the absolute and relative duration PE are both crucial to achieve better speech-aware length control results.
- In real-world video dubbing test set constructed by us, VideoDubber achieves better performance compared with Transformer and the token number control baseline. It proves that in the real task, when the speech and music is consistent, the proposed NMT model with speech-aware length control can achieve better synchronization control ability as well as the translation quality.

VIDEO DUBBING TEST SET

Considering the scarcity of real-world video dubbing dataset, we construct a test set collected from dubbed films to provide comprehensive evaluations of video dubbing systems.

- 42 conversation clips from nine films.
- The clip duration is around 1 - 3 minutes.
- More than 10 sentences are involved in each clip, which contains both long and short sentences.
- The face of speaker is visible mostly during his or her talks, especially visible lips at the end of speech.

Finally, we obtain the test dataset with a total duration of 1 hour, including the original audio, source speech with transcripts, and human-dubbed speech with transcripts.

吴宜函同学在大会会场展示的学术海报 (Poster)

王庆梅是高瓴人工智能学院2021级博士生，为了和国际学者更充分地交流，了解研究领域前沿，在其导师许洪腾准聘副教授的支持下，她与吴宜函一同前往美国华盛顿参加了会议。



En
(<http://ai.ruc.edu.cn/english/index.htm>)

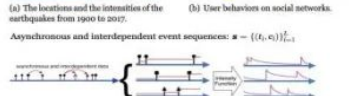
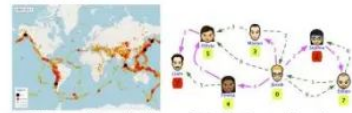
教师入口
(<https://gsai.ruc.edu.cn/index/user/login.html>)

王庆梅同学（左三）与参会学者合影

王庆梅的论文Hierarchical Contrastive Learning for Temporal Point Processes也被会议接收为口头报告（Oral Presentation），在主题为“Time-Series& Data Streams”的环节中介绍并分享了该工作。该工作提出了一种层次化对比学习方法(HCL)来缓解时序点过程学习过程中的过拟合问题。HCL同时考虑了事件级和序列级的噪声对比估计问题，此外，在序列级的噪声对比估计中，该工作没有使用耗时的Ogata方法，而是设计一种基于模型的采样方法来生成正负序列，将计算复杂度从 $O(N^2)$ 降低到 $O(N)$ 。



Motivation



Intensity function: Expected Instantaneous Happening Rate of Events

$$\lambda_u(t) = \frac{\sum_{i=1}^n dN_i(t) | \mathcal{H}_t}{dt}, \mathcal{H}_t = \{(t_i, c_i) | t_i < t, c_i \in C\} \quad (1)$$

Inspired by the EM algorithm, we can learn the TPP model $\{\lambda_u(t)\}_{u=1}^U$ by maximum likelihood estimation (MLE):

$$\max_{\theta} C(\theta; \mathcal{D}) = \max_{\theta} \sum_{u=1}^U \log \lambda_u(t; \theta) - \sum_{u=1}^U \int_0^T \lambda_u(t; \theta) dt \quad (2)$$

High risk of overfitting for imperfect (e.g., sparse, incomplete) event sequence.

Contributions & Future Work

- We present a hierarchical contrastive learning method for temporal point processes, which provides a new regularizer for the scheme of maximum likelihood estimation.
- The proposed method not only considers the event-level contrastive learning, but also design a simple but effective sequence-level contrastive loss.
- The contrastive learning mechanism, especially the sequence-level part is more efficient than the Ogata's thinning-based method. Beyond the scheme of MLE.
- We would like to combine the proposed method with other learning framework in the future, such as the Wasserstein GAN strategy and the reinforcement learning strategy. Additionally, we will try to find theoretical support for our method.

Our Method

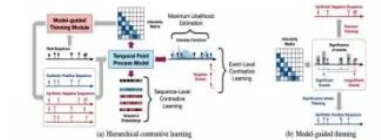


Figure 1: The scheme of our learning method. (a) An illustration of hierarchical contrastive learning. (b) An illustration of model-guided thinning.

In this formulation, the TPP model takes a timestamp t and the historical events before t as its input, and output three terms:

- The multivariate intensity vector at time t , $\lambda_u(t)$.
- The lower-triangular infectivity matrix for events, denoted as G and its element g_{ij} measures the contribution of the j event to the happening of the i event.
- The embedding vector of the historical event sequence till time t .

Hierarchical contrastive learning of TPPs

- Event-level contrastive loss

$$\mathcal{L}_{\text{event}}(\lambda(t)) = \log \frac{\lambda_u(t)}{\mu(t; \mathcal{H}_t)} + \sum_{c \in C} \log \left(1 - \frac{\lambda_u(t)}{\mu(t; \mathcal{H}_t)} \right) \quad (12)$$
- Sequence-level contrastive loss
 - Taking G as an input to obtain the significance of observed events

$$\theta = \{g_{ij}\}_{i,j=1}^K = \text{softmax}(G^{-1} \mathbf{1}) \quad (13)$$
 - Sequence-level contrastive loss:

$$\mathcal{L}_{\text{seq}}(\theta, \theta_{\text{pr}}; \{\mathbf{h}_{u,t}\}_{t=1}^T) = \frac{\mathcal{L}_{\text{event}}(\lambda(t; \theta))}{\log \frac{\exp(\theta^T \mathbf{e}_{e_{j_1}}) + \sum_{j=1}^K \exp(\theta^T \mathbf{e}_{e_{j_2}})}{\exp(\theta^T \mathbf{e}_{e_{j_1}}) + \sum_{j=1}^K \exp(\theta^T \mathbf{e}_{e_{j_2}})}} + \sum_{t=1}^T \log \left(1 - \frac{\exp(\theta^T \mathbf{e}_{e_{j_1}})}{\exp(\theta^T \mathbf{e}_{e_{j_1}}) + \sum_{j=1}^K \exp(\theta^T \mathbf{e}_{e_{j_2}})} \right) \quad (14)$$

Obtain a hierarchical contrastive learning (HCL) method to regularize the MLE of TPPs, our learning problem is

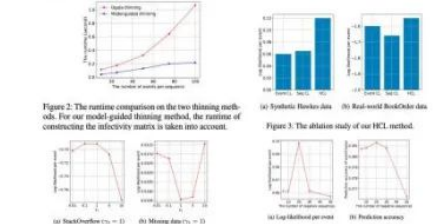
$$\max_{\theta} \frac{C(\theta; \mathcal{D})}{\text{Log likelihood}} + \gamma_1 \sum_{t=1}^T \mathcal{L}_{\text{event}}(\lambda(t; \theta)) + \gamma_2 \mathcal{L}_{\text{seq}}(\theta, \theta_{\text{pr}}; \{\mathbf{h}_{u,t}\}_{t=1}^T) \quad (15)$$

Experiments

Table 2: Comparison for various methods on learning TPPs from different datasets

Models	Data	Metrics	Methods									
			MLE+Reg	MLE+DFA	EM	EM+ASOM	MCE+TPP	MCE+HCL	HCL	HCL+ASOM	HCL+DFA	HCL+TPP
Hawkes	Log-Lik		-0.06 (0.05)	-0.32 (0.34)	-0.61 (1.11)	-0.22 (0.02)	-0.10 (0.00)	-0.04 (0.00)	-0.04 (0.00)	-0.04 (0.00)	-0.04 (0.00)	-0.04 (0.00)
		Type-Acc	0.38 (0.01)	0.38 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)	0.33 (0.01)
		Log-Lik	-0.04 (0.00)	-1.09 (0.00)	-1.38 (0.00)	-0.33 (0.00)	-0.33 (0.00)	-0.33 (0.00)	-0.33 (0.00)	-0.33 (0.00)	-0.33 (0.00)	-0.33 (0.00)
	Missing	Type-Acc	0.42 (0.00)	0.41 (0.01)	0.40 (0.01)	0.38 (0.01)	0.41 (0.01)	0.41 (0.01)	0.41 (0.01)	0.41 (0.01)	0.41 (0.01)	0.41 (0.01)
		Log-Lik	-2.60 (0.00)	-3.28 (0.00)	-3.54 (0.00)	-1.64 (0.00)	-1.64 (0.00)	-1.64 (0.00)	-1.64 (0.00)	-1.64 (0.00)	-1.64 (0.00)	-1.64 (0.00)
		Log-Lik	-0.37 (0.01)	-2.31 (0.12)	-2.96 (0.07)	-0.79 (0.01)	-0.74 (0.04)	-0.74 (0.04)	-0.74 (0.04)	-0.74 (0.04)	-0.74 (0.04)	-0.74 (0.04)
Backdoor	Log-Lik		-0.45 (0.02)	0.43 (0.02)	0.49 (0.01)	0.49 (0.01)	0.49 (0.01)	0.49 (0.01)	0.49 (0.01)	0.49 (0.01)	0.49 (0.01)	
		Type-Acc	0.45 (0.00)	0.43 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	
		Log-Lik	-0.64 (0.20)	-10.97 (2.20)	-	-0.89 (0.11)	-0.92 (0.07)	-0.92 (0.07)	-0.92 (0.07)	-0.92 (0.07)	-0.92 (0.07)	
	Random	Type-Acc	0.60 (0.01)	0.59 (0.00)	0.58 (0.00)	0.61 (0.01)	0.61 (0.01)	0.61 (0.01)	0.61 (0.01)	0.61 (0.01)	0.61 (0.01)	
		Log-Lik	0.11 (0.01)	-1.23 (0.47)	-0.64 (0.13)	0.01 (0.01)	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)	0.12 (0.01)	
		Type-Acc	0.38 (0.01)	0.36 (0.00)	0.34 (0.01)	0.34 (0.01)	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)	
THP	Log-Lik		-0.47 (0.01)	-1.25 (0.21)	-0.70 (0.16)	-1.08 (0.18)	-0.50 (0.02)	-0.54 (0.00)	-0.54 (0.00)	-0.54 (0.00)	-0.54 (0.00)	
		Type-Acc	0.41 (0.01)	0.37 (0.00)	0.41 (0.00)	0.40 (0.00)	0.42 (0.01)	0.41 (0.00)	0.41 (0.00)	0.41 (0.00)	0.41 (0.00)	
		Log-Lik	-1.69 (0.31)	-4.80 (1.79)	-1.84 (0.41)	-1.70 (0.41)	-1.80 (0.30)	-1.80 (0.30)	-1.80 (0.30)	-1.80 (0.30)	-1.80 (0.30)	
	Backdoor	Type-Acc	0.62 (0.00)	0.62 (0.01)	0.62 (0.01)	0.62 (0.01)	0.63 (0.01)	0.63 (0.01)	0.63 (0.01)	0.63 (0.01)	0.63 (0.01)	
		Log-Lik	-0.77 (0.00)	-2.31 (0.12)	-2.96 (0.07)	-0.89 (0.10)	-0.77 (0.02)	-0.77 (0.02)	-0.77 (0.02)	-0.77 (0.02)	-0.77 (0.02)	
		Type-Acc	0.43 (0.00)	0.42 (0.02)	0.40 (0.01)	0.40 (0.01)	0.39 (0.02)	0.39 (0.02)	0.39 (0.02)	0.39 (0.02)	0.39 (0.02)	
Random	Log-Lik	-7.95 (0.35)	-2.14 (0.46)	-	-10.20 (0.51)	-7.51 (0.29)	-7.51 (0.29)	-7.51 (0.29)	-7.51 (0.29)	-7.51 (0.29)		
	Type-Acc	0.53 (0.00)	0.50 (0.01)	0.54 (0.00)	0.53 (0.01)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)		
	Type-Acc	0.53 (0.00)	0.50 (0.01)	0.54 (0.00)	0.53 (0.01)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)	0.54 (0.00)		

“-” means the learning method fails to coverage. The best results are bolded. In each cell, the averaged performance is shown, and the parentheses contain the standard deviation.



Authors:

Qingmei Wang, Minjie Cheng, Shen Yuan, Hongteng Xu

Corresponding: hongtengxu@ruc.edu.cn

王庆梅同学在大会会场展示的学术海报 (Poster)

报告结束后，王庆梅在被围着问问题时感受到了这个工作的意义，而与其他国家研究人员当面交流对她来也是一种特别的体验，备受鼓舞的同时也深切地感受到自己的不足，她说：“通过这次参会，能够感觉到自己的各方面能力都有很大提升空间，离顶尖的科研还有很长的路要走”。



王庆梅同学（左5）与参会学者留念

通过这篇论文的“洗礼”，王庆梅表示，自己对科研工作有了更深入的认识。春风大雅能容物，秋水文章不染尘。论文从idea、代码、实验、写作到报告的slides，都离不开导师许洪腾老师的辛勤指导。王庆梅也表示，很幸运能够在疫情防控进一步“放开”后的第一时间，在导师支持下得以赴美现场参会，长途飞行中邻座的菲律宾夫妇全程照顾，会议现场与世界各地的学者们席地而坐一同讨论，报告环节精彩的talk，被国外友人带着融入新环境，她的工作也收获了建设性意见，这一切都让科研路上的她受益匪浅。

AAAI2023 共收到投稿8777 篇，录用 1721 篇，接收率仅为 19.6%。中国人民大学高瓴人工学院共10篇论文被录用。



En
(<http://ai.ruc.edu.cn/english/index.htm>)



教师入口
(<https://gsai.ruc.edu.cn/index/user/login.html>)

友情链接

- [中国人民大学信息学院 \(http://info.ruc.edu.cn/\)](http://info.ruc.edu.cn/)
- [中国科学技术协会 \(http://www.cast.org.cn/\)](http://www.cast.org.cn/)
- [中国外文出版发行事业局 \(http://www.cipg.org.cn/\)](http://www.cipg.org.cn/)

联系

ai@ruc.edu.cn | 86-10-62511257

北京市海淀区中关村大街59号中国人民大学

copyright 2021 中国人民大学高瓴人工智能学院

关注我们

