CCMT 2020

第十六届全国机器翻译大会

The 16th China Conference on Machine Translation

2020 年 10 月 10 日至 12 日线上召开

会议地址: https://zoom.com.cn/j/66820277556

(ID: 66820277556)

主 办 方：中国中文信息学会

承 办 方：内蒙古大学

社区支持：智源社区

赞 助 商：

目录

前言

第十六届全国机器翻译大会（The 16th China Conference on Machine Translation, CCMT 2020)将于 2020 年 10 月 10 日至 12 日在线上举行。本次会议由中国中文信息学会主办，内蒙古大学承办。

CCMT 旨在为国内外机器翻译界同行提供一个交互平台，加强国内外同行的学术交流，召集各路专家学者针对机器翻译的理论方法、应用技术和评测活动等若干关键问题进行深入的研讨，为促进中国机器翻译事业的发展，起到积极的推动作用。会议已连续成功召开了十五届（前十四届名为全国机器翻译研讨会 CWMT）。其中，共组织过九次机器翻译评测，一次开源系统模块开发（2006）和两次战略研讨（2010、2012）。这些活动对于推动我国机器翻译技术的研究和开发产生了积极而深远的影响。因此，CCMT 已经成为我国自然语言处理领域颇具影响的学术活动。

除学术论文报告外，本次会议将会邀请国内外知名专家进行特邀报告，面向学生和青年学者举行专题讲座，邀请学界和产业界专家举行专题讨论会，面向研究者和用户进行系统展示等，通过丰富多彩的形式和与会者互动探讨机器翻译最炽热的研究论点，揭示机器翻译最前沿的蓝图。同时，CCMT2020 继续组织机器翻译评测，包括机器翻译双语翻译（汉英、英汉、维汉、藏汉和蒙汉），多语言翻译（汉、日、英），语音翻译（汉英）和翻译质量自动评估（汉英、英汉）等多个任务。会上也会就评测工作进行学术交流和专题讨论。

本届会议为期三天。会议热忱欢迎高校、科研机构和 IT 企业的积极参与！愿机器翻译同道，携起手来，共同为机器翻译的研究做出贡献！愿学术界和企业界，能够有更多的交流和合作！

主办单位：中国中文信息学会

承办单位：内蒙古大学

会议组织：

大会主席：

周　明（微软亚洲研究院）

程序委员会主席：

Andy Way（都柏林城市大学）　　李军辉（苏州大学）

评测委员会主席：

杨沐昀（哈尔滨工业大学）

组织委员会主席：

高光来（内蒙古大学）　　侯宏旭（内蒙古大学）

讲习班主席：

刘树杰（微软亚洲研究院）　　肖　桐（东北大学）

学生论坛主席：

黄　非（阿里巴巴）　　黄书剑（南京大学）

前沿趋势论坛主席：

刘　洋（清华大学）　　张家俊（中科院自动化所）

研讨主席：

涂兆鹏（腾讯）　　苏劲松（厦门大学）

出版主席：

杨雅婷（中科院新疆理化所）

赞助主席：

李长亮（金山软件）　　冯　洋（中科院计算所）

宣传主席：

冯　冲（北京理工大学）　　何中军（百度）　　黄国平（腾讯）

日程

中国中文信息学会《前沿技术讲习班》

2020/10/10

09:00 - 11:00 **神经机器翻译中的文本生成方法（周龙）**

14:00 - 16:00 **深度文本生成模型的前沿进展（周浩）**

第十六届全国机器翻译大会

2020/10/11

08:30 - 09:00 **开幕式（主持人：侯宏旭）**

08:30 - 08:40 学会领导孙乐老师致辞

08:40 - 08:45 专委会主任张民老师致辞

08:45 - 08:50 大会主席周明老师致辞

08:50 - 08:55 程序委员会主席李军辉老师介绍大会组织情况

09:00 - 10:00 **特邀报告 1 (主持人：李军辉)**

Graham Neubig, Beyond Likelihood: New Training Objectives for Neural Machine Translation

10:00 - 10:15 **中间休息**

10:15 - 11:30 **英文论文（主持人：刘乐茂）**

10:15 - 10:30 Transfer Learning for Chinese-Lao Neural Machine Translation with Linguistic Similarity

10:30 - 10:45 MTNER: A Corpus for Mongolian Tourism Named Entity Recognition

10:45 - 11:00 Neural Machine Translation based on Back-Translation for Multilingual Translation Evaluation Task

11:00 - 11:15 YuQ: A Chinese-Uyghur Medical-domain Neural Machine Translation Dataset Towards Knowledge-driven

11:15 - 11:30 Quality Estimation for Machine Translation with Multi-granularity Interaction

13:30 - 16:15 **评测论文（主持人：杨沐昀）**

13:30 - 13:50 第 16 届全国机器翻译大会 CCMT 2020 评测报告

13:50 - 14:10 Description and Findings of OPPO's Machine Translation Systems for CCMT 2020

14:10 - 14:20 Tsinghua University Neural Machine Translation Systems for CCMT 2020

14:20 - 14:30 低资源神经机器翻译的关键技术研究

14:30 - 14:45 北京航空航天大学 CCMT2020 翻译质量评测技术报告

14:45 - 15:00 NJUNLP's Machine Translation System for CCMT-2020 Uighur->Chinese Translation Task

NJUNLP's Submission for CCMT20 Quality Estimation Task

15:00 - 15:10 **中间休息**

15:10 - 15:20 中国科学技术信息研究所 CCMT'2020 评测技术报告

15:20 - 15:30 第十六届机器翻译研讨会厦门大学评测报告

15:30 - 15:45 BJTU's Submission to CCMT 2020 Quality Estimation Task

BJTU's Submission to CCMT 2020 Multilingual Translation Evaluation Task

15:45 - 15:55 Tencent Submissions for the CCMT 2020 Quality Estimation Task

15:55 - 16:05 Transformer-based Unified Neural Network for quality estimation and Transformer-based re-decoding model for machine translation

16:05 - 16:15 融合数据增强与多样化解码的神经机器翻译

16:30 - 18:00 **Panel 1 当前机器翻译的瓶颈 (主持人：涂兆鹏；嘉宾：李沐、刘群、刘洋、朱靖波)**

19:30 - 21:10 **短报告论文 (主持人：李茂西；冯冲)**

19:30 - 19:40 维汉神经机器翻译代词性别偏见改进方法

19:40 - 19:50 一种简单的神经机器翻译数据扩充方法

19:50 - 20:00 融合篇章上下文有效识别的篇章级机器翻译

20:00 - 20:10 基于枢轴的汉越联合训练神经机器翻译

20:10 - 20:20 基于掩码机制的非自回归神经机器翻译

20:20 - 20:30 **中间休息**

20:30 - 20:40 神经机器翻译词级别正则化技术研究

20:40 - 20:50 神经机器翻译军事领域英译汉评测及译后编辑研究

20:50 - 21:00 基于联合选择机制的篇章级机器翻译

21:00 - 21:10 基于古籍白话译本的古文机器翻译研究

2020/10/12

09:00 - 10:00 **特邀报告 2 (主持人：张家俊)**

韦福如，自然语言预训练模型进展

10:15 - 11:45 **学生论坛 (主持人：黄非、黄书剑；嘉宾：褚晨翚、顾佳涛、黄轩成、李垠桥、刘宇宸、邵晨泽、郑在翔)**

13:30 - 14:30 **中文论文 (主持人：段湘煜)**

13:30 - 13:45 同源语料增强的印尼语汉语神经机器翻译

13:45 - 14:00 融合语言学知识的神经机器翻译研究进展

14:00 - 14:15 基于数据增强技术的神经机器翻译

14:15 - 14:30 基于迭代知识精炼的对偶学习蒙汉机器翻译

14:45 - 15:30 **前沿趋势论坛 (主持人：刘洋；嘉宾：涂兆鹏)**

15:45 - 17:15 **Panel 2 多模态机器翻译 (主持人：苏劲松；嘉宾：李响、骆卫华、王瑞、肖桐)**

17:30 - 17:50 **闭幕式**

前沿技术讲习班

Tutorial 1 神经机器翻译中的文本生成方法

10 月 10 日 9:00 – 11:00

主　持　人：刘树杰（微软亚洲研究院）

讲习班专家：周　龙（微软亚洲研究院）

内容简介:

采用编码器-解码器框架的神经机器翻译是目前主流的机器翻译模型，如何更好、更快、更准地生成目标语言单词是业界关注的焦点。在神经机器翻译序列生成中，自回归模型采用自左往右的解码方式，其限制了其对未来信息的开发和利用；非自回归模型通过并行计算加快了推理速度，但面临着质量下降等问题；双向推断模型采用从左到右和从右到左同步解码的方式，能有效缓解上述问题。这个报告将首先回顾自回归神经机器翻译的文本生成方法，然后介绍非自回归神经机器翻译近年来的研究进展，最后着重介绍双向神经机器翻译的基本思想与典型应用。

周龙 博士简介:

周龙，微软亚洲研究院自然语言计算组研究员，于中科院自动化所模式识别国家重点实验室获得博士学位。研究方向为自然语言处理，机器翻译，代码智能，自然语言生成等。在国际著名期刊和会议 AIJ、TACL、ACL、EMNLP、AAAI、IJCAI 等发表论文十余篇。曾在汉语浅层篇章分析国际评测 CoNLL-2016、全国机器翻译评测 CWMT-2017、国际口语机器翻译评测 IWSLT-2020 多次斩获第一，获得国际自然语言处理与中文计算会议 NLPCC-2017 最佳论文奖。

前沿技术讲习班

Tutorial 2 深度文本生成模型的前沿进展

10 月 10 日 14:00 – 16:00

主　持　人：肖桐（东北大学）

讲习班专家：周浩（字节跳动人工智能实验室）

内容简介:

文本生成技术是自然语言处理中的一项基础技术，在机器写稿、机器翻译、对话、搜索、在线广告等产品上有很多应用。本次讲座将围绕三个方面介绍文本生成中的深度生成模型。一是序列到序列的生成，包括最新的 Transformer 模型，最新的非自回归的文本生成模型以及它在各种文本生成中的改进。二是比序列生成有更多优势的深度隐变量模型，包括生成与编码结合的变分自编码模型(VAE)与对抗生成网络（GAN）。VAE 要对离散的文本序列学出光滑连续的隐空间，生成时可以从隐空间采样。而 GAN 可以附加一个与任务有关的判别器，可以生成与最终任务更相关的文本。第三类是可控贝叶斯方法，可以生成更大多样性和可解释性的文本。最后，我们讲介绍实际场景中的一些应用，例如数据到文本的生成，问题生成等。

周浩博士简介：

周浩博士是字节跳动 AI Lab 的高级研究员。他于 2017 年获得南京大学博士学位，并于 2019 年获得中国人工智能协会优秀博士论文。周浩博士的研究领域涉及机器学习及其在自然语言处理中的应用。他最近的研究集中在面向自然语言处理的深度生成模型。周浩博士是 ACL、EMNLP、IJCAI、AAAI、NIPS 的程序委员会委员。至今为止，在 ACL, EMNLP, NAACL, TACL, AAAI, IJCAI, NIPS、JAIR 等期刊及会议上发表论文 30 余篇。

特邀报告 1

Beyond Likelihood: New Training Objectives for Neural Machine Translation

10 月 11 日 9:00 – 10:00

主 持 人：李军辉（苏州大学）

特邀专家：Graham Neubig（Carnegie Mellon University）

报告简介：

Maximum likelihood estimation (MLE) is the workhorse of training NMT models, but has a number of issues such as disregard for the actual test-time decoding algorithm, lack of consideration of the inherent ambiguity in the generation of translation references, and inability to deal with training/test-time data distribution mismatch. In this talk I will discuss some new developments in training objectives for machine translation that move beyond the standard paradigm of training with maximum likelihood estimation. Specifically I will first discuss a method that explicitly trains models to maximize the semantic similarity between MT outputs and the human-provided references. Second, I will discuss methods that automatically learn to appropriately weight training data to maximize test-time performance, including in multilingual learning settings.

特邀专家简介：

Graham Neubig is an associate professor at the Language Technologies Institute of Carnegie Mellon University. His work focuses on natural language processing, specifically multi-lingual models that work in many different languages, and natural language interfaces that allow humans to communicate with computers in their own language. Much of this work relies on machine learning, and he is also active in developing methods and algorithms for machine learning over natural language data. He publishes regularly in the top venues in natural language processing, machine learning, and speech, and his work has won awards at EMNLP 2016, EACL 2017, and NAACL 2019.

特邀报告 2

自然语言预训练模型进展

10 月 12 日 9:00 – 10:00

主 持 人：张家俊 (中科院自动化所)

特邀专家：韦福如（微软亚洲研究院）

报告简介:

大规模预训练语言模型很大程度上改变了自然语言处理模型的研究和开发范式，在工业界和学术界都引起了广泛的关注。本报告将对现有的语言模型预训练工作进行总结和比较，然后介绍面向自然语言理解和生成任务的统一预训练语言模型 UniLM 以及多语言预训练模型 InfoXLM，并就未来面临的挑战和进一步的研究方向进行讨论和展望。

特邀专家简介：

韦福如博士，微软亚洲研究院自然语言计算组首席研究员，长期从事自然语言处理的基础研究和技术创新。在自然语言处理领域重要会议和期刊发表论文 100 余篇，被引用 9000 余次，多项研究成果转化到微软重要产品中。入选 2017 年《麻省理工科技评论》中国区“35 岁以下科技创新 35 人”榜单，2019 年第六届世界互联网大会“领先科技成果”奖。近年来，团队开发的预训练模型（UniLM, InfoXLM, LayoutLM, MiniLM 等）被广泛应用于微软的产品中。

Oral 1 英文论文

10 月 11 日 10:15 – 11:30

主持人：刘乐茂（腾讯）

10:15 – 10:30 Transfer Learning for Chinese-Lao Neural Machine Translation with Linguistic Similarity
10:30 – 10:45 MTNER: A Corpus for Mongolian Tourism Named Entity Recognition
10:45 – 11:00 Unsupervised Machine Translation Quality Estimation in Black-box Setting
11:00 – 11:15 YuQ: A Chinese-Uyghur Medical-domain Neural Machine Translation Dataset Towards Knowledge-driven
11:15 – 11:30 Quality Estimation for Machine Translation with Multi-granularity Interaction

Introduction:

Transfer Learning for Chinese-Lao Neural Machine Translation with Linguistic Similarity
Zhiqiang Yu, Zhengtao Yu, Yuxin Huang, Junjun Guo, Zhenhan Wang and Zhibo Man
As a typical low-resource language pair, besides severely limited by the scale of parallel corpus, Chinese-Lao language pair also has considerable linguistic differences, resulting in poor performance of Chinese-Lao NMT task. However, there are considerable cross-lingual similarities between Thai-Lao languages. According to these features, we propose a novel NMT approach. We first train Chinese-Thai and Thai-Lao NMT models wherein Thai is treated as pivot language. Then the transfer learning strategy is used to extract the encoder and decoder respectively from the two trained models. Finally, the encoder and decoder are combined into a new model and then fine-tuned based on a small-scale Chinese-Lao parallel corpus. We argue that the pivot language Thai can deliver more information to Lao decoder via linguistic similarity and help improve translation quality of proposed transfer-based approach. Experimental results demonstrate that our approach achieves significant improvement on Chinese-Lao translation task.

MTNER: A Corpus for Mongolian Tourism Named Entity Recognition
Xiao Cheng, Weihua Wang, Feilong Bao, Guanglai Gao
Name Entity Recognition is the essential tool for machine translation. Traditional Named Entity Recognition focuses on the person, location and organization names. However, there is still a lack of data to identify travel-related named entities, especially in Mongolian. In this paper, we introduce a newly corpus for Mongolian Tourism Named Entity Recognition (MTNER), consisting of 16,000 sentences annotated with 18 entity types. We trained in-domain BERT representations with the 10GB of unannotated Mongolian corpus, and trained a NER model based on the BERT tagging model with the newly corpus. Which achieves an overall 82.09 F1 score on Mongolian Tourism

Named Entity Recog-nition and lead to an absolute increase of +3.54 F1 score over the traditional CRF Named Entity Recognition method.

Unsupervised Machine Translation Quality Estimation in Black-box Setting

Hui Huang, Hui Di , Jin'an Xu1, Kazushige Ouchi and Yufeng Chen

This paper presents the systems developed by Beijing Jiaotong University for the CCMT 2020 multilingual translation evaluation task. For this translation task, we need to build a Japanese-English translation system based on only Japanese-Chinese and English-Chinese data. Our method mainly relies on synthetic data generated by back translation. We implemented three different architectures, namely Transformer-big, Transformer-base and Dynamic-Conv. We also implemented multi-model ensemble technique to further boost the final result. Experiments show that our machine translation system achieved high accuracy without relying on any bilingual training data.

YuQ: A Chinese-Uyghur Medical-domain Neural Machine Translation Dataset Towards Knowledge-driven

Qing Yu, Zhe Li, Jiabao Sheng, Jing Sun and Wushour Slamu

Recent advances of deep learning have been successful in delivering state-of-the-art performance in medical analysis, However, deep neural networks (DNNs) require a large amount of training data with a high-quality annotation which is not available or expensive in the field of the medical domain. The research of medical domain neural machine translation(NMT) is largely limited due to the lack of parallel sentences that consist of medical domain background knowledge annotations. To this end, we propose a Chinese-Uyghur NMT knowledge-driven dataset, YuQ, which refers to a ground medical domain knowledge graphs. Our corpus contains 65K parallel sentences from the medical domain and 130K utterances. By introduce medical domain glossary knowledge to the training model, we can win the challenge of low translation accuracy in Chinese-Uyghur machine translation professional terms. We provide several benchmark models. Ablation study results show that the models can be enhanced by introducing domain knowledge.

Quality Estimation for Machine Translation with Multi-granularity Interaction

Ke Tian and Jiajun Zhang

Quality estimation (QE) for machine translation is the task of evaluating the translation system quality without reference translations. By using the existing translation quality estimation methods, researchers mostly focus on how to extract better features but ignore the translation oriented interaction. In this paper, we propose a QE model for machine translation that integrates multi-granularity interaction on the word and sentence level. On the word level, each word of the target language sentence interacts with each word of the source language sentence and yields the similarity, and the and entropy of the similarity distribution are employed as the word-level interaction score. On the sentence level, the similarity between the source and the target language translation is calculated to indicate the overall translation quality. Finally, we combine the word-level features and the sentence-level features with different weights. We perform thorough experiments with detailed studies and analyses on the English-German dataset in the WMT19 sentence-level QE task, demonstrating the effectiveness of our method.

Oral 2 评测论文

10 月 11 日 13:30 – 16:15

主持人：杨沐昀（哈尔滨工业大学）

13:30 – 13:50 **第 16 届全国机器翻译大会 CCMT 2020 评测报告**

13:50 – 14:10 **Description and Findings of OPPO's Machine Translation Systems for CCMT 2020**

14:10 – 14:20 **Tsinghua University Neural Machine Translation Systems for CCMT 2020**

14:20 – 14:30 **低资源神经机器翻译的关键技术研究**

14:30 – 14:45 **北京航空航天大学 CCMT2020 翻译质量评测技术报告**

14:45 – 15:00 **NJUNLP's Machine Translation System for CCMT-2020 Uighur->Chinese Translation Task**

NJUNLP's Submission for CCMT20 Quality Estimation Task

15:00 – 15:10 **中间休息**

15:10 – 15:20 **中国科学技术信息研究所 CCMT'2020 评测技术报告**

15:20 – 15:30 **第十六届机器翻译研讨会厦门大学评测报告**

15:30 – 15:45 **BJTU's Submission to CCMT 2020 Quality Estimation Task**

Neural Machine Translation based on Back-Translation for Multilingual Translation Evaluation Task

15:45 – 15:55 **Tencent Submissions for the CCMT 2020 Quality Estimation Task**

15:55 – 16:05 **Transformer-based Unified Neural Network for quality estimation and Transformer-based re-decoding model for machine translation**

16:05 – 16:15 **融合数据增强与多样化解码的神经机器翻译**

Introduction:

第 16 届全国机器翻译大会 CCMT 2020 评测报告

杨沐昀，唐煜，汪嘉怿，何中军，孟庆晔，邱瑞玲

本文介绍了 CCMT 2020 机器翻译评测的评测方案、评测数据、参评情况与评测结果。与上届评测（CWMT 2019）相比，CCMT2020 的评测方案包括如下变化：新增语料过滤评测任务；汉英英汉翻译增加新冠相关测试数据；增加了本次评测系统与 CCMT2019 测试集上的性能对比。本次评测还开通年度在线评测，使用的是 CCMT2019 的测试集。本文给出了所有离线参评系统的匿名评测结果，以及截止到会议当前的离线参评系统的匿名评测结果，并在附件中提供了各参评主系统的简要描述。

Description and Findings of OPPO's Machine Translation Systems for CCMT 2020

Tingxun Shi, Qian Zhang, Xiaoxue Wang, Xiaopu Li, Zhengshan Xue and Jie Hao

This paper demonstrates our machine translation systems for the CCMT 2020, which is composed of three parts. Each part respectively focuses on English-Chinese bi-direction translation, Japanese-Chinese-English multi-lingual translation (patent domain), and Chinese minority languages to Mandarin Chinese translation. In each part, we will demonstrate our work on data pre-processing, model training as well as the application of general techniques, such as back-translation, ensemble and reranking. During our experiments, we surprisingly found that simply applying different Chinese word segmentation tools on low-resource corpora could bring obvious benefit across different tasks, and we will separate an independent section to discuss this finding. Among the 7 directions we participated in, we ranked the first in 6 tasks \footnote{For the corpus filtering task, we ranked first in the 500 million words sub-task} and the second for the rest.

Tsinghua University Neural Machine Translation Systems for CCMT 2020

Gang Chen, Shuo Wang, Xuancheng Huang, Zhixing Tan, Maosong Sun and Yang Liu

This paper describes the neural machine translation system of Tsinghua University for the bilingual translation task of CCMT 2020. We participated in the Chinese->English translation tasks. Our systems are based on Transformer architectures and we verified that deepening the encoder can achieve better results. All models are trained in a distributed way. We employed several data augmentation methods, including knowledge distillation, back-translation, and domain adaptation, which are all shown to be effective to improve translation quality. Distinguishing original text from translationese can lead to better results when performing domain adaptation. We found model ensemble and transductive ensemble learning can further improve the translation performance over the individual model. In both Chinese->English and English->Chinese translation tasks, our systems achieved the highest case-sensitive BLEU score among all submissions.

低资源神经机器翻译的关键技术研究

张文博，张新路，杨雅婷，董瑞

本文描述了中国科学院新疆理化技术研究所参加第 16 届全国机器翻译大会（CCMT2020）翻译评测任务总体情况以及采用的技术细节。在评测中，中国科学院新疆理化技术研究所提交了两个翻译任务，分别是蒙汉日常用语机器翻译和维汉新闻领域机器翻译；本文使用所提供的平行数据加上大量的回译单语数据得到伪平行语料来训练最先进的神经机器翻译系统。然后，本文主要采用了目前已被证明最有效的技术，如微调和模型集成来改善翻译效果，最后在该数据集上进行了详细的对比与分析。

北京航空航天大学 CCMT2020 翻译质量评测技术报告

张文超，巢文涵，黄彦

本文介绍了我们参加 CCMT2020 翻译质量评估任务所提交的系统，包括句子级和单词级。而在此次评 测任务中，本研究组所使用的系统是基于预测器-估计器[1]的体系结构，该结构主要是复现了 CWMT2019 质量评估任务中小牛翻译的模型框架 。对于预测器，采用深层 Transformer 以及 Transformer-DLCL (先前层的动态线性组合)作为特征提取模型。并且使用从左到右和从右到左的两个翻译模型来获得双向翻译 的信息。对于估计器，

使用 2 层双向 GRU 来预测句子级任务的 HTER 分数或单词级任务的 OK / BAD 标签。 我们先用大规模双语数据对预测器进行预训练，然后将预测器和估计器与 QE 任务数据一起进行联合训练。 在本文的其余部分介绍了本组参加评测任务的系统框架、处理方法和评测结果。

NJUNLP's Machine Translation System for CCMT-2020 Uighur->Chinese Translation Task

Dongqi Wang, Zihan Liu, Qingnan Jiang, Zewei Sun, Shujian Huang and Jiajun CHEN

This paper describes our submitted systems for CCMT-2020 shared translation tasks. We build our neural machine translation system based on Google's Transformer architecture. We also employ some effective techniques such as back translation, data selection, ensemble translation, fine-tuning and reranking to improve our system.

Neural Machine Translation based on Back-Translation for Multilingual Translation Evaluation Task

Qu Cui, Xiang Geng, Shujian Huang and Jiajun CHEN

Quality Estimation is a task to predict the quality of translations without relying on any references. QE systems are based on neural features but suffer from the limited size of QE data. The best models nowadays transfer bilingual knowledge from parallel data to QE tasks. However, the distribution between parallel data and QE data is different which may lead to that the value of parallel data can not be used for best. More specifically, there are no errors in parallel translations while there may be more than one error in the translations of QE data. To alleviate this problem, we propose a model which will mask some tokens at the target side on parallel data but still need to predict every target token. And based on this model, we propose a variant model that uses a masked language model at the target side to obtain deep bi-directional information. Besides, we also try different ensemble methods to get better performance of the CCMT20 Quality Estimation Task. Our system finally won second place in the ZH-EN language pair and third place in the EN-ZH language pair.

中国科学技术信息研究所 CCMT'2020 评测技术报告

刘文斌，魏家泽，吴振峰，潘优，何彦青

本文详细介绍了中国科学技术信息研究所（ISTIC）参加第十六届全国机器翻译大会机器翻译评测（CCMT'2020）翻译评测任务的总体情况。ISTIC 共参加了 CCMT'2020 中的六个任务，分别是汉英新闻领域的翻译评测、蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测、专利领域的日、汉、英多语言翻译评测以及汉英平行语料过滤任务.本次评测采用谷歌 Transformer 神经网络机器翻译架构。在数据预处理方面，针对评测方发布的数据，采取多种不同语料过滤方法减少语料噪声以提高训练语料的质量。同时为了进一步利用评测方发布的单语数据，在蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测任务上，使用回译方法来构建伪平行语料，补充神经机器翻译模型的训练集。在译文输出过程中，采用了模型平均和集成解码的策略，最后利用译文重排序策略给出最终的译文。在实验中，对比了系统在各任务上不同设置下的表现，并对实验结果进行了分析。

第十六届机器翻译研讨会厦门大学评测报告

张国成，王颖敏，钟恩俊，江秋怡，江舫，章栋，朱宏康，陈毅东，史晓东

本文介绍了厦门大学参加第十六届全国机器翻译研讨会的汉英平行语料过滤任务评测的参评系统情况。在此次评测中，本文参评系统主要利用规则方法对噪音句对进行严格的过滤；同时也设计了五种启发式方法，从不同侧重点对噪音句对平行程度进行度量，尤其是基于单词的编辑距离和基于双语预训练模型的马氏距离在过滤优质平行数据上有良好的表现；最后，我们对表现优异的方法，按照加法和乘法两种方式进行加权融合。最终，本文提交的系统综合排名第二。

BJTU's Submission to CCMT 2020 Quality Estimation Task

Hui Huang, Jinan Xu, Wenjing Zhu, Yufeng Chen and Rui Dang

This paper presents the systems developed by Beijing Jiaotong University for the CCMT 2020 quality estimation task. In this paper, we propose an effective method to utilize pretrained language models to improve the performance of QE. Our model combines three popular pretrained models, which are Bert, XLM and XLM-R, to create a very strong baseline for both sentence-level and word-level QE. We tried different strategies, including further pretraining for bilingual input, multi-task learning for multi-granularities and weighted loss for unbalanced word labels. To generate more accurate prediction, we performed model ensemble for both granularities. Experiment results show high accuracy on both directions, and outperform the winning system of last year on sentence level, demonstrating the effectiveness of our proposed method.

BJTU's Submission to CCMT 2020 Multilingual Translation Evaluation Task

Siyu Lai, Yueting Yang, Jin'an Xu, Yufeng Chen and Hui Huang

This paper presents the systems developed by Beijing Jiaotong University for the CCMT 2020 multilingual translation evaluation tasks. For this translation task, we need to build a Japanese-English translation system based on only Japanese-Chinese and English-Chinese data. Our method mainly relies on synthetic data generated by back translation. We implemented three different architectures, namely Transformer-big, Transformer-base and Dynamic-Conv. We also implemented multi-model ensemble technique to further boost the final result. Experiments show that our machine translation system achieved high accuracy without relying on any bilingual training data.

Tencent Submissions for the CCMT 2020 Quality Estimation Task

Zixuan Wang, Xiaoli Wang, Xinjie Wen, Ruichen Wang and Qingsong Ma

This paper presents our submissions to CCMT 2020 Quality Estimation (QE) sentence-level task for both Chinese-to-English (ZH-EN) and English-to-Chinese (EN-ZH). We propose new methods based on the predictor-estimator architecture. For the predictor, we propose XLM-predictor and Transformer-predictor. XLM-predictor novelly produces two kinds of contextual token representation, i.e., mask-XLM and non-mask-XLM. For the estimator, both RNN-estimator and Transformer-estimator are conducted and two novel strategies, i.e. top K strategy and multi-head attention strategy, are proposed to enhance the sentence feature representation. We also propose new effective ensemble technique for sentence-level predictions.

Transformer-based Unified Neural Network for quality estimation and Transformer-based re-decoding model for machine translation

Cong Chen, Zong Qinqin, Qi Luo, Lianqiu Bai and Maoxi Li

In this paper, we describe our submitted system for CCMT2020 sentence-level quality estimation subtasks and machine translation subtasks. We propose two models: (i) a Transformer-based unified neural network for the quality estimation of machine translation, which consists of a Transformer bottleneck layer and a bidirectional long short-term memory network that are jointly optimized and fine-tuned for quality estimation, and (ii) a Transformer-based re-decoding model for machine translation, which takes the generated machine translation outputs as the approximate contextual environment of the target language and synchronously re-decodes each token in the machine translation outputs. Experimental results on the development set show that the proposed approaches outperform the baseline systems.

融合数据增强与多样化解码的神经机器翻译

张一鸣，刘俊鹏，宋鼎新，黄德根

基于神经机器翻译模型 Transformer，提出一种融合数据增强技术和多样化解码策略的方法来提高机器翻译的性能。首先，对训练语料进行预处理和泛化，提高语料质量并缓解词汇稀疏的现象；其次，通过 back-translation 技术构造伪双语数据，扩充双语平行语料以增强模型；最后，在解码阶段融合检查点平均、模型集成和重打分策略以提高译文质量。CCMT2020 中英新闻领域翻译任务的实验结果显示，改进后的方法较 baseline 的 BLEU 值取得了 4.89%的提升。

Oral 3 短报告论文

---

10 月 11 日 19:30 – 21:10

主持人：李茂西（江西师范大学）、冯冲（北京理工大学）

19:30 – 19:40 维汉神经机器翻译代词性别偏见改进方法
19:40 – 19:50 一种简单的神经机器翻译数据扩充方法
19:50 – 20:00 融合篇章上下文有效识别的篇章级机器翻译
20:00 – 20:10 基于枢轴的汉越联合训练神经机器翻译
20:10 – 20:20 基于掩码机制的非自回归神经机器翻译
20:20 – 20:30 中间休息
20:30 – 20:40 神经机器翻译词级别正则化技术研究
20:40 – 20:50 神经机器翻译军事领域英译汉评测及译后编辑研究
20:50 – 21:00 基于联合选择机制的篇章级机器翻译
21:00 – 21:10 基于古籍白话译本的古文机器翻译研究

Introduction:

维汉神经机器翻译代词性别偏见改进方法

史学文，黄河燕，鉴萍，唐翼琨

在利用神经机器翻译进行维吾尔语到汉语的翻译时，由于维吾尔语代词不区分性别，这给翻译模型在汉语端使用正确的代词带来了很大的挑战。另一方面，由于训练数据本身在不同性别的代词使用场景上频率的偏差，神经机器翻译倾向于输出阳性代词而不是更恰当的代词。为缓解上述问题，本文先利用汉语单语语料构造伪平行数据以扩展原训练集，缓解训练集本身的代词不平衡问题；之后，分别提出引入性别标记和翻译、性别预测联合建模两种方法，将代词性别预测目标显式地融入神经机器翻译模型的训练过程。我们在多个维汉翻译测试集上进行了实验验证，结果表明，我们提出的方法相对于基线系统，在不影响翻译总体效果的情况下缓解了神经机器翻译输出结果的性别偏见问题，在代词性别预测的精度上也有显著提升。

一种简单的神经机器翻译数据扩充方法

刘志东，李军辉，贡正仙

现有的神经机器翻译数据扩充的方法通常需要借助于外部数据资源，如大规模源端或目标端单语数据、双语词典等。本文提出了在不引入外部资源的情况下的一种简单数据扩充方法。具体地，对原始平行数据的目标端进行动态扩充的方法，即在每次加载目标端句子时按照一定策略对句子中单词进行随机噪声化，从而提高目标端语言模型对句子的表达能力。在中-英和英-德数据集上的实验结果表明，相较于 Transformer 基准模型该文提出的方法可以有效提高机器翻译质量。

融合篇章上下文有效识别的篇章级机器翻译

汪浩，贡正仙，李军辉

神经机器翻译在机器翻译领域的发展势头正猛，在很多语言对之间取得的效果已超过传统的统计机器翻译。篇章翻译是近来兴起的研究热点，如何能够在翻译文档时充分利用篇章信息一直是该研究的关键点和难点。在篇章级机器翻译中，如何选取当前句的篇章上下文是非常关键的。虽然相关研究使用的篇章上下文不尽相同，但是却少有在选取之前对上下文信息进行识别筛选。本文提出了一种融合篇章上下文有效识别的篇章级翻译模型，引入判别篇章上下文是否有效的分类任务，并根据判别结果来控制目标端对篇章上下文的利用。在中英、英德翻译任务上，与基准系统相比，都得到了显著的提升。

基于枢轴的汉越联合训练神经机器翻译

高盛祥，刘畅，余正涛，黄继豪

越南语是一种典型的低资源语言，为了缓解汉越机器翻译面临资源稀缺的问题，提出一种基于枢轴的汉越联合训练神经机器翻译方法。首先利用小规模的汉越平行语料训练翻译模型得到汉语和越南语的词向量表征，再将以英语作为枢轴语言的汉语-英语，英语-越南语翻译模型进行联合训练。汉语-英语，英语-越南语翻译模型的汉语、越南语的向量表示与汉越模型得到的汉语、越南语的向量表示计算优化从而进行汉越联合训练。实验结果表明，本文方法将汉越平行语料与汉英、英越平行语料结合起来进行联合训练，

充分利用了英语枢轴语料提升了汉越机器翻译性能。相比基线系统，基于枢轴的汉越联合训练神经机器翻译模型提升了 1.81 个 BLEU 值。

基于掩码机制的非自回归神经机器翻译

贾浩，王煦，季佰军，段湘煜，张民

当前基于自注意力机制的神经机器翻译模型取得了长足的进展，但是采用自回归的神经机器翻译在解码过程中无法并行计算，耗费时间过长。我们提出了一个采用非自回归的神经机器翻译模型，可以实现并行解码，并且只使用一个 Transformer 的编码器模块进行训练，简化了传统的编码器-解码器结构。同时在训练过程中我们引入了掩码机制，减小了与自回归的神经机器翻译的翻译效果差距。相比其他非自回归翻译模型，在 WMT2016 罗马尼亚语-英语翻译任务上我们取得了更好的效果，并且使用跨语言预训练语言模型初始化之后，我们取得了和自回归神经机器翻译模型相当的结果。

神经机器翻译词级别正则化技术研究

邱石贵，章华奥，段湘煜，张民

神经机器翻译是利用大型人工网络对翻译建模的过程，在机器翻译质量上获得很大提升，但是作为深度网络模型，神经机器翻译模型面临泛化能力不足的问题以及在低资源场景下更容易出现过拟合现象。针对此问题，本文研究了词级别上的正则化技术(Word Regularization, WR)，通过对模型输入句子中的单词进行随机的扰动，以此削弱数据的特异性，从而抑制模型对于数据的过度学习防止过拟合，同时提高模型对于未知数据的泛化能力。我们选择 Transformer 模型在中-英数据集和英语-土耳其语数据集上进行相关的实验，结果显示模型在训练收敛后更加稳定不易出现过拟合的情况，并且翻译质量也有明显提升。

神经机器翻译军事领域英译汉评测及译后编辑研究

郭望皓， 胡富茂

尽管神经机器翻译技术取得巨大进步，神经机器翻译系统正在加速推进实用化和商品化，但在垂直领域上的表现还不尽如人意。本研究以国内外主流机器翻译系统军事领域英汉文本翻译为研究对象，在自主构建的 1000 个测试数据集上，谷歌、百度、腾讯、网易有道、搜狗 5 家翻译系统的 BLEU 均值仅为 20.854，较之于通用语料相差超过 130%。实验结果显示，译文在拼写、词汇、句法和语义 4 大类 15 种共 5050 处错误中，军事术语翻译错误占比最高，为 42.83%；其次为普通词语误译和层级结构错误。实验结果表明，目前现有的神经机器翻译系统还不能实现高质量的军事文本翻译，无法满足现实需求，亟需进行译后编辑研究，提升军事文本翻译的准确率。

基于联合选择机制的篇章级机器翻译

陈林卿，李军辉，贡正仙

以往的篇章神经机器翻译研究工作大多将研究重心放在句子级上下文的利用方面，通过不同方式获取句子级上下文并将其与机器翻译模型结合以提高翻译性能。而利用篇章级上下文的研究工作大多需要对篇章语料中的整个文档进行计算，存在计算量过大及信息冗余的情况。该文提出使用软硬结合的上下文选择机制，使用基于句向量的轻量级计算即硬选择机制在全篇章内获取与当前句高度相关的上下文，再通过软选择机制将获取的全局上下文分配给源语言当前语句。实验表明该方法在有效约束模型参数及运算量的前

提下从文档全局获取上下文帮助翻译模型并获得了有意义的性能提升。该文进一步分析了文档中篇章级上下文在当前句子周围的分布情况并观察到一些值得思考的实验现象。

基于古籍白话译本的古文机器翻译研究

魏家泽，何彦青，董诚，洪涛，苏瑞欣

古文献具有极高的文学价值与历史价值，古文机器翻译有助于促进文化传播和古文外译。目前古文机器翻译研究仍然存在语料稀缺、语言风格迥异以及一词多义等问题。本文借助古籍白话译本中蕴含的丰富信息从三个维度来协同提升古文机器翻译效果：从古籍白话译本中提取高质量的句内互译片段用来扩充训练语料实现句内片段协同；对古籍所属朝代信息进行语言分期，借助 Fine-tune 微调技术训练上古期、中古期和近古期三个不同时期的翻译模型实现语言分期协同；从双语注释信息中提取古文词汇的注释信息，构建多义注释词典帮助翻译引擎实现注释信息协同。最后以翻译效果提升为标准对每个协同方法分别进行实验验证，在语料有限的情况下，三种协同方法均可以有效提升翻译效果。

Oral 4 中文论文

---

10 月 12 日 13:00 – 14:30

主持人：段湘煜（苏州大学）

13:00 – 13:45 同源语料增强的印尼语汉语神经机器翻译
13:45 – 14:00 融合语言学知识的神经机器翻译研究进展
14:00 – 14:45 基于数据增强技术的神经机器翻译
14:45 – 14:30 基于迭代知识精炼的对偶学习蒙汉机器翻译

Introduction:

同源语料增强的印尼语汉语神经机器翻译

王琳, 刘伍颖

缺少平行句库的低资源机器翻译面临跨语言语义转述科学问题。围绕具体的低资源印尼语汉语机器翻译问题，我们探索了基于同源语料的语言资源扩建方法，并混合同源语料训练出改进的神经机器翻译模型。这种改进模型在印尼语汉语机器翻译实验中取得了 20.30 的 BLEU4 评分。对实验结果的人工抽样分析发现改进的神经机器翻译效果与同时期的谷歌翻译效果相当。对于某些句子我们的译文质量甚至更优。实验结果证明同源语料能够有效改进低资源神经机器翻译，而这种有效性主要是源于同源语言之间的形态相似性和语义等价性。

融合语言学知识的神经机器翻译研究进展

郭望皓，范江威，张克亮

尽管神经机器翻译已经成为目前机器翻译研究应用中的主流方法与范式，但同时也存在译文流利但不够忠实、罕见词处理困难、低资源语言表现不佳、跨领域适应性差、先验知识利用率低等问题。受统计机器翻译研究启发，在神经机器翻译模型中融入语言学信息，利用已有的语言学知识，缓解神经机器翻译面临的固有困境，提升翻译质量，成为神经机器翻译研究领域的一个热门话题。根据语法单位分类体系，可以将这方面的研究分为三类：分别是融合字词结构信息的神经机器翻译、融合短语结构的神经机器翻译和融合句法结构信息的神经机器翻译，目前的研究也集中在这三个方面。首先，梳理了神经机器翻译面临的主要挑战及原因，然后重点介绍了当前融合语言学知识的神经机器翻译研究现状与主要成果，最后总结归纳现有研究中仍在存在的问题，展望未来的研究方向。

基于数据增强技术的神经机器翻译

韩东旭，叶娜，张桂平

神经机器翻译自兴起以来，不断取得了更好的翻译效果。但是神经网络有一个瓶颈，就是只有在大规模平行语料的前提下才会取得好的效果，对于低资源领域的效果一般。本文从数据增强的角度出发，利用翻译模板的思想，识别并抽取出句子中的名词或术语，保留句子的主干。然后通过将抽取出的术语集合在句子主干框架上进行重组之后生成伪平行语料。最后通过计算句子的困惑度来对生成的伪语料进行把关，生成质量较好的伪语料。该方法有效的缓解了神经网络因为语料不足而导致模型的泛化能力不足的问题。实验结果表明，该方法获得的译文与基线系统相比，BLEU 值提升了 2.32。

基于迭代知识精炼的对偶学习蒙汉机器翻译

孙硕，侯宏旭，乌尼尔，常鑫，贾晓宁，李浩然

深度学习方法凭借对语义的深度理解能力在机器翻译领域取得长足的进步。然而，对于低资源语言，一个难以攻克的问题是大规模双语语料的缺乏导致的模型过拟合。本文针对低资源神经机器翻译数据稀疏的问题，提出了一种迭代知识精炼的对偶学习训练方法，利用回译扩充双语平行语料，通过迭代调整伪语料和真实语料比例在学习语言表征的同时降低噪声风险，最后结合译文质量及流利度奖励在源-目标，目标-源两个方向上优化模型参数，从而达到提升译文质量的目的。我们在 CWMT2019 蒙古语-汉语翻译任务上进行了多项实验，结果表明本文方法相比基线提高显著，充分证明该方法的有效性。

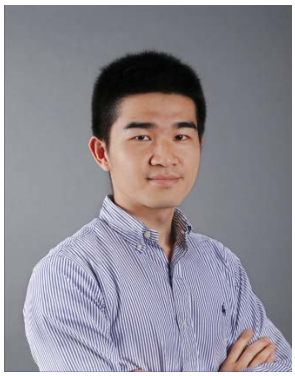
Panel 1 当前机器翻译的瓶颈

10 月 11 日 16:30–18:00

主持人：涂兆鹏（腾讯）

嘉　宾：李沐、刘群、刘洋、朱靖波

嘉宾介绍：



简介：

涂兆鹏，腾讯 AI Lab 专家研究员。研究方向是机器翻译和基于深度学习的自然语言处理，在 ACL、EMNLP、NAACL、AAAI 等国际顶级会议和期刊发表 50 余篇论文，担任或曾担任 Neurocomputing 编委，ACL、EMNLP、NAACL 等会议的机器翻译领域主席，以及 AAAI 高级程序委员会委员。

简介：

李沐，腾讯云与智慧产业事业群智能平台部总监，腾讯公司技术研究通道委员会委员。曾在相关领域国际会议上发表论文 70 余篇，目前研究兴趣包括自然语言处理，人机对话系统与大数据智能等方向。

简介:

刘群，华为诺亚方舟实验室语音语义首席科学家，负责语音和自然语言处理研究。原爱尔兰都柏林城市大学教授、爱尔兰 ADAPT 中心自然语言处理主题负责人、中国科学院计算技术研究所研究员、自然语言处理研究组负责人。分别在中国科学技术大学、中科院计算所、北京大学获得计算机学士、硕士和博士学位。研究方向主要是自然语言理解、语言模型、机器翻译、问答、对话等。研究成果包括汉语词语切分和词性标注系统、基于句法的统计机器翻译方法、篇章机器翻译、机器翻译评价方法等。承担或参与多项中国、爱尔兰和欧盟大型科研项目。在国际会议和期刊发表论文 300 余篇，被引用 8000 多次。培养国内外博士硕士毕业生 50 多人。获得过 Google Research Award、ACL Best Long Paper、钱伟长中文信息处理科学技术奖一等奖、国家科技进步二等奖等奖项。

简介:

刘洋，清华大学计算机科学与技术系长聘教授、人工智能研究所所长，国家杰出青年基金获得者。研究方向是自然语言处理，在自然语言处理和人工智能领域重要国际刊物和国际会议上发表 80 余篇论文，获得 ACL 2017 杰出论文奖和 ACL 2006 优秀亚洲自然语言处理论文奖。获得国家科技进步二等奖、中国电子学会科技进步一等奖、中国中文信息学会钱伟长青年创新一等奖、北京市科学技术奖二等奖等多项科技奖励。担任或曾担任国际计算语言学学会亚太分会执委兼秘书长、Computational Linguistics 编委、ACM TALLIP 副编辑、中国中文信息学会青年工作委员会主任、中国人工智能学会组织工作委员会副秘书长。

简介：

朱靖波，博士，东北大学计算机学院教授、博士生导师、自然语言处理实验室主任、小牛翻译创始人，92 年以来一直从事语言分析和机器翻译基础研究，在国内外期刊杂志和会议上发表了 100 多篇相关学术论文。主持研制了机器翻译开源系统 NiuTrans，于 2016 年荣获钱伟长中文信息处理科学技术一等奖。领导研制的小牛翻译目前支持 140 个语种的翻译能力，属业内支持语种最多的多国语机器翻译商用系统。

Panel 2 多模态机器翻译

---

10 月 12 日 15:45–17:15

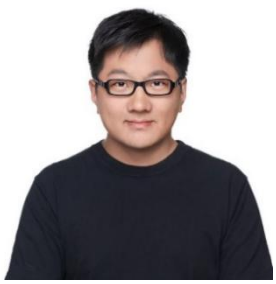
主持人：苏劲松（厦门大学）

嘉　宾：李响、骆卫华、王瑞、肖桐

嘉宾介绍：

简介：

苏劲松，厦门大学信息学院副教授、博导。研究方向：机器翻译，自然语言处理。在 T-PAMI、AI、ACL、AAAI、IJCAI、ACMMM 等 CCF-A 类国际期刊和会议发表论文 30 多篇，长期担任多个国际权威期刊和会议审稿人。目前为中国中文信息学会机器翻译专委会委员、青年工作委员会委员。



简介：

李响，小米高级软件工程师，2018 年博士毕业于中科院计算所，同年加入小米 AI 实验室 NLP 应用组，负责机器翻译业务，目前同声传译，文本翻译，图片翻

译和端上离线翻译等技术已广泛应用于小爱同学，浏览器，扫一扫和小爱老师等多个小米“手机×AIoT”产品中。

简介：

骆卫华，阿里巴巴资深算法专家，目前担任达摩院语言智能实验室翻译平台负责人，组建并带领团队研发面向电商的多语言翻译算法和产品，搭建了阿里巴巴机器翻译引擎和众包翻译平台，成为整个集团的国际化基础设施。加入阿里之前，骆卫华在中科院计算所智能信息重点实验室担任高级工程师，长期从事机器翻译技术的研发和工程化落地，发表数十篇论文，已授权专利数十项，并获得多项省部级科技奖项。

简介：

王瑞博士目前在日本情报通信研究机构（NICT）担任终身研究员。他的研究方向是机器翻译和自然语言处理。目前他的研究兴趣侧重在将传统的语言学知识和先进的机器学习方法融合在机器翻译中。作为第一或通讯作者，他已发表了 30 余篇顶级会议和期刊文章（ACL, EMNLP, ICLR, AAAI, IJCAI, IEEE/ACM transactions 等）。他领导的团队在机器翻译和自然语言处理国际顶级测评中获得多次第一名（ WMT-2018, WMT-2019, WMT-2020, CoNLL-2019 等）。他担任了多个国际会议和期刊的程序委员会委员和审稿人，包括机器学习国际顶级会议 ICLR-2021 的领域主席。

简介：

肖桐，博士，副教授，博士生导师，东北大学自然语言处理实验室副主任，小牛翻译 CEO。2012 年博士毕业于东北大学，中国中文信息学会首届优秀博士论文提名奖获得者，曾先后在日本富士施乐研究中心、微软亚洲研究院访问学习。2013-2014 赴英国剑桥大学开展博士后研究。小牛开源项目的技术负责人（www.niutrans.com），并主持多套机器翻译评测（比赛）系统的研发，在 WMT、CWMT/CCMT 等评测中取得多项任务的第一。至今，NiuTrans 系统已经被来自全世界的两千多个机构下载注册使用，该系统也于 2016 年获得国内自然语言处理领域最高奖 – 钱伟长中文信息处理科学技术奖（一等奖）。小牛翻译产品已支持 119 种语言的互译，同时被大规模应用。在人工智能及自然处理语言领域重要期刊 AI、JAIR、TASL 及顶级会议 AAAI、IJCAI、ACL 发表论文 20 余篇。 社会学术兼职包括：中国中文信息学会青年工作委员会副主任、中国中文信息学会信息检索与内容安全专业委员会委员等、中国计算机学会中文信息技术专委会。和 2014 年在北京交通大学和中国科学院自动化所获得学士和博士学位，并于 2015 年获得中国人工智能学会优秀博士论文。2015 年，他在美国纽约城市大学完成博士后研究。2015 年到 2016 年，他在 IBM Watson 任 Research Staff Member。

博士生论坛

10 月 12 日 10:15 – 11:45

主持人：黄非（阿里巴巴）、黄书剑（南京大学）

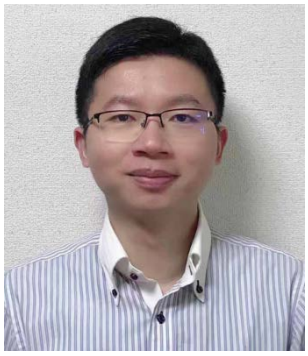
嘉　宾：褚晨翔、顾佳涛、黄轩成、李垠桥、刘宇宸、邵晨泽、郑在翔

简介：

黄非，达摩院机器智能语言技术实验室研究员/资深总监，自然语言基础技术，对话技术和创新翻译团队负责人。他领导 AliNLP 基础技术研发和业务落地，云小蜜对话技术和创新翻译技术，并支持集团内外的国际化业务需求。团队的研究方向包括多语言分词，命名体识别，信息抽取，深度语言模型，知识图谱，文本生成，智能客服和机器翻译，特别是多模态翻译（实时沟通，语音，图像，视频等翻译场景）。对内支持上千个场景日均万亿级调用，对外赋能包括数字政务，城市安全，医疗健康，电力能源，金融科技，海关和电信等多个行业合作伙伴。黄非博士毕业于卡耐基梅隆大学计算机学院。之后在 IBM 和 Facebook 从事自然语言处理的研发和技术管理等职位。他在自然语言处理和人工智能的顶级会议和期刊发表文章 40 多篇，获得美国专利 10 多项，曾担任多个自然语言处理国际会议的领域主席，多个期刊会议论文的审稿人和资深程序委员。

简介：

黄书剑，博士，南京大学计算机科学与技术系副教授，博士生导师。于 2006 年和 2012 年于南京大学获得工学学士和博士学位。主要研究方向包括机器翻译、计算机辅助翻译、文本分析与理解、知识发掘等。发表论文三十余篇，其中包括 ACL，AAAI，IJCAI 等顶级国际会议。曾担任 ACL，AAAI，IJCAI，NAACL，EMNLP 等会议的 PC 或审稿人，担任 CCMT2019 程序委员会主席，NLPCC2016 机器翻译领域主席，CWMT2017、2018 评测委员会主席等。现任中文信息学会青年工作委员会执行委员，中文信息学会机器翻译专委会副主任。2017 年受江苏省自然科学基金优秀青年基金和江苏省青年科技人才托举工程资助。联合指导的博士生获得中国人工智能学会优秀博士生奖（2019）。



简介：

Chenhui Chu received his B.S. in Software Engineering from Chongqing University in 2008, and M.S., and Ph.D. in Informatics from Kyoto University in 2012 and 2015, respectively. He is currently a program-specific associate professor at Kyoto University. His research won the MSRA collaborative research 2019 grant award, 2018 AAMT Nagao award, and CICLing 2014 best student paper award. He is on the editorial board of the Journal of Natural Language Processing, and Journal of Information Processing. His research interests center on natural language processing, particularly machine translation and language and vision understanding.

简介：

Jiatao Gu is currently a research scientist at the Facebook AI Research in New York City. His general research interests lie in applying deep learning approaches to natural language processing (NLP) problems. He obtained his Ph.D. degree at the department of Electrical and Electronic Engineering, University of Hong Kong in 2018 and he was supervised by Prof. Victor O.K. Li. He spent a wonderful time visiting the CILVR Lab, New York University working with Prof. Kyunghyun Cho. Before that, he obtained his Bachelor's degree at the Electronic Engineering Department, Tsinghua University in 2014 with the guidance of Prof. Ji Wu.

简介：

李垠桥，东北大学自然语言处理实验室 18 级博士研究生。 研究方向为网络结构搜索、语言建模、机器翻译等，同时组织实验室开源项目 NiuTensor 的系统开发，参与 WMT、CWMT 等机器翻译评测任务，在 ACL、IJCAI、NLPCC 等会议期刊发表论文若干，社会学术兼职上担任中国中文信息学会青年工作委员会学生委员，参与 AAAI、CCL 等自然语言处理会议审稿工作。

简介：

刘宇宸，就读于中科院自动化所模式识别国家重点实验室，师从宗成庆研究员。研究方向是自然语言处理，主要涉及文本机器翻译、语音翻译、多模态学习。曾在 EMNLP、AAAI、Interspeech 等自然语言处理和语音识别领域国际会议上发表论文多篇。

简介：

邵晨泽，本科毕业于中国科学院大学，现为中国科学院计算技术研究所 2018 级直博生，导师为冯洋研究员。研究方向为机器翻译，在领域顶级会议 ACL、AAAI、EMNLP 均发表一作论文，2018 年和 2019 年参加全国机器翻译评测分别获得第三名和第一名。

简介：

黄轩成，清华大学自然语言处理实验室在读博士生，师从刘洋教授，研究兴趣是机器翻译、自然语言处理，曾以第一作者身份在自然语言处理/机器学习的会议 EMNLP、IJCAI 上发表学术论文。2020 年参加全国机器翻译评测获得第一名。

简介：

郑在翔，南京大学自然语言处理实验室在读博士生，导师为陈家骏教授和黄书剑副教授，曾在英国爱丁堡大学自然语言处理组进行一年的学术访问，现在是字节跳动 AI Lab 研究实习生。他的主要研究兴趣为神经机器翻译、文本生成和深度生成模型，并以第一作者/主要作者在 ICLR、TACL、EMNLP、IJCAI、TASLP 等自然语言处理/机器学习的期刊和会议上发表论文数篇，同时也是 ICLR、ACL、EMNLP、AAAI、IJCAI 等会议的审稿人。

前沿趋势论坛

10 月 12 日 14:45 – 15:15

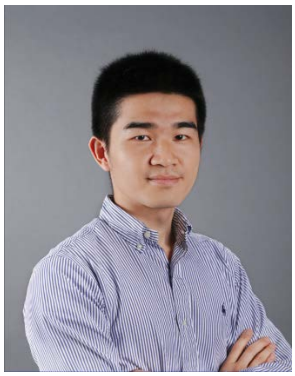
主持人：刘洋（清华大学）

嘉　宾：涂兆鹏

嘉宾介绍：

简介:

刘洋，清华大学计算机科学与技术系长聘教授、人工智能研究所所长，国家杰出青年基金获得者。研究方向是自然语言处理，在自然语言处理和人工智能领域重要国际刊物和国际会议上发表 80 余篇论文，获得 ACL 2017 杰出论文奖和 ACL 2006 优秀亚洲自然语言处理论文奖。获得国家科技进步二等奖、中国电子学会科技进步一等奖、中国中文信息学会钱伟长青年创新一等奖、北京市科学技术奖二等奖等多项科技奖励。担任或曾担任国际计算语言学学会亚太分会执委兼秘书长、Computational Linguistics 编委、ACM TALLIP 副编辑、中国中文信息学会青年工作委员会主任、中国人工智能学会组织工作委员会副秘书长。



简介：

涂兆鹏，腾讯 AI Lab 专家研究员。研究方向是机器翻译和基于深度学习的自然语言处理，在 ACL、EMNLP、NAACL、AAAI 等国际顶级会议和期刊发表 50 余篇论文，担任或曾担任 Neurocomputing 编委，ACL、EMNLP、NAACL 等会议的机器翻译领域主席，以及 AAAI 高级程序委员会委员。