Trends of Machine Translation in China

—— NiuTrans for case study

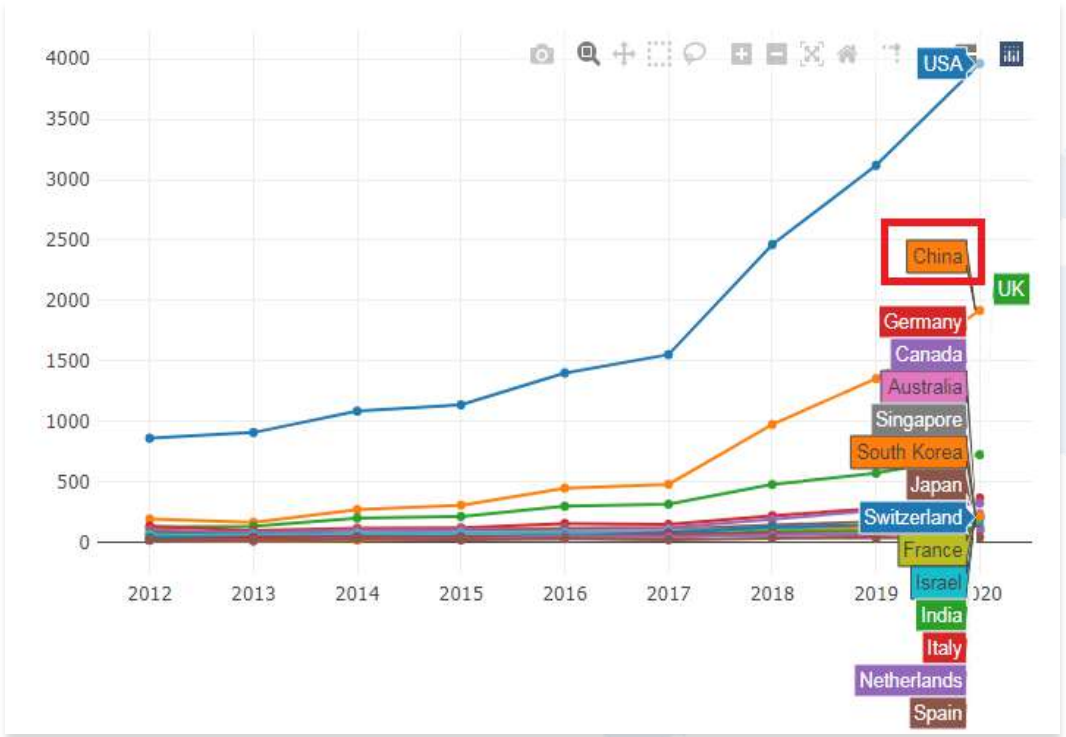
Chunliang Zhang

NiuTrans Co.
NLP Lab @ Northeastern University

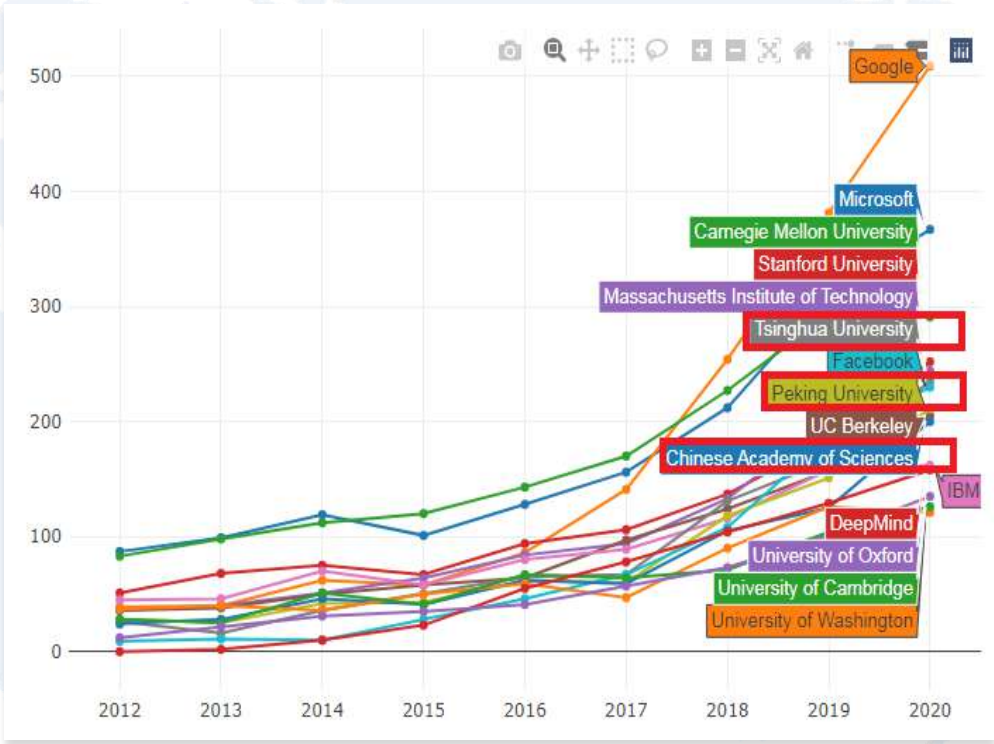
1 NLP R&D power Comparison Worldwide

小牛翻译 NiuTrans.com

- By organization: 3 Chinese institutes among the top 15
- By country: The US taking the lead, followed by China and the UK



2012-2020 Paper publication by countries

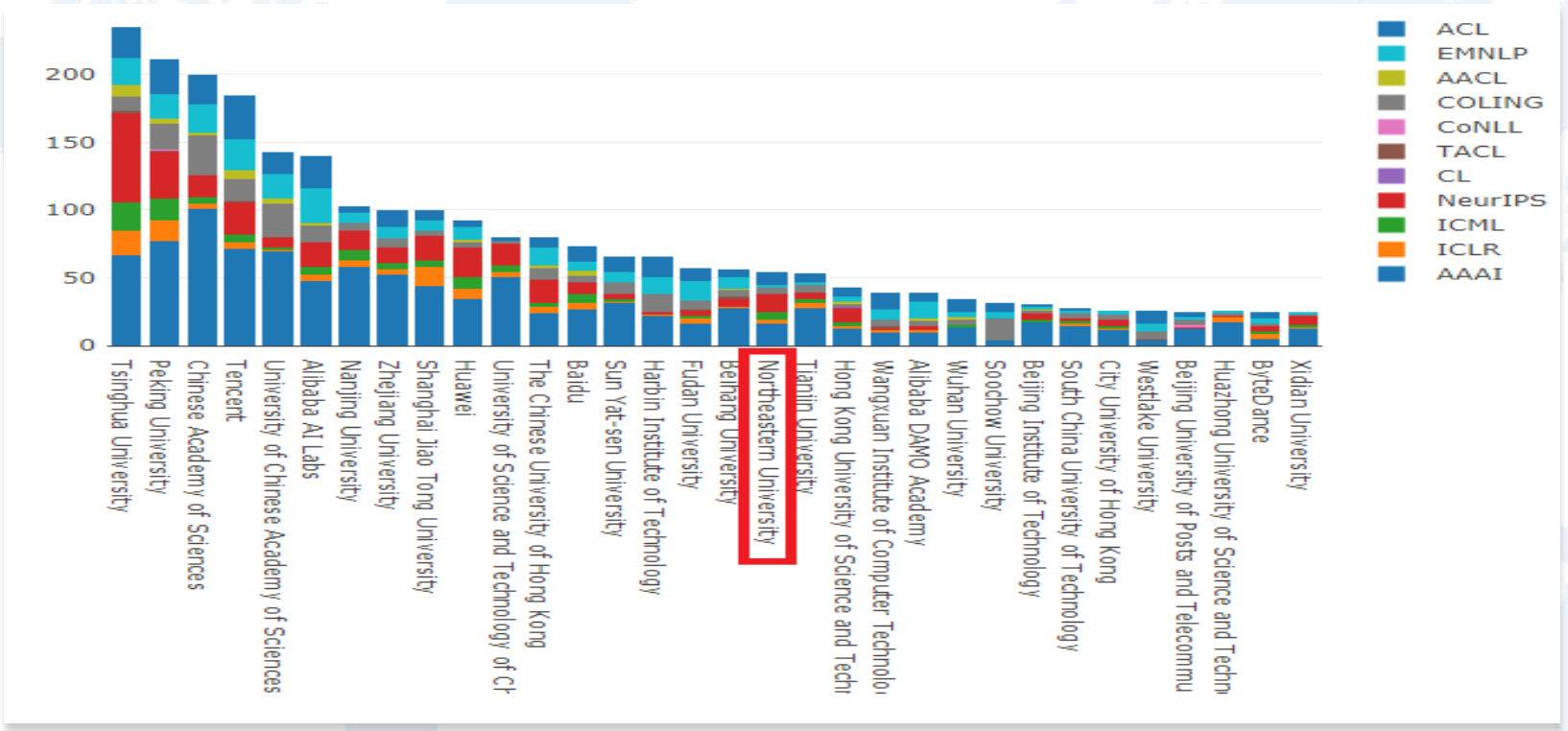


2012-2020 Paper publication by organizations

https://www.marekrei.com/blog/ml-and-nlp-publications-in-2020/

2 NLP R&D power Comparison in China

- Most on the top 30 of the list are Universities and research institutes funded by government
- Only one university focus on Machine Translation: Northeastern University



The Top NLP Research Organizations in China （2012-2020）

https://www.marekrei.com/blog/ml-and-nlp-publications-in-2020/

3 Major MT Business Players in China

- Almost all the Internet-based service giants develop MT
- At least 3 teams are linked to universities/national institutes
- Some companies build hybrid business model: machine+human
- The focus is shifting from Tech R&D to MT engine, products & Marketing

Internet/AI Giants' Team

university-related Co.

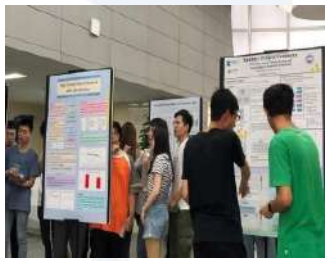
the others (to name a few)

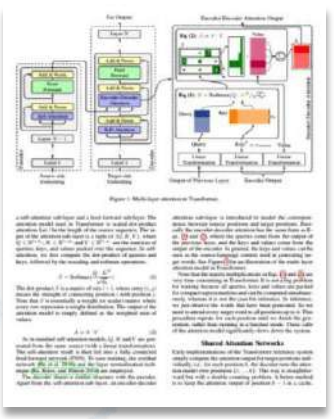
4 NiuTrans Team out of NEUNLP Lab

小牛翻译 NiuTrans.com

41 year for one thing: MT R&D

- NEUNLP lab was founded in 1980，a NLP and MT research pioneer in China
- A leader of MT R&D into MT industry in China
- 150+ members, 80% are PhDs and MSs
- 35 China central government research projects





5 Our History

小牛翻译 NiuTrans.com

304 languages

100+ major partners

- NENLP Lab founded

- built NiuTrans co.

- 22 languages, including 7 minority languages in China
- Write our own NiuTrans.NMT

- 118 languages
- NiuTrans Cloud Online
- Hangzhou, Shanghai, Shenzhen Offices

Dec. 1980 | **Oct. 2007** | **May 2012** | **2015** | **2016** | **2017** | **2018** | **2019-2020**

- NiuTrans.SMT @Lab

- 4 languages, CN, JN, KR, RU
- Angel investment by iFLYTEK

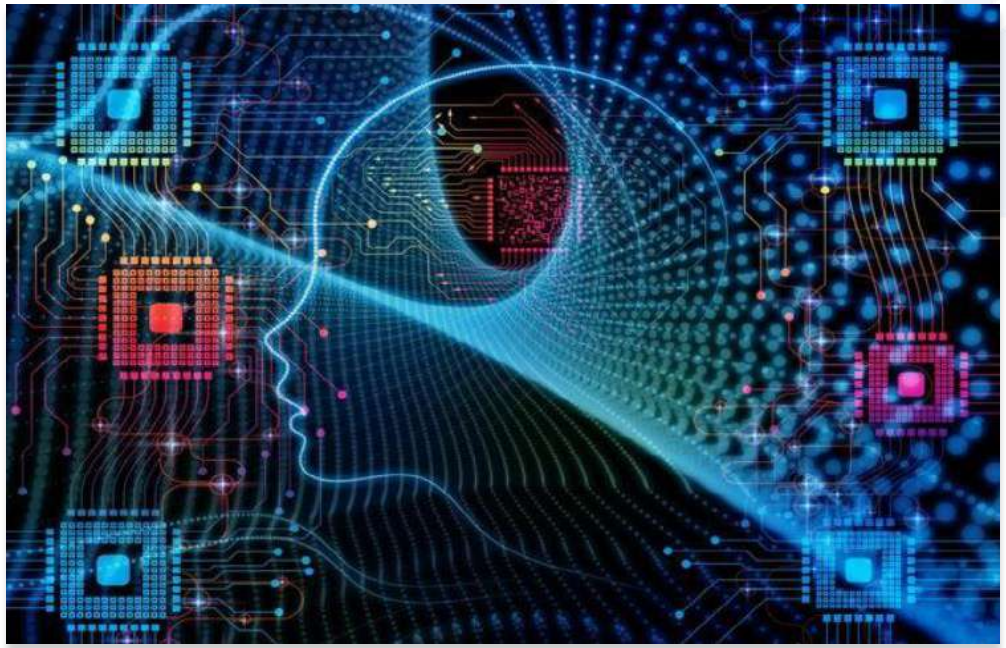
- 67 languages
- Beijing Office

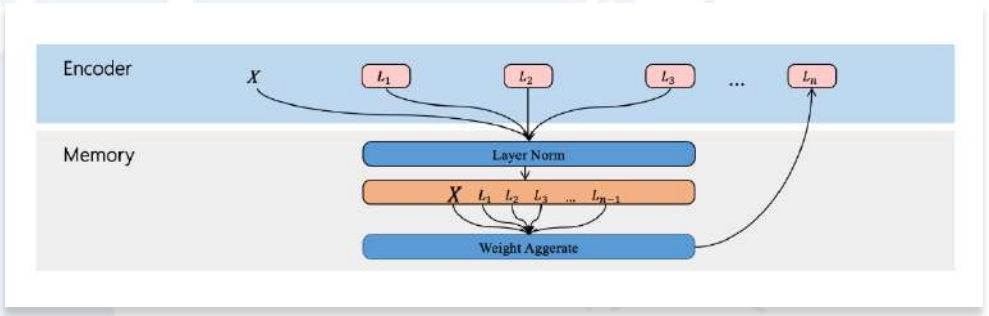
- 304 languages
- Business partnership & go-international strategies

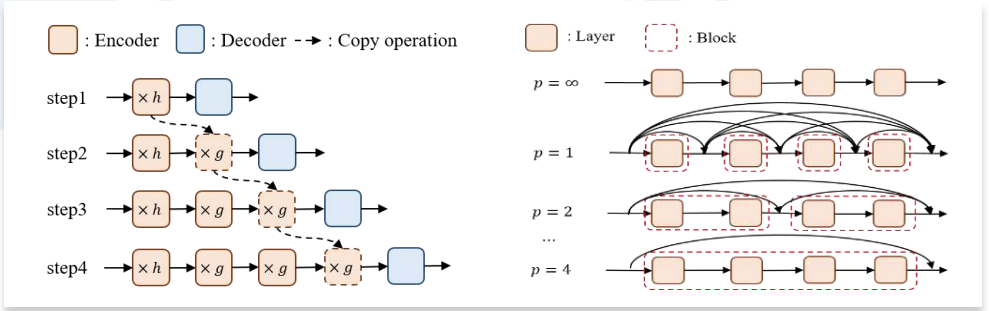
NiuTrans R&D: Deep Neural Network Training

小牛翻译 NiuTrans.com

- **Deepening Transformer**
  - Achieve **better performance** by deepening the neural networks
  - Our model remains **stable** when it is especially deep (over 35 layers)
- **Efficient training methods for Deep Transformer**
  - Shallow-to-Deep Training: a simple yet **efficient** method



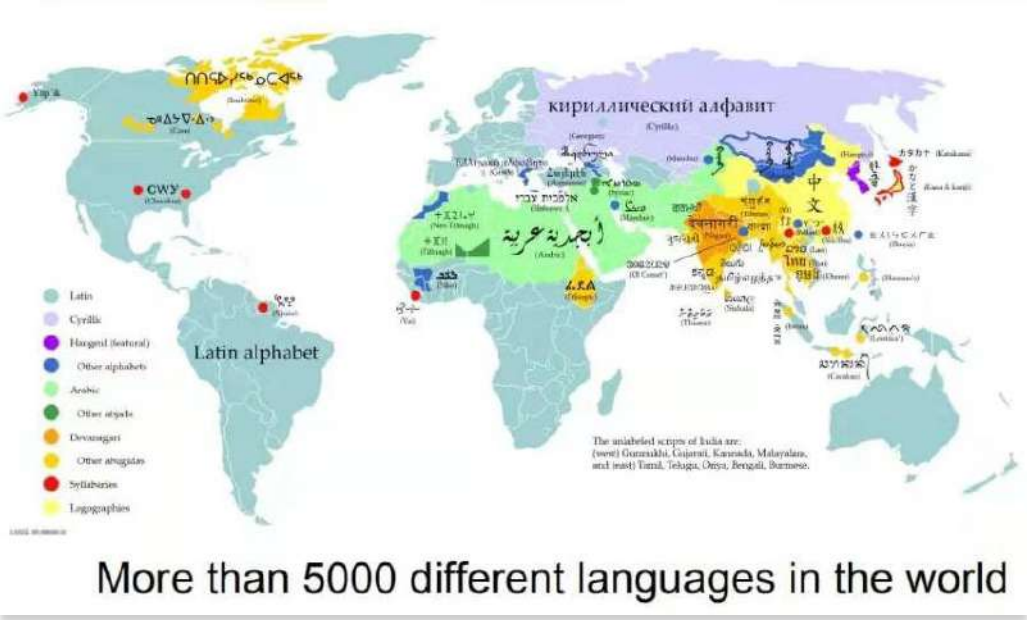




7 NiuTrans R&D: Low-resource Machine Translation

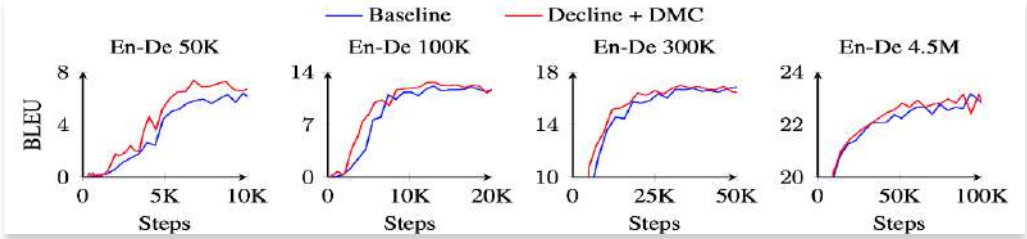
小牛翻译 NiuTrans.com

- **99% languages pairs do not have enough training data**
  - Corpus Building Technology: Bilingual Parallel corpus, Bilingual Comparable corpus, Monolingual corpus.
- **Bilingual Dictionary Induction and Dynamic Curriculum Learning**
  - Finding the corresponding words between two language by unsupervised method automatically.
  - Choosing the appropriate training samples for the current model.



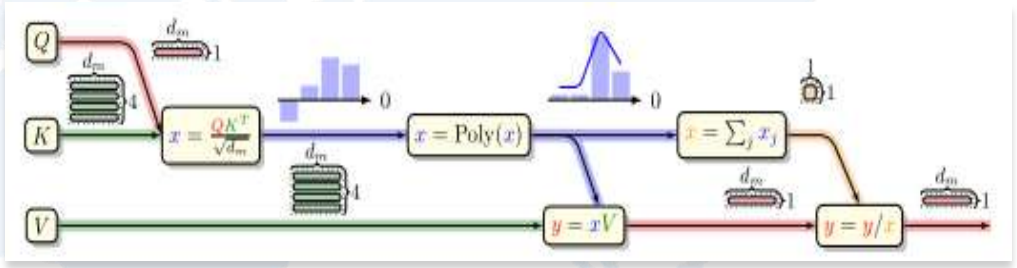
Algorithm 1 Iterative Dimension Reduction

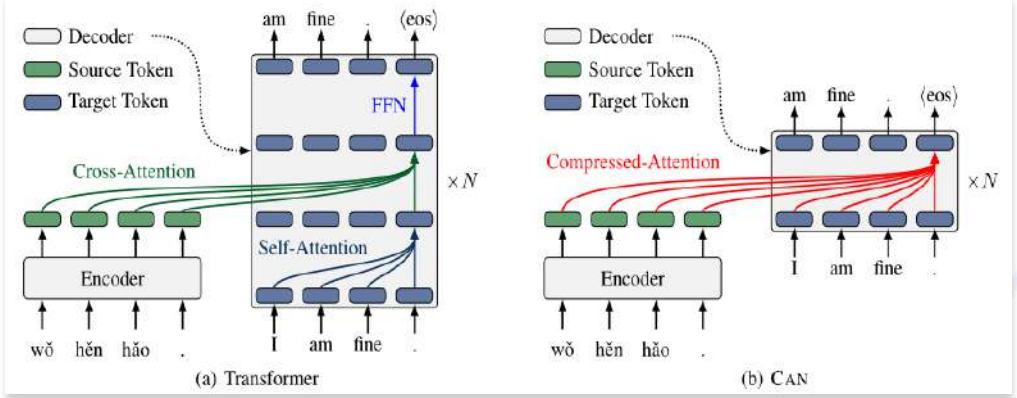
procedure IDR(E, n) ▷ E is the raw embeddings, n is the initial dimension
  $D \leftarrow \varnothing$ ▷ Set the dictionary to empty
  while $n \leq 300$ **do** ▷ 300 is the dimension of the raw embeddings
    Reduce E to $\bar{E}$ with dimension $\min(n, 300)$ using PCA and dropmax
    if $D = \varnothing$ **then**
      Run the initialization and the self-learning on $\bar{E}$
    else
      Run the self-learning on $\bar{E}$ with D as the initial dictionary
    end if
    Translate 4K most frequent words and store the results in D
    $n \leftarrow n \times 2$
  end while
  return W_X and W_Y
end procedure



NiuTrans R&D: Decoding methods for Deep Transformer

- **Faster Transformer with less storage in decoding**
  - Traditional Transformer relies on **massive computing power and storage (GPU)**.
  - It’ s hard to apply the deep model in **mobile SoC (CPU)**.
  - 8-bit Integer Inference for the Transformer Model + **lightweight model** architecture



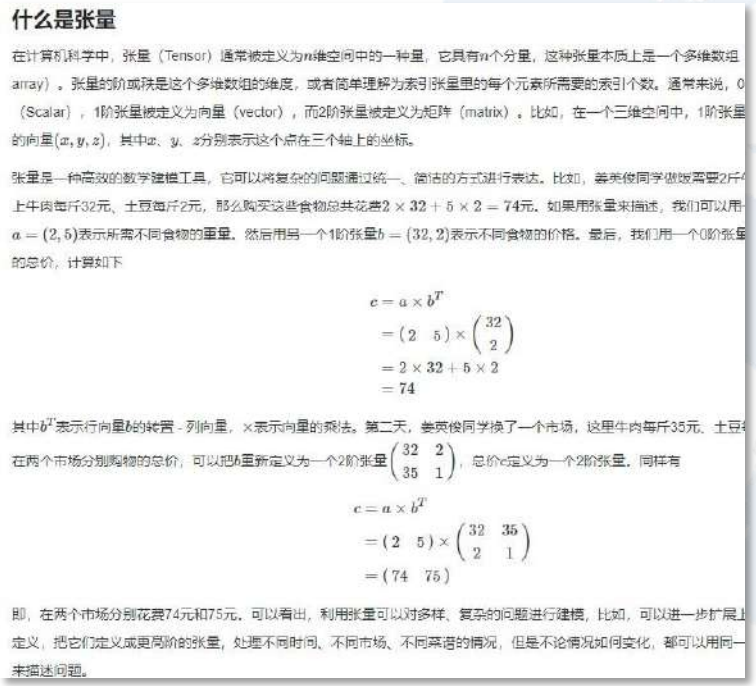


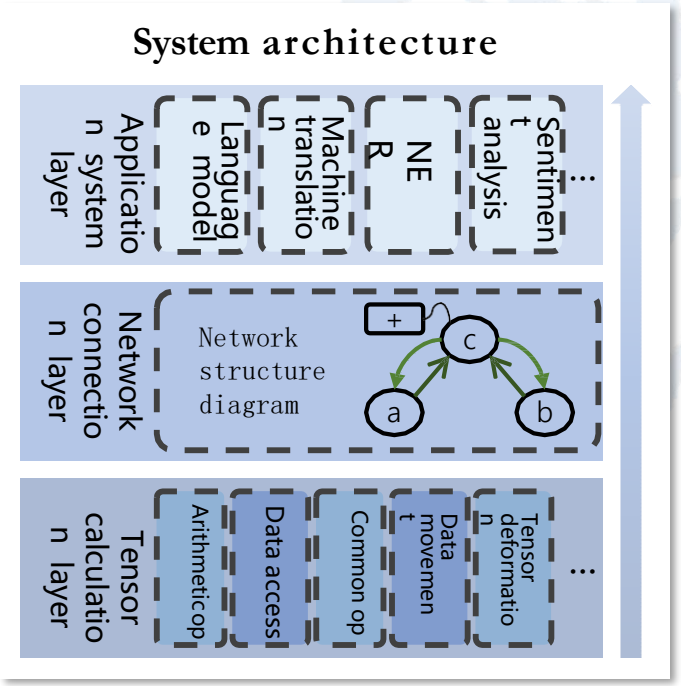
9 NiuTrans Opensource

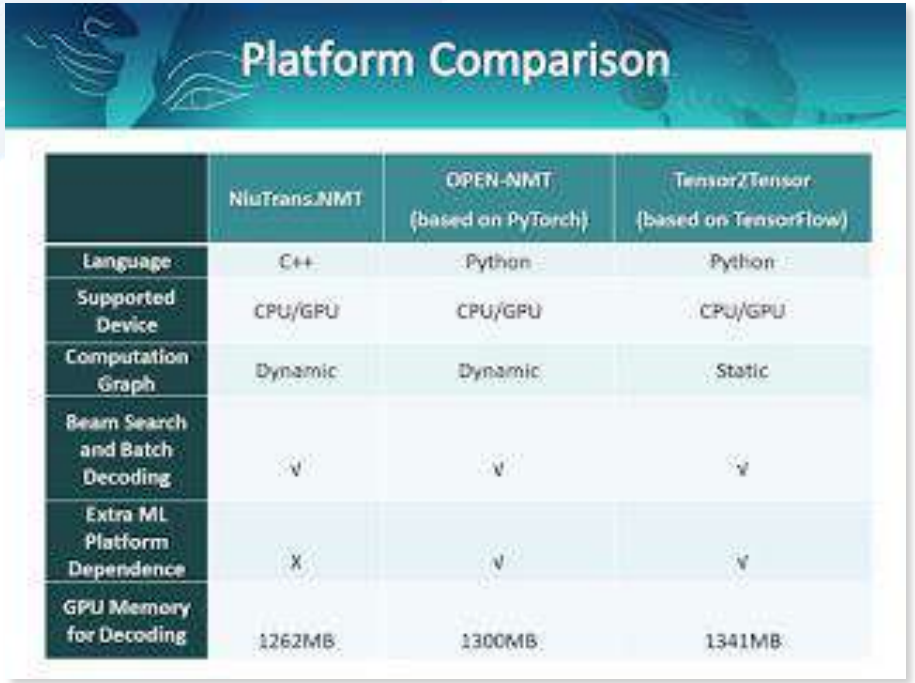
小牛翻译 NiuTrans.com

MT Opensource, downloaded by 2000+ institutions in 70+ countries

- Started in 2007, NiuTrans.SMT v1.0 released in 2011
- The 1st successful Open source MT in China
- NiuTensor & NiuTrans.NMT released in Dec. 2019





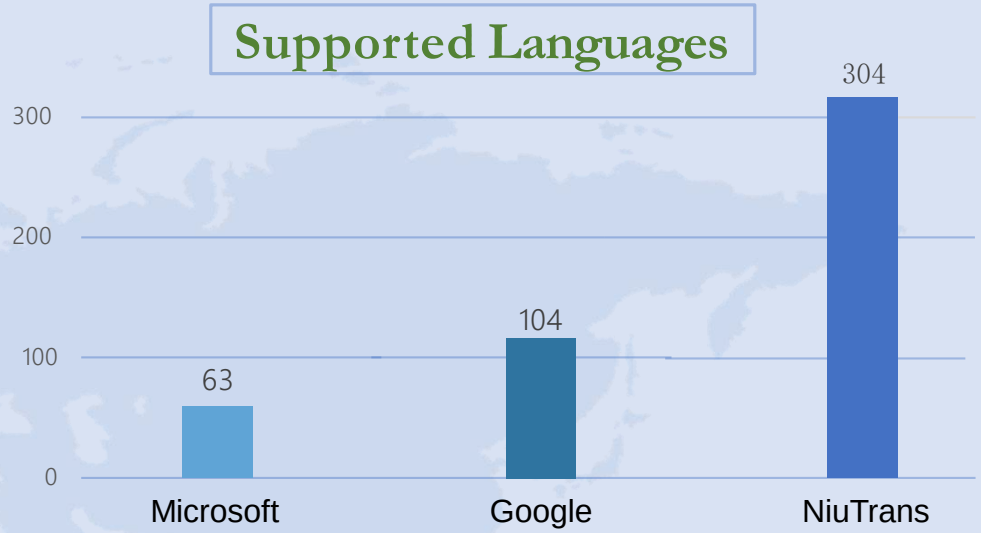


	NiuTrans.NMT	OPEN-NMT (based on PyTorch)	Tensor2Tensor (based on TensorFlow)
Language	C++	Python	Python
Supported Device	CPU/GPU	CPU/GPU	CPU/GPU
Computation Graph	Dynamic	Dynamic	Static
Beam Search and Batch Decoding	√	√	√
Extra ML Platform Dependence	X	√	√
GPU Memory for Decoding	1262MB	1300MB	1341MB

10 NiuTrans: the largest number of languages supported

小牛翻译 NiuTrans.com

- Support 304 languages, translate between any two of them
- Most of them are low-resource languages, covering all China neighbor countries' official languages



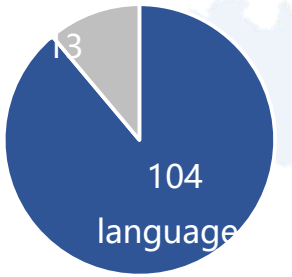
China neighbors

Countries covered

UN member countries

Language coverage in the official languages of the States members of the United Nations

Language covered



NiuTrans: Performance Evaluation

小牛翻译 NiuTrans.com

International Competition

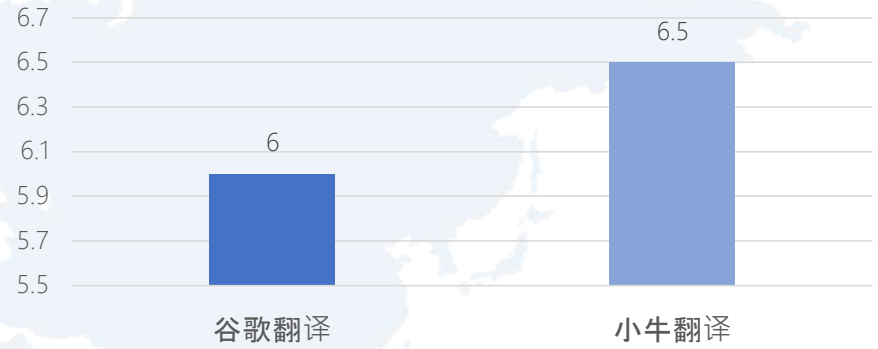
WMT2018/2019/2020

NTCIR2011-Japan

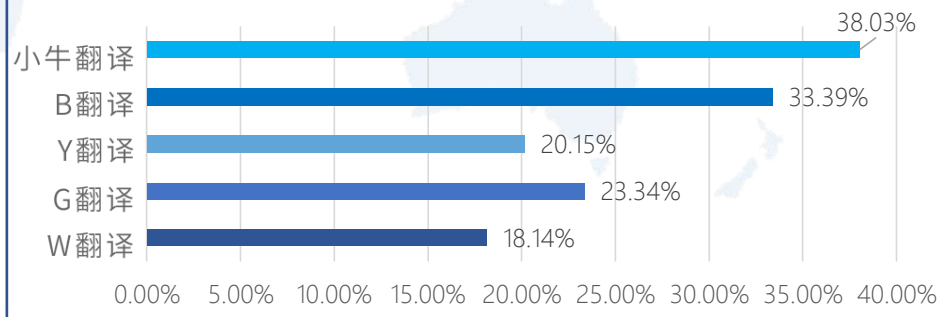
China MT Competition

CCMT2011/2018/2019

Human Evaluation 1



Human Evaluation 2



MT for Winter Olympics 2022 (iFLYTEK+NiuTrans)

小牛翻译 NiuTrans.com

News Reports

Sports Commentary

A bite of local food

Hotel Service

Travel assistant

Medical service

Robot concierge

Easy transport

羽生結弦

フィギュアスケート男子シングルの日本選手。 2018年2月、平昌五輪フィギュアスケート男子シングルで優勝

短道速滑

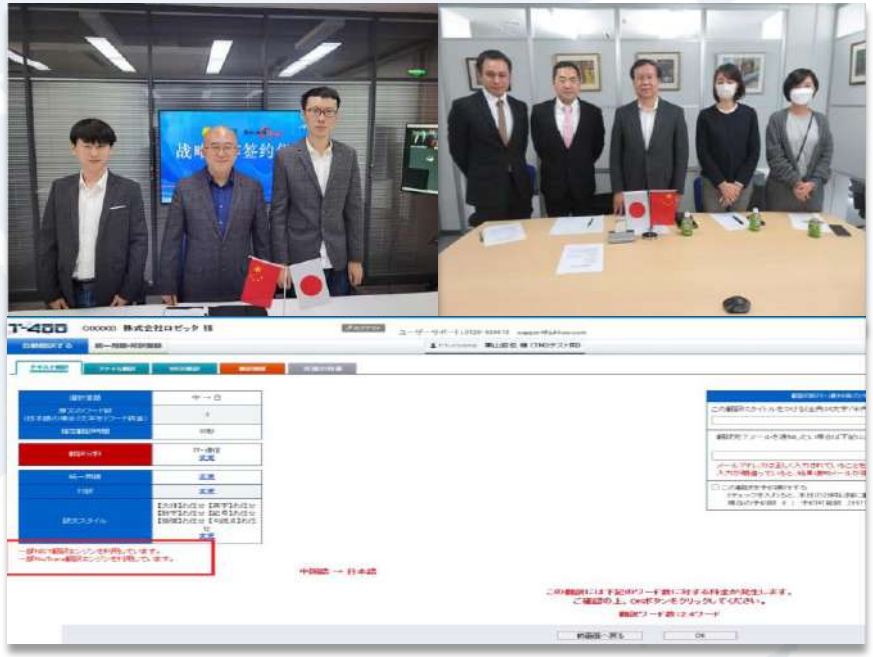
短道速滑是冬季奥运会项目，比赛采用淘**汰制，以**预、次、半**决、决**赛的比赛方式进行。

NiuTrans Going International

小牛翻译 NiuTrans.com

- 2016: Strategic Partnership with SYSTRAN Korea
- 2017: Release NiuTrans Plugin for Trados & memoQ
- 2020: Business Cooperation with 5 companies in Japan, including Rozetta

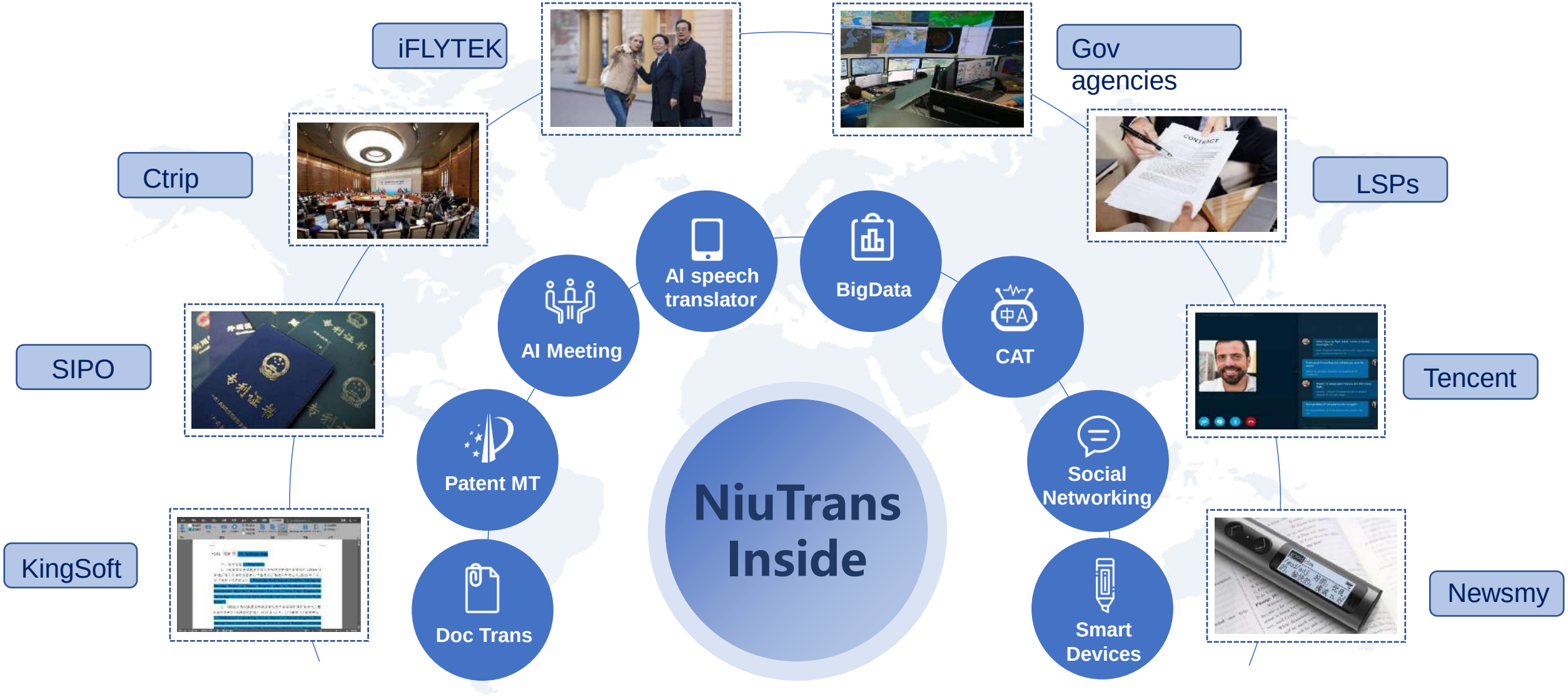
NiuTrans Cloud User Map



Partnership with Rozetta & RC Japan

NiuTrans Inside: a MT engines provider

小牛翻译 NiuTrans.com

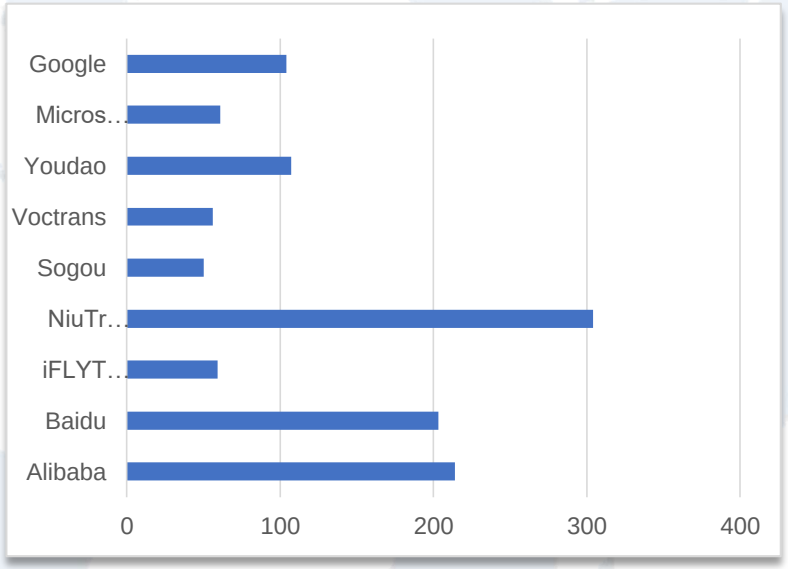


Post-pandemic MT Trends in China

小牛翻译 NiuTrans.com

- **Much better MT quality is demanded for the low-resource languages**
  - Hi-quality Corpus matter much more!

China's Belt&Road Map



MTs' Supported Language Statistics in 2020

High demand of language data

Post-pandemic MT Trends in China

小牛翻译 NiuTrans.com

➢ **More custom MT for certain businesses or industries**

- Generic MT can't work the same good in all the fields

Medicine

Finance

Law

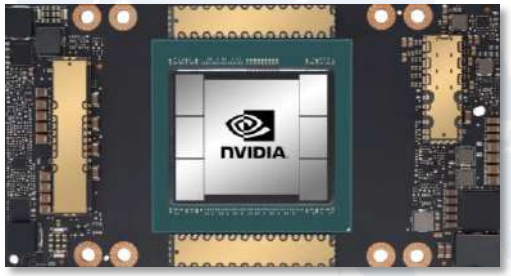
Engineering

Post-pandemic MT Trends in China

小牛翻译 NiuTrans.com

➢ **AI Translation engine embedded into more devices**

- integrate with other language tech: text + speech + image
- run on both GPU and CPU chips, especially CPU-powered smart mobile devices



小牛翻译
NiuTrans.com

Make the best MT engine

Thanks