Trends of Machine Translation in China
—— NiuTrans for case study

Chunliang Zhang
NiuTrans Co.
NLP Lab @ Northeastern University
NLP R&D power Comparison Worldwide

- 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

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)

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

- Baidu
- Sogou
- Tencent
- Youdao
- Alibaba
- ByteDance

University-related Co.

- NiuTrans.com
- CloudTranslation
- nylangTech

The others (to name a few)

- ATMAN
- ATMAN
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
Our History

- NENLP Lab founded
- built NiuTrans co.

304 languages

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

100+ major partners

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

- 67 languages
- Beijing Office

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

- 304 languages
- Business partnership & go-international strategies
NiuTrans R&D: Deep Neural Network Training

- **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
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.
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
NiuTrans Opensource

**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

System architecture

Platform Comparison

<table>
<thead>
<tr>
<th></th>
<th>NiuTrans.NMT</th>
<th>OPEN-NMT (based on PyTorch)</th>
<th>Tensor2Tensor (based on TensorFlow)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language</strong></td>
<td>C++</td>
<td>Python</td>
<td>Python</td>
</tr>
<tr>
<td><strong>Supported Device</strong></td>
<td>CPU/GPU</td>
<td>CPU/GPU</td>
<td>CPU/GPU</td>
</tr>
<tr>
<td><strong>Computation Graph</strong></td>
<td>Dynamic</td>
<td>Dynamic</td>
<td>Static</td>
</tr>
<tr>
<td><strong>Beam Search and Batch Decoding</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>Extra ML Platform Dependence</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>GPU Memory for Decoding</strong></td>
<td>126MB</td>
<td>1300MB</td>
<td>1341MB</td>
</tr>
</tbody>
</table>
NiuTrans: the largest number of languages supported

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

**Supported Languages**

<table>
<thead>
<tr>
<th>Language Coverage</th>
<th>Microsoft</th>
<th>Google</th>
<th>NiuTrans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>63</td>
<td>104</td>
<td>304</td>
</tr>
</tbody>
</table>

**Countries Covered**

- China neighbors: 20 countries
- UN member countries: 193 countries

**Language Coverage in the Official Languages of the States Members of the United Nations**

- Covered: 104 languages
- Uncovered: 13 languages
NiuTrans: Performance Evaluation

**International Competition**

**No. 1**

WMT2018/2019/2020

ECTSIR2011-Japan

**China MT Competition**

**No. 1**

CCMT2011/2018/2019

**Human Evaluation 1**

- 谷歌翻译: 6.7
- 小牛翻译: 6.5

**Human Evaluation 2**

- 小牛翻译: 38.03%
- B翻译: 33.39%
- Y翻译: 23.34%
- G翻译: 20.15%
- W翻译: 18.14%
A bite of local food
Hotel Service
Travel assistant
Medical service
Robot concierge
Easy transport

Sports Commentary

News Reports

MT for Winter Olympics 2022 (iFLYTEK+NiuTrans)

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

羽生結弦

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

- 2016: Strategic Partnership with SYSTRAN Korea
- 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

- AI speech translator
- BigData
- CAT
- Social Networking
- Smart Devices
- LSPs
- Gov agencies
- iFLYTEK
- Ctrip
- SIPO
- KingSoft
- Tencent
- Newsmy

NiuTrans Inside
Much better MT quality is demanded for the low-resource languages
- Hi-quality Corpus matter much more!
More custom MT for certain businesses or industries

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

Thanks