



Low-Resource Neural Machine Translation

Mieradilijiang Maimaiti

2019.11.12, Beijing (Minzu University of China)

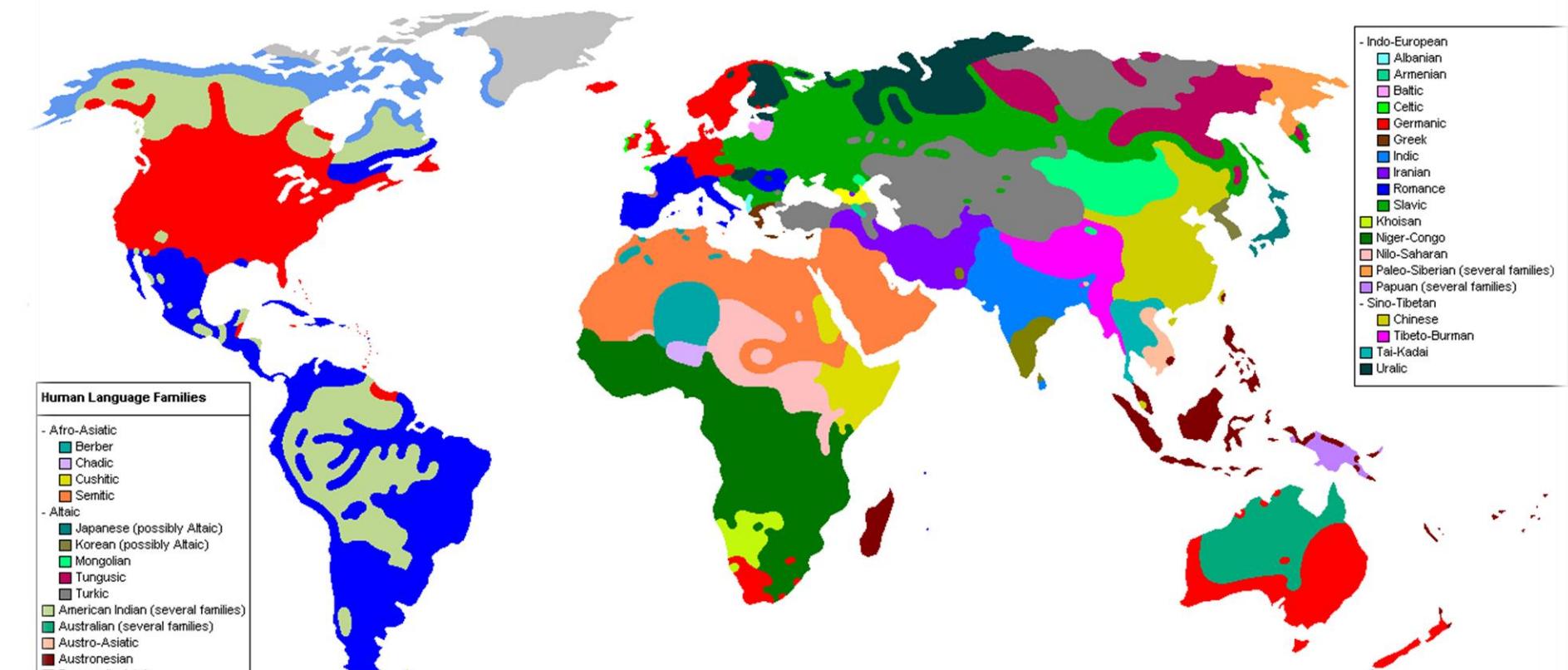


Outline

- Machine Translation
- Related Work and Current State for LRLs NMT
- Motivation
- Our works
- Projects
- Copyrights & Patents
- Conclusions



Cross Lingual Environment



Machine Translation (MT)

- Machine Translation: let the computer translate the human language, as well as it is a technique that uses a computer to automatically convert one natural language (Source language) to another natural language (target language).



Typical MT Systems





Current State of MT Community



清华大学
Tsinghua University



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

Tencent 腾讯

Baidu 翻译



S 搜狗搜索

阿里巴巴
Alibaba.com™

Microsoft
Research
微软 未来 计算 开放 宽容

SAMSUNG

有道 youdao



Demo (SMT)

清华大学多语种翻译系统

1.5.245:9662/index.html

清华大学多语种翻译系统

چىخۇا ئۇنىۋېرسىتى كۆپ تىللەق تەرجىمە سىستېمىسى

维文 > 汉语 > 通用领域 翻译

ئىنقلابىي قۇربانلارنى خاتىرىلەش كۈندە خالق قەھرىمانلىرىغا گۈل سېۋىتى تەقدىم قۇرمۇسى بېيجىڭدا داغدۇغلىق ئۆتكۈزۈلدى.

革命先烈纪念日人民英雄敬献花篮的仪式在北京隆重举行

ts.cn/system/2018/10/01/035399943.shtml

天山网 - 新疆维吾尔自治区人民政府网

中文 Türkçe Русский язык English Уйгурچا uygurche قازاقشا قىزىزچىمە

ئەم سەھىپىنىڭ مەسىھىسىدە ئەم سەھىپىنىڭ مەسىھىسىدە

پارتىيە 19-قۇرۇسالى روحىنى چوڭقۇر ئۆگىنەيلى، ئىزجىلاشتۇرالىلى

ئىنقلابىي قۇربانلارنى خاتىرىلەش كۈندە خالق قەھرىمانلىرىغا گۈل سېۋىتى تەقدىم قىلىش مۇراسىمى بېيجىڭدا داغدۇغلىق ئۆتكۈزۈلدى

最多可以输入500个字符

2018/10/01 15:37

مؤەممەد مەركىزى | مەنەنە: شىنجاڭ كېزىتى
شى جىنبىك، لى كېچىپاڭ، لى جىنۇش، والك يالا، والك خۇنىك، جاۋ لېچ، خەن جىڭ، والك چىشمۇن قاتاناشى

.NLP&CSS group, Tsinghua University :2011-2018 ©

Email: miradel51@126.com Tel: 13051308938 Wechat: 821777278



Demo (NMT)

The screenshot shows a web-based multilingual translation system. At the top, there are several browser tabs in Chinese, including '清华大学多语...', '清华大学跨语...', '多语种翻译系统', '天山网_百度搜...', '天山网 - 新疆...', 'ئەڭتەڭ ئۇرۇنىڭ ئەپلىك...', and '晒照为大s庆'. The main page title is '多语种翻译系统' (Multilingual Translation System). Below it, a sub-section title is 'كۆپ تىللېق تەرىجىمە سىستېمىسى' (A multi-language translation system). The interface includes language selection buttons: '维吾尔语' (Uyghur), '汉语' (Chinese), '通用领域' (General field), and '翻译' (Translation). A large text area contains Chinese text about the 'National People's Congress Standing Committee, State Council, National Committee of the Chinese People's Political Consultative Conference, Central Military Commission, various democratic parties, all工商联 and non-party爱国人士, various people's organizations and各界群众, veterans, old revolutionaries and their families' arranged in a formation facing the Central Committee of the Communist Party of China. It also mentions the '少先队' (Young Pioneers) named after Chinese heroes. Below this text is a URL: 'uy.ts.cn/system/2018/10/01/035399943.shtml'. A red arrow points from the right side of the page towards this URL. Another red arrow points from the bottom right towards a highlighted box containing a portion of the Chinese text. This highlighted box is enclosed in a red border and contains the following text:

ئامما تۈزگەن چاسا ئەتتەن ئالدىغا جۇڭگو كومىنىستىك پارتىيەس مەركىزىي كومىتېتى، مەملىكتىك خەلق قۇرۇلتىنى دائىمىي كومىتېتى، گۇۋۇيۇن، مەملىكتىك سىياسى كېڭىش، مەركىزىي ھەربىي كومىتېت، ھرقايسى ديمۆراتىك پارتىيە-گۈزۈلەر، مەملىكتىك سودا-ساناھىتچىلەر بىرلەشىمىسى ۋە پارتىيە-گۈزۈھەسىز ۋە تەنپەرەۋەر زاتلار، ھرقايسى خەلق تەشكىلاتلىرى ۋە ھەر ساھە ئاممىسى، پىشىقەدمە جەڭچىلەر، پىشىقەدمە بولداشلار ۋە ئىقلابىي قىدا، ئازىلما ئاش ئادىل ئايدا ئەتتىنىڭ ئەتتىنىڭ ئامىدا تەقدىم قىلغانچان چوك قاتار تىزلىغاندى.



他 喜欢 北京 。

He likes Beijing .

他 在 东京 居住 。

He lives in Tokyo .

日本 的 首都 是 东京 。

The capital of Japan is Tokyo .

北京 是 中国 的 首都 。

Beijing is the capital of China .

他 来自 日本 。

He is from Japan .

日本 临近 中国 。

Japan is near China .

中国 是 亚洲 国家 。

China is an Asian Country .

北京 位于 中国 的 北方 。

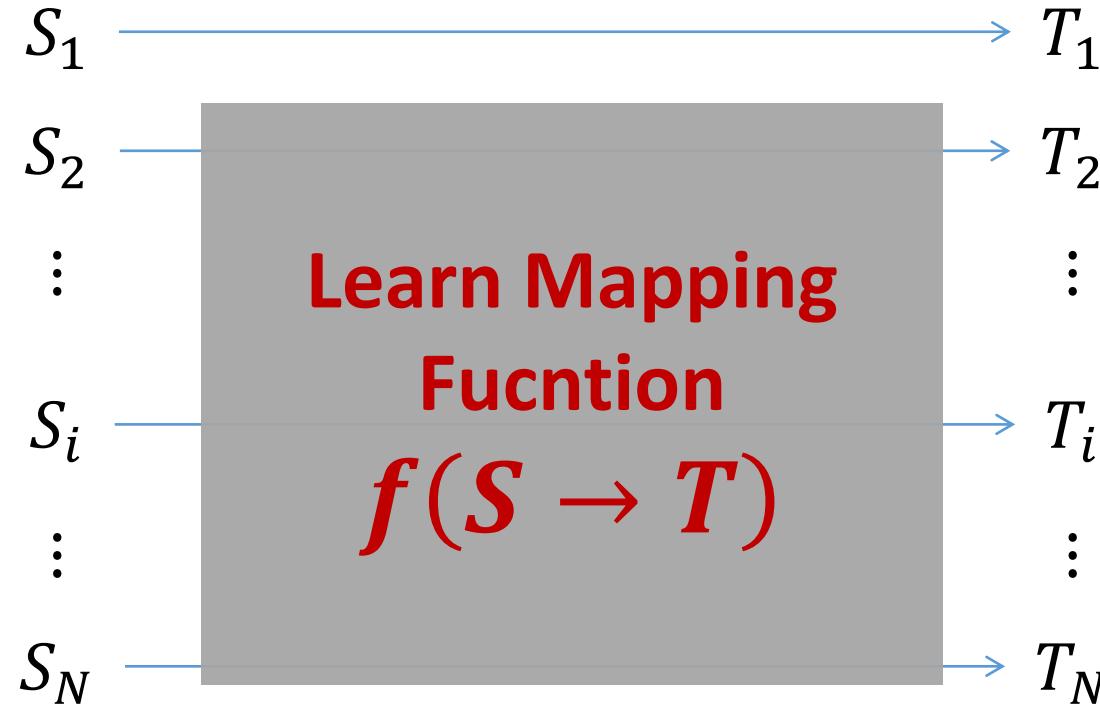
Beijing is located in the North of China .

... ...

... ...

(Jiajun Zhang, CCL2018)

Mapping function from source to target language



$$S_{New} \xrightarrow{f(S \rightarrow T)} T_{New}$$

(Jiajun Zhang, CCL2018)

Mapping function from source to target language

Chinese:

我 在 三星 做了 报告



Mapping
function
 $f(S \rightarrow T)$

English:

I gave a talk in Samsung



X

布什

与

沙龙

举行

了

会谈



$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{z}} \frac{\exp(\boldsymbol{\theta} \cdot \phi(\mathbf{x}, \mathbf{y}, \mathbf{z}))}{\sum_{\mathbf{y}'} \sum_{\mathbf{z}'} \exp(\boldsymbol{\theta} \cdot \phi(\mathbf{x}, \mathbf{y}', \mathbf{z}'))}$$

**y**

Bush

held

a

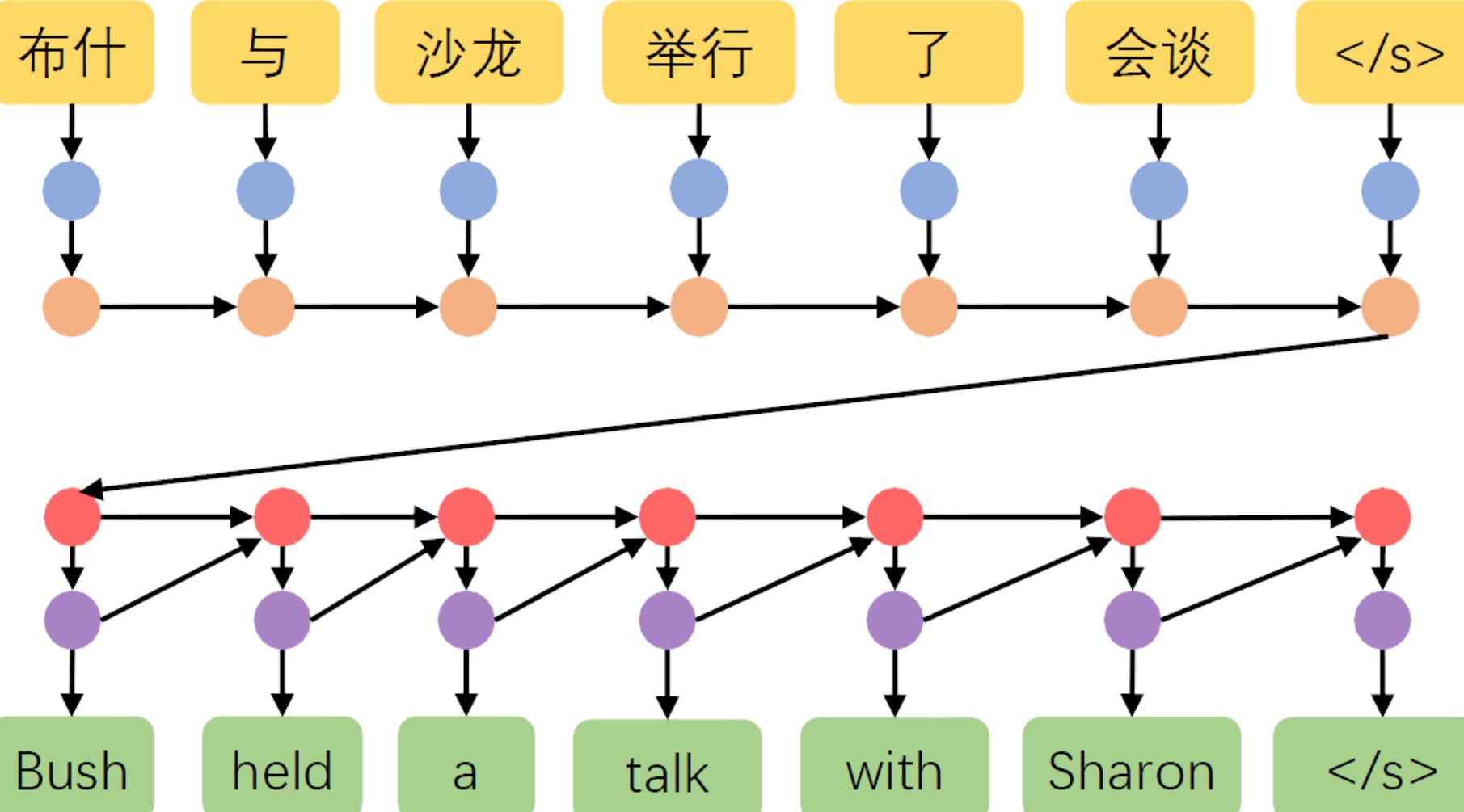
talk

with

Sharon

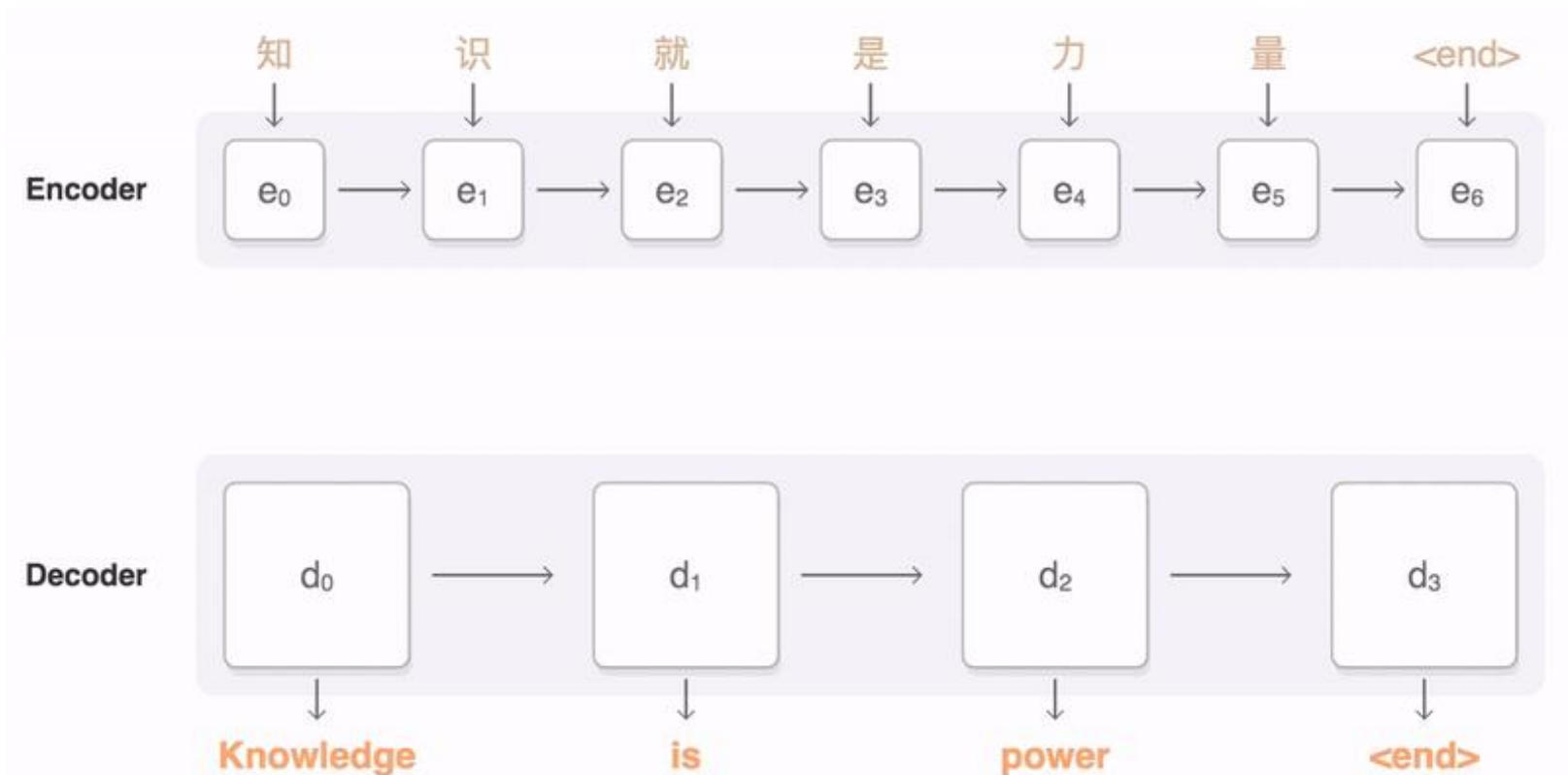


(Och and Ney., 2002)



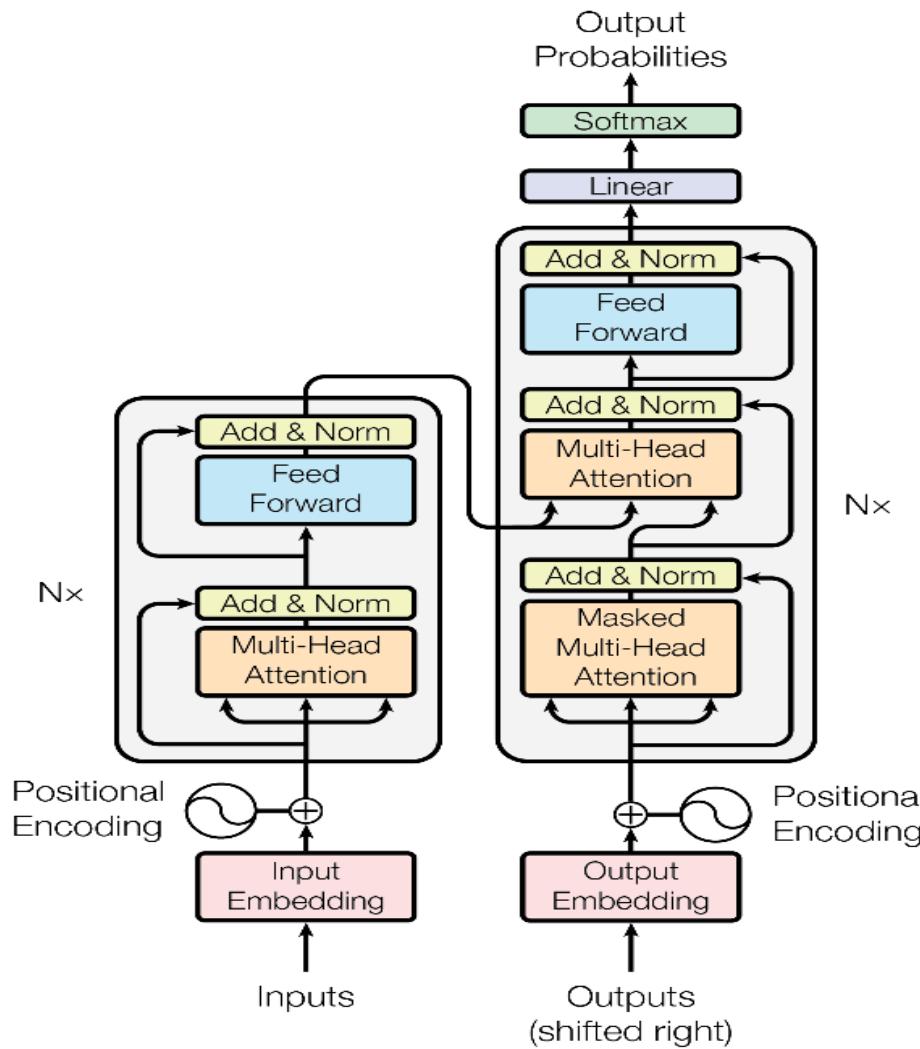
(Sutskever et al., 2014)

NMT - Attention



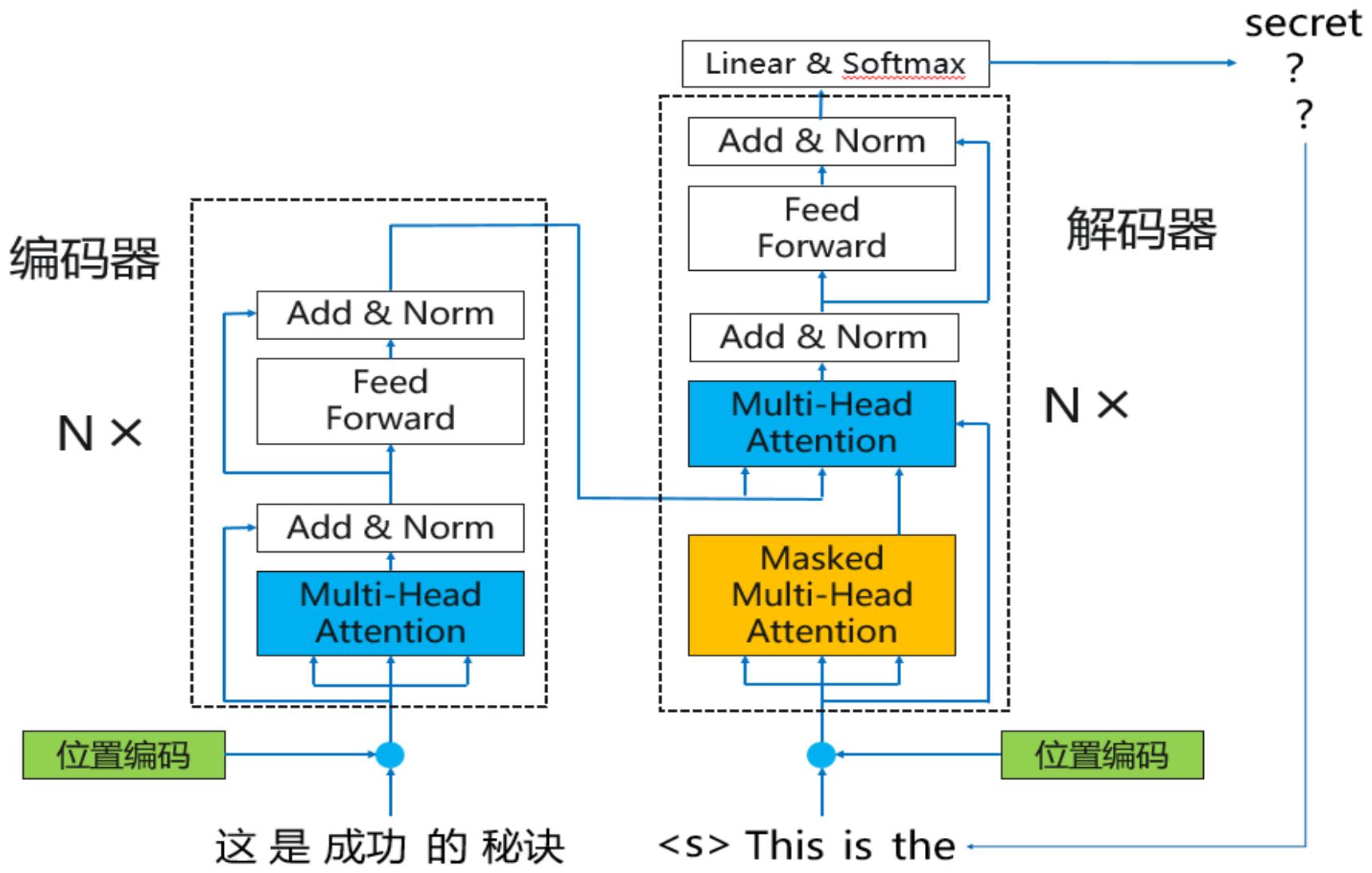
(Wu, et al. 2016)

NMT - Transformer



(Vaswani, et al. 2017)

NMT - Transformer



(Guoping Huang, Qcon2018)

Outline

✓ Machine Translation

● Related Work and Current State for LRLs NMT

● Motivation

● Our works

● Projects

● Copyrights && Patents

● Conclusions





MT Everywhere

清华大学多... 清华大学跨... 多语种翻译... 多语种翻译... 天山网_百度... 天山网 - 新... تەڭرىتاخ تۈرى شى جىنلىك ئۆپەت... لىنچلاسى قۇرغانلا... 1.5.245:9662/index.html

清华大学多语种翻译系统

چىخخۇا ئۇنىۋېرسىتى كۆپ تىللېق تەرجىمە سىستېمىسى

维文 >> 汉语 通用领域 翻译

维文
汉语
藏文
蒙文

ئىنقلابىي قۇربانلارنى خاتىرلەش كۈندە خلق قەھrimانلىرىغا گۈل سېۋىتى تقدىم
مۇراسىمى بېرىجىدا داغدۇغلىق ئۆتكۈزۈلدى.

革命先烈纪念日人民英雄敬献花篮的仪式在北京隆重举行

最多可以输入500个字符

.NLP&CSS group, Tsinghua University :2011-2018 ©
Email: miradel51@126.com Tel: 13051308938 Wechat: 821777278

MT Everywhere



MT Everywhere



تەرجىمانغا بۇ خەتنى چىكىپ كىرىڭ

ئىزدەش تەرجىمانغا يېڭى لۇغەت قوشۇلدى!

中国移动 4G

上午7:49

2%

选择小程序

只能查看已发布小程序的数据

管理员绑定小程序

izdaxFilim 1.9万
3353581344@qq.com 累计用户

汉维翻译 70.0万
3066993171@qq.com 累计用户

izdax翻译 8.8千
3228907610@qq.com 累计用户



Some Issues

天山网讯 10月7日19时15分许

通用领域 翻译

英文 中文简体 阿拉伯语 土耳其语 维语

TianShan Network News October 07 19: 15 Split
Tianshan Network Haber Ekimde 07: 15: 15
91 / 51 7 تشرين الأول / أكتوبر الجلسة 91 / 51 7月8日

“新闻”翻译成“会议” “月份”重翻译
日期漏了，时间也不对
7月8日



Significance of Low Resource MT Research

- Academic value

- Machine translation is a data-driven task, and it relies heavily on parallel corpora. Therefore, the performance is really good on high resource language pairs, but inferior accuracy on low-resource language pairs.

- Application value

- In the “One Belt One Road” work, there is an urgent need for policy communication, smooth trade, financial access, facility connectivity, shared citizenship and cultural exchanges between 65 countries along the route, especially in Central/East/Western Asia.



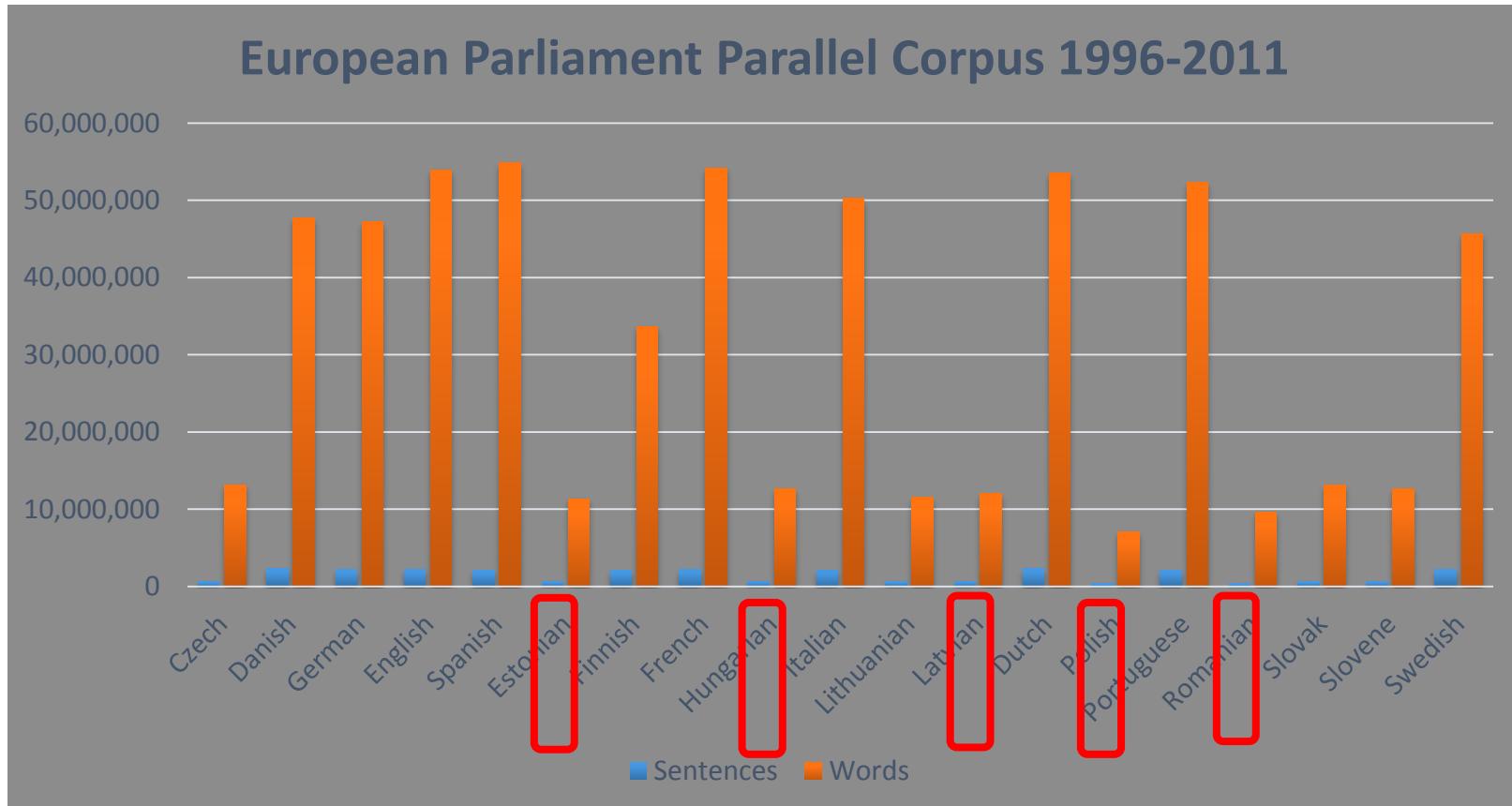
Significance of Low Resource MT Research

- Hot topics concerned by **academia and industry**
- The **national languages** of many countries are almost morphologically rich low-resource languages.



From: <http://www.mrcjcn.com/n/224527.html>

Low-Resource Languages (LRLs)



(Koehn, 2005)

Related Work and Current State

- High-resource Languages
- Domain Adaptation
- Morphology Analyzer
- Data Augmentation
- Transfer Learning
- Zero-Shot Learning
-



High-resource Languages

- Attention Based Multilingual Encoder
(Marton *et al.*, 2009; Nakov, *et al.*, 2012; Dong *et al.*, 2015;
Zoph and Knight *et al.*, 2016; Thanh-Le *et al.*, 2016 ; Schwenk
et al., 2017 ; Johnson *et al.*, 2016 ; Get *al.*, 2018)
- Transfer Learning
(Wang, *et al.*, 2012; Zoph *et al.*, 2016 ; Zoph *et al.*, 2017b ;
Nguyen *et al.*, 2017 ; Chu *et al.*, 2017 ; Passban *et al.*, 2017 ;
Dabre *et al.*, 2017; Wang *al.*, 2018)



Domain Adaptation

- Data Selection
 - (Foster *et al.*, 2007 ; Zhao *et al.*, 2007; Lü *et al.*, 2007 ; Moore *et al.*, 2010 Axelrod *et al.*, 2011; Lewis *et al.*, 2011; Duh *et al.*, 2013; Chen *et al.*, 2016 ; Ruder *et al.*, 2017)
- Context Information
 - (Tiedemann *et al.*, 2010; Gong *et al.*, 2011; Carpuat1 *et al.*, 2013)
- Topic Information
 - (Tam *et al.*, 2007 ; Xiao *et al.*, 2012 ; Su *et al.*, 2012 ; Eidelman *et al.*, 2012; Hewavitharana, *et al.*, 2013; Zhang *et al.*, 2014 ; Zhang *et al.*, 2016)



Morphology Analyzer

- Rule Based

(Maddox *et al.*, 2003; Daybelge *et al.*, 2007; Lignos *et al.*, 2009;
Hatem, *et al.*, 2011; Kessikbayeva *et al.*, 2015)

- Traditional Statistical Method

(Kudo *et al.*, 2004; Creutz *et al.*, 2006; Virpioja *et al.*, 2013; Stig-Arne *et al.*, 2014 ; Kohonen *et al.*, 2010 ; Ruokolainen *et al.*, 2014; Sennrich *et al.*, 2016a)

- Neural Network Method

(Stuskever *et al.*, 2014; Bahdanau *et al.*, 2015; Wu *et al.*, 2016;
Vaswani *et al.*, 2017; Belinkov *et al.*, 2017; Vania *et al.*, 2017;
Rajana *et al.*, 2017)



Data Augmentation

- Monolingual Based
 - (Koehn *et al.*, 2002; Quirk *et al.*, 2004; Ueffing, *et al.*, 2006; Marta, *et al.*, 2006; Wubben *et al.*, 2012; Gulchere *et al.*, 2015; Sennrich *et al.*, 2016b ; Cheng *et al.*, 2016b ; Zhang *et al.*, 2018)
- Word Level Replacement
 - (Francis *et al.*, 2009; Fadaee *et al.*, 2017 ; Huang *et al.*, 2016; Sennrich *et al.*, 2016c; Ribeiro *et al.*, 2018; Wang *et al.*, 2018)



Challenges for LRLs MT

- Challenges
 - Unable to make efficient use of high resource language issue
 - Specific domain adaptive problem
 - Unable to analyze morphological problems efficiently
 - Syntactic and Semantic error problem after data augmentation

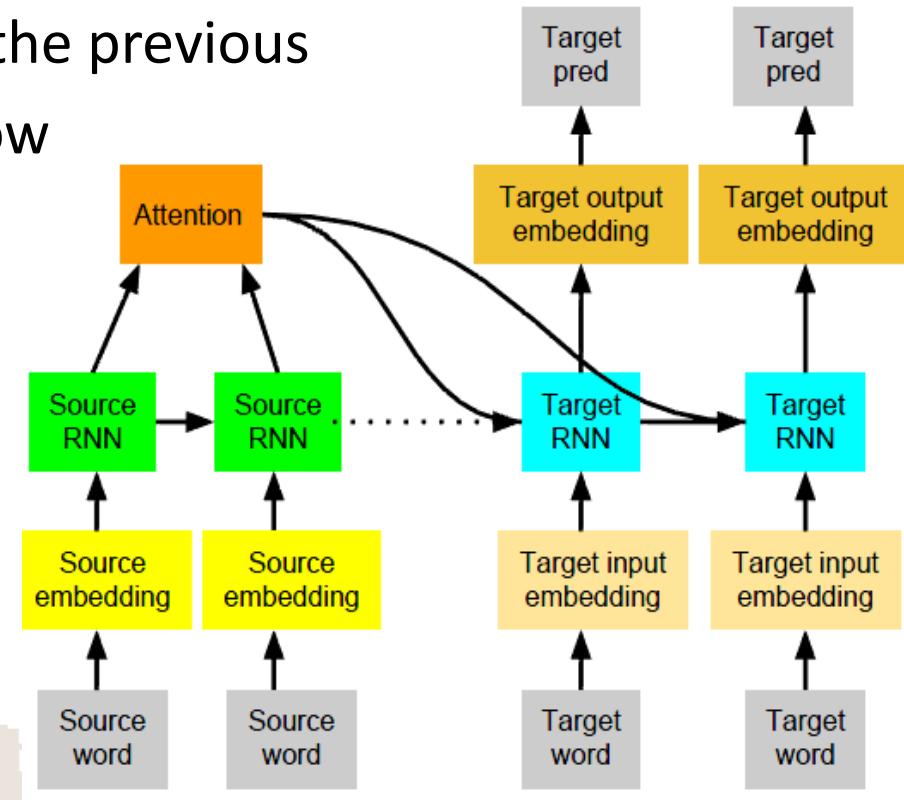


Challenge 1-Using High-resource Languages

- (Zoph *et al.*, 2016) exploited the transfer learning on high-resource languages to help LRLs.
- (Passban *et al.*, 2017) based on the previous idea, by using of one HRLs from two different domains.

Issue:

- Both of them unable to **use efficiently** multiple HRLs.
- Ignored **character level similarity**.

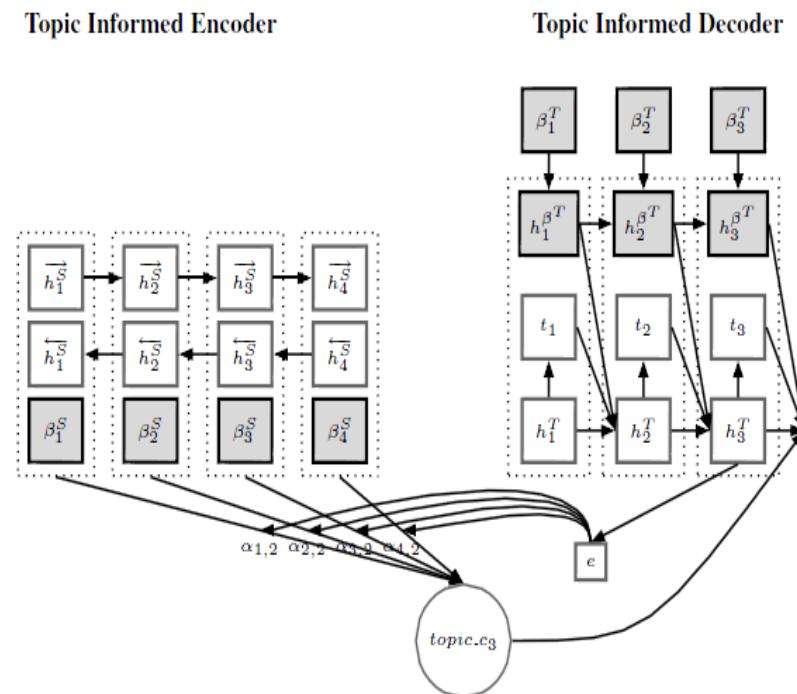


Challenge2-Specific Domain Adaptive issue

- (Zhang *et al.*, 2016) we can regard their idea was fully-supervised method, first learn the topic

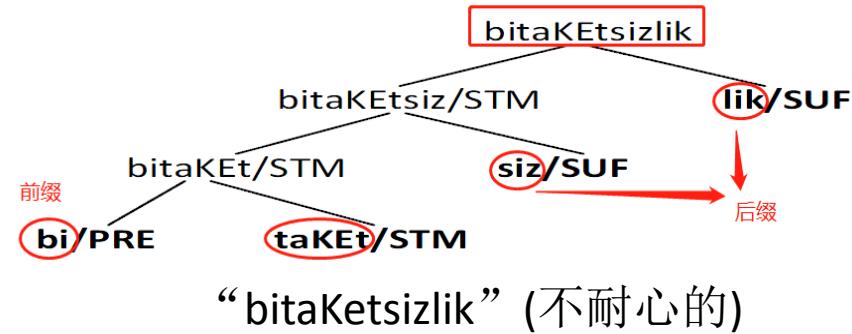
Information by using LDA, then feed them into NMT.

- **Issue:**
 - Unable to **jointly train** the topic information and NMT.
 - Dependently train the **topic model** and time consuming.
 - Hard to **set the topic number**, heuristically set the topic number both on encoder and decoder of NMT.



Challenge3-Inferior Accuracy of Morphological Analyzer

- (Virpioja *et al.*, 2013) exploit the semi-supervised method to segment, and proposed language independent lexicon analyzer Morfessor2.0
- (Sennrich *et al.*, 2016a) used the greedy algorithm to compute the state of each characters to be connected with others, and proposed the BPE.



- Issue :
 - Unable to **coverage** the language knowledge
 - Still exist **over /imperfect /non** segmentation

r ·	→	r ·
l o	→	l o
l o w	→	l o w
e r ·	→	e r ·

{‘low’, ‘lowest’, ‘newer’, ‘wider’}



Challenge 4-Syntactic and Semantic Errors

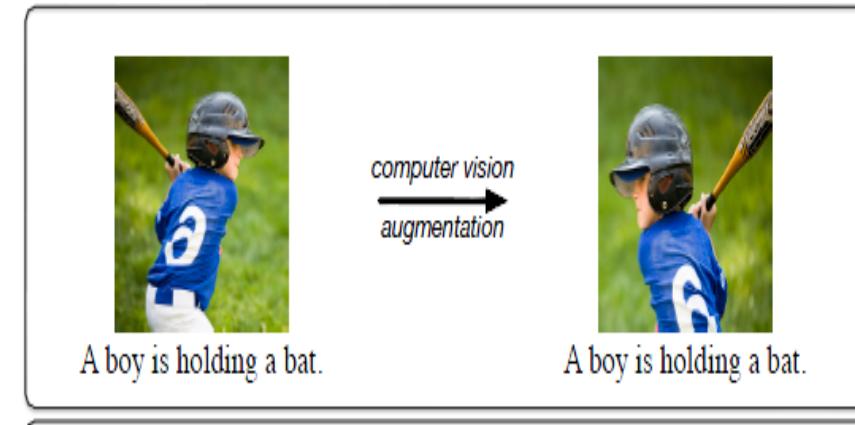
- (Fadaee *et al.*, 2017) exploited the very common augmentation method used in CV, namely data augmentation.
- (Cheng *et al.*, 2016b) expand the corpus size via back translation on monolingual data.

Semantics: John waters the [Plant/Bike]

Syntax: I have three [bags/pencil]

- Issue:

- Relied on existed translation model.
- Unable to address the syntactic and semantic errors efficiently after data augmentation.



original pair	augmented pair
$S : s_1, \dots, s_i, \dots, s_n$	$S' : s_1, \dots, s'_i, \dots, s_n$
$T : t_1, \dots, t_j, \dots, t_m$	$T' : t_1, \dots, t'_j, \dots, t_m$

Outline

✓ Machine Translation

✓ Related Work and Current State for LRLs NMT

● Motivation

● Our works

● Projects

● Copyrights && Patents

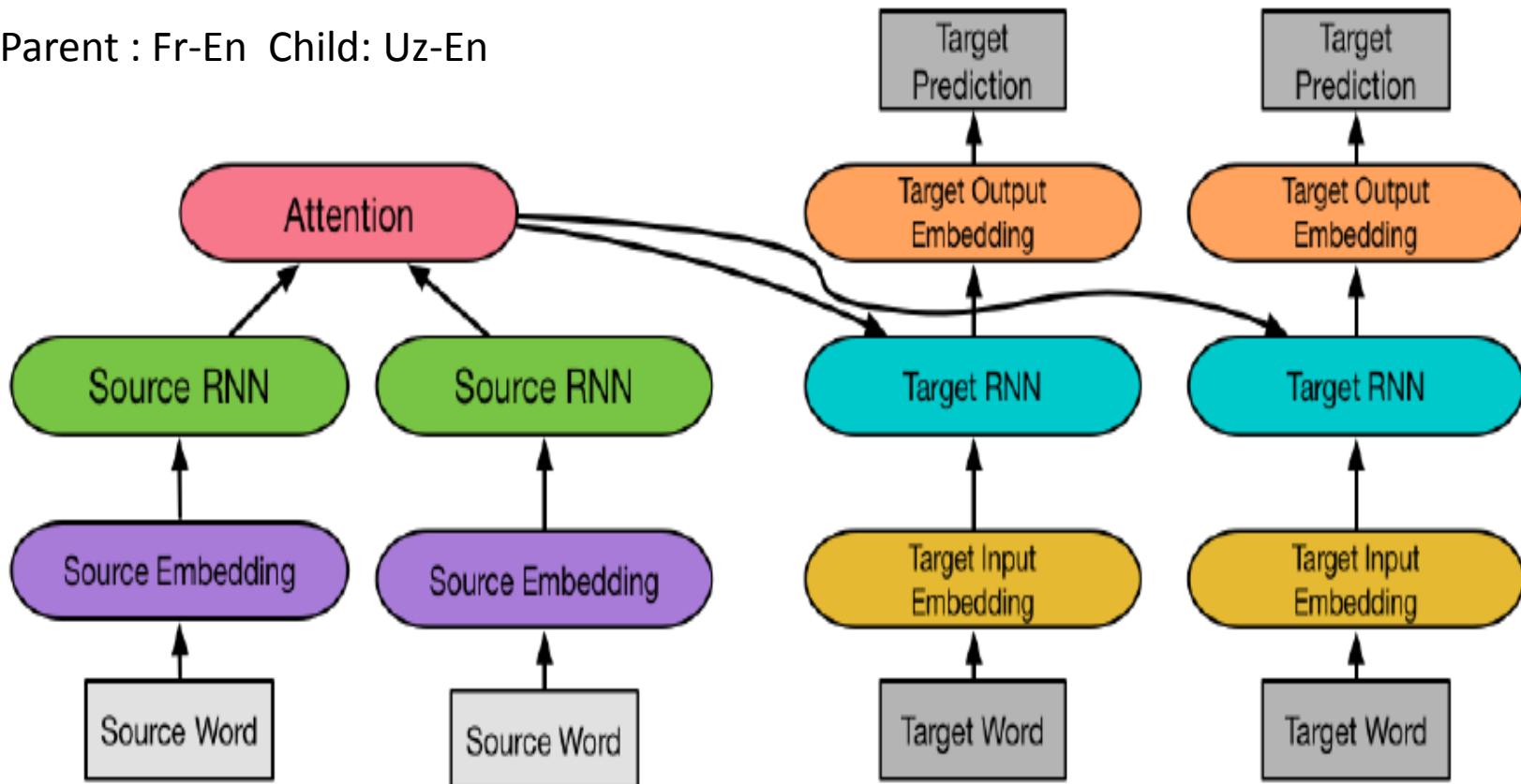
● Conclusions



Motivation – Transfer Learning

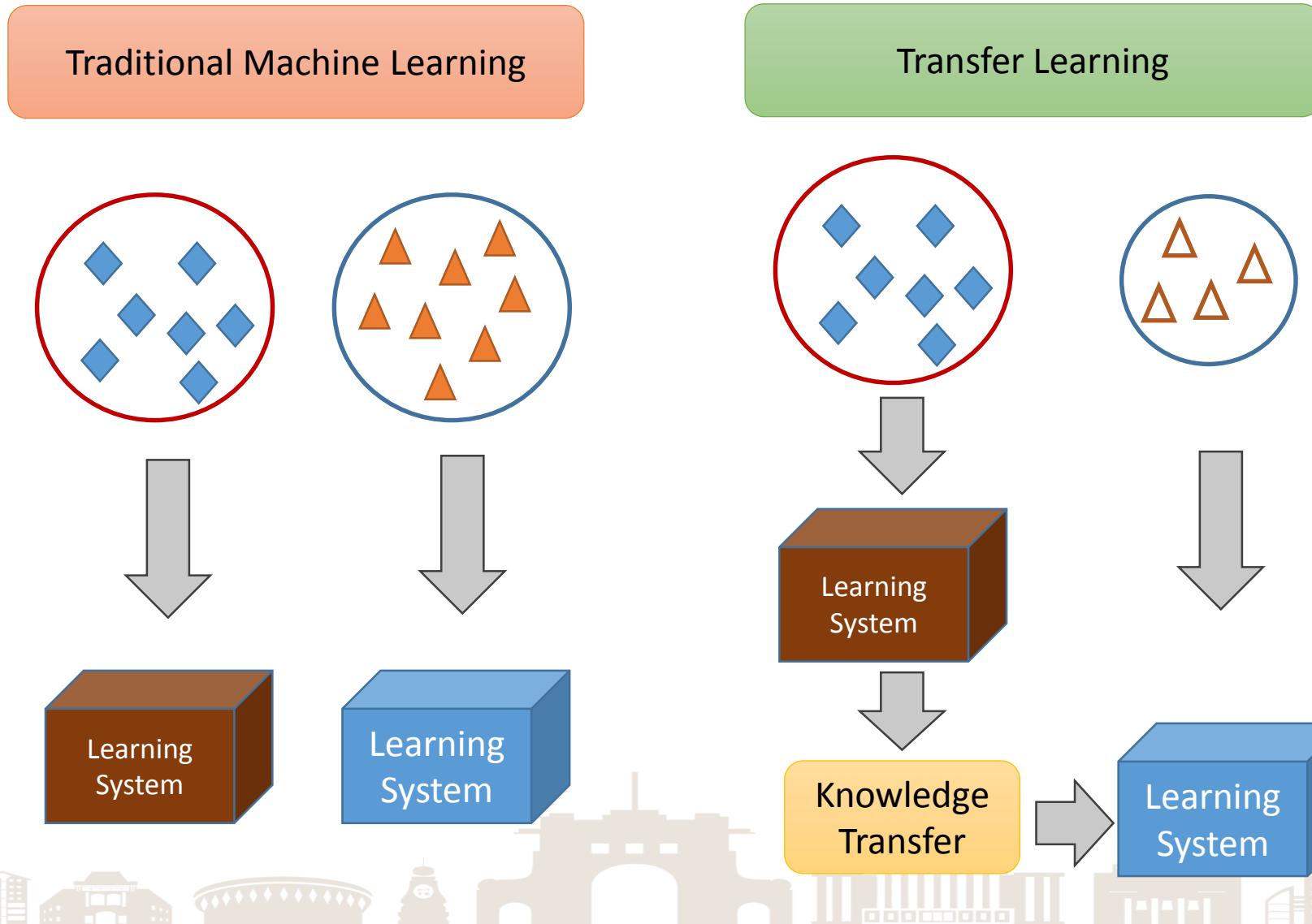
Optimal setting for transferring from **parent** model to **child** model.

Parent : Fr-En Child: Uz-En



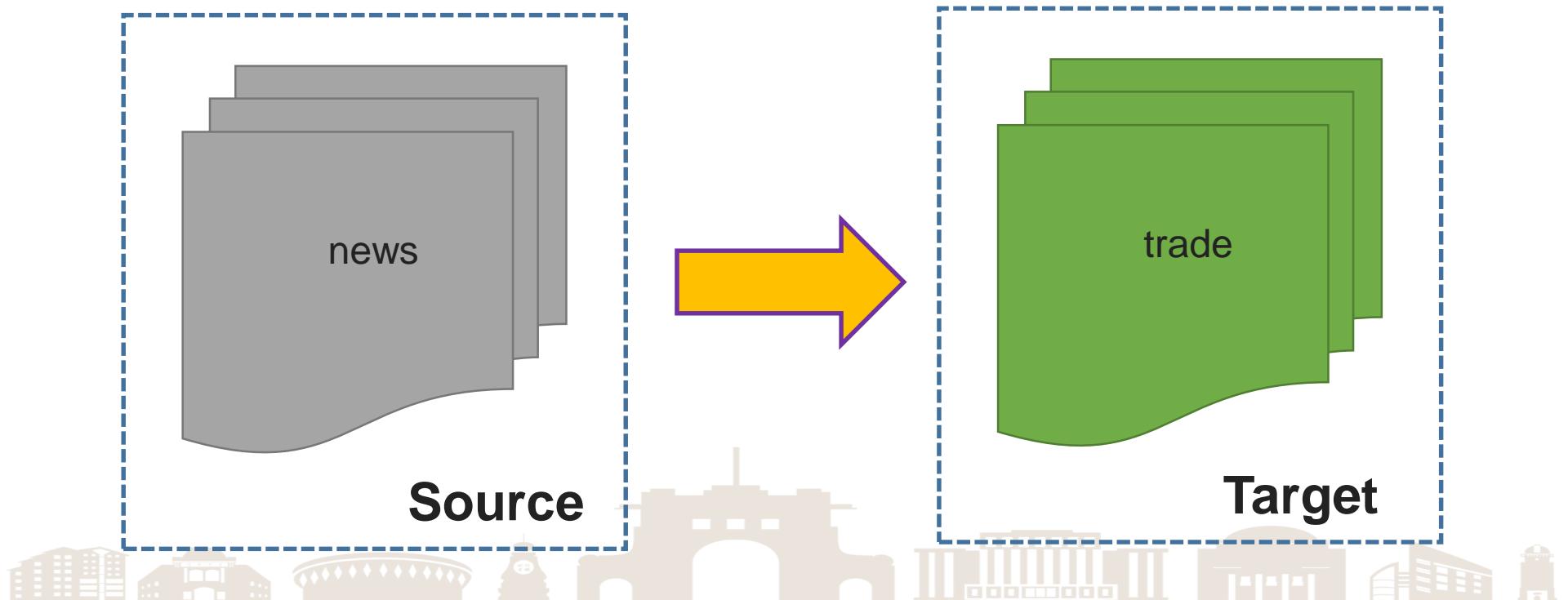
[Barret Zoph et al., 2016]

Transfer Learning & Machine Learning



Transfer Learning – domain adaptation

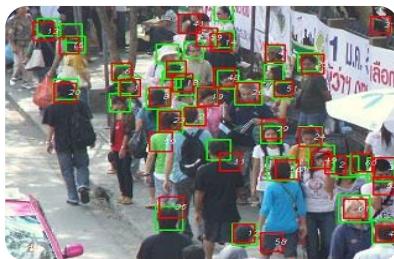
This scenario arises when we **aim at** learning from a **source** data distribution a well performing model on a **different** (but related) **target** data distribution.



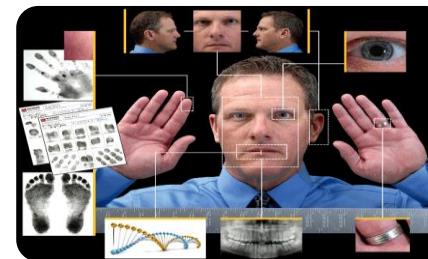
Transfer Learning – domain adaptation

In Natural Language Processing (NLP), train a system on some language data, retune && apply it to specific different task.

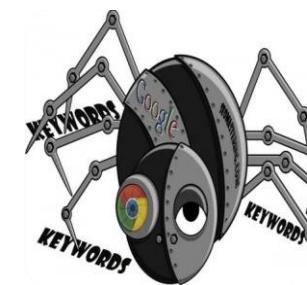
Build speech recognition system using **recorded phone calls**, then tune it to use as an **airline reservation hotline**.



CV



ER



IR



ASR



Outline

✓ Machine Translation

✓ Related Work and Current State for LRLs NMT

✓ Motivation

● Our works

● Projects

● Copyrights && Patents

● Conclusions





Multi-Round Transfer Learning for Low-Resource NMT Using Multiple High-Resource Languages

Mieradilijiang Maimaiti¹, Yang Liu¹, Huanbo Luan¹, Maosong Sun¹

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China



Methodology - NMT

- X, Y ; Raw source and target sentences.
- Given source sentences $X = x_1, \dots, x_i, \dots, x_I$ and target sentence $Y = y_1, \dots, y_i, \dots, y_I$
- Standard NMT models usually factorize the sentence-level translation probability as a product of word-level probabilities:
- $P(y|x; \theta) = \prod_{j=1}^J P(y_i|x, y_{<j}; \theta)$
- θ is model parameters, $y_{<j}$ is partial translation.



Methodology – Transfer Learning

- We take the $L_3 \rightarrow L_2$ as parent and $L_1 \rightarrow L_2$ as child language pair. L_3 and L_1 are source languages of parent and child ,respectively, L_2 is the target language for both.

- $\theta_{L_3 \rightarrow L_2} = \{<e_{L_3}, W, e_{L_2}>\}$ while e_{L_3} and e_{L_3} source and target embedding of parent model, W is parameters.

- $\hat{\theta}_{L_3 \rightarrow L_2} = \operatorname{argmax}_{\theta_{L_3 \rightarrow L_2}} \{L(D_{L_3}, \theta_{L_3 \rightarrow L_2})\}$ train the parent model $M_{L_3 \rightarrow L_2}$

- Then fine-tune the child model $M_{L_1 \rightarrow L_2}$ with parent model $M_{L_3 \rightarrow L_2}$:

- $\theta_{L_1 \rightarrow L_2} = f(\hat{\theta}_{L_3 \rightarrow L_2})$, while f is initialization function.

Main Idea

- we aim to deal with the problem of how to make full use of these corpora of highly related **multiple languages**, to increase the translation quality of the child model.
- Increase the similar even identical words between parent and child language by using **unified transliteration method**.



Original Transfer Learning

- The original TL transfers parameters of parent model into child model.

$$\theta_{L_3} = \{\langle e_{L_3}, W, e_{L_3} \rangle\}$$

$$\hat{\theta}_{L_3 \rightarrow L_2} = \operatorname{argmax}_{\theta_{L_3 \rightarrow L_2}} \{L(D_{L_3 \rightarrow L_2}, \theta_{L_3 \rightarrow L_2})\}$$

$$\theta_{L_1 \rightarrow L_2} = f(\hat{\theta}_{L_3 \rightarrow L_2})$$



[Barret Zoph et al., 2016]

Modified Transfer Learning

- The modified TL transfers parameters of parent model into child model from one parent with different domains.

$$\theta_{\hat{L}_3 \rightarrow L_2} = f(\hat{\theta}_{L_3 \rightarrow L_2})$$

$$\hat{\theta}_{\hat{L}_3 \rightarrow L_2} = \operatorname{argmax}_{\hat{L}_3 \rightarrow L_2} \left\{ L(D_{\hat{L}_3 \rightarrow L_2}, \theta_{\hat{L}_3 \rightarrow L_2}) \right\}$$

$$\theta_{L_1 \rightarrow L_2} = f(\hat{\theta}_{\hat{L}_3 \rightarrow L_2})$$



[Passban et al., 2017]

Multi-round Transfer Learning

- The **central idea** of our proposed MRTL is to encourage the child model receive more information from different parent models.

$$\theta_{L_4 \rightarrow L_2} = f(\hat{\theta}_{L_3 \rightarrow L_2})$$

$$\hat{\theta}_{L_4 \rightarrow L_2} = \operatorname{argmax}_{\theta_{L_4 \rightarrow L_2}} \{L(D_{L_4 \rightarrow L_2}, \theta_{L_4 \rightarrow L_2})\}$$

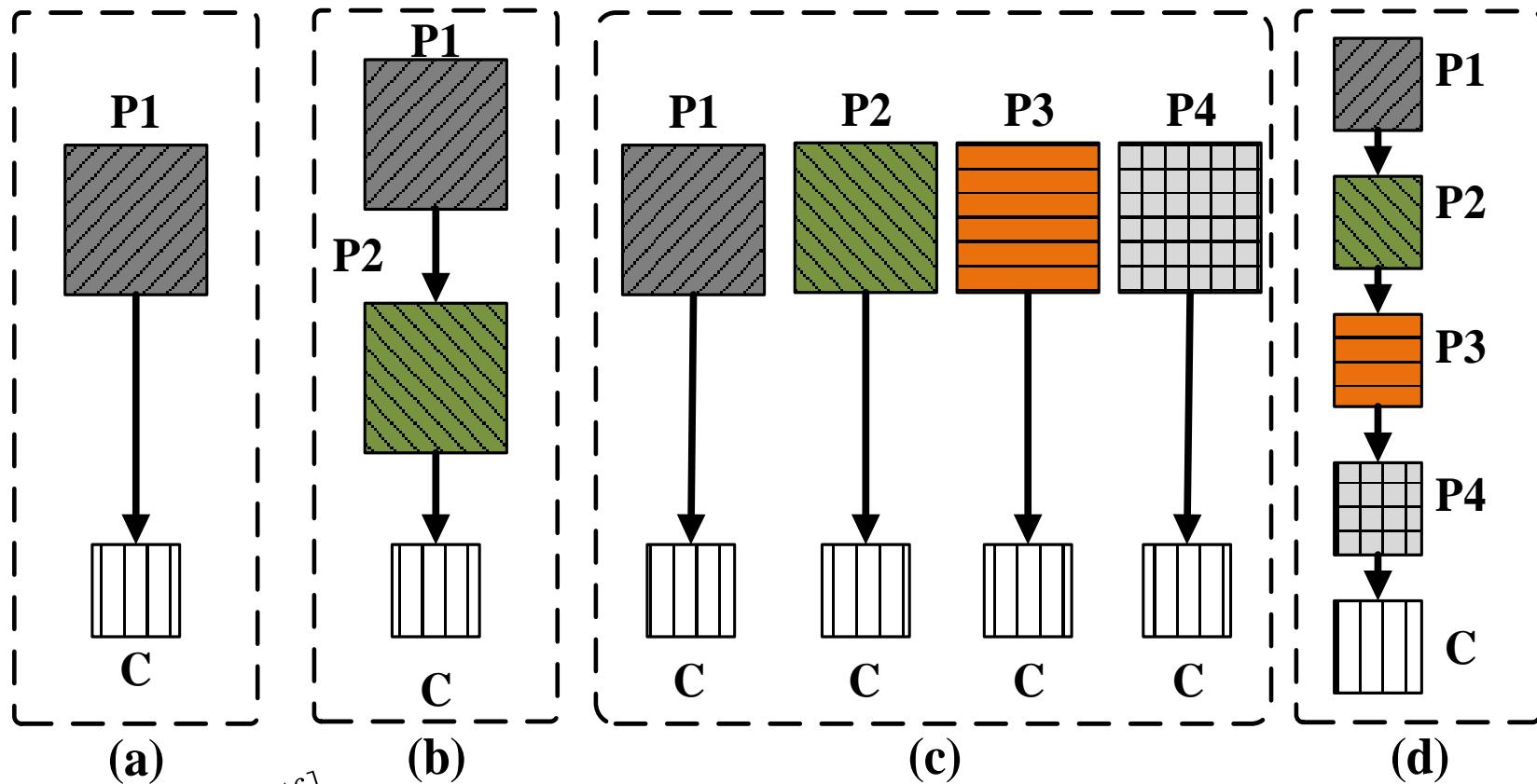
$\vdots = \vdots$

$$\theta_{L_{k+1} \rightarrow L_2} = f(\hat{\theta}_{L_k \rightarrow L_2})$$

$$\hat{\theta}_{L_{k+1} \rightarrow L_2} = \operatorname{argmax}_{\theta_{L_{k+1} \rightarrow L_2}} \{L(D_{L_{k+1} \rightarrow L_2}, \theta_{L_{k+1} \rightarrow L_2})\}$$

$$\theta_{L_1 \rightarrow L_2} = f(\hat{\theta}_{L_{k+1} \rightarrow L_2})$$

Multi-round Transfer Learning



This work

Language pairs

Language features of all languages used in our experiments

Language		Family	Group	Branch	Order	Unit	Inflection
Arabic	(Ar)	Hamito-Semitic	Semitic	South	VSO	Word	High
Farsi	(Fa)	Indo-European	Indic	West	SOV	Word	Moderate
Urdu	(Ur)		Iranian	Iranian	SOV	Word	Moderate
Finnish	(Fi)	Uralic	Finno-Ugric	Finnish	SVO	Word	Moderate
Hungarian	(Hu)			Ugric	SVO	Word	Moderate
Turkish	(Tr)	Altaic	Turkic	Oghuz	SOV	Word	Moderate
Uyghur	(Uy)			Qarluq	SOV	Word	Moderate
Chinese	(Ch)	Sino-Tibetan	Chinese	Sinitic	SVO	Character	Light



Corpora

Characteristics of all languages parallel corpora.

Languages	Train	Dev	Test	Source		Target	
				Vocab.	# Word	Vocab.	# Word
Ar → Ch	5.1M	2.0K	2.0K	1.0M	32.2M	0.5M	37.4M
Fa → Ch	1.4M	2.0K	1.0K	0.2M	10.4M	0.2M	10.0M
Ur → Ch	78.0K	1.0K	1.0K	17.6K	2.6M	12.7K	2.4M
Fi → Ch	2.8M	2.0K	1.0K	0.7M	18.4M	0.3M	23.1M
Hu → Ch	4.1M	2.0K	1.0K	1.0M	30.4M	0.5M	32.5M
Tr → Ch	4.4M	2.0K	1.0K	0.7M	30.6M	0.5M	35.9M
Uy → Ch	46.3K	1.0K	1.0K	73.5K	1.1M	42.1K	11.2M

The “Vocab.” and “# Word” represent vocabulary (word type) and word token, respectively



Shared words

The shared words between each languages

	Ar	Fa	Ur	Fi	Hu	Tr	Uy
Ar		11.49%	8.31%	0.52%	0.45%	0.73%	0.77%
Fa	2.34%		8.29%	0.27%	0.30%	0.32%	0.57%
Ur	0.15%	0.75%		0.01%	0.01%	0.03%	0.11%
Fi	0.36%	0.94%	0.53%		2.74%	3.80%	0.50%
Hu	0.45%	1.46%	0.70%	3.85%		5.07%	0.75%
Tr	0.57%	1.22%	1.14%	4.22%	4.01%		2.47%
Uy	0.06%	0.21%	0.47%	0.05%	0.06%	0.24%	

The Ar, Fa, Ur, and Uy are converted with proposed unified transliteration method. Besides, shared word rate calculated as the division of shared word numbers to word type counts of the language in each column.



Methodology – Unified Transliteration

The shared words between each languages

Language	Original	Latin	Chinese	English	Unified
Ar	مدرسة	maktab			
Uy	مەكتەپ	mektep	学校	School	mektep
Tr	okul	mektep			
Fa	باغ وحش	bağça			
Uy	باغچا	bağça	果园	Orchard	bagça
Tr	bahçesi	bahçe			
Ar	غرفة القراءة	qiraaaatxana			
Fa	اتاق مطالعه	qiraaaat xana	阅览室	Reading room	qiraetxana
Tr	Okuma odası	kuraathane			
Uy	قراءة تخانى	qiraetxana			

The second column “Original” represents prototype scripts of corresponding languages. Besides, the last column “Unified” stands for the converted format with unified transliteration method.



Methodology – Unified Transliteration

Algorithm 1: Unified Transliteration Method

```

Input: the source side monolingual sentences  $D_{sm} = \{x_{sm}^m\}_{m=1}^M$  of parent (child).
Output: the transliterated word sequence in current sentences  $D'_{sm}$ .
/* Initialize the variables.
Currentl ← the word sequence in current line among  $D_{sm}$ ;
Outputc ← the transliterated word sequence in current line among  $D'_{sm}$ ;
Read the source side monolingual sentences  $D_{sm}$  of parent (child);
for each line in  $D_{sm}$  do
    /* the current line should be decoded as ‘‘utf-8’’.
    Currentl ← each line.decode('utf-8');
    /* split the current line with white space and save them as a list.
    Currentl ← Currentl.strip().split();
    for each word in Currentl do
        for each char in each word do
            /* check each char from the manually prepared mapping table.
            each charlatin ← each char;
        end
        /* check the each word if contains same repeated char continually.
        if Is Contain repeated char in each charlatin then
            compare the length of Currentl and the length of Latinized Currentl;
            remove repeated each charlatin from each word;
        end
        convert them into unified form sequentially;
        each word ← sums of unified chars after removing repeated chars;
        final word ← each word ;                                /* reserve the final word.
    end
    Outputc ← the joint sequence of final words with white-space
end

```

Experiment

- System : THUMT
- Parameters :
 - Dropout 0.1
 - Word Embedding 620
 - Hidden State 1000
 - Vocabulary : **source 3w, target 29k**
- Other parameter we use the default parameters of THUMT
- Preprocess
 - Clear data use [NiuTrans](#) , Chinese segmenter use [THULAC](#)
 - Some processing tools designed by us



Experiment

- GPU: 4
- GPU type: NVIDIA TITAN X (Pascal)
- Training time : less than 3-4 days (include fine-tuning)
- Corpus :
 - Parent : Open Subtitle2016 and Tanzil corpora
 - Child : Chinese-LDC (CLDC) corpus
- UNK replace : NO
- BPE : Yes



Experiment

The Effect of unified transliteration method for MRTL.

Method	Round	Parent	Child	BLEU
TRANSFORMER	R=0	N/A		28.28
MRTL (Original)		Ur → Ch		10.29
		Fa → Ch		28.83
	R=1	Ar → Ch	Uy → Ch	30.64 ⁺⁺
		Ur → Ch		10.93*
MRTL (Unified)		Fa → Ch		29.96 ^{****}
		Ar → Ch		31.64 ^{++*}



Experiment

The Effect of corpus size to child model in single fine-tuning.

Method	Parent	Child	BLEU
TRANSFORMER	N/A		28.28
MRTL (R=1)	Tr → Ch (0.5M)	Uy → Ch	29.89 ⁺⁺
	Tr → Ch (2.4M)		30.88 ^{++b}
	Tr → Ch (4.4M)		32.74 ^{++h}



Experiment

The effect of parent language pairs to

Method	Parent	Child	BLEU
TRANSFORMER	N/A		28.28
	<i>Ur</i> → <i>Ch</i>		10.93
	<i>Fa</i> → <i>Ch</i>		29.96 ⁺⁺
	<i>Fi</i> → <i>Ch</i>	<i>Uy</i> → <i>Ch</i>	30.85 ⁺⁺
MRTL (R=1)	<i>Tr</i> → <i>Ch</i> (2.4M)		30.88 ^{++*}
	<i>Ar</i> → <i>Ch</i>		31.64 ^{++*}
	<i>Hu</i> → <i>Ch</i>		32.41 ^{++o}
	<i>Tr</i> → <i>Ch</i> (4.4M)		32.74 ^{++†}



Experiment

Parent Language Selection

Different language **family**

MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Uy → Ch	28.28
R=1	Hu → Ch Tr → Ch	Uralic Altaic	Open Subtile	4.1M		32.41 ⁺⁺ 32.58 ⁺⁺

Different **domain**

MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Uy → Ch	28.28
R=1	Ur → Ch Fa → Ch	Indo-European	Tanzil Open Subtitle	78.0K		10.93 24.27

Different **corpus size**

MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Uy → Ch	28.28
R=1	Fi → Ch Hu → Ch	Uralic	Open Subtile	2.8M 4.1M		30.85 ⁺⁺ 32.41 ⁺⁺



Experiment

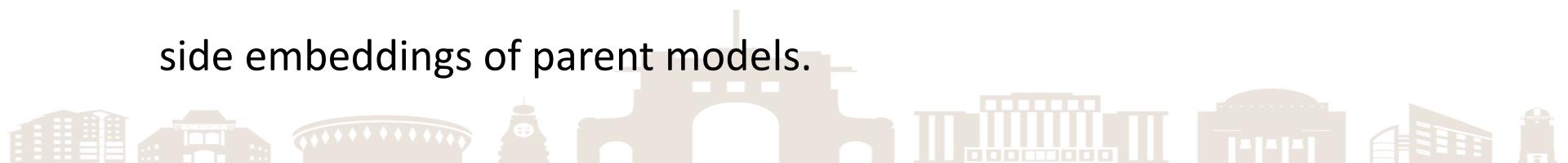
The Effect of MRTL.

Method	Round	Parent	Child	BLEU
TRANSFORMER	R=0	N/A		28.28
MANY-to-ONE				32.43 ⁺⁺
MRTL	R=1	Tr (4.4M) → Ch		32.03 ⁺⁺
	R=2	Tr (4.4M), (2.4M) → Ch		32.54 ⁺⁺
	R=3	Tr (4.4M), (2.4M), Fi → Ch	Uy → Ch	33.54 ^{++‡*}
		Tr (4.4M), (2.4M), Fi, Hu → Ch		33.66 ^{++**}
	R=4	Ar (Unified), Tr (4.4M), Hu, Fi → Ch		33.73 ^{++**}
		Tr (4.4M), Ar (Unified), Hu, Fi → Ch		33.91 ^{++**}



Conclusions

- We address the **drawbacks of TL**, which exploits **only** one parent to optimize the child model at a time.
- We **mitigate** the gap between parent and child language pairs at the **character** level.
- We achieve **transparency** in network architectures, as well as in our method for neural network architecture.
- We observe meaningful discovery **by sharing both** source side and target side embeddings of parent models.





Discussion on Bilingual Cognition in International Exchange Activities

Mieradilijiang Maimaiti¹ and Xiaohui Zou¹

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

² Sino-American Saerle Research Center, Beijing, China

ICIS2018, Beijing





How to Understand: Three Types of Bilingual Information Processing?

Mieradilijiang Maimaiti¹, Shunpeng Zou², Xiaoqun Wang³ and Xiaohui Zou²³⁴

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

² China University of Geosciences (Beijing) 29 Xueyuan Road, 100083, China

³ Audio-Visual Building, Room 415, Peking University, 100871 Beijing, China

⁴Sino-American Searle Research Center, UC Berkeley 94720-3840, USA

ICCSIP2018, Beijing



Motivation: human-machine translation

Source sentence:

At the meeting on UN Operational Activities for Development, Wang also stressed that developed countries should bear the primary responsibility for financing for development.

Machine translation:

联合国发展业务活动的一次会议上，王汉斌还强调，发达国家应发展筹资问题负有主要责任。



Final translation:

在联合国发展业务活动的会议上，王还强调，发达国家在发展筹资问题上应负主要责任。



Motivation: human-machine translation

Source sentence:

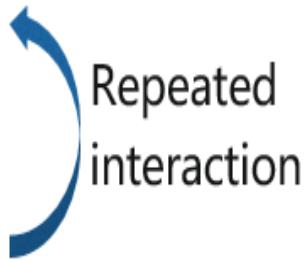
At the meeting on UN Operational Activities for Development, Wang also stressed that developed countries should bear the primary responsibility for financing for development.

Human partial input:

在联合国发展业务活动的

Human-machine interact translation:

在联合国发展业务活动的一次会议上，王汉斌还强调，发达国家应发展筹资问题负有主要责任。



Final translation:

在联合国发展业务活动的会议上，王还强调，发达国家在发展筹资问题上应负主要责任。



(Guoping Huang, Qcon2018)

Methodology – three types of bilingualism

- Narrow bilingualism

- Chinese – English

- Alternative bilingualism

- Terminologies and popular sayings

- Generalized bilingualism

- Mathematical language of arithmetic numbers and natural language of Chinese characters



Methodology

- Replacing
- Switching
- Dropping
- Editing
- Adding

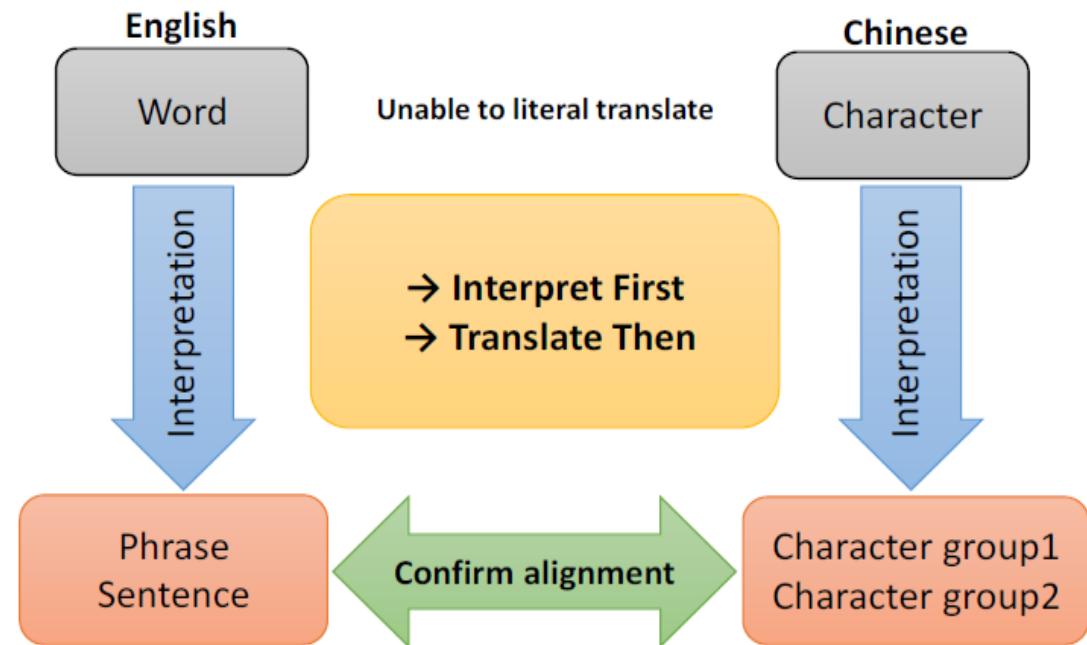


Fig. 1. The sketching graph for narrow bilingual information processing model.



Methodology

Algorithm 1 Bilingual Information Processing.

Input: Given bilingual corpus $L, L = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^N$

Output: translation result y

- 1: **for** $t = 1, 2, 3, 4, 5, \dots, N$ **do** **do**
 - 2: Generate “Phrase Set” and modify sentence structures
 - 3: Select some words and modify its position among sentence S
 - 4: Remove the original input from the testset
 - 5: Feed the revised sentence S' to the testset
 - 6: Translate final input
 - 7: **end for**
-



Our work (Word segmentation)

[This slide intentionally left blank]

Not good as we expected



Our work (Data Augmentation)

[This slide intentionally left blank]

We have been doing corresponding experiments...



Outline

- ✓ Machine Translation
- ✓ Related Work and Current State for LRLs NMT
- ✓ Motivation
- ✓ Our works
- Projects
- Copyrights & Patents
- Conclusions



Projects

- Project1: “少数民族网络舆情综合分析与云服务关键技术研究及应用示范”
- Project2: “跨语言社会舆情分析基础理论与关键技术
- Project3: “面向三元空间的互联网中文信息处理理论与方法”
- Project4: “基于深度学习的维汉机器翻译”





清华大学多语种翻译系统

چىخۇا ئۇنىۋېرسىتې كۆپ تىللۇق تەرىجىمە سىستېمىسى

维文 >> 汉语 ▾ 通用领域 翻译

ئىنقالابى قۇرۇبانلارنى خاتىرىلەش كۈنىدە خەلق قەھرىمانلىرىغا گۈل سېۋىتى تەقدىم مەسىم سىجىخىدا دادىعە غۇلەتىكە ئۆلە

革命先烈纪念日人民英雄敬献花篮的仪式在北京隆重举行



ئېنقالابىسى قۇربانلارنى خاتىرىلەش كۈندە خالق قەھىرىمانلىرىغا گول سېۋىتى تەقدىم قىلىش مۇراسىمى بېبىجىڭدا داغلىقىلىق ئەتكىنلىك ئەتكىنلىك

2018/10/01 15:37

مُؤْهَهِ رَدِّي : قُوَّةُ بَانِجَانِ، قَبْيُوم

حنسیک، لم کیحالا، لم جدنشه، ۋالك بالك، ۋالك خننیك، حاڭ لىحر، خەن حىڭ، ۋالك حىشدىن قاتىاشتە.

最多可以输入500个字符

.NLP&CSS group, Tsinghua University :2011-2018 ©

Email: mirade@126.com Tel: 13051308938 Wechat: 821777278



清华大学跨语言信息检索系统

查询语言

中文

文档语言

维吾尔语 藏语 蒙古语

请输入关键词...



点击率	日期	标题	编号
261 ئاۋانلىقى:	2018-07-26	شى جىنپىڭ كېسەك ئالىن دۈلەتلىرى سودا سىلاتەت مۇنىرىگە ئاتىتى ھەم مۇھەم سۆز قىلدى	1
229 ئاۋانلىقى:	2018-07-26	شى جىنپىڭ راما قۇسا زۇڭۇق ئۆتكۈزگەن فارشى ئېلىش ۋە جۈڭگۈ - جەنۇنى ئافرۇقا دىيلەمانىڭ...	2
173 ئاۋانلىقى:	2018-07-26	سىرىيەن شەق كىچىك ئەزىز بوللازىدۇغا كىرسى مۇمكىن	3
149 ئاۋانلىقى:	2018-07-26	داۋ ئىنس سۇغۇرۇنى مالىيە ياردىم بىللى ئۇلۇممى كىشى بېشىغا يېڭىدىن 40 يۇمن قوشۇلدى	4
123 ئاۋانلىقى:	2018-07-26	گۈزۈلۈن ئەكتەرۈش گۈرۈيىسى حىلىن ئۆلکەسىدە بىرلىك بىلەن ئەكتەرۈپ بىر ئەرىب قاش خىزمىتلى...	5

5.196.9666/index.html



多语种翻译系统

كۆپ تىللېق تەرجىمە سىستېمىسى

维吾尔语



汉语

通用领域

翻译

ئامما تۈزگەن چاسا ئەترەت ئالدىغا جۇڭگۇ كومەنىسىنىك پارتىيەسى مەركىزىي كومىتېتى، مەملەكتىلىك خەلق قۇرۇلتىبىي دائىمىي كومىتېتى، گۇۋۇزىوەن، مەملەكتىلىك سىياسىي كېڭىش، مەركىزىي ھەربىي كومىتېت، ھەرقايىسى دەمۆكراشىك پارتىيە-گۇરۇھلار، مەملەكتىلىك سودا-ساناھىتچىلار بىرلەشمىسى ۋە پارتىيە-گۇرۇھىسىز ۋە تەنپەرەرەر زاتلار، ھەرقايىسى خەلق تەشكىلاتلىرى ۋە ھەر ساھە ئاممىسى، پىشىقىدەم جەڭچىلەر، پىشىقىدەم يولداشلار ۋە ئىنقلاپى قوربانلارنىڭ ئائىلە تاۋابىتالرى، جۇڭگۇ پىيونپىلار ئەترىتىنىڭ نامىدا تەقدم قىلىنغان چوك تېتىكى توققۇز گۈل سېۋىتى قاتار تىزىلغانىدى.

uy.ts.cn/system/2018/10/01/035399943.shtml

群众组成方队，面对的是中共中央，全国人民代表大会常务委员会、国务院；，全国政协；中央军事委员会；各民主党派、全国工商联及无党派爱国人士，各人民团体和各界群众、老战士、老同志和革命先烈的家属。以中国少先队命名的九个大型花束排列。

(ھۇممەت قازاۋۇلىرى تېپىللىكىلار)، دېكەن بۇرۇق بېرىلىشى بىلەن، ئۆج تاربىيەنىك ھۇممەت قازاۋۇلىرى مەرداň، مەزۇت، ھۇممەت قەدم بىلەن خاتمە ئۇنارنىڭ ئالدىغا كېلىپ، مىشقاىرىنى تۇنۇپ تىك تۇردى.

دەل سائەت 10:10، خەلق قەھرىمانلىرىنىڭ كۆل سېۋىتىنى تەقدم قىلىش مۇزاسىمىي رەسمىي باشلاندى. ھەربىي ئۇركىپستىر «پىشىلار مارشى»نى ئۇرۇنىدى، بۇئۇن مەيداندىكىلەر جۇڭخوا خەلق جۇمھۇرىيەتنىنىڭ دۆلەت شىئىرىنى ئۇنلاۋ ئۇرۇقىدى.

دۆلەت شىئىرى ئۇقۇلۇپ بولغاندىن كېپىن، بۇئۇن مەيداندىكىلەر سۈكۈتتە تۇرۇپ جۇڭگۇ خەلقنىڭ ئازادىق ئىشلىرى ۋە جۇمھۇرىيەتنىش قۇرۇلۇش ئىشلىرى ئۈچۈن قەھرىمانلارچە ئۇزۇنى بېغىشلىغان ئىشلىلىرى قوربانلارغا تەزىبىي بىلدۈردى.

تەزىبىي بىلدۈرۈش ئاباغلاشقاندىن كېپىن، گۈل تۇنغان ئۆسۈلەر، باللار خەلق قەھرىمانلىرى خاتمە ئۇنارنىغا بۇزىلىنىپ تۇرۇپ، «بىز كومۇنىزم ئىزىتساڭلىرى»نى ئۇرۇقىدى ھەم پىيونپىلار ئەترىتىنىڭ ئەترەت سالىمنى بەردى.

ئامما تۈزگەن چاسا ئەترەت ئالدىغا جۇڭگۇ كومەنىسىنىك پارتىيەسى مەركىزىي كومىتېتى، گۇۋۇزىوەن، مەملەكتىلىك سىياسىي كېڭىش، مەركىزىي ھەربىي كومىتېت، ھەرقايىسى دەمۆكراشىك پارتىيە-گۇرۇھلار، مەملەكتىلىك سودا-ساناھىتچىلار بىرلەشمىسى ۋە پارتىيە-گۇرۇھىسىز ۋە تەنپەرەرەر زاتلار، ھەرقايىسى خەلق تەشكىلاتلىرى ۋە ھەر ساھە ئاممىسى، پىشىقىدەم جەڭچىلەر، پىشىقىدەم يولداشلار ۋە ئىنقلاپى قوربانلارنىڭ ئائىلە تاۋابىتالرى، جۇڭگۇ پىيونپىلار ئەترىتىنىڭ نامىدا تەقدم قىلىنغان چوك تېتىكى توققۇز گۈل سېۋىتى قاتار تىزىلغانىسىدە، گۈل سېۋىتلىك قىلىپ لېتىنسىسا «خەلق قەھرىمانلىرى مەتكۈھا ھايات» دېكەن ئالىشۇن رەكىكىت چوڭ خەلتەر بېرىشىلەندى.

ھەربىي ئۇركىپستىر جۇڭقۇر مۇھەببەتكە تولغان گۈل تەقدم قىلىش مۇزىكىسىنى تۇرۇلىغاندا، 18 ھۇممەت قازاۋۇلى گۈل سېۋەتلىرىنى كۆتۈرۈپ، ئاستا قەددىمەر بىلەن خەلق قەھرىمانلىرى خاتمە ئۇنارنىغا قاربىكىم، گۈل بىۋەتلىرىنى ئاستە ئۇنارنىڭ ئۆل كەچىسىكە قۇبىدى.

شى جىنپىلاڭ قاتارلىق بارتىيە ۋە دۆلەت رەھىبلىرى ئارقىدىن خاتمە ئۇنارنى ئۆل كەچىسىكە چىقىپ، گۈل سېۋىتى ئالدىدا توخناب، خېلى ئۇزاق سۈكۈتتە ئۇردى، ئۇنقاشەك خۇنجاڭا، بۇرۇكلىپ ئېچىلغان گۈللىمىسى، چىراپلىق ئاساسى مەۋاپىتىگىلەك خەلق قەھرىمانلىرىنى چوقۇر ئەسلىش ۋە ئالىي ئېتىتمەن

最多可以输入500个字符

中文 ▾



维文 ▾

10月6日晚，“我爱你中国”主题灯光秀把夜色中的乌鲁木齐装点得分外绚丽。各族市民和众多来疆游客观灯赏景，欢声笑语，为中华人民共和国69周年华诞送上衷心的祝福。

ئايىنك 6 - كۈنى كەچتە، « سىزنى سۈپىمەن » دېگەن تېمىدىكى چىراق نۇرى كەچىدىكى ئۇرۇمچى شەھىرىدە - 10 نومۇر ئىلىس، گۈزەل تۈسکە كىرىدى. ھەر مىللەت شەھەر ناھاللىرى ۋە شىنجاڭغا كەلگەن نۇرغۇن ساياھەتچىلەرنىڭ بايۆس سەپىسى، كۈلكە سادالىرى جىڭىخۇا خەلق جۇمھۇرىتى قۇرۇلغانلىقىنىڭ 69 يىللەقنى قىرغىن تەبرىكلىدى



The screenshot shows a news article from Tsinghua University's website. A purple arrow points from the Chinese text above to the headline in the screenshot.

清华大学多... X 清华大学跨... X 多语种翻译... X 多语种翻译... X 天山网_百度... X 天山网 - 新... X 天山网 - 新... X 天山时评... X

/system/2018/10/07/035403688.shtml

设为首页 加入收藏

讲文明 树新风 公益广告展播 龙卡分期 龙卡分期 梦想成真 民族团结一家亲

天山网 www.ts.cn 地州 ▾ 频道 ▾ 导航 ▾

您当前的位置 : 天山网 >> 新闻中心 >> 时政新闻

【天山时评】华灯似锦激荡家国情怀

2018年10月07日 10:56 来源：新疆日报

樊虹壹

10月6日晚，“我爱你中国”主题灯光秀把夜色中的乌鲁木齐装点得分外绚丽。各族市民和众多来疆游客观灯赏景，欢声笑语，为中华人民共和国69周年华诞送上衷心的祝福。璀璨灯光照亮夜空，展示新疆发

TS.CN 在这里读懂新疆

Outline

- ✓ Machine Translation
- ✓ Related Work and Current State for LRLs NMT
- ✓ Motivation
- ✓ Our works
- ✓ Projects
- Copyrights & Patents
- Conclusions



Copyrights

- 已发布

- 米尔阿迪力江·麦麦提, “维吾尔语人工词性标注及语料库构建系统” [简称: 维吾尔语人工词性标注平台], 登记号: 2016SR031180
- 米尔阿迪力江·麦麦提, “维吾尔语自动词性标注系统” [简称: 维语词性标注平台], 登记号: 2016SR052763
- 米尔阿迪力江·麦麦提, “维吾尔语自动词干提取与词性标注系统” [简称: 维吾尔语形态分析器], 登记号: 2016SR379408
- 孙茂松, 米尔阿迪力江·麦麦提, 刘洋, 栾焕博, “基于Python的多语种多文种编码转换工具软件(1.0)” [简称: 维蒙藏-汉(小语种)在线翻译系统], 登记号: 2019SR0110291
- 孙茂松, 米尔阿迪力江·麦麦提, 刘洋, 栾焕博, “面向低资源语言的多语种在线机器翻译系统(1.0)” [简称: 小语种在线翻译系统], 登记号: 2019SR0108620
- 孙茂松, 米尔阿迪力江·麦麦提, 刘洋, 栾焕博, “跨语言社会舆情分析在线信息检索系统” [简称: 小语种跨语言信息检索系统], 在申请



- 已授权
 - 米尔阿迪力江·麦麦提, 刘洋, 栾焕博, 孙茂松, “一种基于无监督领域自适应的神经网络机器翻译方法”, 申请日: 2017年03月09日, 申请号: 201710139214.0, 授权公布日: 2017年08月11日, 授权号: CN 107038159 A
- 在审
 - 孙茂松, 米尔阿迪力江·麦麦提, 刘洋, 栾焕博, “神经网络机器翻译模型的训练方法和装置”, 申请日: 2018年07月27日, 申请号: 201810845896.1



Outline

- ✓ Machine Translation
- ✓ Related Work and Current State for LRLs NMT
- ✓ Motivation
- ✓ Our works
- ✓ Projects
- ✓ Copyrights & Patents
- Conclusions



Conclusions

- Semi-supervised Learning
- Unsupervised Learning
- Pivot-based Methods
- Data Augmentation
- Data Selection
- Transfer Learning
- Meta Learning
- Morphological analyses
- Pos-tagging
- NER



شكرا لك

شكريا

ئىخالىخانى

הודות

謝 謝！

رەخەمەت!

Kiitos

köszönöm

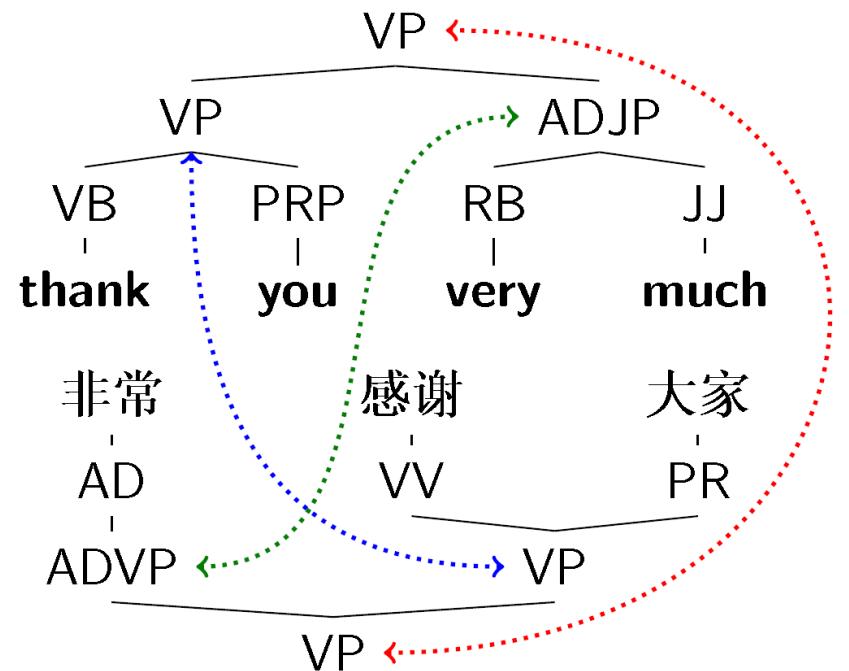
සාර්කීරුණු

Teşekkür



Any Questions ?

Questions diverses ?



This inspiration comes from Dzmitry Bahdanau @ ICLR2014 .