



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



Outline

- International Exchange Activities
- Demands for Machine/Human Translation
- Background
- Motivation
- Methodology
- Experiments
- Conclusions



Outline

- International Exchange Activities
- Demands for Machine/Human Translation
- Background
- Motivation
- Methodology
- Experiments
- Conclusions



International Exchange Activities (IEAs)



How?

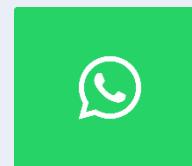


Weak ideas

- How could we achieve higher quality in international exchange activities?
- How about hire a multilingual human translator?
 - Money
 - Time
 - Efficiency
- How about AI?
 - Machine Translation System
 - Quality ?
 -



Some efficient ways for IEAs

Different channels	Examples		
Social Media Networking	  		
BLOGS	  		
Activities Sharing and Storage	   		
Telecommunication Options	   		

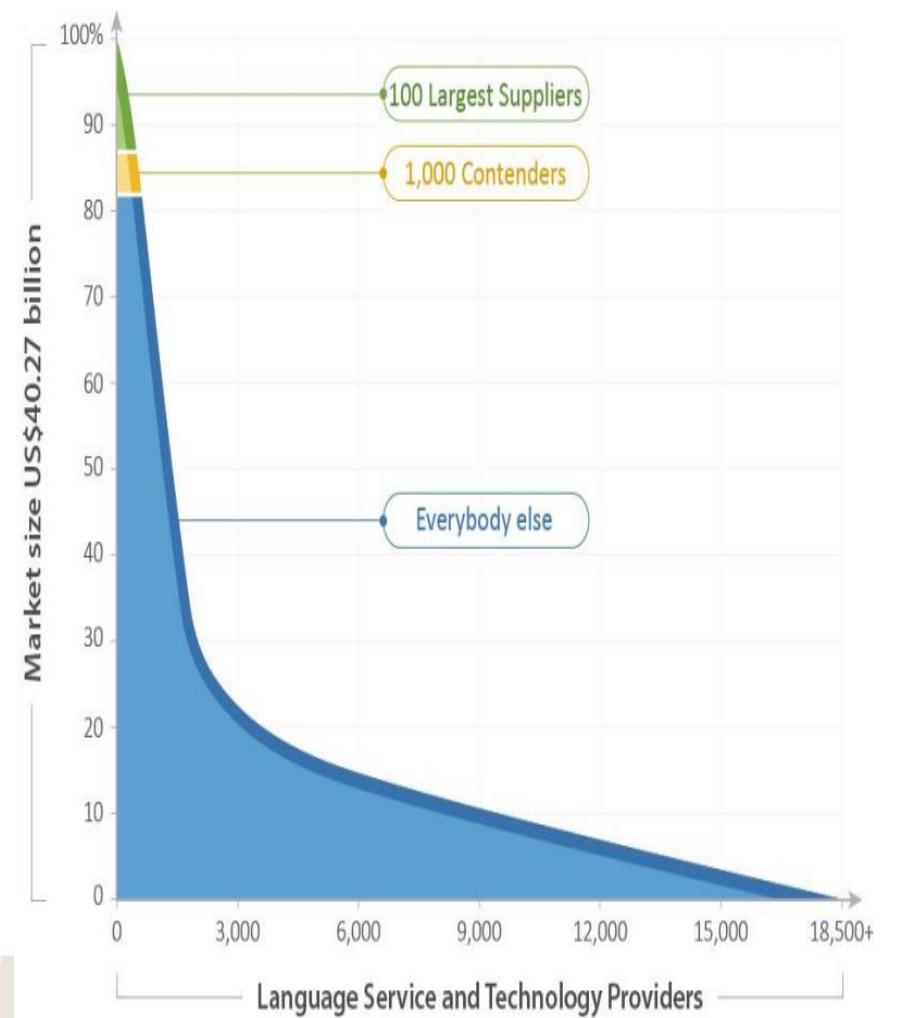
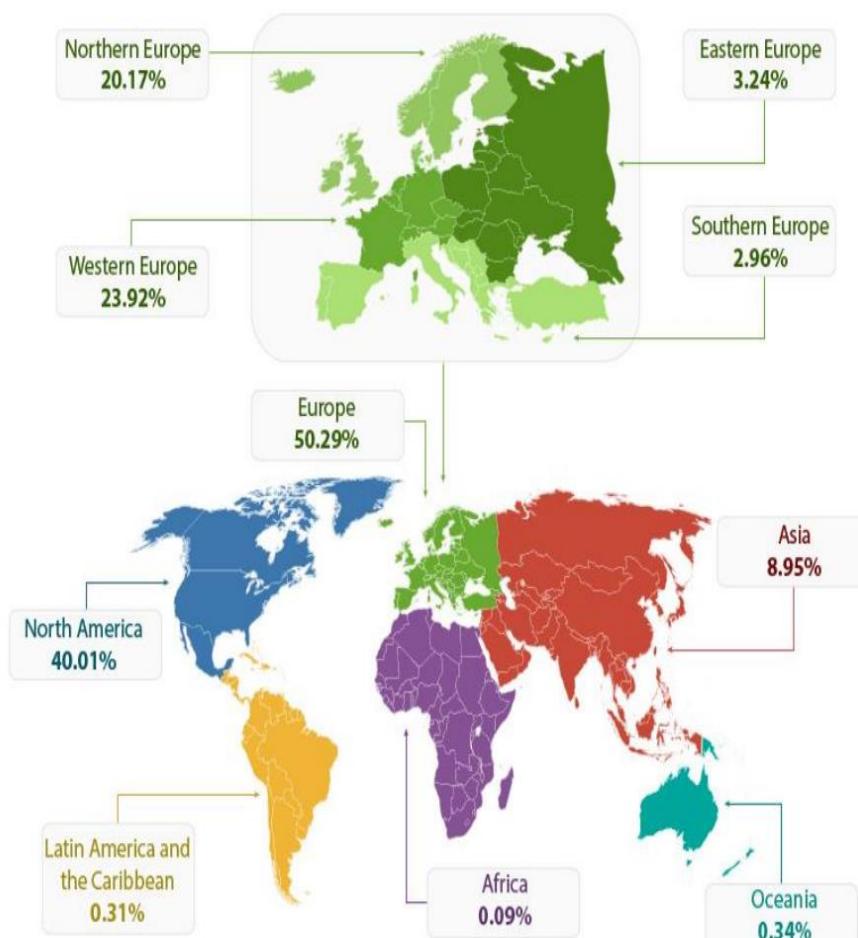


Outline

- ✓ International Exchange Activities
- Demands for Machine/Human Translation
- Background
- Motivation
- Methodology
- Experiments
- Conclusions



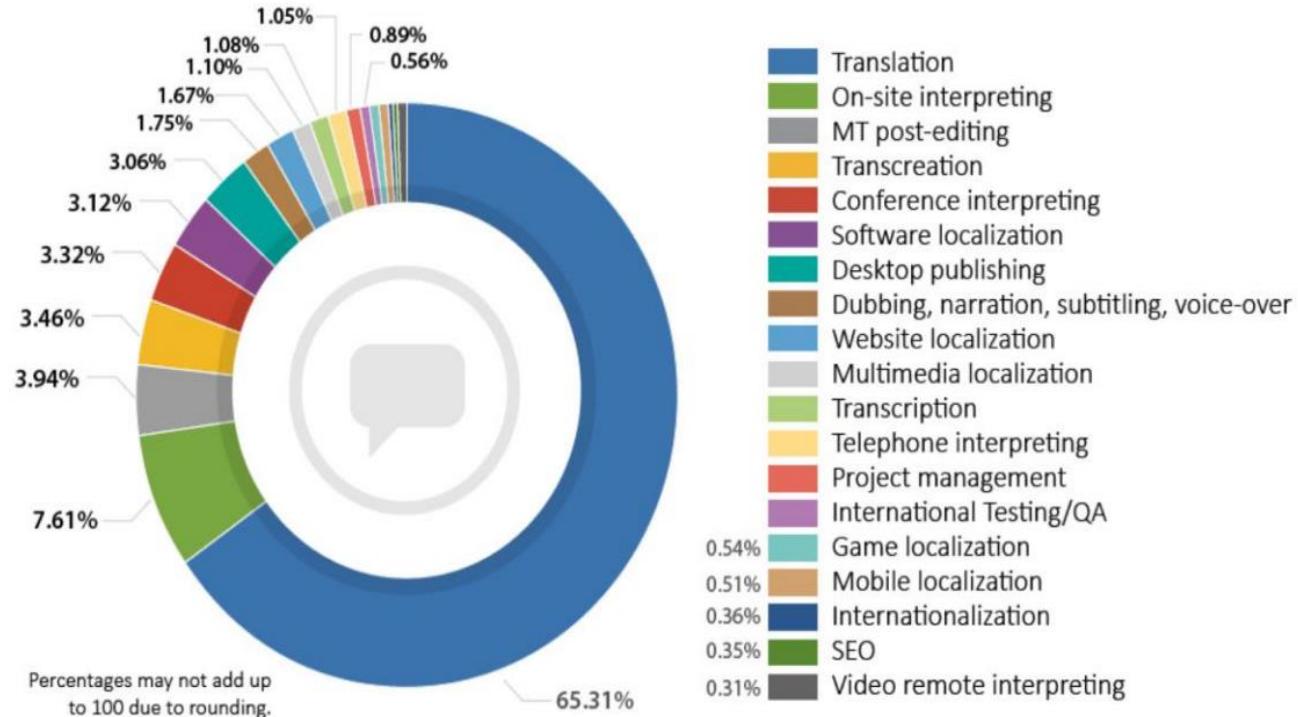
Demands for Machine/Human Translation



(Guoping Huang, Qcon2018)



Demands for Machine/Human Translation



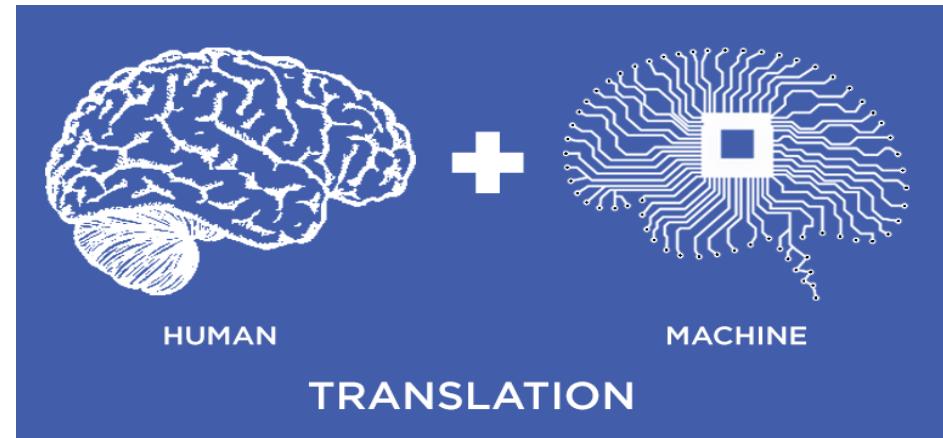
- Human translator is the main
- Tech less than original
- MT has low percentage, but growing faster

(Guoping Huang, Qcon2018)



Demands for Machine/Human Translation

- Human translation fields
≈ 666 million words / day
[Pym et al., 2012]
- Machine translation files
>> 100 billion words / day
[Turovsky, 2016]



Demand for translation far outpaces what is humanly possible to produce.



Outline

- ✓ International Exchange Activities
- ✓ Demands for Machine/Human Translation

- Background

- Motivation

- Methodology

- Experiments

- Conclusions



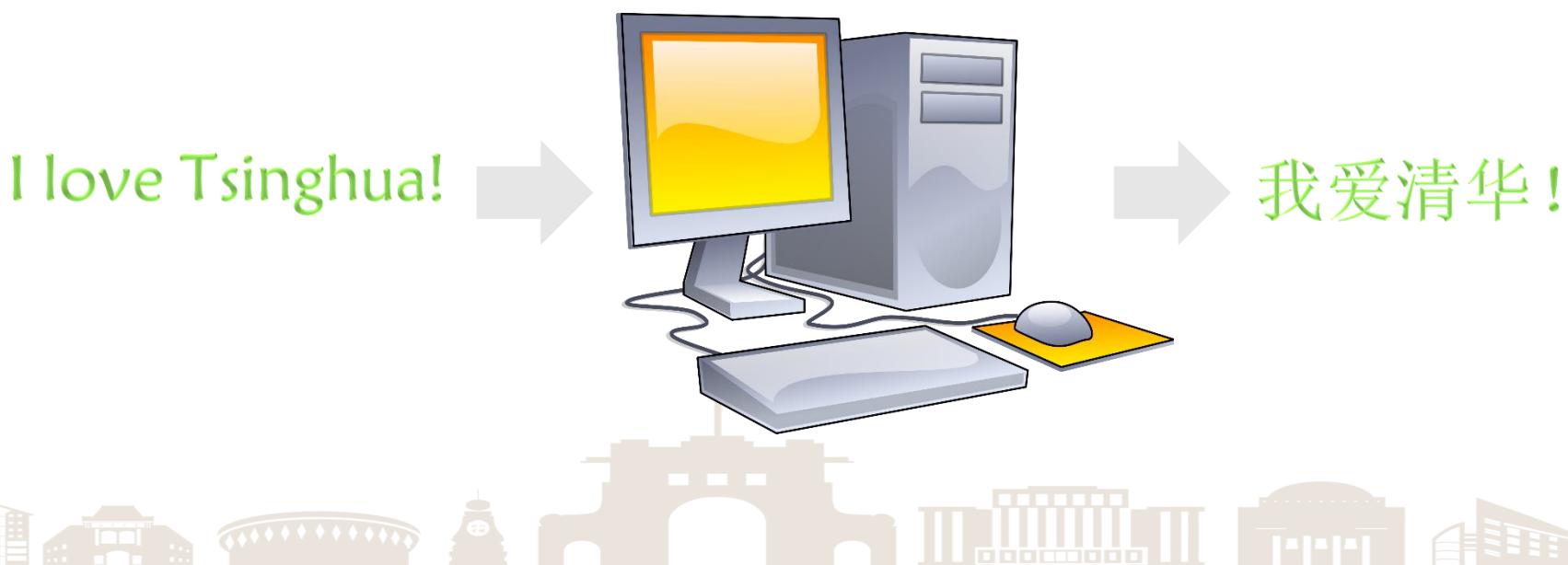
Background

- Machine Translation (MT)
 - Statistical Phrase-based Machine Translation (SMT)
 - Neural Machine Translation (NMT)



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



Demo (SMT)

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... 1.5.245:9662/index.html

清华大学多语种翻译系统

چىخخۇا ئۇنىۋېرسىتى كۆپ تىللەق تەرجىمە سىستېمىسى

维文 > 汉语 > 通用领域 翻译

ئىنقلابىي قۇربانلارنى خاتىرىلەش كۈندە خالق قەھرىمانلىرىغا گۈل سېۋىتى تەقدىم قۇرمۇسىنىڭ دەغدۇغلىق 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

中文 Türkçe Русский язык English Уйгурچا uygurche قازاقشا ئۇيغۇرچە

خانىم - قىزلاز ساغلاملىق سايادەت قانۇن مۇھىم سىز ئۆكىنىش مەملىكتە شەنھائە مۇھىم خەۋەر شەنھائە ئەقىنسىدە خەلقئارا

ئۇرمۇش ئوغىنى ئامىلە تەلتارىيە ئاتالمازار ئۇرمۇز ئەزىزىيە مەھىسىن سەھىپە ئەقىنسىدە خەلقئارا

پارتىيە 19-قۇرۇس اى روحىنى چوڭقۇر ئۆگىنەيلى، ئىزجىلاشتۇرالىم!

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

مؤەممەر ئەركىزى < مۇھىم خەۋەر < تەڭرىتىغۇ ئورى

最多可以输入500个字符

2018/10/01 15:37

مؤەممەر ئەركىزى | مەنبىي: شىنجاڭ كېزىتى
شى جىنىڭ، لى كېچىڭلە، لى جىنىڭ، والى يالا، والى خۇنىڭ، جاۋ لېچ، خەن جىڭ، والى چىشمۇن قاتاناشى

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

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

Mieradilijiang.M and Xiaohui.Z ICIS2018 2021/6/20 16

Demo (NMT)

The screenshot shows a web-based multilingual translation system. At the top, there are several browser tabs in Chinese and Uyghur. The main content area has a blue header with the text "多语种翻译系统" (Multilingual Translation System) and "كۆپ تىللېق تارجىمە سىستېمىسى" (Multilingual Translation System). Below the header, there are four language selection buttons: "维吾尔语" (Uyghur), "汉语" (Chinese), "通用领域" (General Field), and "翻译" (Translation). The "翻译" button is highlighted in blue.

The main content area displays two parallel texts:

Uyghur Text:

ئاما تۈزگەن چاسا ئەترتەت ئالدىغا جۇڭگۇ كومىمۇنىستىك پارتىيەس مەركىزىي كومىتېتى، مەملىكتىك خلق قۇرۇلتىن دائىمىي كومىتېتى، گۇۋۇپۇون، مەملىكتىك سىياسى كېڭىش، مەركىزىي ھەربىي كومىتېت، ھەرقاپسى دەمۆكراتىك پارتىيە-گۇزۇھار، مەملىكتىك سودا-ساناڭتىچىلەر بىرلەشىسى ۋە پارتىيە-گۇزۇھىز ۋەنەنەرۋەر زانلار، ھەرقاپسى خلق تاشكىلاتلىرى ۋە ھەر ساھە قامىسى، پېشقەددەم جەڭچىلەر، پېشقەددەم يۈلەشلار ۋە ئىنقلابى قىق، بازىلە ئاڭ ئايلى ئازادىتائىلىرى، جۇڭگۇ بىبىنپىرلار ئەترىتىنىڭ نامىدا ئەقدىم قىلىغان چوك قاتار تىزىلغاندى.

Chinese Text:

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。

A large purple arrow points from the Chinese text down to the Uyghur text, indicating the direction of translation. A red box highlights the Uyghur text at the bottom right, and another red box highlights the Chinese text at the bottom left.

最多可以输入500个字符



MT Parallel Corpus

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 .

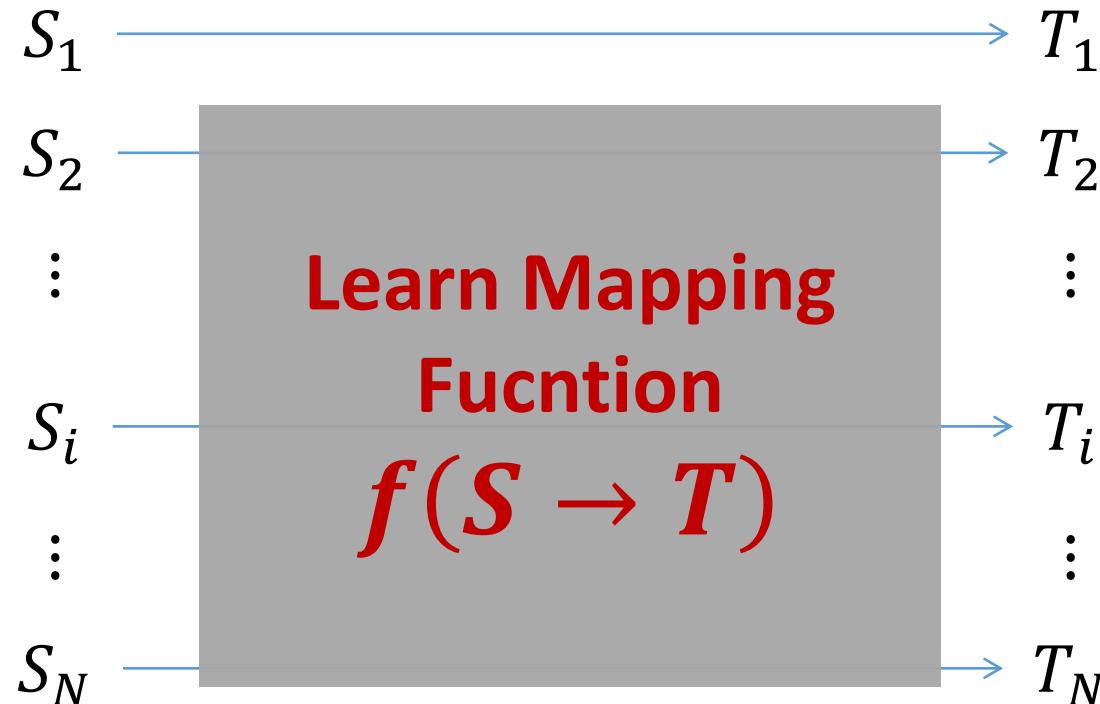
... ...

... ...

(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 北京 大学 做了 报告



Mapping
function
 $f(S \rightarrow T)$

English:

I gave a talk in Peking University



(Jiajun Zhang, CCL2018)

SMT

x

布什

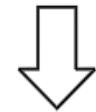
与

沙龙

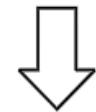
举行

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

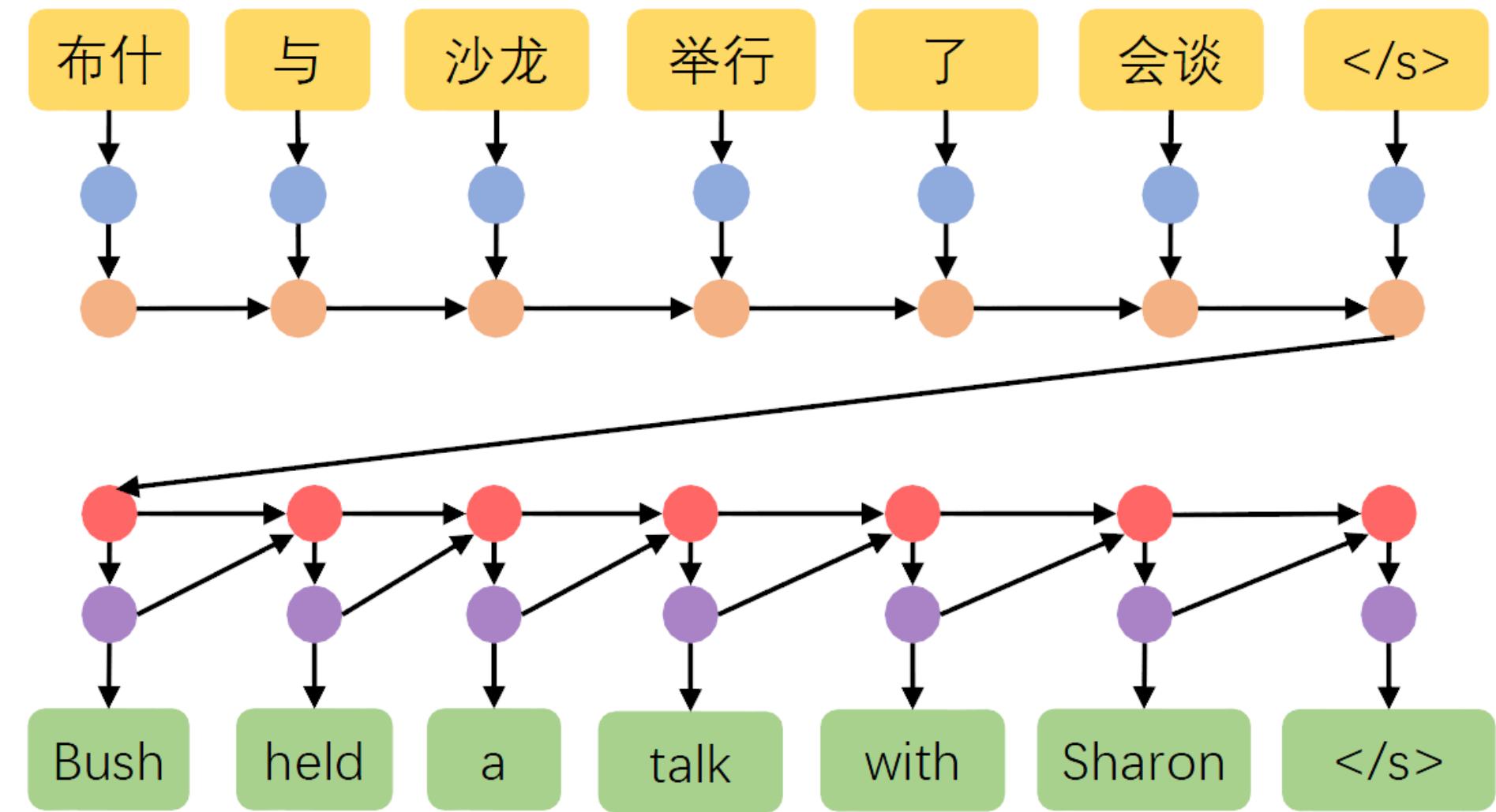
with

Sharon



(Och and Ney., 2002)

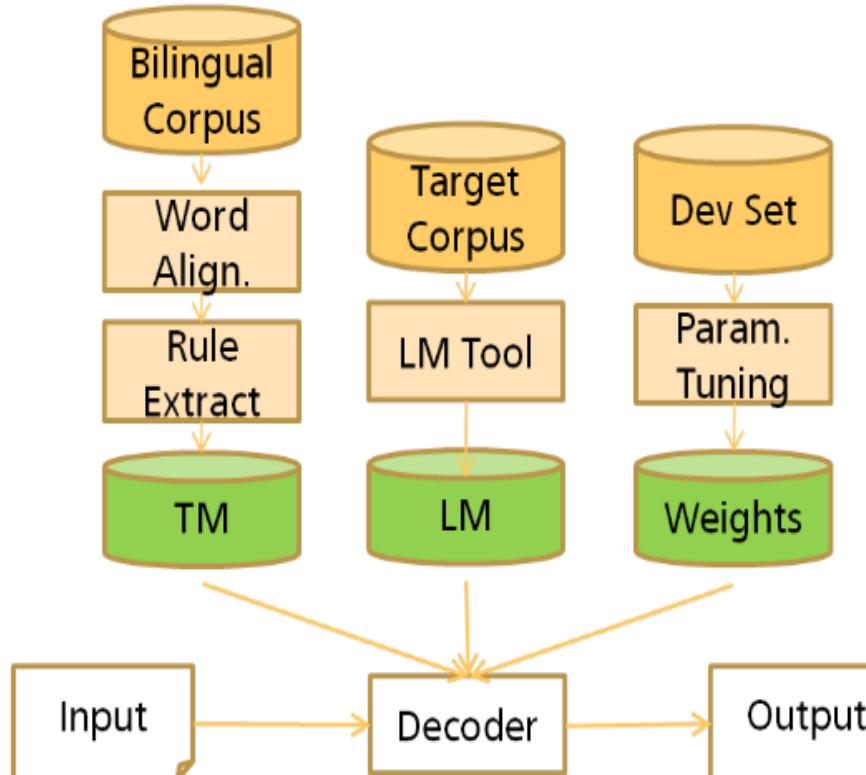
NMT



(Sutskever et al., 2014)

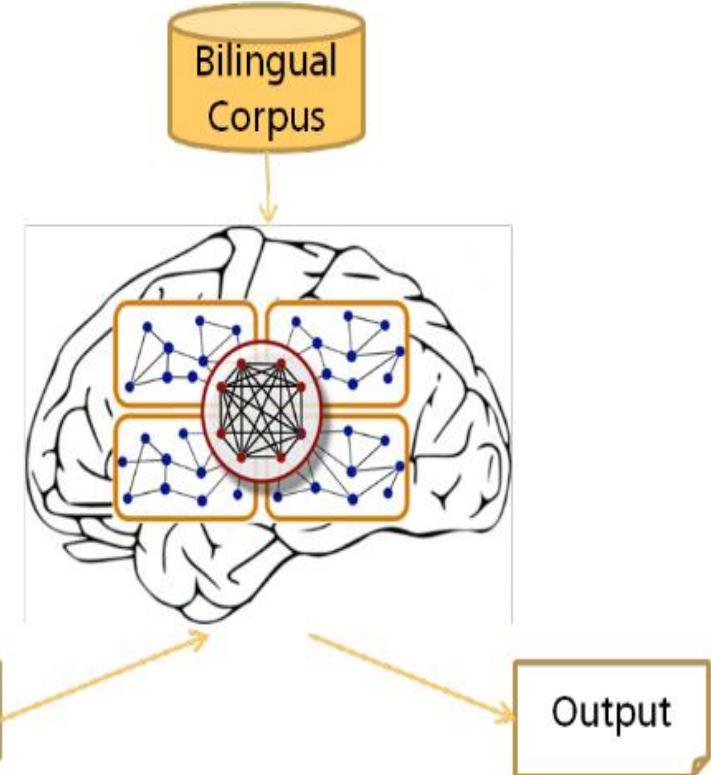
SMT && NMT

Many **sub-components** are tuned separately



SMT (1993 ~)

single , large neural network



NMT (2014~)

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:
- $P(y|x; \theta) = \prod_{j=1}^J P(y_j|x, y_{<j}; \theta)$
- θ is model parameters, $y_{<j}$ is partial translation.

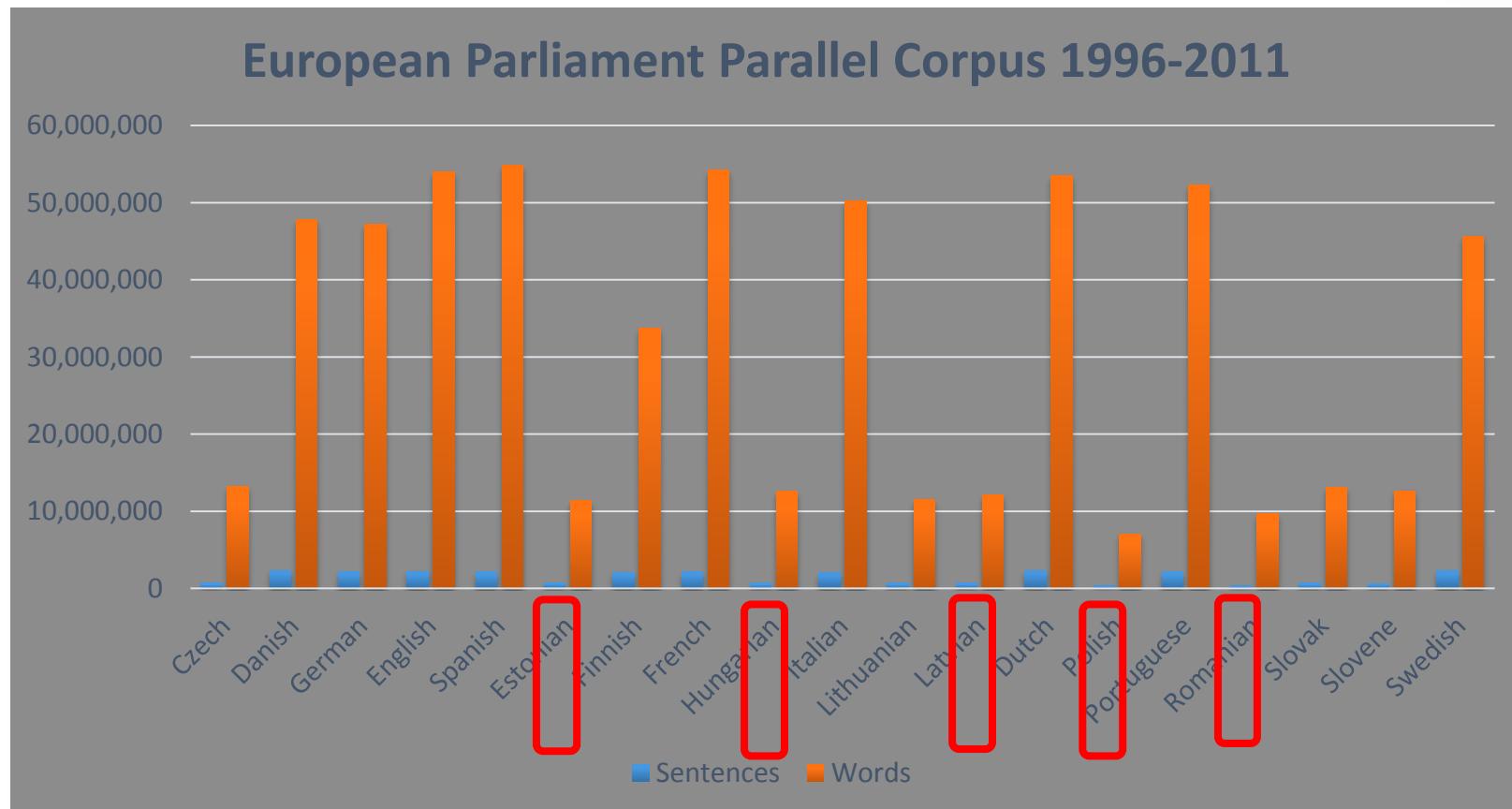


NMT

- NMT models usually rely on an encoder-decoder scenario.
- Let $\langle X, Y \rangle = \{x^{(n)}, y^{(n)}\}_{n=1}^N$ be a training corpus. The log-likelihood of the training parallel data is maximized by the standard training objective function:
$$\hat{\theta} = \operatorname{argmax}_{\theta} \left\{ \sum_{n=1}^N \log P(y^{(n)} | x^{(n)}; \theta) \right\}$$
- $\hat{y} = \operatorname{argmax}_y \{P(y|x; \hat{\theta})\}$ $\hat{y}_j = \operatorname{argmax}_y \{P(y|x, \hat{y}_{<j}; \hat{\theta})\}$



Low-Resource Languages (LRLs)



Outline

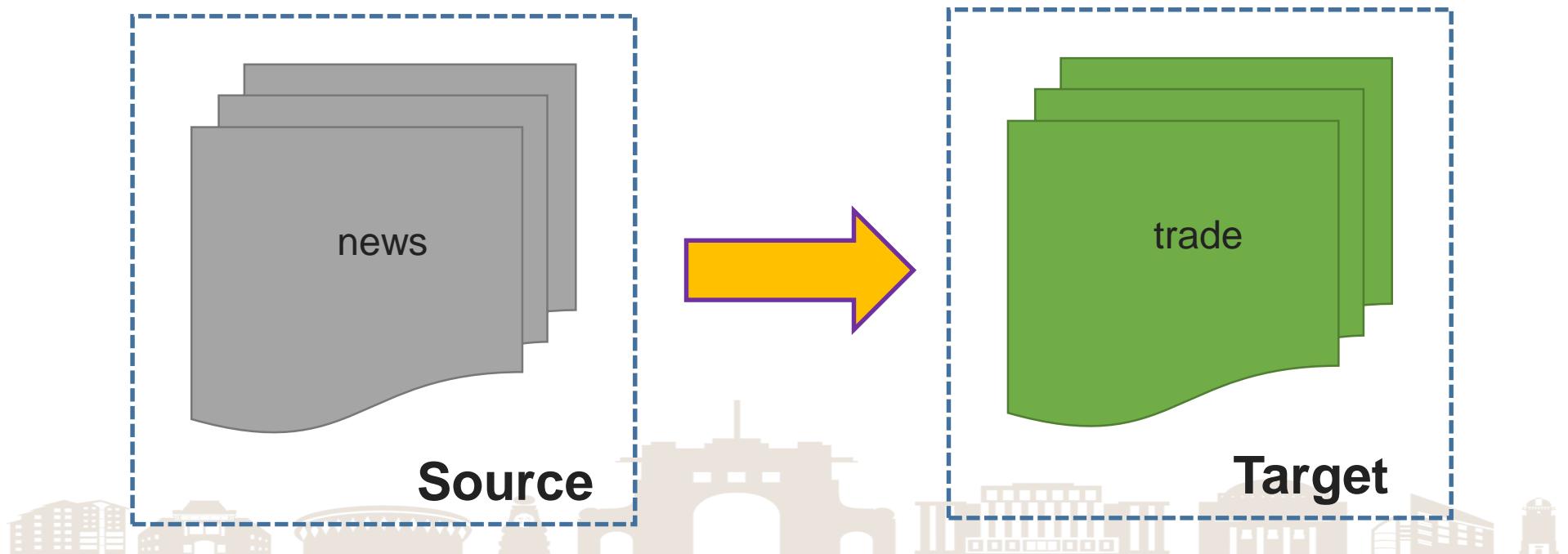
- ✓ International Exchange Activities
- ✓ Demands for Machine/Human Translation
- ✓ Background
- Motivation
- Methodology
- Experiments
- Conclusions



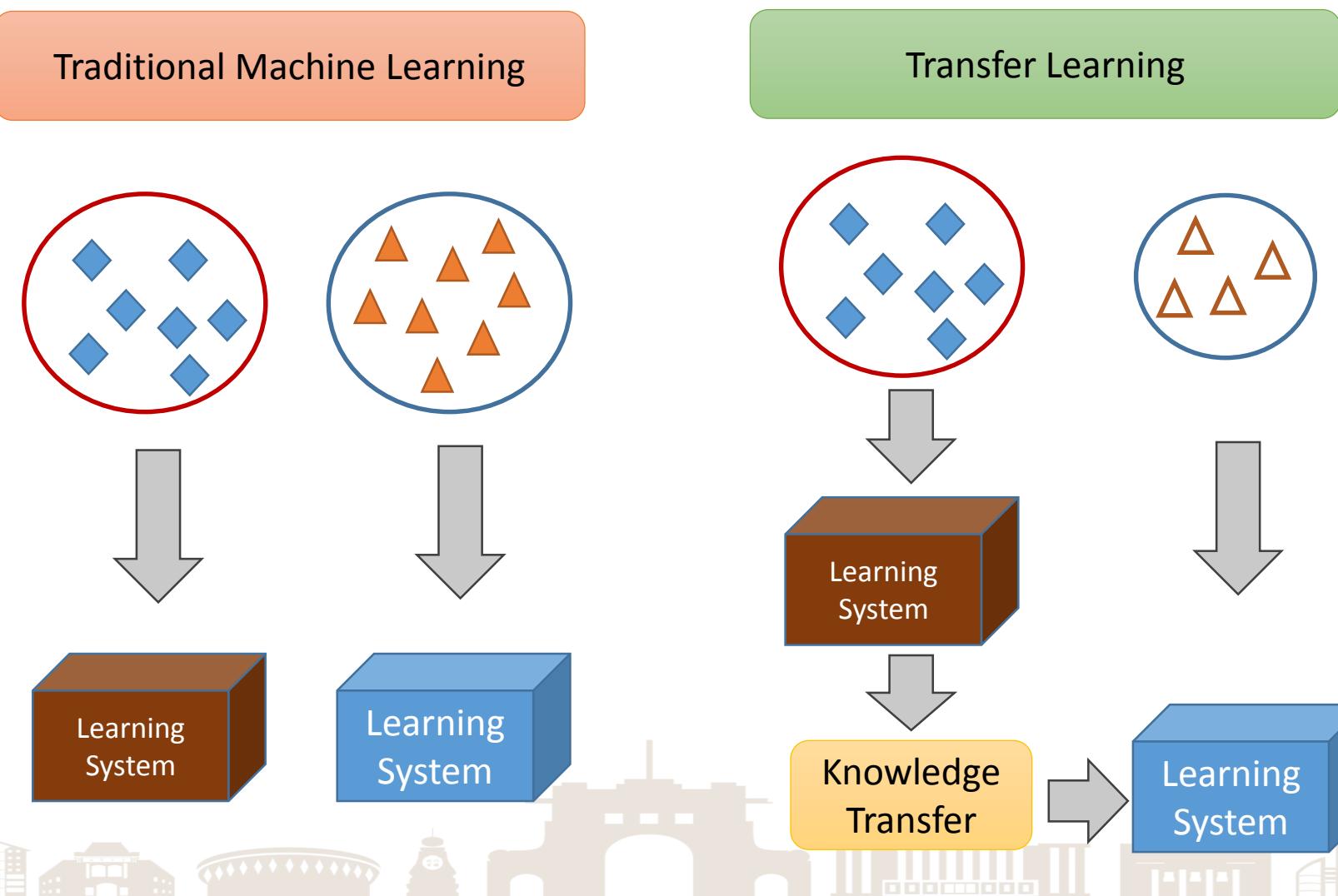
Motivation

- Transfer Learning (TL)

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.



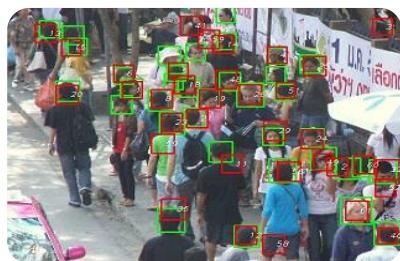
Transfer Learning



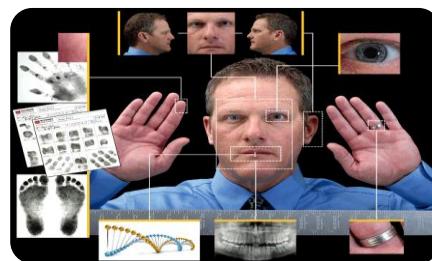
Transfer Learning

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

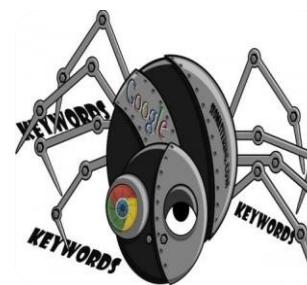
Build speech recognition system using recorded phone calls, then tune it to use as an airline reservation hotline.



CV



ER



IR



ASR

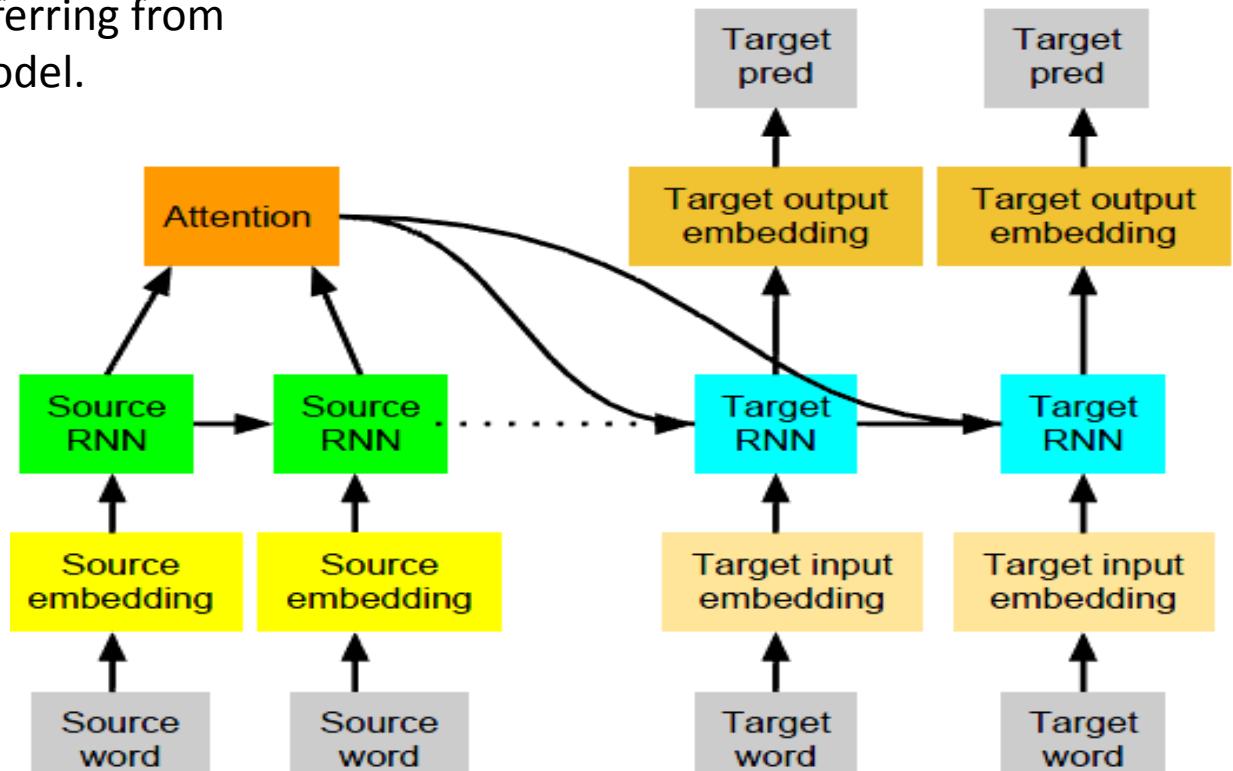
Transfer Learning

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

Parent : Fr-En , De-En
Child: Uz-eEn

Transfer learning;
High-resource language;

Drawback:
Ignore the word sharing;



(Barret Zoph et al., 2016)

Outline

- ✓ International Exchange Activities
- ✓ Demands for Machine/Human Translation
- ✓ Background
- ✓ Motivation
- Methodology
- Experiments
- Conclusions

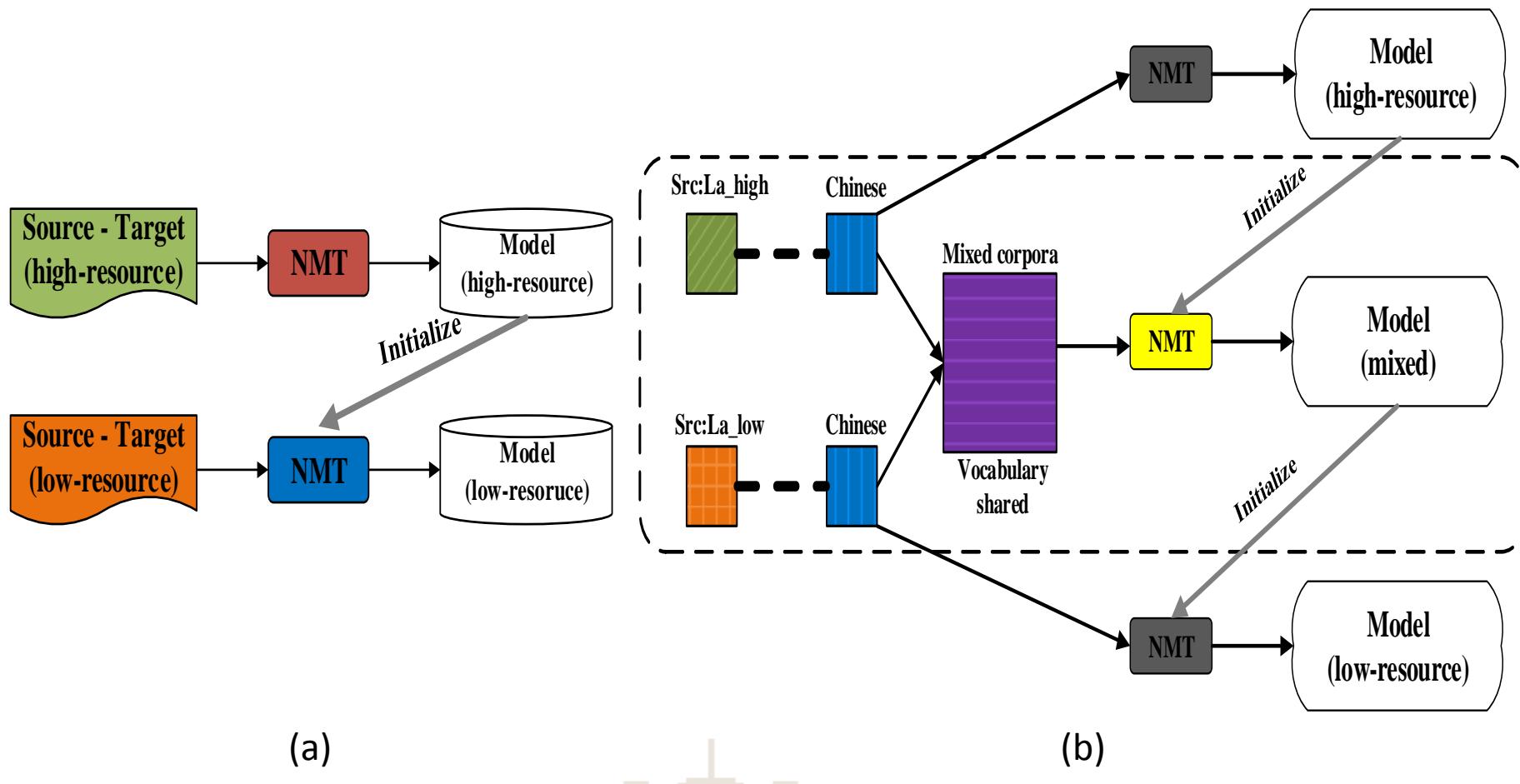


Methodology

- 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.



Methodology



Original Transfer Learning

Mixed Transfer Learning

Outline

- ✓ International Exchange Activities
- ✓ Demands for Machine/Human Translation
- ✓ Background
- ✓ Motivation
- ✓ Methodology
- Experiments
- Conclusions



Experiment

- System : [DL4MT](#)
- Parameters :
 - Dropout 0.1
 - Word Embedding 620
 - Hidden State 1000
 - Vocabulary 3w
- Other parameter we use the default parameters of DL4MT
- Preprocess

- Clear data use [NiuTrans](#) , Chinese segmenter use [THULAC](#)

Experiment

Experiment data is available [here](#) and [here](#).

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
Chinese	(Ch)	Sino-Tibetan		Chinese		Sinitic		SVO		Character	Light

Languages	Train	Dev	Test	Source			Target		
				Vocab.	#Word	#CoV.[%]	Vocab.	#Word	#CoV.[%]
Ar → Ch	5.1M	2.0K	2.0K	1.0M	32.2M	88.30	0.5M	37.4M	96.80
Fa → Ch	1.4M	2.0K	1.0K	0.2M	10.4M	96.80	0.2M	10.0M	96.80
Ur → Ch	78.0K	2.0K	1.0K	17.6K	2.6M	100.00	12.7K	2.4M	100.00



Experiment

	Arabic	Farsi	Urdu
Arabic	N/A	12.36%	8.61%
Farsi	2.51%	N/A	8.62%
Urdu	0.16%	0.78%	N/A

Method	Parent	Child	BIEU
RnnSearch	N/A	Ur → Ch	18.31
Mixed	Fa → Ch (Non-Shared)		19.73
	Ar→ Ch (Non-Shared)		20.04
	Fa → Ch (Shared)		21.69 ⁺⁺
	Ar→ Ch (Shared)		22.44 ⁺⁺



Outline

- ✓ International Exchange Activities
- ✓ Demands for Machine/Human Translation
- ✓ Background
- ✓ Motivation
- ✓ Methodology
- ✓ Experiments
- Conclusions



Conclusions

- Mitigate the gap between a non-native speaker and native speaker by exploiting prepared or necessary draft
- Make full use of the combination of bilingual cognition and artificial translation system in IEAs.
- Provide some Efficient and effective channels for IEAs.
- The proposed NMT training approach for LRLs in IEAs is transparent to neural network architecture.



شكرا لك

شكريا

ଧ୍ୟାନମ୍ବର

הודות

謝 謝！

ره خمهت!

Kiitos

köszönöm

ଘାର୍ଦ୍ଧିକ୍ଷେ

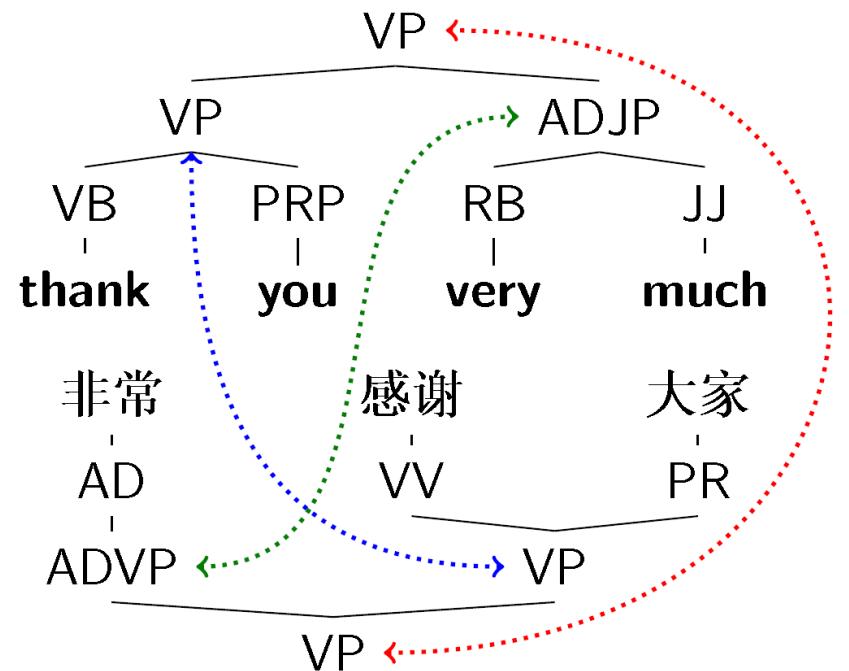
Teşekkür



Any Questions ?



Questions diverses ?



This inspiration comes from Dzmitry Bahdanau @ ICLR2014 .