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

Mieradilijiang Maimaiti[1], Yang Liu[1], Huanbo Luan[1] , Maosong Sun[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

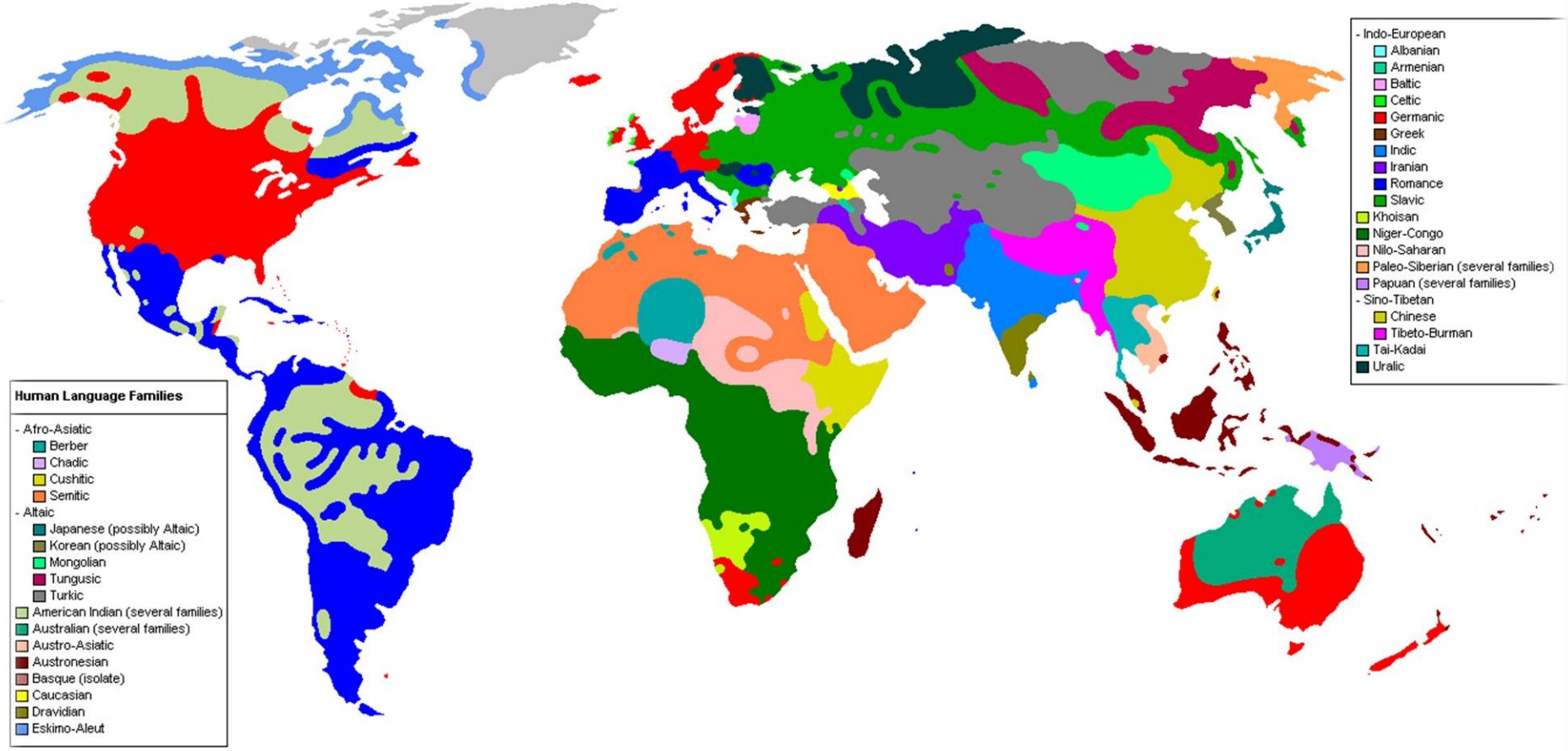
2019.03.16,Beijing

Accepted by ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2019.

Outline

- Machine Translation
- Demands for Machine/Human Translation
- Related Work and Current State for LRLs NMT
- Motivation
- Methodology
- Experiments
- 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).

I love Tsinghua!

我爱清华！

Typical MT Systems

Current State of MT Community

Bilingual Information

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 。

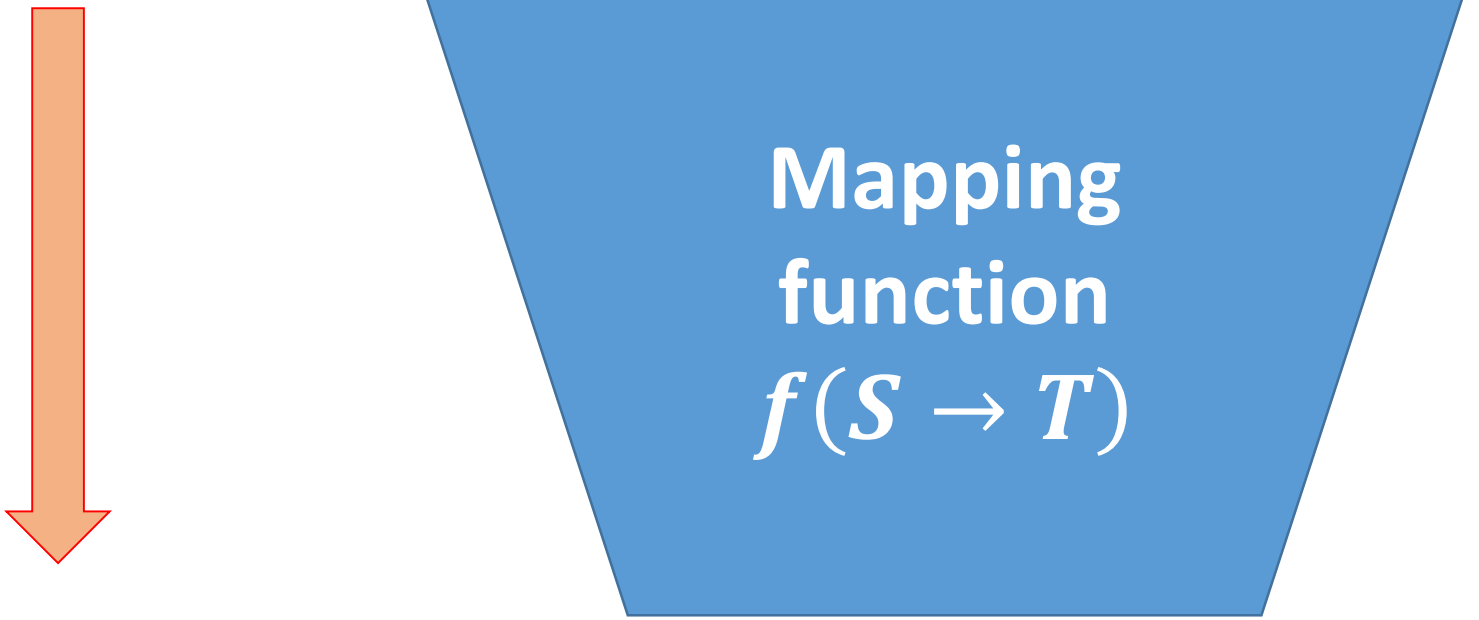
Beijing is located in the North of China .

… …

(Jiajun Zhang, CCL2018)

Mapping function from source to target language

Chinese: 我 在 清华 大学 做了 报告



English: I gave a talk in Tsinghua University

SMT

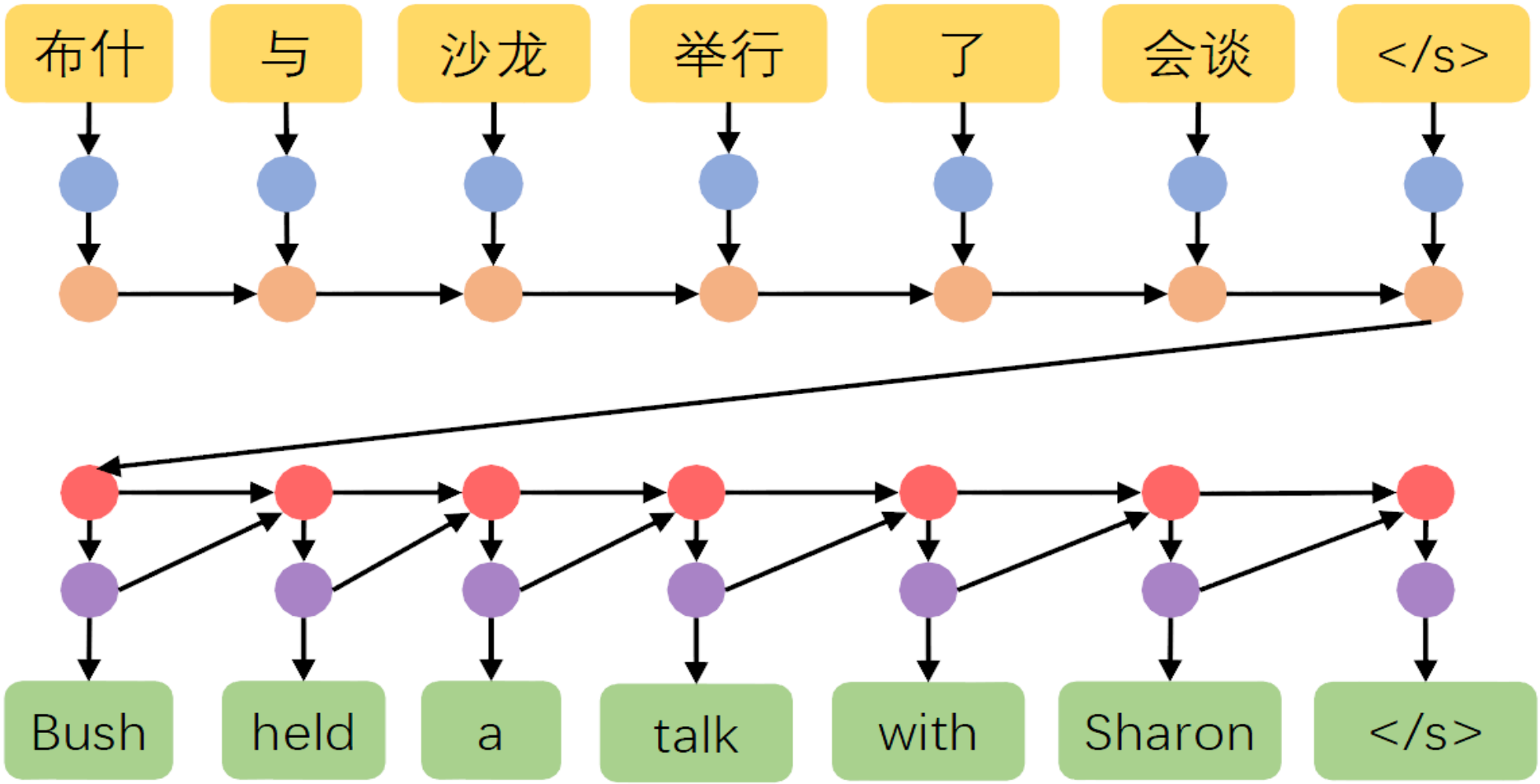
x 布什 与 沙龙 举行 了 会谈

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

y Bush held a talk with Sharon

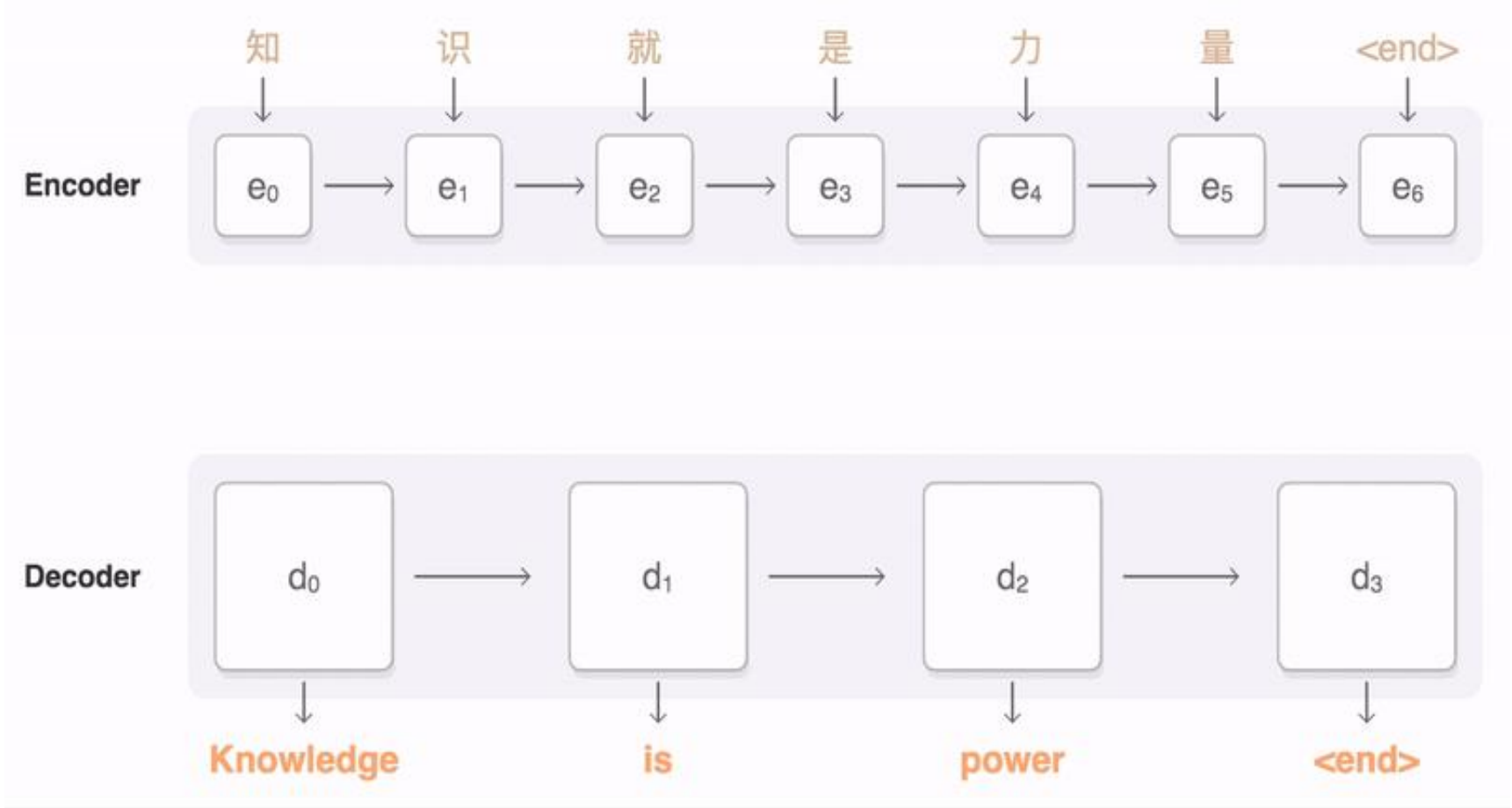
(Och and Ney., 2002)

NMT – Encoder Decoder



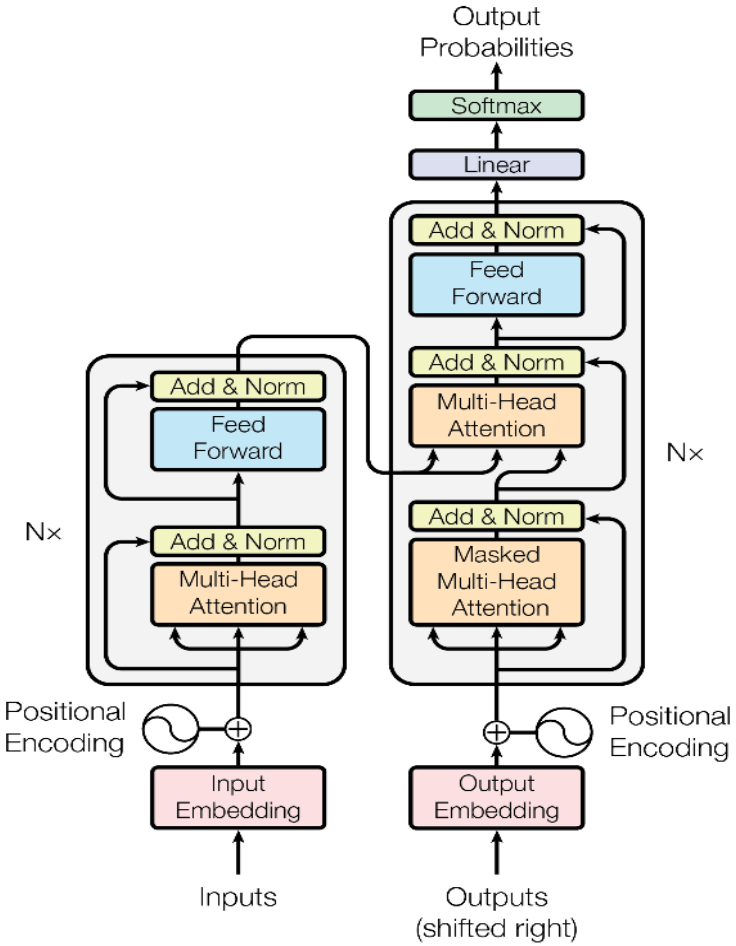
(Sutskever et al., 2014)

NMT – Attention



(Wu,et al. 2016)

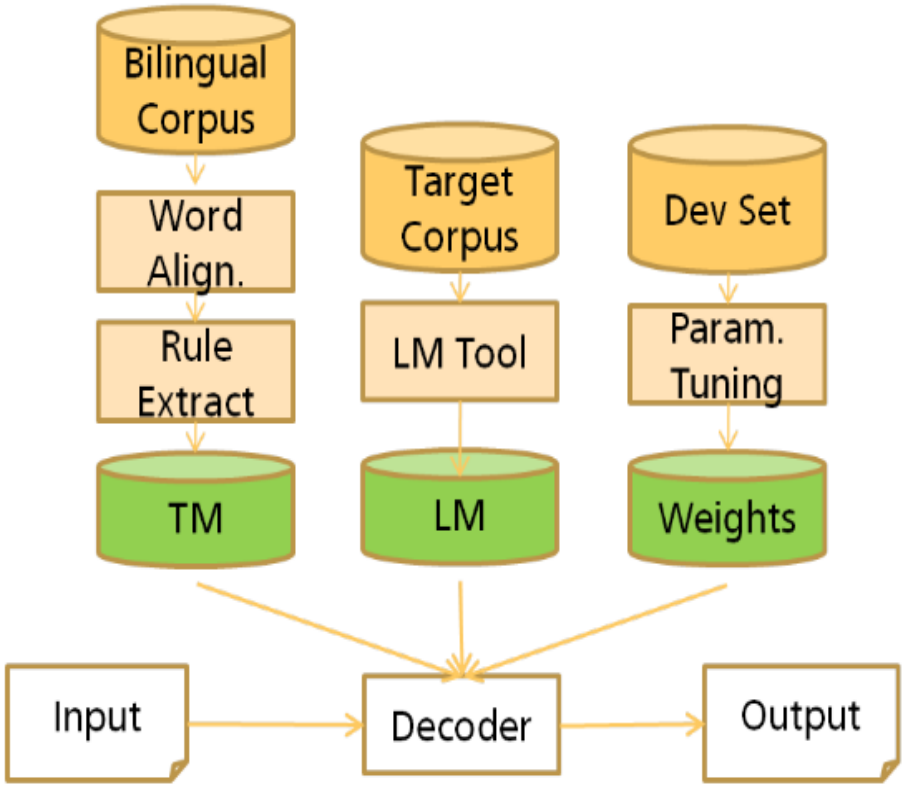
NMT – Transformer



(Vaswani,et al. 2017)

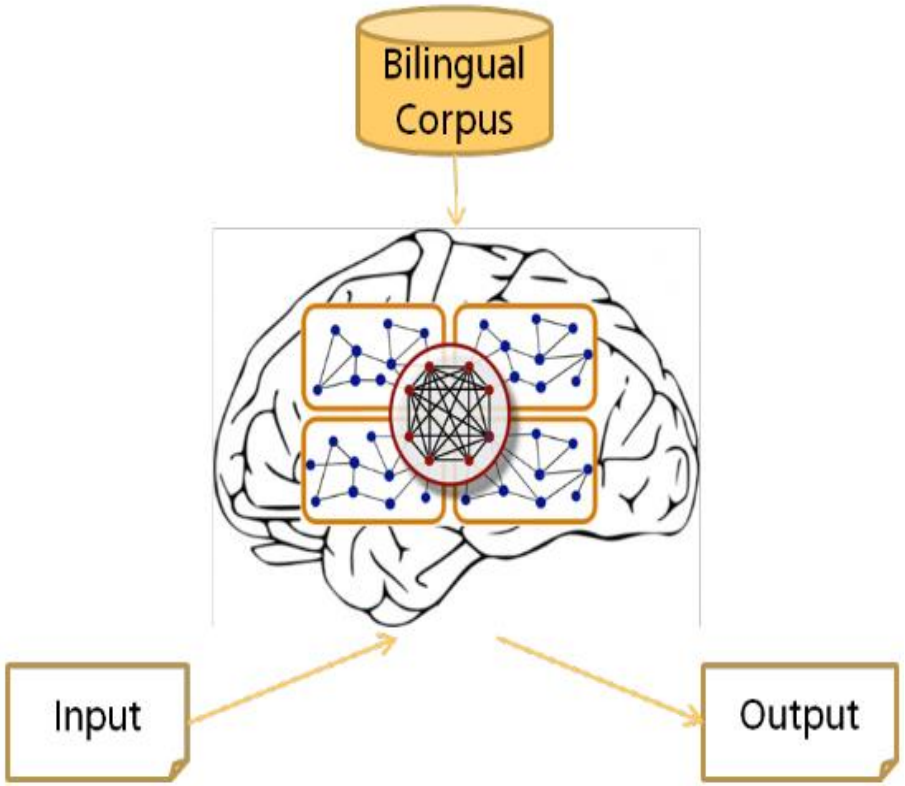
SMT && NMT

Many sub-components are tuned separately



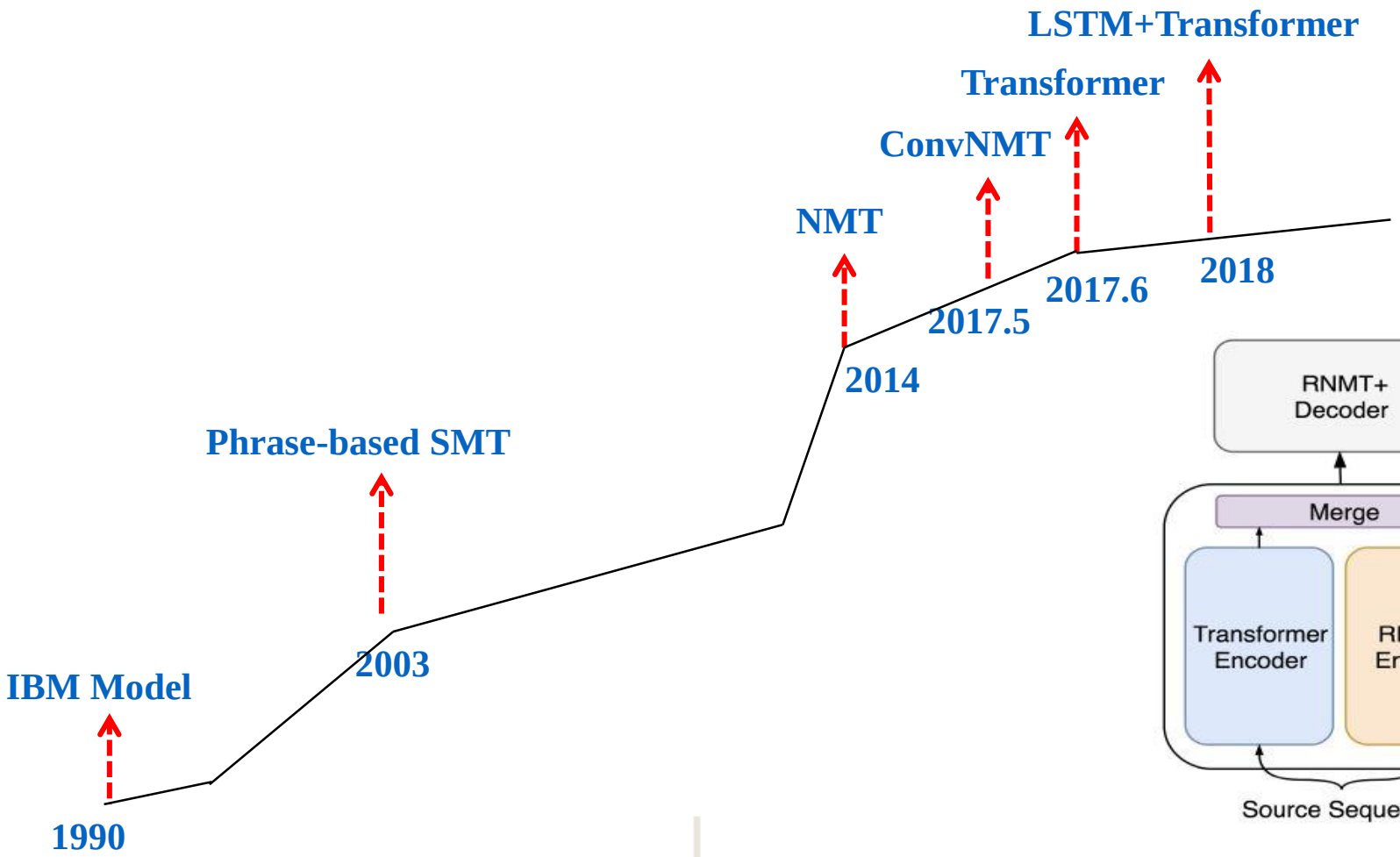
SMT (1993 ~)

single , large neural network



NMT (2014~)

History of MT



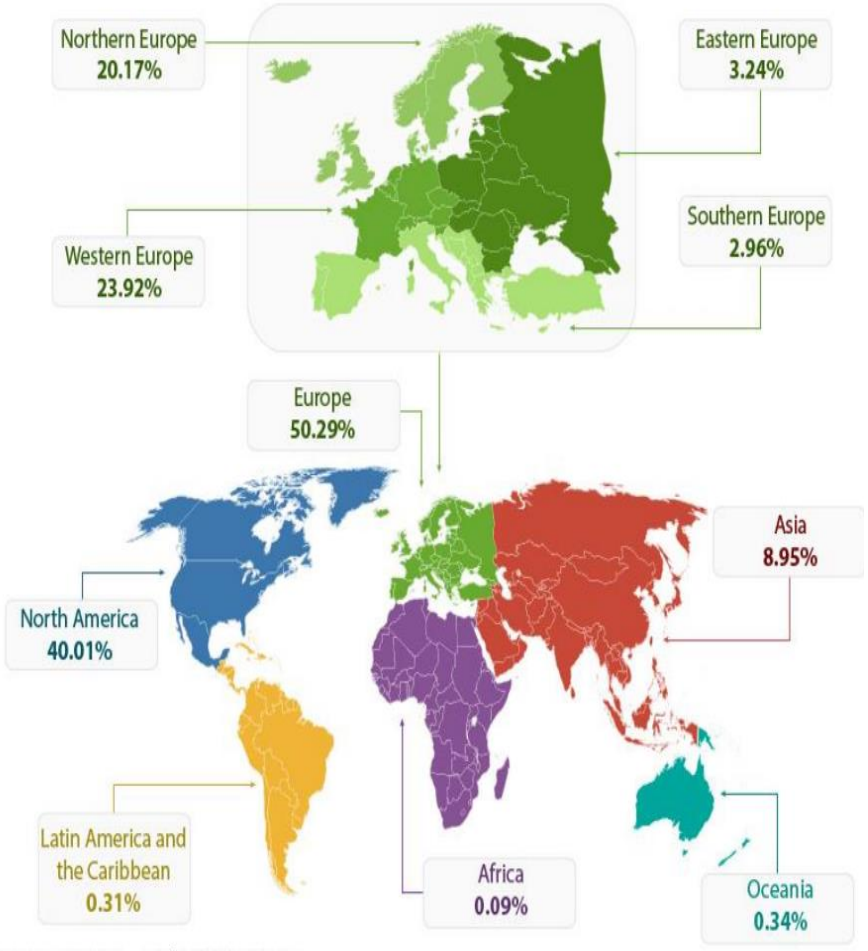
(Jiajun Zhang, CCL2018)

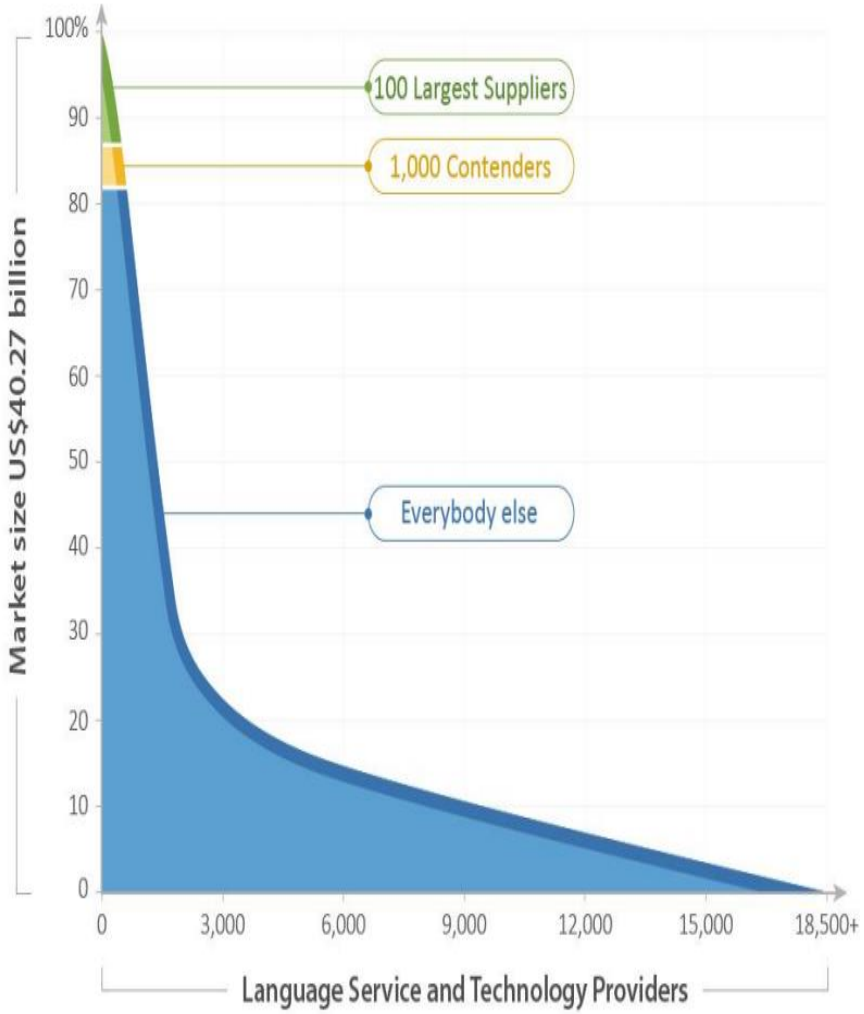
Outline

✓Machine Translation

- Demands for Machine/Human Translation
- Related Work and Current State for LRLs NMT
- Motivation
- Methodology
- Experiments
- Conclusions

Demands for Machine/Human Translation



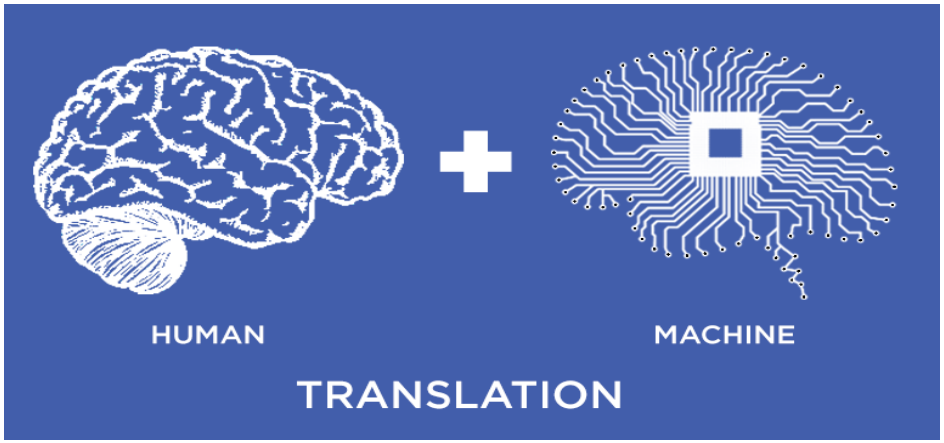


(Guoping Huang, Qcon2018)

Demands for Machine/Human Translation

- Human translation fields
  ≈ 666 million words / day
  [Pym et al., 2012]

- Machine translation fields
  >> 100 billion words / day
  [Turovsky, 2016]



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

Outline

✓Machine Translation

✓Demands for Machine/Human Translation

- Related Work and Current State for LRLs NMT
- Motivation
- Methodology
- Experiments
- Conclusions

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.

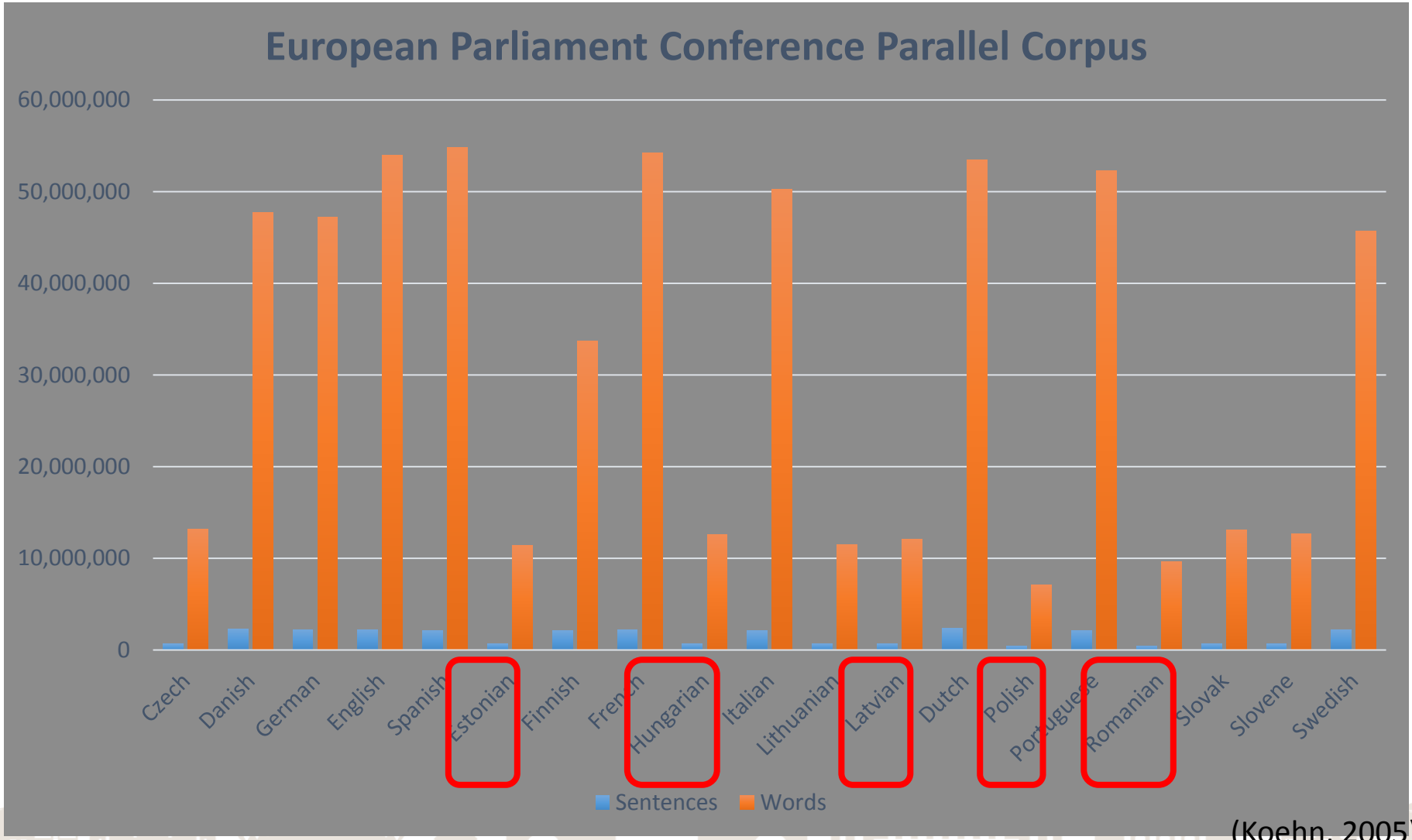
- Application value
  - In the “One Belt One Road” work, there is an urgent need for policy communication, smooth trade, financial access, facility connectivity, shared citizenship and cultural exchanges between 65 countries along the route, especially in Central/East/Western Asia.

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

Significance of Low Resource MT Research



(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 ; G*et 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

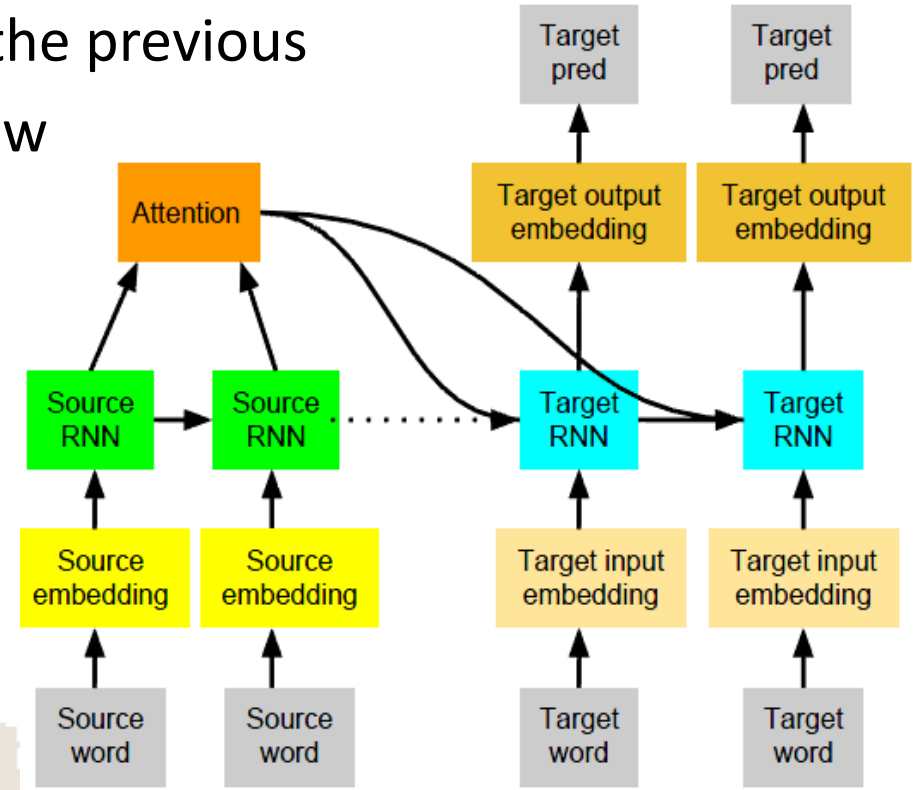
  (Francis *et al.*, 2009; Fadaee *et al.*, 2017 ; Huang *et al.*, 2016; Sennrich *et al.*, 2016c; Ribeiro *et al.*, 2018; Wang *et al.*, 2018)

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

Challenge1-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 tow Different domains.



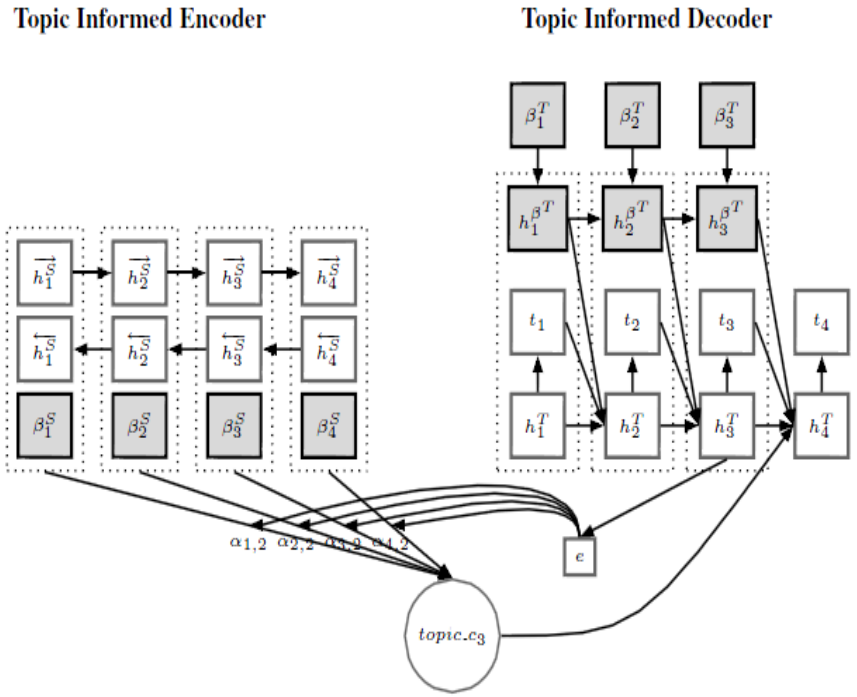
- Issue：
  - Both of them unable to use efficiently multiple HRLs.
  - Ignored character level similarity.

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

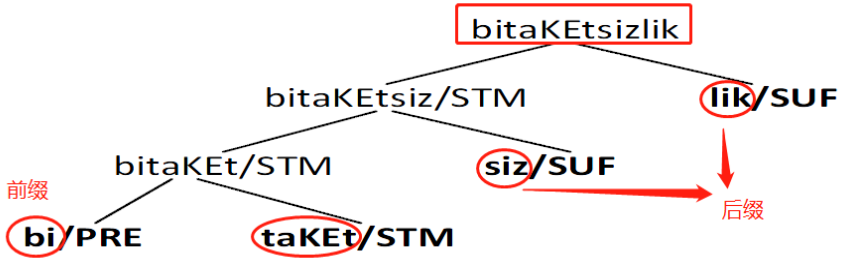
    Information and NMT.
  - Dependently train the topic model

    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.



“bitaKetsizlik”(不耐心的)

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

r · → r·
l o → lo
lo w → low
e r· → er·

{‘low’, ‘lowest’, ‘newer’, ‘wider’}

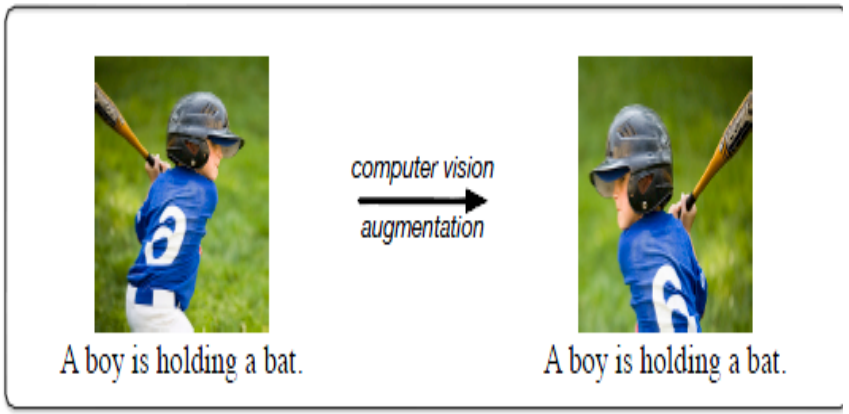
Challenge4-Syntactic and Semantic Errors

- (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

$S : s_1, ..., s_i, ..., s_n$

$T : t_1, ..., t_j, ..., t_m$

augmented pair

$S' : s_1, ..., s'_i, ..., s_n$

$T' : t_1, ..., t'_j, ..., t_m$

Outline

✓Machine Translation

✓Demands for Machine/Human Translation

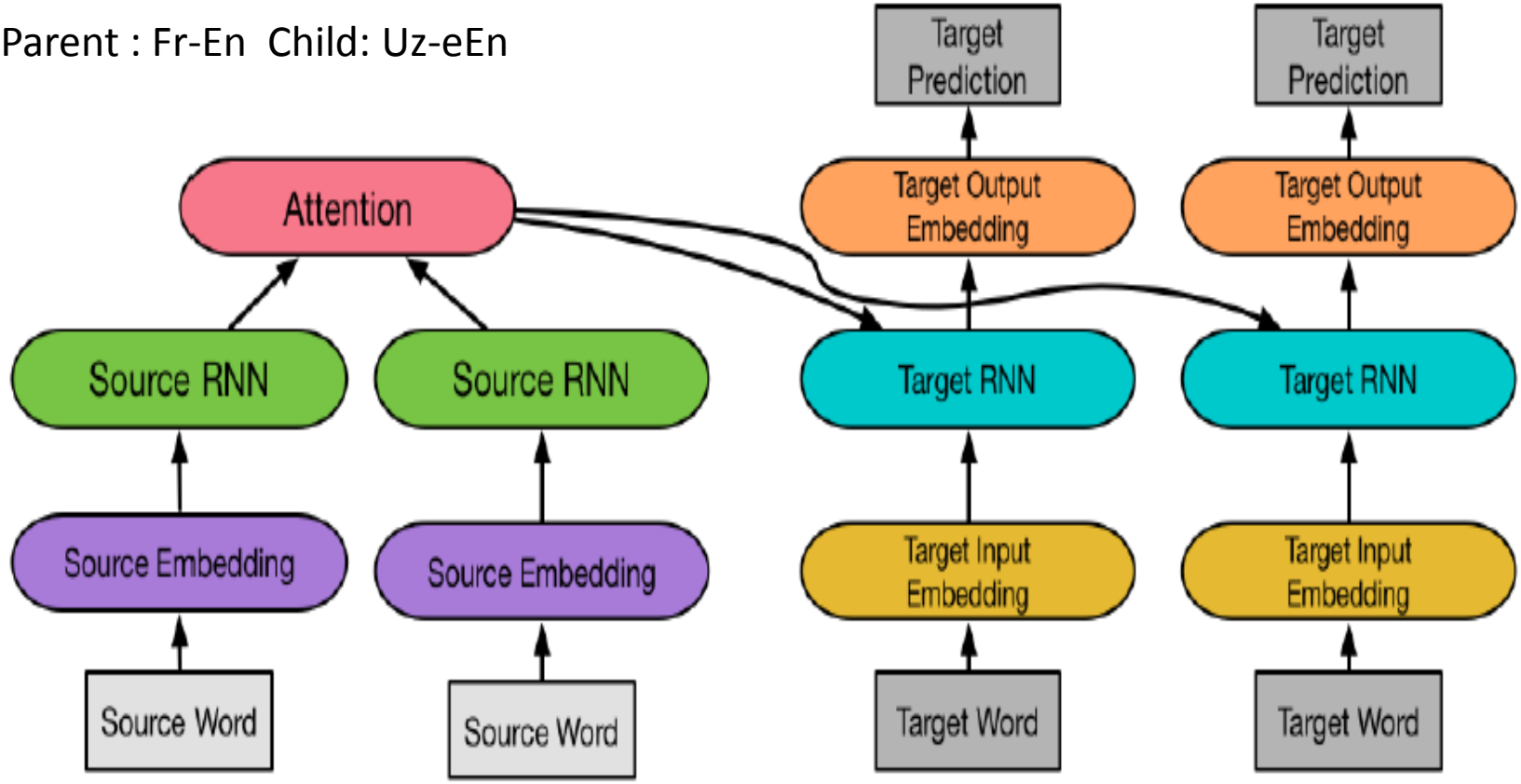
✓Related Work and Current State for LRLs NMT

- Motivation
- Methodology
- Experiments
- Conclusions

Motivation – Transfer Learning

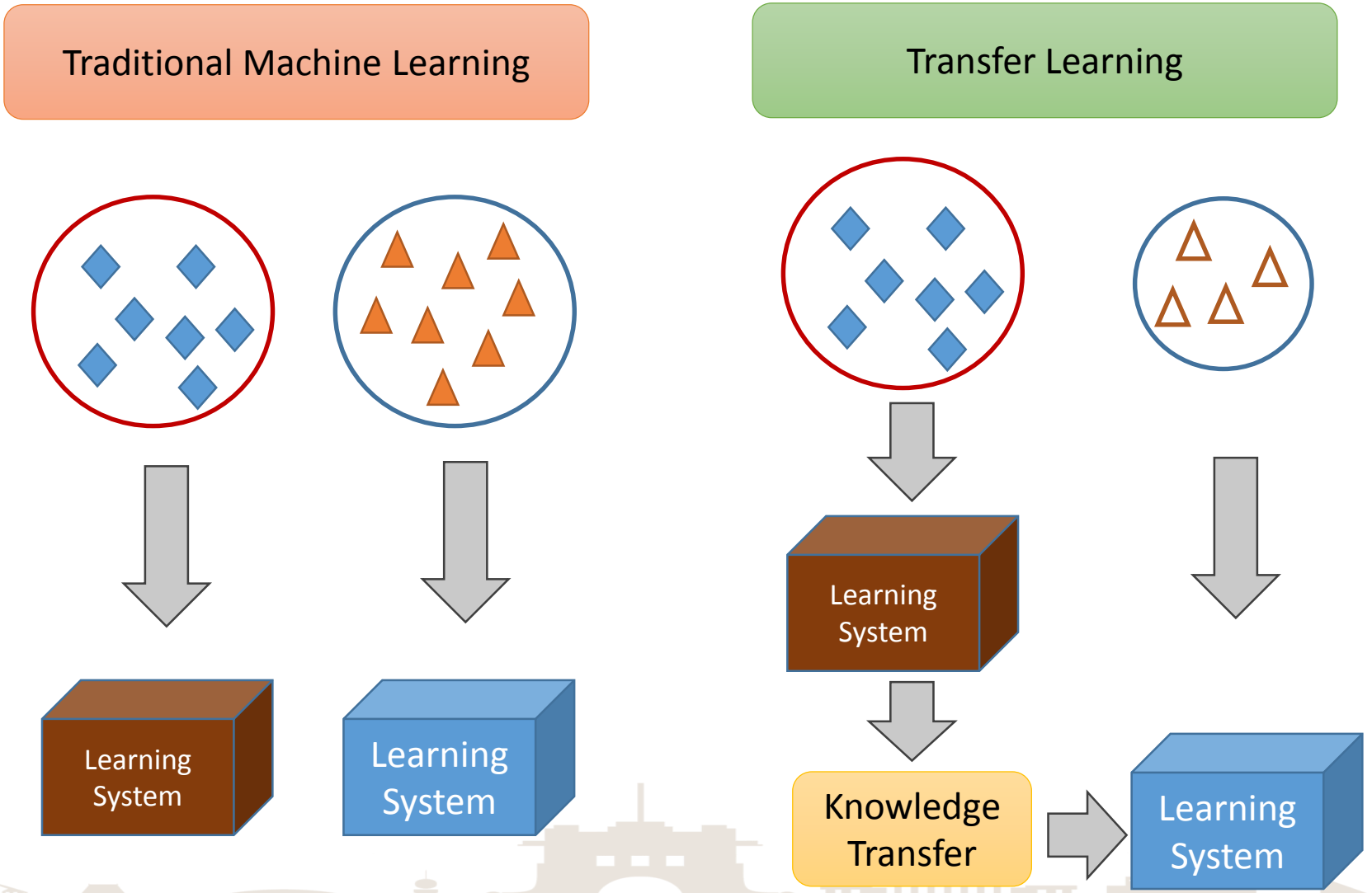
Optimal setting for transferring from **parent** model to **child** model.

Parent : Fr-En Child: Uz-eEn



[Barret Zoph et al., 2016]

Transfer Learning && Machine Learning



Transfer Learning – domain adaptation

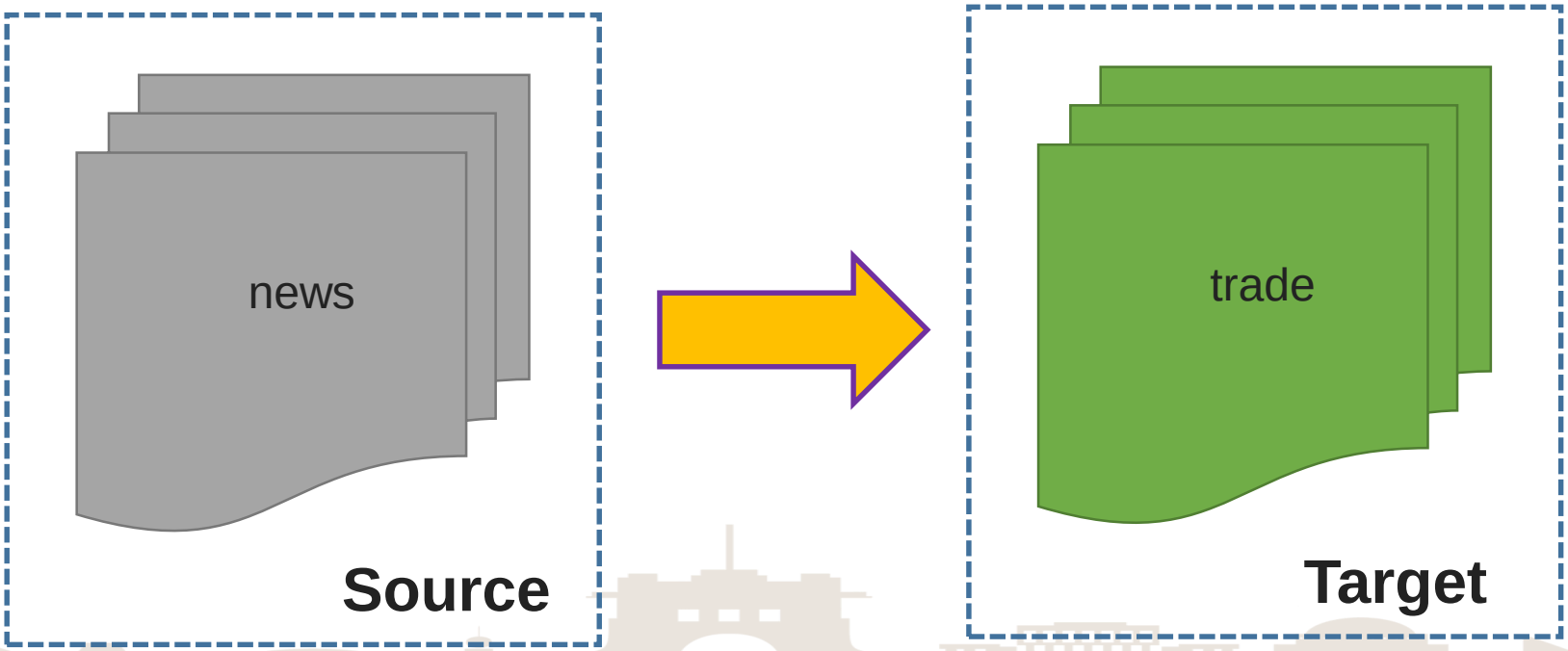
Domain Adaptation (DA) is a field associated with machine learning and transfer learning.

DA is one of the branches of transfer learning.

DA build a system on **one kind of data** and **adjust** it to apply to another.

Transfer Learning – domain adaptation

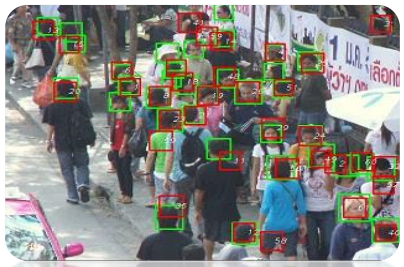
This scenario arises when we aim at learning from a source data distribution a well performing model on a different (but related) target data distribution.



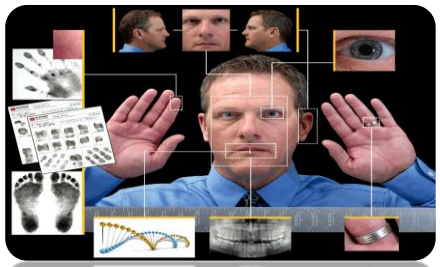
Transfer Learning – domain adaptation

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

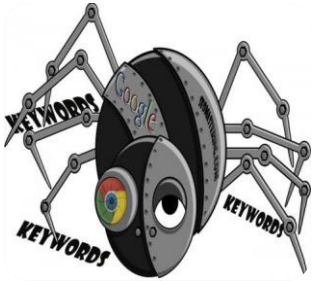
Build speech recognition system using recorded phone calls, then tune it to use as an airline reservation hotline.



CV



ER



IR

ASR

Outline

- ✓Machine Translation
- ✓Demands for Machine/Human Translation
- ✓Related Work and Current State for LRLs NMT
- ✓Motivation
- ●Methodology
- ●Experiments
- ●Conclusions

Methodology - NMT

- X, Y;Raw source and target sentences.
- Given source sentences $X = x_1, \cdots, x_i, \cdots, x_I$ and target sentence Y= $y_1, \cdots, y_i, \cdots, y_I$
- 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 - 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:
- $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\{\sum_{n=1}^{N} logP(y^{(n)}|x^{(n)};\theta)\}$
- $\hat{y} = \underset{y}{\operatorname{argmax}}\{P(y|x;\hat{\theta})\} \quad \hat{y}_j = \underset{y}{\operatorname{argmax}}\{P(y|x,\hat{y}_{<j};\hat{\theta})\}$

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.

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

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

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

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

[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.

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

$$\hat{\theta}_{L_3^{`}\rightarrow L_2} = \underset{L_3^{`}\rightarrow L_2}{\operatorname{argmax}}\left\{L(D_{L_3^{`}\rightarrow L_2}, \theta_{L_3^{`}\rightarrow L_2})\right\}$$

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

[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.

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

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

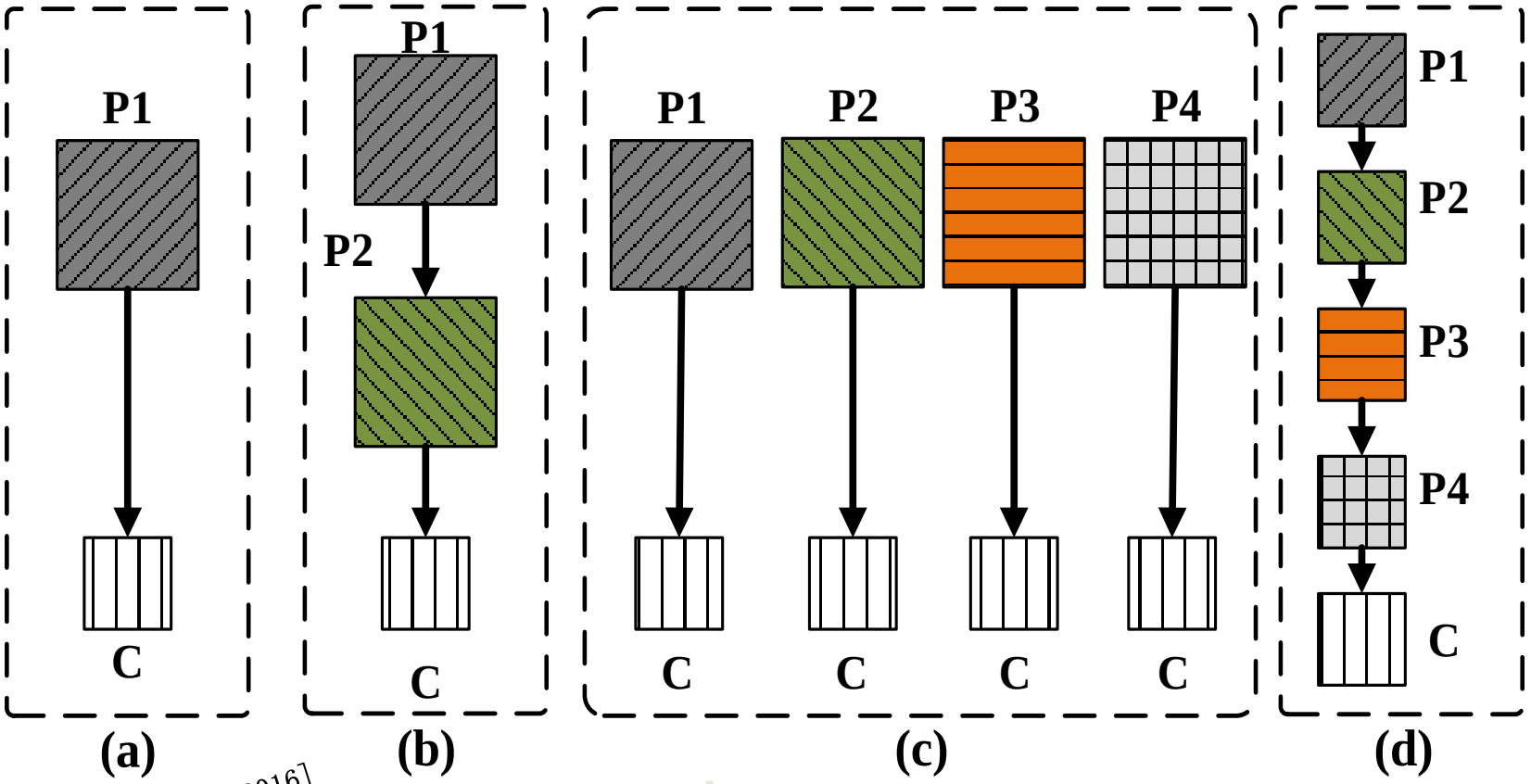
$$\vdots \quad = \quad \vdots$$

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

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

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

Multi-round Transfer Learning



[Barret Zoph et al., 2016]

[Passban et al., 2017]

This work

Methodology – Unified Transliteration

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

Methodology – Unified Transliteration

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	学校	School	**mektep**
Uy	مەكتەپ	**mektep**			
Tr	okul	**mektep**			
Fa	باغ وحش	bağça	果园	Orchard	**bağça**
Uy	باغچا	bağça			
Tr	bahçesi	**bahçe**			
Ar	غرفة القراءة	qiraaaatxana	阅览室	Reading room	**qiraetxana**
Fa	اتاق مطالعه	qiraaaat xana			
Tr	Okuma odası	**kıraathane**			
Uy	قىرائەتخانا	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. */
$Current_l \leftarrow$ the word sequence in current line among D_{sm};
$Output_c \leftarrow$ 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''. */
  $Current_l \leftarrow$ each line.decode('utf-8');
  /* split the current line with white space and save them as a list. */
  $Current_l \leftarrow Current_l$.strip().split();
  for *each word in* $Current_l$ **do**
    for *each char in each word* **do**
      /* check each char from the manually prepared mapping table. */
      each $char_{latin} \leftarrow$ each char;
    end
    /* check the each word if contains same repeated char continually. */
    if *Is Contain repeated char in each* $char_{latin}$ **then**
      compare the length of $Current_l$ and the length of Latinized $Current_l$;
      remove repeated each $char_{latin}$ from each word;
    end
    convert them into unified form sequentially;
    each word $\leftarrow$ sums of unified chars after removing repeated chars;
    final word $\leftarrow$ each word ; /* reserve the final word. */
  end
  $Output_c \leftarrow$ the joint sequence of final words with white-space
end

Outline

- ✓Machine Translation
- ✓Demands for Machine/Human Translation
- ✓Related Work and Current State for LRLs NMT
- ✓Motivation
- ✓Methodology
- ●Experiments
- ●Conclusions

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)	R=1	Ur → Ch	Uy → Ch	10.29
		Fa → Ch		28.83
		Ar → Ch		30.64^{++}
MRTL (Unified)		Ur → Ch		10.93^{*}
		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	Uy → Ch	28.28
MRTL (R=1)	Tr → Ch (0.5M)		29.89^{++}
	Tr → Ch (2.4M)		$30.88^{++\flat}$
	Tr → Ch (4.4M)		$32.74^{++\natural}$

Experiment

The effect of parent language pairs to child language pair

Method	Parent	Child	BLEU
TRANSFORMER	N/A		28.28
MRTL (R=1)	*Ur* → *Ch*	Uy → Ch	10.93
	Fa → *Ch*		29.96^{++}
	Fi → *Ch*		30.85^{++}
	Tr → Ch (2.4M)		$30.88^{++\star}$
	Ar → *Ch*		$31.64^{++\star}$
	Hu → *Ch*		$32.41^{++\circ}$
	Tr → Ch (4.4M)		$32.74^{++\dagger}$

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	Uralic	Open Subtile	4.1M		32.41^{++}
	Tr → Ch	Altaic				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	Indo-European	Tanzil	78.0K		10.93
	Fa → Ch		Open Subtitle			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	Uralic	Open Subtile	2.8M		30.85^{++}
	Hu → Ch			4.1M		32.41^{++}

Experiment

The Effect of MRTL.

Method	Round	Parent	Child	BLEU
TRANSFORMER	R=0	N/A	Uy → Ch	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		$33.54^{++‡*}$
		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^{++**}

Experiment

The Translation example of Uy→Ch between various methods

Method	Translation result
Source	*muvapiq sürük İçide yolğa qoyu@@ lmisa dölet heqsiz yolğa qoysa bolidu .*
Reference	*zai **heli qixian** nei meiyou shishi de , guo jia keyi **wuchang** shishi .* 在 合理 期限 内 没有 实施 的 ， 国 家 可以 无偿 实施 。
TRANSFORMER	*shidang xianqi **shishi** .* 适当 限期 实施 。
MANY-to-ONE	*zai shidang qixian nei **bu neng** shixing guo jia mianfei shishi .* 在 适当 期限 内 不 能 实行 国 家 免费 实施 。
$R = 1$	*zai shidang qixian nei **bu neng** you mianfei shishi **guo jia** .* 在 适当 期限 内 不 能 有 免费 实施 国 家 。
$R = 2$	*zai **heli** qijian nei shishi xi ze , guo jia keyi textbfmianfei shishi .* 在 合理 期间 内 实施 细 则 ， 国 家 可以 免费 实施 。
$R = 3$	*zai **heli qijian** nei wei shixing guo jia ke mianfei shishi .* 在 合理 期间 内 未 实行 国 家 可 免费 实施 。
$R = 4$	*dui **heli qijian** nei wei shixing de , guo jia keyi **wuchang** shishi .* 对 合理 期间 内 未 实行 的 ， 国 家 可以 无偿 实施 。

Outline

- ✓Machine Translation
- ✓Demands for Machine/Human Translation
- ✓Related Work and Current State for LRLs NMT
- ✓Motivation
- ✓Methodology
- ✓Experiments
- ●Conclusions

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.

شكرا لك

شـكريا

köszönöm

谢 谢！

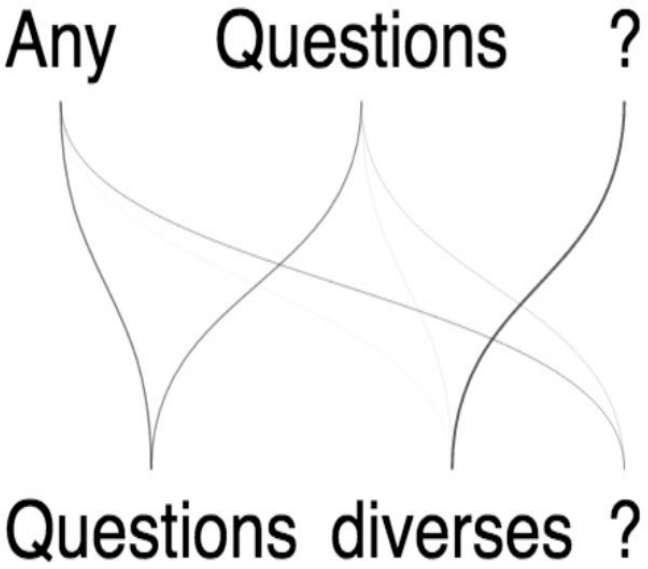
བཀའ་དྲིན་ཆེ།

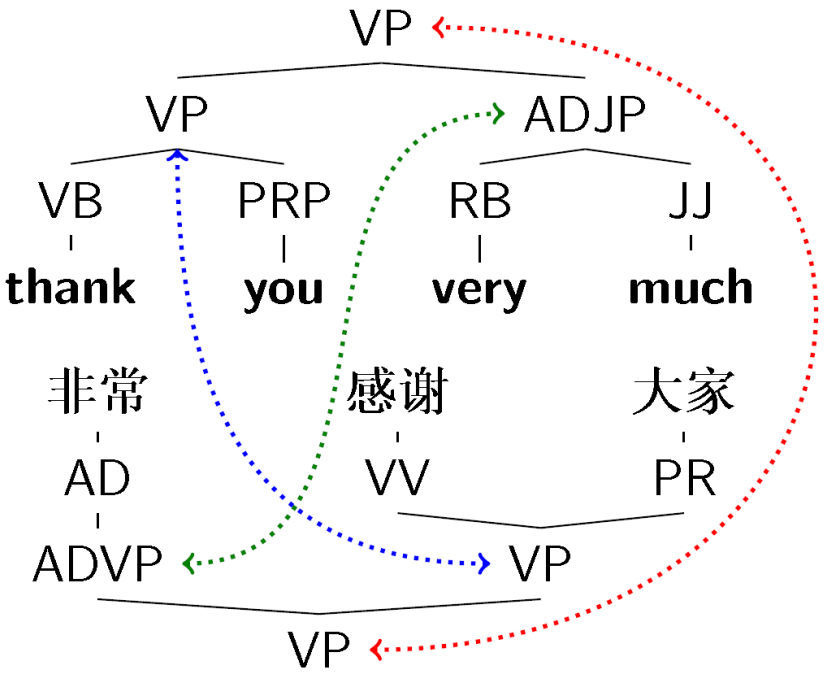
رەخمەت!

תודה

Teşekkür

Kiitos





This inspiration comes from Dzmitry Bahdanau @ *ICLR2014* .