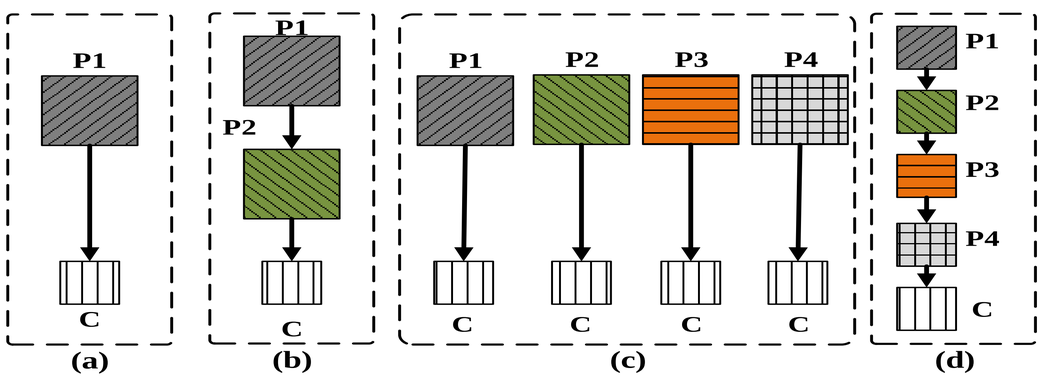
Multi-Round Transfer Learning for Low-Resource NMT Using Multiple High-Resource Languages

Mieradilijiang Maimaiti[1], Yang Liu[1], Huanbo Luan[1] , Maosong Sun[1]

[1]Department of Computer Science and Technology, Tsinghua University

Motivation and Background

- NMT has made remarkable progress in recent years, but the performance of NMT suffers from a data sparsity problem since large-scale parallel corpora are only readily available for high-resource languages (HRLs).
- Transfer learning (TL) has been used widely in low-resource languages machine translation; while TL is becoming one of the vital directions in low-resource (LR) NMT.



- However, leveraging the original TL to LR models is neither able to make full use of highly related multiple HRLs nor receive different parameters from the same parents.
- To address this issue, we present a language-independent multi-round transfer learning (MRTL) which aims to exploit HRLs effectively. Besides, to reduce the differences between HRLs and LRLs at the character level, we introduce a unified transliteration method for various language families.

Methodology

- **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.
- **Multi round fine-tuning**
  - The original TL transfers parameters of parent model into child model.

$$\theta_{L_3} = \{\langle e_{L_3}, W, e_{L_3}\rangle\} \quad \hat{\theta}_{L_3\to L_2} = \underset{\theta_{L_3\to L_2}}{\operatorname{argmax}}\{L(D_{L_3\to L_2}, \theta_{L_3\to L_2})\} \quad \theta_{L_1\to L_2} = f(\hat{\theta}_{L_3\to L_2})$$

$$\theta_{L_3^{`}\to L_2} = f(\hat{\theta}_{L_3\to L_2}) \quad \hat{\theta}_{L_3^{`}\to L_2} = \underset{L_3^{`}\to L_2}{\operatorname{argmax}}\left\{L(D_{L_3^{`}\to L_2}, \theta_{L_3^{`}\to L_2})\right\} \quad \theta_{L_1\to L_2} = f(\hat{\theta}_{L_3^{`}\to L_2})$$

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

$$\theta_{L_4\to L_2} = f(\hat{\theta}_{L_3\to L_2}) \quad \hat{\theta}_{L_4\to L_2} = \underset{\theta_{L_4\to L_2}}{\operatorname{argmax}}\{L(D_{L_4\to L_2}, \theta_{L_4\to L_2})\} \quad \theta_{L_1\to L_2} = f(\hat{\theta}_{L_{k+1}\to L_2})$$

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

- **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 ← the word sequence in current line among D_sm;
Output_c ← 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 ← each line.decode('utf-8');
    /* split the current line with white space and save them as a list. */
    Current_l ← 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 ← 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 ← sums of unified chars after removing repeated chars;
        final word ← each word ;                  /* reserve the final word. */
    end
    Output_c ← the joint sequence of final words with white-space
end
```

Experiments

Datasets

Language features

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

Characteristics of 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.

Unified transliterations

Method	Round	Parent	Child	BLEU
TRANSFORMER	R=0	N/A	Uy → Ch	28.28
MRTL (Original)	R=1	Ur → 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^{++*}

Effect of parent models

Method	Parent	Child	BLEU
TRANSFORMER	N/A	Uy → Ch	28.28
MRTL (R=1)	*Ur → 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		$32.41^{++\diamond}$
	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^{++}

Effect of MRTL method

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		$33.54^{++\ddagger *}$
		Tr (4.4M), (2.4M), Fi, Hu → Ch		$33.66^{++\star\star}$
	R=4	Ar (Unified), Tr (4.4M), Hu, Fi → Ch		$33.73^{++\star\star}$
		Tr (4.4M), Ar (Unified), Hu, Fi → Ch		$33.91^{++\star\star}$

Examples

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 。

Conclusion

- 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.

Email: meadljmm15@mails.tsinghua.edu.cn