A Teacher-Student Framework for Zero-Resource Neural Machine Translation

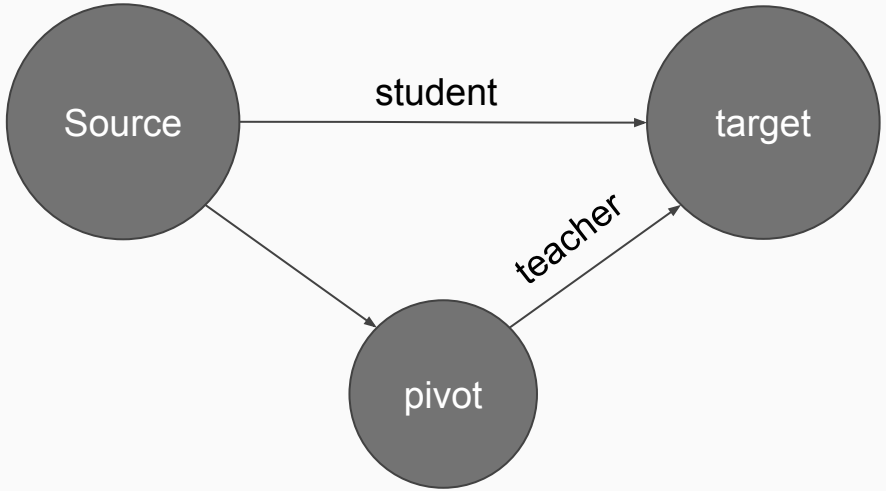
Authors: Yun Chen, Yang Liu, Yong Cheng, Victor O.K. Li
Conference: ACL - 2017
Link: http://nlp.csai.tsinghua.edu.cn/~ly/papers/acl2017_cy.pdf

Abstract

- Despite of recent remarkable progress in end-to-end NMT, the field still struggles to produce good results for languages with less data.

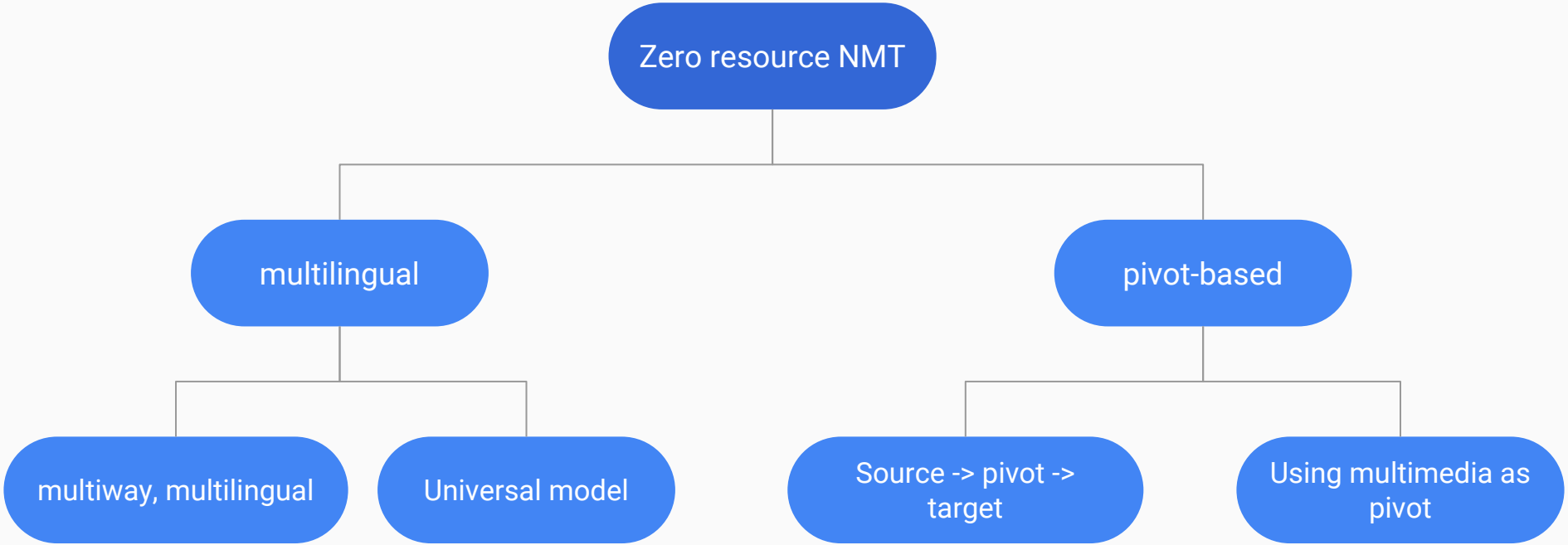
- This paper proposes a method for zero-resource (without parallel corpora for source-target) NMT by assuming that parallel sentences have close probabilities of generating a sentence in a third language.

- Based on this assumption, their method Is able to train a source-to-target NMT model (“student”) without Parallel corpora available, guided by an existing Pivot-to-target NMT model (“teacher”) on a source-pivot parallel corpus.



- Proposed method significantly improves over a baseline pivot-based model by +3.0 BLEU points across various language pairs.

Introduction



Structure of methods used in zero resource NMT

Multilingual

- Firat et al. (2016b) present a multi-way, multilingual model with shared attention to achieve zero-resource translation. They fine-tune the attention part using pseudo bilingual sentences for the zero-resource language pair.

- Another direction is to develop a universal NMT model in multilingual scenarios (Johnson et al., 2016; Ha et al., 2016). Use parallel corpora of multiple languages to train one single model, which is then able to translate a language pair without parallel corpora available

- Although these approaches prove to be effective, the combination of multiple languages in modeling and training leads to increased complexity compared with standard NMT.

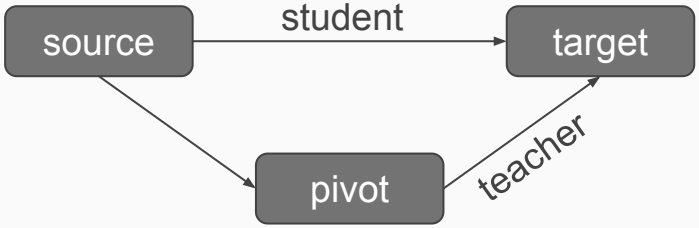
Pivot-based

- **Achieve source-to-target NMT without parallel data via a *pivot*, which is either text or image.**
- Cheng et al. (2016a) propose a pivot-based method for zero-resource NMT:

- Use of multimedia information as pivot also benefits zero-resource translation (Nakayama and Nishida, 2016)
- However, pivot-based approaches usually need to divide the decoding process into two steps, which is not only more computationally expensive, but also potentially suffers from the error propagation problem (Zhu et al., 2013).

Proposed System

- The proposed system assumes that parallel sentences should have close probabilities of generating a sentence in a third language.
- An existing “teacher” model is leveraged to guide the learning of “student” model on a source-pivot parallel corpus.



- Compared with pivot based approaches (Cheng et al., 2016a), this method allows direct parameter estimation of the intended NMT model, without the need to divide decoding into two steps.
- This strategy improves efficiency and avoids error propagation in decoding.
- Experiments on the Europarl and WMT datasets show that this approach achieves significant improvements in terms of both translation quality and decoding efficiency over a baseline pivot-based approach to zero-resource NMT on Spanish-French and German-French translation tasks.

Background

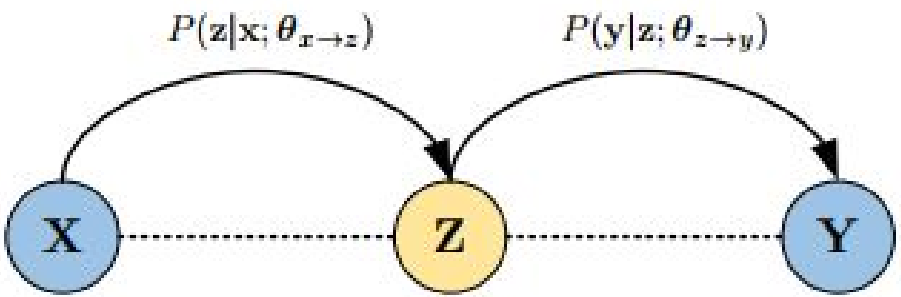
- Let x be a source-language sentence and y be a target-language sentence. We use $P(y|x; \theta_{x\to y})$ to denote a source-to-target neural translation model, where $\theta_{x\to y}$ is a set of model parameters.

- Given a source-target parallel corpus $D_{x,y}$, which is a set of parallel source-target sentences, the model parameters can be learned by maximizing the log-likelihood of the parallel corpus:

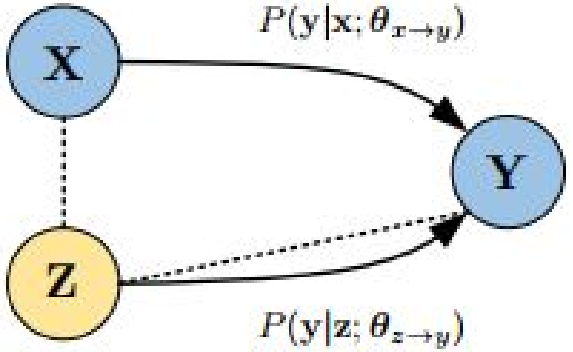
$$\hat{\boldsymbol{\theta}}_{x\to y} = \underset{\boldsymbol{\theta}_{x\to y}}{\operatorname{argmax}} \left\{ \sum_{\langle \mathbf{x},\mathbf{y}\rangle \in D_{x,y}} \log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\to y}) \right\}$$

- Given learned model parameters $\hat{\boldsymbol{\theta}}_{x\to y}$, the decision rule for finding the translation with the highest probability for a source sentence x is given by

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x\to y}) \right\}. \qquad (1)$$



(a)



(b)

Figure 1: (a) The pivot-based approach and (b) the teacher-student approach to zero-resource neural machine translation. $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ denote source, target, and pivot languages, respectively. We use a dashed line to denote that there is a parallel corpus available for the connected language pair. Solid lines with arrows represent translation directions. The pivot-based approach leverages a pivot to achieve indirect source-to-target translation: it first translates $\mathbf{x}$ into $\mathbf{z}$, which is then translated into $\mathbf{y}$. Our training algorithm is based on the translation equivalence assumption: if $\mathbf{x}$ is a translation of $\mathbf{z}$, then $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ should be close to $P(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_{z\to y})$. Our approach directly trains the intended source-to-target model $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ ("student") on a source-pivot parallel corpus, with the guidance of an existing pivot-to-target model $P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y})$ ("teacher").

What's the problem with this pivot based system?

- Due to the exponential search space of pivot sentences, the decoding process of translating an unseen source sentence x has to be divided into two steps:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \left\{ P(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x \to z}) \right\}, \quad (3)$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}_{z \to y}) \right\}. \quad (4)$$

- The above two-step decoding process potentially suffers from the error propagation problem (Zhu et al., 2013): the translation errors made in the first step (i.e., source-to-pivot translation) will affect the second step (i.e., pivot-to-target translation).

- Therefore, it is necessary to explore methods to directly model source-to-target translation without parallel corpora available.

So why the pivot?

- Simple and easy-to-implement, **pivot-based methods have been widely used in SMT for translating zero-resource language pairs** (de Gispert and Marino˜ , 2006; Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008; Wu and Wang, 2009; Zahabi et al., 2013; Kholy et al., 2013).

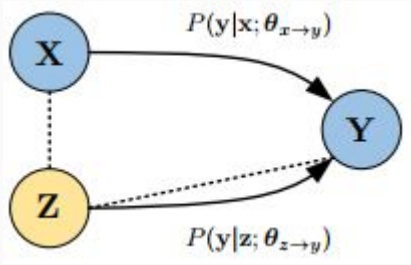
- As pivot based methods are **agnostic to model structures**, they have been adapted to NMT recently (Cheng et al., 2016a; Johnson et al., 2016).

Approach

- The basic idea is to use a pre-trained pivot-to-target model (“teacher”) to guide the learning process of a source-to-target model (“student”) without training data available on a source-pivot parallel corpus.

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x \to y}) \right\}. \qquad (1)$$

- One advantage of this approach is that Equation (1) can be used as the decision rule for decoding, which avoids the error propagation problem faced by two-step decoding in pivot-based approaches.



- As shown in Figure, we still assume that a source-pivot parallel corpus $D_{x,z}$ and a pivot-target parallel corpus $D_{z,y}$ are available. Unlike pivot-based approaches, we first use the pivot-target parallel corpus $D_{z,y}$ to obtain a teacher model $P(y|z; \hat{\theta}_{z \to y})$, where $\hat{\theta}_{z \to y}$ is a set of learned model parameters. Then, the teacher model “teaches” the student model $P(y|x; \theta_{x \to y})$ on the source-pivot parallel corpus $D_{x,z}$ based on the following assumptions.

Assumption 1

If a source sentence x is a translation of a pivot sentence z, then the probability of generating a target sentence y from x should be close to that from its counterpart z.

We can further introduce a word-level assumption:

Assumption 2

If a source sentence x is a translation of a pivot sentence z, then the probability of generating a target word y from x should be close to that from its counterpart z, given the already obtained partial translation $y_{<j}$.

These two assumptions are empirically verified in our experiments

Sentence level teaching

Given a source-pivot parallel corpus $D_{x,z}$, the training objective based on Assumption 1 is defined as follows:

$$\mathcal{J}_{\text{SENT}}(\boldsymbol{\theta}_{x\to y}) = \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \text{KL}\Big(P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y})\Big\|P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big), \quad (5)$$

where the KL divergence sums over all possible target sentences:

$$\text{KL}\Big(P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y})\Big\|P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big) = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}) \log \frac{P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y})}{P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})}. \quad (6)$$

As the teacher model parameters are fixed, the training objective can be equivalently written as

$$\mathcal{J}_{\text{SENT}}(\boldsymbol{\theta}_{x\to y}) = -\sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}}\Big[\log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big]. \quad (7)$$

Sentence level training

In training, our goal is to find a set of source-to-target model parameters that minimizes the training objective:

$$\hat{\boldsymbol{\theta}}_{x \to y} = \underset{\boldsymbol{\theta}_{x \to y}}{\operatorname{argmin}} \left\{ \mathcal{J}_{\text{SENT}}(\boldsymbol{\theta}_{x \to y}) \right\}. \quad (8)$$

- With learned source-to-target model parameters $\hat{\boldsymbol{\theta}}_{x \to y}$, they use the standard decision rule as shown in Equation (1) to find the translation ŷ for a source sentence x.
- However, a major difficulty faced by their approach is the intractability in calculating the gradients because of the exponential search space of target sentences.
- To address this problem, it is possible to construct a subspace by either sampling (Shen et al., 2016), generating a k-best list (Cheng et al., 2016b) or mode approximation (Kim and Rush, 2016). Then, standard stochastic gradient descent algorithms can be used to optimize model parameters.

Word level training

Instead of minimizing the KL divergence between the teacher and student models at the sentence level, we further define a training objective at the word level based on Assumption 2:

$$\mathcal{J}_{\text{WORD}}(\boldsymbol{\theta}_{x\to y}) = \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}}\Big[J(\mathbf{x},\mathbf{y},\mathbf{z},\hat{\boldsymbol{\theta}}_{z\to y},\boldsymbol{\theta}_{x\to y})\Big], \quad (9)$$

Where:

$$J(\mathbf{x},\mathbf{y},\mathbf{z},\hat{\boldsymbol{\theta}}_{z\to y},\boldsymbol{\theta}_{x\to y}) = \sum_{j=1}^{|\mathbf{y}|} \text{KL}\Big(P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y})\Big\| P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y})\Big). \quad (10)$$

Equation (9) suggests that the teacher model P(y|z, y<j ; $\hat{\theta}_{z\to y}$) “teaches” the student model P(y|x, y<j ; $\theta_{x\to y}$) in a word-by-word way. Note that the KL-divergence between the two models is defined at the word level:

$$\text{KL}\Big(P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y})\Big\|P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y})\Big) = \sum_{y\in\mathcal{V}_y} P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y}) \log \frac{P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y})}{P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y})},$$

where $\mathbf{v}_y$ is the target vocabulary

Word level training

As the parameters of the teacher model are fixed, the training objective can be equivalently written as:

$$\mathcal{J}_{\text{WORD}}(\boldsymbol{\theta}_{x\to y}) = -\sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}}\left[S(\mathbf{x},\mathbf{y},\mathbf{z},\hat{\boldsymbol{\theta}}_{z\to y},\boldsymbol{\theta}_{x\to y})\right], \quad (11)$$

Where:

$$S(\mathbf{x},\mathbf{y},\mathbf{z},\hat{\boldsymbol{\theta}}_{z\to y},\boldsymbol{\theta}_{x\to y}) = \sum_{j=1}^{|\mathbf{y}|}\sum_{y\in\mathcal{V}_y} P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y}) \times \log P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y}). \quad (12)$$

They use similar approaches as described in sentence level training for approximating the full search space with sentence-level teaching. After obtaining $\hat{\theta}_{x\to y}$, the same decision rule as shown in Equation (1) can be utilized to find the most probable target sentence ŷ for a source sentence x

Conclusion

Conclusion

- This paper proposes a novel framework to train the student model without parallel corpora available under the guidance of the pre-trained teacher model on a source-pivot parallel corpus.

- They introduce sentence-level and word-level teaching to guide the learning process of the student model.

- They also analyze zero-resource translation with small source-pivot data, and combine our word level sampling method with initialization and parameter freezing suggested by (Zoph et al., 2016). The experiments on the Europarl corpus show that our approach obtains an significant improvement over the pivot-based baseline.