Machine Translation with Weakly Paired Documents

[1]Lijun Wu, [2]Jinhua Zhu, [3]Fei Gao, [4]Di He,
[5]Tao Qin, [1]Jianhuang Lai and [5]Tie-Yan Liu
[1]Sun Yat-sen University; [2]University of Science and Technology of China;
[3]Institute of Computing Technology, Chinese Academy of Sciences;
[4]Peking University, [5]Microsoft Research Asia

1. Introduction

- NMT achieves strong performance in rich-resource language pairs with large amount of parallel data.
- **Low-resource** language pairs have much lower translation accuracy due to the lack of bilingual sentence pairs.
- **Unsupervised** machine translation has been explored with monolingual data only.
- In reality, large amount of **weakly paired bilingual documents** can be leveraged.
- *We propose to boost the unsupervised machine translation with weakly paired documents using two innovated components.*
- *We achieve strong performances in various language pairs and reduce the gap between supervised and unsupervised translation up to 50%.*

2. Approach

- We propose to leverage weakly paired bilingual documents from **Wikipedia**.
- Notations:
  - $D = \{(d_i^X, d_i^Y)\}$ as the set of weakly paired documents (e.g., two cross-lingual linked Wikipedia pages)
  - n_i^X, n_i^Y are the number of sentences in paired documents d_i^X, d_i^Y, usually $n_i^X != n_i^Y$
  - x, y are the sentences of language X, Y

♦ Mining implicitly aligned sentence pairs

- e_w, cross-lingual word embedding from MUSE
- p_w, the estimated frequency from the document
- a, predefined parameter and $\widehat{e_s}$ is the weighted sentence embedding
- u_1, the first principal component from all sentence embedding

$$\begin{aligned} \hat{e}_s &= \sum_{w \in s} \frac{a}{a + p(w)} e_w, \\ u_1 &\leftarrow PCA(E), \\ e_s &= \hat{e}_s - u_1 u_1^T \hat{e}_s. \end{aligned}$$

- Select sentence pairs by $\mathrm{sim}(s^X, s^Y) = \frac{<e_{s^X}, e_{s^Y}>}{\|e_{s^X}\| \|e_{s^Y}\|}$ larger than c_1, also ensure this pair is larger than others pairs by c_2
- The implicitly aligned sentence training loss of two-sides is

$$\begin{aligned} L_p(S;\theta) = &-\frac{1}{|S|} \sum_{(s^X, s^Y) \in S} \log P_{X \to Y}(s^Y | s^X; \theta) \\ &- \frac{1}{|S|} \sum_{(s^X, s^Y) \in S} \log P_{Y \to X}(s^X | s^Y; \theta). \end{aligned}$$

♦ Aligning Topic Distribution

- Translate d_i^X to $\widehat{d_i^Y}$
- Evaluate the word distribution between d_i^Y and $\widehat{d_i^Y}$
- Feed pair $(s_{i,k}^X, \widehat{s_{i,k}^Y})$ into NMT model and calculate $P(w^Y; d_i^X)$ by

$$P(w_{i,k,t}^Y | s_{i,k}^X, \hat{s}_{i,k,<t}^Y) \sim P_{X \to Y}(w_t^Y | s_{i,k}^X, \hat{s}_{i,k,<t}^Y; \theta),$$
$$P(w^Y; d_i^X, \theta) \propto \sum_{i,k,t} P(w_t | \hat{s}_{i,k,<t}^Y),$$

- The ground-truth document word distribution is

$$P(w^Y; d_i^Y) = \frac{\#w \text{ in } d_i^Y}{\#token \text{ in } d_i^Y}.$$

- The document alignment loss of $X \to Y$ is

$$L_d(D;\theta, X \to Y) = \frac{1}{|D|} \sum_{(d_i^X, d_i^Y) \in D} KL(P(w^Y; d_i^Y) || P(w^Y; d_i^X, \theta)).$$

- The detailed two-sides loss

$$L_d(D;\theta) = L_d(D;\theta, X \to Y) + L_d(D;\theta, Y \to X).$$

3. Algorithm

- The overall loss function is

$$L = L_m(M;\theta) + \alpha L_p(S;\theta) + \beta L_d(D;\theta).$$

Algorithm 1 Training Algorithm

Require: Initial translation model with parameter θ; monolingual dataset M, implicitly aligned sentence pairs dataset S, weakly paired documents dataset D; optimizer *Opt*

1: **while** not converged **do**
2: Randomly sample a mini-batch monolingual sentences from M, implicitly aligned sentence pairs from S and weakly paired documents from D
3: Calculate loss L_m, L_p and L_d
4: Update θ by minimizing the overall objective L using optimizer *Opt*
5: **end while**

- L_m is the original unsupervised NMT training loss

4. Experiments

- **Data Statistics**

Language	#Wiki Documents
English	5,684,240
German	2,201,782
Spanish	1,389,469
Romanian	387,627

Task	#Document Pairs
English-German	948,631
English-Spanish	836,564
English-Romanian	87,289

- **Overall Results**

Unsupervised Method	En→De	De→En	En→Es	Es→En	En→Ro	Ro→En
Lample et al. (2017)	9.6	13.3	-	-	-	-
Yang et al. (2018)	10.9	14.6	-	-	-	-
NMT (Lample et al., 2018)	17.2	21.0	19.7	20.0	21.2	19.5
PBSMT (Lample et al., 2018)	17.9	22.9	-	-	22.0	23.7
PBSMT + NMT (Lample et al., 2018)	20.2	25.2	-	-	25.1	23.9
NMT + First Wiki Sentence	16.3	19.3	17.3	18.3	19.4	18.1
NMT + Document Translation	12.0	14.9	14.5	15.3	16.8	15.7
Ours	**24.2**	**30.3**	**28.1**	**27.6**	**30.1**	**27.6**
Supervised NMT	33.6	38.2	33.2	32.9	32.8	35.4

5. Studies

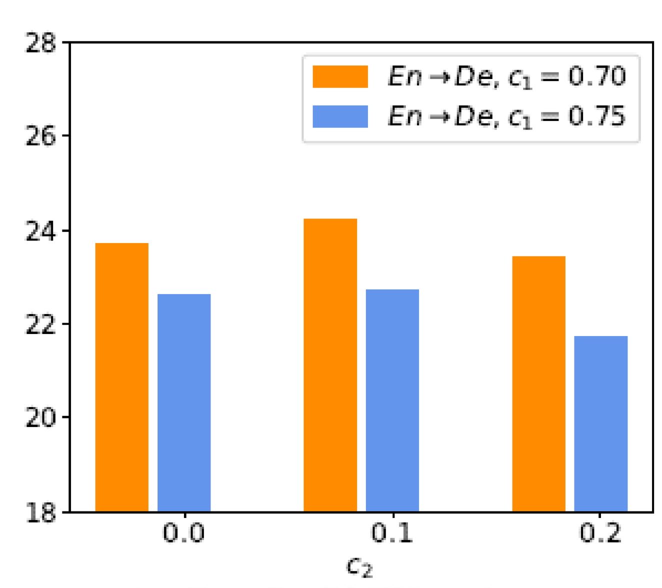
- **Analysis**

Ablation Study

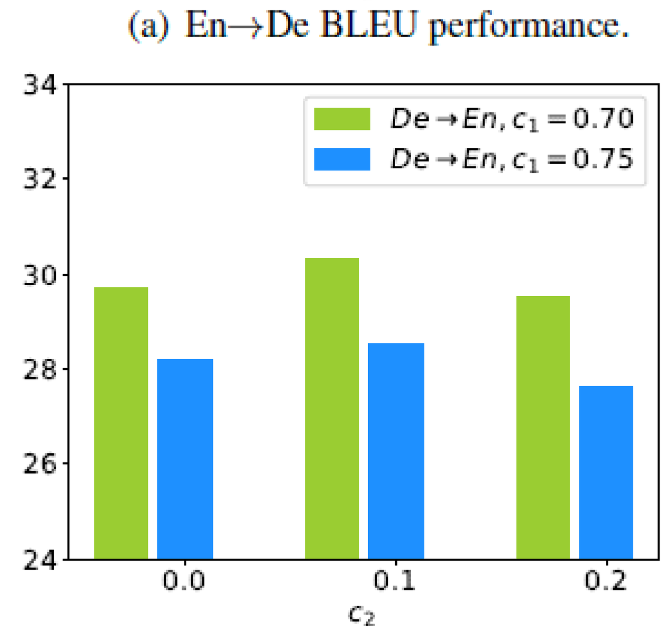
Our Method	En→De	De→En
with L_p and L_d	24.2	30.3
without L_d	22.9	28.7
without L_p	18.5	23.3

Impact of Sentence Quality

	English-German		
c_1/c_2	0.0	0.1	0.2
0.70	257,947	199,965	132,403
0.75	100,497	84,271	58,814



(a) En→De BLEU performance.



(b) De→En BLEU performance.

Contact
- wulijun3@mail2.sysu.edu.cn