Word translation without parallel data

Ким Алёна

НИУ ВШЭ

22 марта, 2019

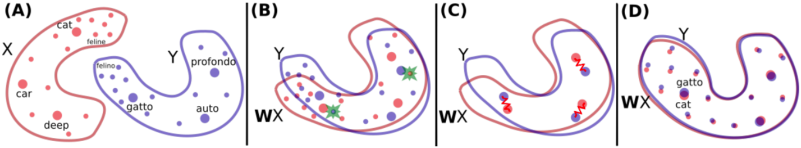
План

- машинный перевод: необходимы данные
- два уровня: word & sentence
    - adversarial training
    - метрика Cross-Domain Similarity Local Scaling (CSLS)
    - unsipervised validation criterion
- результаты

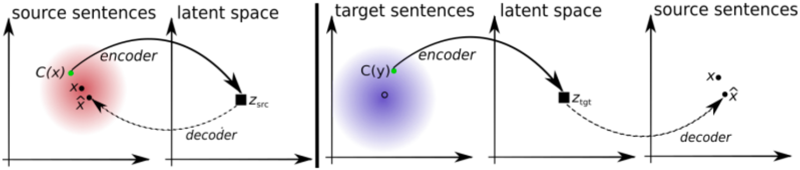
Word level

$$W^* = \underset{W \in M_d(\mathbb{R})}{\text{argmin}} \; ||WX - Y||_F$$

translation t of any word s: $t = \text{argmax}_t \, cos(Wx_s, y_t)$



Sentence level



Adversarial Training. Word Level

source: $X = x1, ..., xn$, **target**: $Y = y1, ..., ym$

Discriminator objective θ_D

probability $P_{\theta_D}(source = 1|z)$ that z is a vector from source language

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(source = 1|Wx_i) -$$

$$-\frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(source = 0|y_i)$$

Mapping objective W

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(source = 0|Wx_i) -$$

$$-\frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(source = 1|y_i)$$

Cross-Domain Similarity Local Scaling (CSLS)

- K-NN is asymetric
- $r_T(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_T(Wx_s)} \cos(Wx_s, y_t)$
- $CSLS(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_t(Wx_s) - r_s(y_t)$

Sentence level

$\mathcal{Z}^S = (x^s_1, \dots, x^s_{|\mathcal{W}_S|})$
$\mathcal{Z}^N = (x^t_1, \dots, x^t_{|\mathcal{W}_T|})$
input sentence: $\mathbf{x} = (x_1, \dots, x_m)$, **language** l
encoder: $e_{\theta_{enc},\mathcal{Z}}(\mathbf{x}, l) = e(\mathbf{x}, l) \Longrightarrow \mathrm{z} = (\mathrm{z}_1, \dots, z_m)$
decoder: $d_{\theta_{dec},\mathcal{Z}}(\mathbf{x}, l) = d(\mathbf{x}, l) \Longrightarrow \mathrm{y} = (\mathrm{y}_1, \dots, y_k)$
encoder and decoder - Bi-LSTM

Objective functions

Denoising Auto-Encoding

$$\mathcal{L}_{auto}(\theta_{enc}, \theta_{dec}, \mathcal{Z}, l) = \mathbb{E}_{x\sim\mathcal{D}_l, \hat{x}\sim d(e(C(x),l),l)}[\Delta(\hat{x}, x)]$$

Cross Domain Training

$$\mathcal{L}_{cd}(\theta_{enc}, \theta_{dec}, \mathcal{Z}, l_1, l_2) = \mathbb{E}_{x\sim\mathcal{D}_{l_1}, \hat{x}\sim d(e(C(M(x)),l_2),l_1)}[\Delta(\hat{x}, x)]$$

Adversarial Training

$$\mathcal{L}_{adv}(\theta_{enc}, \mathcal{Z}|\theta_D) = -\mathbb{E}_{(x_i,l_i)}[\log p_D(l_j|e(x_i, l_i))]$$

$$\mathcal{L}_D(\theta_D|\theta, \mathcal{Z}) = -\mathbb{E}_{(x_i,l_i)}[\log p_D(l_i|e(x_i, l_i))]$$

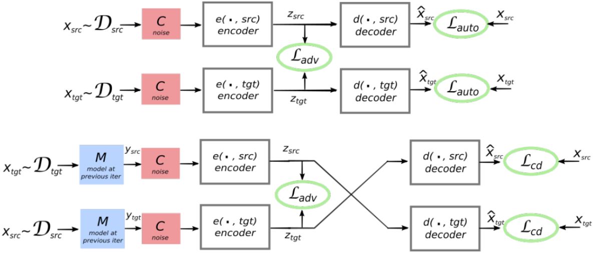
Final Objective

Final Objective Function

$$\mathcal{L}(\theta_{enc},\theta_{dec},\mathcal{Z}) = \lambda_{adv}\mathcal{L}_{adv}(\theta_{enc},\mathcal{Z}|\theta_D)+$$

$$+\lambda_{cd}[\mathcal{L}_{cd}(\theta_{enc},\theta_{dec},\mathcal{Z},src,tgt)+\mathcal{L}_{cd}(\theta_{enc},\theta_{dec},\mathcal{Z},tgt,src)]+$$

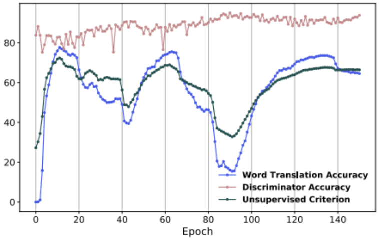
$$+\lambda_{auto}[\mathcal{L}_{auto}(\theta_{enc},\theta_{dec},\mathcal{Z},src)+\mathcal{L}_{auto}(\theta_{enc},\theta_{dec},\mathcal{Z},tgt)]$$

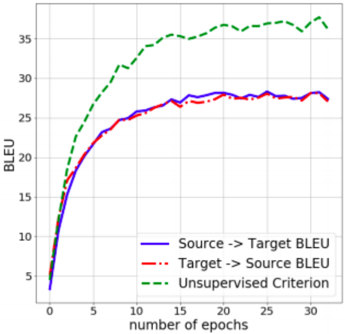


Validation criterion

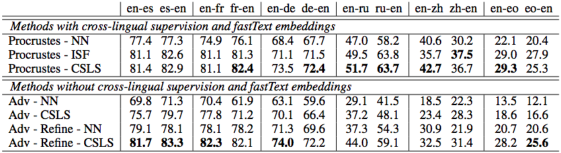
$$\begin{gathered}
M_{src->tgt}(x) = d(e(x, src), tgt) \\
MS(e, d, \mathcal{D}_s rc, \mathcal{D}_t gt) = \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{src}}[BLEU(x, M_{src->tgt} \circ M_{tgt->src}(x))] + \\
+\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{tgt}}[BLEU(x, M_{tgt->src} \circ M_{src->tgt}(x))]
\end{gathered}$$

Validation criterion

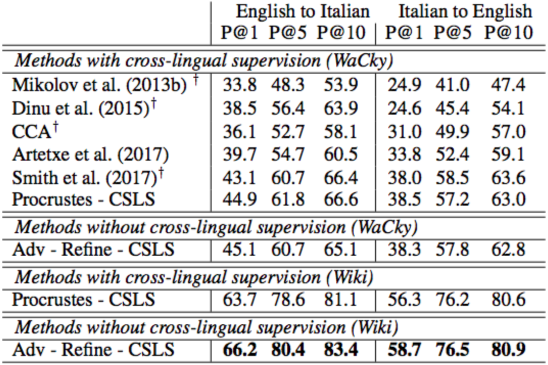




Experiments



	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-eo	eo-en
Methods with cross-lingual supervision and fastText embeddings												
Procrustes - NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	22.1	20.4
Procrustes - ISF	81.1	82.6	81.1	81.3	71.1	71.5	49.5	63.8	35.7	**37.5**	29.0	27.9
Procrustes - CSLS	81.4	82.9	81.1	**82.4**	73.5	**72.4**	**51.7**	**63.7**	**42.7**	36.7	**29.3**	25.3
Methods without cross-lingual supervision and fastText embeddings												
Adv - NN	69.8	71.3	70.4	61.9	63.1	59.6	29.1	41.5	18.5	22.3	13.5	12.1
Adv - CSLS	75.7	79.7	77.8	71.2	70.1	66.4	37.2	48.1	23.4	28.3	18.6	16.6
Adv - Refine - NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	54.3	30.9	21.9	20.7	20.6
Adv - Refine - CSLS	**81.7**	**83.3**	**82.3**	82.1	**74.0**	72.2	44.0	59.1	32.5	31.4	28.2	**25.6**



	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision (WaCky)						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Dinu et al. (2015)†	38.5	56.4	63.9	24.6	45.4	54.1
CCA†	36.1	52.7	58.1	31.0	49.9	57.0
Artetxe et al. (2017)	39.7	54.7	60.5	33.8	52.4	59.1
Smith et al. (2017)†	43.1	60.7	66.4	38.0	58.5	63.6
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0
Methods without cross-lingual supervision (WaCky)						
Adv - Refine - CSLS	45.1	60.7	65.1	38.3	57.8	62.8
Methods with cross-lingual supervision (Wiki)						
Procrustes - CSLS	63.7	78.6	81.1	56.3	76.2	80.6
Methods without cross-lingual supervision (Wiki)						
Adv - Refine - CSLS	**66.2**	**80.4**	**83.4**	**58.7**	**76.5**	**80.9**

References

[1]. https://arxiv.org/pdf/1710.04087.pdf
[2]. https://arxiv.org/abs/1711.00043