Meta-Learning for Low-Resource Neural Machine Translation

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho & Victor O.K. Li
EMNLP 2018

Objectif

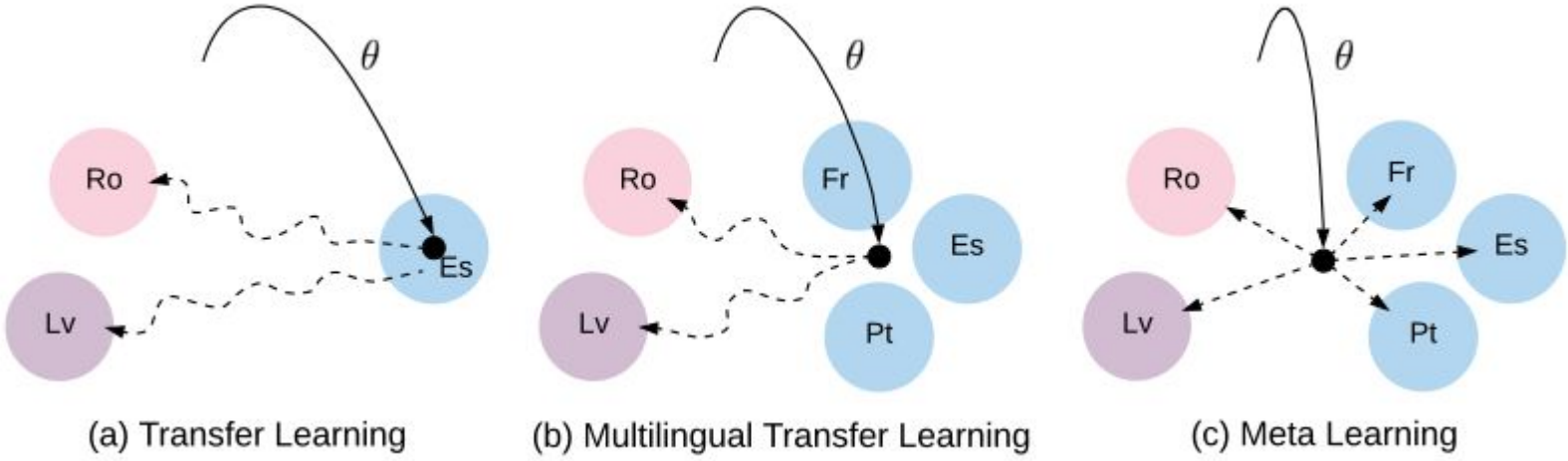
- traduire dans des langues avec peu de données

Par rapport aux travaux de Guillaume & Alexis :

→ ici **on utilise des données appairées**

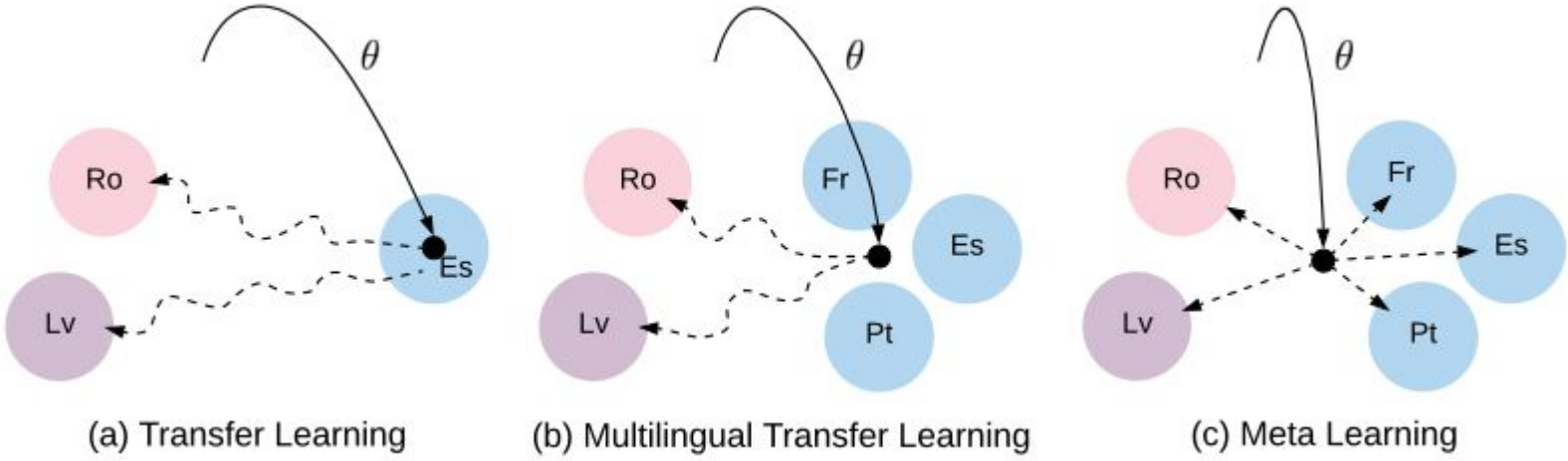
Meta-learning

Apprendre **un état initial** tel qu'il est possible de fine-tune sur d'autres tâches (langues) *facilement*



Meta-learning

Apprendre **un état initial** tel qu'il est possible de fine-tune sur d'autres tâches (langues) *facilement*



Cet état n'est pas forcément un bon modèle !

MAML

Utiliser des paires avec beaucoup de données pour *meta-learn* et les paires rares pour *apprendre*

$$\theta^* = \text{Learn}(\mathcal{T}^0; \underbrace{\text{MetaLearn}(\mathcal{T}^1, \ldots, \mathcal{T}^K)}_{\theta^0}).$$

θ^0 Initialisation meta-apprise

On **simule** K épisodes de traduction

A chaque épisode k sur une tâche T^k

- passe **NMT classique** à partir du θ actuel sur $D_{\mathcal{T}^k}$

gradient

$$\theta'_k = \text{Learn}(D_{\mathcal{T}^k}; \theta) = \theta - \eta \nabla_\theta \mathcal{L}^{D_{\mathcal{T}^k}}(\theta)$$

$$\text{Learn}(D_{\mathcal{T}}; \theta^0) = \arg\max_\theta \mathcal{L}^{D_{\mathcal{T}}}(\theta)$$

$$= \arg\max_\theta \sum_{(X,Y) \in D_{\mathcal{T}}} \log p(Y|X, \theta)$$

MAML

Utiliser des paires avec beaucoup de données pour *meta-learn* et les paires rares pour *apprendre*

$$\theta^* = \text{Learn}(\mathcal{T}^0; \underbrace{\text{MetaLearn}(\mathcal{T}^1, \ldots, \mathcal{T}^K)}_{\theta^0 \text{ Initialisation meta-apprise}}).$$

On **simule** K épisodes de traduction

A chaque épisode k sur une tâche T^k

- passe **NMT classique** à partir du θ actuel sur $D_{\mathcal{T}^k}$ **gradient**

$$\theta'_k = \text{Learn}(D_{\mathcal{T}^k}; \theta) = \theta - \eta \overbrace{\nabla_\theta \mathcal{L}^{D_{\mathcal{T}^k}}(\theta)}$$

- évaluation du nouveau θ'_k **méta gradient**

$$\theta \leftarrow \theta - \eta' \sum_k \overbrace{\nabla_\theta \mathcal{L}^{D'_{\mathcal{T}^k}}(\theta'_k)}$$

Lexique universel

- utilisent MUSE de Conneau pour initialiser les représentations dans toutes les langues $\epsilon_u \in \mathbb{R}^{M \times d}$
- dans chaque langue on apprend $\epsilon^k_{\text{query}} \in \mathbb{R}^{|V_k| \times d}$

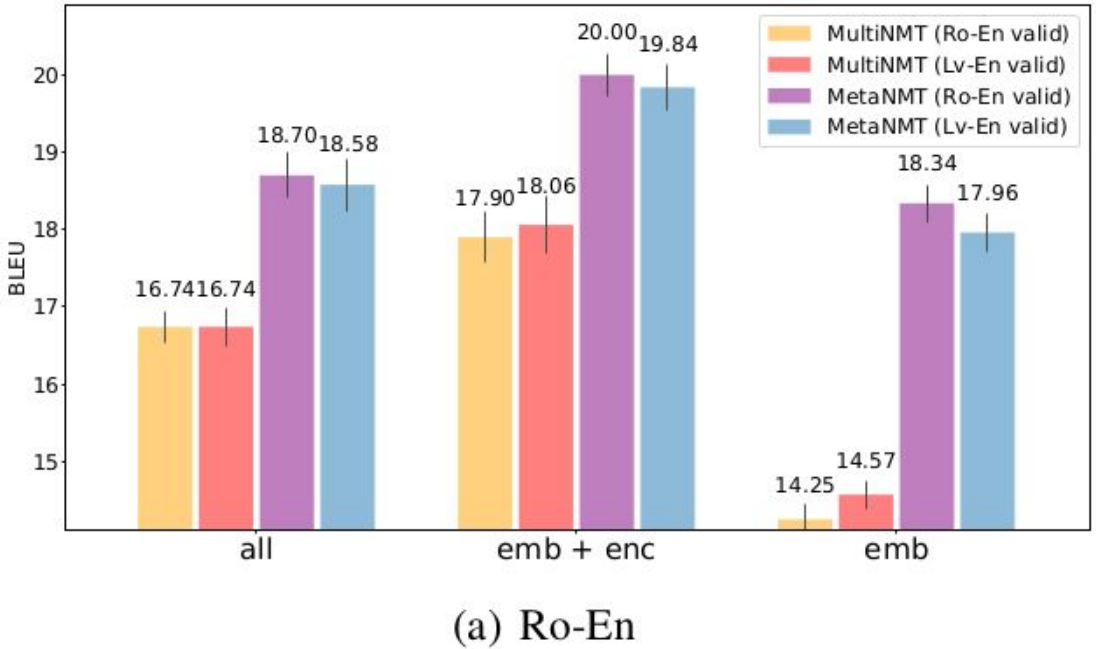
Puis pour chaque mot x l'embedding est donné par :

$$\epsilon^0[x] = \sum_{i=1}^{M} \alpha_i \epsilon_u[i],$$

$$\text{where } \alpha_i \propto \exp\left\{-\tfrac{1}{\tau}\epsilon_{\text{key}}[i]^\top A \epsilon^k_{\text{query}}[x]\right\}$$

Exp

Modèle Transformer + lexique universel

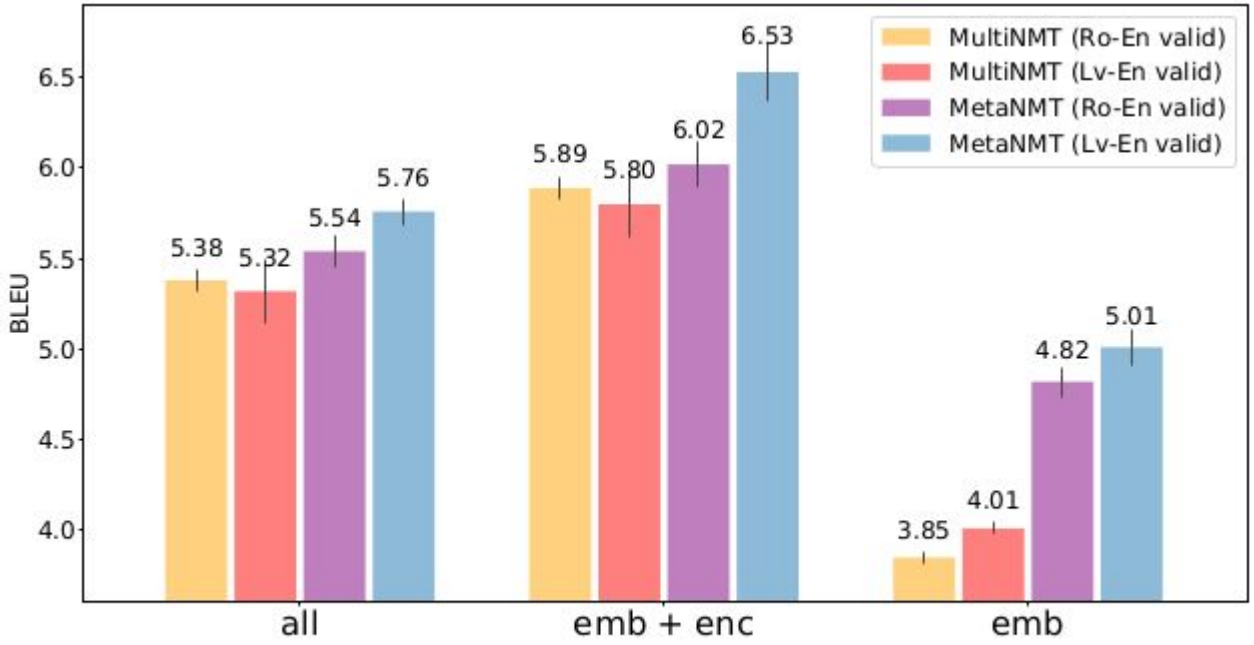


(a) Ro-En

Résultats sur Ro-En

Chaque courbe correspond à un set de validation différent

Exp


MultiNMT (Ro-En valid)
MultiNMT (Lv-En valid)
MetaNMT (Ro-En valid)
MetaNMT (Lv-En valid)
BLEU
6.5
6.0
5.5
5.0
4.5
4.0
5.38
5.32
5.54
5.76
5.89
5.80
6.02
6.53
3.85
4.01
4.82
5.01
all
emb + enc
emb

(d) Tr-En