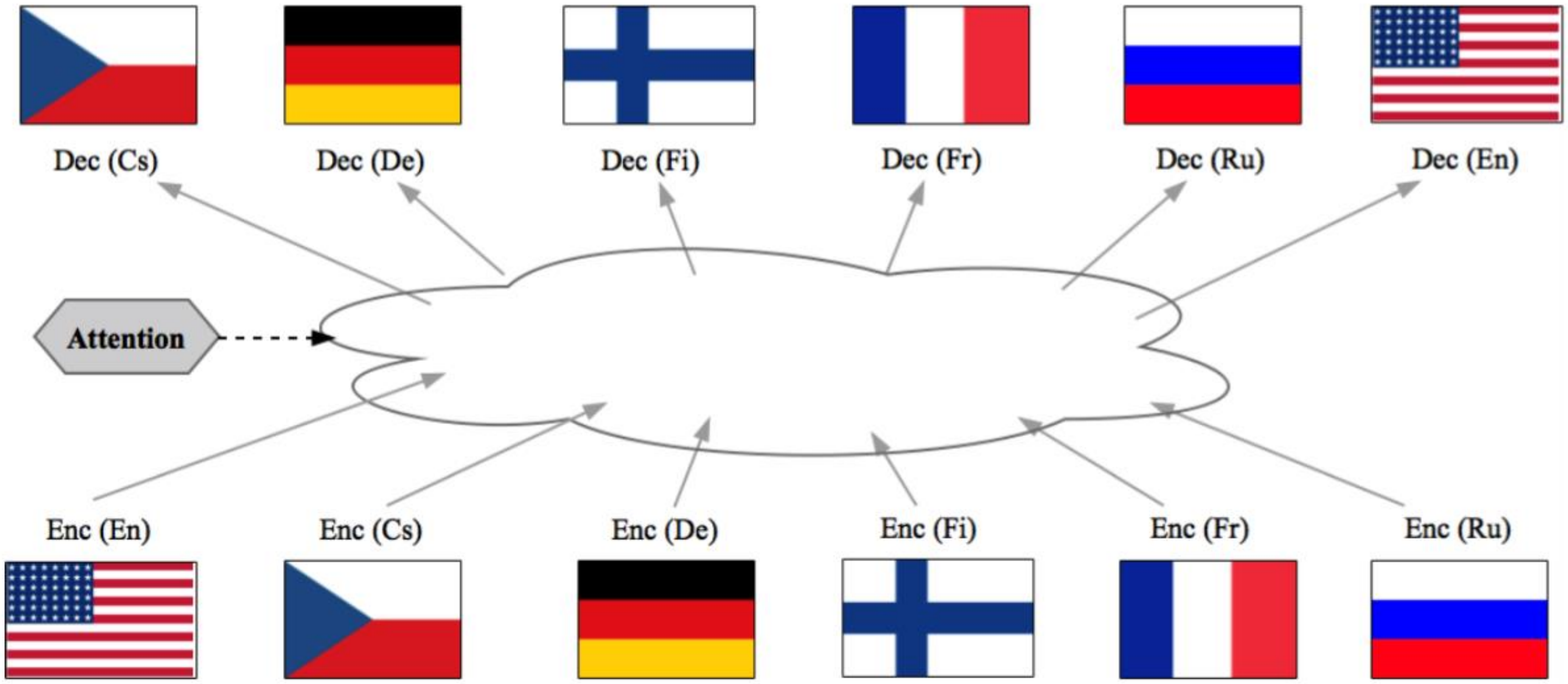
Multilingual Translation – (1)

- Traditionally,
  - If a parallel corpus exists, one system for each language pair.
    - Parallel corpus: $D^{a \to b} = \left\{ (X_1^a, Y_1^b), \ldots, (X_N^a, Y_N^b) \right\}$
    - Translation system: $\log p(Y^b | X^a)$
  - If no direct parallel corpus exists, a pivot-based translation.
    - No direct parallel corpus: $D^{a \to b} = \emptyset$
    - But, $|D^{a \to c}| > 0, |D^{c \to b}| > 0$
    - Then, $\log p(Y^b | \hat{X}^c), \text{where } \hat{X}^c = \arg\max_X \log p(X^c | X^a)$
    - *c* is a pivot language (often, English.)
  - No knowledge transfer between different language pairs.

Multilingual Translation as Multitask Learning – (2)

- Now, [Firat et al., 2016a; Firat et al., 2016b; Johnson et al., 2016; Ha et al., 2016; Lee et al., 2017]



Multilingual Translation as Multitask Learning – (3)

- Separate encoder/decoders
  - [Firat et al., 2016a; Firat et al., 2016b]
- One encoder per source *l*

$$f^l_{\mathrm{enc}} : V_l \times \cdots \times V_l \to \mathbb{R}^d \times \cdots \times \mathbb{R}^d$$

- One decoder per target *l'*

$$\log p^{l'}(Y^{l'}|H)$$

- For each pair (l, l'),

$$\log p^{l'}(Y^{l'}|H = f^l_{\mathrm{enc}}(X^l))$$

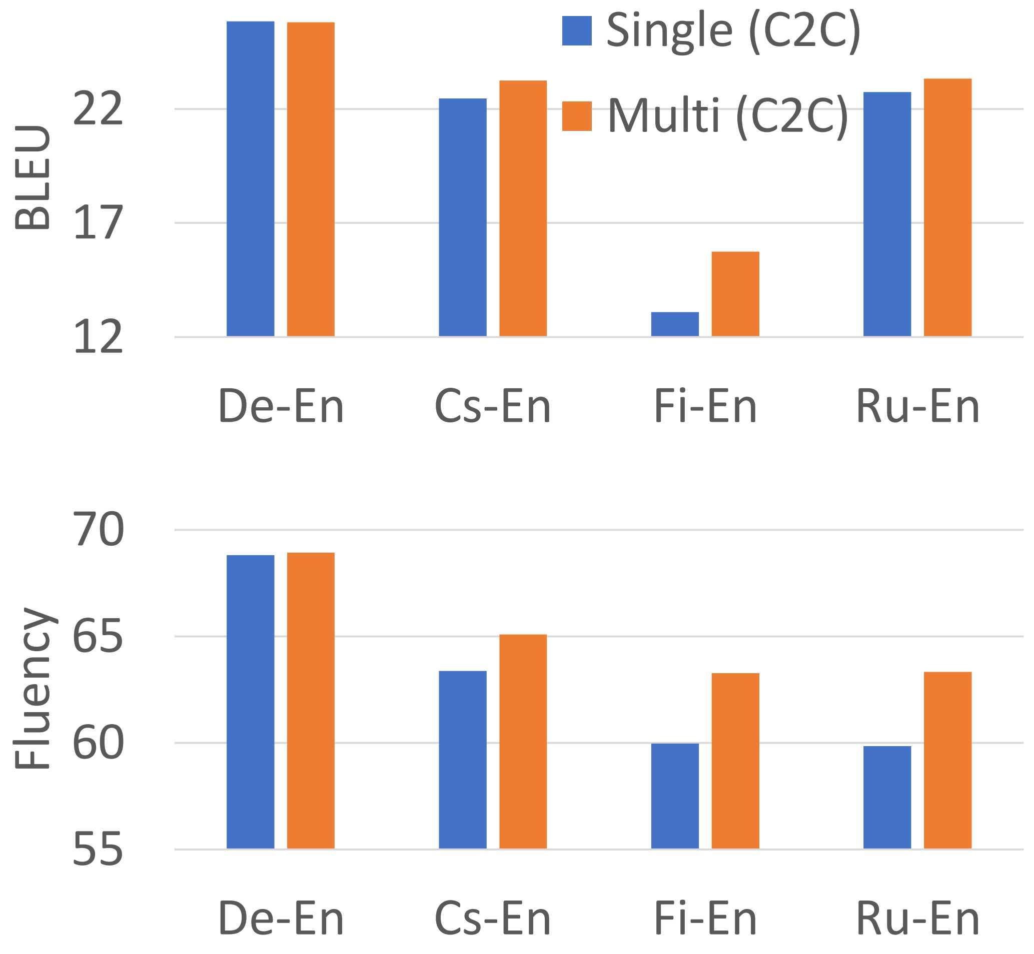
- Train using all available language pairs

- Universal encoder/decoders
  - [Johnson et al., 2016; Ha et al., 2016; Lee et al., 2017; Gu et al., 2018]
- Shared lexicons $f^l_{\mathrm{lex}} : V_l \to V$
  - A shared vocabulary of language-agnostic tokens [J, 2016; H, 2016; L, 2017]
  - Universal lexical representation [G, 2018]
- One encoder-decoder for all pairs

$$f^{-l'}_{\mathrm{lex}}(\arg\max_Y \log p(Y|f^l_{\mathrm{lex}}(X^l)))$$

Multilingual Translation as Multitask Learning – (4)

- Does it work?
- Single-pair Systems
  De→En, Cs→En, Fi→En, Ru→En
- Multilingual System
  {De, Cs, Fi, Ru}→En
- The latter has 1/4x parameters
- Better translation quality on low-resource languages (Fi & Ru)



Lee, Cho and Hoffman (2017)

Multilingual Translation as Multitask Learning – (5)

- Does it work? – **Yes!***
- Single-pair Systems vs. Multilingual System
- Works with intra-sentence code-switching

(e) Multilingual

Multi src	Bei der Metropolitního výboru pro dopravu für das Gebiet der San Francisco Bay erklärten Beamte , der Kongress könne das Problem банкротство доверительного Фонда строительства шоссейных дорог einfach durch Erhöhung der Kraftstoffsteuer lösen .
EN ref	At the Metropolitan Transportation Commission in the San Francisco Bay Area , officials say Congress could very simply deal with the bankrupt Highway Trust Fund by raising gas taxes .
bpe2char	During the Metropolitan Committee on Transport for San Francisco Bay , officials declared that Congress could solve the problem of bankruptcy by increasing the fuel tax bankrupt .
char2char	At the Metropolitan Committee on Transport for the territory of San Francisco Bay , officials explained that the Congress could simply solve the problem of the bankruptcy of the Road Construction Fund by increasing the fuel tax .

** It often fails to translate between a pair of languages not seen during training*

Lee, Cho and Hoffman (2017)

Limitations of Multitask Learning – (1)

- Tricky when the availability of data drastically differs across languages.
  - *overfitting* on low-resource pairs, while *underfitting* on high-resource pairs.

$$L(\theta) = \sum_l \frac{1}{N^l} \sum_{n=1}^{N^l} \log p_\theta(Y_n^l | X_n^l)$$

  - Extremely low-resource pairs can easily be *ignored*.

$$L(\theta) = \sum_l \sum_{n=1}^{N^l} \log p_\theta(Y_n^l | X_n^l)$$

  - See [Firat et al., 2016a] and [Lee et al., 2017] for more discussion.
  - *It is really horrible to figure out how to tackle this in practice…*

Limitations of Multitask Learning – (2)

- Assumes the availability of all language pairs in advance.
  - The entire model must be re-trained each time a new language is introduced.
- Transfer Learning [Zoph et al., 2016; Nguyen & Chiang, 2017]
  - Only re-train a subset of parameters on a new language pair.
  - Many possible strategies, but no clear winning strategy.

Setting	Dev BLEU	Dev PPL
No retraining	0.0	112.6
Retrain source embeddings	7.7	24.7
+ source RNN	11.8	17.0
+ target RNN	14.2	14.5
+ target attention	**15.0**	13.9
+ target input embeddings	14.7	**13.8**
+ target output embeddings	13.7	14.4

Zoph et al., 2016

Limitation of Multitask Learning – (3)

- Inconvenient truths about multitask+transfer learning
  - Relies on our intuition that all languages/tasks share common underlying structures: *true?*
  - Assumes multitask learning can capture those underlying structures and share across multiple languages/tasks: *true?*
  - Assumes multitask-learned parameters are a good initialization for further training: *true?*

- Is there a more satisfying approach?

Meta-Learning: MAML [Finn et al., 2018] – (1)

- Model-agnostic meta-learning [Finn et al., 2018]
- Two-stage learning
  1. Simulated learning

$$\begin{aligned}\mathrm{Learn}(D_{\mathcal{T}};\theta^0) &= \arg\max_{\theta} \mathcal{L}^{D_{\mathcal{T}}}(\theta) \\ &= \arg\max_{\theta} \sum_{(X,Y)\in D_{\mathcal{T}}} \log p(Y|X,\theta) - \beta\|\theta-\theta^0\|^2,\end{aligned}$$

  2. Meta-learning

$$\mathcal{L}(\theta) = \mathbb{E}_k \mathbb{E}_{D_{\mathcal{T}^k}, D'_{\mathcal{T}^k}} \left[\sum_{(X,Y)\in D'_{\mathcal{T}^k}} \log p(Y|X;\mathrm{Learn}(D_{\mathcal{T}^k};\theta)) \right],$$

Meta-Learning: MAML [Finn et al., 2018] – (2)

1. Simulated learning
   - Given a small subset $D_{\mathcal{T}}$ of the training set of task $\mathcal{T}$, update the model parameters $N = 1$ times.

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

$$= \arg\max_{\theta} \sum_{(X,Y) \in D_{\mathcal{T}}} \log p(Y|X, \theta) - \beta \|\theta - \theta^0\|^2,$$

$$= \theta_0 - \eta \nabla_\theta \mathcal{L}^{D_{\mathcal{T}^k}}(\theta_0)$$

   - Clip the update so that $\eta \nabla_\theta \mathcal{L}^{D_{\mathcal{T}^k}}(\theta_0)$ does not deviate too much from θ_0.
   - *It simulates finetuning on a target task with a limited resource.*

Meta-Learning: MAML [Finn et al., 2018] – (3)

2. Meta-Learning
   - Randomly select a task k and select a training subset $D = D_{\mathcal{T}^k}$.
   - Randomly select a validation subset $D' = D'_{\mathcal{T}^k}$ for evaluation.
   - Update the meta-parameter θ_0 by gradient descent:

     $$\theta_0 \leftarrow \theta_0 + \eta_0 \nabla_\theta \mathcal{L}^{D'}(\theta')$$

     where

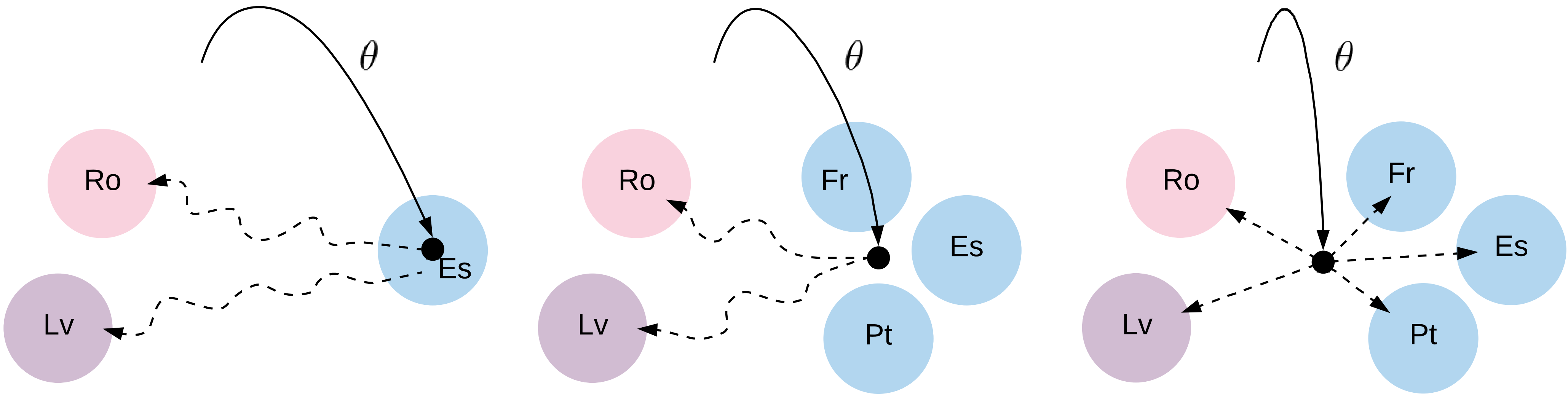
     $$\begin{aligned} \nabla_\theta \mathcal{L}^{D'}(\theta') &= \nabla_{\theta'} \mathcal{L}^{D'}(\theta') \nabla_\theta (\theta - \eta \nabla_\theta \mathcal{L}^D(\theta)) \\ &= \nabla_{\theta'} \mathcal{L}^{D'}(\theta') - \eta \nabla_{\theta'} \mathcal{L}^{D'}(\theta') H_\theta(\mathcal{L}^D(\theta)) \end{aligned}$$

   - *Update the meta-parameter so that N-step GD on the k-th task works well.*

Meta-Learning: MAML [Finn et al., 2018] – (4)

3. Fast adaptation to a new task
   - Given a small training set D of the new target task, SGD starting from the meta-parameter θ_0 .
   - Early stopping based on $\|\theta - \theta_0\|^2$.

Multitask learning vs. Meta-learning



(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

a) Transfer learning does not take into account subsequent learning.
b) Multilingual learning does not take into account new, future tasks.
c) Meta-learning considers subsequent learning on new, future tasks.

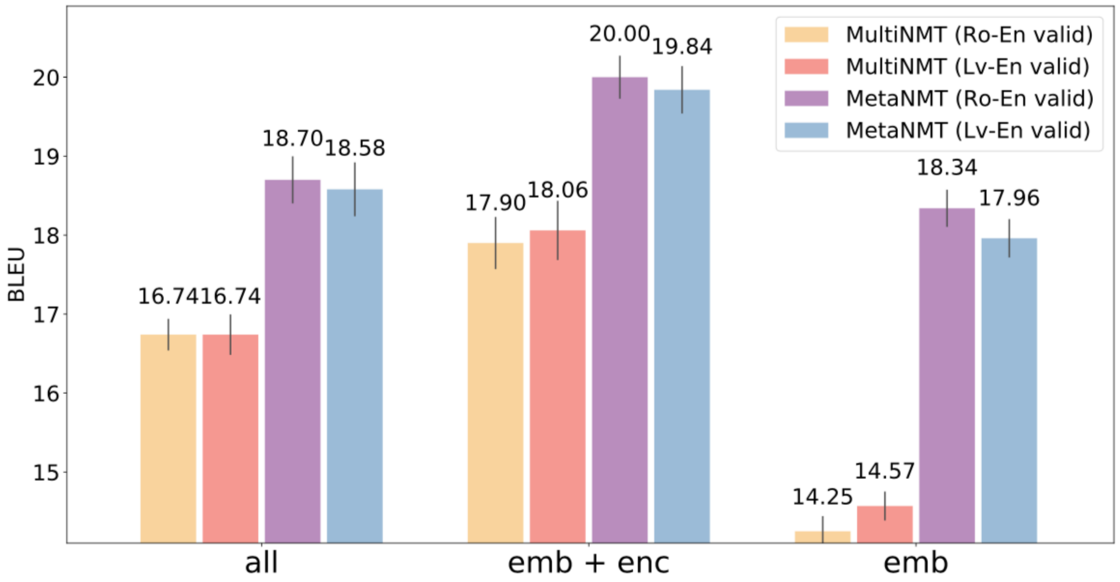
Extension to Neural Machine Translation

- I/O mismatch between different tasks
  - Vocabulary mismatch among different languages
- Multilingual word embedding [Artetxe et al., 2017; Conneau et al., 2018; and more]
  - Project each token into a continuous vector space $f^l : V^l \to \mathbb{R}^d$
  - Ensure that they are compatible:
    $\|f^l(v^l) - f^{l'}(v^{l'})\|^2 < \epsilon, \text{iff } v^l \text{ and } v^{l'} \text{ have the same meaning.}$
- Universal lexical representation [Gu et al., 2018]

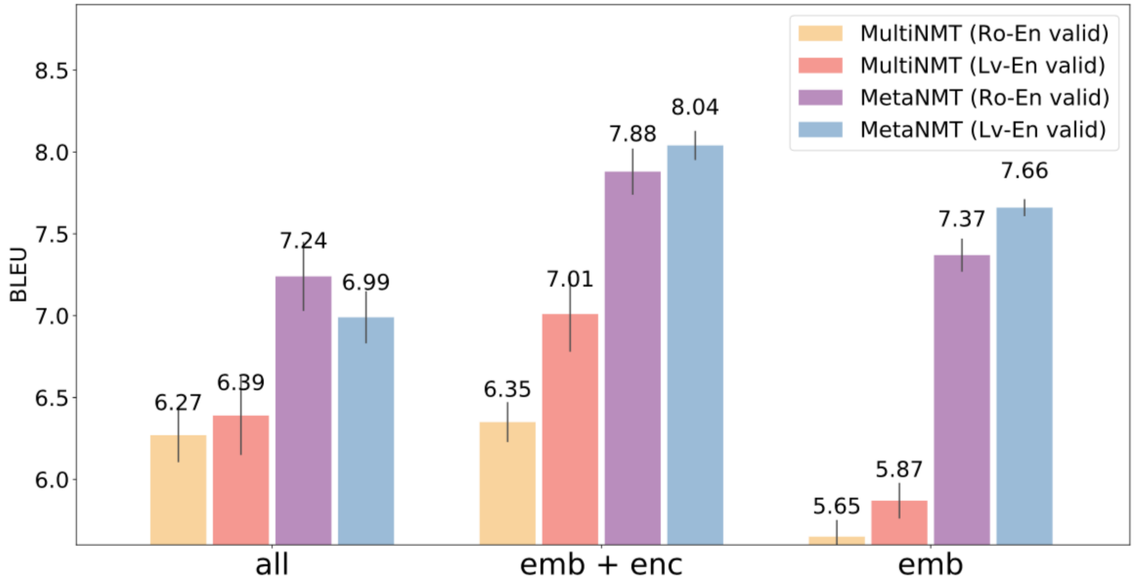
- *Meta-NMT!*

Experiments

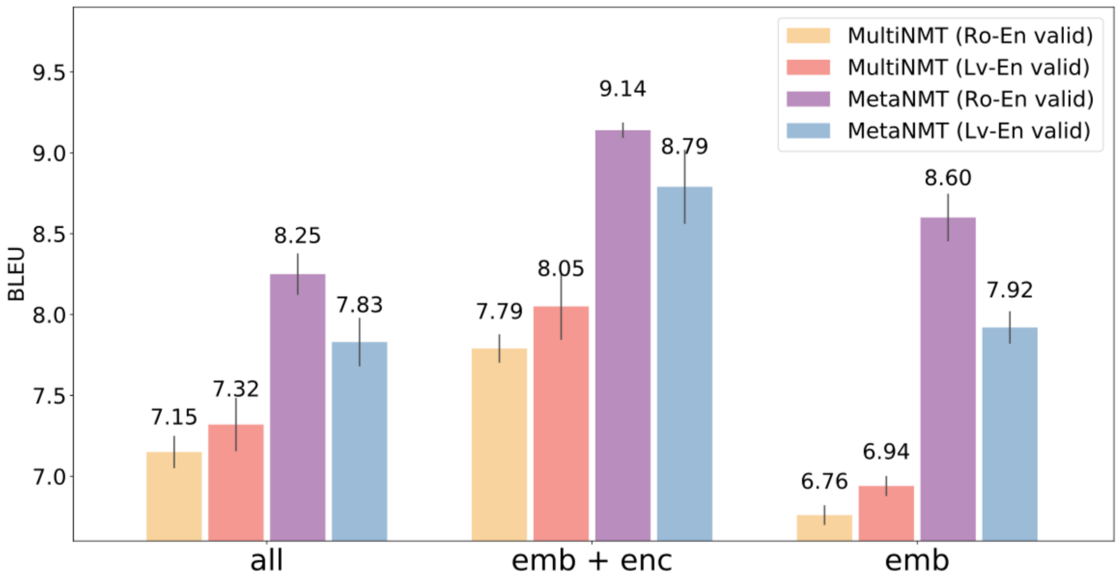
- Source tasks: all the languages from Europarl + Russian
  - Bg→En, Cs→En, Da→En, De→En, El→En, Es→En, Et→En, Fr→En, Hu→En, It→En, Lt→En, Nl→En, Pl→En, Pt→En, Sk→En, Sl→En, Sv→En and Ru→En.
  - Reasonable high-resource language pairs.
- Target tasks: (simulated) low-resource language pairs
  - Ro→En, Lv→En, Fi→En, Tr→En and Ko→En
  - Approximately 16k target tokens (English side): roughly 800 sentence pairs.
- Universal lexical representation: obtained from Wikipedia.
- Early stopping of meta-learning: either Ro-En or Lv-En



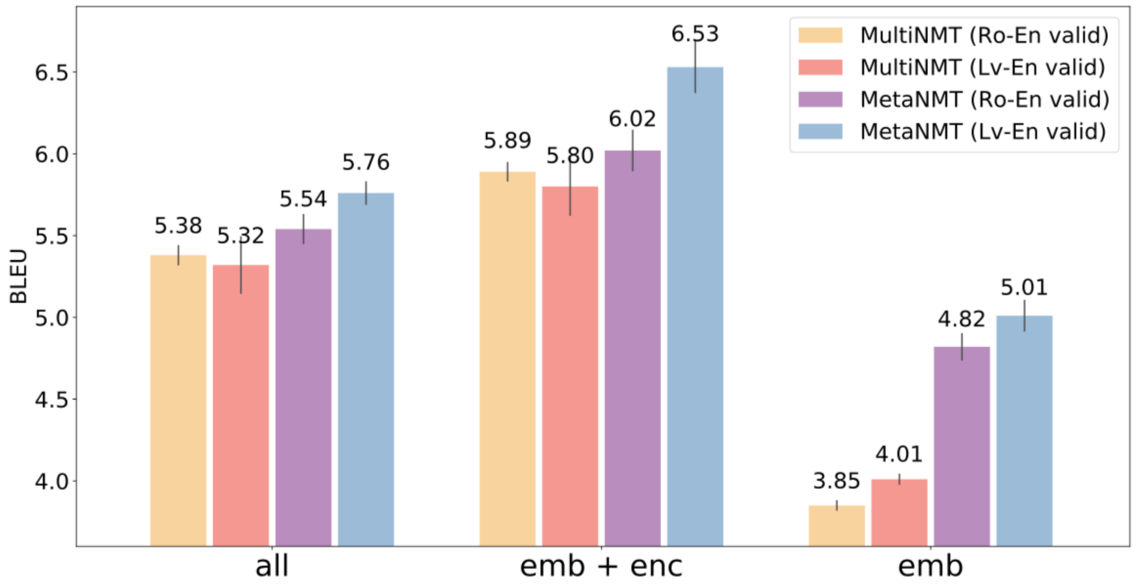
(a) Ro-En



(b) Lv-En



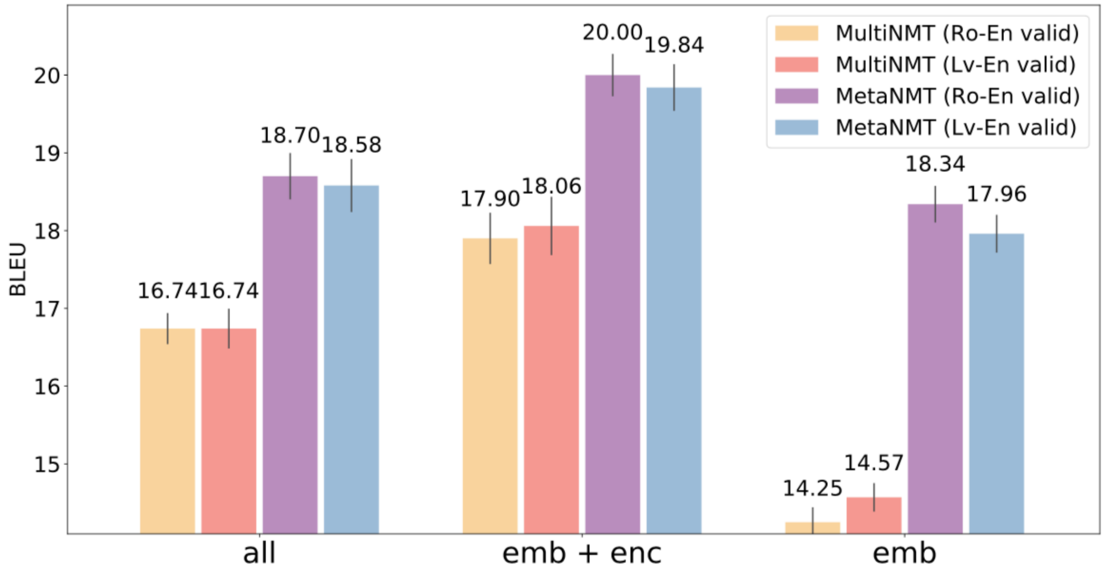
(c) Fi-En



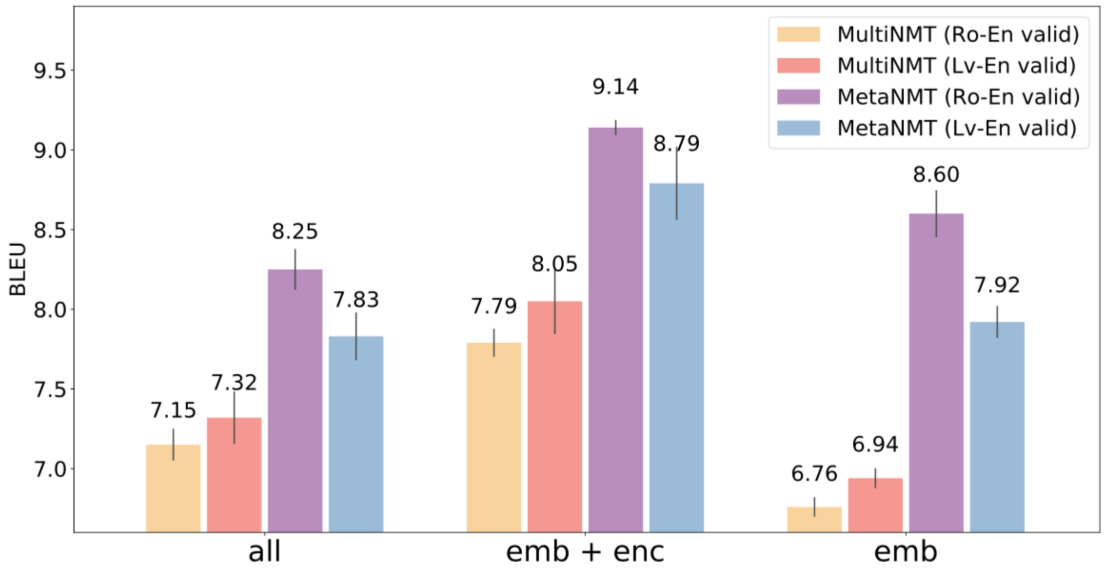
(d) Tr-En

Experiments – (1)

- Meta-learning outperforms multitask learning across all the target languages and across different finetuning strategies.
- Using only 800 examples, reaches up to 65% of fully-supervised models in terms of BLEU.



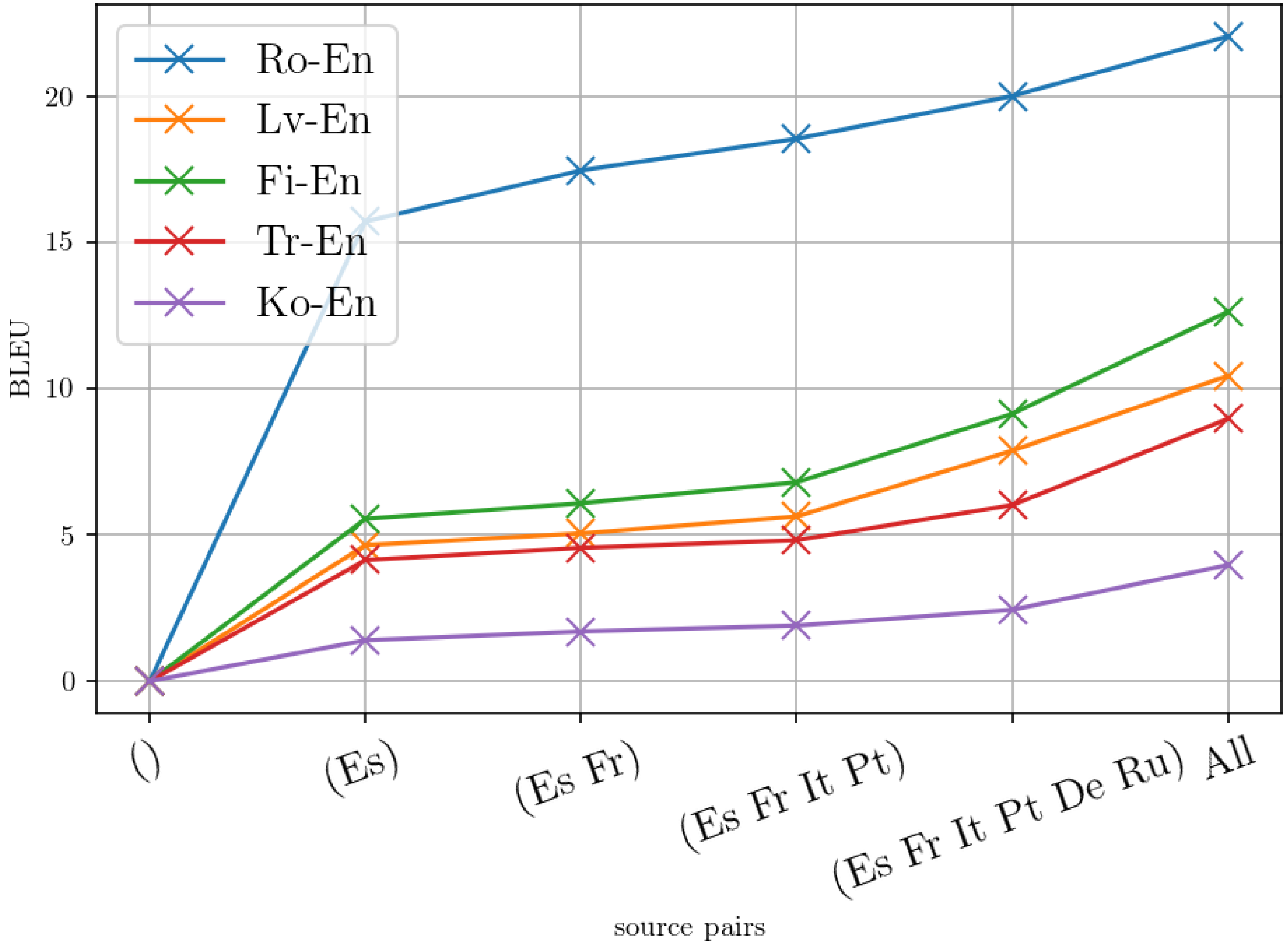
(a) Ro-En



(c) Fi-En

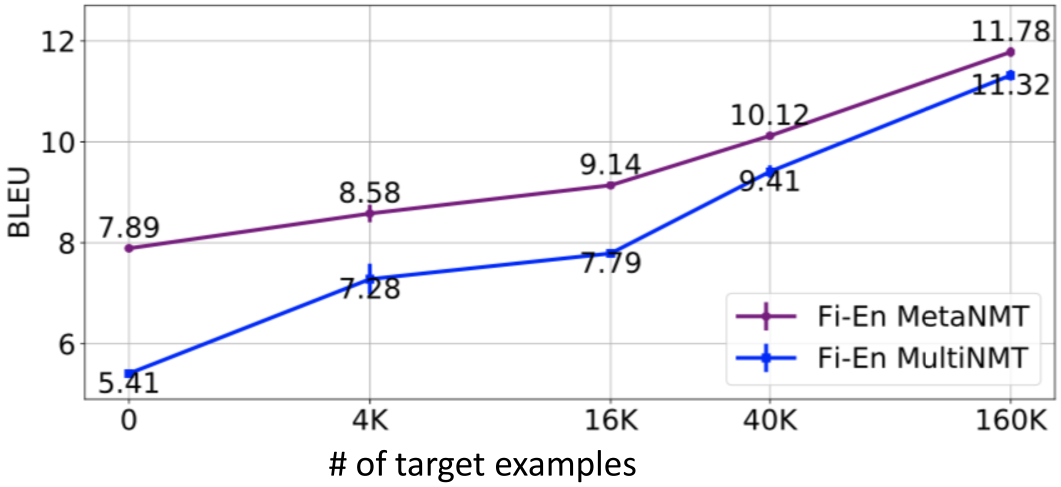
Experiments – (2)

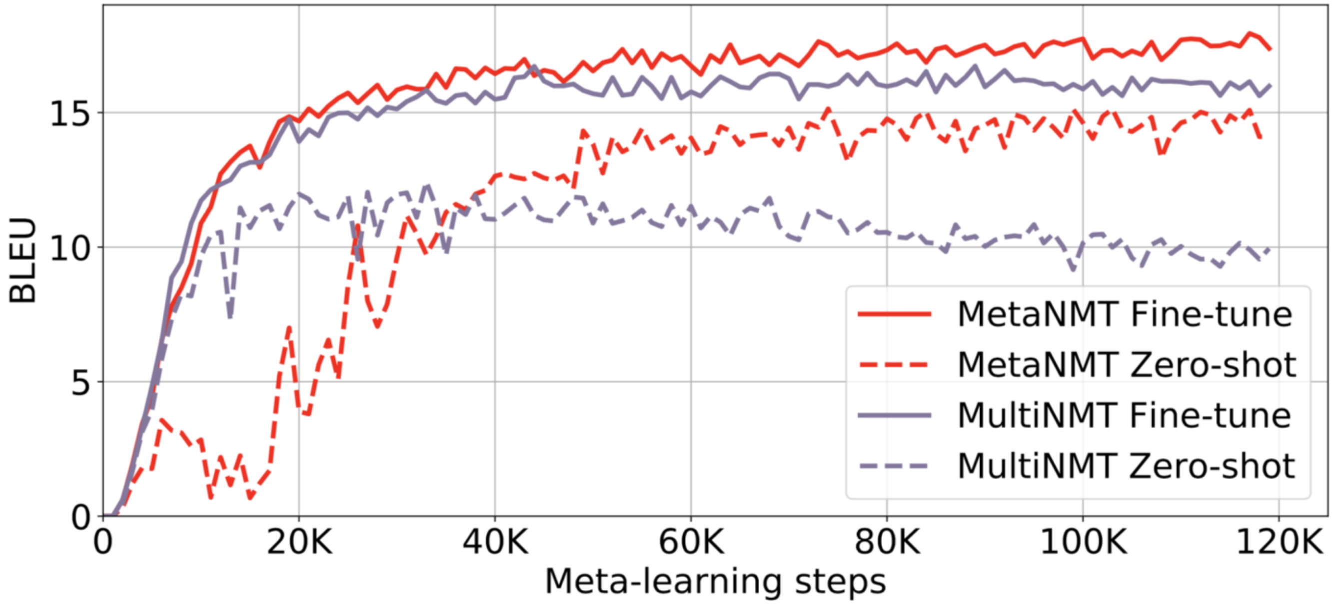
- More source tasks lead to greater improvements.
- The similarity between source and target asks matters.



Experiments – (3)

- Multi-task learning over-adapts to the source tasks.
  - Performance on the target task degrades with longer multi-task learning.
- Meta-learning does not over-adapt.
  - The meta-learning objective explicitly takes into account finetuning on a target task.
  - It requires less target examples.





Experiments – (4) Sample Translations

Source (Tr)	google mülteciler için 11 milyon dolar toplamak üzere bağış eşleştirme kampanyasını başlattı .
Target	google launches donation-matching campaign to raise $ 11 million for refugees .
Meta-0	google refugee fund for usd 11 million has launched a campaign for donation .
Meta-16k	google has launched a campaign to collect $ 11 million for refugees .
Source (Ko)	이번에 체포되어 기소된 사람들 중에는 퇴역한 군 고위관리 , 언론인 , 정치인 , 경제인 등이 포함됐다
Target	among the suspects are retired military officials , journalists , politicians , businessmen and others .
Meta-0	last year , convicted people , among other people , of a high-ranking army of journalists in economic and economic policies , were included .
Meta-16k	the arrested persons were included in the charge , including the military officials , journalists , politicians and economists .

Conclusion

- Meta-learning allows us to exploit many high-resource tasks for *extremely low-resource* target tasks.
- Gradual shift toward higher-order learning
  - Learning to optimize [Andrychowicz et al., 2017; and others]
  - Multi-agent modelling (theory of mind) [Foerster et al., 2018 LOLA; and others]
  - Neural architecture search [Zoph & Le, 2016; and others]
  - Hyperparameter search [Luketina et al., 2016; and others]
  - And more on the horizon…