Multi-task Learning for Multiple Languages Translation

Daxiang Dong Hua Wu Wei He Dianhai Yu Haifeng Wang

Baidu Inc.

Background

Consider the problem of translating one source language into multiple target languages.

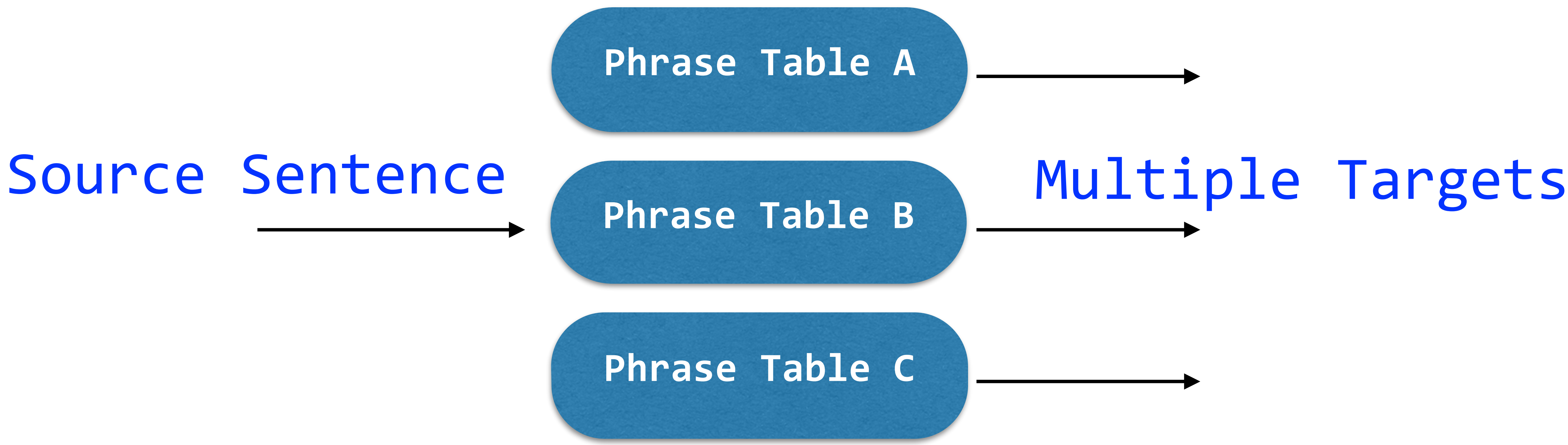
- Practical Usages :
  - Web pages translation
  - Product introduction for global scale users

  …

- Modern machine translation system solution:
  - Build up translation service in pairwise manner
  - Translation quality may not be acceptable in some directions when the size of training corpora is small

Statistical Machine Translation

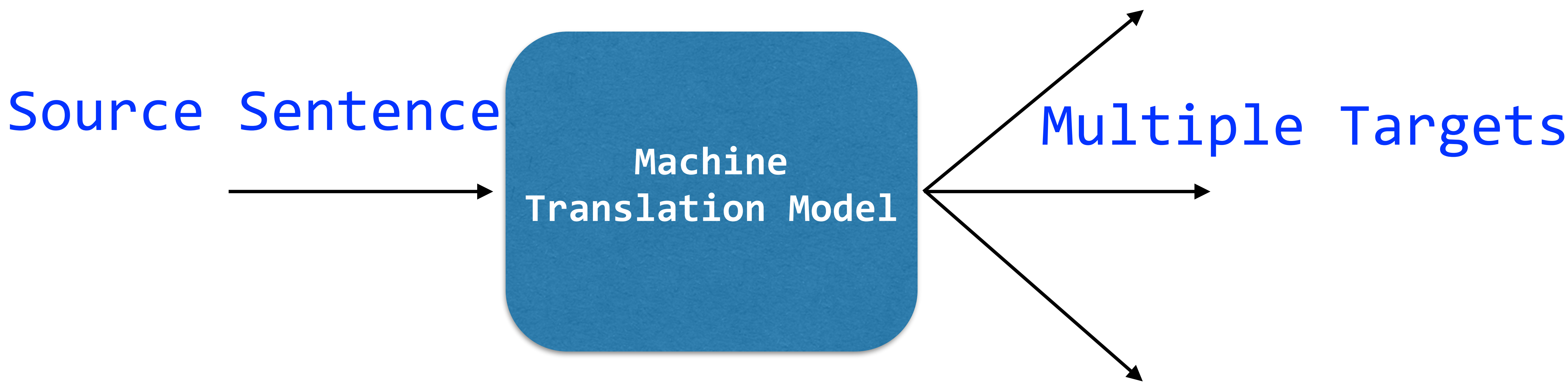
Frequently used in commercialized system



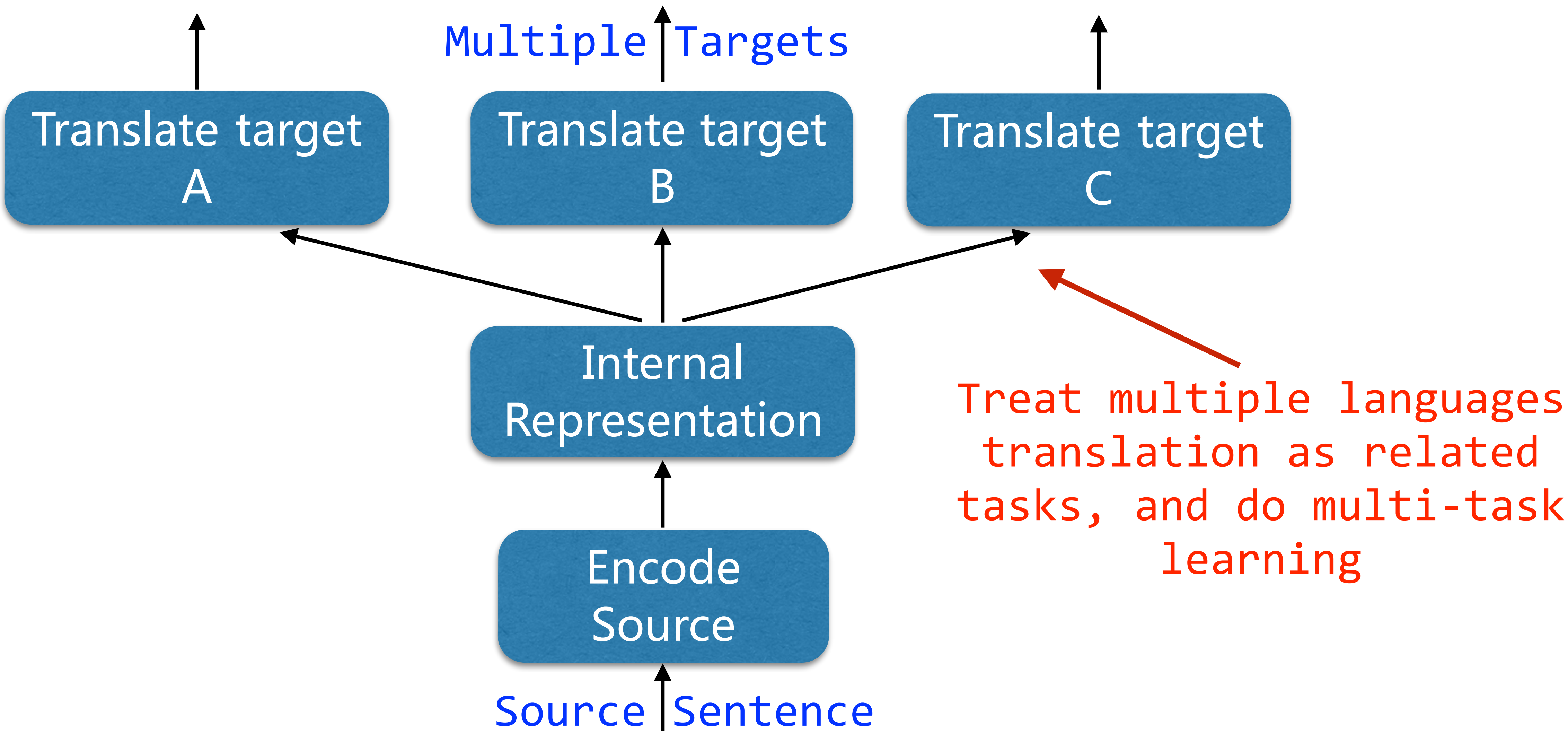
- Phrased-based MT generates multiple phrase tables
- Data sparsity problem is severe in resource-poor parallel corpora
- It is hard for phrase tables to share corpora information

Motivation

How can we share data information of multiple parallel corpora and translate a source sentence into multiple targets within a unified model?



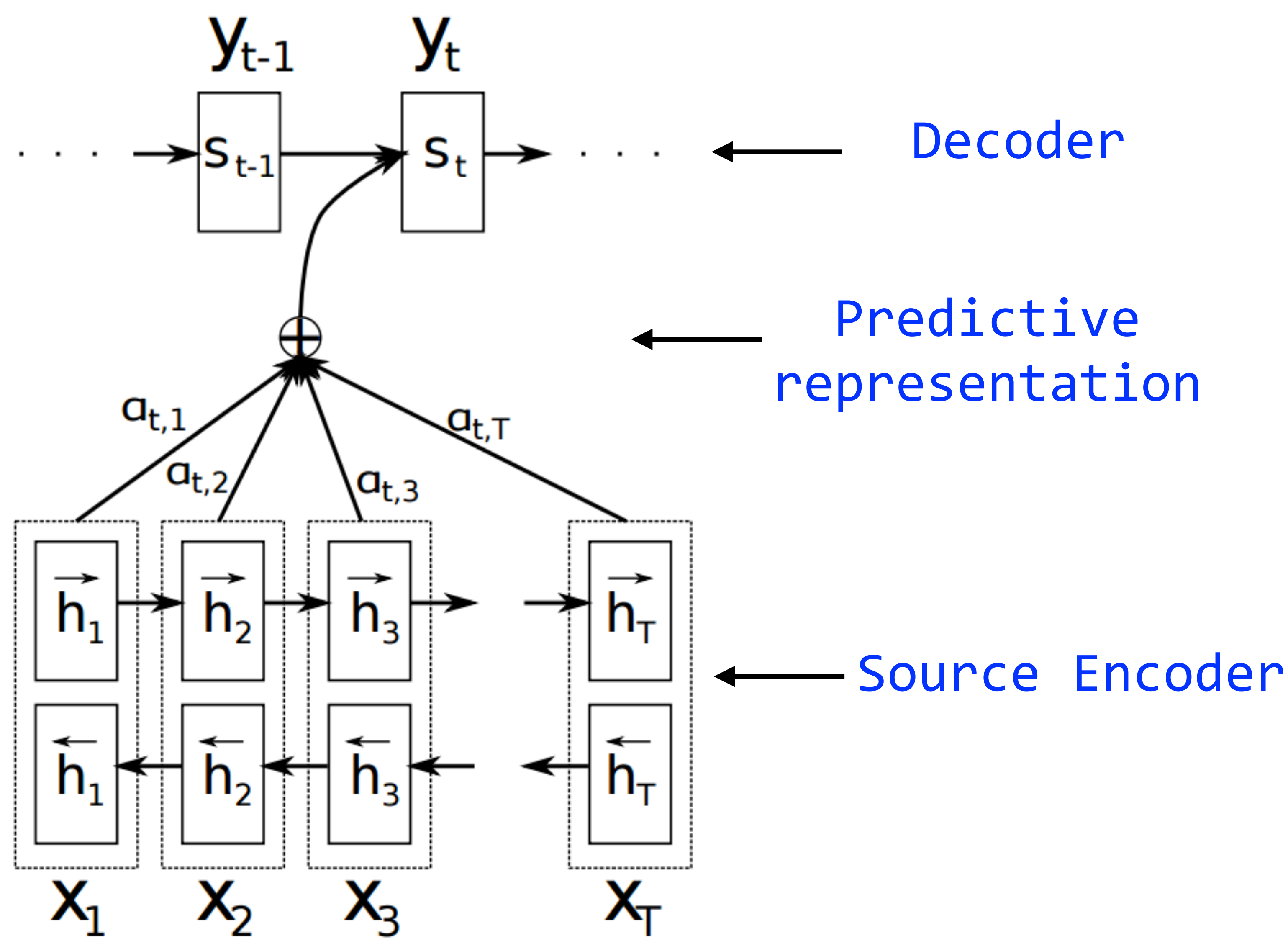
Our Solution



- Share source language information within a shared encoder.
- Do multi-task learning with multiple parallel corpora in a unified model

Neural Machine Translation

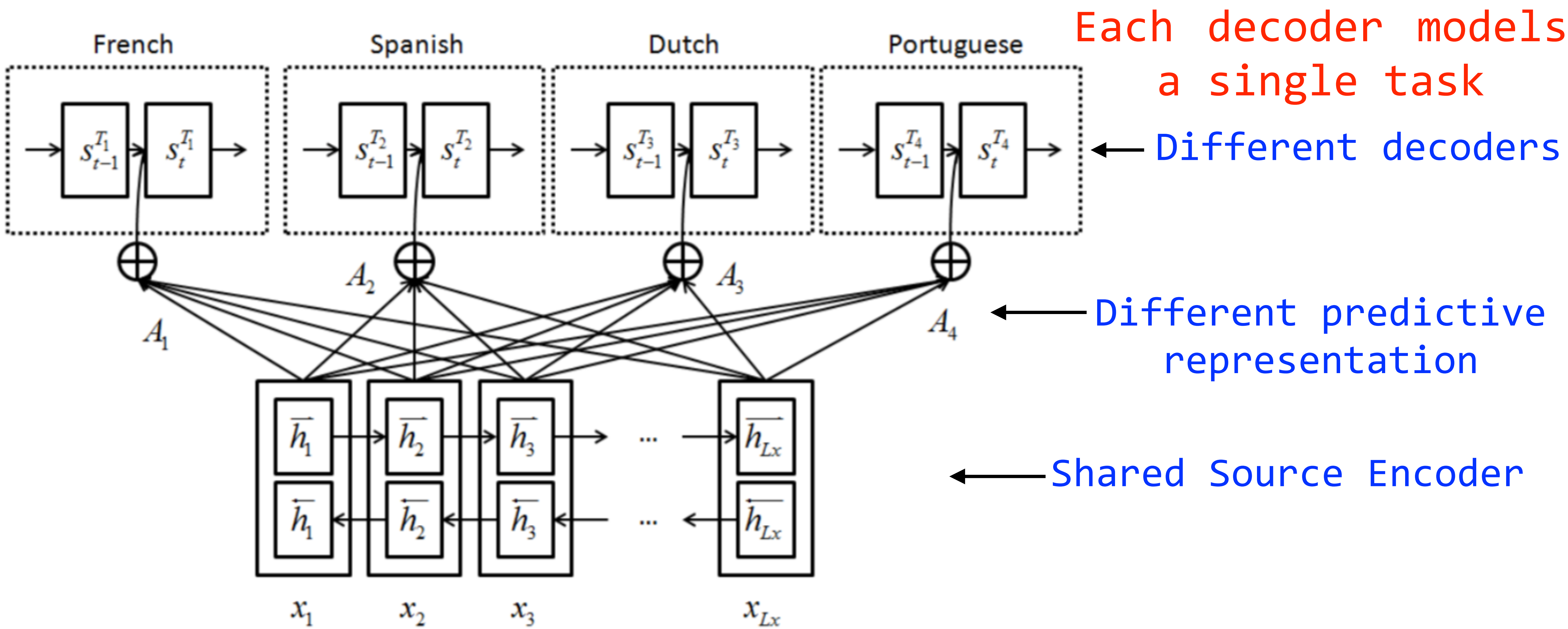
Base Model: NMT



- Source sentences and target sentences are modeled with encoder and decoder, each of which is a gated recurrent neural network.
- Soft alignment model is applied between encoder and decoder.

* figure is cited from: NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Multi-task Learning Framework



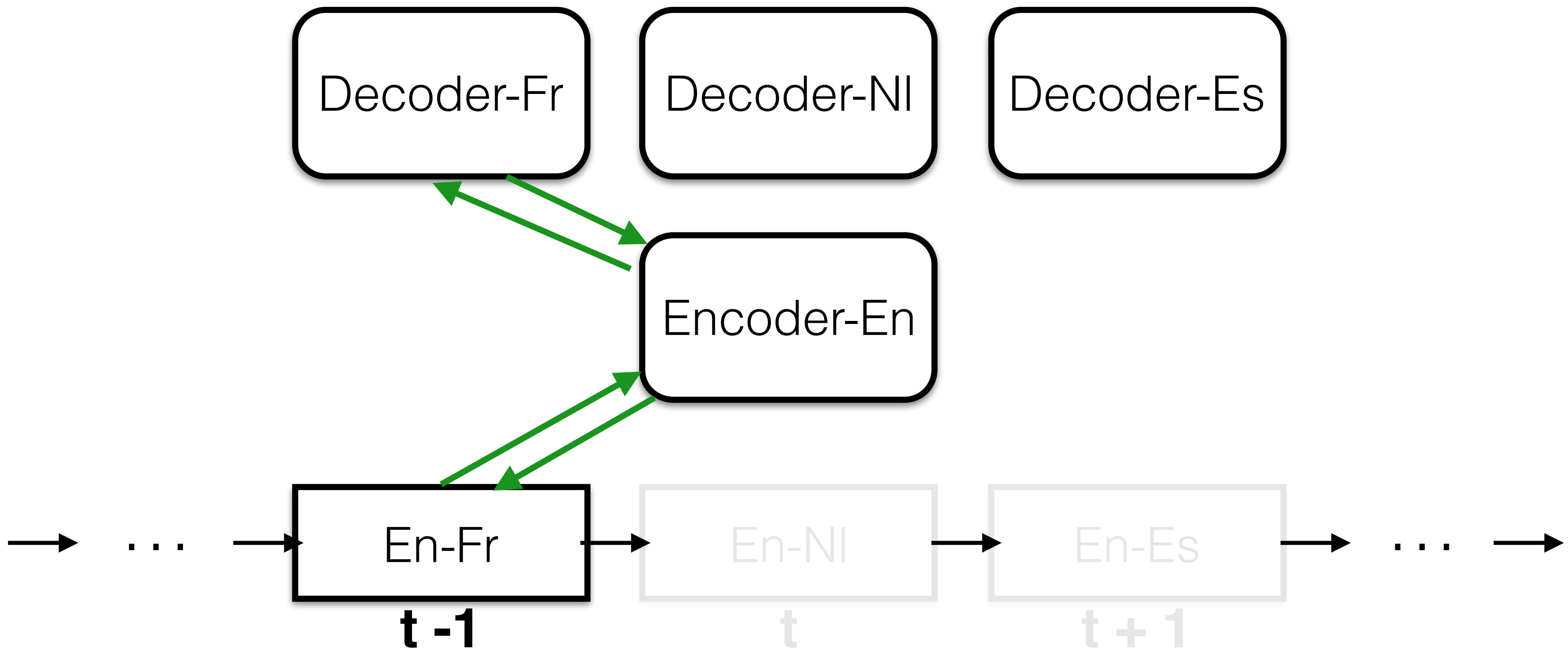
- Share encoder across different language pairs
- Decoders and soft-alignment models are separated on different target languages

Training Objective

$$L(\Theta) = \operatorname*{argmax}_{\Theta}\left(\sum_{T_p}\left(\frac{1}{N_p}\sum_{i}^{N_p}\log p(\mathbf{y_i}^{T_p}|\mathbf{x_i}^{T_p};\Theta)\right)\right)$$

- Maximize the summation of log-likelihood of all language pairs
- Log-likelihood of each parallel sentence is the log of conditional probability of sequence y_i given sequence x_i
- T_p is the language pair index, and N_p is the size of parallel corpora. Θ denotes all model parameters we want to learn

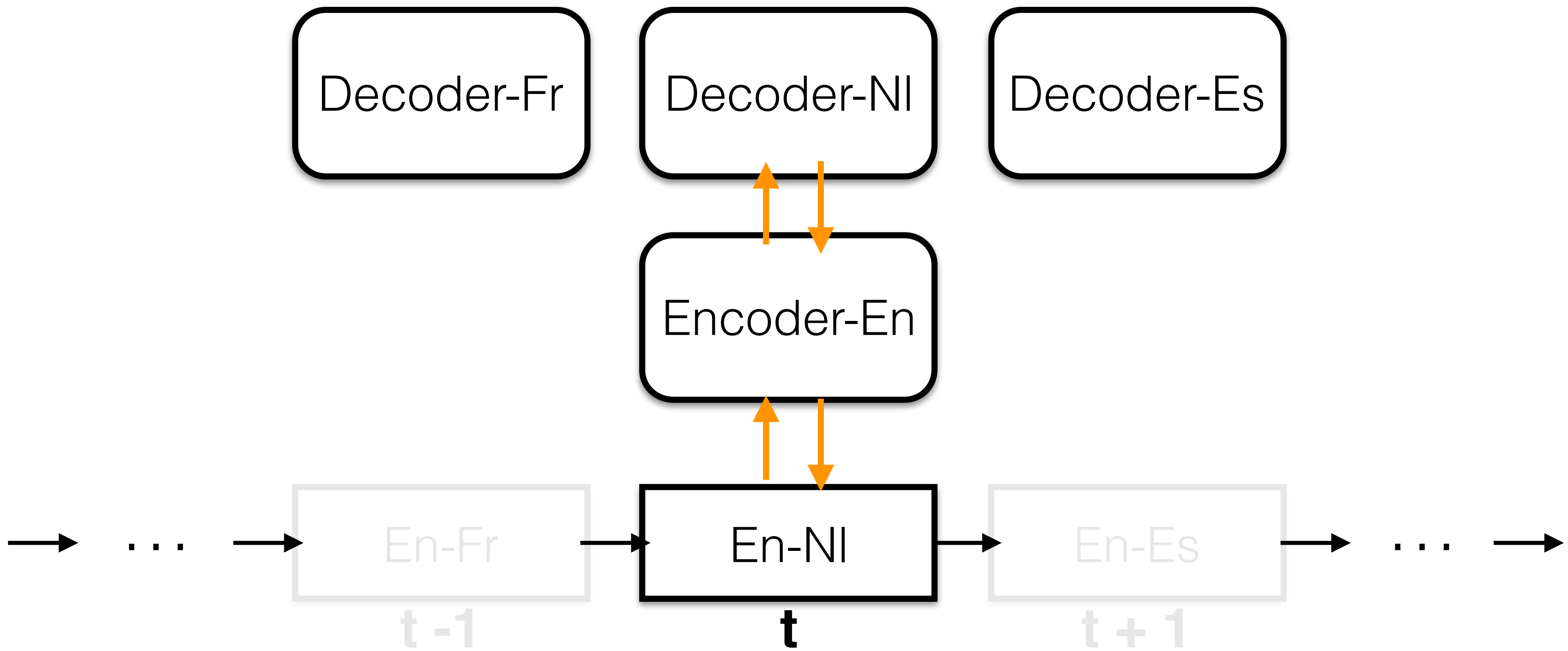
Optimization



Several mini-batches between language pairs

- Learning with mini-batch stochastic gradient descent
- Synchronize encoder parameters every several mini batches
- Train several mini batches between language pairs for speedup

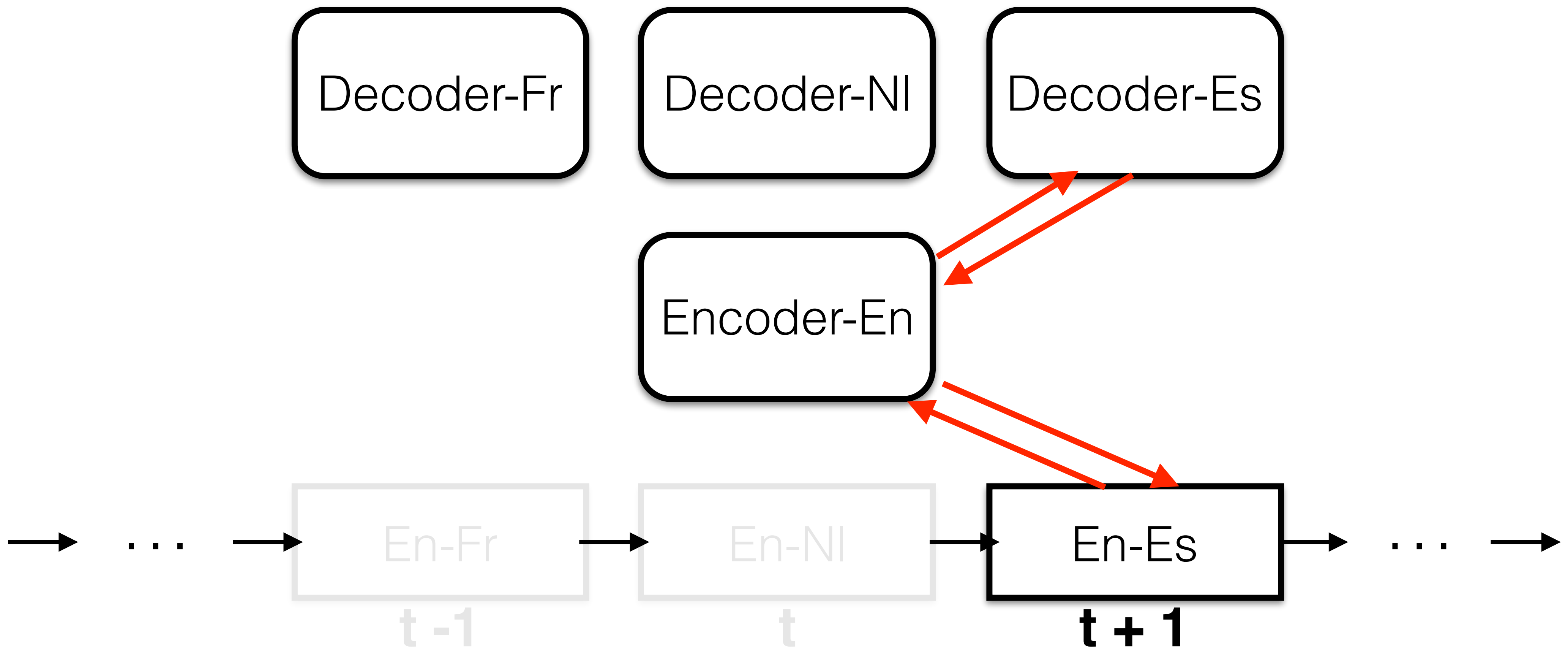
Optimization



Several mini-batches between language pairs

- Learning with mini-batch stochastic gradient descent
- Synchronize encoder parameters every several mini batches
- Train several mini batches between language pairs for speedup

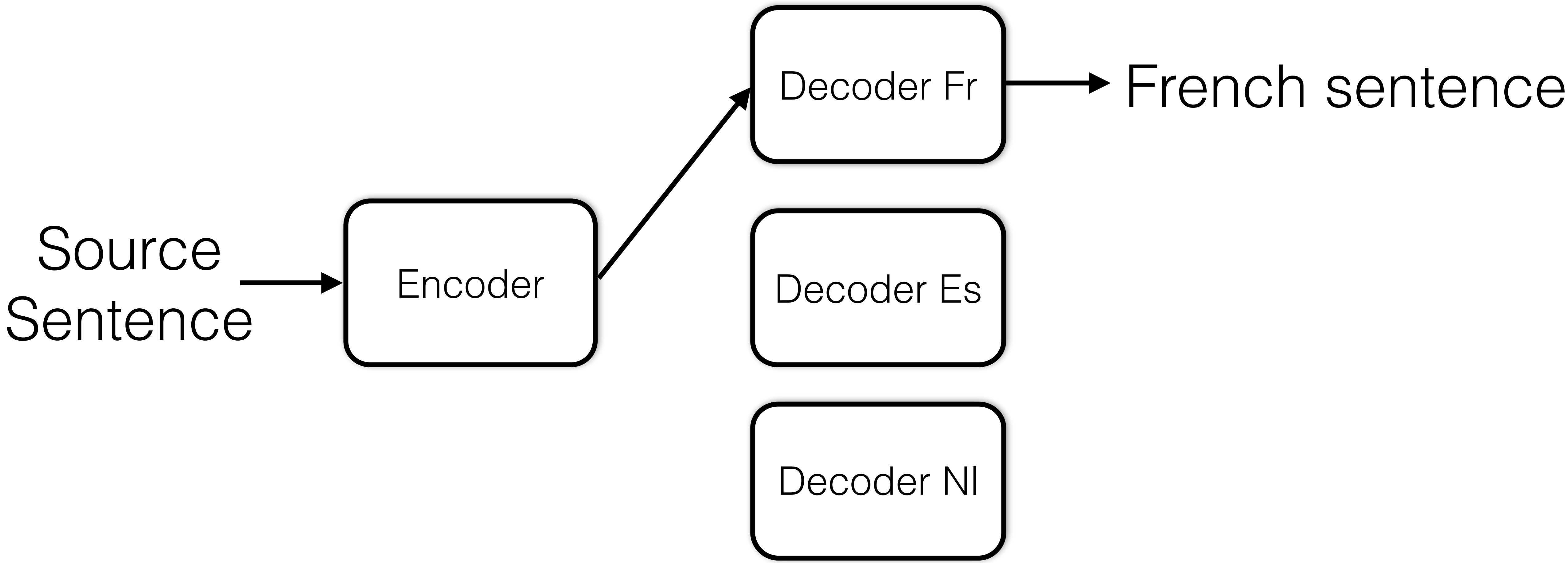
Optimization



Several mini-batches between language pairs

- Learning with mini-batch stochastic gradient descent
- Synchronize encoder parameters every several mini batches
- Train several mini batches between language pairs for speedup

Translation



`Beam Search with beam size = 10`

Experiments

Validate our framework with two experiments

- Resource-Poor setting: Multi-task learning NMT helps to alleviate data sparsity problem of resource-poor language pairs.
- Resource-Rich setting: Multi-task learning NMT also improves translation performance of resource-rich language pairs.

Model analysis

- Comparison with Moses
- Qualitative analysis of results on why multitask learning works for machine translation

Datasets

Training Data: Europarl dataset

Lang	En-Es	En-Fr	En-Nl	En-Pt	En-Nl-sub	En-Pt-sub
Sent Size	1,965,734	2,007,723	1,997,775	1,960,407	300,000	300,000
Src Tokens	49,158,635	50,263,003	49,533,217	49,283,373	8,362,323	8,260,690
Trg Tokens	51,622,215	52,525,000	50,611,711	54,996,139	8,590,245	8,334,454

- Notes: En-Nl-sub and En-Pt-sub are sub-sampled to about 15% of full parallel corpus

Testing data: Europarl Common Test Set, WMT 2013

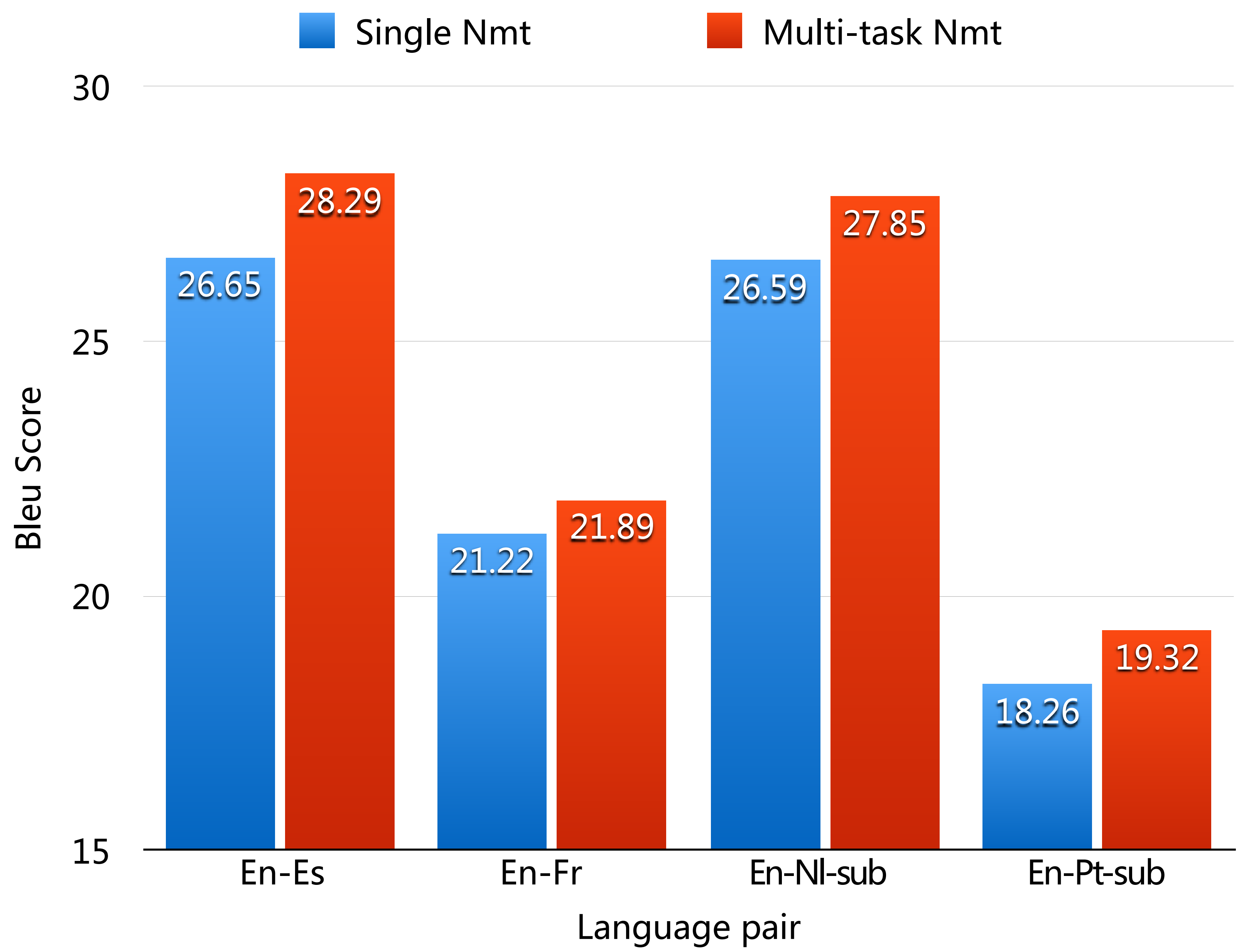
Language Pair	En-Es	En-Fr	En-Nl	En-Pt
Common Test	1755	1755	1755	1755
WMT 2013	3000	3000	-	-

- Notes: En-Nl and En-Pt test sets are not available in WMT dataset

Preprocessing

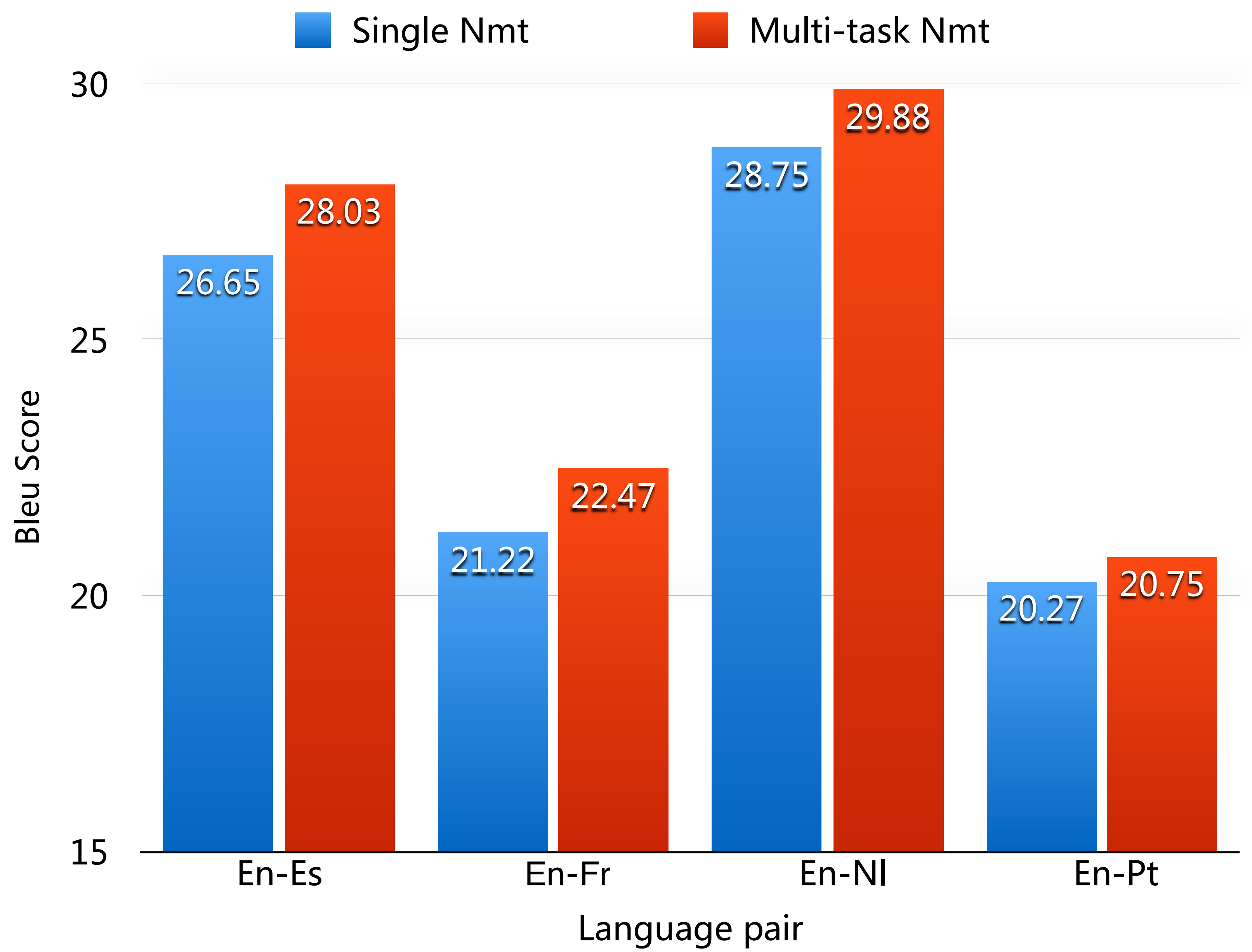
- 30k words vocabulary for source language
- 30k words vocabulary for every target language
- OOV words are marked with UNK

Resource-Poor Setting



- Translation performance of resource-poor language pairs benefit from multi-task learning.

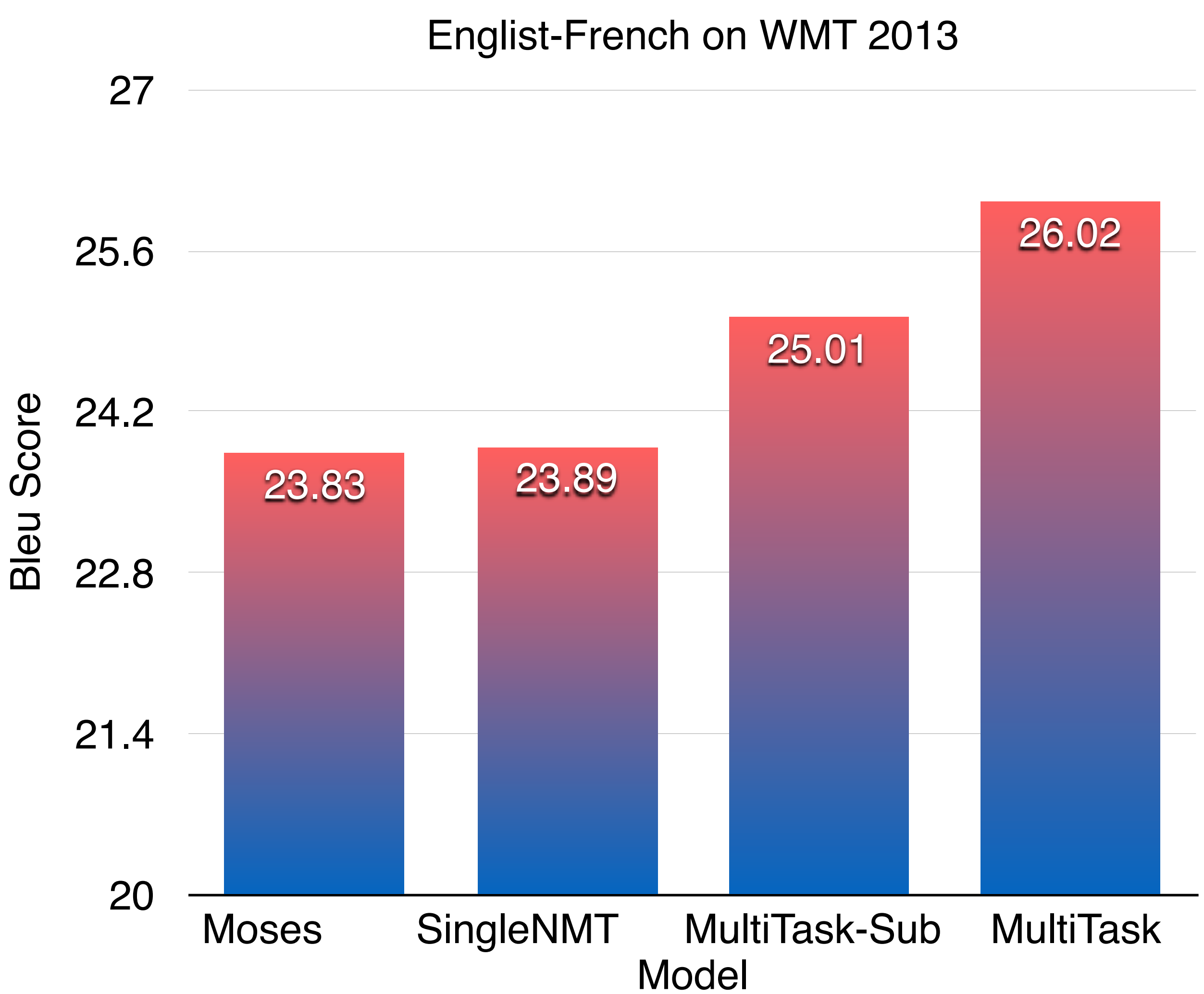
Resource-Rich Setting

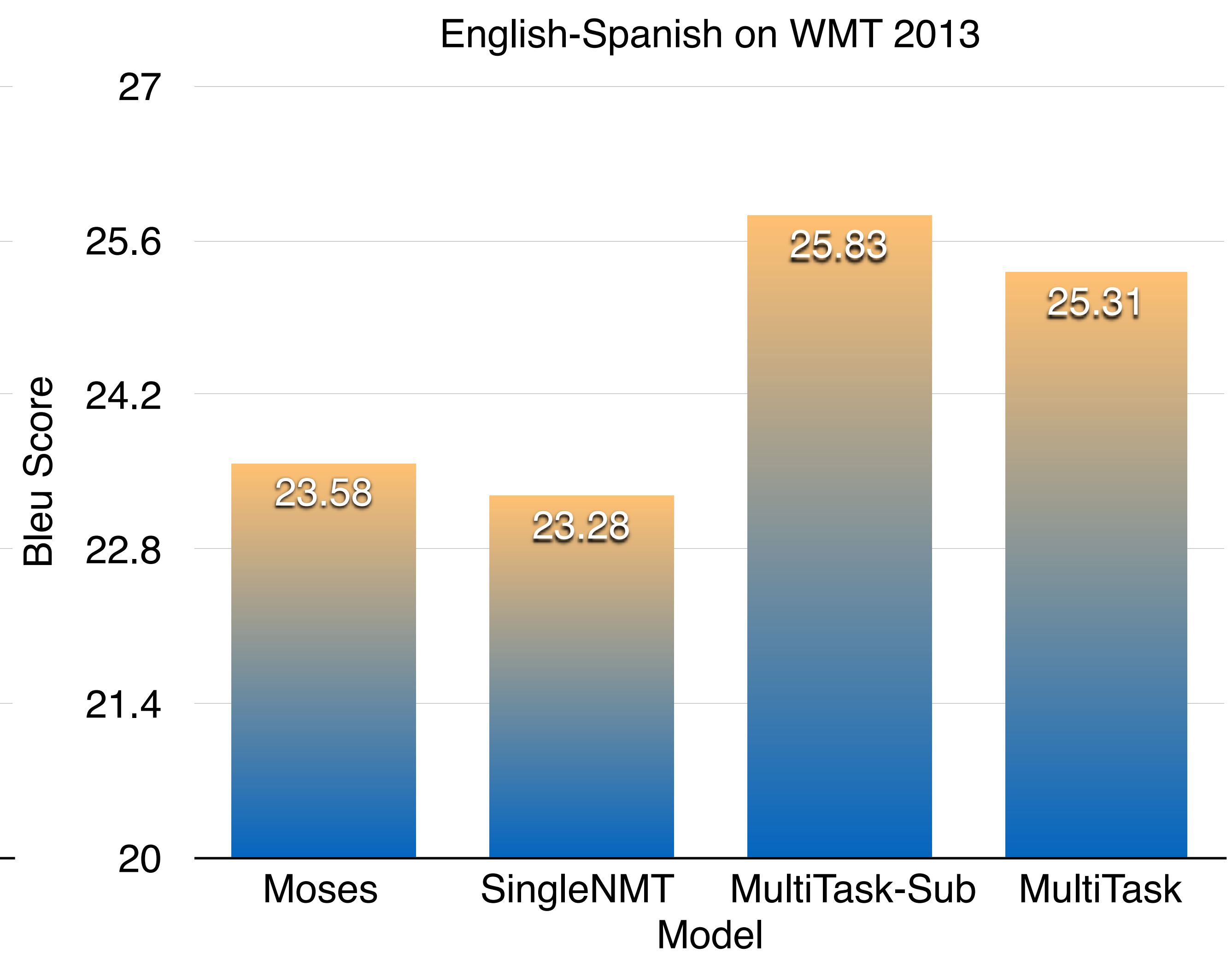


- Translation performance can also be improved given full training corpora

Comparison with Moses

Compare single NMT, Multi NMT, Multi-sub NMT with Moses model





- Single NMT is comparable with Moses.
- Multi-task learning outperforms single NMT and Moses

Why does multi-task learning work in machine translation?

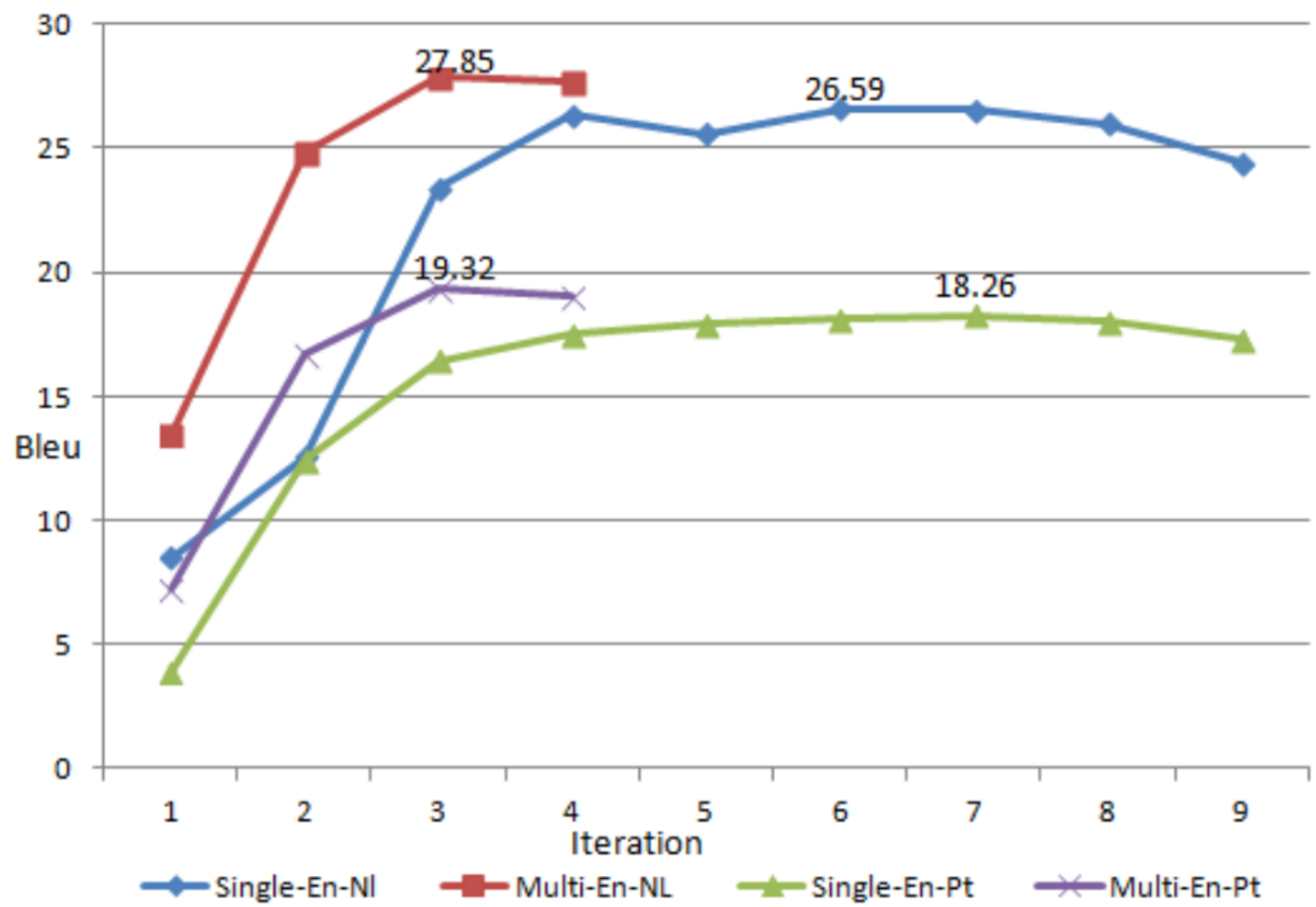
Multitask Model	Source word nearest neighbor
provide	deliever(0.78), providing(0.74), give(0.72)
crime	terrorism(0.66), criminal(0.65), homeless(0.65)
regress	condense(0.74), mutate(0.71), evolve(0.70)
six	eight(0.98), seven(0.96), 12(0.94)

NMT resource-poor Model	Source word nearest neighbor
provide	though(0.67), extending(0.56), parliamentarians
crime	care(0.75), remember(0.56), three(0.53)
regress	committing(0.33), accuracy(0.3), longed-for
six	eight(0.87), three(0.69), thirteen(0.65)

- The sharing of source information between different tasks helps to learn better source word representation

Why does multi-task learning work in machine translation?

Convergence comparison between Multi-NMT and Single NMT under resource-poor setting



- Better source word representation will help translation performance converge faster and better

Summary

- We propose a novel multi-task learning framework for machine translation
- Our framework is able to translate one source language into many different target languages within a unified model
- Experiments show that our approach can boost translation performance in every target language in both resource-poor setting and resource-rich setting.

Future work

- Extend the modeling of multiple languages into multiple domains translation.
- Consider modeling the correlation between different target decoders as well.

Thanks!