CS11-731
Machine Translation and
Sequence-to-Sequence Models

Semisupervised and Unsupervised Methods

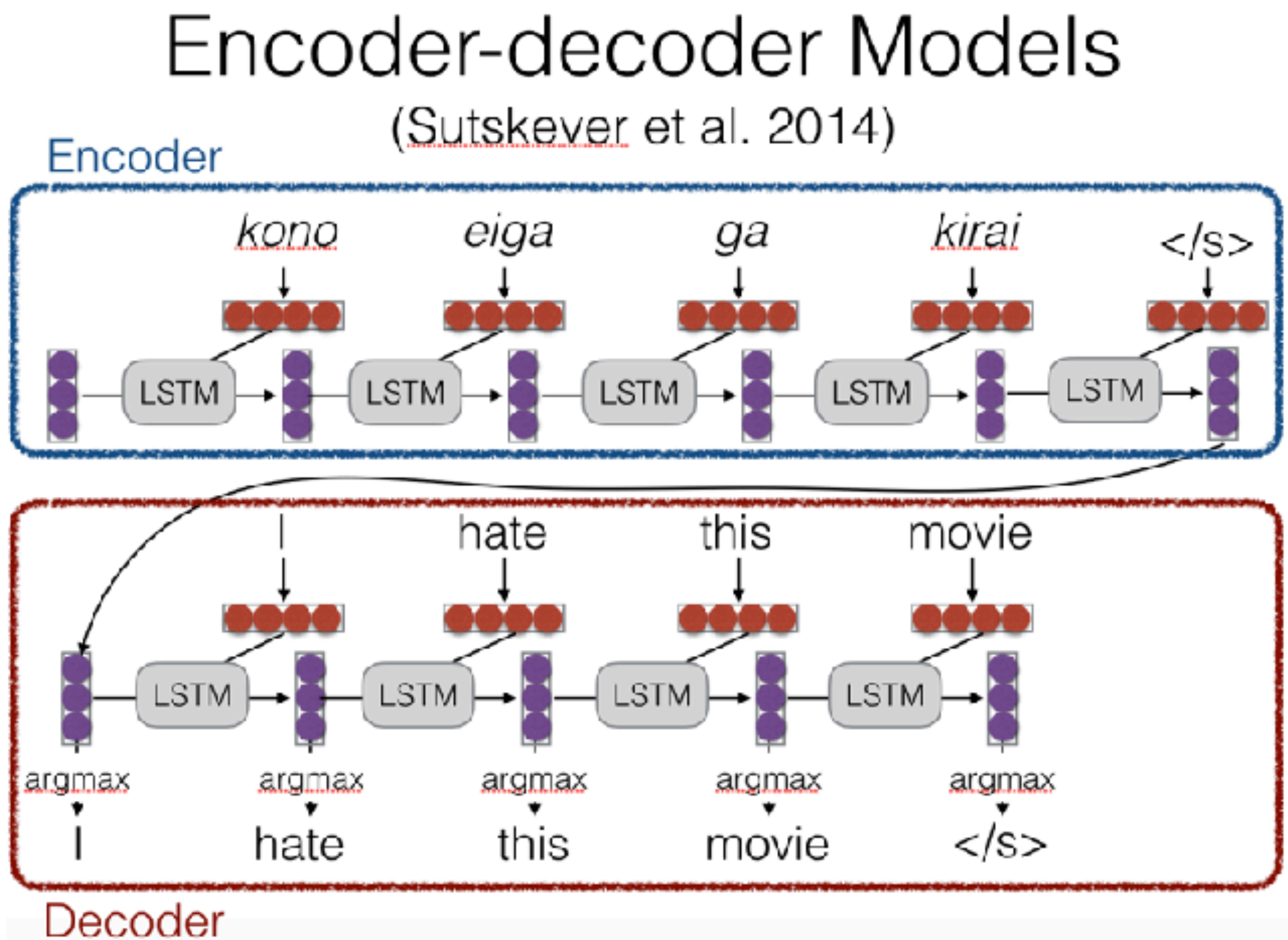
Antonis Anastasopoulos

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/mtandseq2seq2019/

Supervised Learning

We are provided the **ground truth**



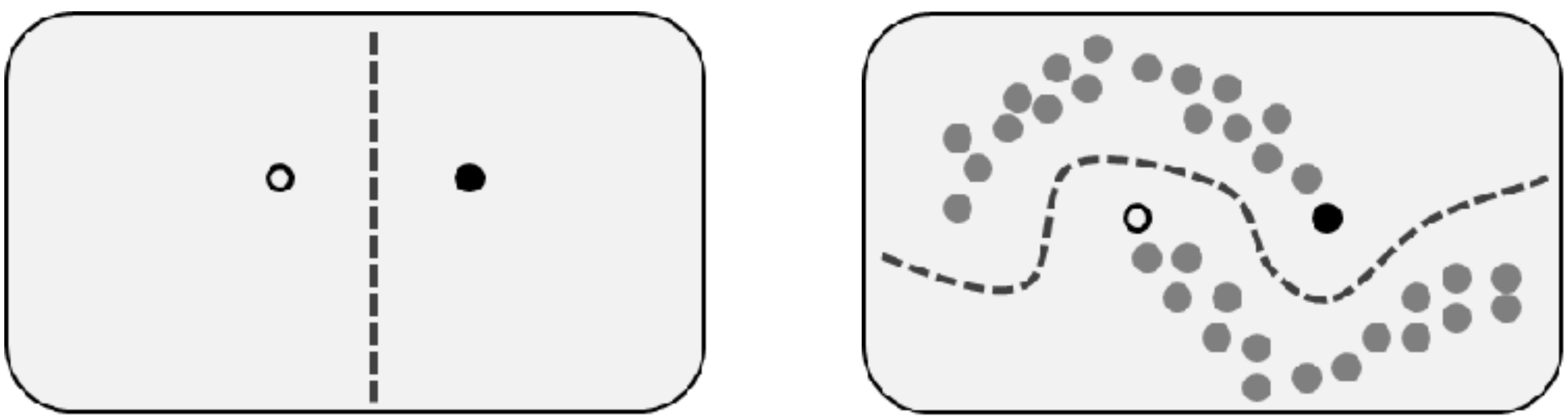
Unsupervised Learning

No ground labels:

the task is to uncover latent structure

Semi-supervised Learning

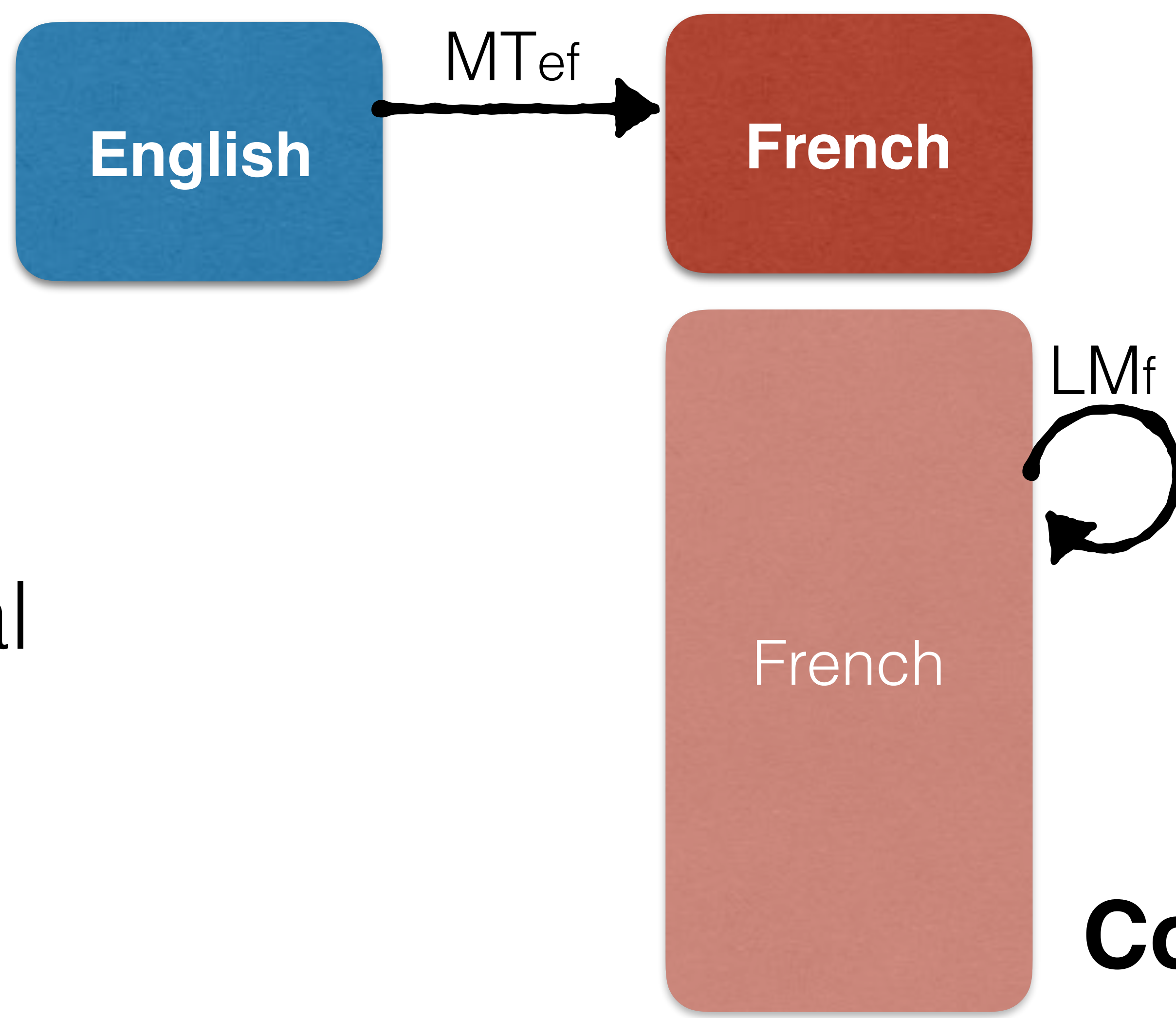
A happy medium:
use both annotated and unannotated data



Incorporating Monolingual Data

On Using Monolingual Corpora in Neural Machine Translation (Gulcehre et al. 2015)

Parallel



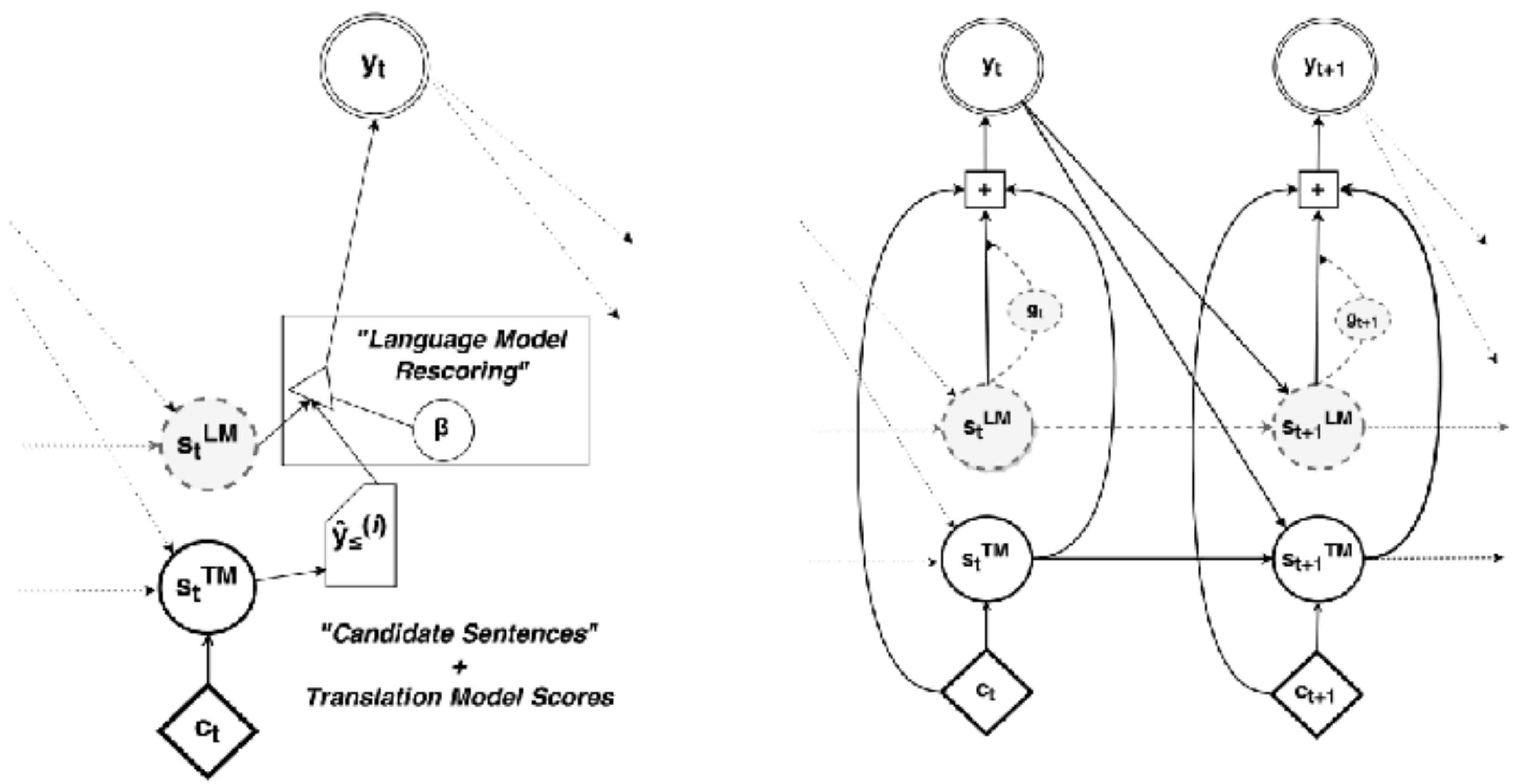
Train NMT

Monolingual

Train LM

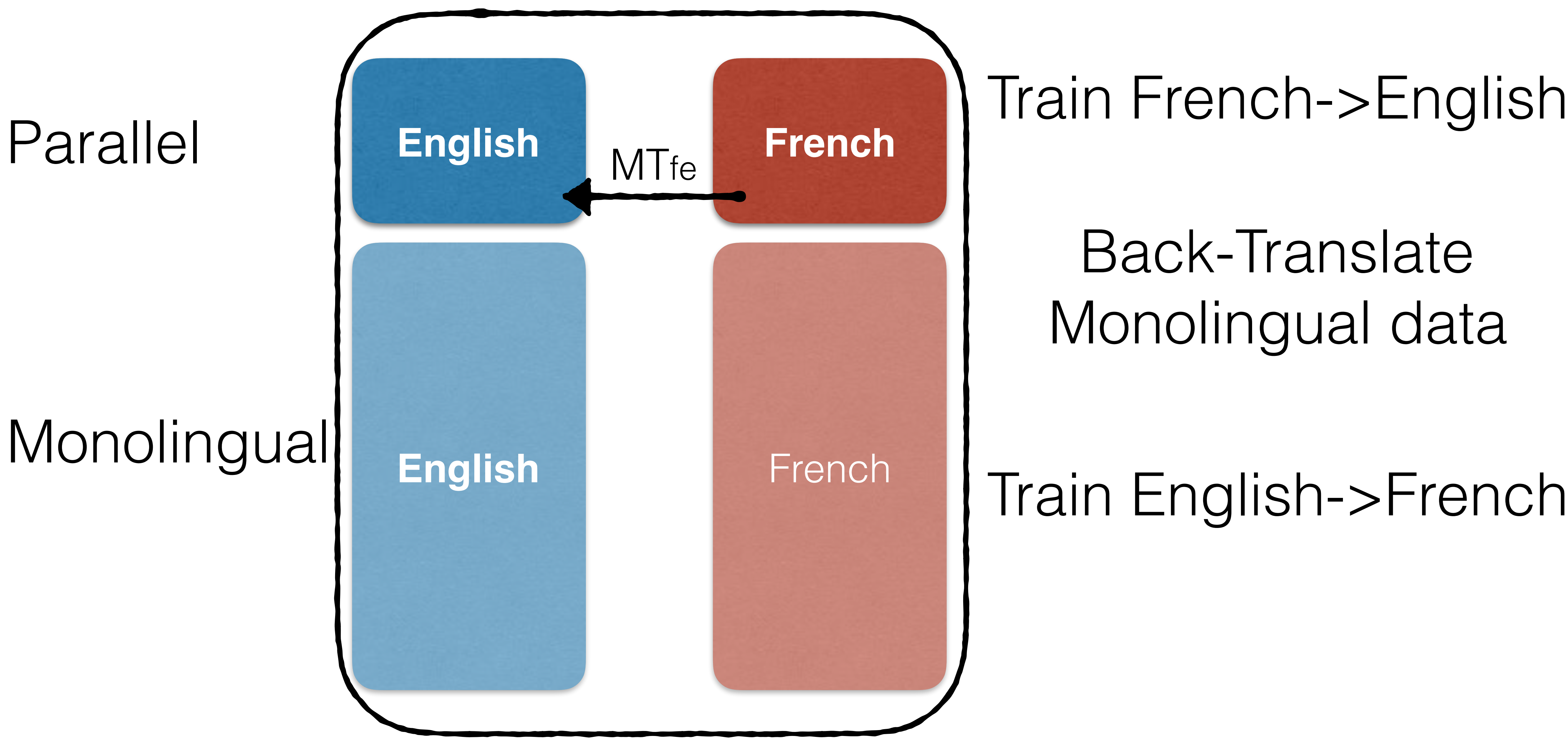
Combine the two!

On Using Monolingual Corpora in Neural Machine Translation (Gulcehre et al. 2015)



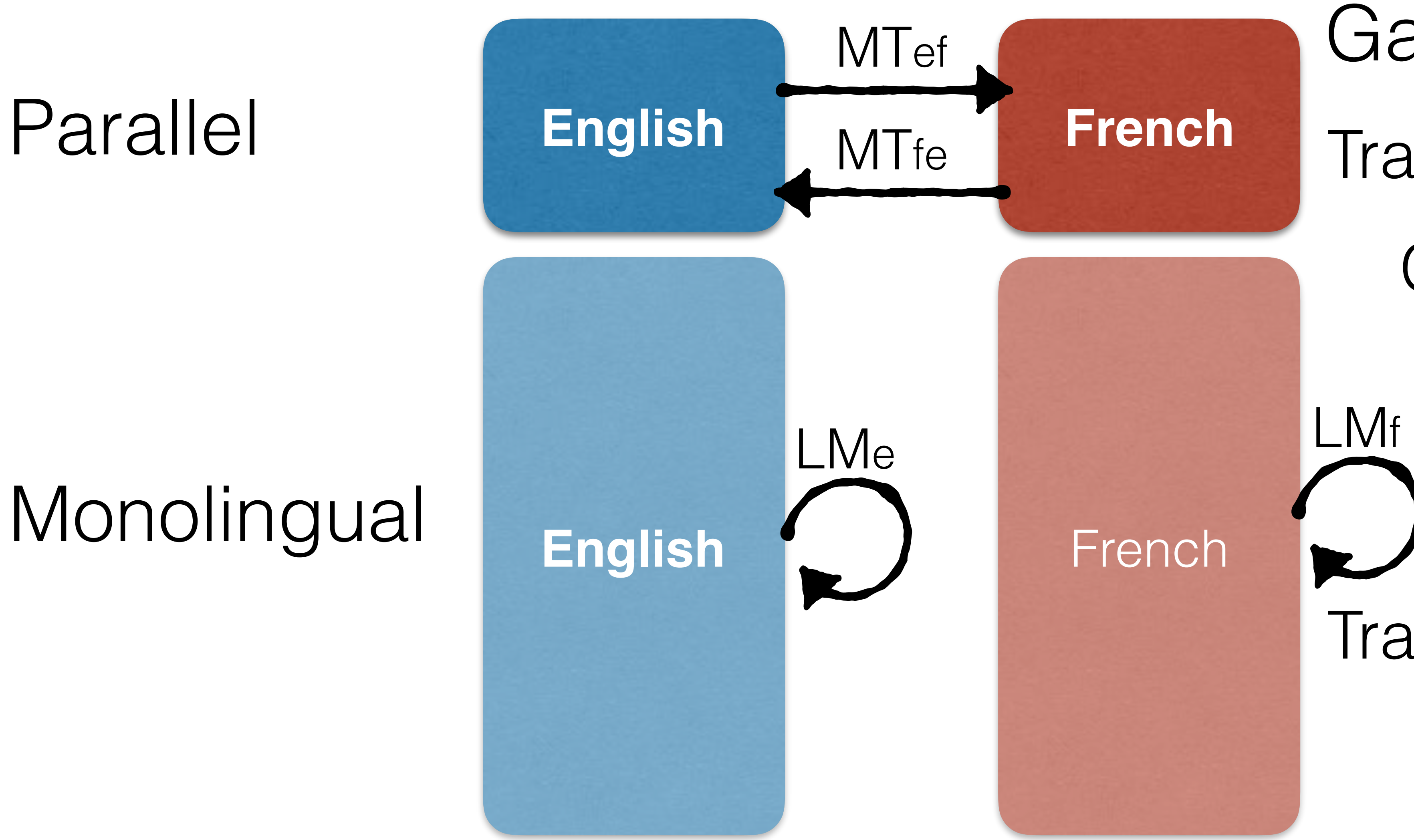
(a) Shallow Fusion (Sec. 4.1)   (b) Deep Fusion (Sec. 4.2)

Back-translation (Sennrich et al. 2016)



Dual Learning (He et al. 2016)

Assume MT_{ef}, MT_{fe}, LM_e, LM_f



Game:

Translate sample with MT_{ef}

Get reward with LM_f

Translate sample with MT_{fe}

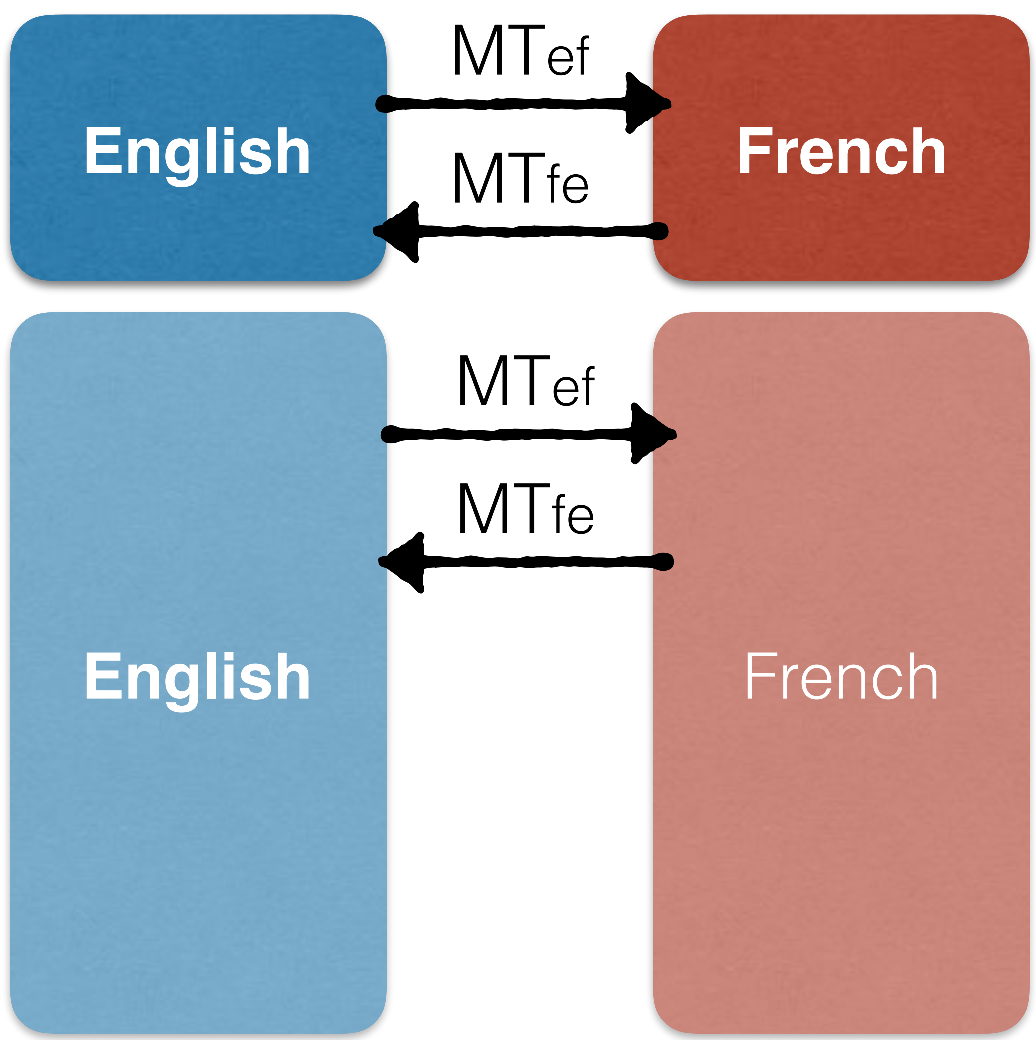
Get reward with LM_e

Semi-Supervised Learning for MT (Cheng et al. 2016)

Round-trip translation for supervision

Parallel

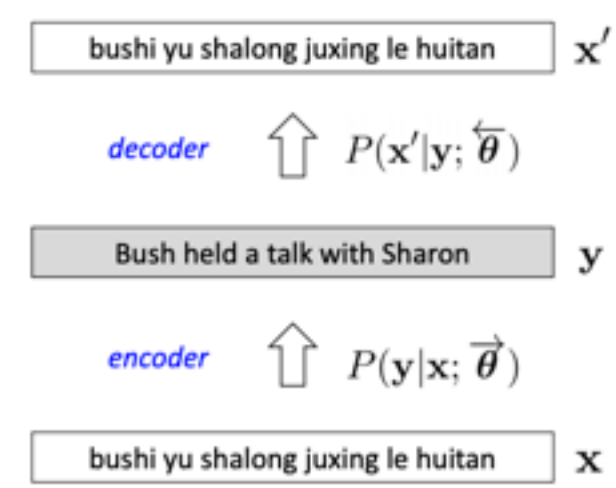
Monolingual



Translate *e* to *f'* with MT_{ef}

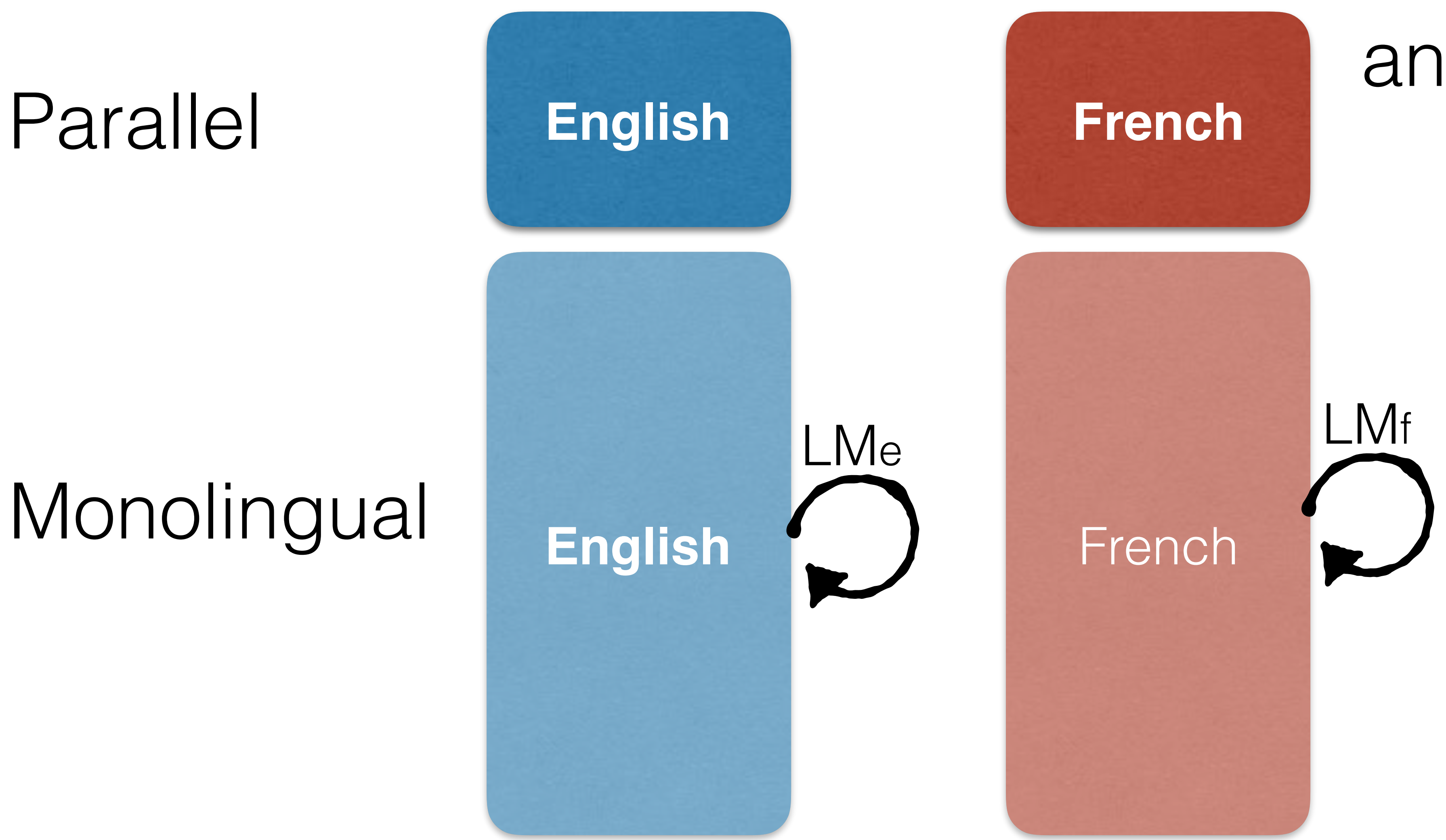
Translate *f'* to *e'* with MT_{fe}

Loss from *e* and *e'*

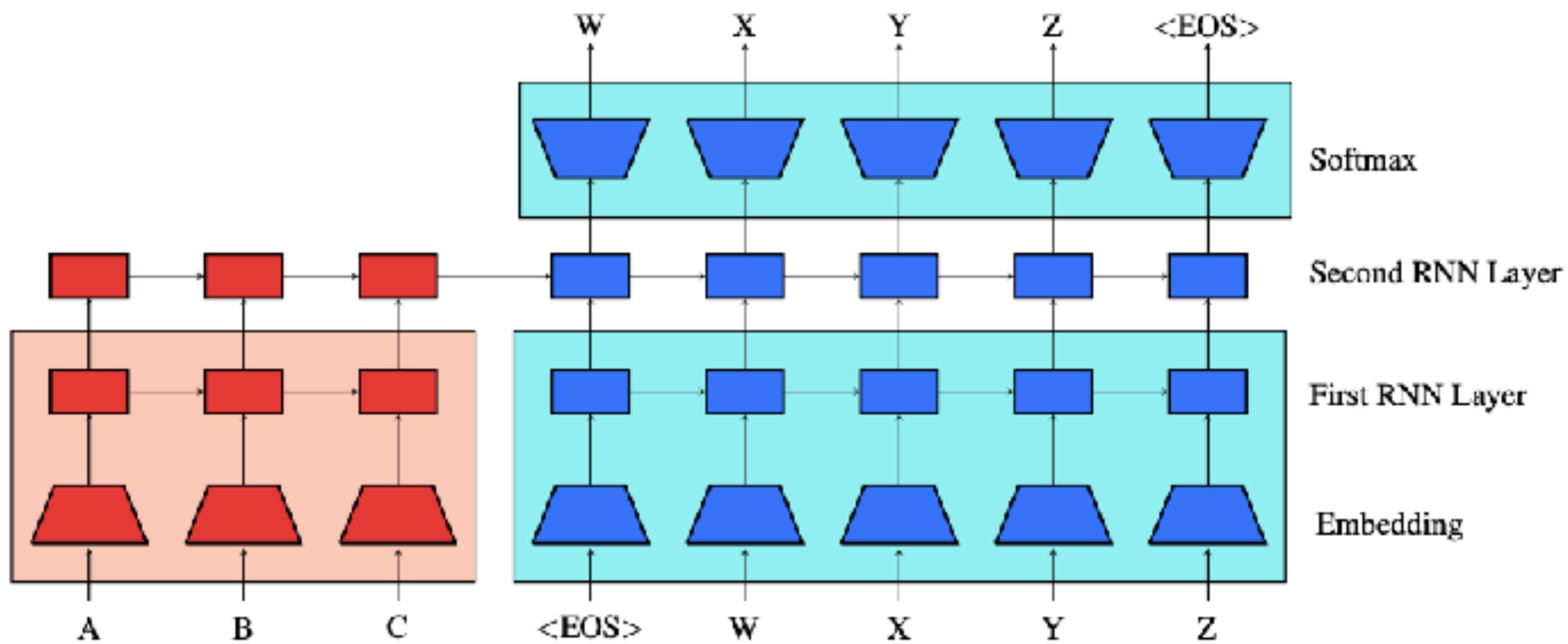


Another idea: use monolingual data to pretrain model components

Use the monolingual data to train the encoder and the decoder.



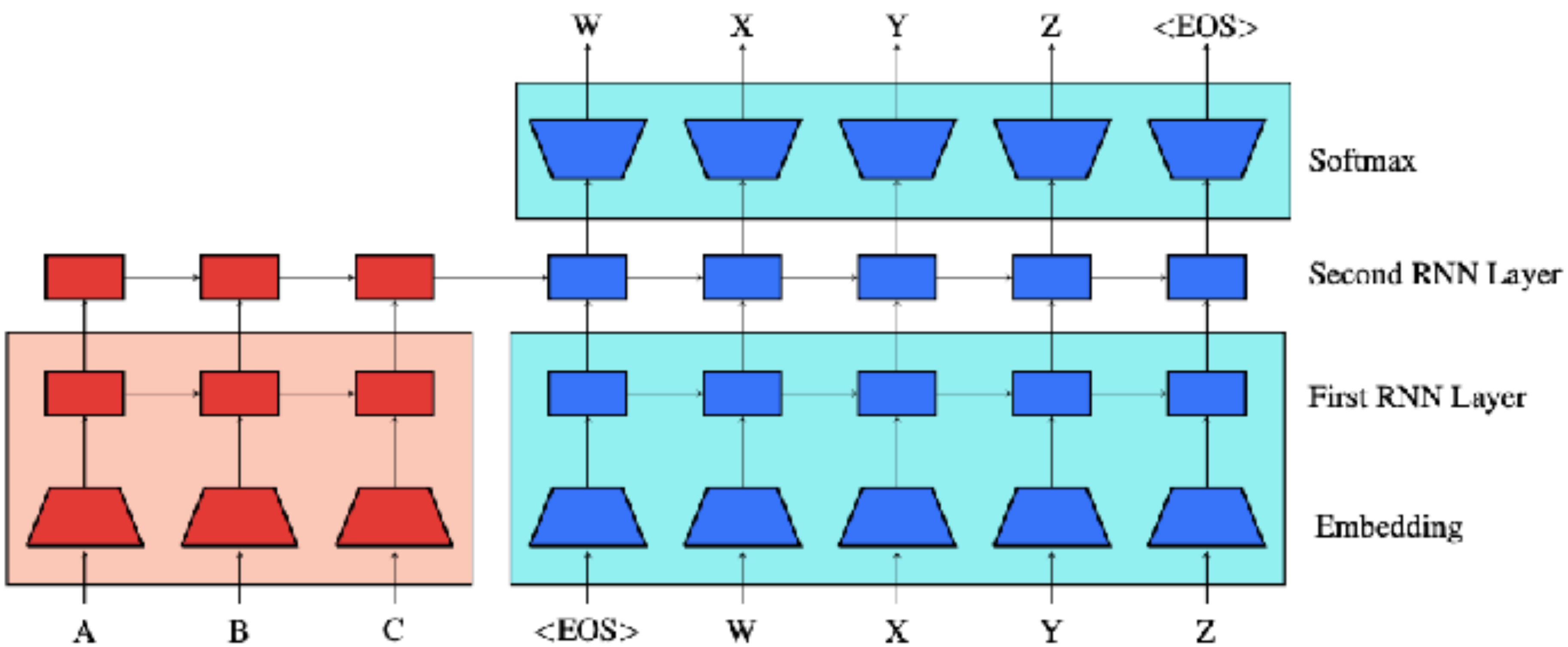
Another idea: use monolingual data to pretrain model components



Shaded regions are pre-trained

From "Unsupervised Pretraining for Sequence to Sequence Learning", Ramachadran et al. 2017.

Another idea: use monolingual data to pretrain model components



Shaded regions are pre-trained

From "Unsupervised Pretraining for Sequence to Sequence Learning", Ramachadran et al. 2017.

Another idea: use monolingual data to pretrain model components

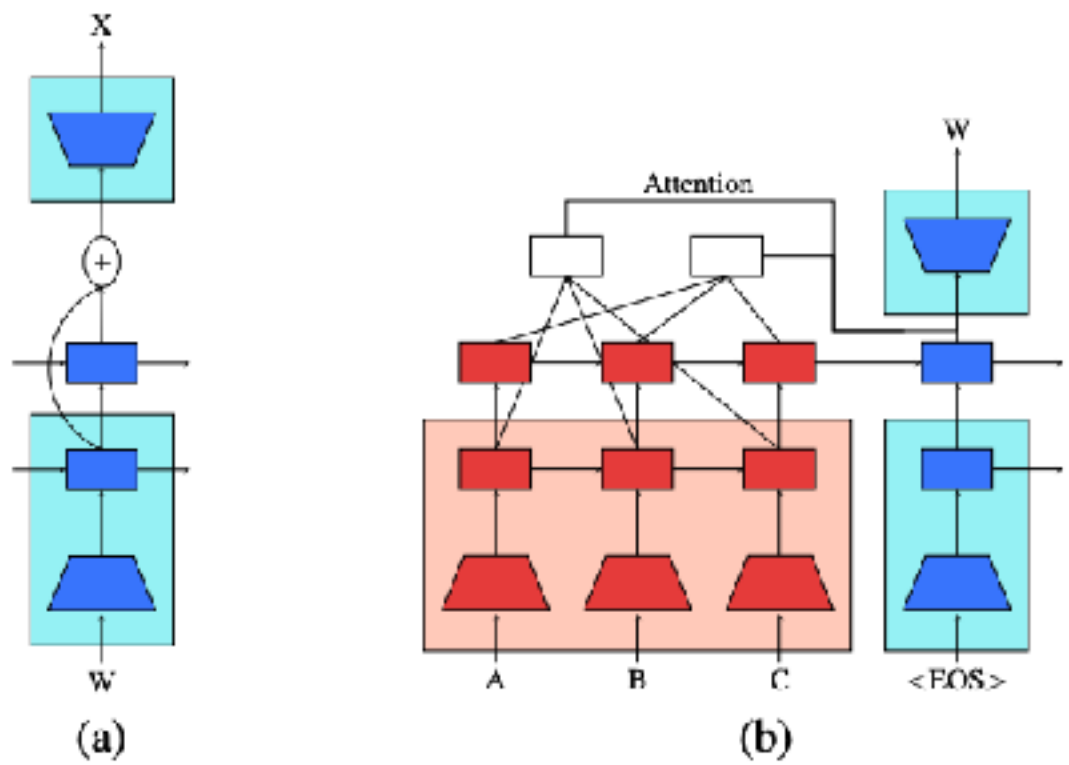


Figure 2: Two small improvements to the baseline model: (a) residual connection, and (b) multi-layer attention.

From "Unsupervised Pretraining for Sequence to Sequence Learning", Ramachadran et al. 2017.

Another idea: use monolingual data to pretrain model components

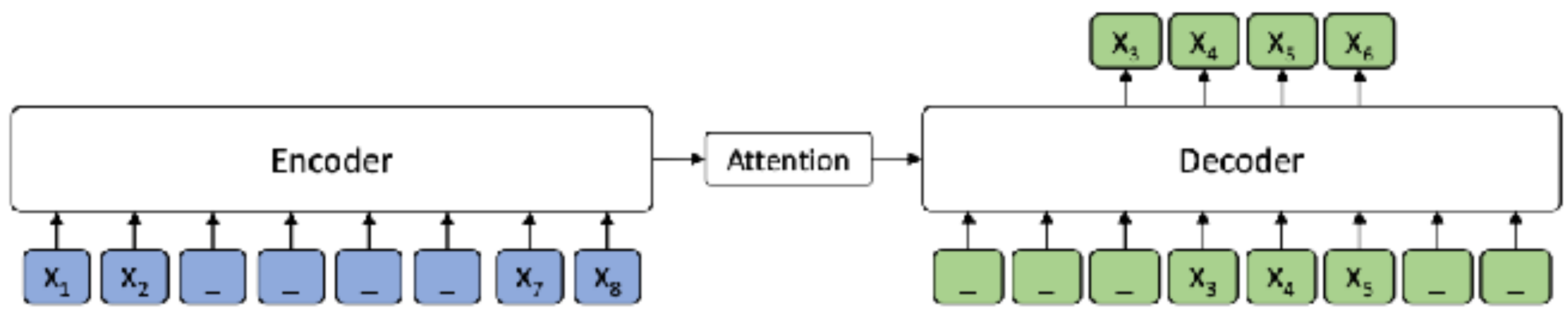
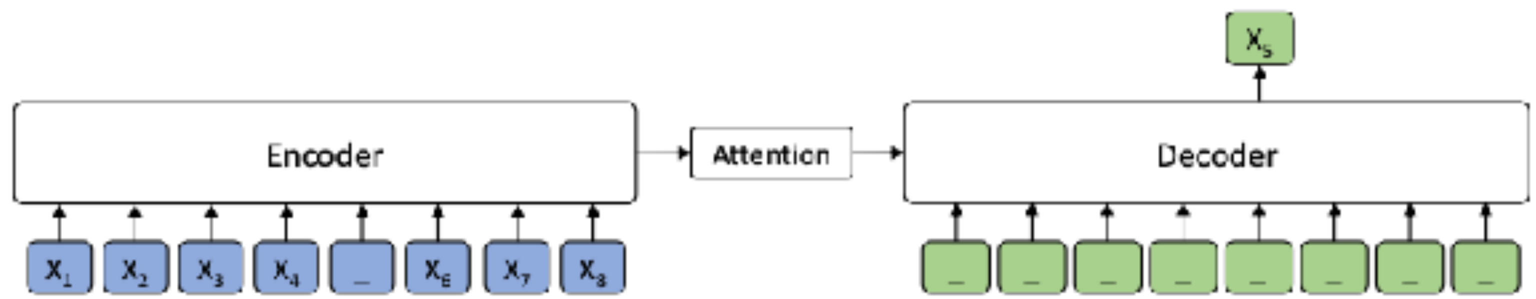


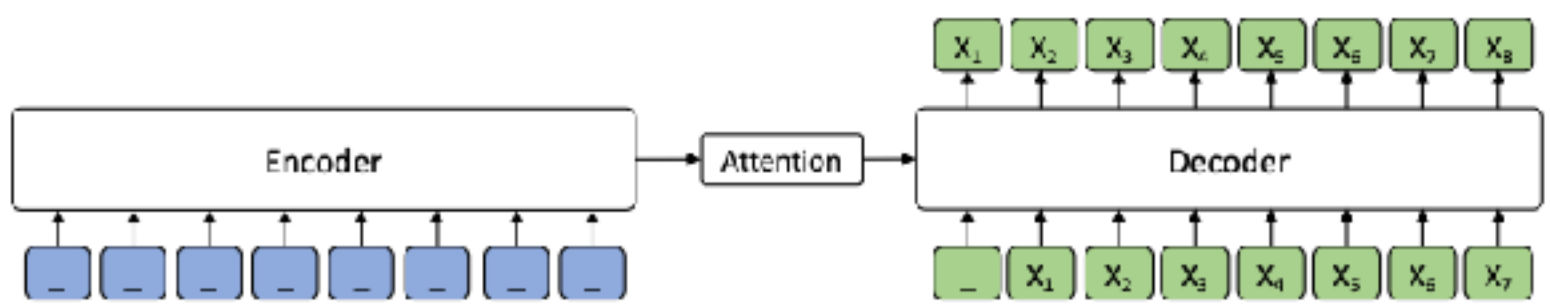
Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [M].

From "MASS: Masked Sequence to Sequence Pre-training for Language Generation", Song et al. 2019.

Another idea: use monolingual data to pretrain model components



(a) **Masked language modeling in BERT** ($k = 1$)



(b) **Standard language modeling** ($k = m$)

From "MASS: Masked Sequence to Sequence Pre-training for Language Generation", Song et al. 2019.

Another idea: use monolingual data to pretrain model components

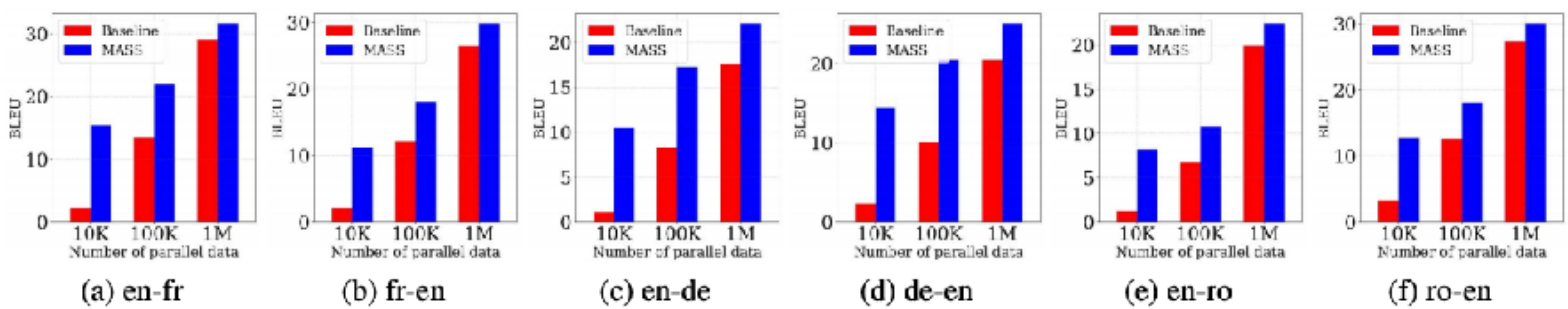


Figure 3. The BLEU score comparisons between MASS and the baseline on low-resource NMT with different scales of paired data.

From "MASS: Masked Sequence to Sequence Pre-training for Language Generation", Song et al. 2019.

Pre-trained Word Embeddings in NMT

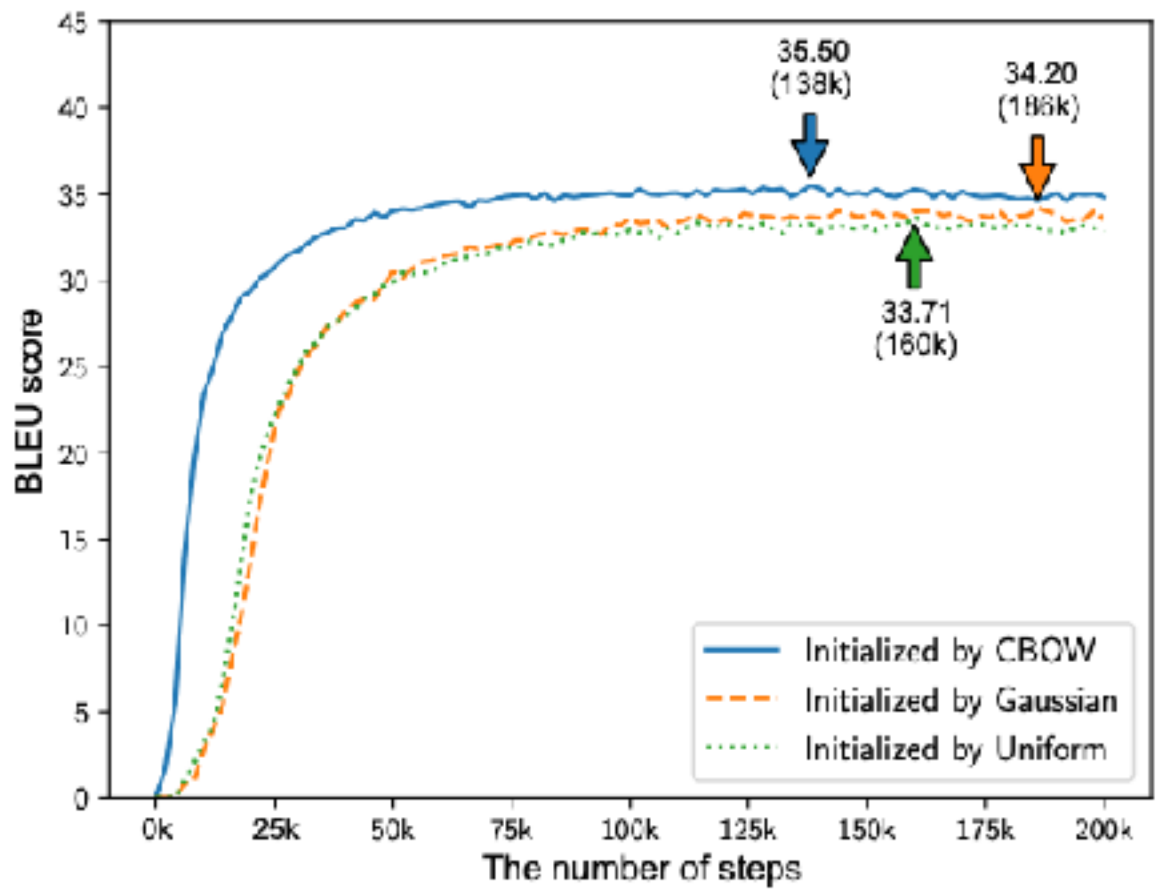
Modern neural embeddings (Mikolov et al, 2014)

Skip-gram model: predict a word’s context

CBOW model: predict a word from its context

Others: GLoVe, fastText, etc

Pre-trained embeddings



From "A Bag of Useful Tricks for Practical Neural Machine Translation: Embedding Layer Initialization and Large Batch Size", Neishi et al. 2017.

Pre-trained embeddings: when are they useful?

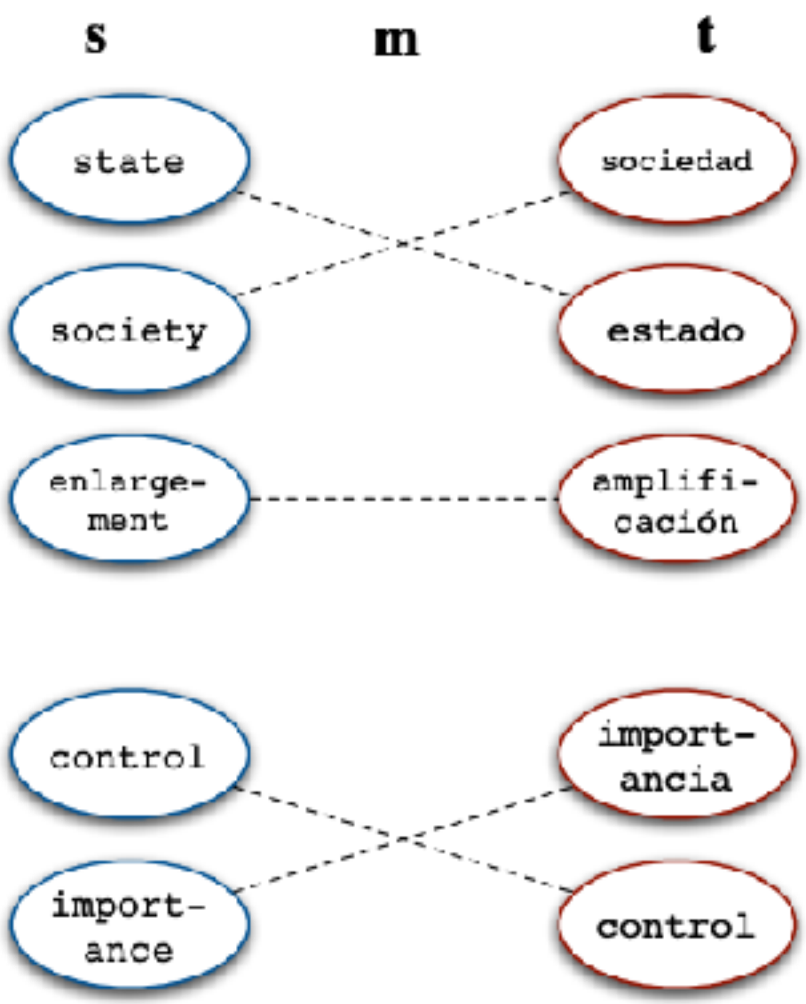
Src → / → Trg	std / std	pre / std	std / pre	pre / pre
GL → EN	2.2	**13.2**	2.8	12.8
PT → EN	26.2	**30.3**	26.1	**30.8**
AZ → EN	1.3	**2.0**	1.6	**2.0**
TR → EN	14.9	17.6	14.7	**17.9**
BE → EN	1.6	2.5	1.3	**3.0**
RU → EN	18.5	**21.2**	18.7	**21.1**

Table 2: Effect of pre-training on BLEU score over six languages. The systems use either random initialization (std) or pre-training (pre) on both the source and target sides.

From "When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?", Qi et al. 2017.

Bilingual Lexicon Induction

What is Bilingual Lexicon Induction?



From "Learning Bilingual Lexicons from Monolingual Corpora", Haghighi et al. 2008.

What is Bilingual Lexicon Induction?

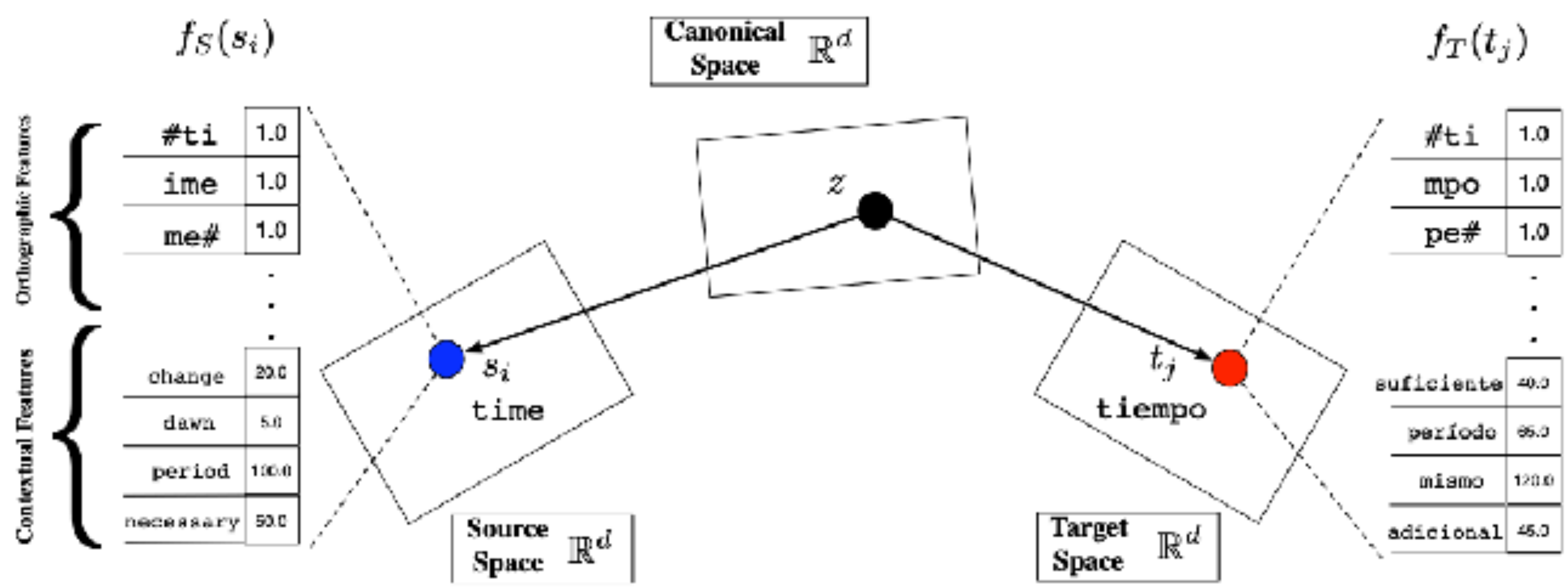
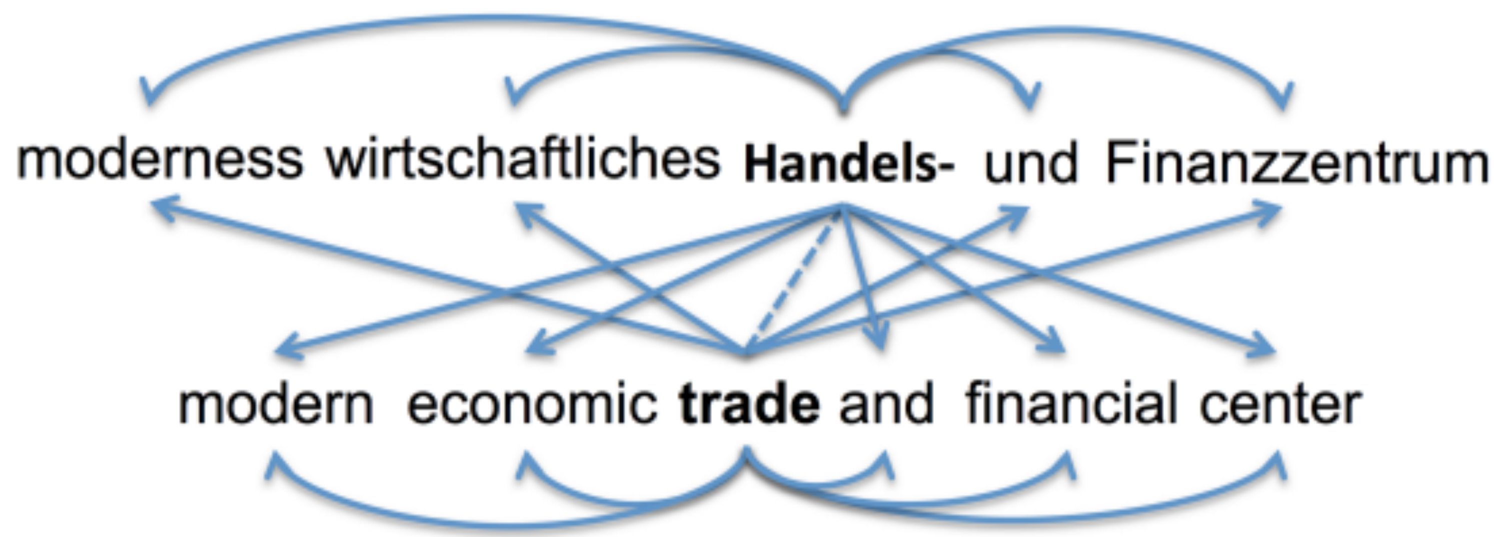


Figure 2: Illustration of our MCCA model. Each latent concept $z_{i,j}$ originates in the canonical space. The observed word vectors in the source and target spaces are generated independently given this concept.

From "Learning Bilingual Lexicons from Monolingual Corpora", Haghighi et al. 2008.

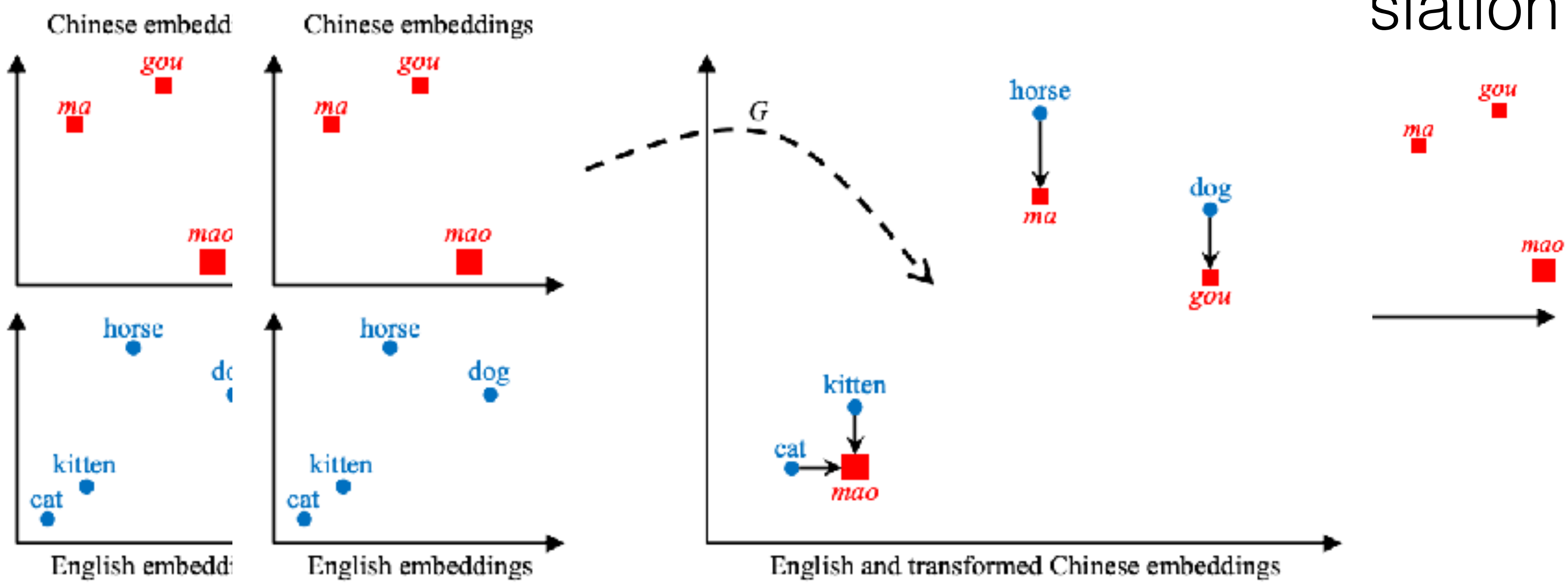
Bilingual Skip-gram model: Using translations and alignments



From "Bilingual Word Representations with Monolingual Quality in Mind", Luong et al. 2015.

Mapping two monolingual embedding spaces

slation



From "Earth Mover’s Distance Minimization for Unsupervised Bilingual Lexicon Induction", Zhang et al. 2015.

Finding the best mapping

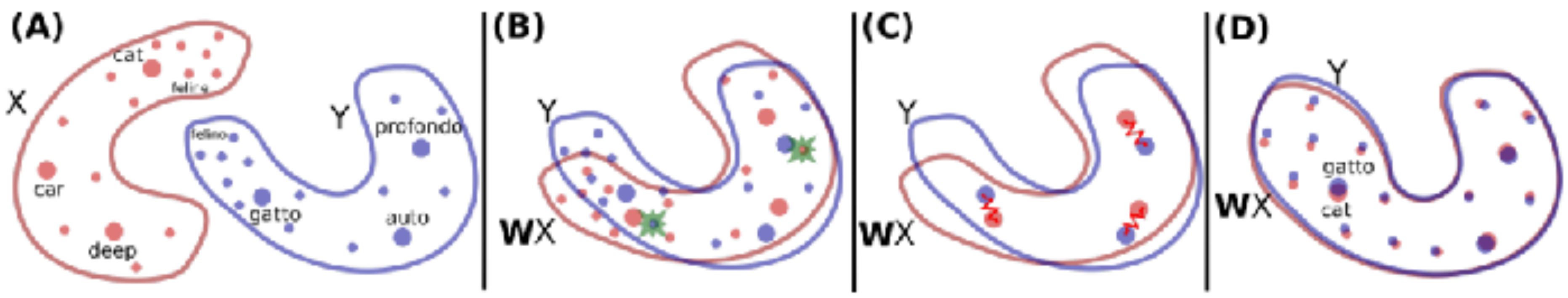
The orthogonality assumption is important!

$$W^\star = \operatorname*{argmin}_{W \in O_d(\mathbb{R})} \|WX - Y\|_{\mathrm{F}} = UV^T, \text{with } U\Sigma V^T = \mathrm{SVD}(YX^T).$$

What about if we don’t have a seed lexicon?

From "Word Translation Without Parallel Data", Conneau et al. 2018.

Unsupervised Mapping + Refinement



From "Word Translation Without Parallel Data", Conneau et al. 2018.

Issues with mapping methods

(a) Top 10 most frequent English words

(b) German translations of top 10 most frequent English words

(c) Top 10 most frequent English nouns

(d) German translations of top 10 most frequent English nouns

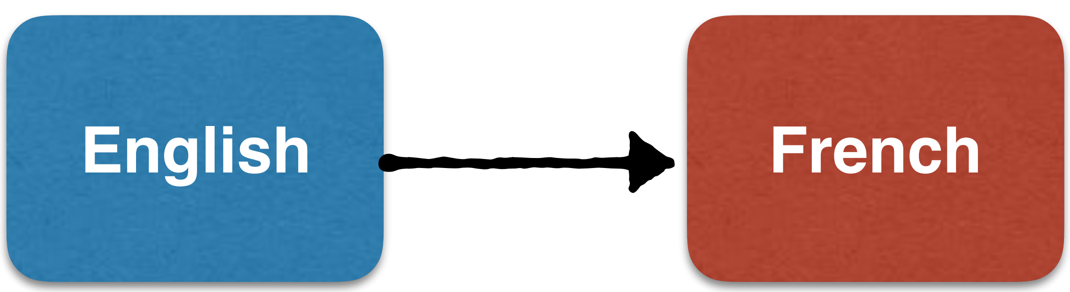
From "On the Limitations of Unsupervised Bilingual Dictionary Induction", Søgaard et al. 2018.

Unsupervised Translation

… at the core of it all: decipherment

$$\arg\max_{\theta} \prod_{f} P_{\theta}(f)$$

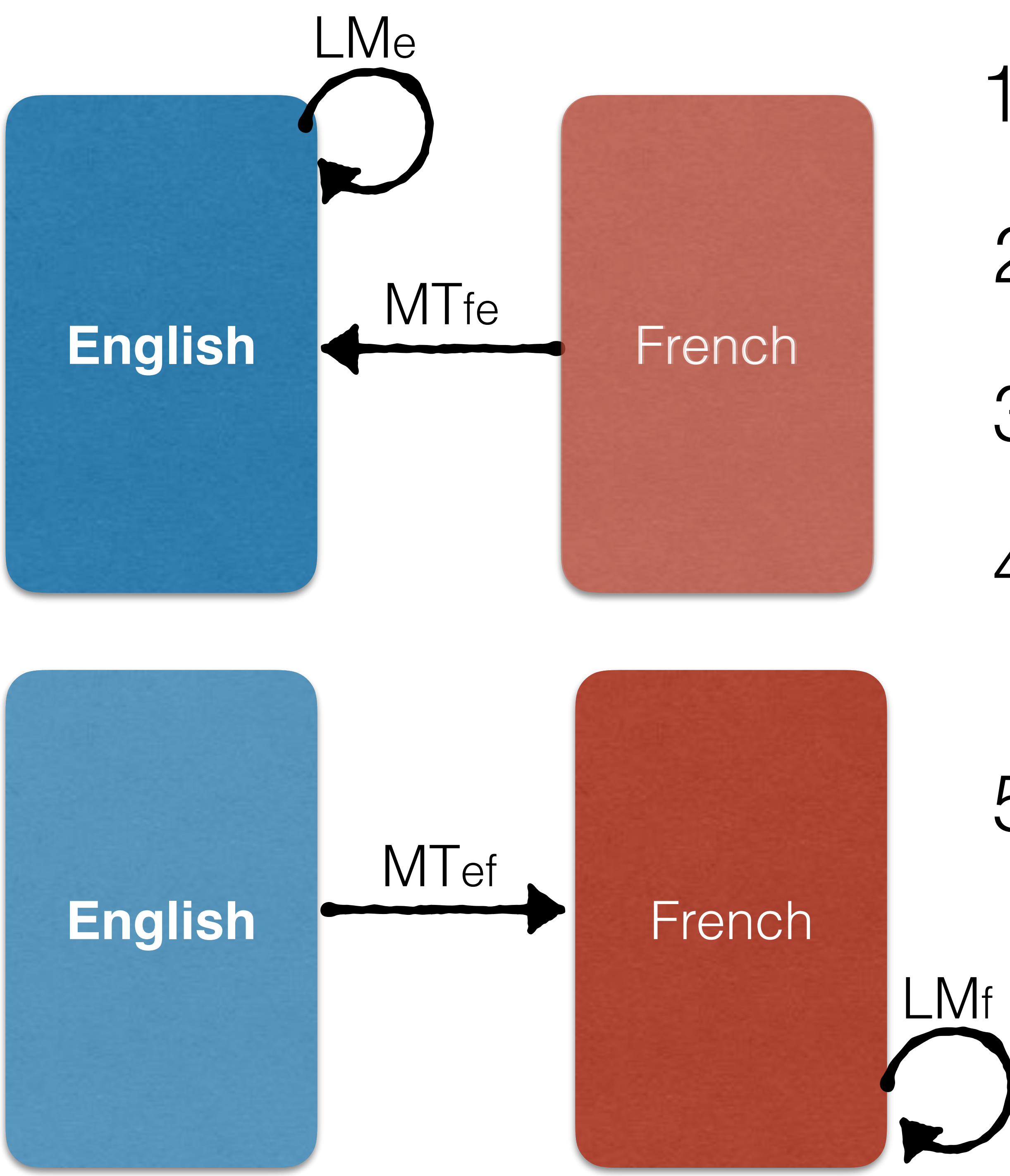
Weaver (1955): *This is really English, encrypted in some strange symbols*



$$\arg\max_{\theta} \prod_{f} \sum_{e} P(e) \cdot P_{\theta}(f|e)$$

From "Deciphering Foreign Language", Ravi and Knight 2011.

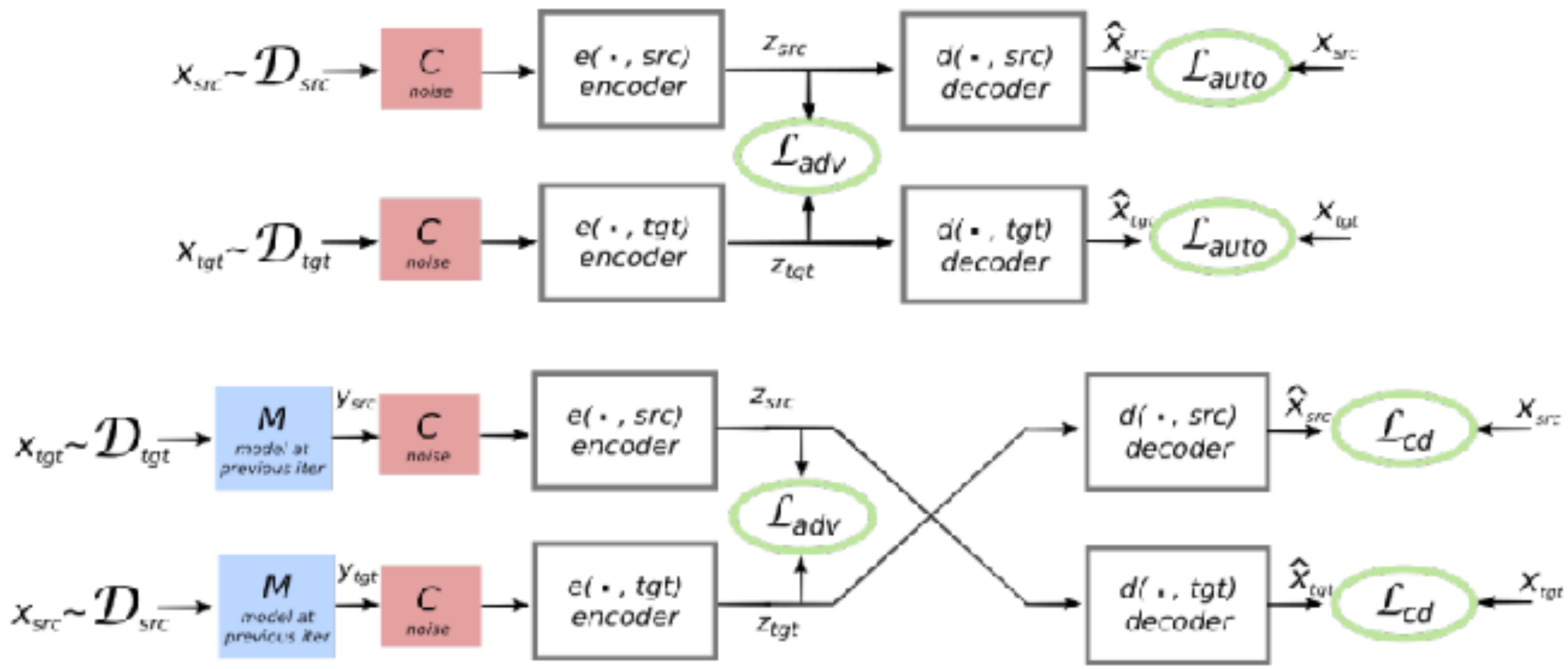
Unsupervised MT (Lample et al. and Artetxe et al. 2018)



1. Embeddings + Unsup. BLI
2. BLI —> Word Translations
3. Train MT_{fe} and MT_{ef} systems
4. Meanwhile, use unsupervised objectives (denoising LM)
5. Iterate

Unsupervised MT (Lample et al. 2018)

Also add an adversarial loss for the intermediate representations:



From "Unsupervised MT Using Monolingual Corpora Only", Lample et al 2018.

Unsupervised MT (Lample et al. 2018)

Source	un homme est debout près d' une série de jeux vidéo dans un bar .
Iteration 0	a man is seated near a series of games video in a bar .
Iteration 1	a man is standing near a closeup of other games in a bar .
Iteration 2	a man is standing near a bunch of video video game in a bar .
Iteration 3	a man is standing near a bunch of video games in a bar .
Reference	**a man is standing by a group of video games in a bar .**
Source	une femme aux cheveux roses habillée en noir parle à un homme .
Iteration 0	a woman at hair roses dressed in black speaks to a man .
Iteration 1	a woman at glasses dressed in black talking to a man .
Iteration 2	a woman at pink hair dressed in black speaks to a man .
Iteration 3	a woman with pink hair dressed in black is talking to a man .
Reference	**a woman with pink hair dressed in black talks to a man .**
Source	une photo d' une rue bondée en ville .
Iteration 0	a photo a street crowded in city .
Iteration 1	a picture of a street crowded in a city .
Iteration 2	a picture of a crowded city street .
Iteration 3	a picture of a crowded street in a city .
Reference	**a view of a crowded city street .**

From "Unsupervised MT Using Monolingual Corpora Only", Lample et al 2018.