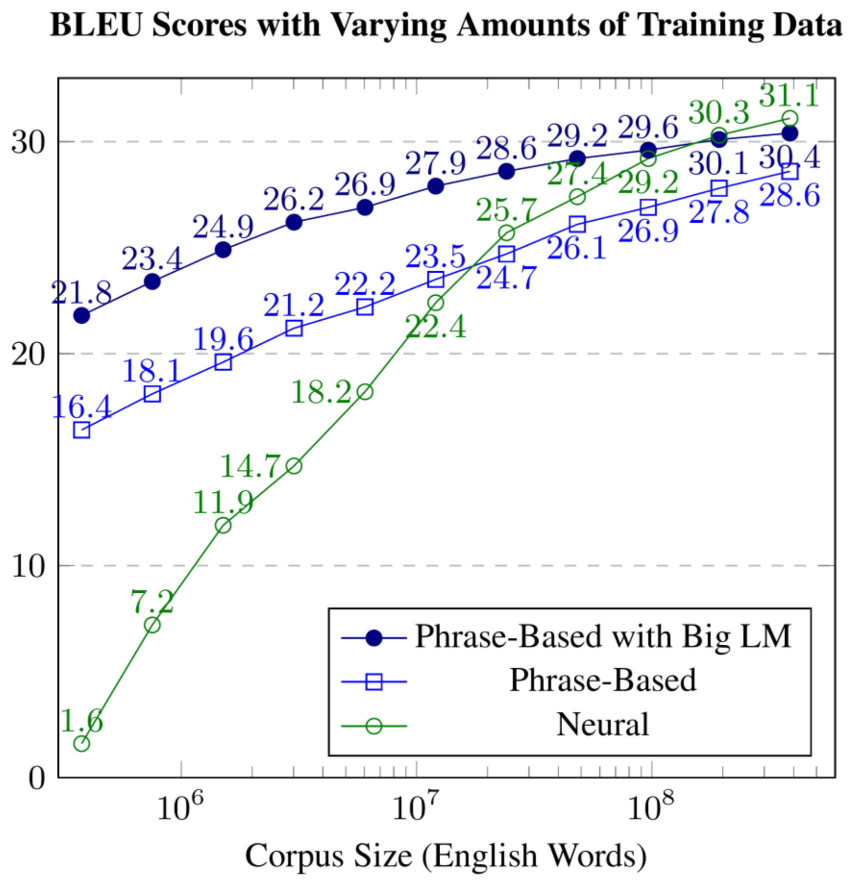
COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Neural machine translation with less supervision

CMSC 470

Marine Carpuat

Neural MT only helps in high-resource settings



[Koehn & Knowles 2017]

Ongoing research

- Learn from other sources of supervision than pairs (E,F)
    - **Monolingual text**
    - Multiple languages

Neural Machine Translation
Standard Training is **Supervised**

- We are provided with pairs (**x**,**y**) where **y** ts the ground truth for each sample **x**

  x = Chinese sentence

  y = translation of **x** in English written by a human

- What is the training loss?

Unsupervised learning

- No labels for training samples
    - E.g., we are provided with Chinese sentences **x**, or English sentences **y**, but no **(x,y)** pairs

- Goal: uncover latent structure in unlabeled data

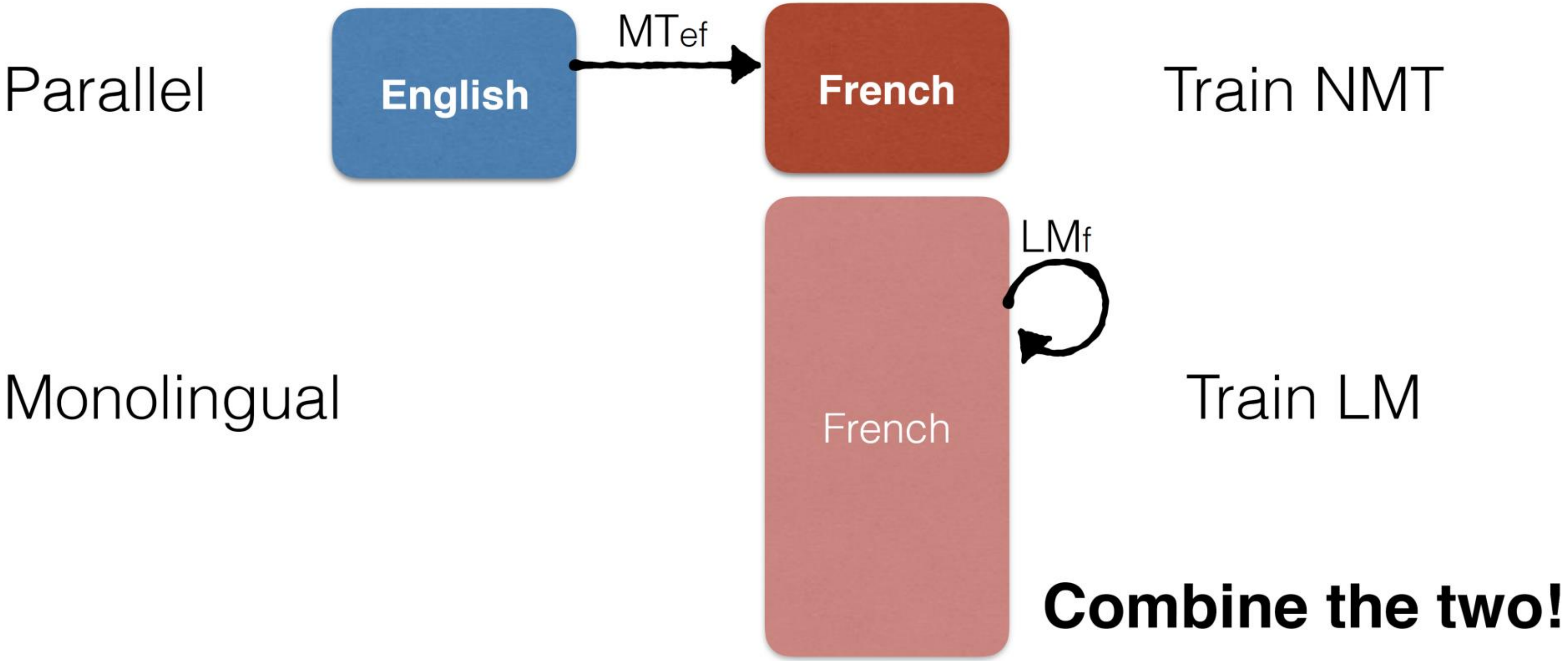
Semi-supervised learning

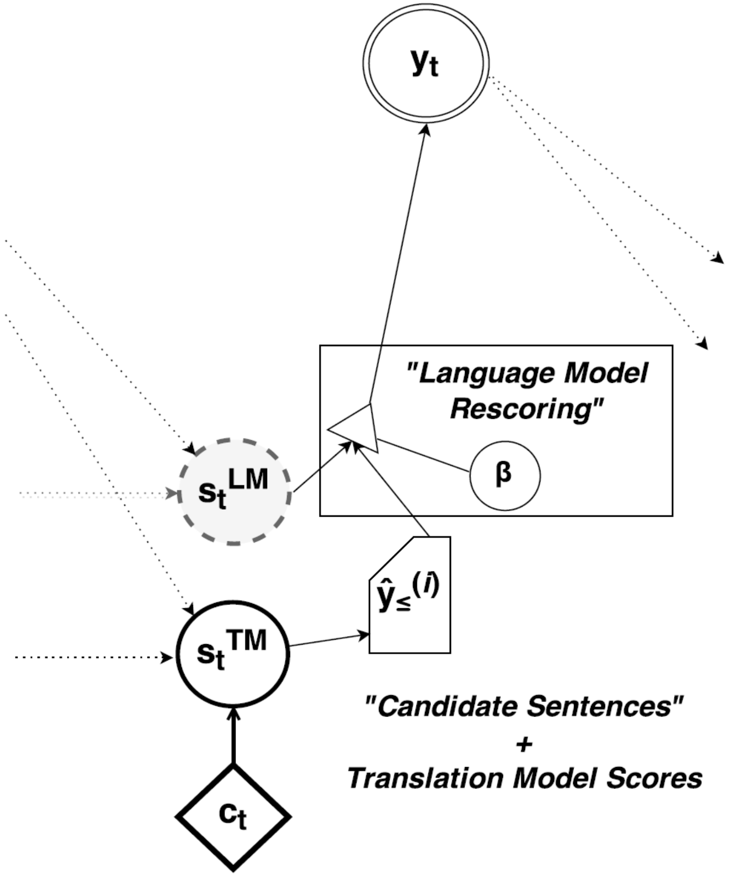
- Uses both annotated and unannotated data
  - **(x,y)** Chinese-English pairs
  - Chinese sentences **x**, and/or English sentences **y**

- Combines
  - Direct optimization of supervised training objective
  - Better modeling of data with cheaper unlabeled examples

Semi-supervised NMT

Using Monolingual Corpora in Neural Machine Translation [Gulcehre et al. 2015]





(a) Shallow Fusion (Sec. 4.1)

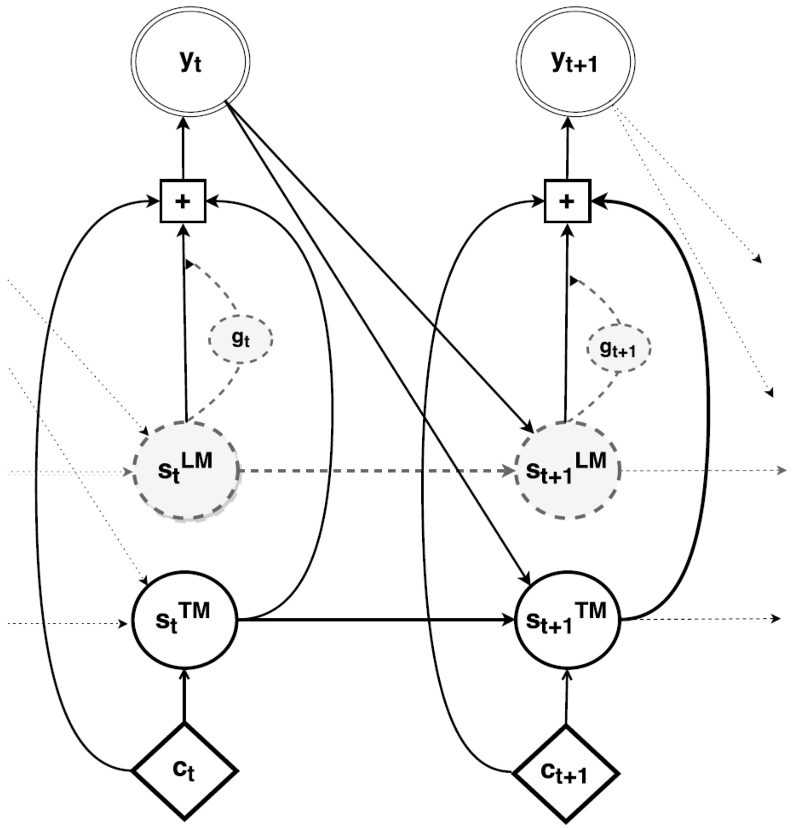
Approach 1: Shallow Fusion

Use a language model to rescore translation candidates from the NMT decoder

$$\begin{aligned}\log p(\mathbf{y}_t = k) &= \log p_{\text{TM}}(\mathbf{y}_t = k) \\ &\quad + \beta \log p_{\text{LM}}(\mathbf{y}_t = k),\end{aligned}$$

Approach 2: Deep Fusion

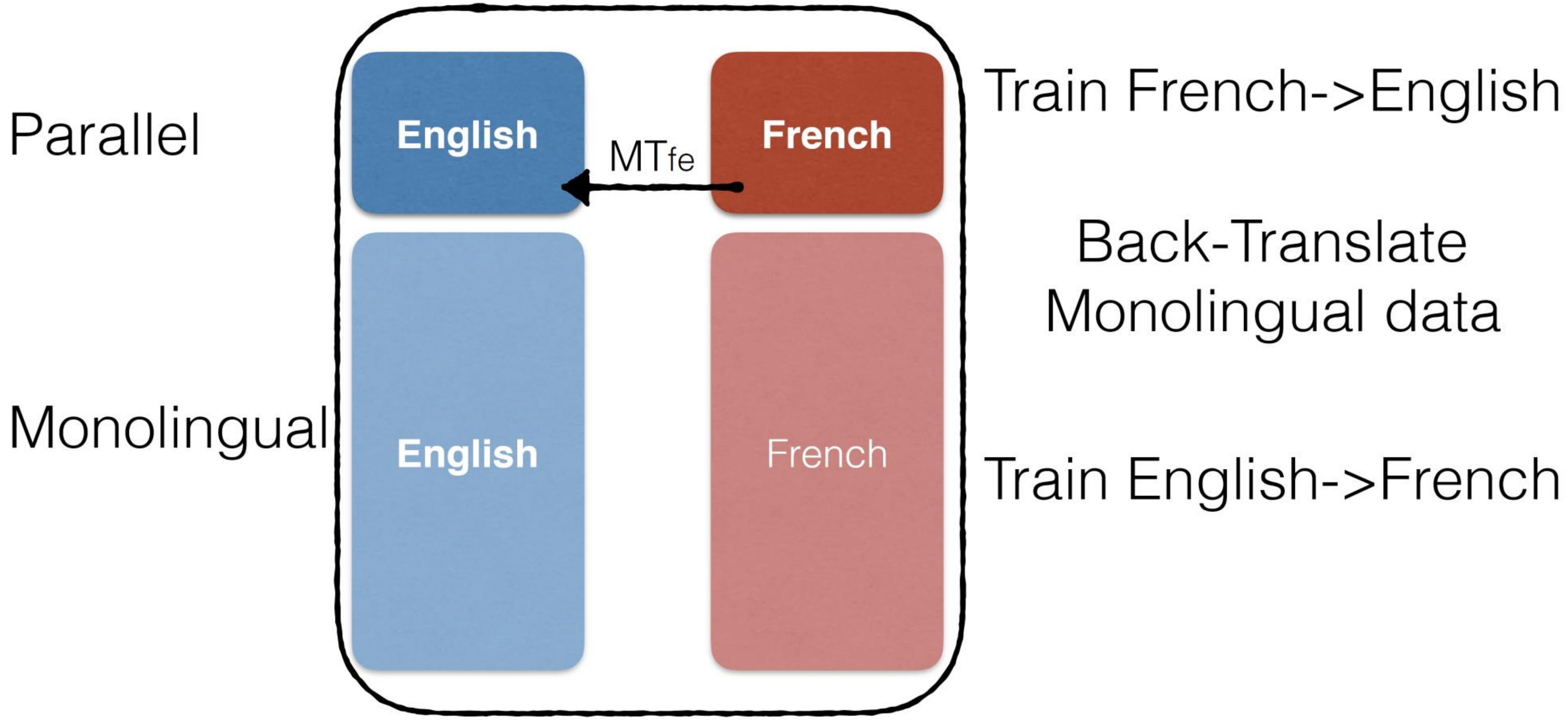
Integrate RNN language model and NMT model by concatenating their hidden states



(b) Deep Fusion (Sec. 4.2)

$$p(\mathbf{y}_t|\mathbf{y}_{<t},\mathbf{x}) \propto \exp(\mathbf{y}_t^\top(\mathbf{W}_o \mathrm{f}_o(\mathbf{s}_t^{\mathrm{LM}},\mathbf{s}_t^{\mathrm{TM}},\mathbf{y}_{t-1},\mathbf{c}_t)+\mathbf{b}_o))$$

Using Monolingual Corpora via Backtranslation [Sennrich et al. 2015]



Backtranslation

- Pros
  - Simple approach
  - No additional parameters

- Cons
  - Computationally expensive
    - to train an auxiliary NMT model for back-translation
    - to translate large amounts of monolingual corpora

Combining Multilingual Machine Translation and Backtranslation [Niu et al. 2018]

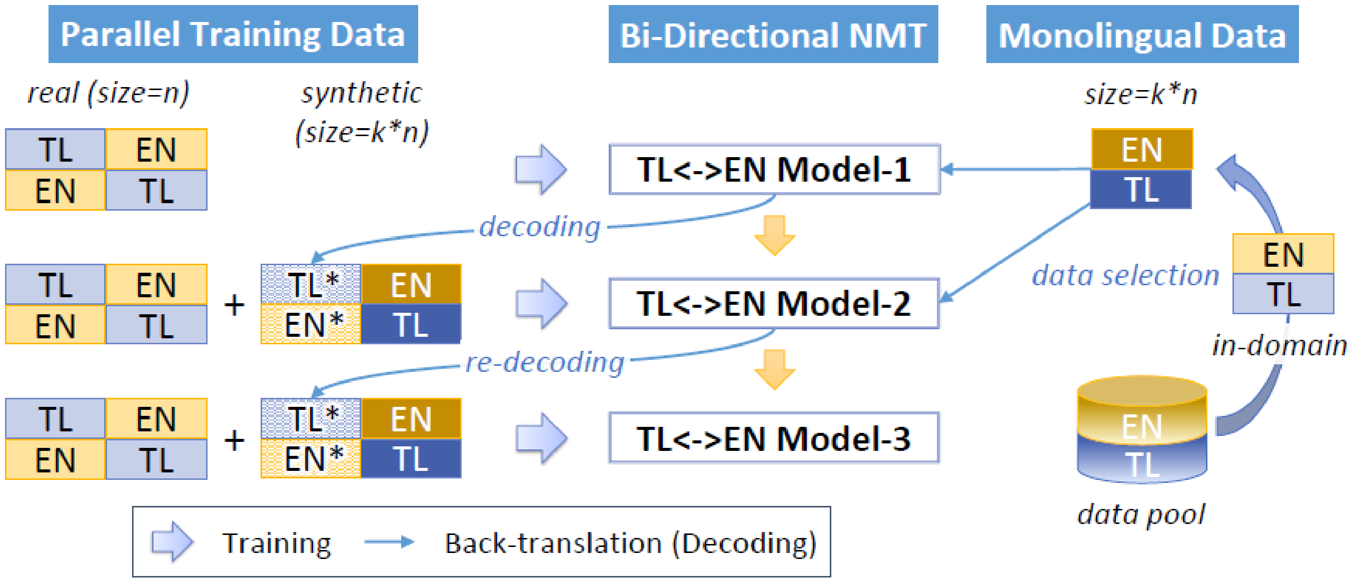


Figure 5.1: The framework of bi-directional NMT with synthetic parallel data. A bi-directional model (Model-1) is initialized on parallel data, and it translates select source and target monolingual data. Training is then continued on the augmented parallel data, leading to a cycle of improvement ($\rightarrow$ Model-2 $\rightarrow$ Model-3).

Experiments: 3 language pairs x 2 directions

Type	Dataset	# Sentences
High-resource: German↔English		
Training	Common Crawl + Europarl v7 + News Comm. v12	4,356,324
Dev	Newstest 2015+2016	5,168
Test	Newstest 2017	3,004
Mono-DE	News Crawl 2016	26,982,051
Mono-EN	News Crawl 2016	18,238,848
Low-resource: Tagalog↔English		
Training	News/Blog	50,705
Dev/Test	News/Blog	491/508
Dev/Test	Bible	500/500
Sample	Bible	61,195
Mono-TL	Common Crawl	26,788,048
Mono-EN	ICWSM 2009 blog	48,219,743
Low-resource: Swahili↔English		
Training	News/Blog	23,900
Dev/Test	News/Blog	491/509
Dev/Test	Bible	500/500
Sample	Bible	14,699
Mono-SW	Common Crawl	12,158,524
Mono-EN	ICWSM 2009 blog	48,219,743

Experiments: impact on BLEU

ID	Training Data	TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
Uni-directional models							
U-1	L1→L2	31.99	31.28	32.60	39.98	29.51	23.01
U-2	L1→L2 + L1*→L2	**24.21**	**29.68**	**25.84**	**38.29**	**33.20**	**25.41**
U-3	L1→L2 + L1→L2*	22.13	27.14	24.89	36.53	30.89	23.72
U-4	L1→L2 + L1*→L2 + L1→L2*	23.38	29.31	25.33	37.46	33.01	25.05
Bi-directional models							
ID	L1=EN	L2=TL		L2=SW		L2=DE	
B-1	L1↔L2	32.72	31.66	33.59	39.12	28.84	22.45
B-2	L1↔L2 + L1*↔L2	32.90	32.33	33.70	39.68	29.17	24.45
B-3	L1↔L2 + L2*↔L1	32.71	31.10	33.70	39.17	31.71	21.71
B-4	L1↔L2 + L1*↔L2 + L2*↔L1	33.25	32.46	**34.23**	38.97	30.43	22.54
B-5	L1↔L2 + L1*→L2 + L2*→L1	**33.41**	**33.21**	34.11	**40.24**	**31.83**	**24.61**
B-5*f*	L1↔L2 + L1*→L2 + L2*→L1	33.79	32.97	34.15	40.61	31.94	24.45
B-6*f*	L1↔L2 + L1*→L2 + L2*→L1	**34.50**	**33.73**	**34.88**	**41.53**	**32.49**	**25.20**

Table 5.2: BLEU scores for uni-directional models (ID=U-k) and bi-directional NMT models (ID=B-k) trained on different combinations of real and synthetic parallel data. Models in B-5f are fine-tuned from base models in B-1. Best models in B-6f are fine-tuned from precedent models in B-5f and underscored synthetic data is re-decoded using precedent models. The highest score within each box is highlighted.

Experiments: impact on training updates

Model		TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
Uni-directional	Baseline	76	78	63	66	41	48
	Synthetic	177	176	137	104	88	75
	TOTAL		507		371		252
Bi-directional	Baseline		125		93		61
	Synthetic		285		218		113
	TOTAL	↓ 19%	410	↓ 14%	311	↓ 31%	174
(fine-tuning)	Synthetic	↓ 23%	219	↓ 44%	122	↓ 24%	86

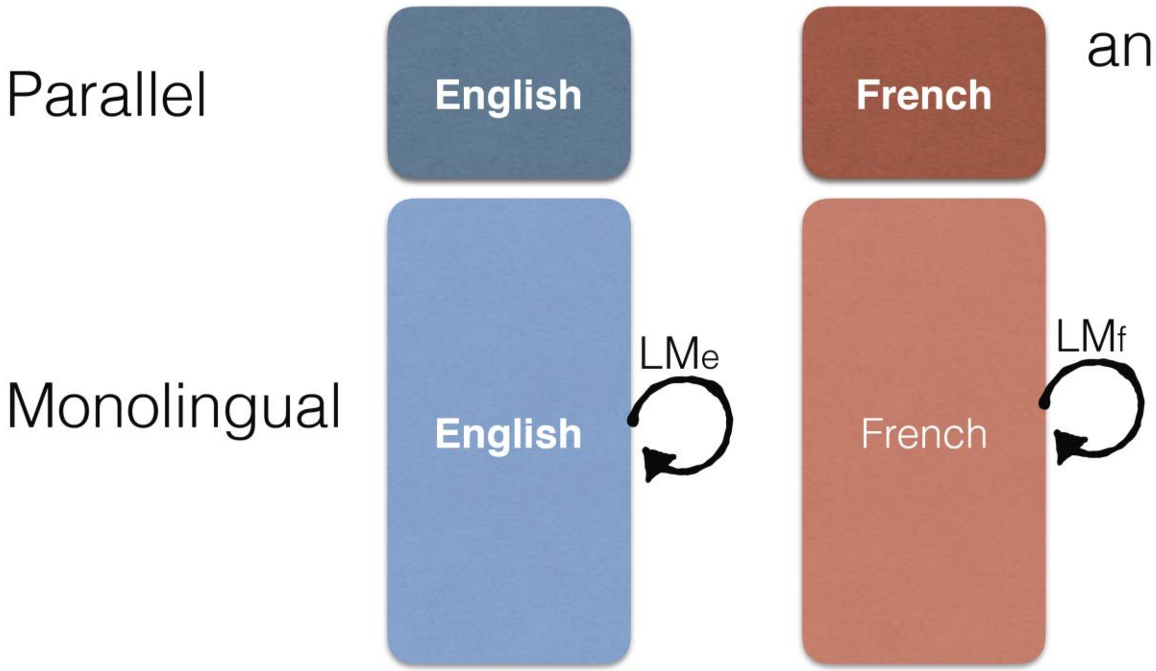
Table 5.3: Number of checkpoints (= $|\texttt{updates}|/1000$ for TL/SW↔EN or $|\texttt{updates}|/10{,}000$ for DE↔EN) used by various NMT models. Bi-directional models (with fine-tuning) reduce training time significantly.

Combining Multilingual Machine Translation and Backtranslation [Niu et al. 2018]

- A single NMT model with standard architecture performs both forward and backward translation during training
- Significantly reduces training costs compared to uni-directional systems
- Improves translation quality for low-resource language pairs

Another idea: use monolingual data to pre-train model components

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



Another idea: use monolingual data to pre-train model components

- Encoder can be pre-trained as language model
- Decoder can be pre-trained as language model
- Word embeddings can be pre-trained using word2vec or other objectives
- But impact is mixed in practice because of mismatch between pre-training and NMT objectives

3 strategies for semi-supervised neural MT

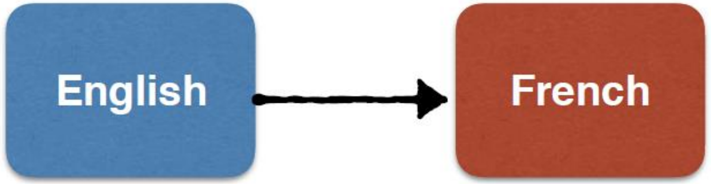
- Incorporate a target language model p(y) via shallow or deep fusion
- Create synthetic pairs (x*,y) via backtranslation
- Pre-train encoder, decoder or embeddings on monolingual data x or y

Unsupervised NMT

Translation as decipherment

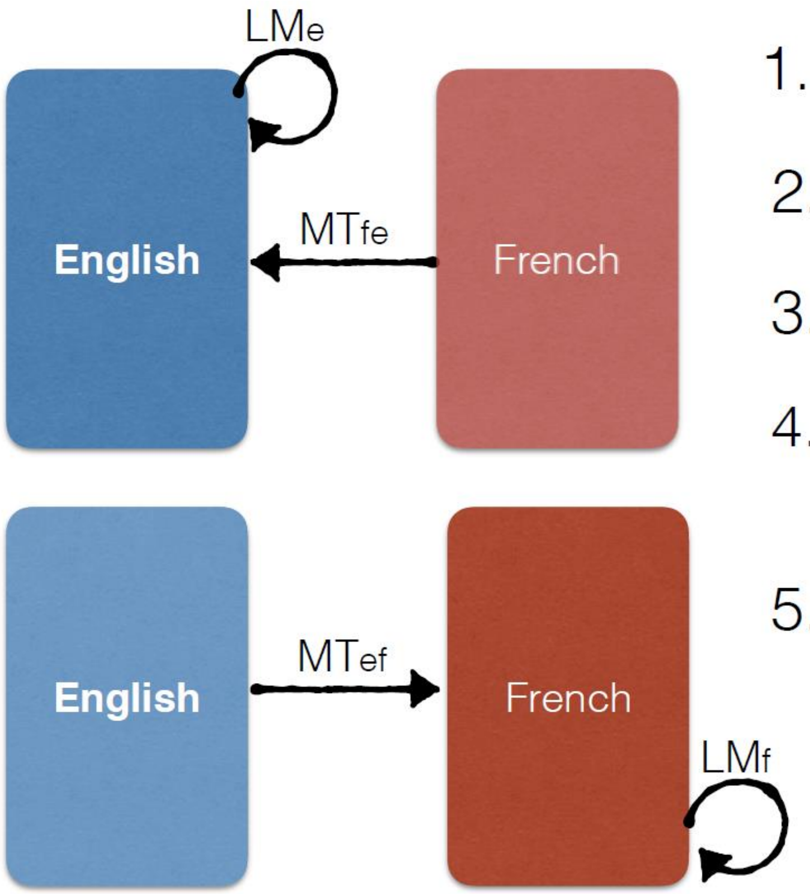
$$\arg\max_{\theta} \prod_{f} P_{\theta}(f)$$

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



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

Unsupervised Machine Translation [Lample et al.; 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

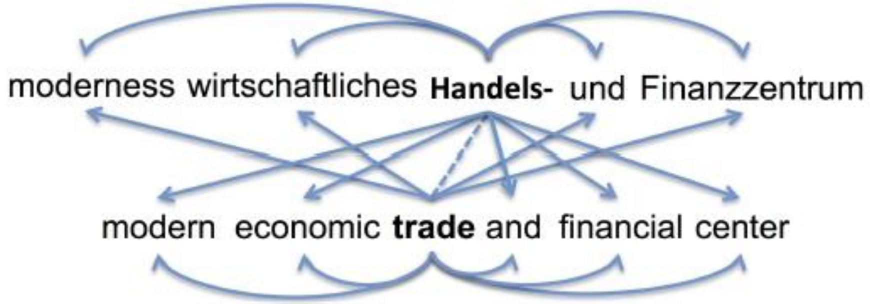
Slides credit: Antonis Anastasopoulos (CMU)

Aside: (noisy) bilingual lexicons can be induced from bilingual embeddings

- One method: bilingual skipgram model

  - put words from 2 (or more) languages into the same embedding space
  - cosine similarity can be used to find translations in the 2nd language, in addition to similar/related words in the 1st language

Aside: (noisy) bilingual lexicons can be induced from bilingual embeddings



One approach: bilingual skipgram model

Requires word aligned parallel data

Skipgram embeddings are trained to predict
- Neighbors of words w1 in language 1 (e.g., German)
- Neighbors of words w2 in language 2 (e.g., English)
- Language 1 neighbors of word w1
- Language 1 neighbors of word w2

Luong et al. (2015): https://nlp.stanford.edu/~lmthang/bivec/

Unsupervised objectives intuition: auto-encoding + back-translation

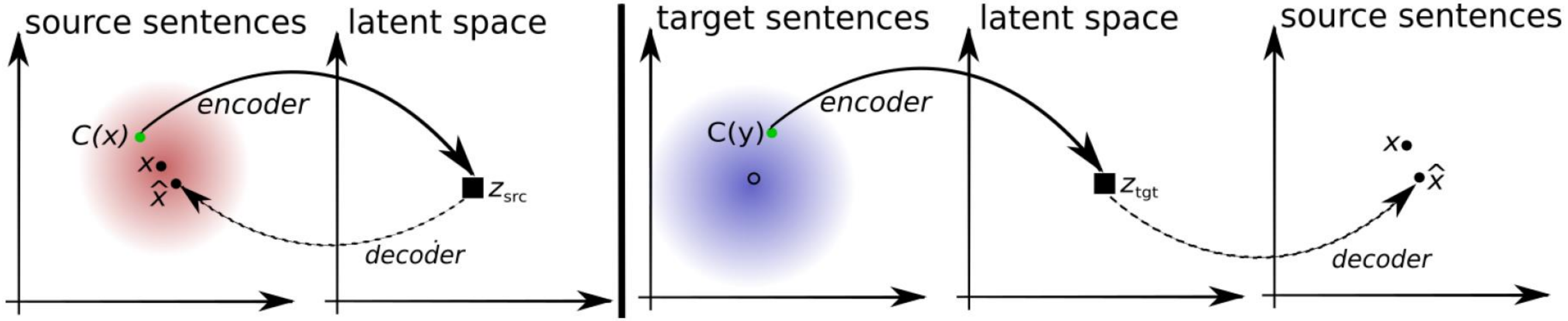


Figure from Lample et al. ICLR 2018

Experiments

	MMT1 en-fr	MMT1 de-en	WMT en-fr	WMT de-en
Monolingual sentences	14.5k	14.5k	15M	1.8M
Vocabulary size	10k / 11k	19k / 10k	67k / 78k	80k / 46k

Table 1: **Multi30k-Task1 and WMT datasets statistics.** To limit the vocabulary size in the WMT en-fr and WMT de-en datasets, we only considered words with more than 100 and 25 occurrences, respectively.

Figure from Lample et al. ICLR 2018

Experiments

	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70
oracle word reordering	11.62	24.88	18.27	6.79	10.12	20.64	19.42	11.57
Our model: 1st iteration	27.48	28.07	23.69	19.32	12.10	11.79	11.10	8.86
Our model: 2nd iteration	31.72	30.49	24.73	21.16	14.42	13.49	13.25	9.75
Our model: 3rd iteration	32.76	32.07	26.26	22.74	15.05	14.31	13.33	9.64

Table 2: **BLEU score on the Multi30k-Task1 and WMT datasets** using greedy decoding.

Figure from Lample et al. ICLR 2018

Experiments

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 .**

Table 3: **Unsupervised translations.** Examples of translations on the French-English pair of the Multi30k-Task1 dataset. Iteration 0 corresponds to word-by-word translation. After 3 iterations, the model generates very good translations.

Figure from Lample et al. ICLR 2018

Unsupervised neural MT

- Given a bilingual embeddings / translation lexicon, it is possible to train a neural MT system without examples of translated sentences!

- But current evidence is limited to simulations on high resource languages, and sometimes parallel data
  - Unclear how well results port to realistic low-resource scenarios