89688: Statistical Machine Translation

Unsupervised Neural Machine Translation

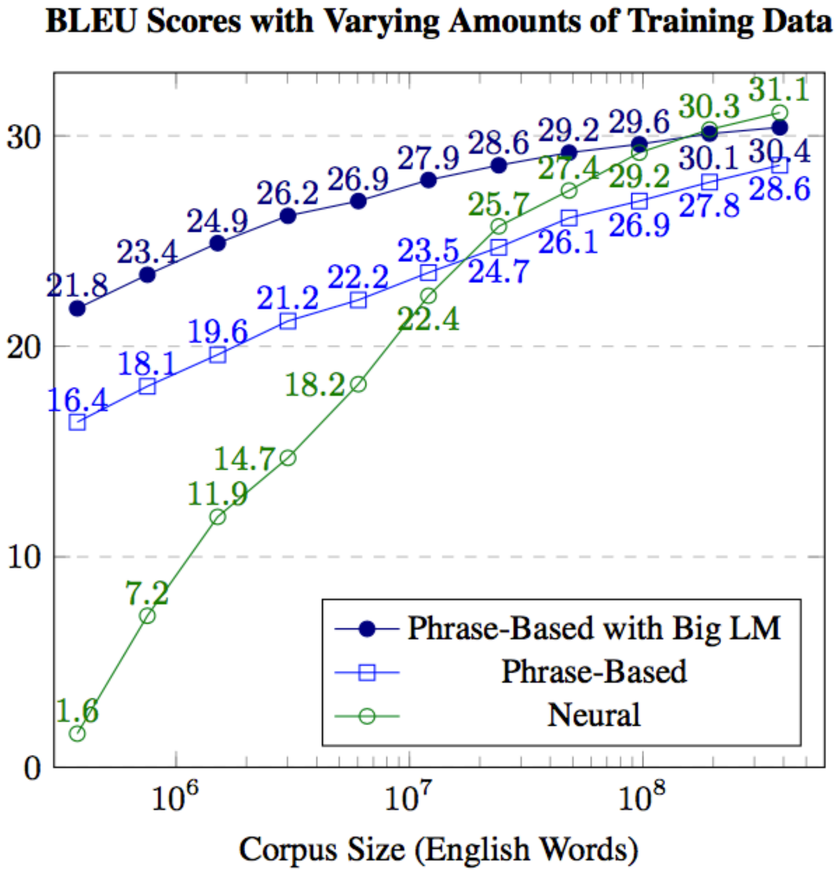
June 2020

Roee Aharoni
Computer Science Department
Bar Ilan University

NMT vs. SMT: Sample Efficiency

NMT vs. SMT: Sample Efficiency

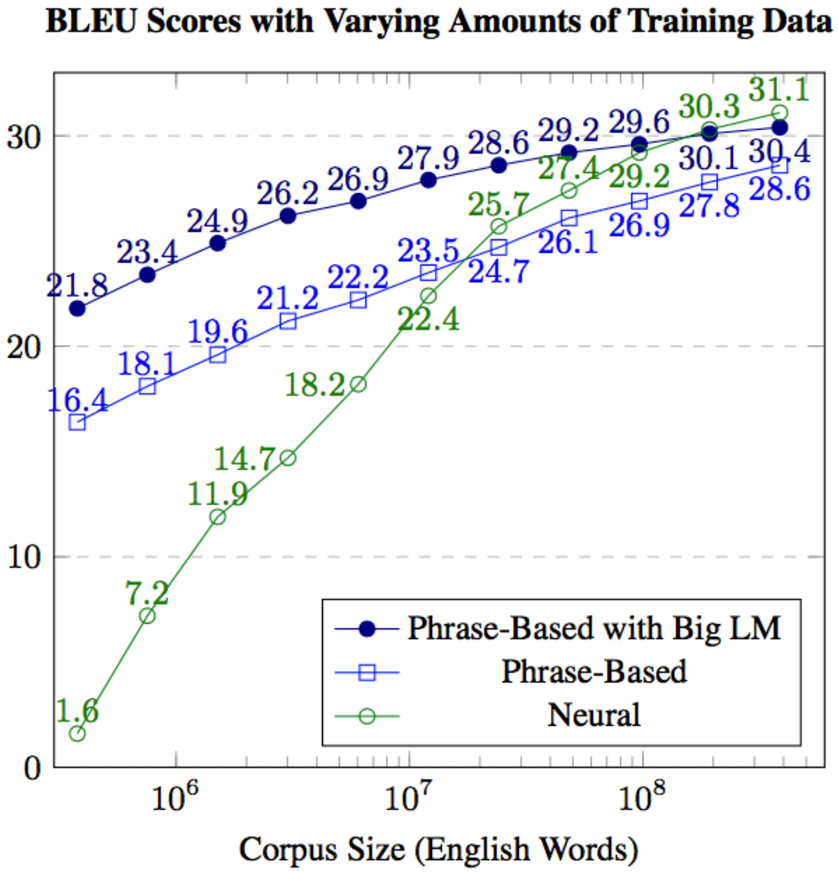
- NMT is better than SMT only when given >10m parallel words



Koehn & Knowles, 2017

NMT vs. SMT: Sample Efficiency

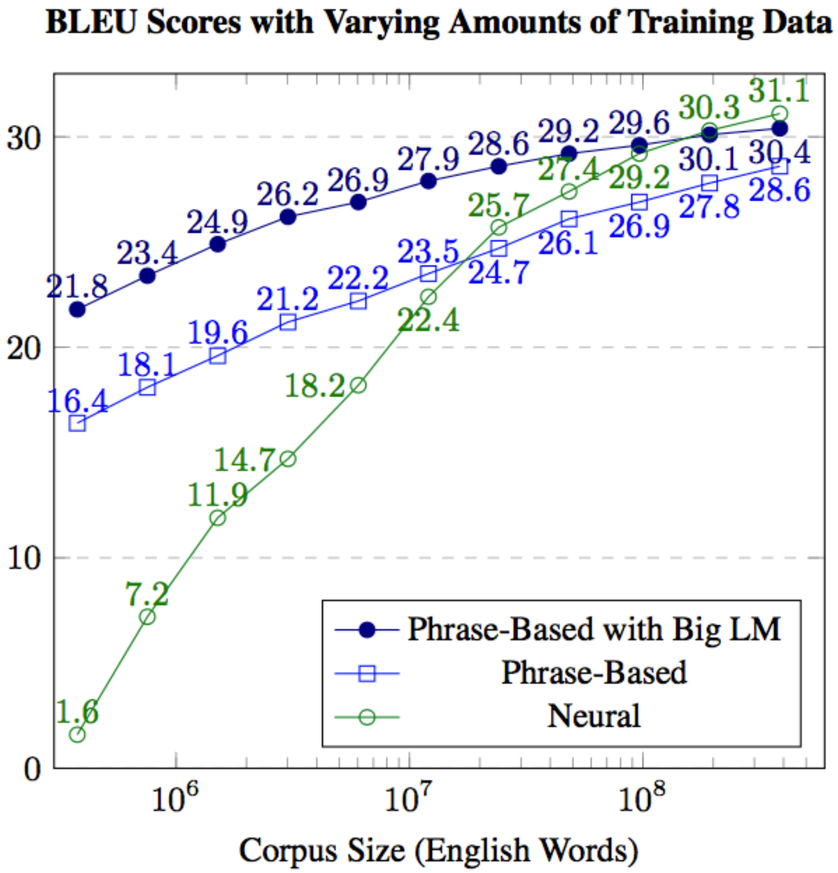
- NMT is better than SMT only when given >10m parallel words
- NMT is better than “Semi Supervised” SMT (SMT + a large language model) only when given >100m parallel words



Koehn & Knowles, 2017

NMT vs. SMT: Sample Efficiency

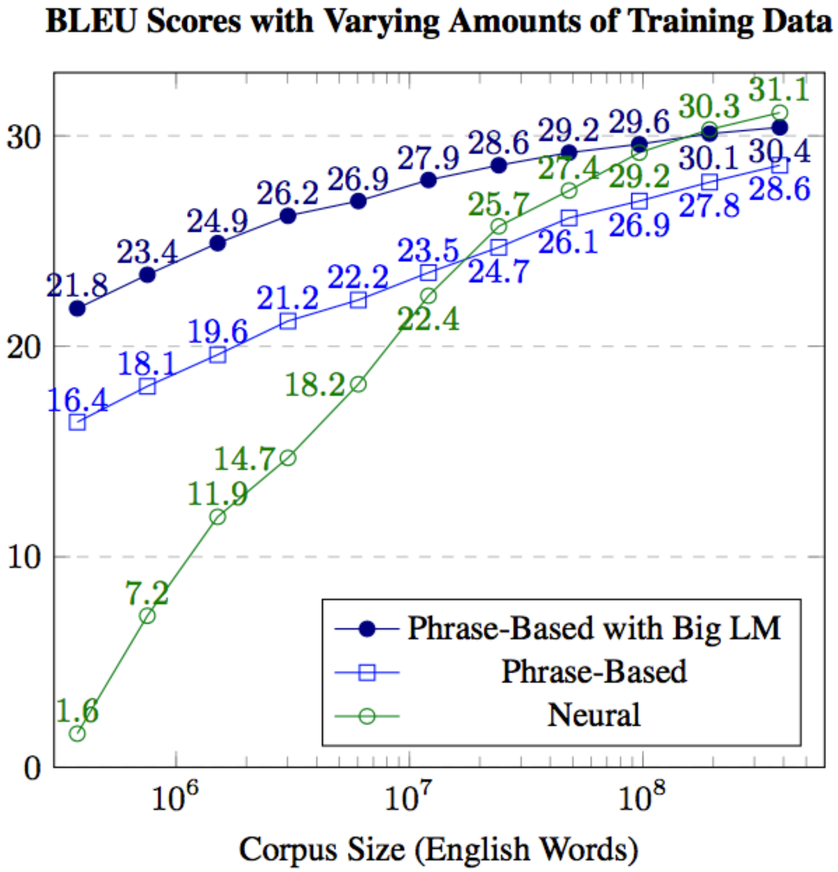
- NMT is better than SMT only when given >10m parallel words
- NMT is better than “Semi Supervised” SMT (SMT + a large language model) only when given >100m parallel words
- But getting parallel data is expensive!



Koehn & Knowles, 2017

NMT vs. SMT: Sample Efficiency

- NMT is better than SMT only when given >10m parallel words
- NMT is better than “Semi Supervised” SMT (SMT + a large language model) only when given >100m parallel words
- But getting parallel data is expensive!
- Can we do well using only **monolingual data?**



Koehn & Knowles, 2017

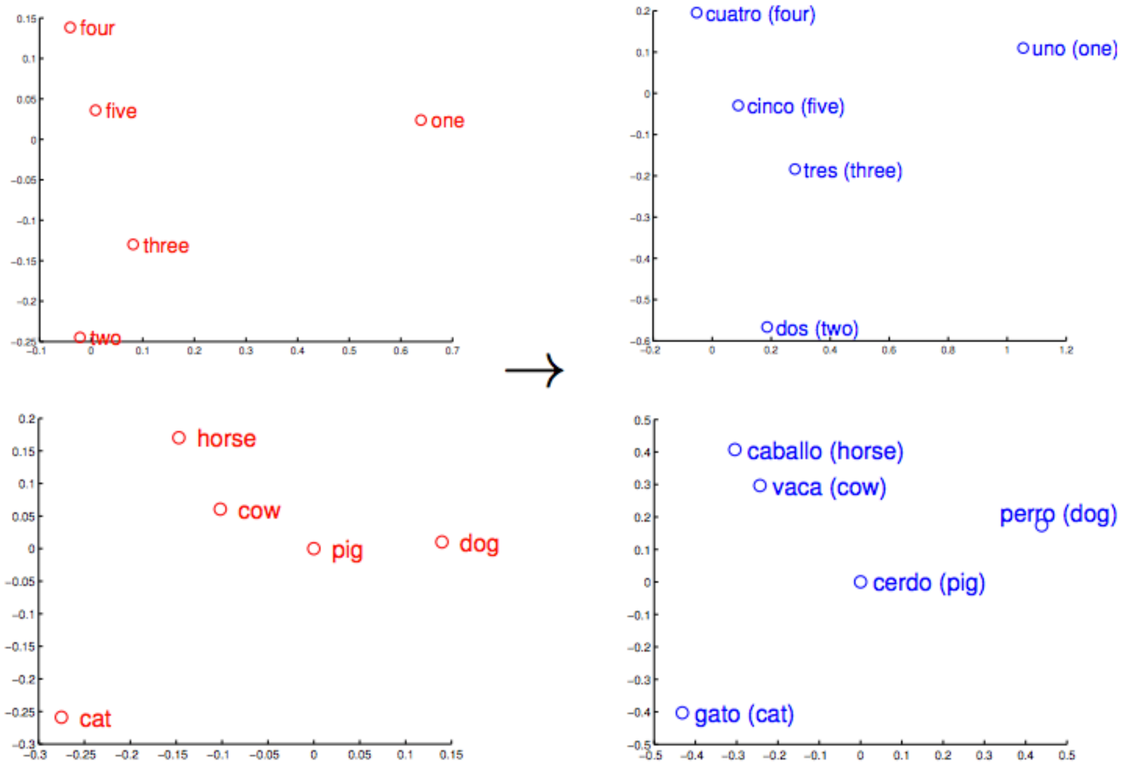
Motivation: Mikolov et al. 2013

Motivation: Mikolov et al. 2013

- “Exploiting Similarities among Languages for Machine Translation” - Mikolov, Le & Sutskever, 2013

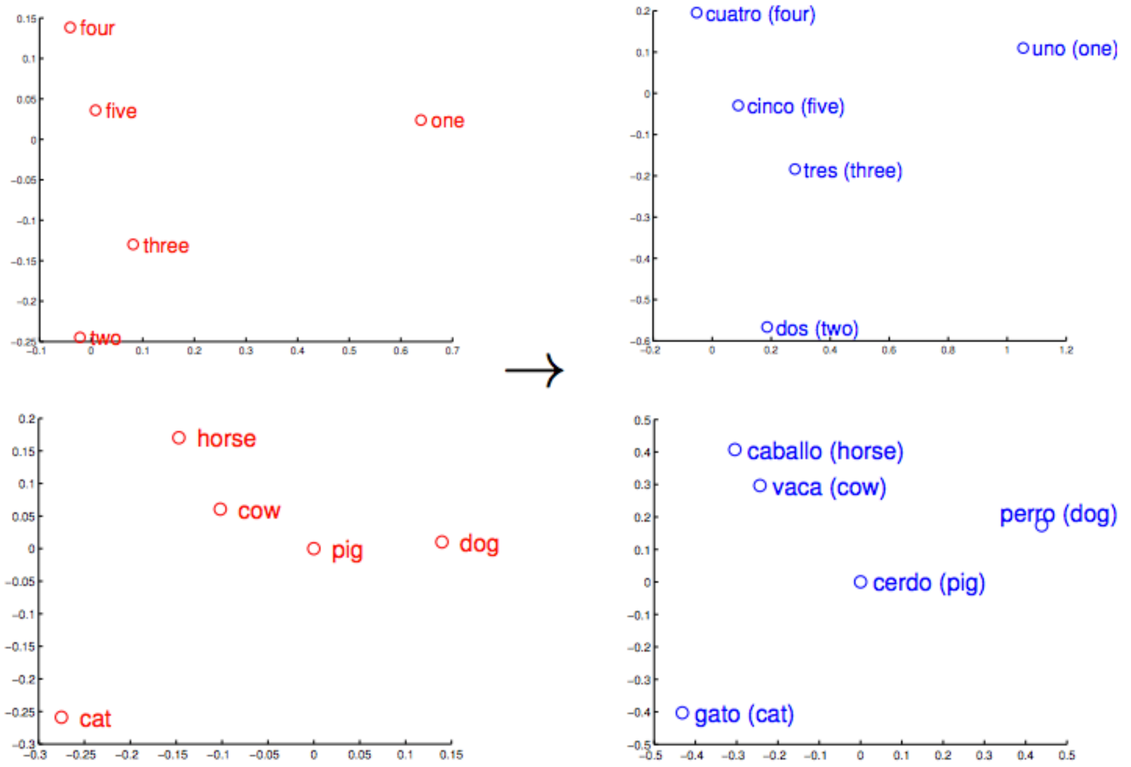
Motivation: Mikolov et al. 2013

- “Exploiting Similarities among Languages for Machine Translation” - Mikolov, Le & Sutskever, 2013
- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation



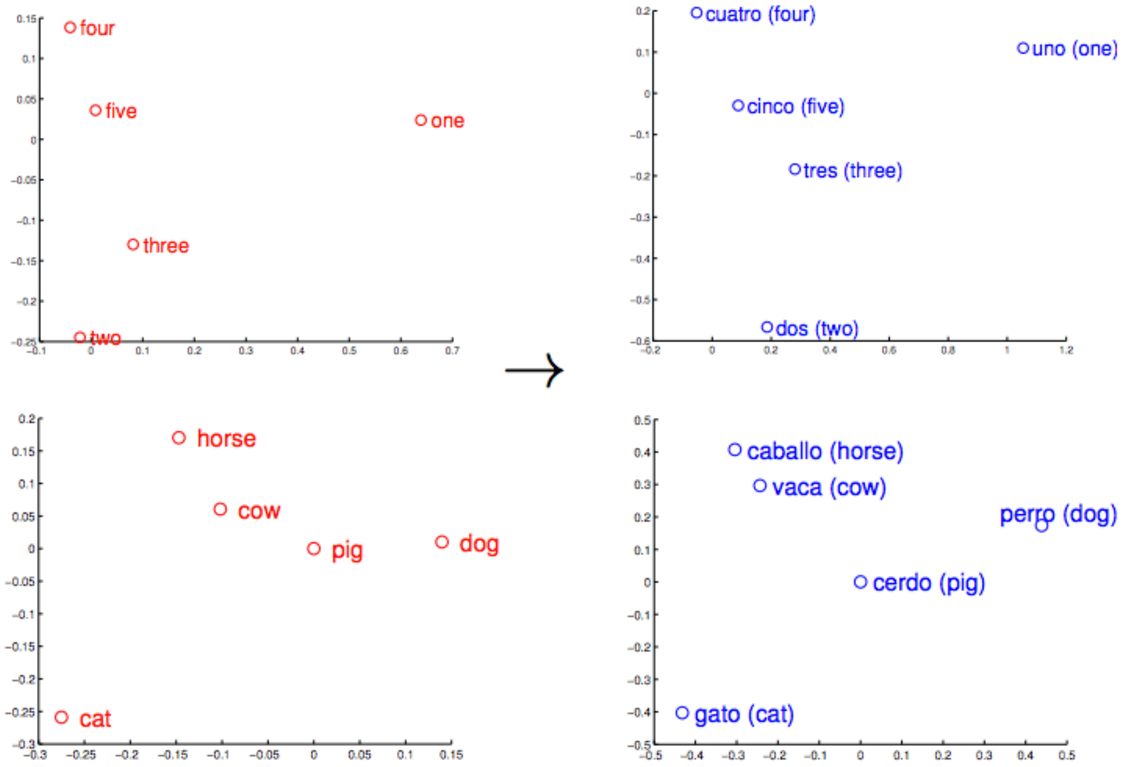
Motivation: Mikolov et al. 2013

- “Exploiting Similarities among Languages for Machine Translation” - Mikolov, Le & Sutskever, 2013
- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation
- Learned a rotation matrix to translate words from one embedding space to another with some success



Motivation: Mikolov et al. 2013

- “Exploiting Similarities among Languages for Machine Translation” - Mikolov, Le & Sutskever, 2013
- Observed a **similar structure in unsupervised embedding spaces of different languages**, after rotation
- Learned a rotation matrix to translate words from one embedding space to another with some success
- Weakly supervised - requires a small dictionary (5000 entries)



Unsupervised NMT: A Tale of Two Papers

Unsupervised NMT: A Tale of Two Papers

- Both submitted to ICLR 2018 with critical acclaim (October 2017)

Unsupervised NMT: A Tale of Two Papers

- Both submitted to ICLR 2018 with critical acclaim (October 2017)
- Similar motivations - both try to tackle:

Unsupervised NMT: A Tale of Two Papers

- Both submitted to ICLR 2018 with critical acclaim (October 2017)
- Similar motivations - both try to tackle:
  - **Structure/Fluency** - how to determine the correct word order in the output?

Unsupervised NMT: A Tale of Two Papers

- Both submitted to ICLR 2018 with critical acclaim (October 2017)
- Similar motivations - both try to tackle:
  - **Structure/Fluency** - how to determine the correct word order in the output?
  - **Semantics/Adequacy** - how to pick the correct translations given the source?

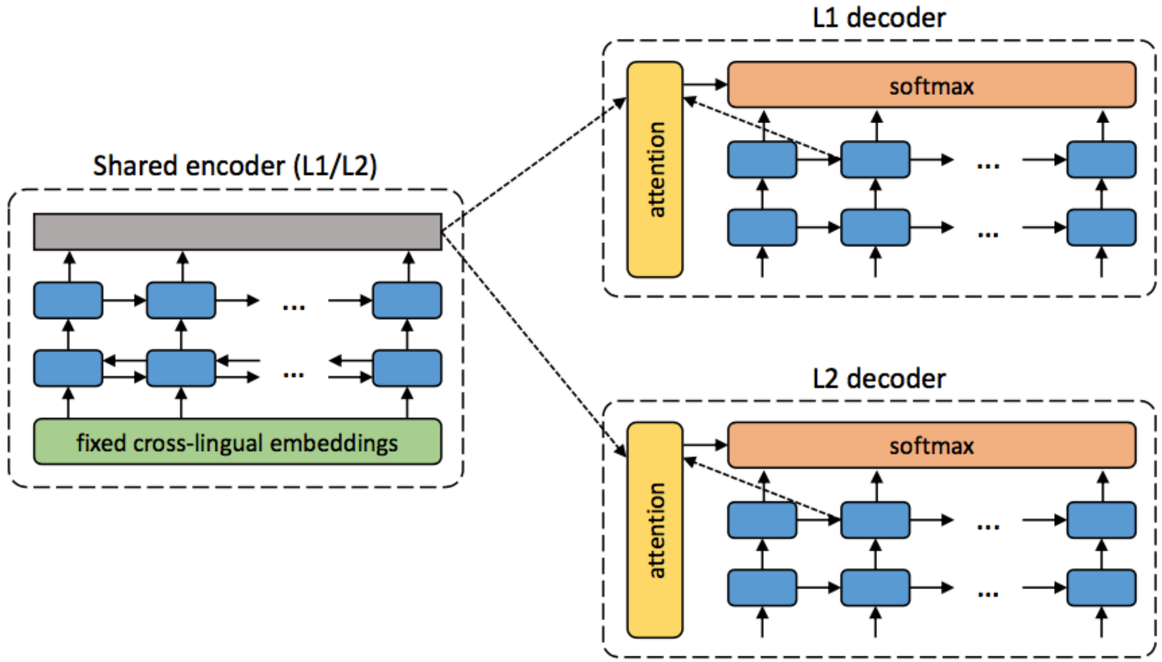
Unsupervised NMT: A Tale of Two Papers

- Both submitted to ICLR 2018 with critical acclaim (October 2017)
- Similar motivations - both try to tackle:
  - **Structure/Fluency** - how to determine the correct word order in the output?
  - **Semantics/Adequacy** - how to pick the correct translations given the source?
- Very similar modeling tricks (with slight differences)

Paper I: Artetxe, Agirre & Cho

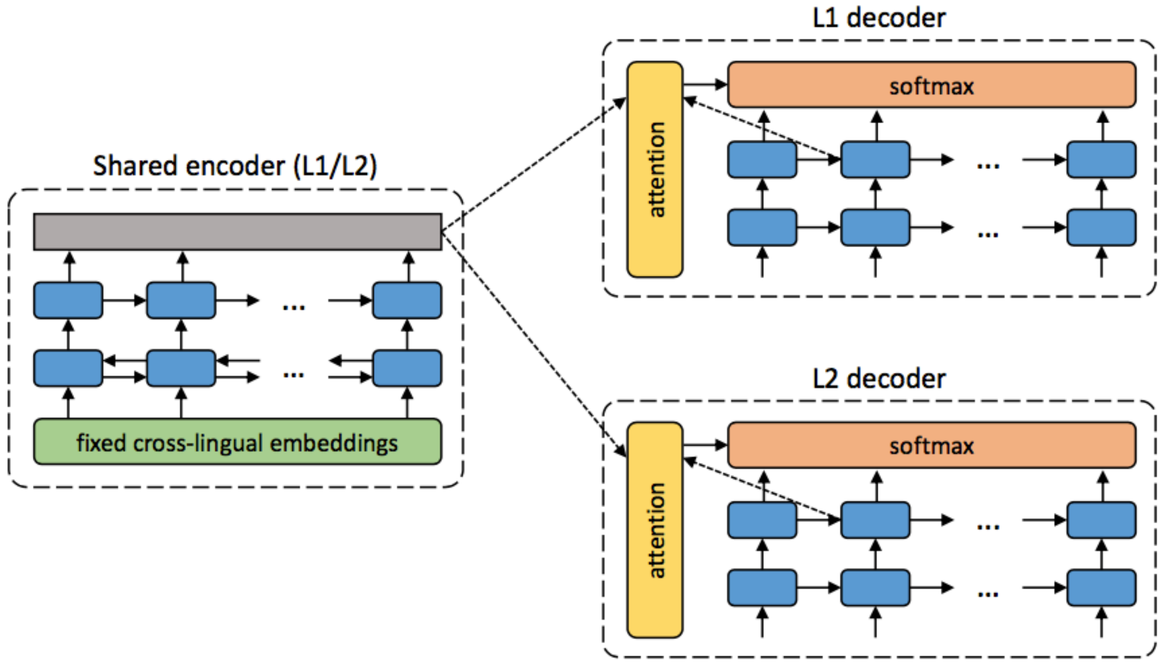
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**



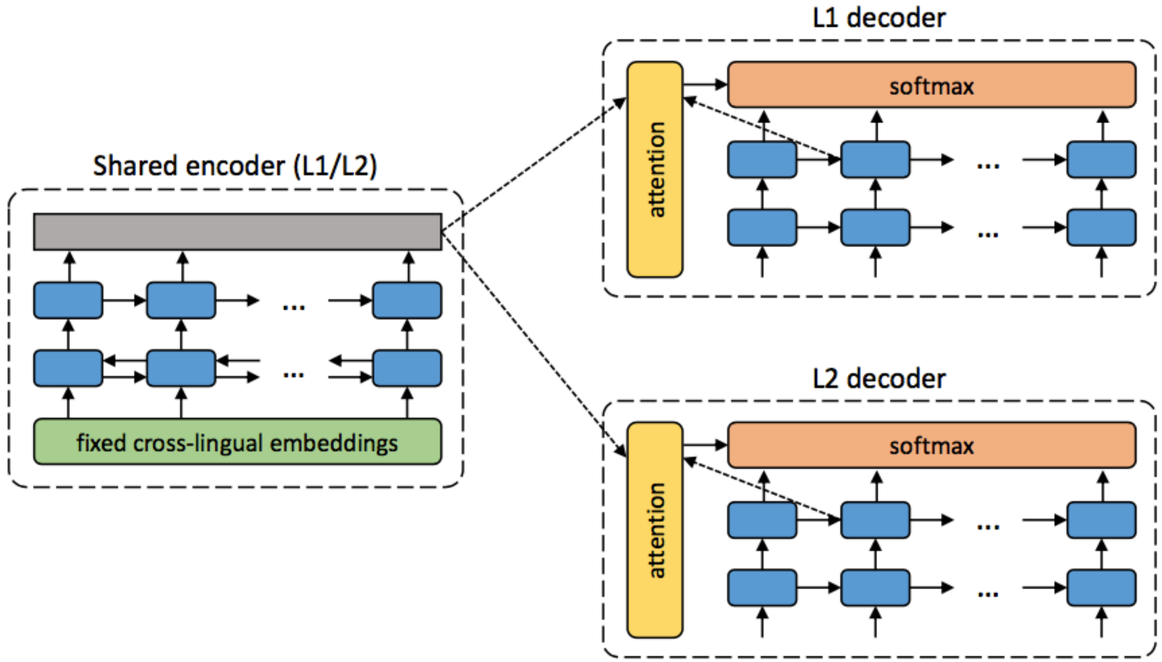
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders



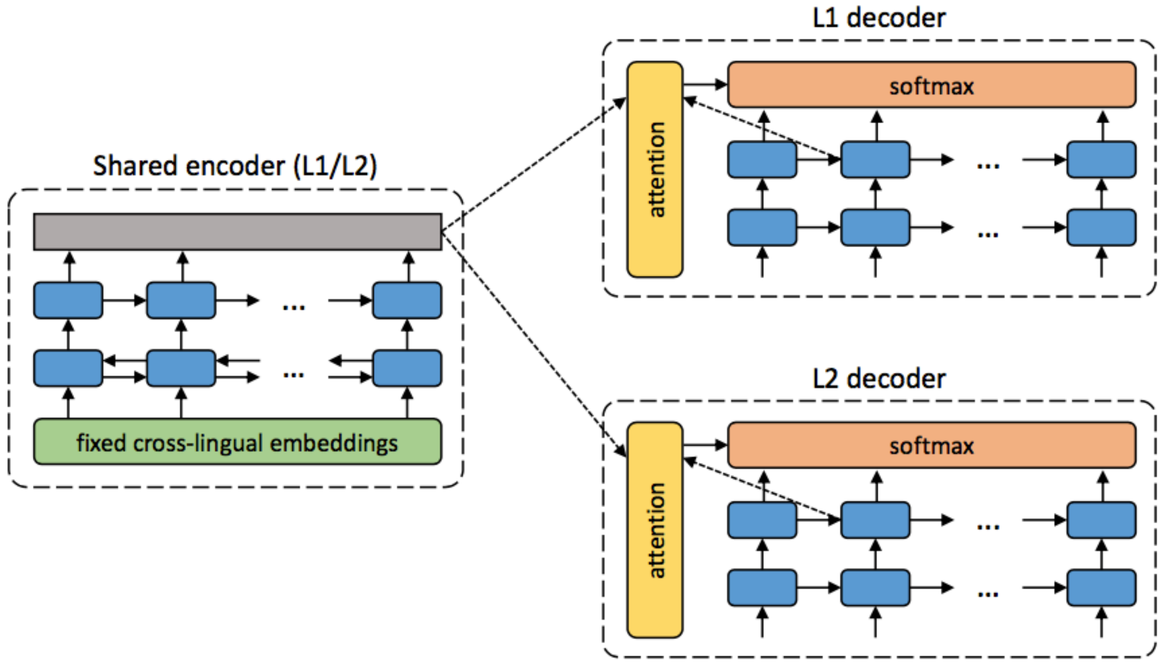
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders
  - Attention



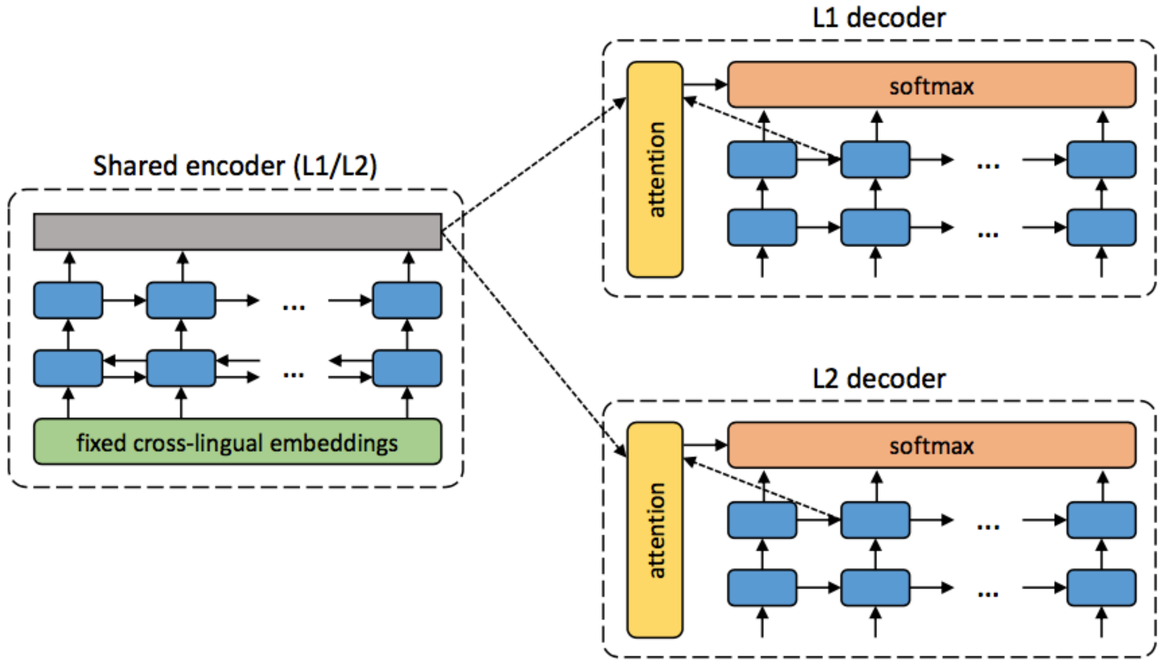
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders
  - Attention
- **Main “Tricks”:**



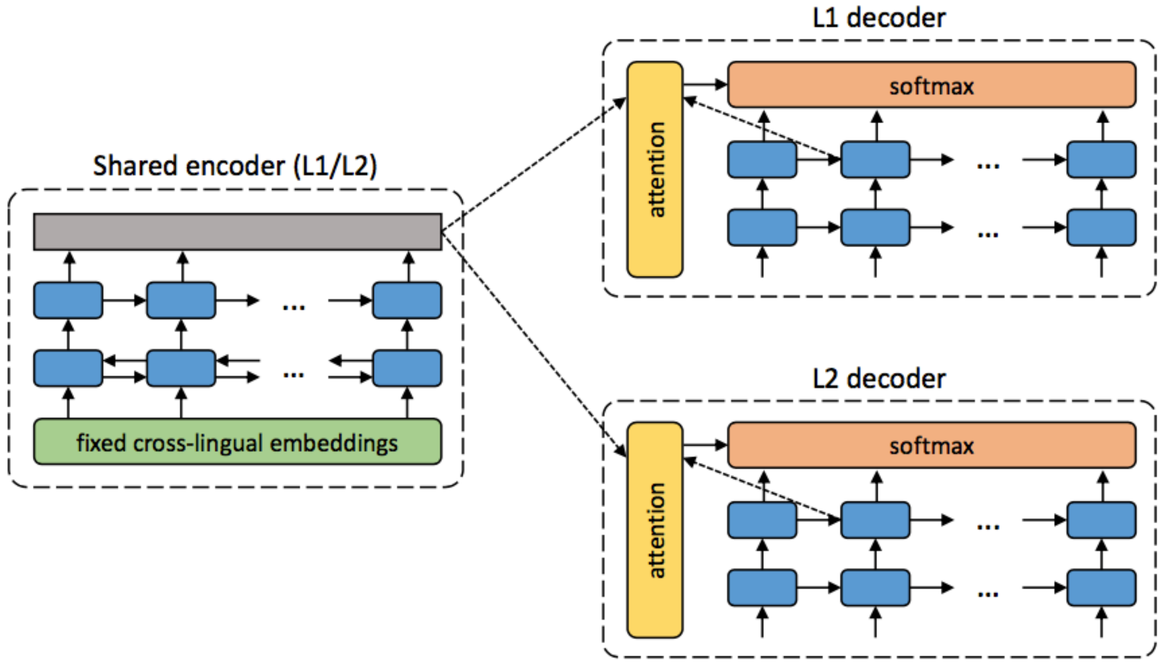
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders
  - Attention
- **Main "Tricks":**
  - Fixed, unsupervised cross-lingual embeddings (**Adequacy**)



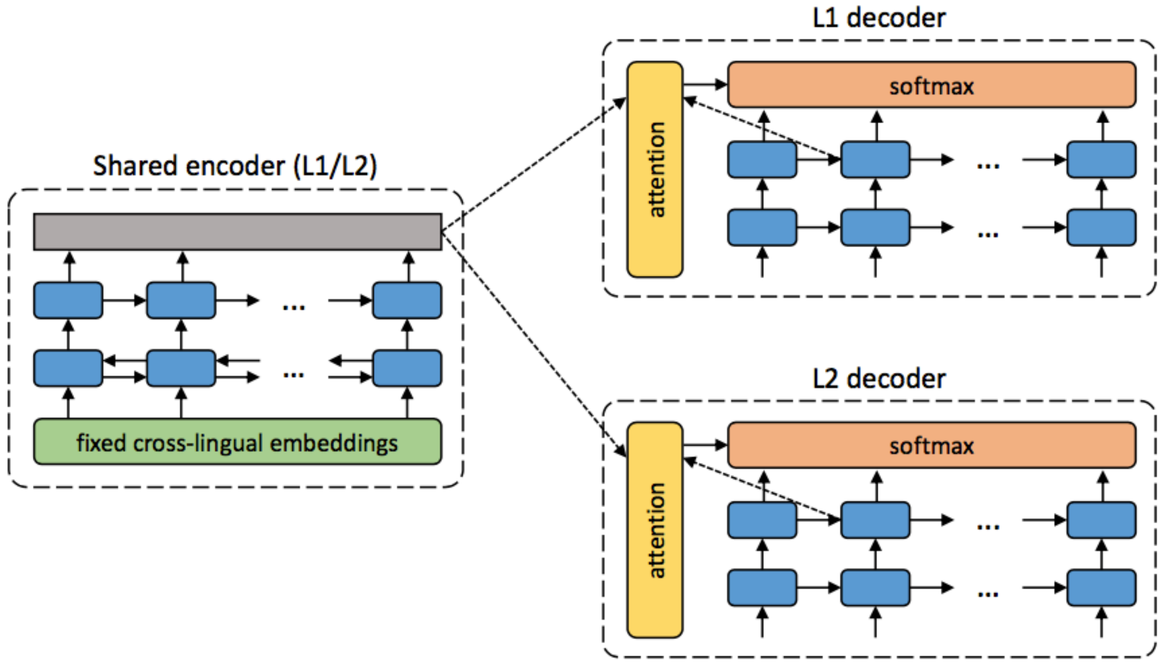
Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders
  - Attention
- **Main "Tricks":**
  - Fixed, unsupervised cross-lingual embeddings (**Adequacy**)
  - Backtranslation loss (**Adequacy**)



Paper I: Artetxe, Agirre & Cho

- **Model Architecture:**
  - **Shared** GRU encoder, **Separate** GRU decoders
  - Attention
- **Main "Tricks":**
  - Fixed, unsupervised cross-lingual embeddings (**Adequacy**)
  - Backtranslation loss (**Adequacy**)
  - Denoising auto-encoder loss (**Fluency**)



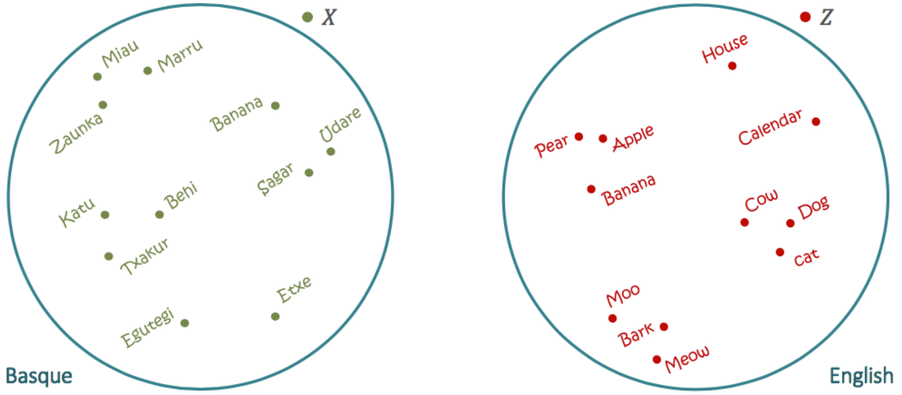
Unsupervised Cross-Lingual Word Embeddings

Unsupervised Cross-Lingual Word Embeddings

- Artetxe, Labake & Agirre, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

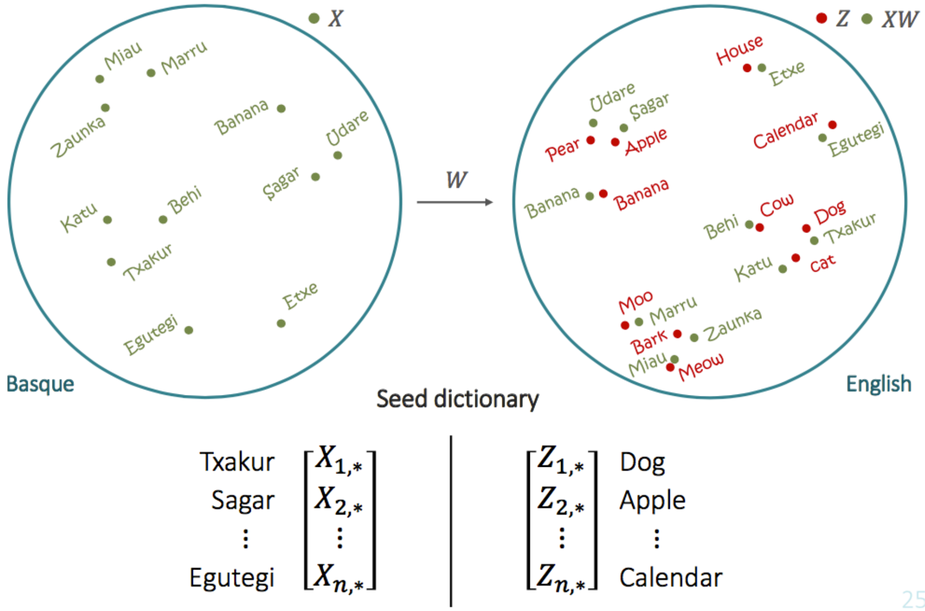
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

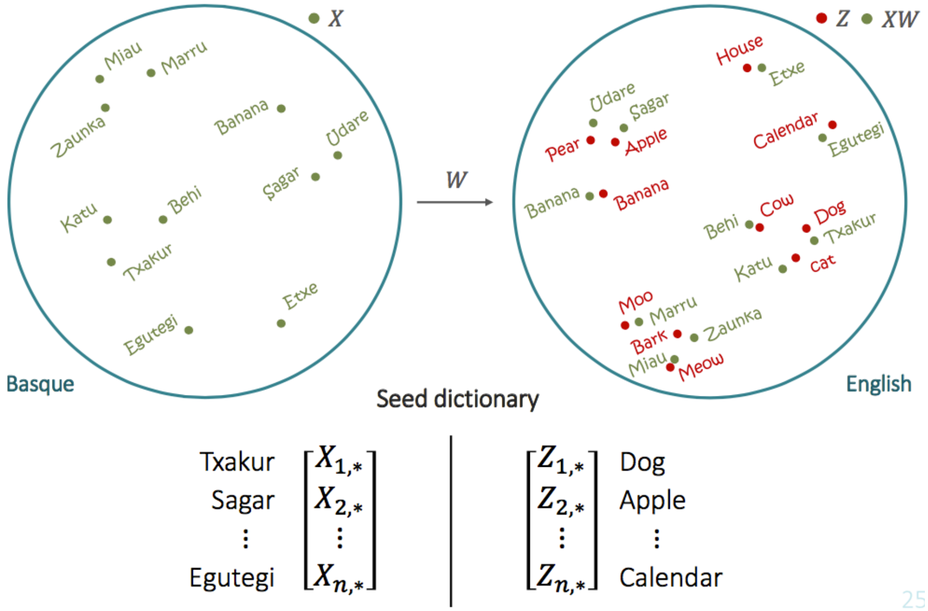
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

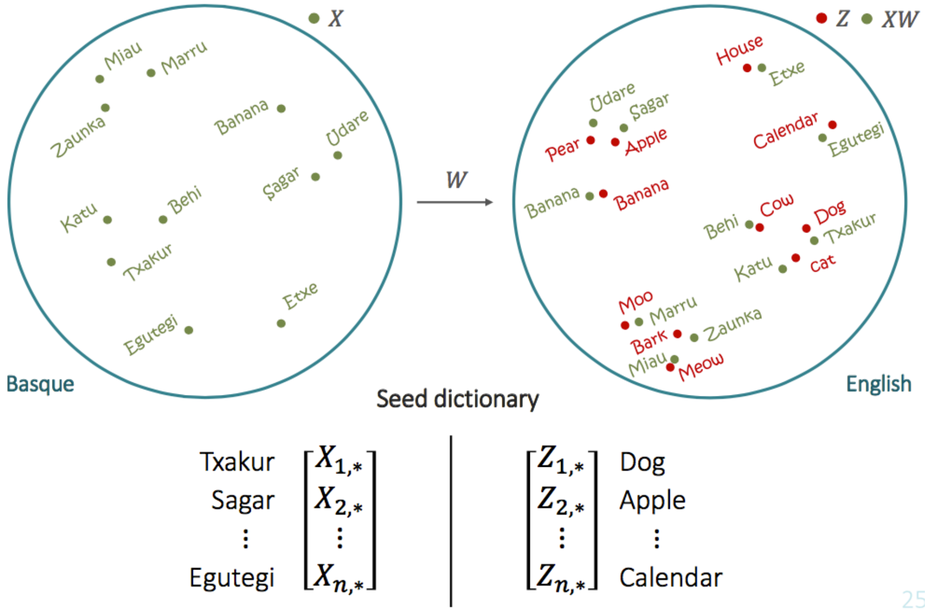
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:
  - Start with a **seed dictionary**. Clever idea: use numerals (5-5, 1989-1989…) as seed dictionary - fully unsupervised
  - Optimize the mapping W w.r.t the dictionary: $\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

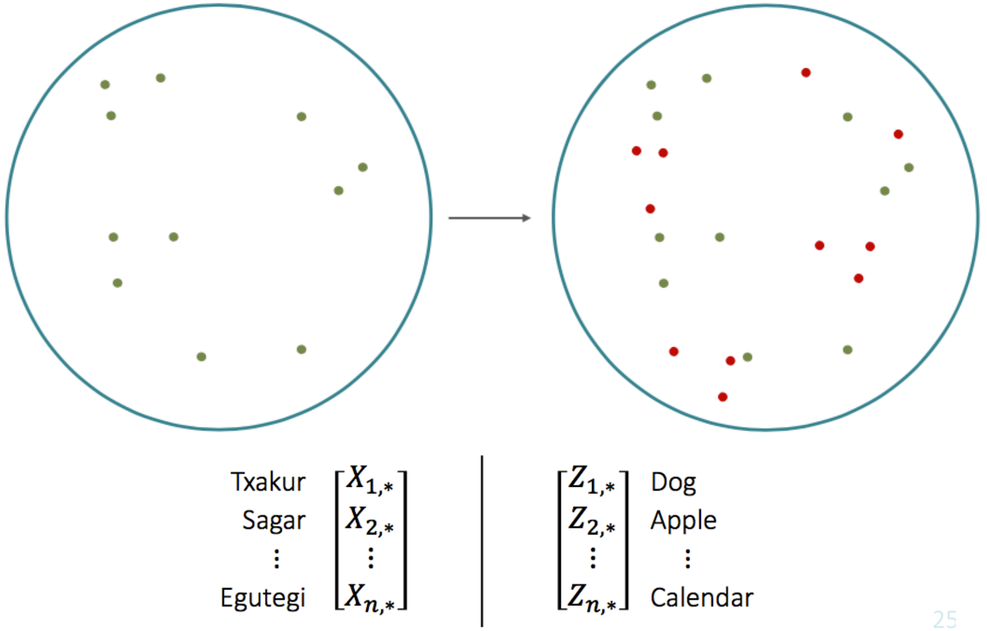
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:
  - Start with a **seed dictionary**. Clever idea: use numerals (5-5, 1989-1989…) as seed dictionary - fully unsupervised
  - Optimize the mapping W w.r.t the dictionary: $\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$
- Extract a new dictionary and **repeat iteratively** until a convergence threshold is met



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

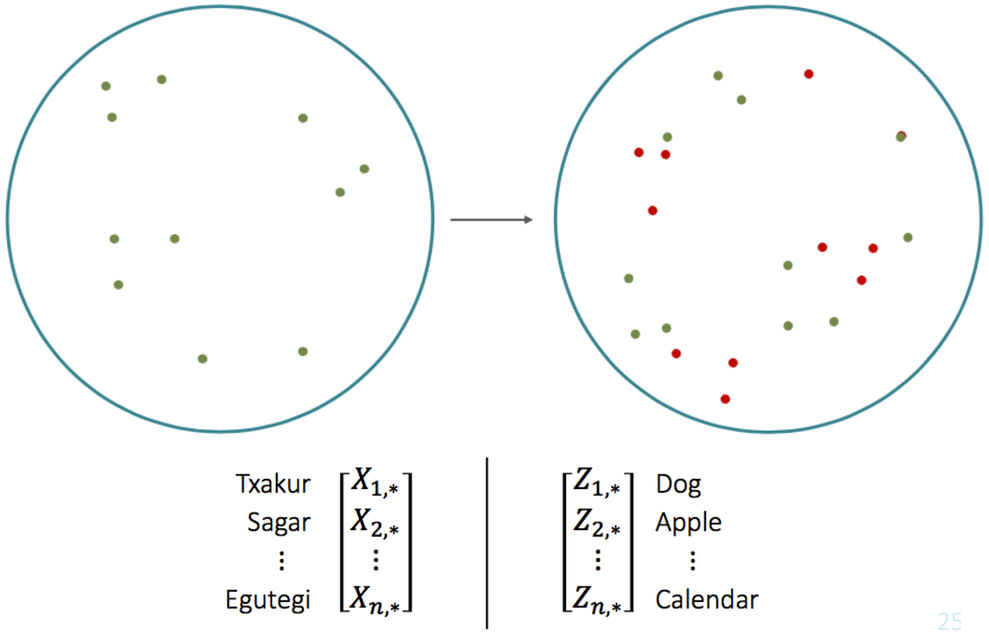
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:
  - Start with a **seed dictionary**. Clever idea: use numerals (5-5, 1989-1989…) as seed dictionary - fully unsupervised
  - Optimize the mapping W w.r.t the dictionary: $\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$
- Extract a new dictionary and **repeat iteratively** until a convergence threshold is met



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

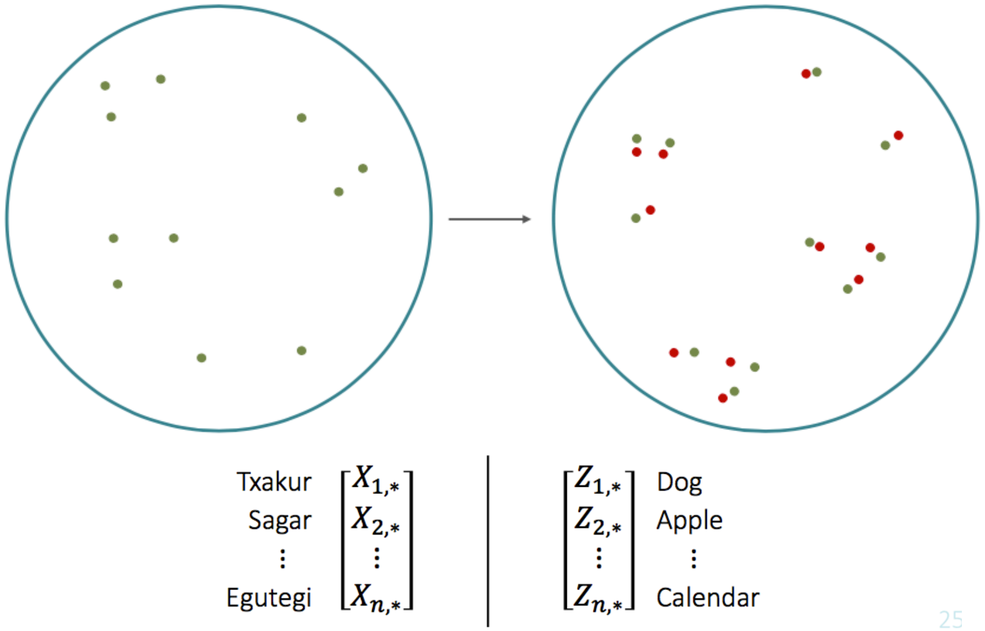
- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:
  - Start with a **seed dictionary**. Clever idea: use numerals (5-5, 1989-1989…) as seed dictionary - fully unsupervised
  - Optimize the mapping W w.r.t the dictionary: $\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$
- Extract a new dictionary and **repeat iteratively** until a convergence threshold is met



From Artetxe, ACL 2017

Unsupervised Cross-Lingual Word Embeddings

- Artetxe, Labake & Agirre, ACL 2017
- Start with monolingual embedding spaces in two languages (trained using word2vec)
- Learn a linear mapping from one language to the other:
  - Start with a **seed dictionary**. Clever idea: use numerals (5-5, 1989-1989…) as seed dictionary - fully unsupervised
  - Optimize the mapping W w.r.t the dictionary: $\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$
- Extract a new dictionary and **repeat iteratively** until a convergence threshold is met



From Artetxe, ACL 2017

Learning Semantics: Back-Translation

Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don’t have parallel data

Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don’t have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)

Learning Semantics: Back-Translation

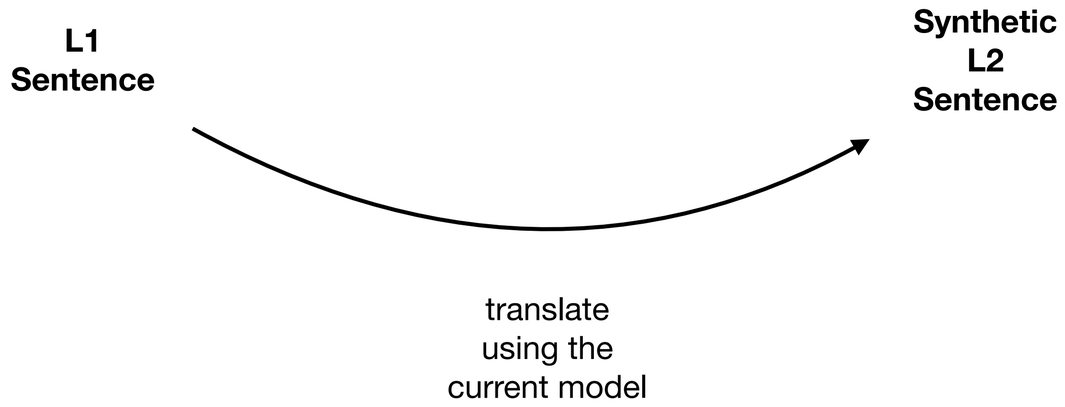
- We need to learn a mapping from one language (L1) into another (L2), but we don’t have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)
- Use the synthetic data for training using cross entropy loss

Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don’t have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)
- Use the synthetic data for training using cross entropy loss

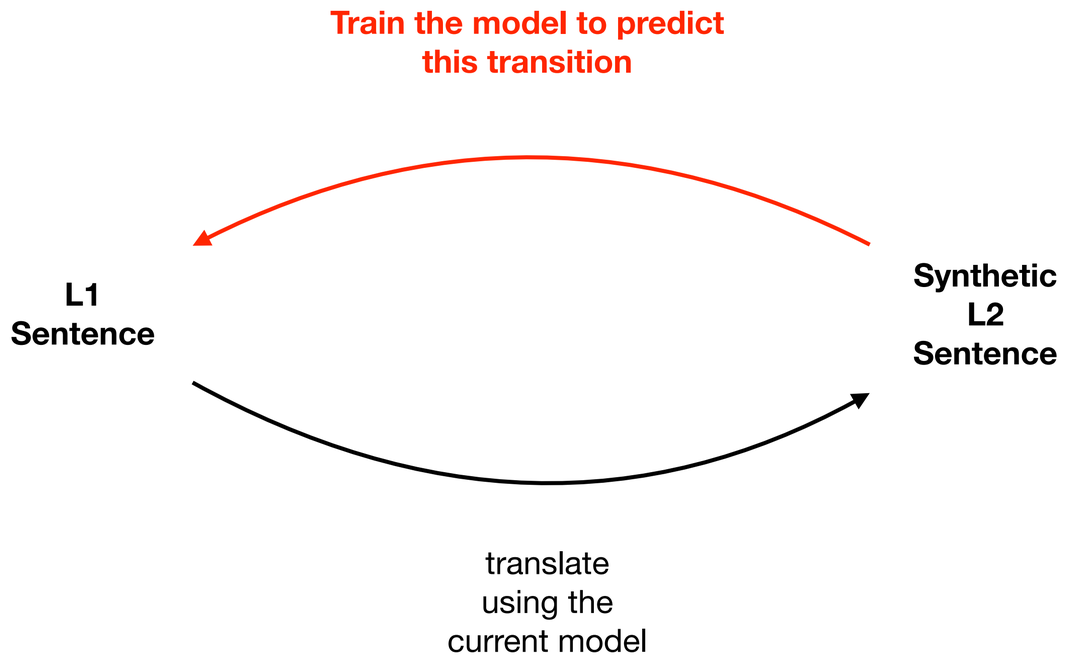
Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don't have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)
- Use the synthetic data for training using cross entropy loss



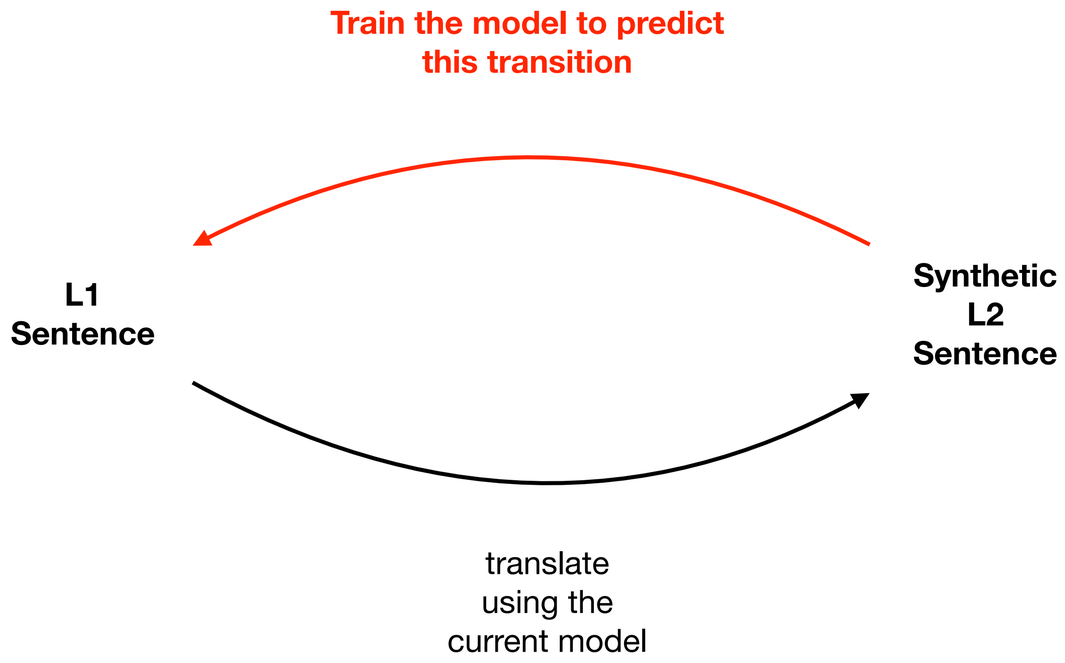
Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don't have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)
- Use the synthetic data for training using cross entropy loss



Learning Semantics: Back-Translation

- We need to learn a mapping from one language (L1) into another (L2), but we don't have parallel data
- Solution: create **synthetic** parallel data by translating with the current model (possible since the model is bidirectional)
- Use the synthetic data for training using cross entropy loss
- **This is not entirely useless** since the cross-lingual embeddings do carry some alignment signal



Learning Structure: De-noising Auto-Encoders

Learning Structure: De-noising Auto-Encoders

- The decoder needs to learn how to organize the words on the target side

Learning Structure: De-noising Auto-Encoders

- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding

Learning Structure: De-noising Auto-Encoders

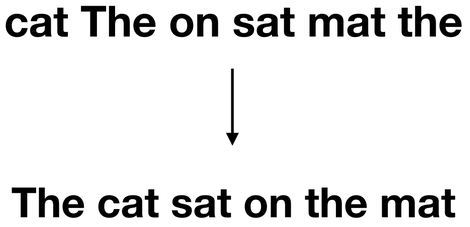
- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding
    - But this would lead it to learn trivial copying!

Learning Structure: De-noising Auto-Encoders

- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding
  - But this would lead it to learn trivial copying!
- **Introduce “noise”** by randomly swapping adjacent words (N/2 times) in the input, to force the decoder to learn word ordering

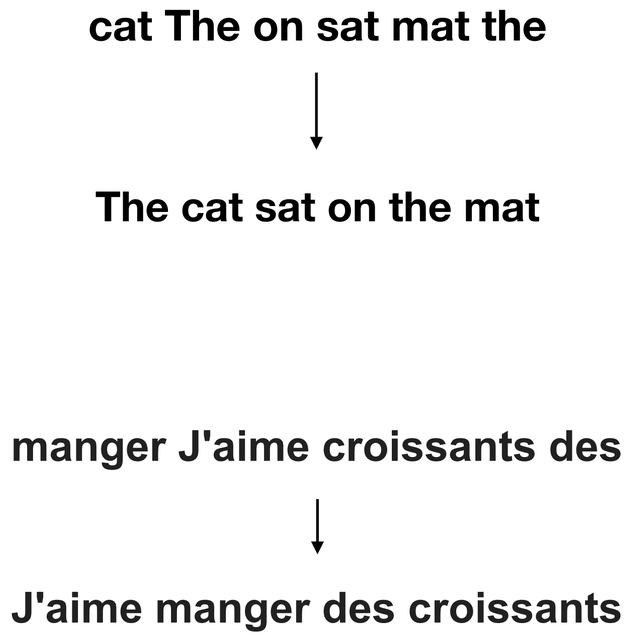
Learning Structure: De-noising Auto-Encoders

- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding
  - But this would lead it to learn trivial copying!
- **Introduce “noise”** by randomly swapping adjacent words (N/2 times) in the input, to force the decoder to learn word ordering



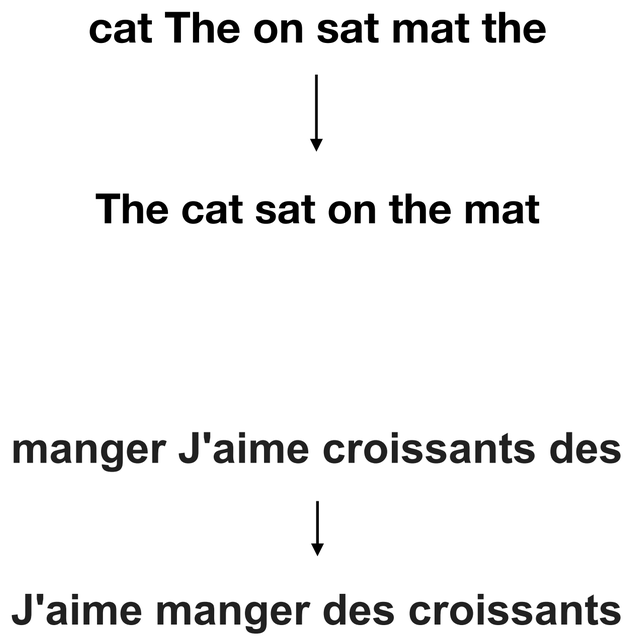
Learning Structure: De-noising Auto-Encoders

- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding
  - But this would lead it to learn trivial copying!
- **Introduce “noise”** by randomly swapping adjacent words (N/2 times) in the input, to force the decoder to learn word ordering



Learning Structure: De-noising Auto-Encoders

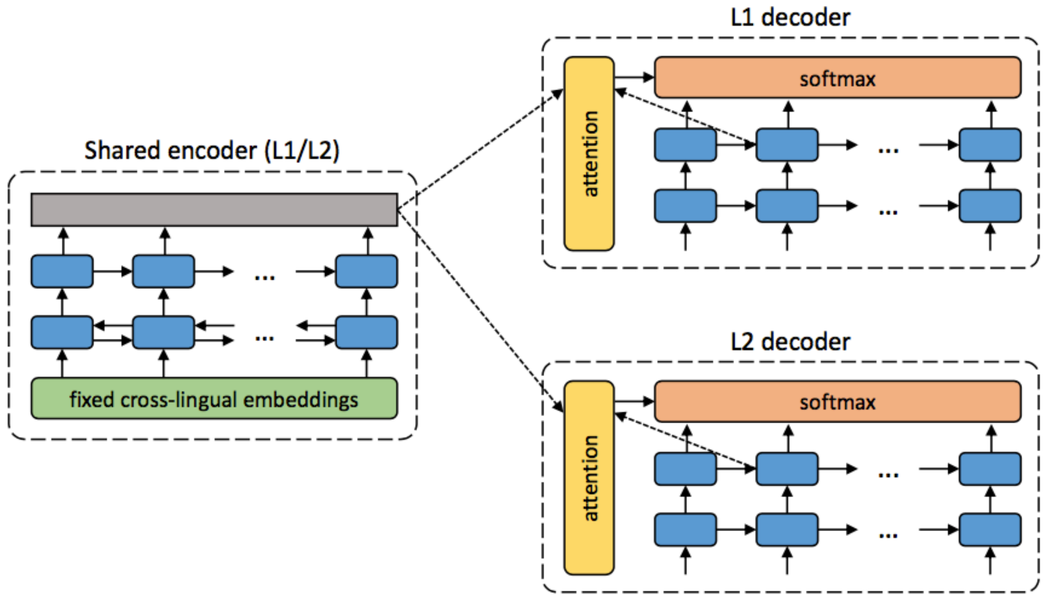
- The decoder needs to learn how to organize the words on the target side
- We could train it to predict a sentence given itself - auto-encoding
  - But this would lead it to learn trivial copying!
- **Introduce “noise”** by randomly swapping adjacent words (N/2 times) in the input, to force the decoder to learn word ordering
- Train using conventional cross entropy loss



Putting It All Together: Iterative Training

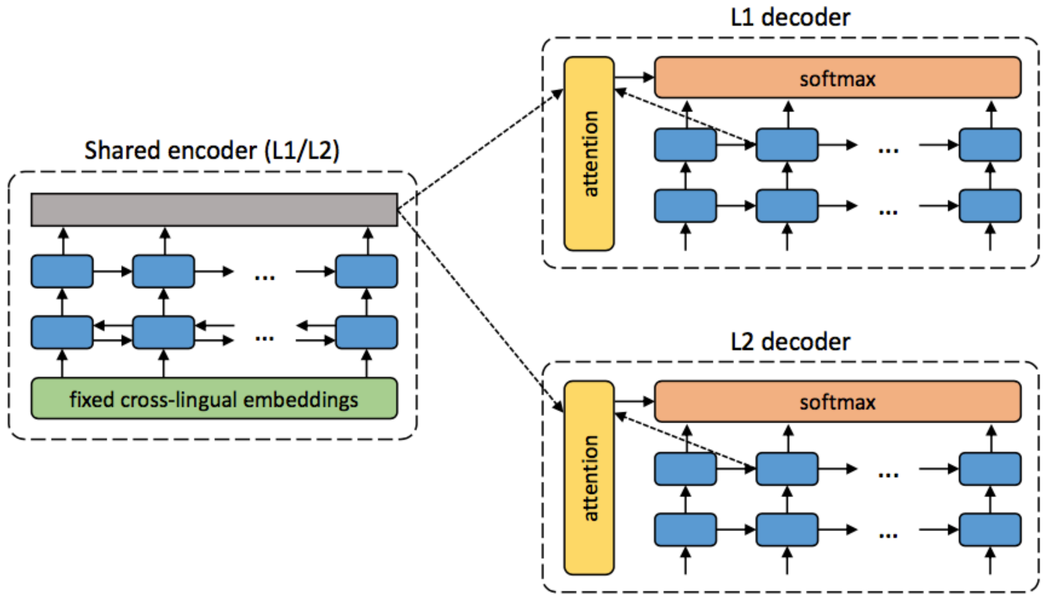
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation



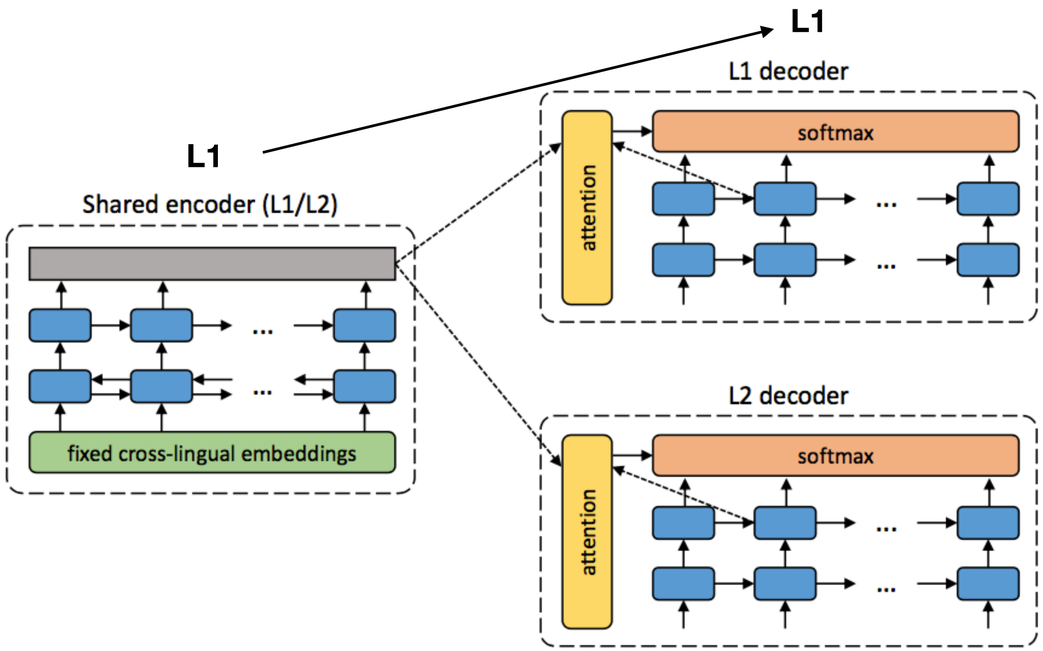
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:



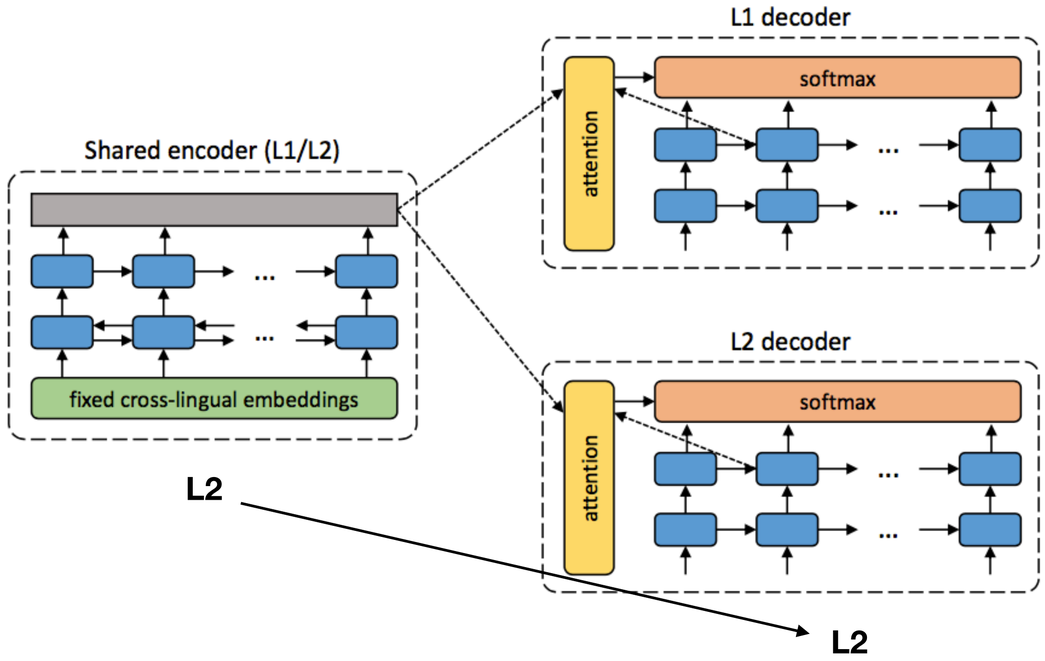
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1



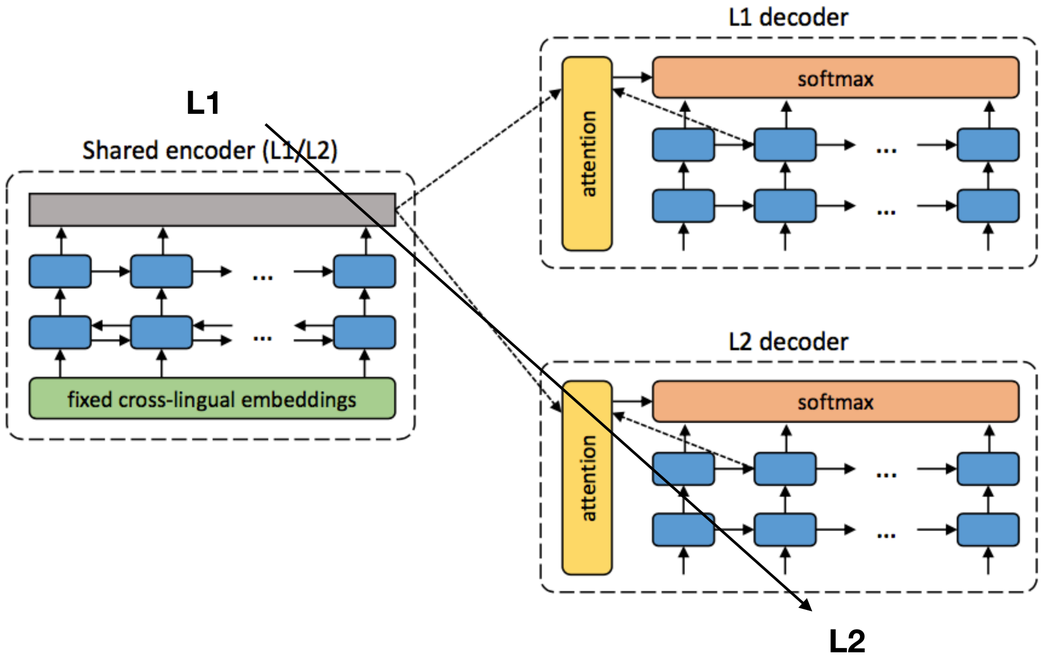
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1
  - Denoising batch: L2 to L2



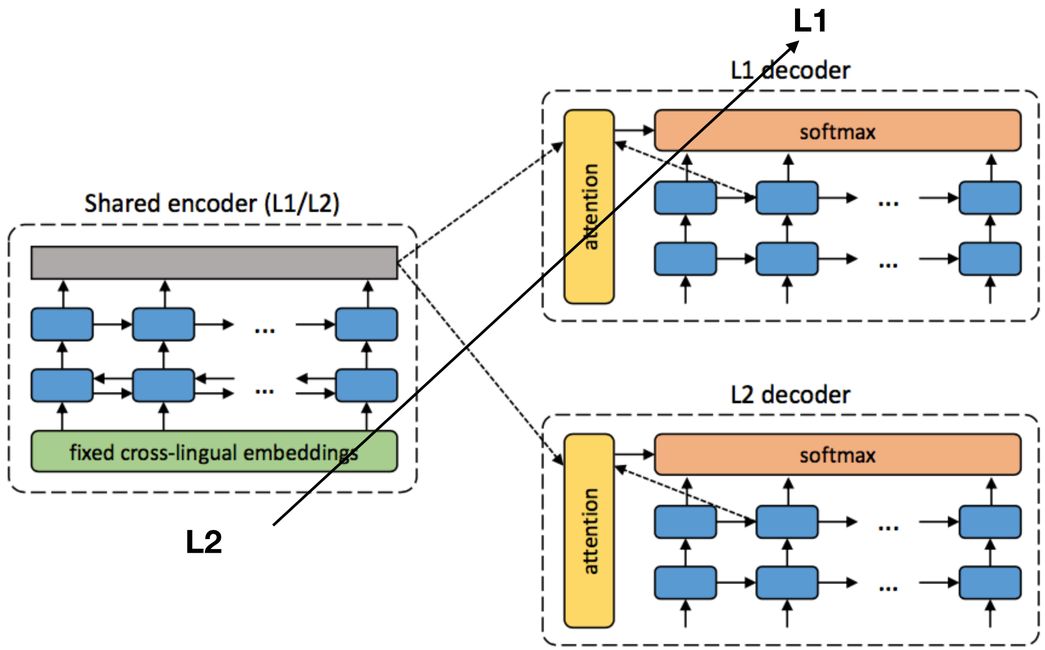
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1
  - Denoising batch: L2 to L2
  - Back-translation batch: L1 to L2



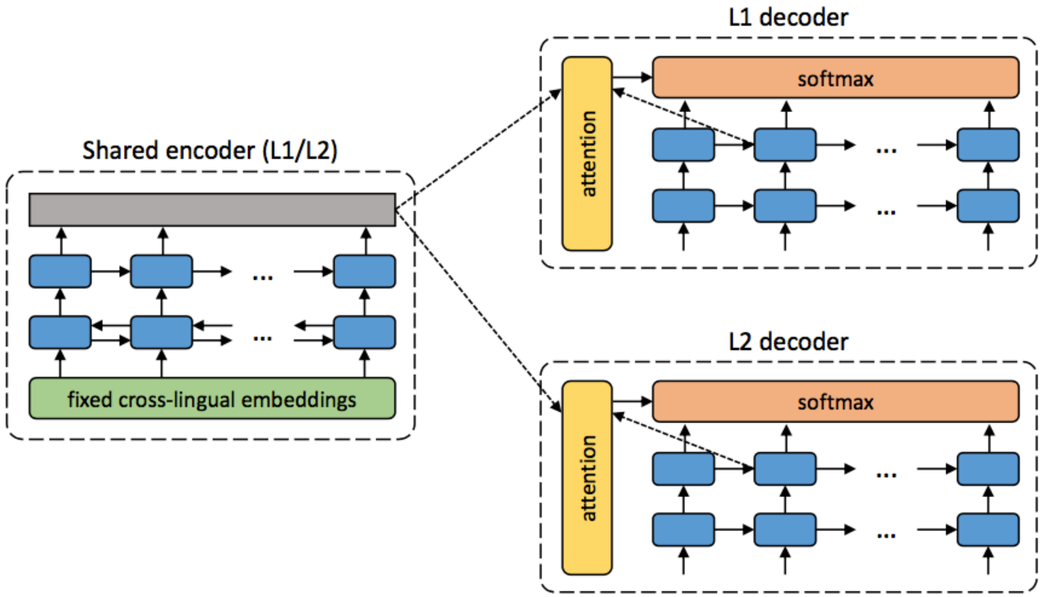
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1
  - Denoising batch: L2 to L2
  - Back-translation batch: L1 to L2
  - Back-translation batch: L2 to L1



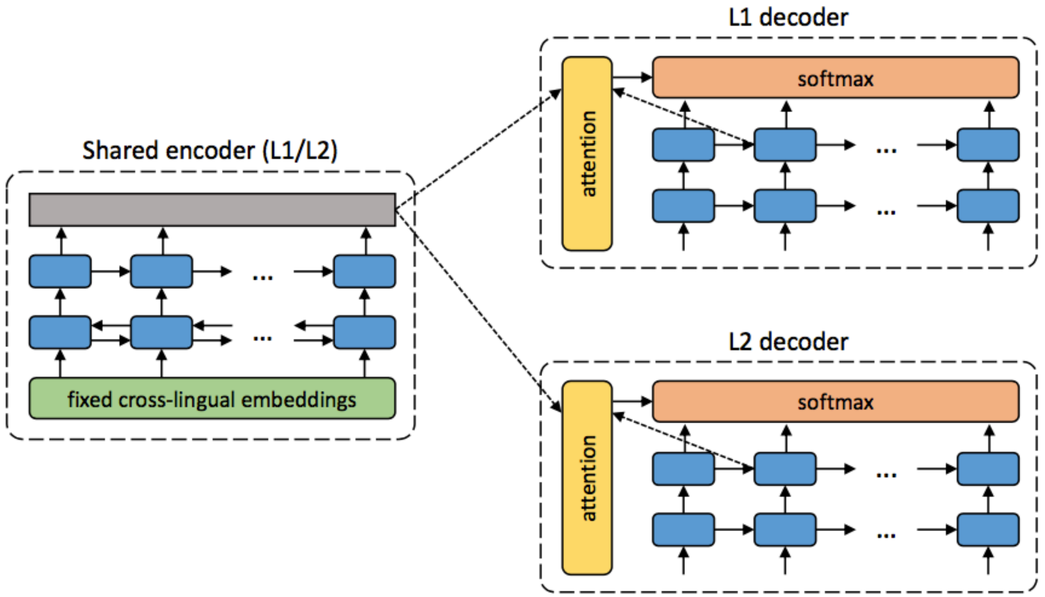
Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1
  - Denoising batch: L2 to L2
  - Back-translation batch: L1 to L2
  - Back-translation batch: L2 to L1
- When do we stop? Can't use parallel validation set!



Putting It All Together: Iterative Training

- Training iterations alternate between denoising and back-translation
- Each training iteration is composed of:
  - Denoising batch: L1 to L1
  - Denoising batch: L2 to L2
  - Back-translation batch: L1 to L2
  - Back-translation batch: L2 to L1
- When do we stop? Can't use parallel validation set!
  - Train for a fixed amount of iterations (300k)



Results

Results

- Denoising alone is weaker than the nearest-neighbor baseline

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Results

- Denoising alone is weaker than the nearest-neighbor baseline
- Denoising+Back-translation significantly improves results

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Results

- Denoising alone is weaker than the nearest-neighbor baseline
- Denoising+Back-translation significantly improves results
- No clear benefit from BPE (harder to learn embeddings?)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Results

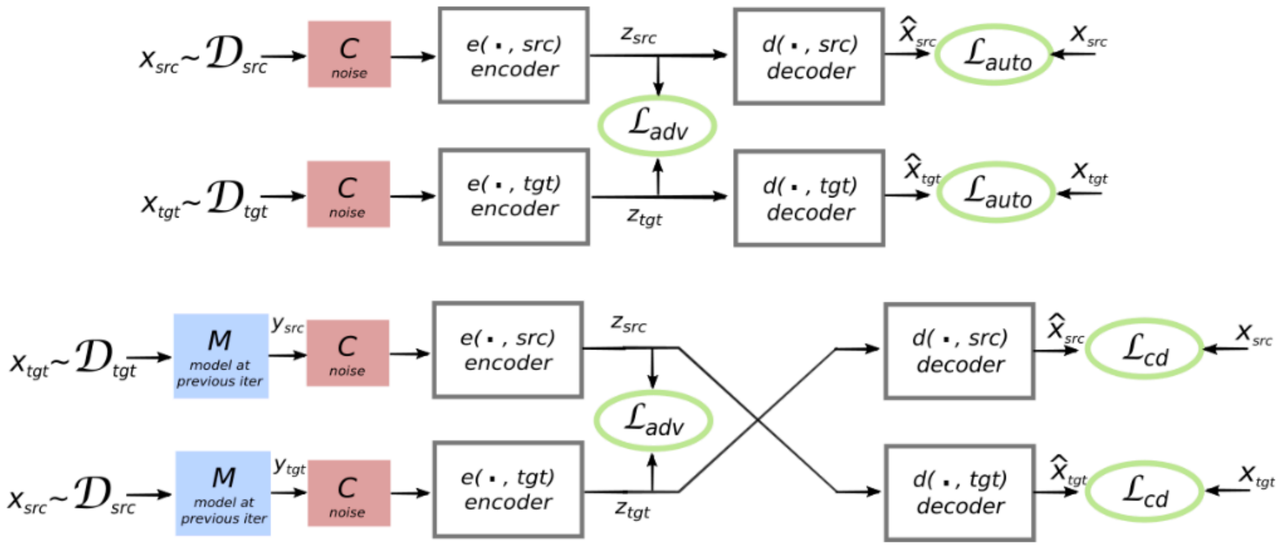
- Denoising alone is weaker than the nearest-neighbor baseline
- Denoising+Back-translation significantly improves results
- No clear benefit from BPE (harder to learn embeddings?)
- Semi supervised learning can also use this framework with notable gains
- **Still a very large gap from the supervised approach (but a nice start nonetheless)**

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	6. Comparable NMT	20.48	19.89	15.04	11.05
	7. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Paper II: Lample, Denoyer and Ranzato (FAIR)

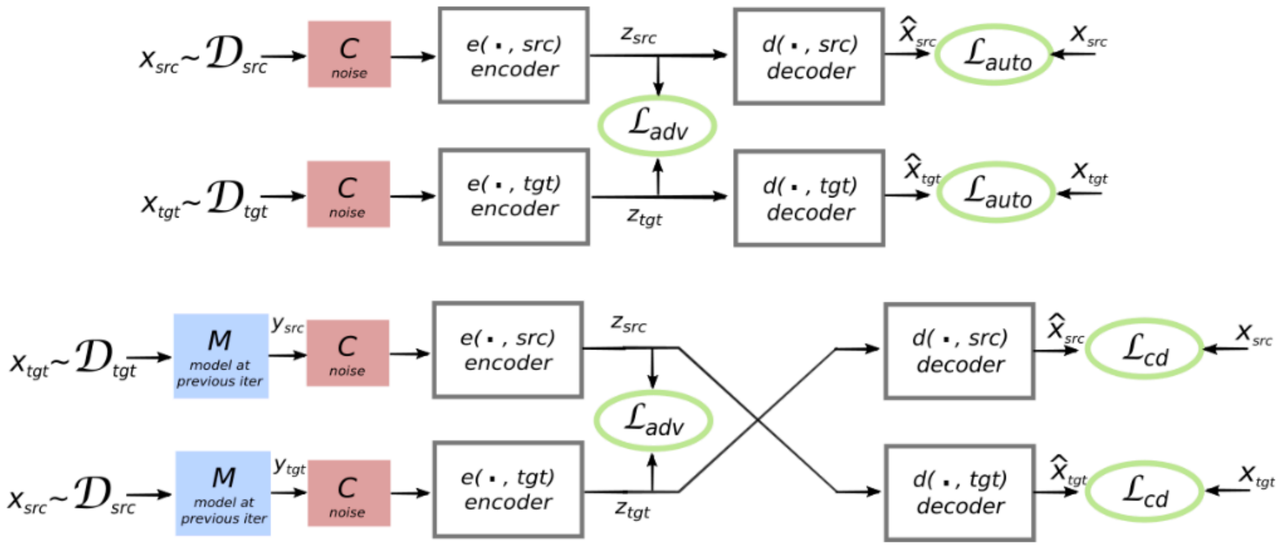
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**



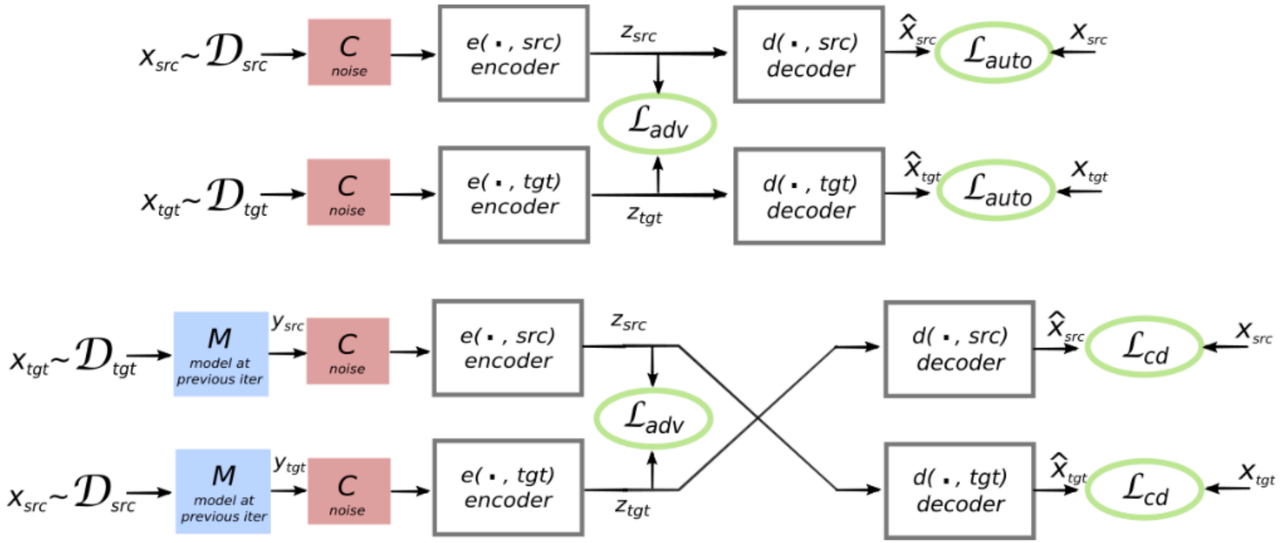
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder



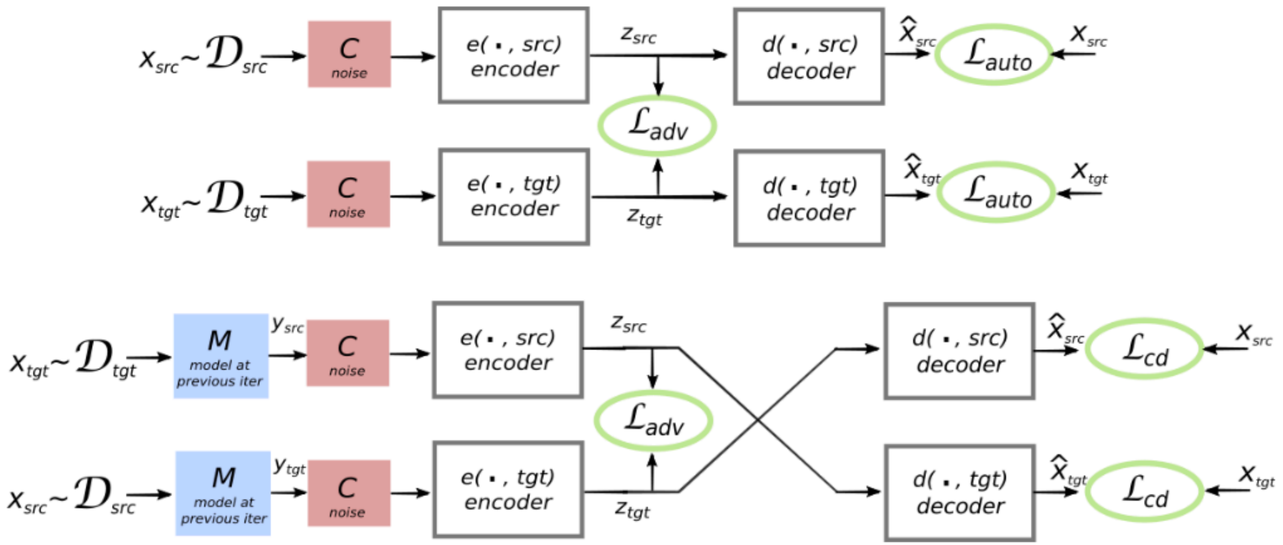
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention



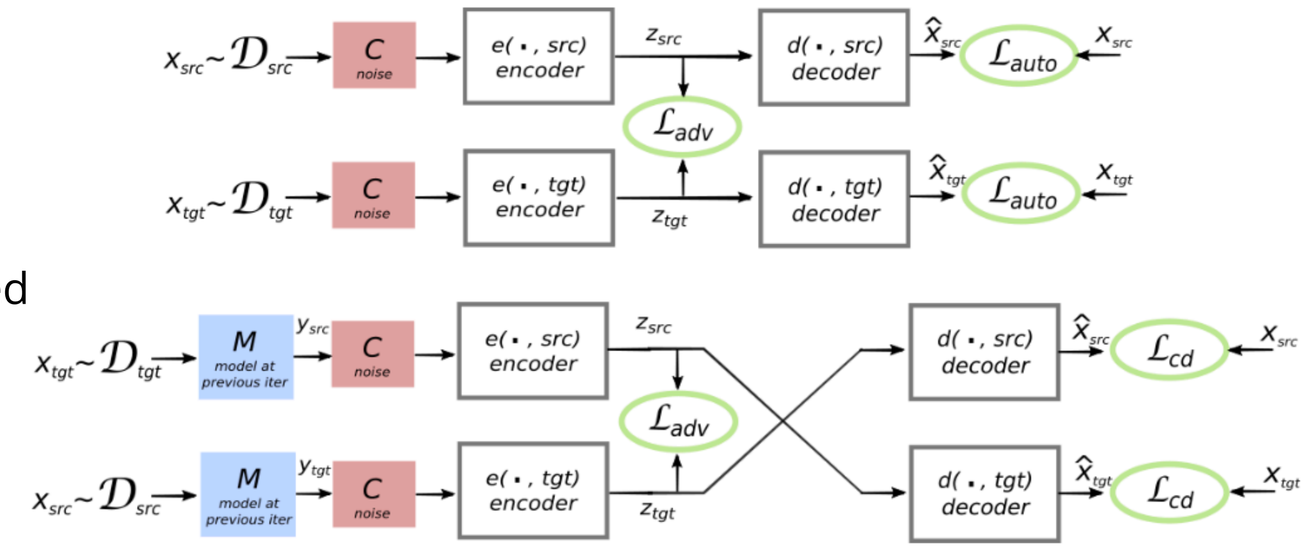
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention
- **Main “Tricks”:**



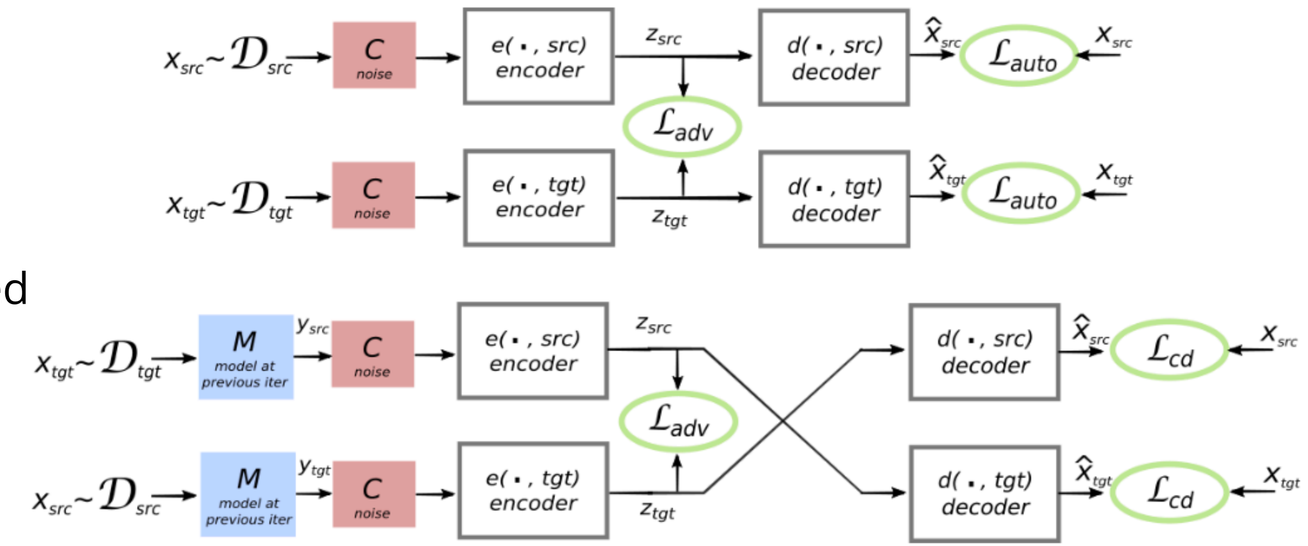
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention
- **Main "Tricks":**
  - **Changing, adversarially trained** unsupervised cross-lingual embeddings (**Adequacy**)



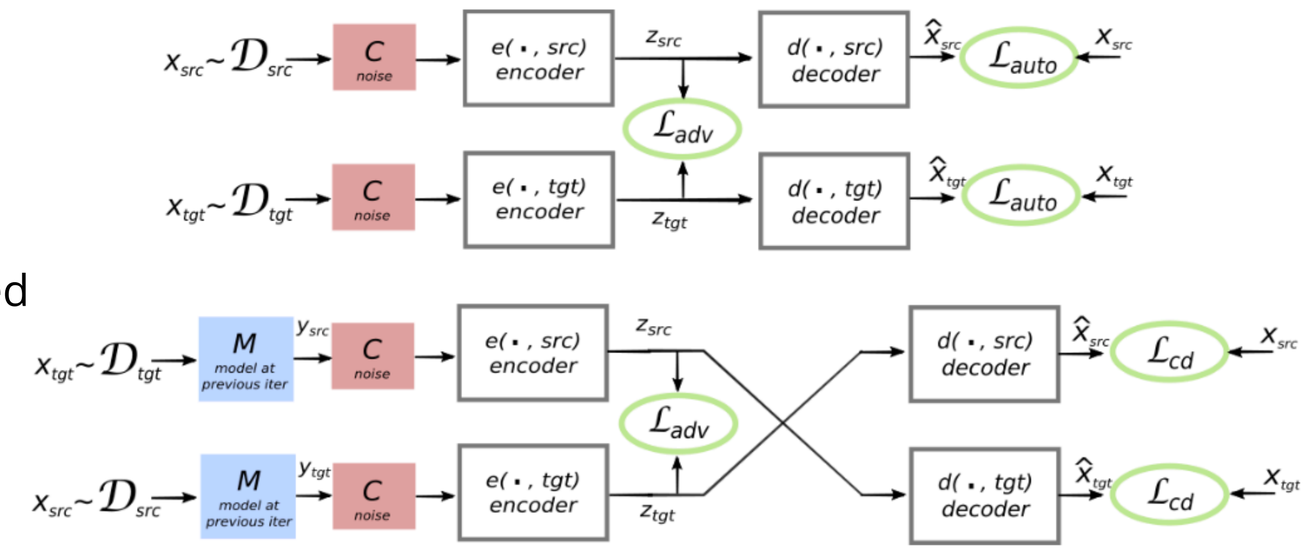
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention
- **Main "Tricks":**
  - **Changing, adversarially trained** unsupervised cross-lingual embeddings (**Adequacy**)
  - Backtranslation loss (**Adequacy**)



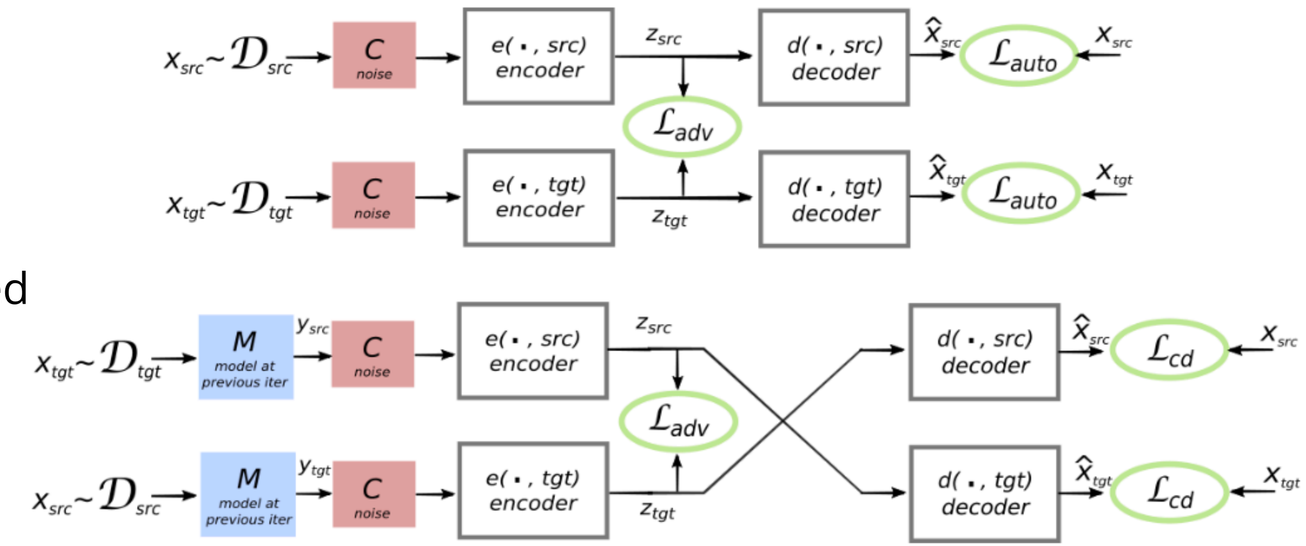
Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention
- **Main "Tricks":**
  - **Changing, adversarially trained** unsupervised cross-lingual embeddings (**Adequacy**)
  - Backtranslation loss (**Adequacy**)
  - Denoising auto-encoder loss (**Fluency**)



Paper II: Lample, Denoyer and Ranzato (FAIR)

- **Model Architecture:**
  - **Shared** GRU encoder, **Shared** GRU decoder
  - Attention
- **Main "Tricks":**
  - **Changing, adversarially trained** unsupervised cross-lingual embeddings (**Adequacy**)
  - Backtranslation loss (**Adequacy**)
  - Denoising auto-encoder loss (**Fluency**)
  - **Adversarial loss**



Adversarial Training

Adversarial Training

- Introduced by Ganin et al., 2016 for domain adaption in computer vision

Adversarial Training

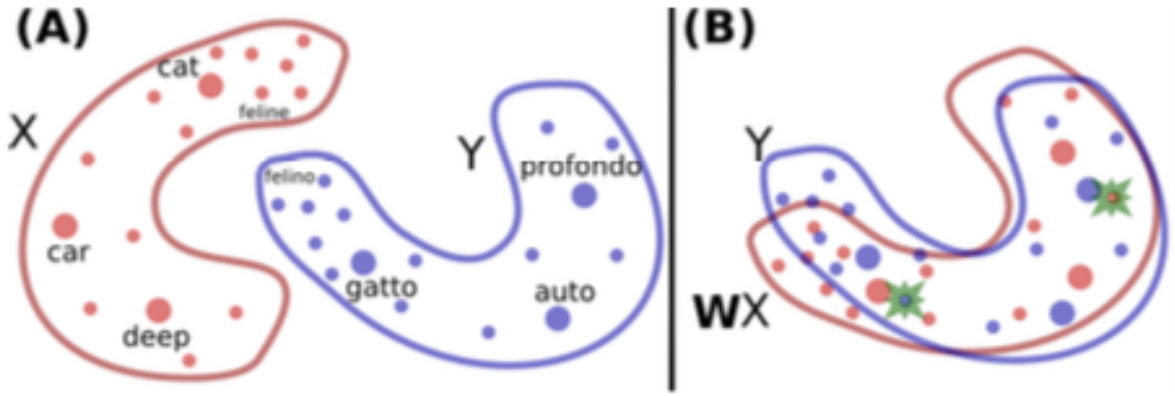
- Introduced by Ganin et al., 2016 for domain adaption in computer vision
- The general idea: force the model to “unlearn” a specific objective to make it learn better representation for the target objective

Adversarial Training

- Introduced by Ganin et al., 2016 for domain adaption in computer vision
- The general idea: force the model to “unlearn” a specific objective to make it learn better representation for the target objective
- Used twice here:

Adversarial Training

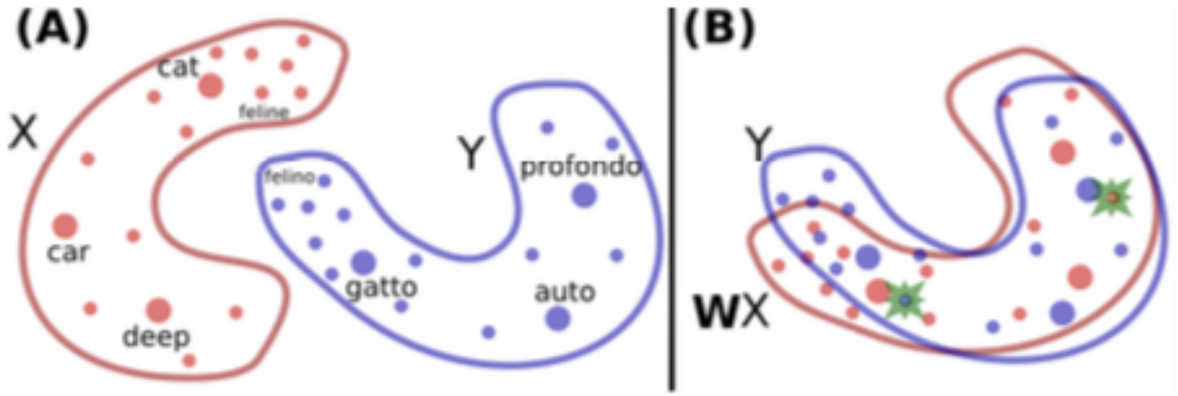
- Introduced by Ganin et al., 2016 for domain adaption in computer vision
- The general idea: force the model to “unlearn” a specific objective to make it learn better representation for the target objective
- Used twice here:
  - In the cross-lingual embedding learning - to learn a mapping from one embedding space to the other:



Conneau et al. 2017

Adversarial Training

- Introduced by Ganin et al., 2016 for domain adaption in computer vision
- The general idea: force the model to “unlearn” a specific objective to make it learn better representation for the target objective
- Used twice here:
  - In the cross-lingual embedding learning - to learn a mapping from one embedding space to the other:
  - In the NMT training - to “push” the representations from the two languages to a shared “semantic” space



Conneau et al. 2017

$$p_D(l|z_1, ..., z_m) \propto \prod_{j=1}^{m} p_D(\ell|z_j),$$

$$\mathcal{L}_{adv}(\theta_{enc}, \mathcal{Z}|\theta_D) = -\mathbb{E}_{(x_i,\ell_i)}[\log p_D(\ell_j|e(x_i, \ell_i))]$$

Unsupervised Model Selection Criterion

Unsupervised Model Selection Criterion

- When do we stop training without a validation set? can we do better than fixed amount of updates?

Unsupervised Model Selection Criterion

- When do we stop training without a validation set? can we do better than fixed amount of updates?
- Measure “corruption” when translating a sentence back and forth using the model (in both directions), using BLEU

$$MS(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}) =$$

$$\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{src}}\left[BLEU(x, M_{src\to tgt} \circ M_{tgt\to src}(x))\right] +$$

$$\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[BLEU(x, M_{tgt\to src} \circ M_{src\to tgt}(x))\right]$$

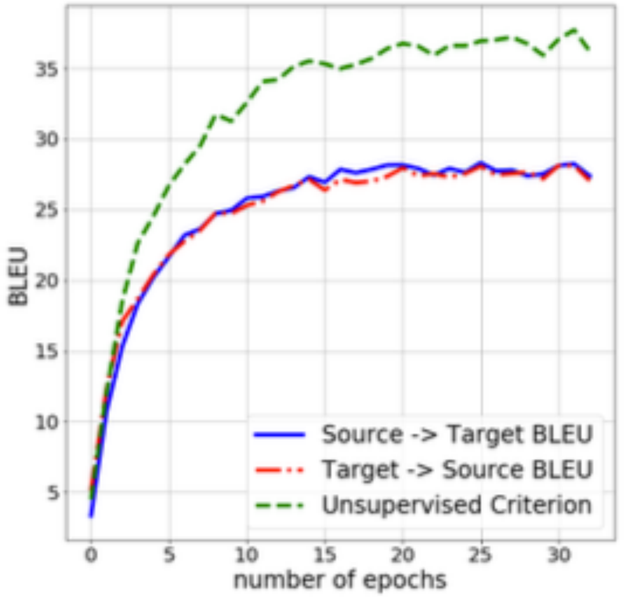
Unsupervised Model Selection Criterion

- When do we stop training without a validation set? can we do better than fixed amount of updates?
- Measure “corruption” when translating a sentence back and forth using the model (in both directions), using BLEU
- Correlates well with “supervised” BLEU, no need for parallel sentences

$$MS(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}) =$$

$$\frac{1}{2}\mathbb{E}_{x \sim \mathcal{D}_{src}}\left[BLEU(x, M_{src \to tgt} \circ M_{tgt \to src}(x))\right] +$$

$$\frac{1}{2}\mathbb{E}_{x \sim \mathcal{D}_{tgt}}\left[BLEU(x, M_{tgt \to src} \circ M_{src \to tgt}(x))\right]$$



Results

Results

- Model significantly outperforms word-by-word baselines, showing the importance of the back-translation + denoising + adversarial approach

	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

Results

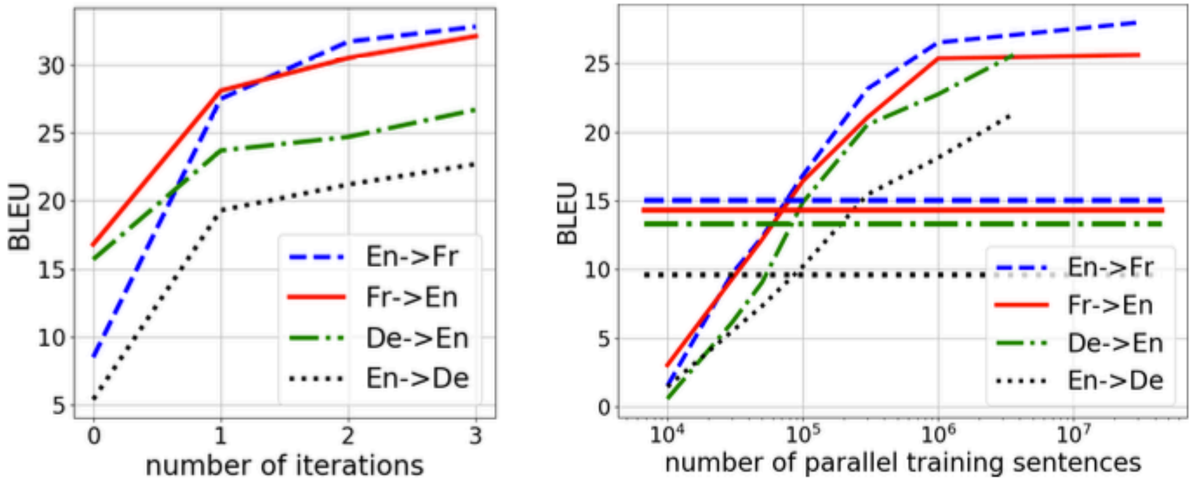
- Model significantly outperforms word-by-word baselines, showing the importance of the back-translation + denoising + adversarial approach
- Supervised models are still significantly better

	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

Results

- Model significantly outperforms word-by-word baselines, showing the importance of the back-translation + denoising + adversarial approach
- Supervised models are still significantly better
- Unsupervised models performance is equivalent to a supervised model with ~100k parallel sentences

	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



Ablation Study

Ablation Study

- Back-translation, pre-trained word vectors and de-noising are crucial

Ablation Study

- Back-translation, pre-trained word vectors and de-noising are crucial
- Adversarial loss gives a nice boost of ~3-6 points

Ablation Study

- Back-translation, pre-trained word vectors and de-noising are crucial
- Adversarial loss gives a nice boost of ~3-6 points
- Best model obtained using all components together

Ablation Study

- Back-translation, pre-trained word vectors and de-noising are crucial
- Adversarial loss gives a nice boost o ~3-6 points
- Best model obtained using all components together

	en-fr	fr-en	de-en	en-de
$\lambda_{cd} = 0$	25.44	27.14	20.56	14.42
Without pretraining	25.29	26.10	21.44	17.23
Without pretraining, $\lambda_{cd} = 0$	8.78	9.15	7.52	6.24
Without noise, $C(x) = x$	16.76	16.85	16.85	14.61
$\lambda_{auto} = 0$	24.32	20.02	19.10	14.74
$\lambda_{adv} = 0$	24.12	22.74	19.87	15.13
Full	**27.48**	**28.07**	**23.69**	**19.32**

Comparison

Comparison

- **Both models:**

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)
  - Different decoder per language (Artetxe et al.) vs. shared encoder (Lample et al.)

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)
  - Different decoder per language (Artetxe et al.) vs. shared encoder (Lample et al.)
  - Adversarial training (Lample et al.)

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)
  - Different decoder per language (Artetxe et al.) vs. shared encoder (Lample et al.)
  - Adversarial training (Lample et al.)
  - Slightly different noise method (Lample et al.) - swapping and dropping words, also adding noise before back-translation

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)
  - Different decoder per language (Artetxe et al.) vs. shared encoder (Lample et al.)
  - Adversarial training (Lample et al.)
  - Slightly different noise method (Lample et al.) - swapping and dropping words, also adding noise before back-translation
  - Use unsupervised model selection criterion (Lample et al.) vs. fixed amount of updates

Comparison

- **Both models:**
  - Heavily rely on bilingual word embeddings
  - Heavily rely on de-noising and back-translation using a shared encoder
- **Notable Differences:**
  - Perform back-translation once per epoch (Lample et al.) vs. after every update (Artetxe et al.)
  - BPE and word-based modeling (Artetxe et al.) vs. word based alone (Lample et al.)
  - Fixed embeddings (Artetxe et al.) vs. changing (Lample et al.)
  - Different decoder per language (Artetxe et al.) vs. shared encoder (Lample et al.)
  - Adversarial training (Lample et al.)
  - Slightly different noise method (Lample et al.) - swapping and dropping words, also adding noise before back-translation
  - Use unsupervised model selection criterion (Lample et al.) vs. fixed amount of updates
  - Initialize back-translation using nearest-neighbor word-by-word translation (Lample et al.)

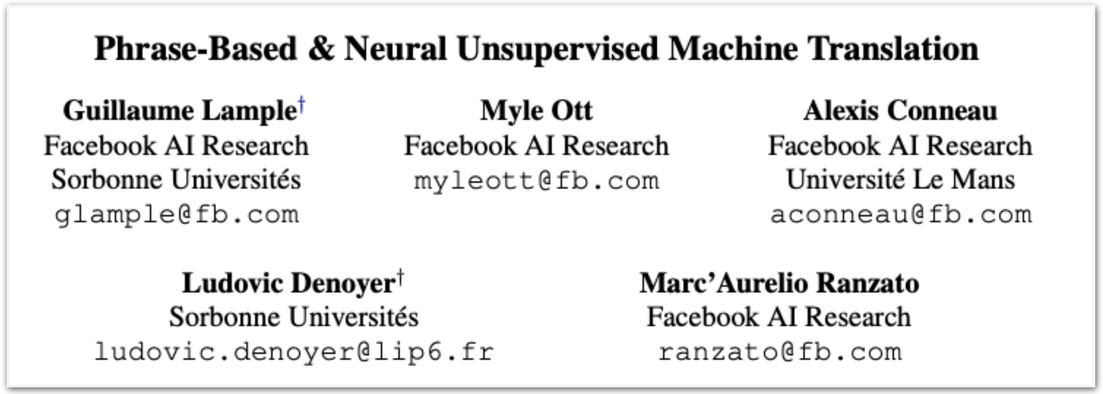
Phrase-Based Unsupervised NMT

Phrase-Based Unsupervised NMT

- A second wave of works integrated phrase-based models for unsupervised NMT (also from the same authors):

Phrase-Based Unsupervised NMT

- A second wave of works integrated phrase-based models for unsupervised NMT (also from the same authors):
  - Lample et al. (2018)



Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

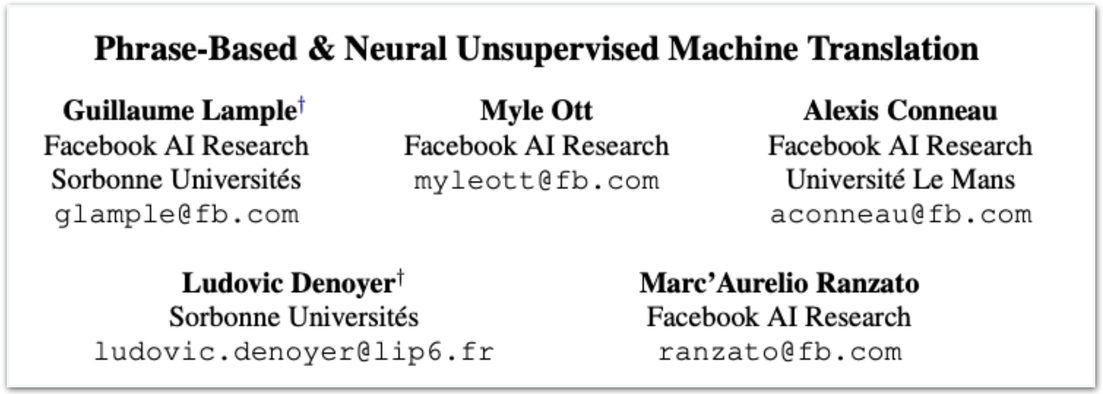
Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

Phrase-Based Unsupervised NMT

- A second wave of works integrated phrase-based models for unsupervised NMT (also from the same authors):
  - Lample et al. (2018)
  - Artetxe et al. (2018)



Phrase-Based & Neural Unsupervised Machine Translation

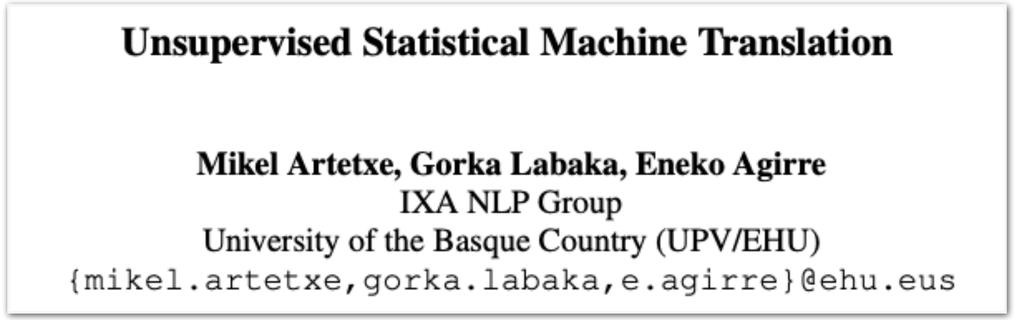
Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

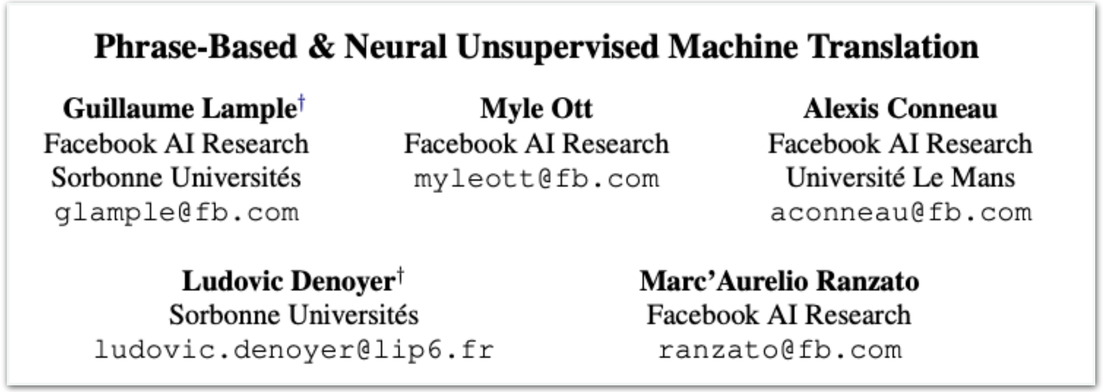


Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Phrase-Based Unsupervised NMT

- A second wave of works integrated phrase-based models for unsupervised NMT (also from the same authors):
  - Lample et al. (2018)
  - Artetxe et al. (2018)
  - Artetxe et al (2019)



Phrase-Based & Neural Unsupervised Machine Translation

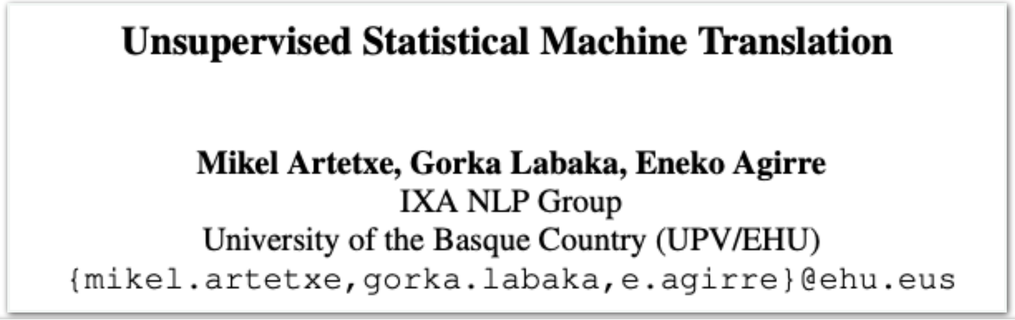
Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

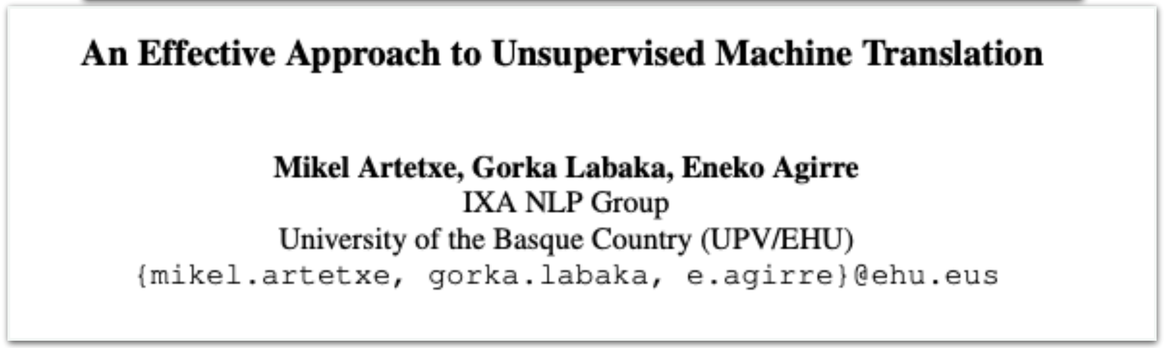
Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com



Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

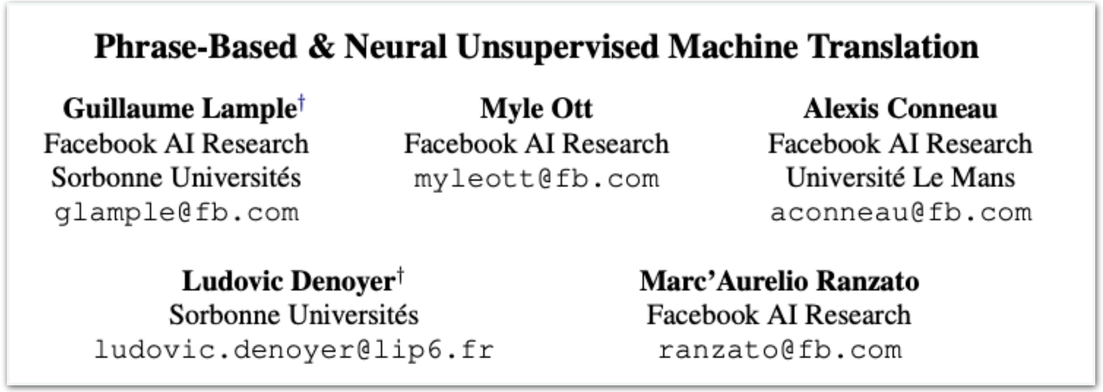


An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Phrase-Based Unsupervised NMT

- A second wave of works integrated phrase-based models for unsupervised NMT (also from the same authors):
  - Lample et al. (2018)
  - Artetxe et al. (2018)
  - Artetxe et al (2019)
- Makes sense, as SMT was shown to work better in low resource scenarios



Phrase-Based & Neural Unsupervised Machine Translation

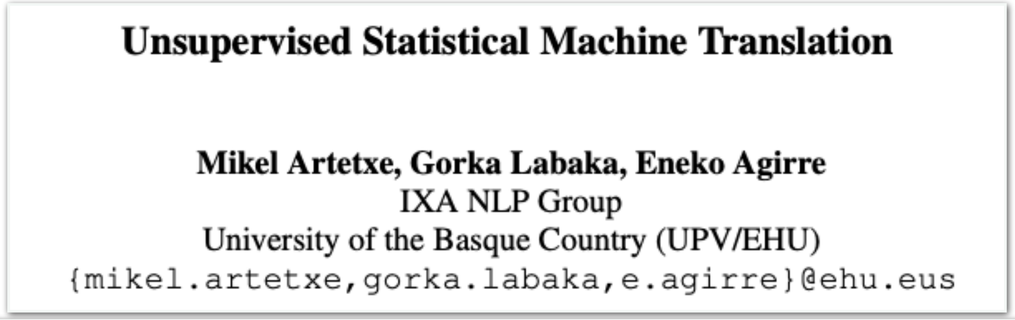
Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

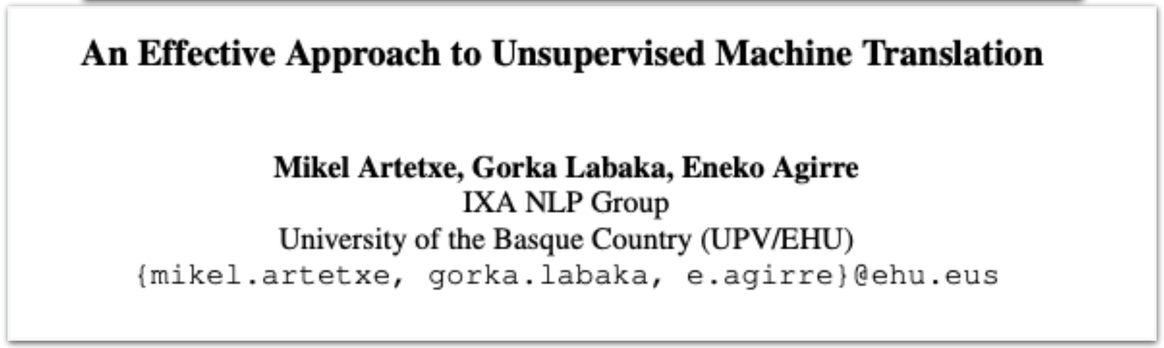
Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com



Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



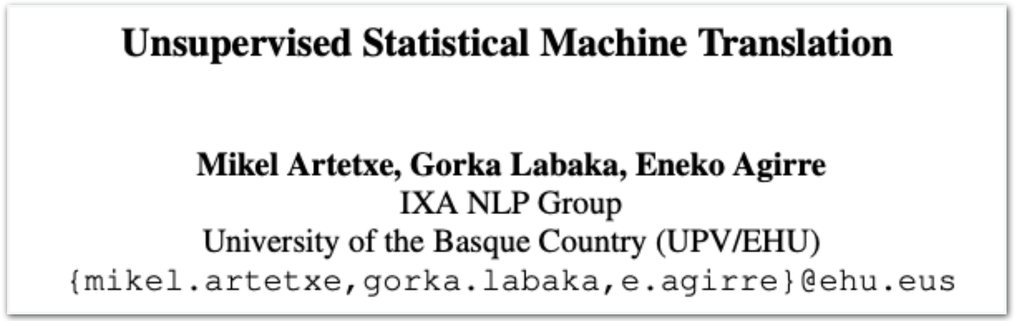
An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Unsupervised Statistical Machine Translation

Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018



Unsupervised Statistical Machine Translation

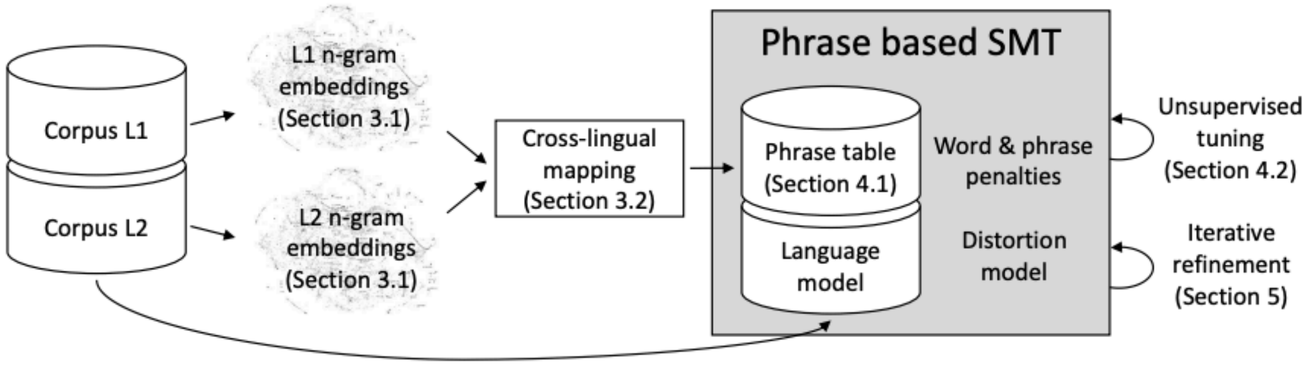
Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018
- Main idea: use SMT instead of NMT

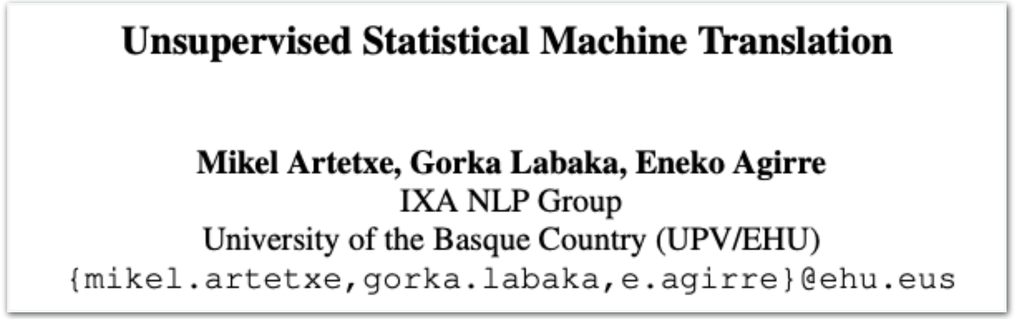
Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



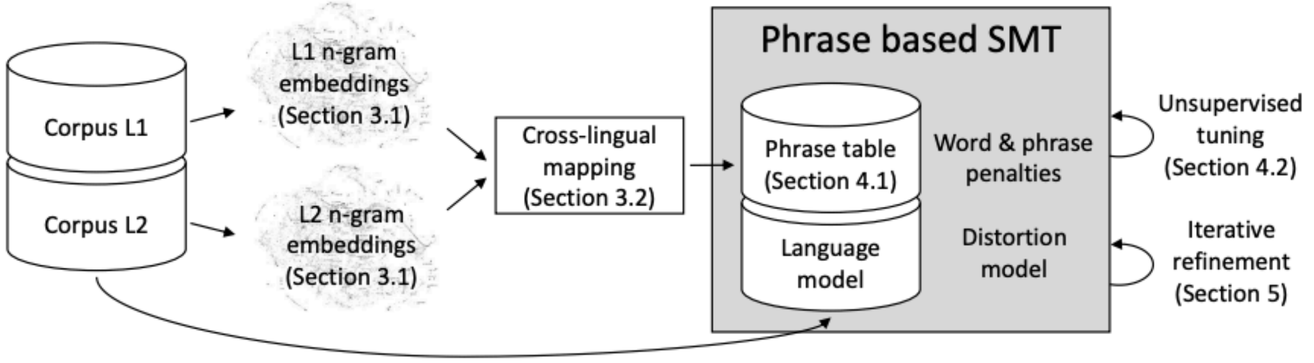
Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018
- Main idea: use SMT instead of NMT
- Train n-gram embeddings using a variation of skip-gram



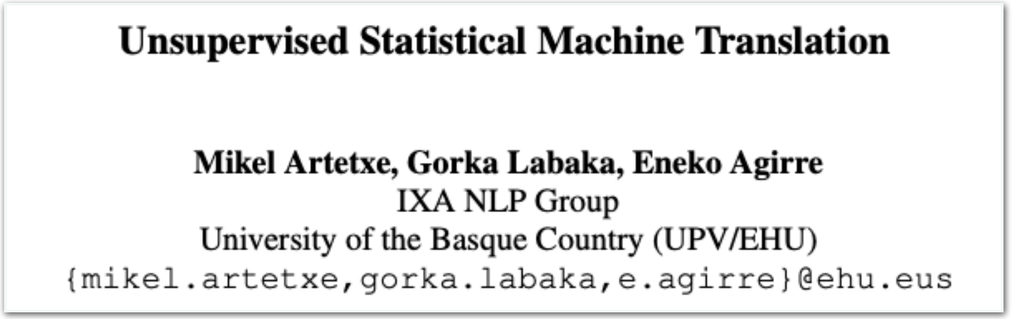
Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



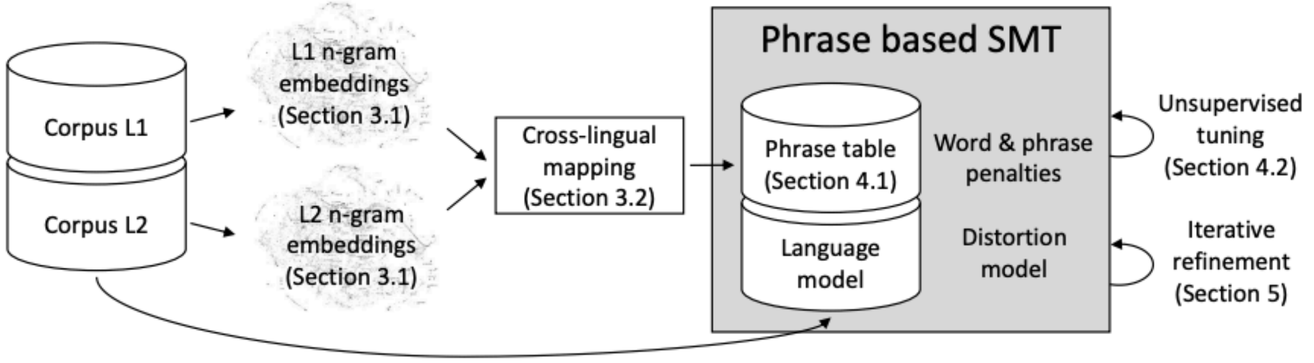
Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018
- Main idea: use SMT instead of NMT
- Train n-gram embeddings using a variation of skip-gram
- Learn a mapping between the embedding spaces



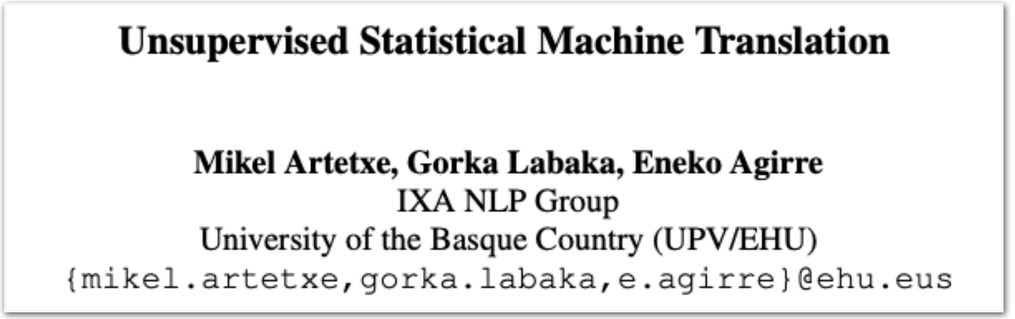
Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



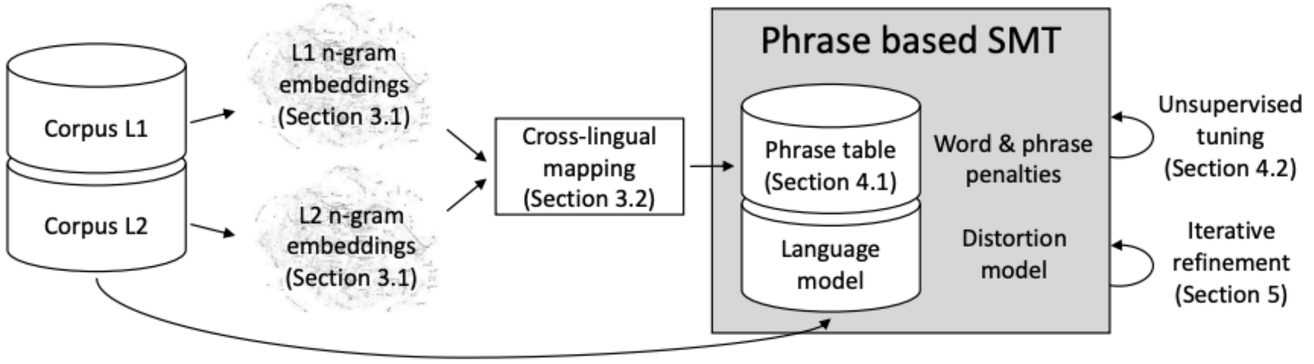
Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018
- Main idea: use SMT instead of NMT
- Train n-gram embeddings using a variation of skip-gram
- Learn a mapping between the embedding spaces
- Create a phrase table by computing translation probabilities using softmax over the cosine-similarities to the 100 nearest neighbours



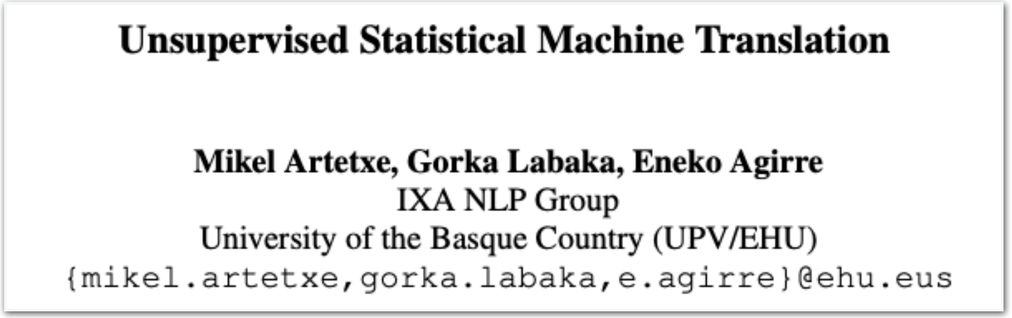
Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



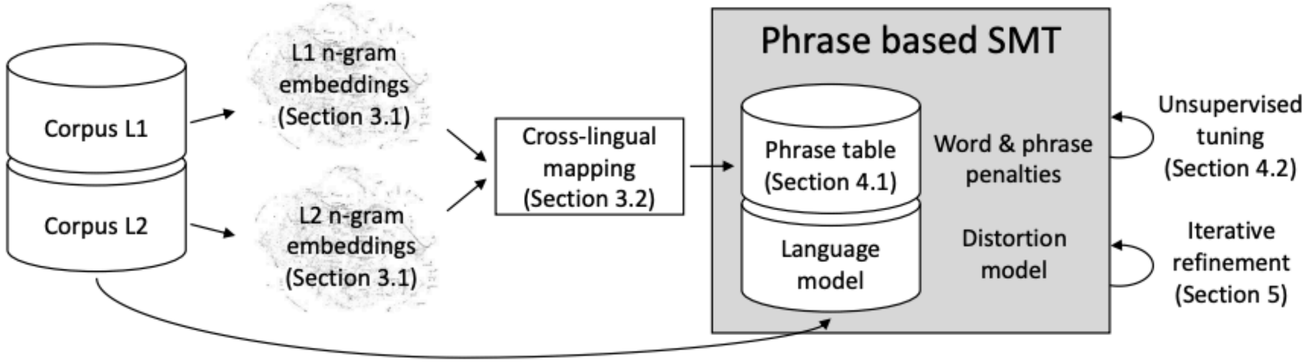
Unsupervised Statistical Machine Translation

- Artetxe, Labaka and Agirre, 2018
- Main idea: use SMT instead of NMT
- Train n-gram embeddings using a variation of skip-gram
- Learn a mapping between the embedding spaces
- Create a phrase table by computing translation probabilities using softmax over the cosine-similarities to the 100 nearest neighbours
- Tune the resulting system using iterative back-translation



Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus



Results

Results

- Much better than previous unsupervised NMT approaches across 6 language pairs

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Artetxe et al. (2018c)	15.56	15.13	10.21	6.55	-	-
Lample et al. (2018)	14.31	15.05	-	-	13.33	9.64
Yang et al. (2018)	15.58	16.97	-	-	14.62	10.86
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**	**23.05**	**18.23**

Results

- Much better than previous unsupervised NMT approaches across 6 language pairs
- Unsupervised Tuning and Iterative Back-Translation are important

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Artetxe et al. (2018c)	15.56	15.13	10.21	6.55	-	-
Lample et al. (2018)	14.31	15.05	-	-	13.33	9.64
Yang et al. (2018)	15.58	16.97	-	-	14.62	10.86
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**	**23.05**	**18.23**

	WMT-14				WMT-16	
	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**	**DE-EN**	**EN-DE**
Unsupervised SMT	21.16	20.13	13.86	10.59	18.01	13.22
+ unsupervised tuning	22.17	22.22	14.73	10.64	18.21	13.12
+ iterative refinement (it1)	24.81	26.53	16.01	13.45	20.76	16.94
+ iterative refinement (it2)	**26.13**	**26.57**	17.30	13.95	22.80	18.18
+ iterative refinement (it3)	25.87	26.22	**17.43**	**14.08**	**23.05**	**18.23**

Results

- Much better than previous unsupervised NMT approaches across 6 language pairs
- Unsupervised Tuning and Iterative Back-Translation are important
- Still far from the supervised approaches

	WMT-14				WMT-16	
	FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Artetxe et al. (2018c)	15.56	15.13	10.21	6.55	-	-
Lample et al. (2018)	14.31	15.05	-	-	13.33	9.64
Yang et al. (2018)	15.58	16.97	-	-	14.62	10.86
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**	**23.05**	**18.23**

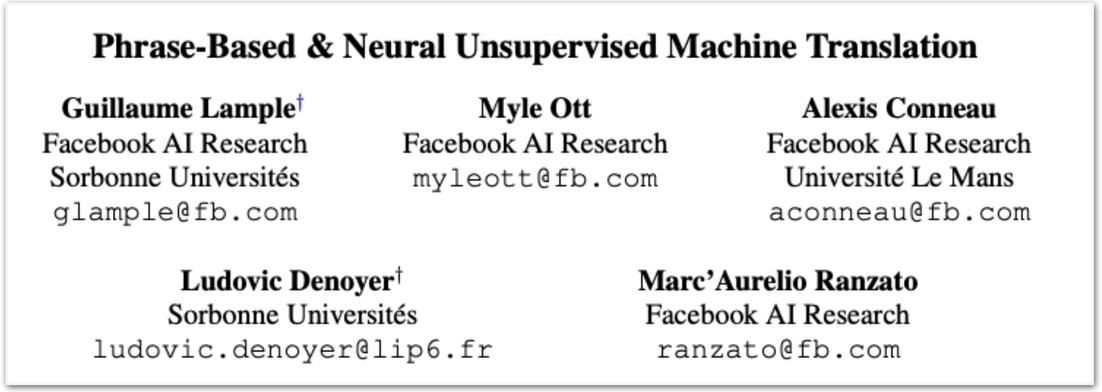
	WMT-14				WMT-16	
	FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Unsupervised SMT	21.16	20.13	13.86	10.59	18.01	13.22
+ unsupervised tuning	22.17	22.22	14.73	10.64	18.21	13.12
+ iterative refinement (it1)	24.81	26.53	16.01	13.45	20.76	16.94
+ iterative refinement (it2)	**26.13**	**26.57**	17.30	13.95	22.80	18.18
+ iterative refinement (it3)	25.87	26.22	**17.43**	**14.08**	**23.05**	**18.23**

		WMT-14				WMT-16	
		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Supervised	NMT (transformer)	-	41.8	-	28.4	-	-
	WMT best	35.0	35.8	29.0	20.6	40.2	34.2
	SMT (europarl)	30.61	30.82	20.83	16.60	26.38	22.12
	+ w/o lexical reord.	30.54	30.33	20.37	16.34	25.99	22.20
	+ constrained vocab.	30.04	30.10	19.91	16.32	25.66	21.53
	+ unsup. tuning	29.32	29.46	17.75	15.45	23.35	19.86
Unsup.	Proposed system	25.87	26.22	17.43	14.08	23.05	18.23

Hybrid Unsupervised Machine Translation

Hybrid Unsupervised Machine Translation

- First proposed by Lample et al. (2018) - proposed similar ideas with SMT, and a joint approach by training NMT on SMT outputs



Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

Hybrid Unsupervised Machine Translation

- First proposed by Lample et al. (2018) - proposed similar ideas with SMT, and a joint approach by training NMT on SMT outputs
- Improved by Artetxe et al (2019) with better tuning of the SMT model and a gradual mixing of SMT and NMT back-translations

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Hybrid Unsupervised Machine Translation

- First proposed by Lample et al. (2018) - proposed similar ideas with SMT, and a joint approach by training NMT on SMT outputs
- Improved by Artetxe et al (2019) with better tuning of the SMT model and a gradual mixing of SMT and NMT back-translations
- Pure SMT systems perform better than pure NMT systems, yet the best results are obtained by initializing an NMT system with an SMT system

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

An Effective Approach to Unsupervised Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

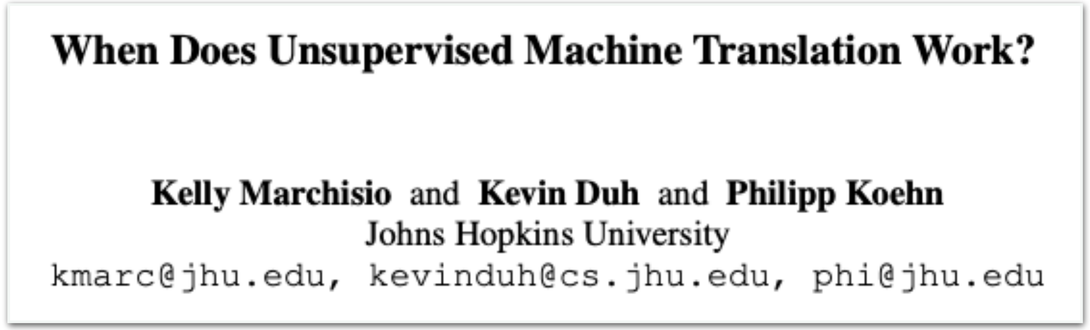
When Does It Really Work?

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful



When Does Unsupervised Machine Translation Work?

Kelly Marchisio and **Kevin Duh** and **Philipp Koehn**
Johns Hopkins University
kmarc@jhu.edu, kevinduh@cs.jhu.edu, phi@jhu.edu

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful

When Does Unsupervised Machine Translation Work?

Kelly Marchisio and **Kevin Duh** and **Philipp Koehn**
Johns Hopkins University
kmarc@jhu.edu, kevinduh@cs.jhu.edu, phi@jhu.edu

When and Why is Unsupervised Neural Machine Translation Useless?

Yunsu Kim **Miguel Graça**[†] **Hermann Ney**
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@cs.rwth-aachen.de

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful
- Both found that:

When Does Unsupervised Machine Translation Work?

Kelly Marchisio and **Kevin Duh** and **Philipp Koehn**
Johns Hopkins University
kmarc@jhu.edu, kevinduh@cs.jhu.edu, phi@jhu.edu

When and Why is Unsupervised Neural Machine Translation Useless?

Yunsu Kim **Miguel Graça**[†] **Hermann Ney**
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@cs.rwth-aachen.de

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful
- Both found that:
  - A critical condition is having the data drawn from a **similar domain**

When Does Unsupervised Machine Translation Work?

Kelly Marchisio and **Kevin Duh** and **Philipp Koehn**
Johns Hopkins University
kmarc@jhu.edu, kevinduh@cs.jhu.edu, phi@jhu.edu

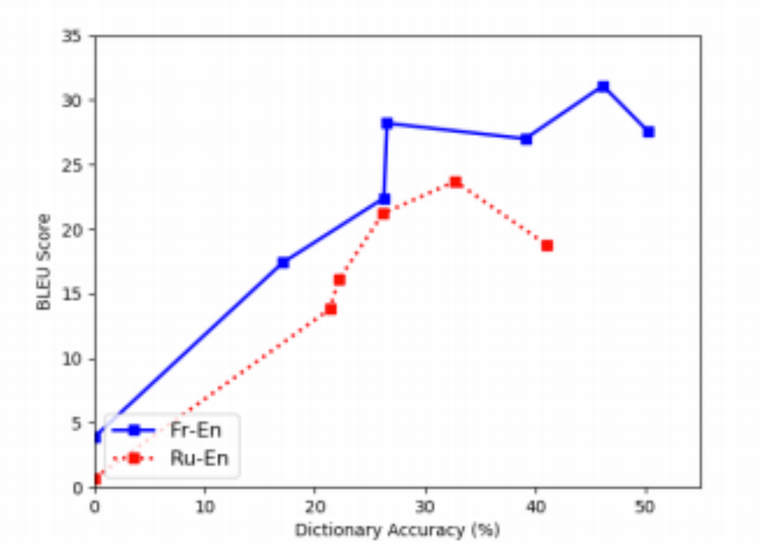
When and Why is Unsupervised Neural Machine Translation Useless?

Yunsu Kim **Miguel Graça**† **Hermann Ney**
Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@cs.rwth-aachen.de

Domain (en)	Domain (de/ru)	BLEU [%] de-en	en-de	ru-en	en-ru
Newswire	Newswire	23.3	19.9	11.9	9.3
	Politics	11.5	12.2	2.3	2.5
	Random	18.4	16.4	6.9	6.1

When Does It Really Work?

- Two recent works (2020) analyzed what are the conditions required to make unsupervised NMT useful
- Both found that:
  - A critical condition is having the data drawn from a **similar domain**
  - The **pretraining quality** has a strong effect on the final model

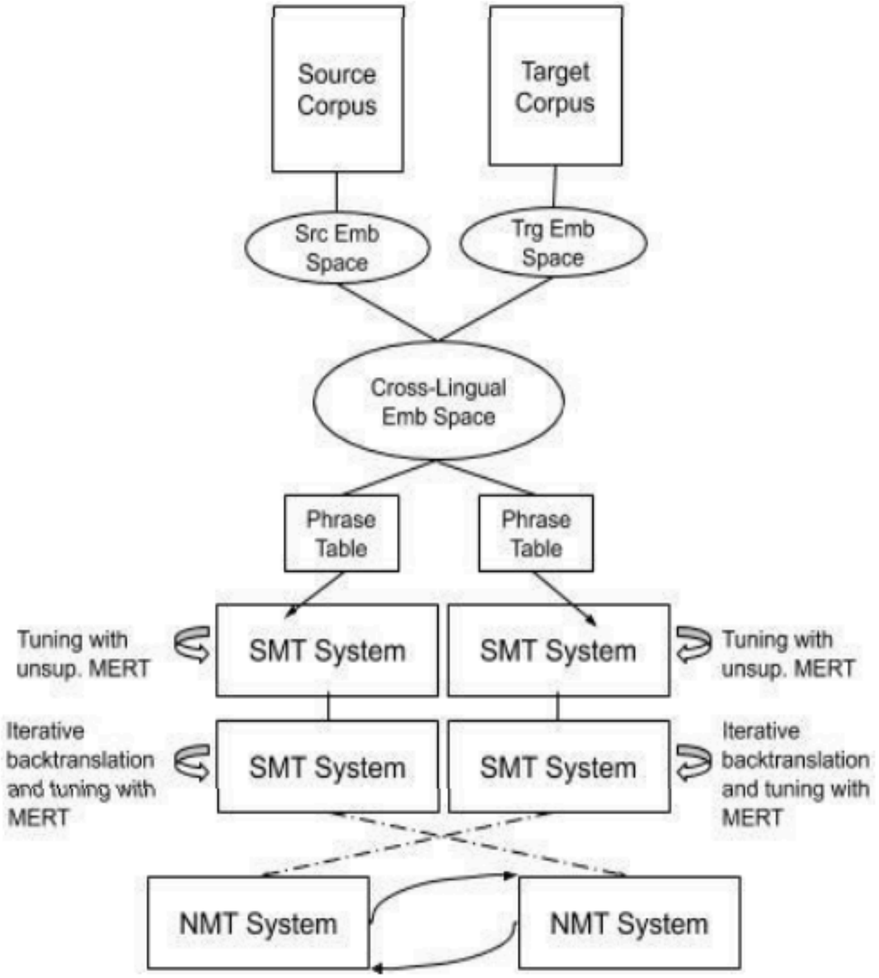


Domain (en)	Domain (de/ru)	BLEU [%] de-en	en-de	ru-en	en-ru
Newswire	Newswire	23.3	19.9	11.9	9.3
	Politics	11.5	12.2	2.3	2.5
	Random	18.4	16.4	6.9	6.1

Summary

Summary

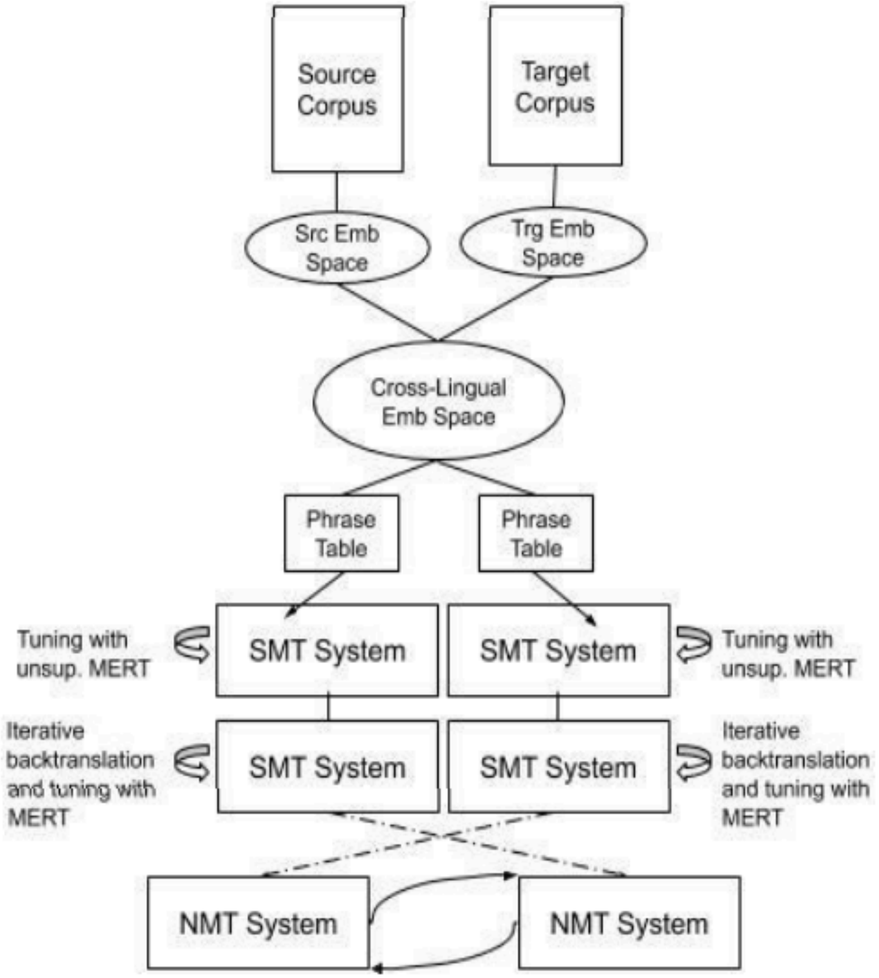
- Unsupervised NMT is possible!



From Marchisio et al. (2020)

Summary

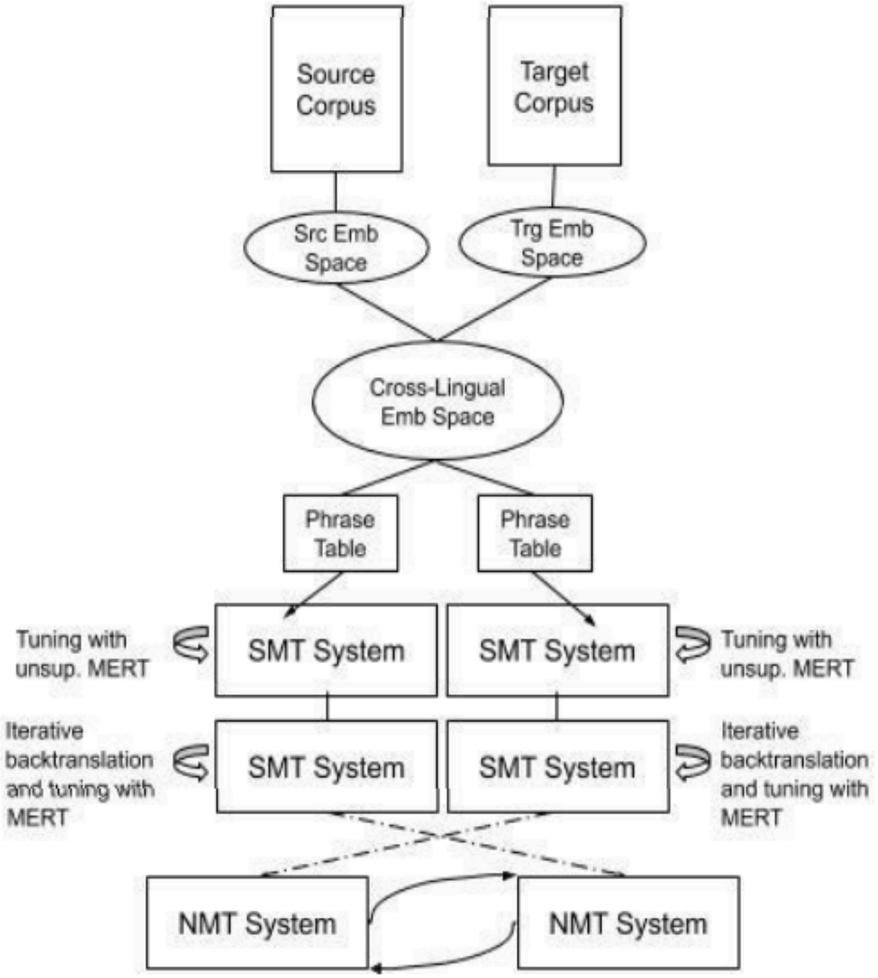
- Unsupervised NMT is possible!
  - Cross-lingual embeddings



From Marchisio et al. (2020)

Summary

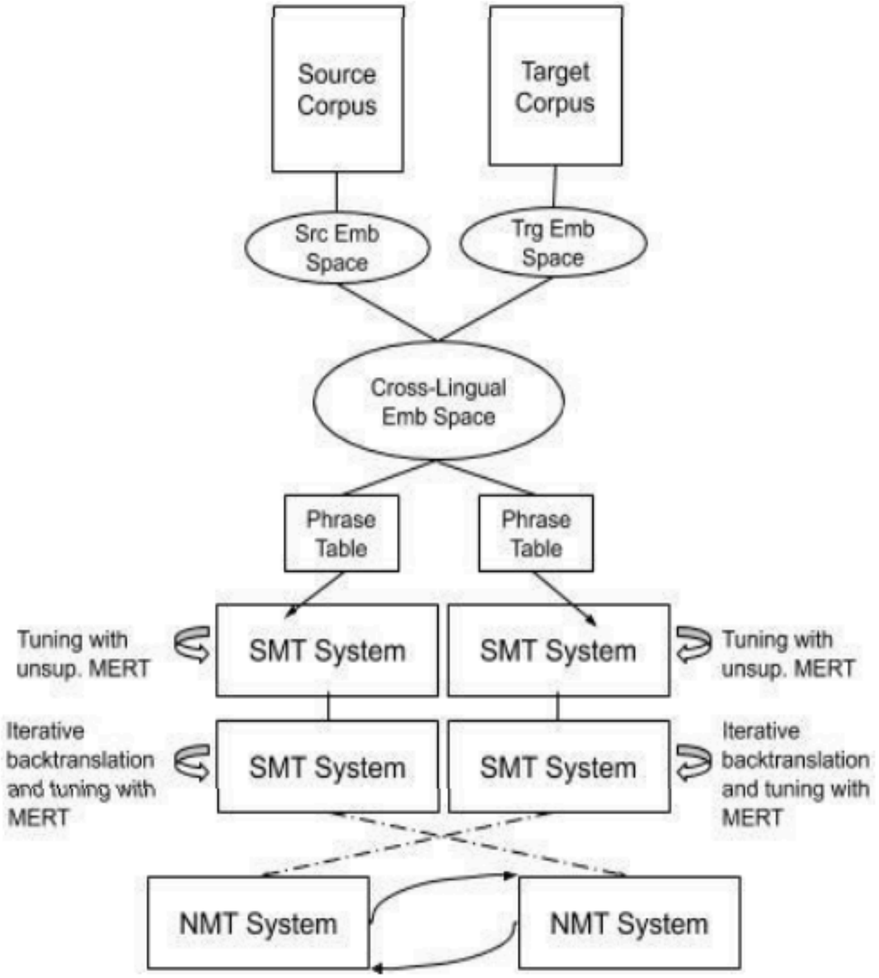
- Unsupervised NMT is possible!
  - Cross-lingual embeddings
  - Iterative Back-Translation



From Marchisio et al. (2020)

Summary

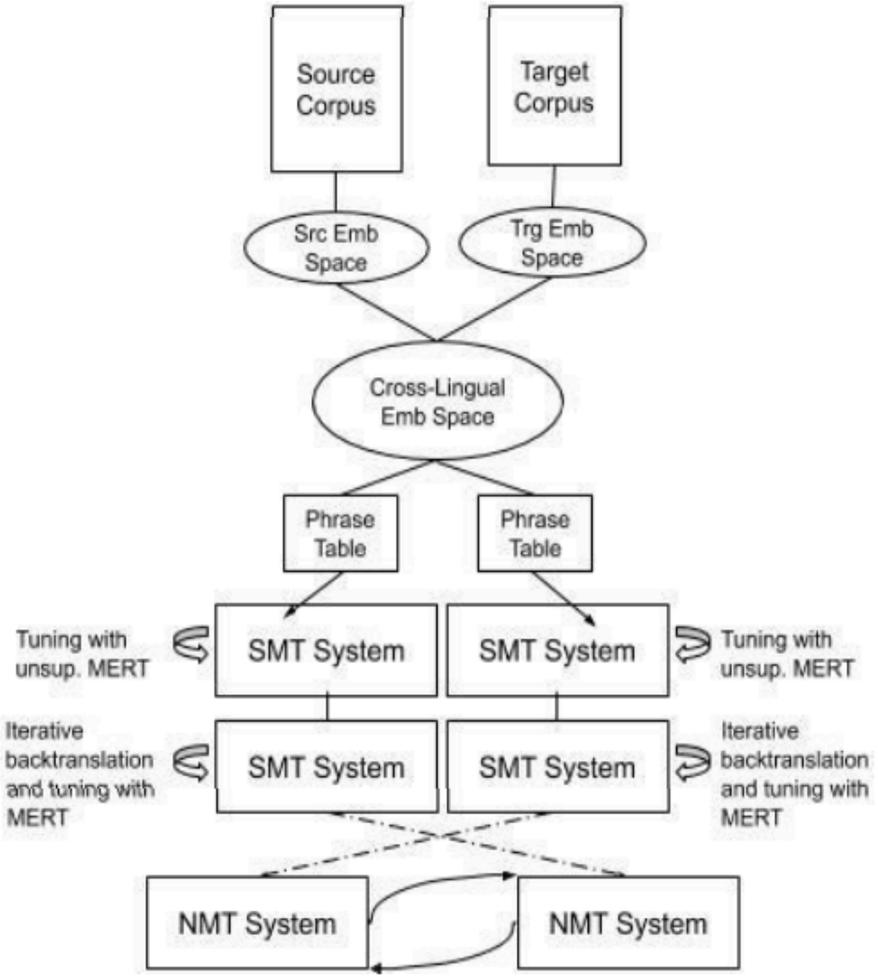
- Unsupervised NMT is possible!
  - Cross-lingual embeddings
  - Iterative Back-Translation
- SOTA - “Hybrid”: SMT bootstrapping, then NMT



From Marchisio et al. (2020)

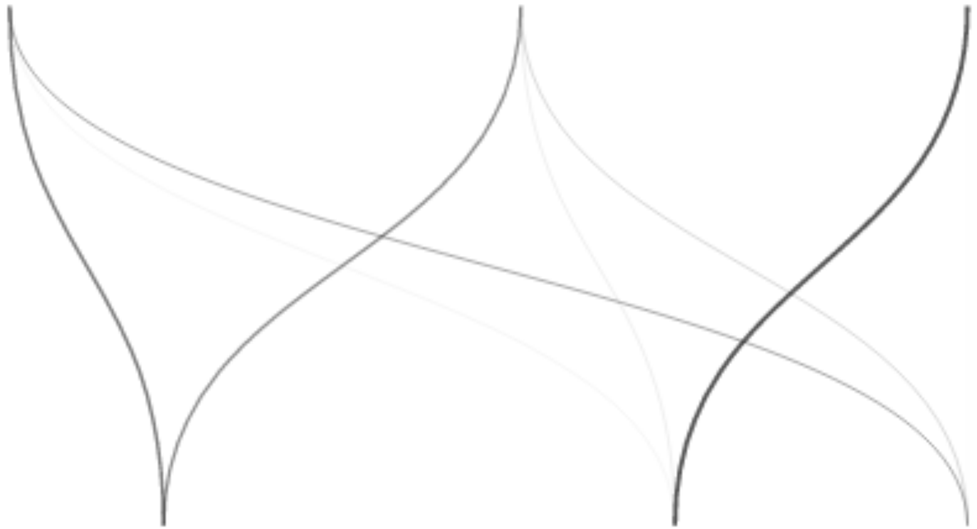
Summary

- Unsupervised NMT is possible!
  - Cross-lingual embeddings
  - Iterative Back-Translation
- SOTA - “Hybrid”: SMT bootstrapping, then NMT
- When does it work - domains, pretraining matters!



From Marchisio et al. (2020)

Any Questions ?



Questions diverses ?