Unsupervised Machine Translation

Guillaume Lample

Facebook AI Research, Université Pierre-et-Marie-Curie

Joint work with Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou

Motivation

- Neural machine translation works well for language pairs with a lot of parallel data (English-French, English-German, etc.)

Motivation

- Neural machine translation works well for language pairs with a lot of parallel data (English-French, English-German, etc.)
- Performance drops when parallel data is scarce
  - Vietnamese, Norwegian, Basque, Ukrainian, Serbian

Motivation

- Neural machine translation works well for language pairs with a lot of parallel data (English-French, English-German, etc.)
- Performance drops when parallel data is scarce
  - Vietnamese, Norwegian, Basque, Ukrainian, Serbian
- The creation of parallel data is difficult, and costly

Motivation

- Neural machine translation works well for language pairs with a lot of parallel data (English-French, English-German, etc.)
- Performance drops when parallel data is scarce
  - Vietnamese, Norwegian, Basque, Ukrainian, Serbian
- The creation of parallel data is difficult, and costly
- Most language pairs use English as a pivot

Motivation

- Neural machine translation works well for language pairs with a lot of parallel data (English-French, English-German, etc.)
- Performance drops when parallel data is scarce
  - Vietnamese, Norwegian, Basque, Ukrainian, Serbian
- The creation of parallel data is difficult, and costly
- Most language pairs use English as a pivot
- However, monolingual data is much easier to find

Questions

- Can we use monolingual data to improve a MT system?

Questions

- Can we use monolingual data to improve a MT system?
- Can we reduce the amount of supervision?

Questions

- Can we use monolingual data to improve a MT system?
- Can we reduce the amount of supervision?
- Can we even learn WITHOUT ANY supervision?

Questions

- Can we use monolingual data to improve a MT system?
- Can we reduce the amount of supervision?
- Can we even learn WITHOUT ANY supervision?

Word translation without parallel data

Guillaume Lample *, Alexis Conneau *, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou (*ICLR 2018)*
Code: https://github.com/facebookresearch/MUSE

Unsupervised Machine Translation Using Monolingual Corpora Only

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato (*ICLR 2018)*

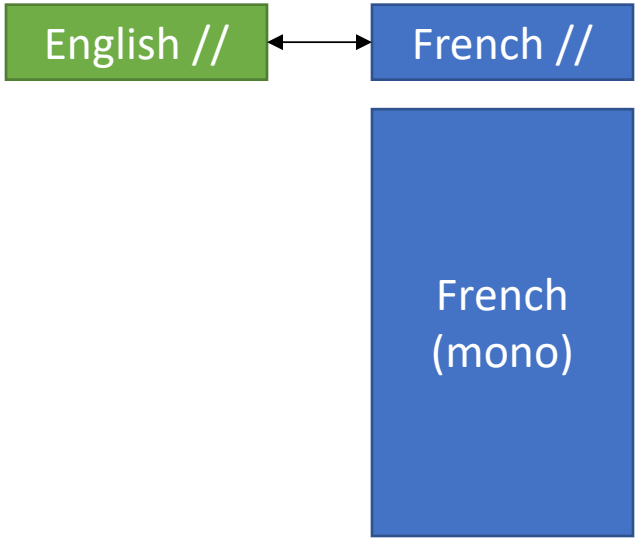
Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)

Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)

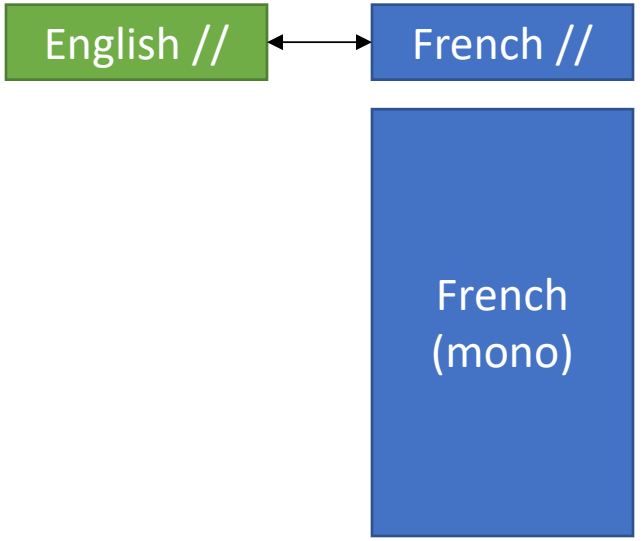
- Small parallel dataset
- Huge monolingual corpus in the target language



Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)

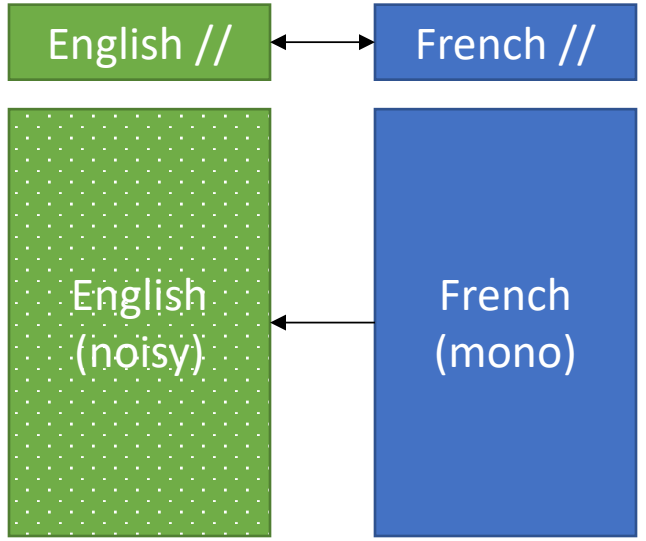
- Small parallel dataset
- Huge monolingual corpus in the target language
- Train a (target → source) model $\boldsymbol{M_{t2s}}$



Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)

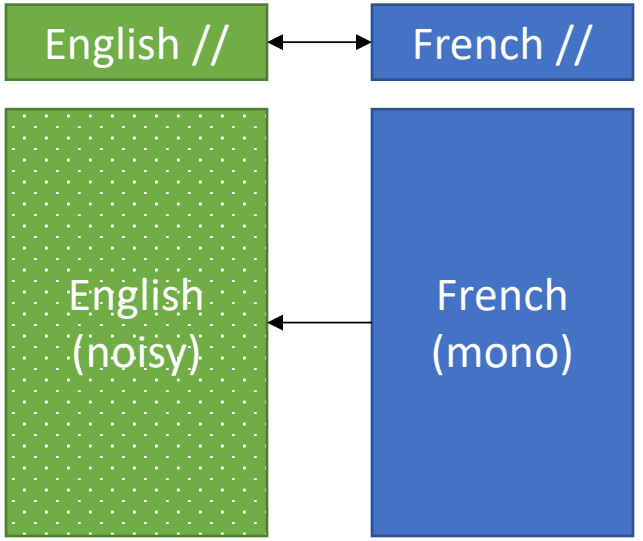
- Small parallel dataset
- Huge monolingual corpus in the target language
- Train a (target → source) model $\boldsymbol{M_{t2s}}$
- Use $\boldsymbol{M_{t2s}}$ to translate the target monolingual corpus



Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)

- Small parallel dataset
- Huge monolingual corpus in the target language
- Train a (target → source) model $\boldsymbol{M_{t2s}}$
- Use $\boldsymbol{M_{t2s}}$ to translate the target monolingual corpus
- Use the two parallel datasets to train $\boldsymbol{M_{s2t}}$



Prior work

- Semi-supervised
  - Back-translation (Sennrich et al., 2015)
  - Dual learning (He et al., 2016)

    - (source → target → source) $\boldsymbol{M}_{t2s}(\boldsymbol{M}_{s2t}(x_s)) = x_s$
    - (target → source → target) $\boldsymbol{M}_{s2t}(\boldsymbol{M}_{t2s}(x_t)) = x_t$

Our approach

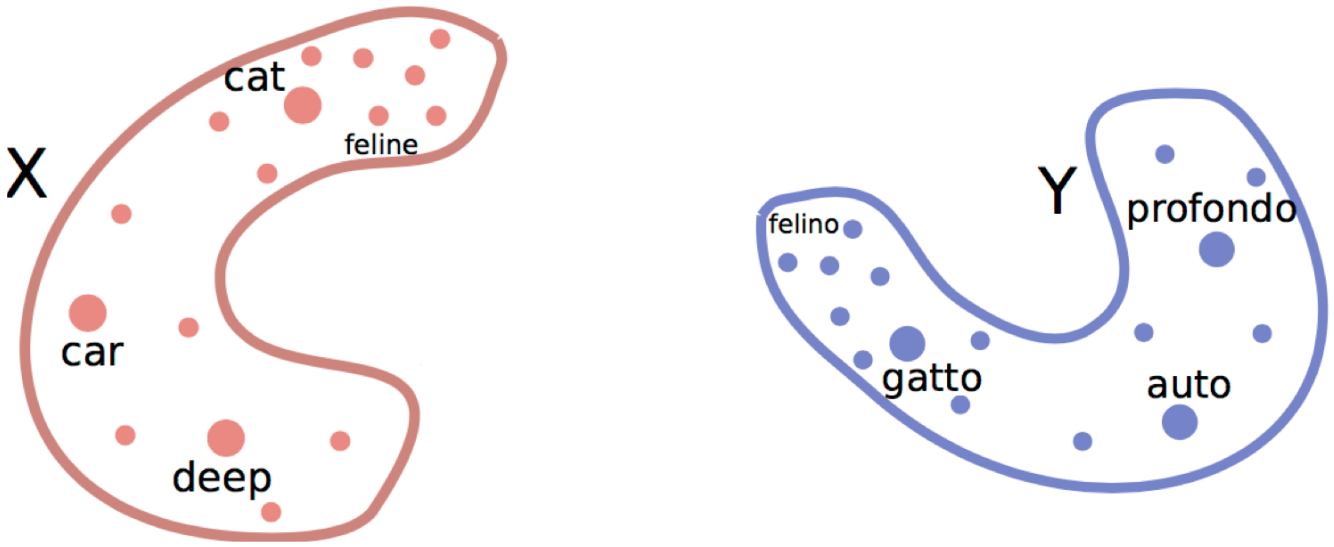
- Start with unsupervised word translation
  - Easier task to start with
  - Already insights of why it could work
  - Can be used as a first step towards unsupervised sentence translation

Weakly-supervised word translation

- Exploiting similarities among languages for machine translation (Mikolov et al., 2013)

Weakly-supervised word translation

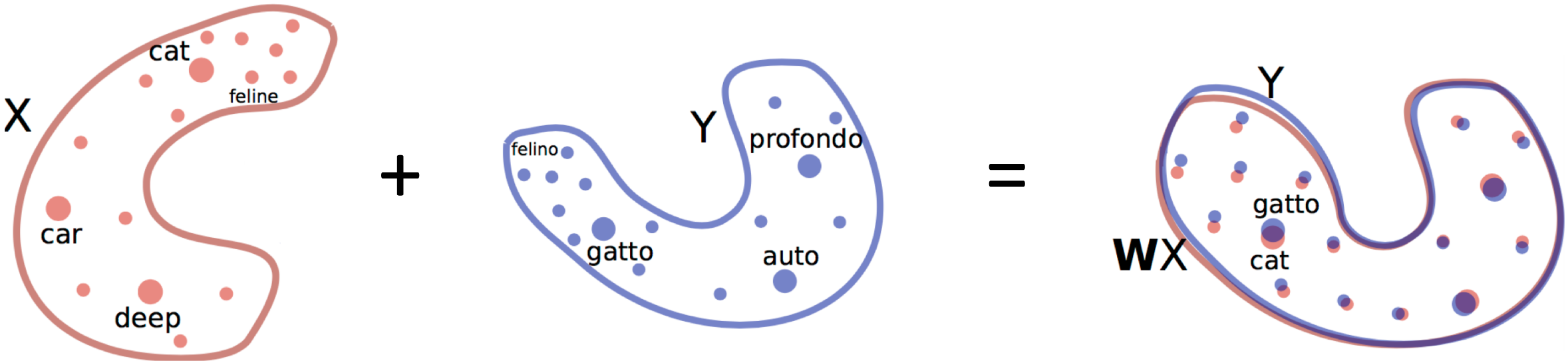
- Exploiting similarities among languages for machine translation (Mikolov et al., 2013)
  - Start from two pre-trained monolingual spaces (word2vec)



- Totally unsupervised
- Widely used
- Strong systems for monolingual embeddings
- Semantically and syntactically relevant
- Not task-specific, useful across domains

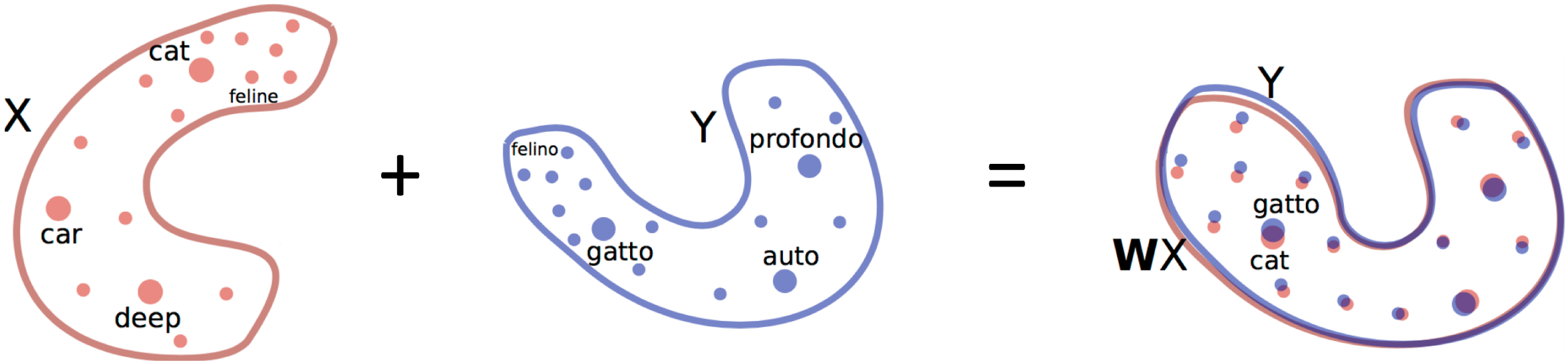
Weakly-supervised word translation

- Exploiting similarities among languages for machine translation (Mikolov et al., 2013)
  - Start from two pre-trained monolingual spaces (word2vec)
  - Project the source space onto the target space using a small dictionary



Weakly-supervised word translation

- Exploiting similarities among languages for machine translation (Mikolov et al., 2013)
  - Start from two pre-trained monolingual spaces (word2vec)
  - Project the source space onto the target space using a small dictionary



- Feed-forward network does not improve over linear mapping (Mikolov et al., 2013)
- Orthogonal projection works best Xing et al. (2015), Smith et al. (2017)

Weakly-supervised word translation

- Linear projection – Mikolov et al. (2013)

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}}$$

Weakly-supervised word translation

- Linear projection – Mikolov et al. (2013)

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}}$$

- Orthogonal projection – Xing et al. (2015), Smith et al. (2017) – **Procrustes**

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

Weakly-supervised word translation

- Linear projection – Mikolov et al. (2013)

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}}$$

- Orthogonal projection – Xing et al. (2015), Smith et al. (2017) – **Procrustes**

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

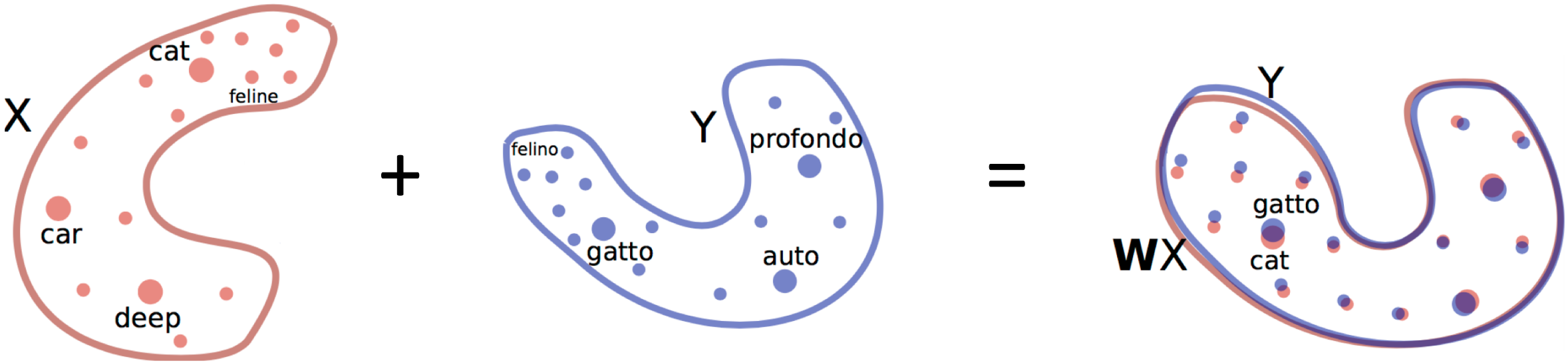
- Given a source word s, define the translation as:

$$t = \operatorname{argmax}_t \cos(Wx_s, y_t)$$

(nearest neighbor according to the cosine distance)

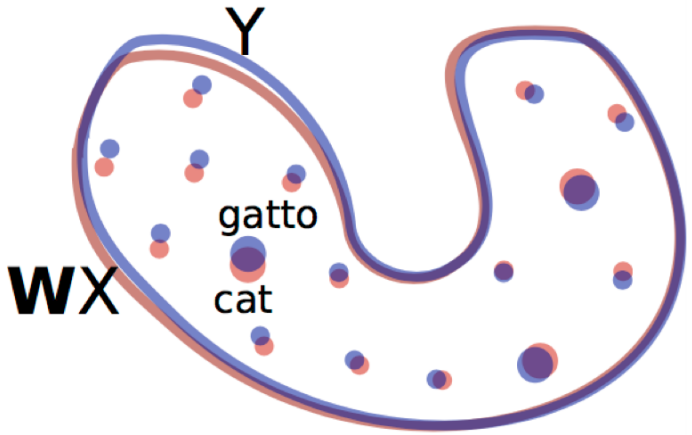
Unsupervised word translation

- Can we find the mapping **W** in an unsupervised way?



Adversarial training

- If ***WX*** and *Y* are perfectly aligned, these spaces should be undistinguishable



Adversarial training

- If ***WX*** and *Y* are perfectly aligned, these spaces should be undistinguishable
- Train a discriminator *D* to discriminate elements from ***WX*** and *Y*

Discriminator training

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right)$$

Adversarial training

- If ***WX*** and *Y* are perfectly aligned, these spaces should be undistinguishable
- Train a discriminator *D* to discriminate elements from ***WX*** and *Y*
- Train ***W*** to unable the discriminator from making accurate predictions

Discriminator training

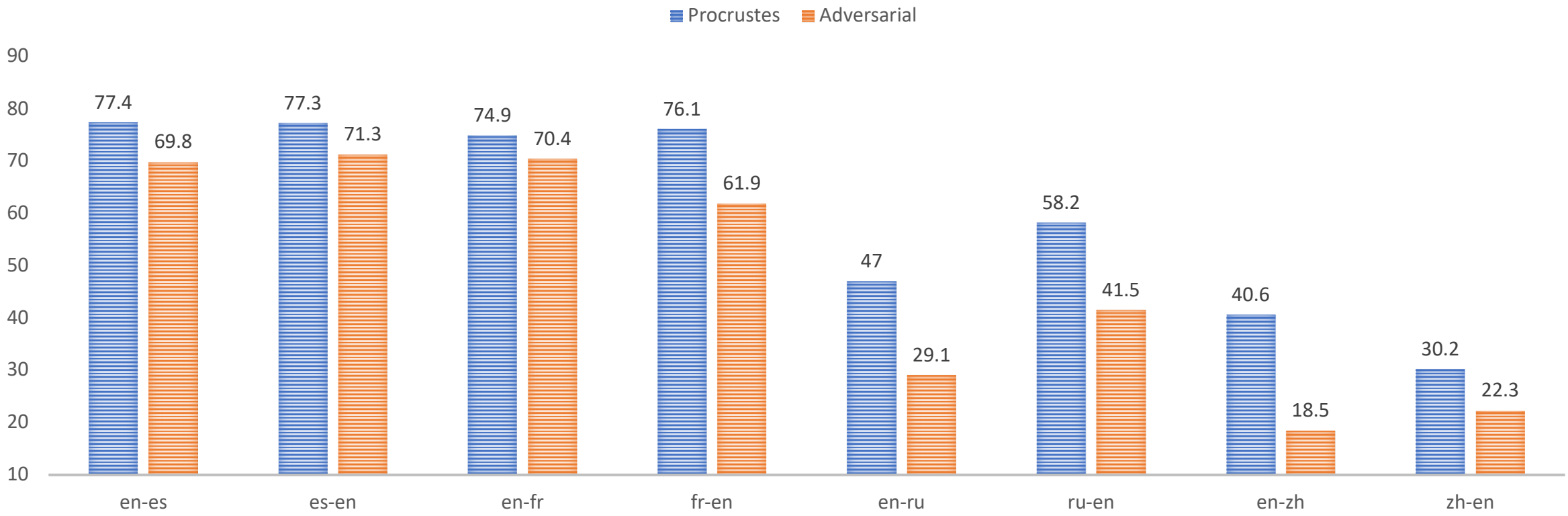
$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right)$$

Mapping training

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 0\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 1\middle|y_i\right)$$

Results on word translation – Adversarial

Results on word translation – Adversarial



Word translation retrieval – P@1 – Adversarial

1.5k source queries, 200k target keys (vocabulary of 200k words for all languages)

Refinement

- With adversarial, the embedding spaces are not properly aligned
  - Nearest neighbors of source words are source words
  - Nearest neighbors of target words are target words

Refinement

- With adversarial, the embedding spaces are not properly aligned
  - Nearest neighbors of source words are source words
  - Nearest neighbors of target words are target words

- Procrustes is better at aligning embedding spaces
  - Cross-lingual nearest neighbors are much closer
  - Uses a dictionary of 5,000 word pairs

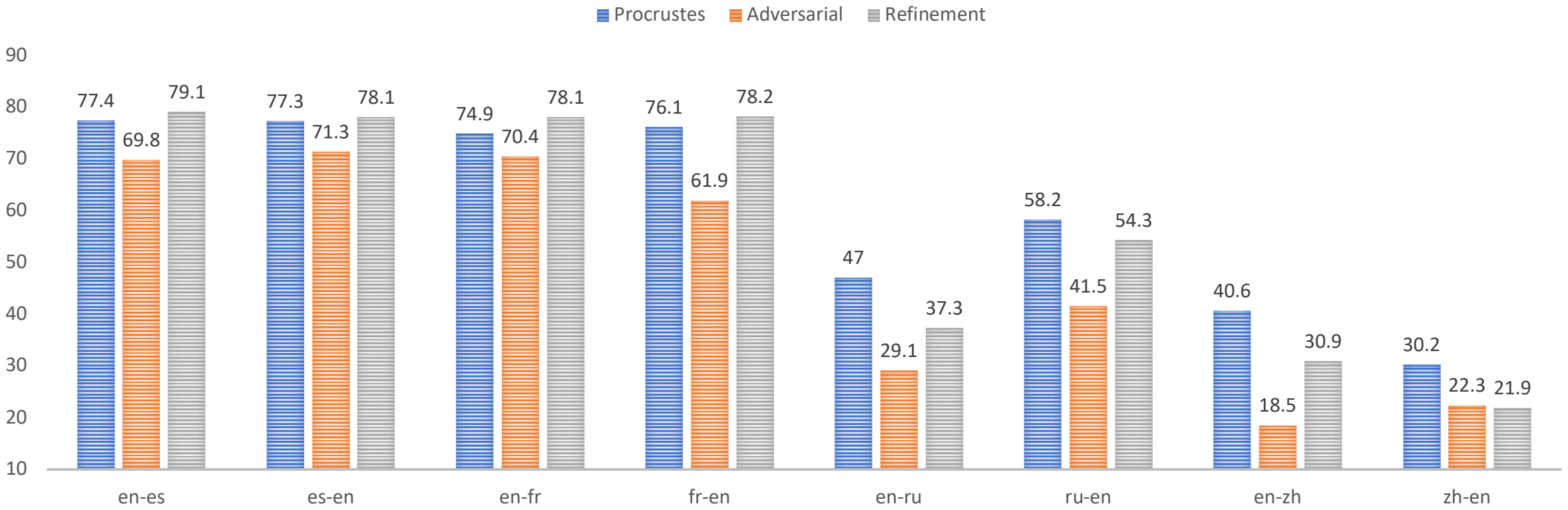
Refinement

- With adversarial, the embedding spaces are not properly aligned
  - Nearest neighbors of source words are source words
  - Nearest neighbors of target words are target words

- Procrustes is better at aligning embedding spaces
  - Cross-lingual nearest neighbors are much closer
  - Uses a dictionary of 5,000 word pairs

- Refinement step:
  - Generate a dictionary of confident pairs (with adversarial)
  - Apply Procrustes on this dictionary
  - Possibly iterate over the refinement step

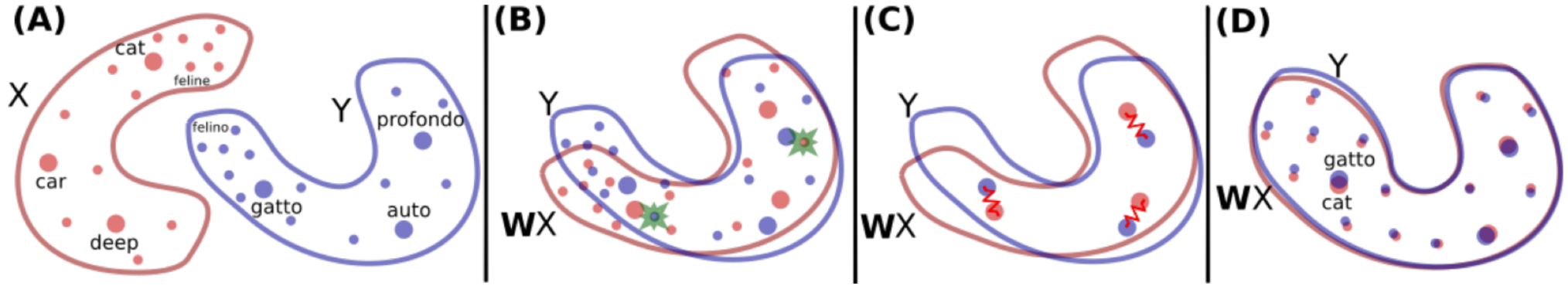
Results on word translation – Refinement



Word translation retrieval – P@1 – Adversarial + Refinement

1.5k source queries, 200k target keys (vocabulary of 200k words for all languages)

Summary



- **(A)** Train monolingual embeddings
- **(B)** Align them using adversarial training
- Refinement step
  - **(C)** Select high-confidence translation pairs
  - **(D)** Apply Procrustes on the generated dictionary
- **Generate translations using CSLS**

Mitigating hubness – CSLS

- Cross-Domain Similarity Local Scaling

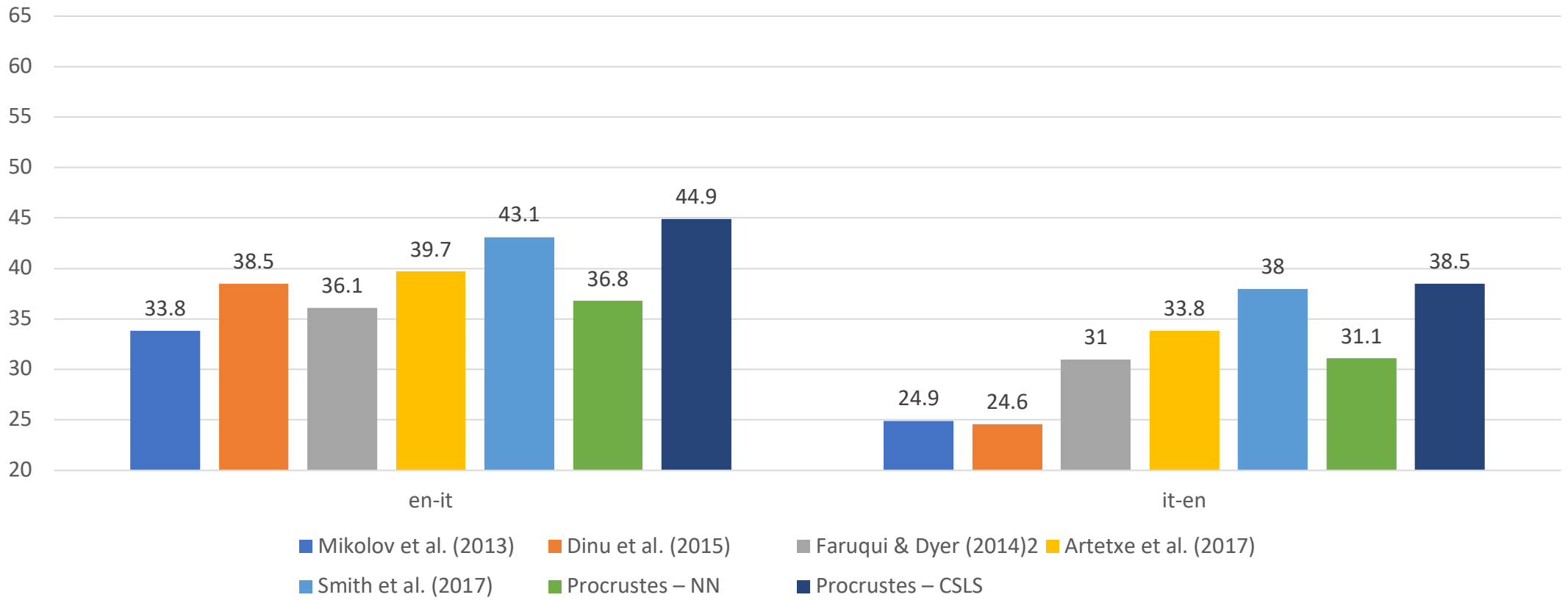
$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t)$$

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t)$$

  - Encourage nearest neighbors reciprocity (empirical observation)
  - Increases the similarity associated with isolated word vectors

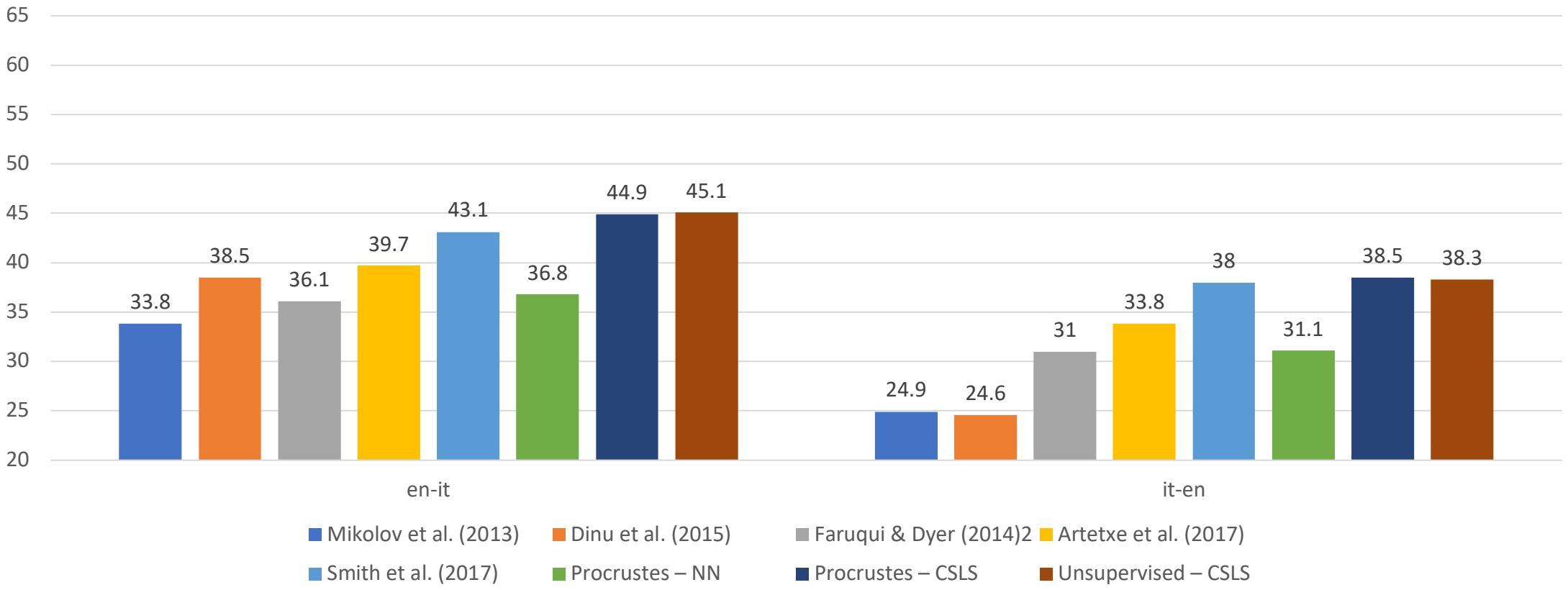
Source word	un	deux	trois	quatre	cinq	six	sept	huit	neuf
Translation – NN	one	two	two	four	two	two	two	four	two
Translation – CSLS	one	two	three	four	five	six	seven	eight	nine

Comparison with previous approaches



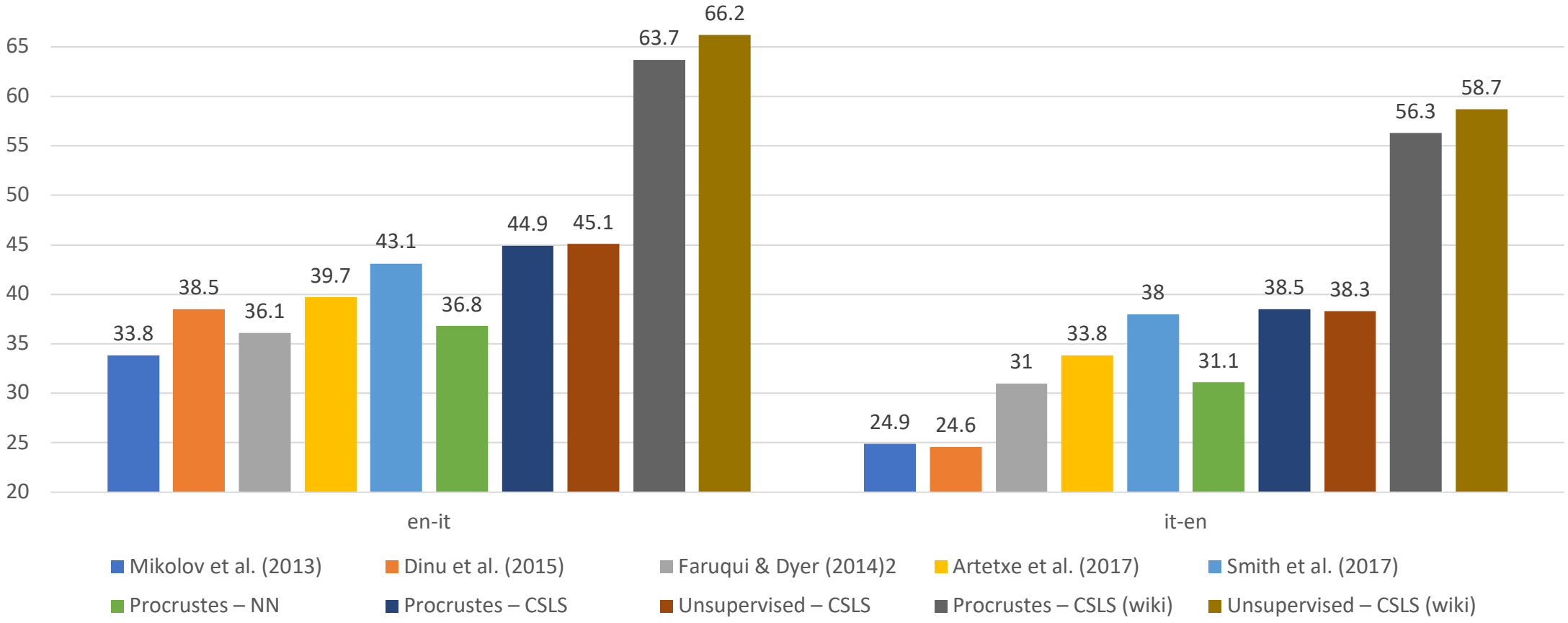
Word retrieval accuracy (P@1)

Comparison with previous approaches



Word retrieval accuracy (P@1)

Comparison with previous approaches

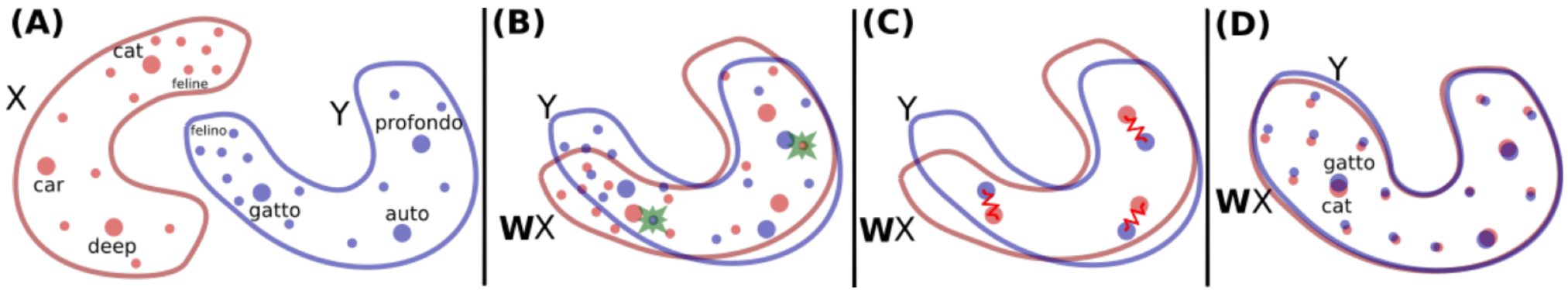


Word retrieval accuracy (P@1)

Unsupervised word translation – Summary

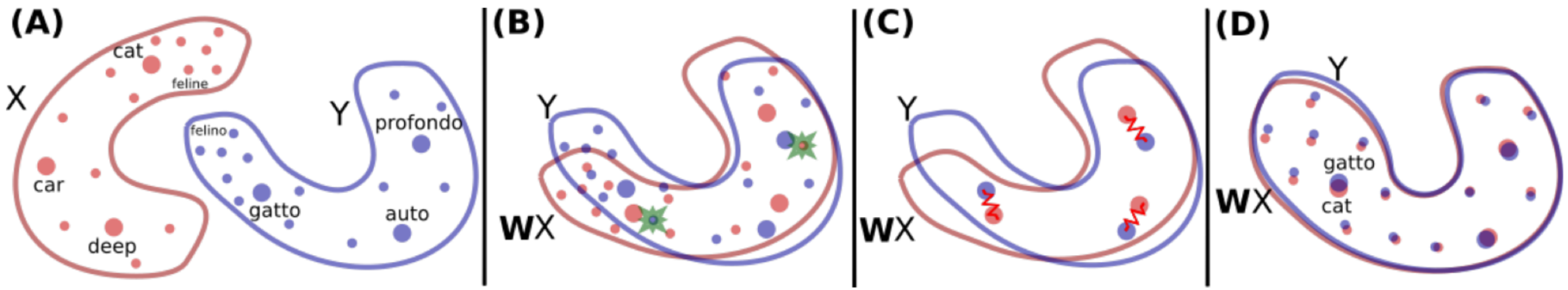
- **Given independent monolingual datasets in a source and a target language:**
  - **We can create high-quality cross-lingual dictionaries**
  - **We can create high-quality cross-lingual embeddings**

Unsupervised sentence translation



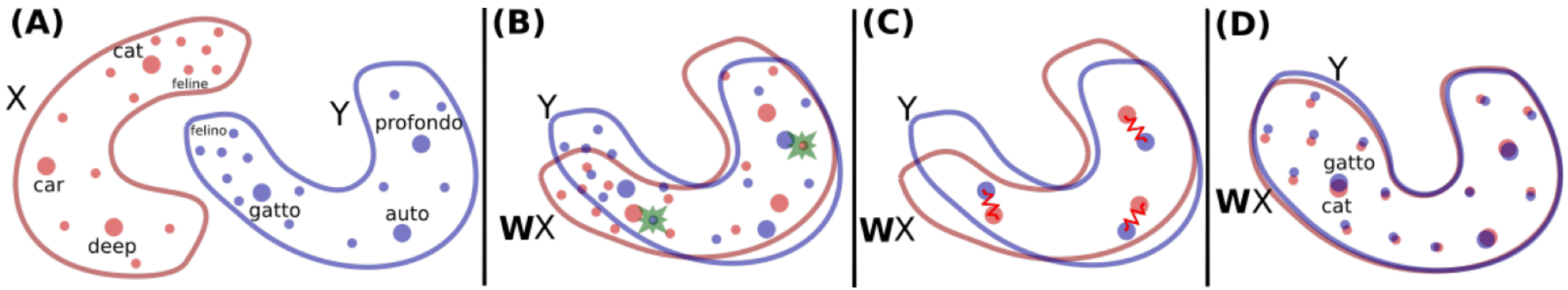
- Could we apply the same unsupervised training procedure to sentences?

Unsupervised sentence translation



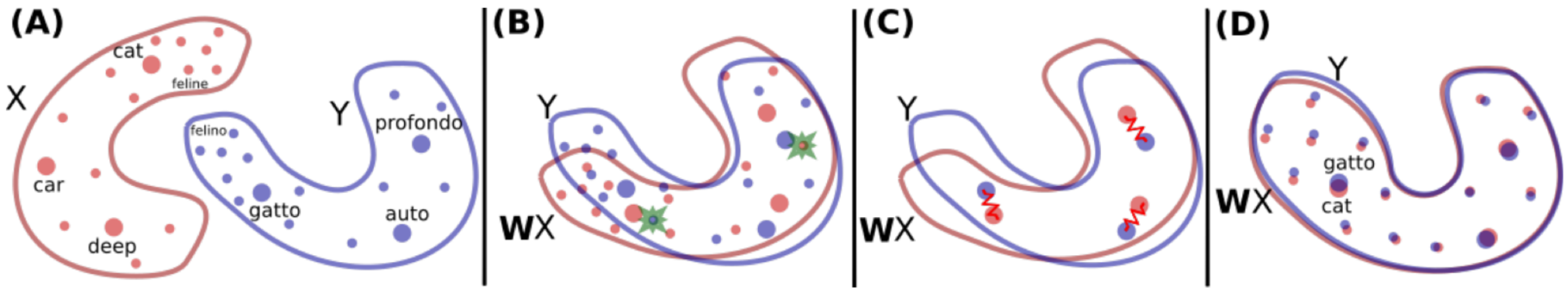
- Could we apply the same unsupervised training procedure to sentences?
  - Number of points grows exponentially with sentence length

Unsupervised sentence translation



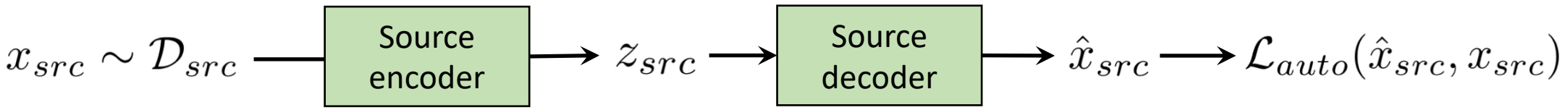
- Could we apply the same unsupervised training procedure to sentences?
  - Number of points grows exponentially with sentence length
  - No similar embedding structures across languages

Unsupervised sentence translation



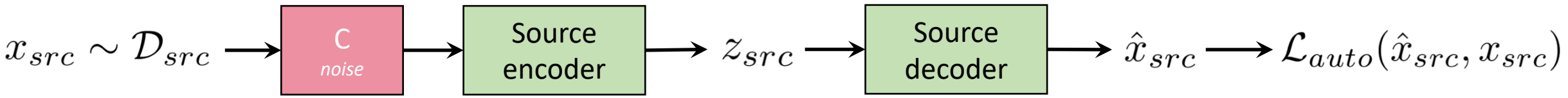
- Could we apply the same unsupervised training procedure to sentences?
  - Number of points grows exponentially with sentence length
  - No similar embedding structures across languages
  - Direct application does not work (even in a supervised setting)

Proposed architecture – Denoising Auto-Encoding



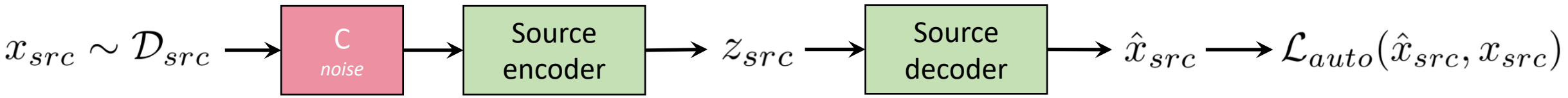
- Train a source → source denoising autoencoder (DAE)

Proposed architecture – Denoising Auto-Encoding



- Train a source → source denoising autoencoder (DAE)
- Critical to add noise to avoid trivial reconstructions

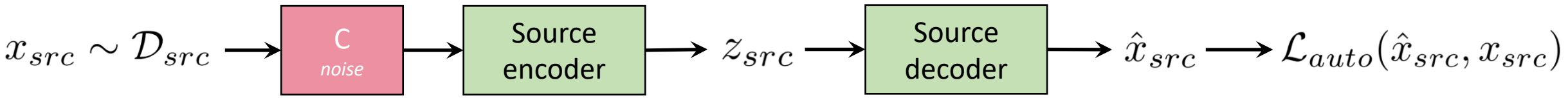
Proposed architecture – Denoising Auto-Encoding



- Train a source → source denoising autoencoder (DAE)
- Critical to add noise to avoid trivial reconstructions
- Two sources of noise
  - Word dropout: each word is removed with a probability ***p*** (usually 0.1)

Ref: *Arizona was the first to introduce such a requirement .*
Arizona was the first to such a requirement .
Arizona was first to introduce such a requirement .

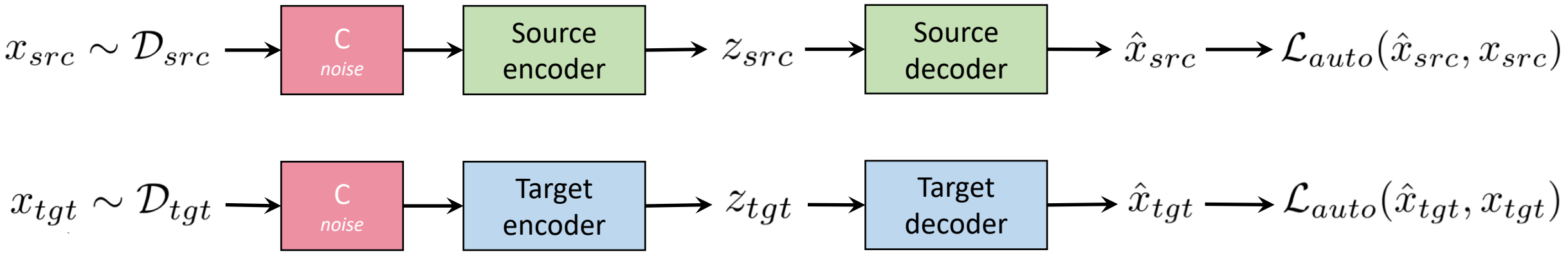
Proposed architecture – Denoising Auto-Encoding



- Train a source → source denoising autoencoder (DAE)
- Critical to add noise to avoid trivial reconstructions
- Two sources of noise
  - Word dropout: each word is removed with a probability ***p*** (usually 0.1)
  - Word shuffle: word order is (slightly) shuffled inside sentences $|\sigma(i) - i| \leq k$

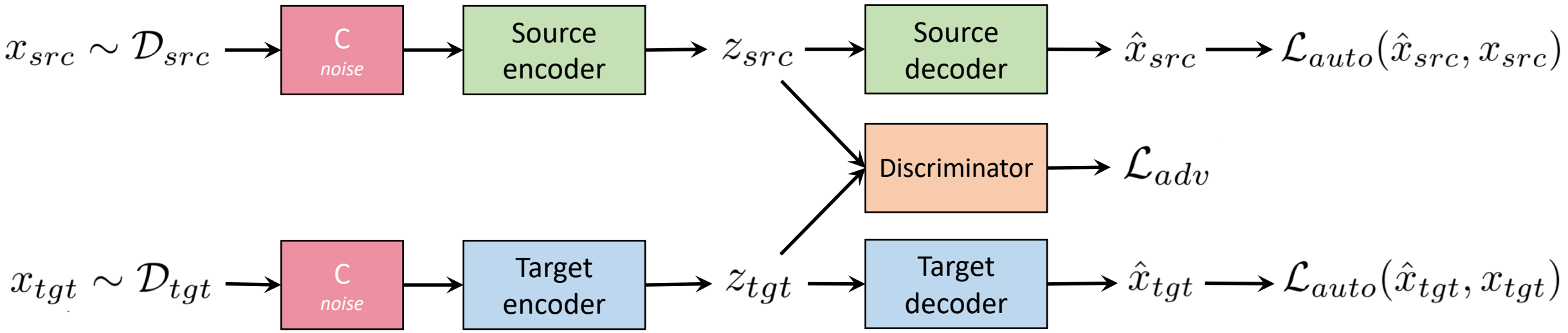
Ref: *Arizona was the first to introduce such a requirement .*
Arizona the first was to introduce a requirement such.
Arizona was the to introduce first such requirement a .

Proposed architecture – Denoising Auto-Encoding



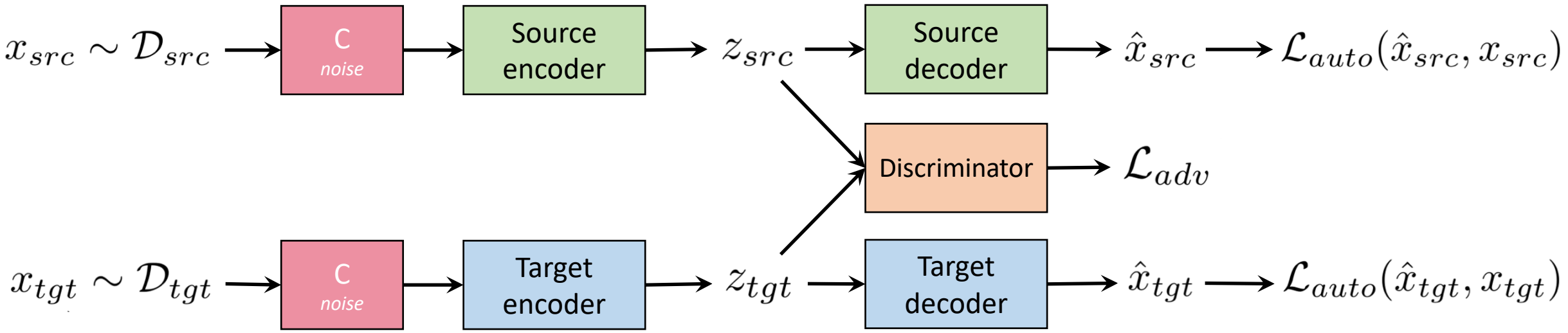
- Train a source → source denoising autoencoder (DAE)
- Train a target → target denoising autoencoder (DAE)

Proposed architecture – Denoising Auto-Encoding



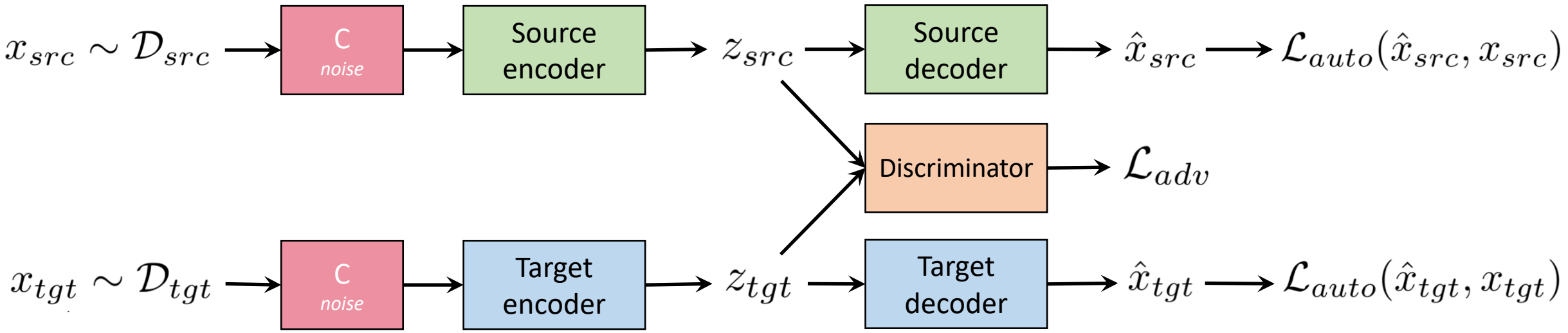
- Train a source → source denoising autoencoder (DAE)
- Train a target → target denoising autoencoder (DAE)
- Make source and target latent states indistinguishable using adversarial training

Proposed architecture – Denoising Auto-Encoding



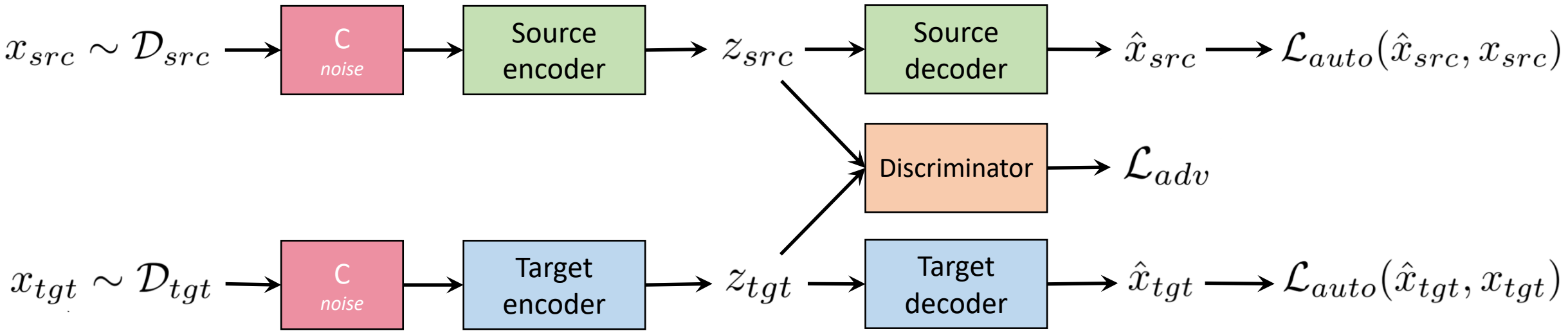
- Train a source → source denoising autoencoder (DAE)
- Train a target → target denoising autoencoder (DAE)
- Make source and target latent states indistinguishable using adversarial training
- We want decoders to operate in the same space → share parameters between encoders

Proposed architecture – Denoising Auto-Encoding



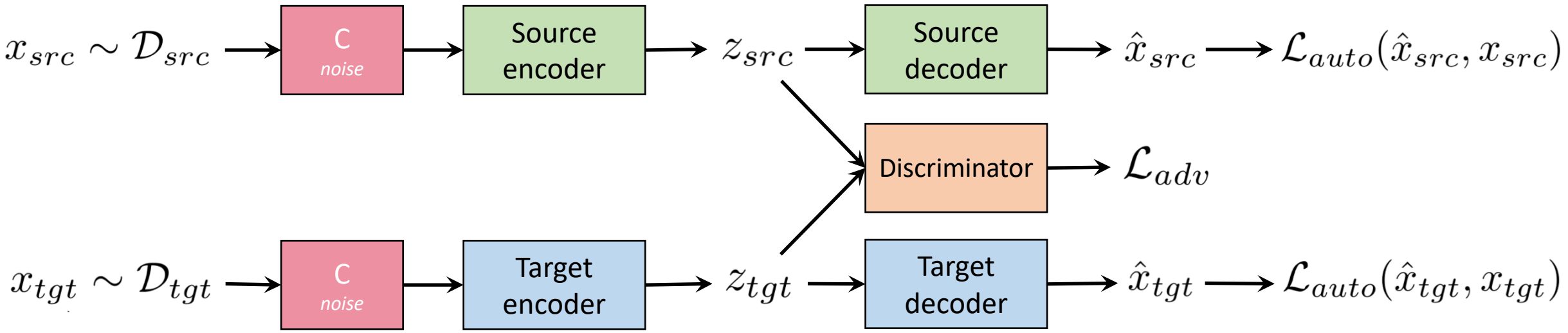
- Works on simple / small datasets, with short sentences or small vocabulary

Proposed architecture – Denoising Auto-Encoding



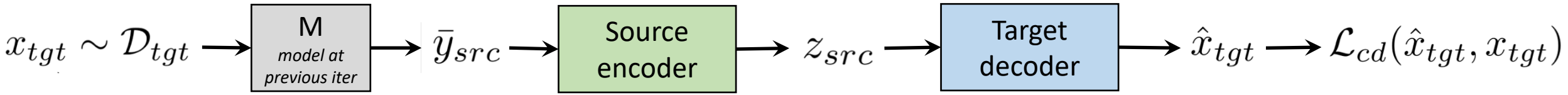
- Works on simple / small datasets, with short sentences or small vocabulary
- Problem: at test time we want (source → target) or (target → source)

Proposed architecture – Denoising Auto-Encoding



- Works on simple / small datasets, with short sentences or small vocabulary
- Problem: at test time we want (source → target) or (target → source)
- Cross-Domain training: train the model to perform actual translations
  - We do not have parallel data → generate artificial translations for training

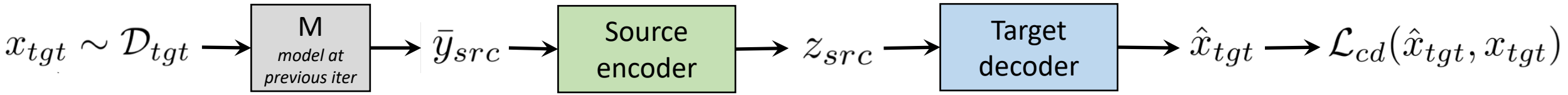
Proposed architecture – Cross-Domain Training



- Train on pairs generated using a stale version of the model
  - Start with word-by-word translation

x_{tgt} *une photo d' une rue bondée en ville .* **(sentence from monolingual corpus)**

Proposed architecture – Cross-Domain Training

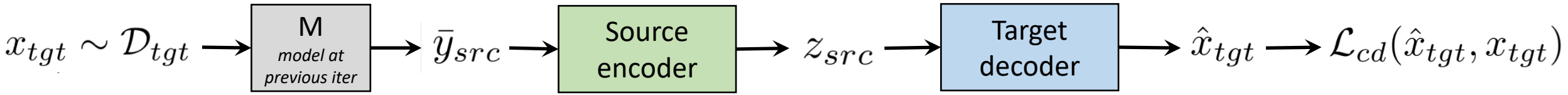


- Train on pairs generated using a stale version of the model
  - Start with word-by-word translation

x_{tgt} *une photo d' une rue bondée en ville .* **(sentence from monolingual corpus)**

$\bar{y}_{src}$ *a photo of a street crowded in a city .* **(word-by-word translation)**

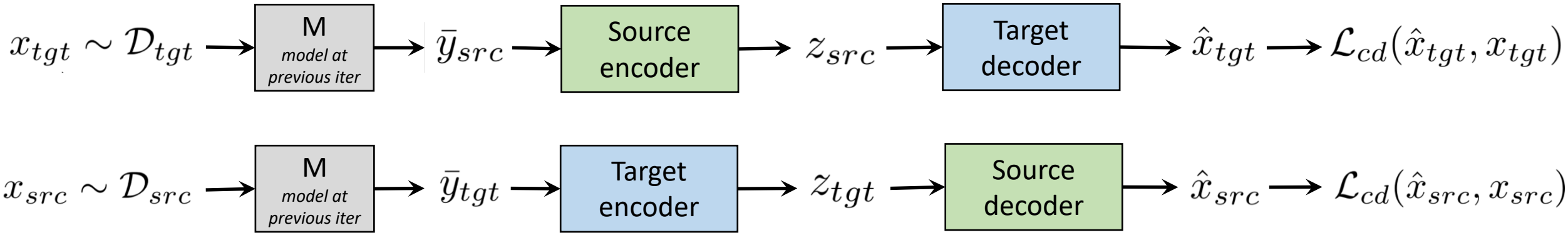
Proposed architecture – Cross-Domain Training



- Train on pairs generated using a stale version of the model
  - Start with word-by-word translation

x_{tgt}	*une photo d' une rue bondée en ville .*	**(sentence from monolingual corpus)**
$\bar{y}_{src}$	*a photo of a street crowded in a city .*	**(word-by-word translation)**
y_{src}	*a view of a crowded city street .*	**(gold translation)**

Proposed architecture – Cross-Domain Training



- Train on pairs generated using a stale version of the model
  - Start with word-by-word translation
- Symmetric training

Recap

- Denoising autoencoding to learn good sentence representations $\mathcal{L}_{auto}$

Recap

- Denoising autoencoding to learn good sentence representations $\mathcal{L}_{auto}$
- Match distributions of latent features across the two domains
  - Adversarial training $\mathcal{L}_{adv}$
  - Parameter sharing

Recap

- Denoising autoencoding to learn good sentence representations $\mathcal{L}_{auto}$
- Match distributions of latent features across the two domains
  - Adversarial training $\mathcal{L}_{adv}$
  - Parameter sharing
- Cross-lingual training to learn to translate $\mathcal{L}_{cd}$
  - Trick: use stale version of the model to produce a noisy source
  - Use a word-by-word translation model to initialize the algorithm

Recap

- Denoising autoencoding to learn good sentence representations $\mathcal{L}_{auto}$
- Match distributions of latent features across the two domains
  - Adversarial training $\mathcal{L}_{adv}$
  - Parameter sharing
- Cross-lingual training to learn to translate $\mathcal{L}_{cd}$
  - Trick: use stale version of the model to produce a noisy source
  - Use a word-by-word translation model to initialize the algorithm
- Pretrain word embeddings with aligned embeddings

Unsupervised model selection

Unsupervised model selection

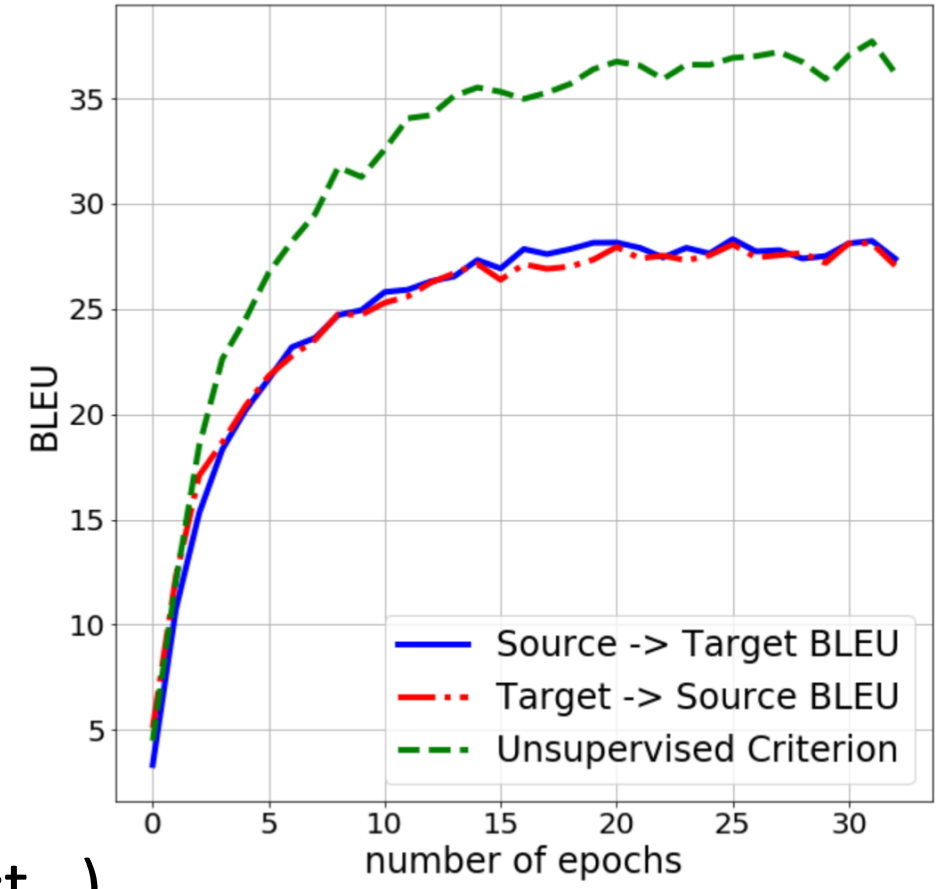
- We can not use a validation set with parallel sentences
- We use monolingual datasets instead: $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$
- We define the following unsupervised criterion:

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

Unsupervised model selection

- We can not use a validation set with parallel sentences
- We use monolingual datasets instead: $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$
- We define the following unsupervised criterion:

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

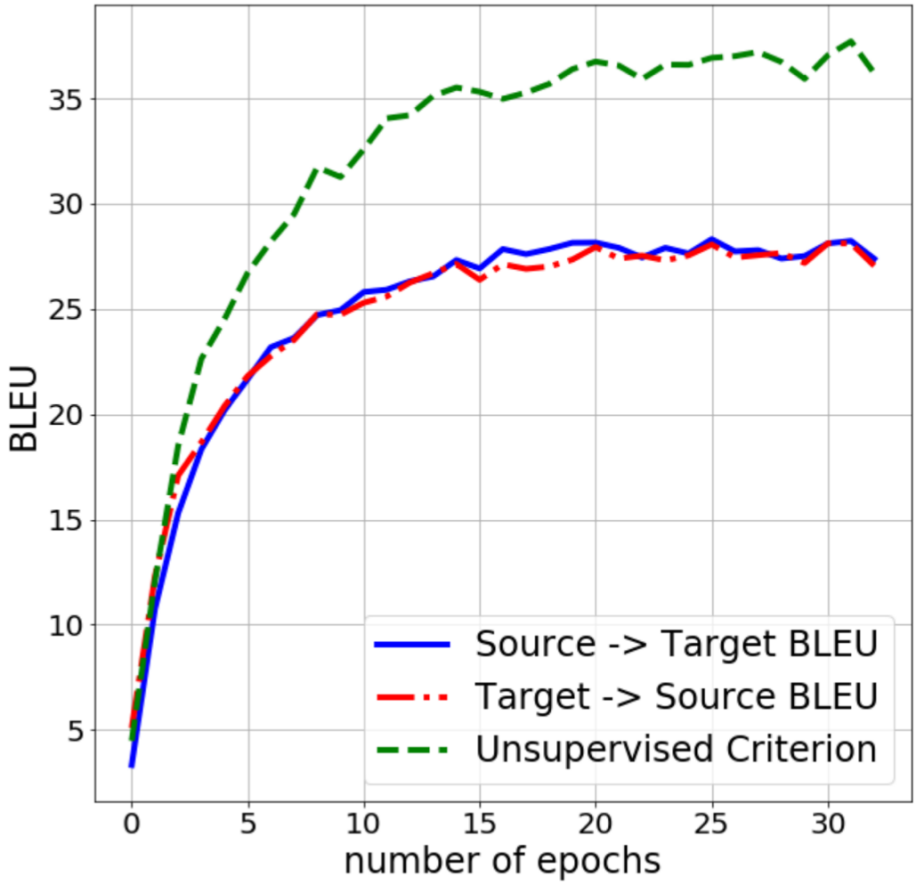


- Correlation is good (but far from perfect…)

Unsupervised model selection

- We can not use a validation set with parallel sentences
- We use monolingual datasets instead: $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$
- We define the following unsupervised criterion:

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



Datasets

- Multi30k-Task1 (MMT1) (en-fr / en-de)
  - 29k images with English, French and German captions
  - 14.5k monolingual datasets
  - 1k sentences in the evaluation set

Datasets

- Multi30k-Task1 (MMT1) (en-fr / en-de)
  - 29k images with English, French and German captions
  - 14.5k monolingual datasets
  - 1k sentences in the evaluation set

- WMT 2014 (en-fr)
  - Sentences of length ≤ 50 (30M pairs)
  - 15M monolingual datasets
  - Evaluation on the full newstest2014

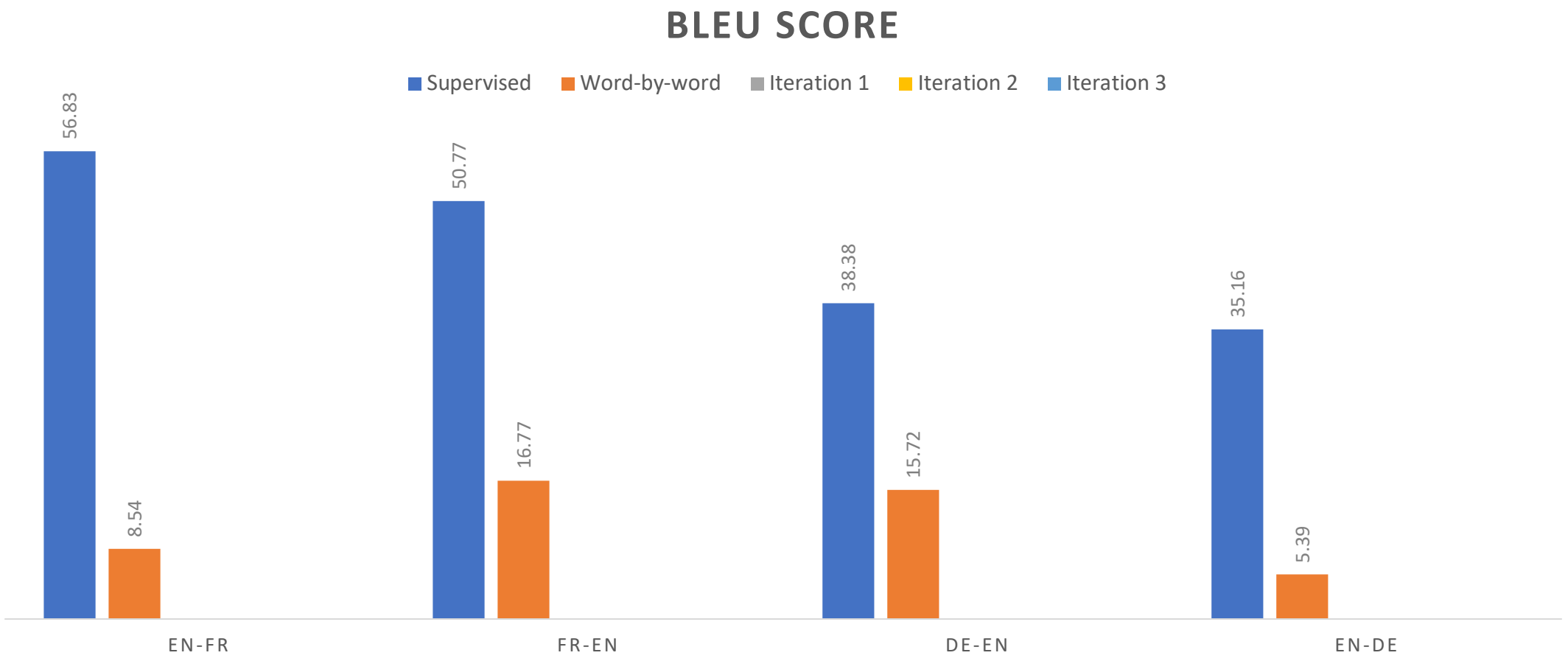
Datasets

- Multi30k-Task1 (MMT1) (en-fr / en-de)
  - 29k images with English, French and German captions
  - 14.5k monolingual datasets
  - 1k sentences in the evaluation set

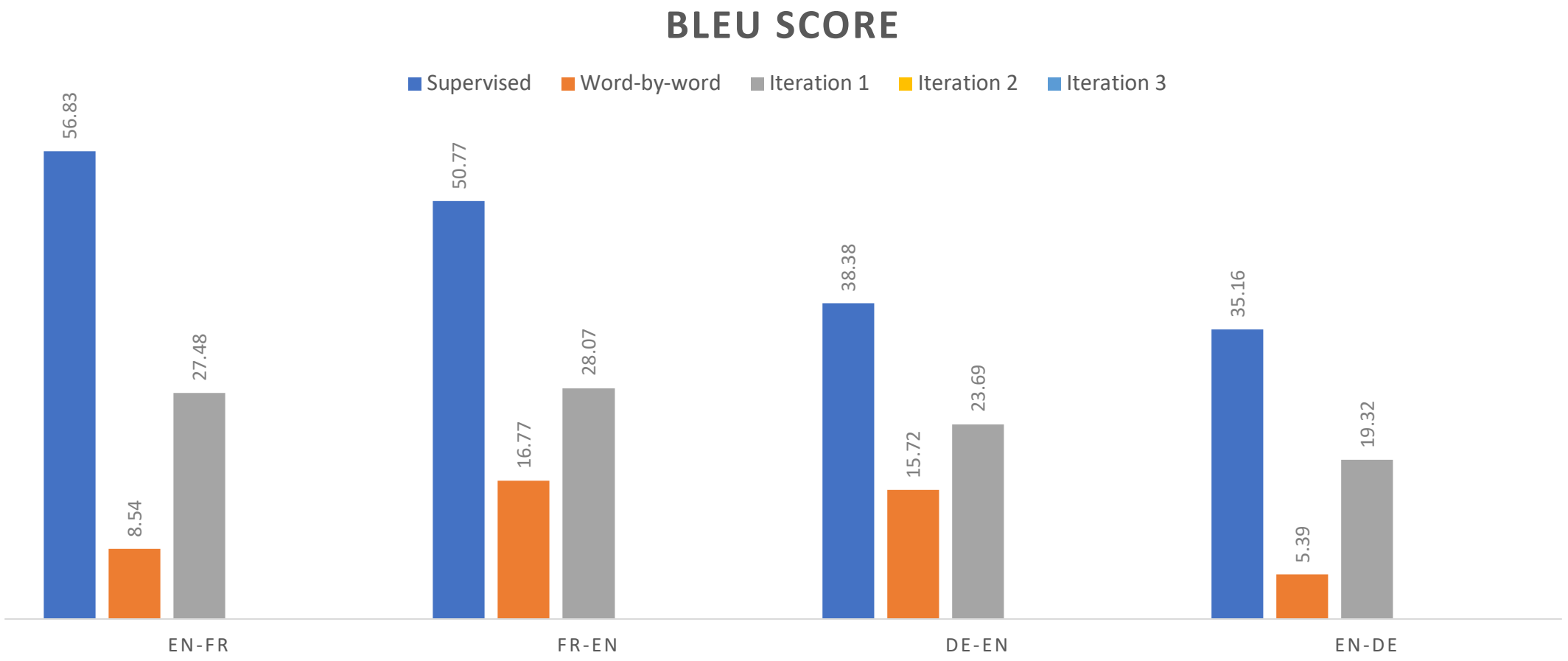
- WMT 2014 (en-fr)
  - Sentences of length ≤ 50 (30M pairs)
  - 15M monolingual datasets
  - Evaluation on the full newstest2014

- WMT 2016 (en-de)
  - Same setting, 3.6M pairs
  - 1.8M monolingual datasets

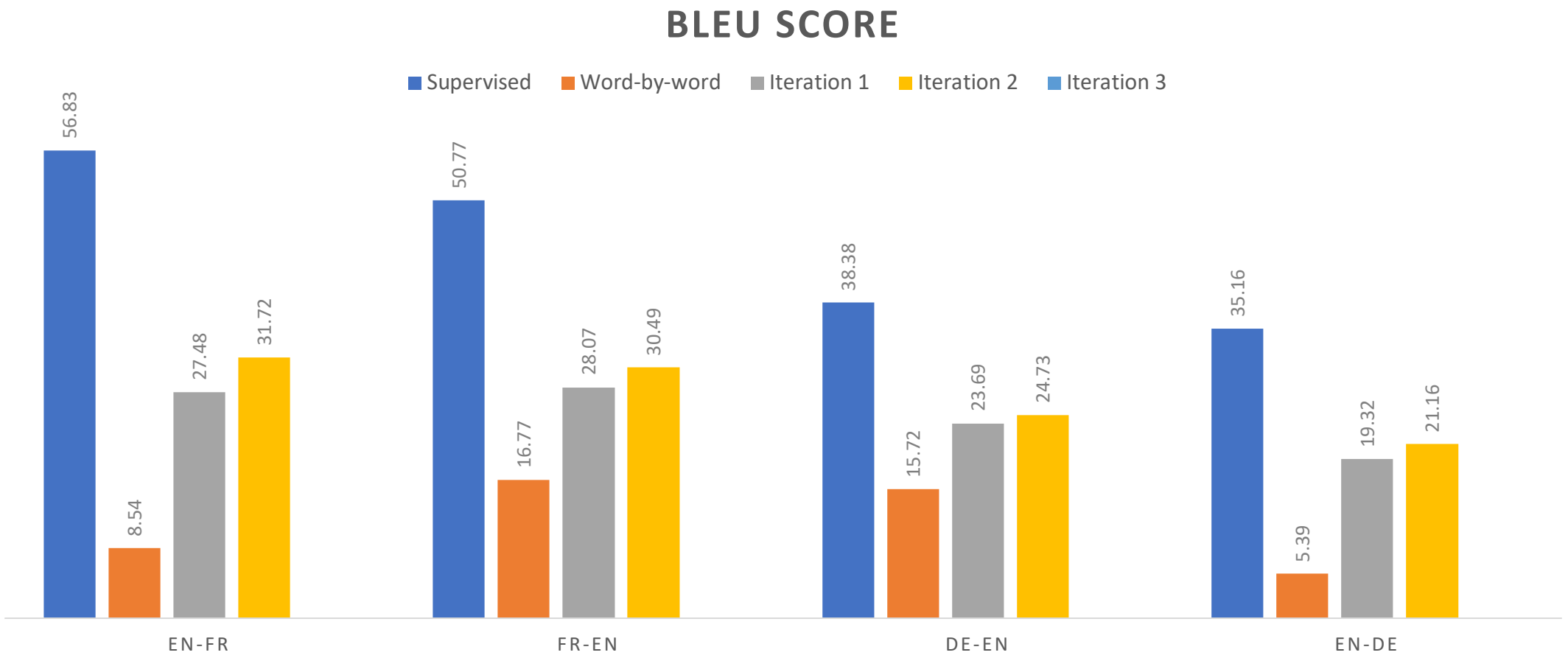
Results on MMT1 - Baseline



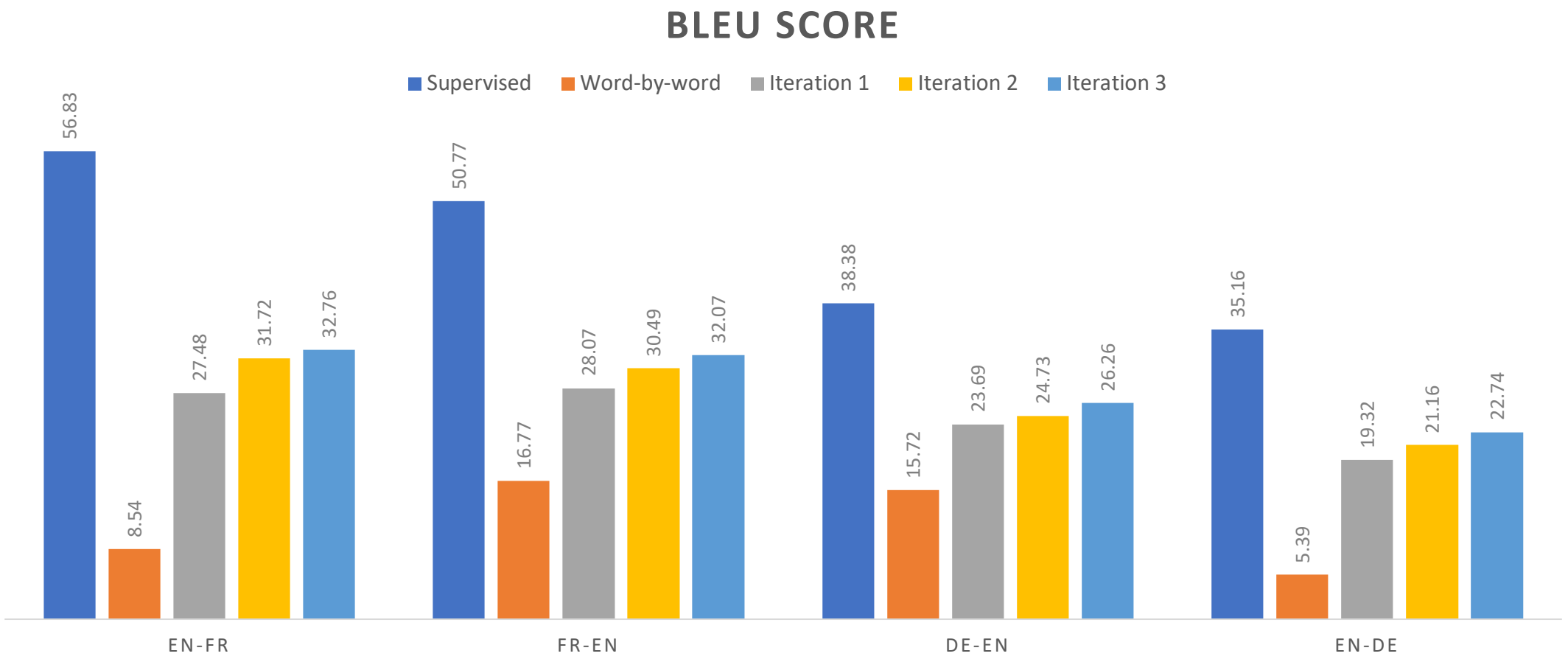
Results on MMT1



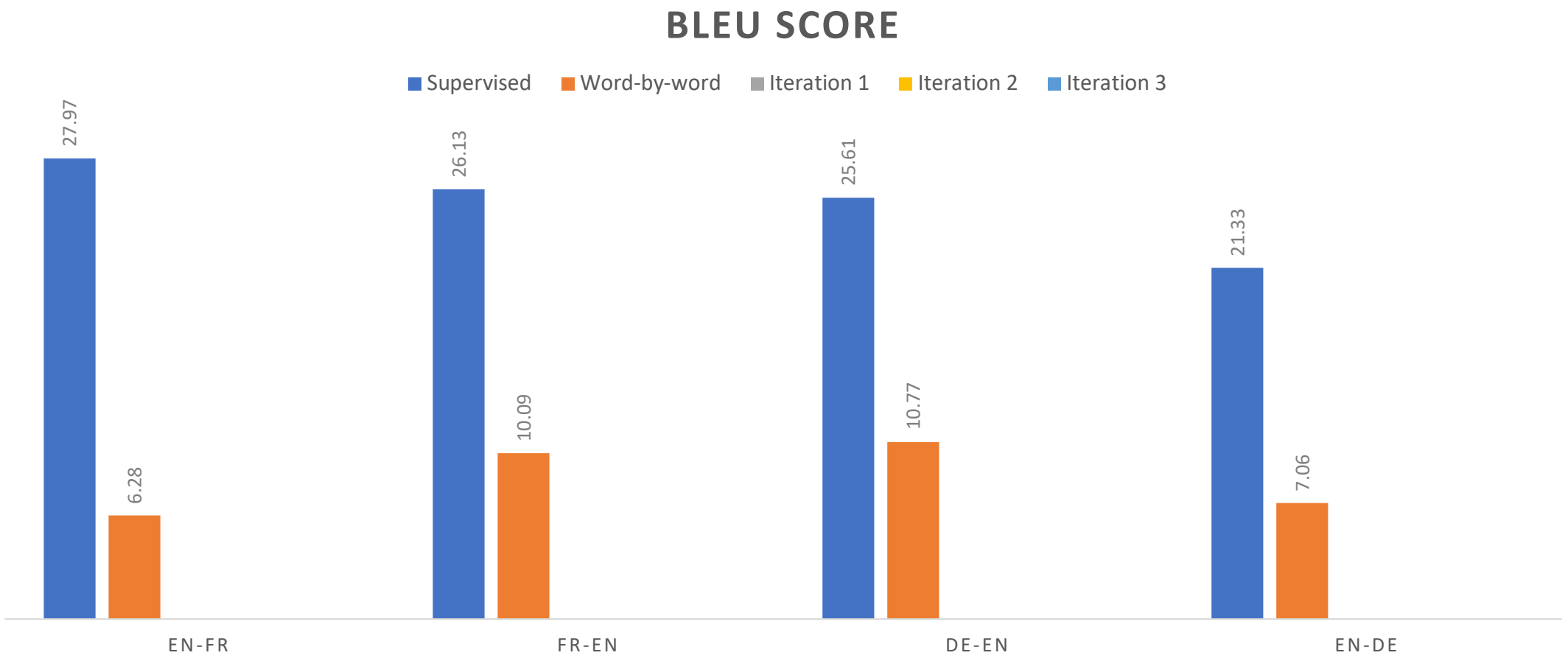
Results on MMT1



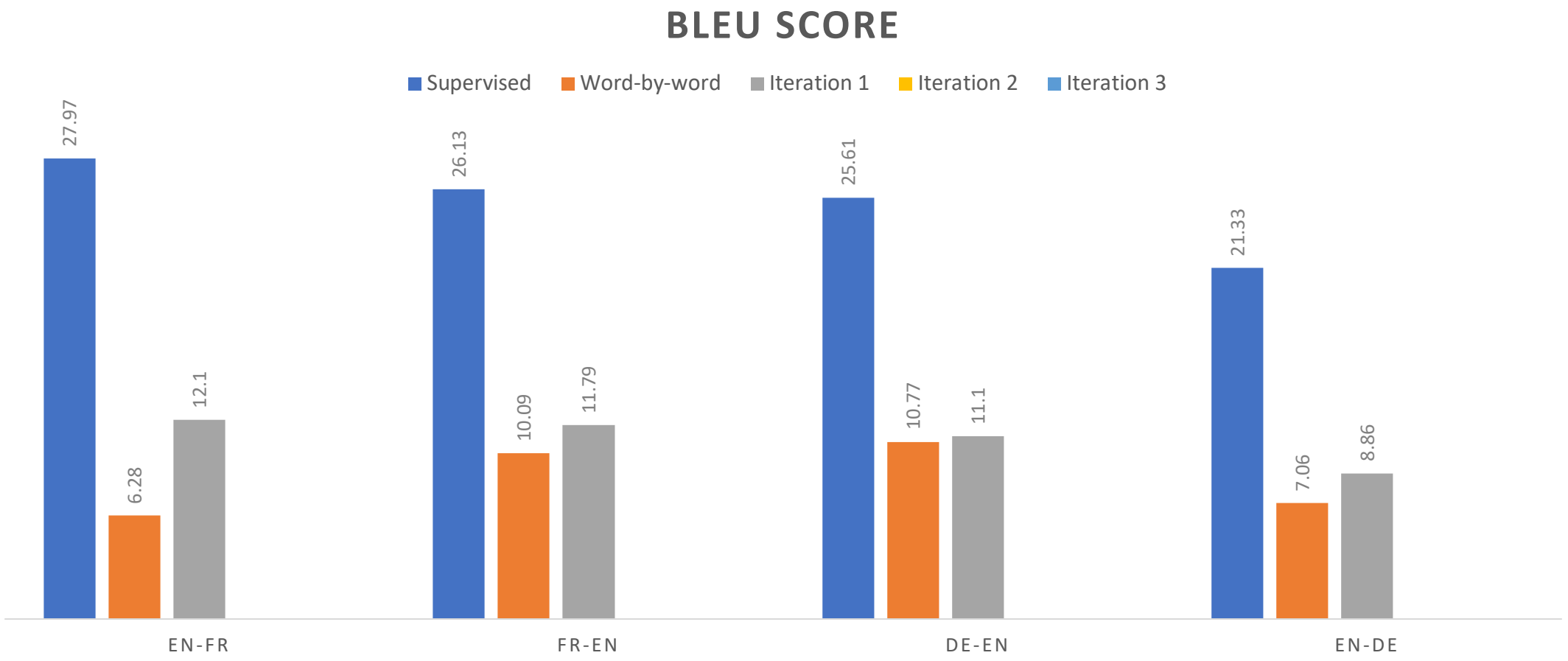
Results on MMT1



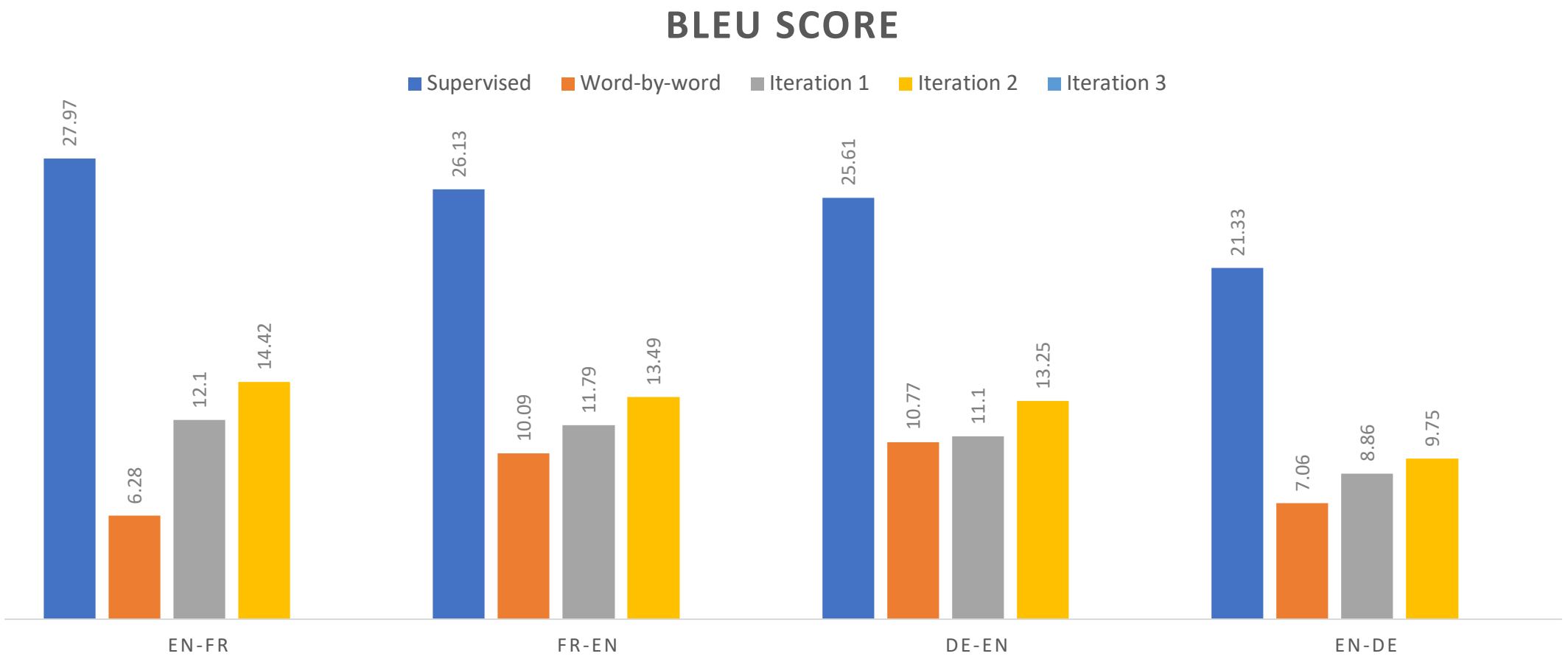
Results on WMT - Baseline



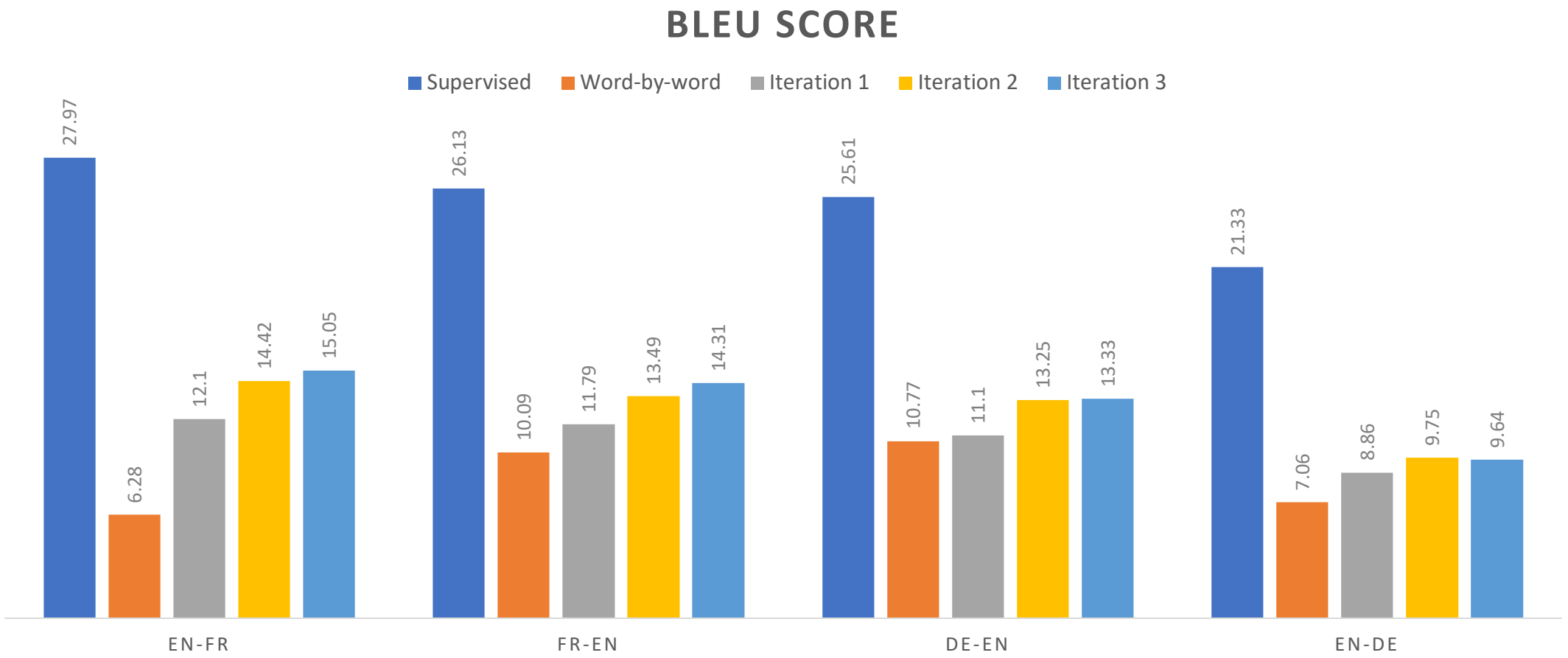
Results on WMT



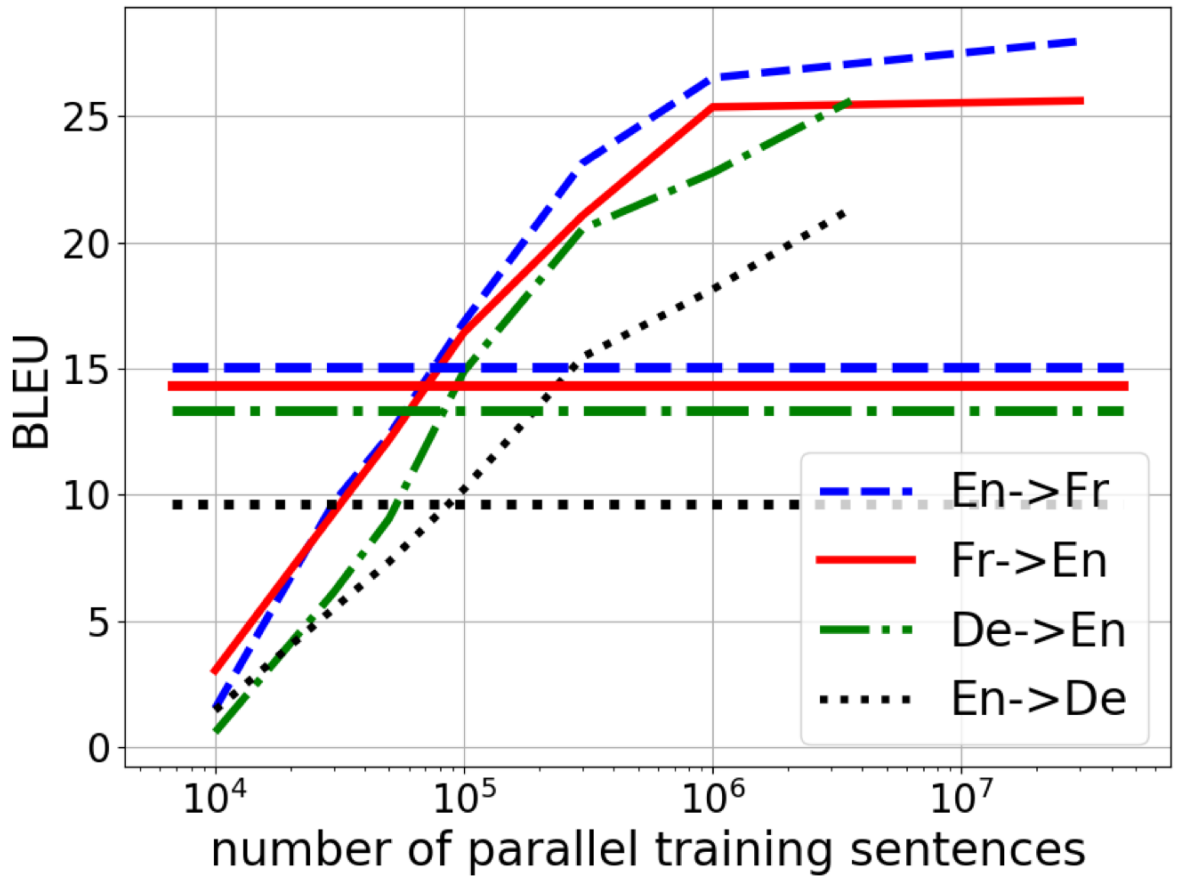
Results on WMT



Results on WMT



Comparison with supervised methods



Examples of unsupervised translations

Source	**une femme aux cheveux roses habillée en noir parle à un homme .**
Iteration 0	
Iteration 1	
Iteration 2	
Iteration 3	
Reference	**a woman with pink hair dressed in black talks to a man .**

Examples of unsupervised translations

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	
Iteration 2	
Iteration 3	
Reference	**a woman with pink hair dressed in black talks to a man .**

Examples of unsupervised translations

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	
Iteration 3	
Reference	**a woman with pink hair dressed in black talks to a man .**

Examples of unsupervised translations

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	
Reference	**a woman with pink hair dressed in black talks to a man .**

Examples of unsupervised translations

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 .**

Results on MMT1 – Ablation study

	en-fr	fr-en	de-en	en-de
Full	27.48	28.07	23.69	19.32

Results on MMT1 – Ablation study

	en-fr	fr-en	de-en	en-de
$\lambda_{cd} = 0$	25.44	27.14	20.56	14.42
Full	27.48	28.07	23.69	19.32

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)

Results on MMT1 – Ablation study

	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
Full	27.48	28.07	23.69	19.32

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)

Results on MMT1 – Ablation study

	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
Full	27.48	28.07	23.69	19.32

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)

Results on MMT1 – Ablation study

	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

Full	27.48	28.07	23.69	19.32

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)

Results on MMT1 – Ablation study

	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
Full	27.48	28.07	23.69	19.32

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)
- λ_{auto} auto-encoding loss coefficient (λ_{auto} = 0 : no denoising auto-encoding)

Results on MMT1 – Ablation study

	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

- λ_{cd} cross-domain loss coefficient (λ_{cd} = 0 : no cross domain denoising)
- λ_{auto} auto-encoding loss coefficient (λ_{auto} = 0 : no denoising auto-encoding)
- λ_{adv} adversarial loss coefficient (λ_{adv} = 0 : no adversarial loss)

Ablation Study: Conclusion

- It is necessary to use either cross-lingual embeddings and/or cross-domain training
- Adding noise to input sentences is critical
- The adversarial loss and the autoencoding loss are equally important

Possible improvements

Possible improvements

- Use beam search

Possible improvements

- Use beam search
- Use BPE
  - Currently a lot of UNK words (up to 10% of UNK in German source references)
  - Supervised WMT en→fr BLEU
    35 with BPE, 28 with word-level on the used vocabulary

Possible improvements

- Use beam search
- Use BPE
  - Currently a lot of UNK words (up to 10% of UNK in German source references)
  - Supervised WMT en→fr BLEU
    35 with BPE, 28 with word-level on the used vocabulary
- Use a better validation criterion
  - WMT: between 1 and 3 BLEU lost at every iteration
  - In practice, use a small parallel validation set

Possible improvements

- Use beam search
- Use BPE
  - Currently a lot of UNK words (up to 10% of UNK in German source references)
  - Supervised WMT en→fr BLEU
    35 with BPE, 28 with word-level on the used vocabulary
- Use a better validation criterion
  - WMT: between 1 and 3 BLEU lost at every iteration
  - In practice, use a small parallel validation set
- Do semi-supervised learning

Thank you

Word translation without parallel data

Guillaume Lample *, Alexis Conneau *, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou (*ICLR 2018)*
Code: https://github.com/facebookresearch/MUSE

Unsupervised Machine Translation Using Monolingual Corpora Only

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato (*ICLR 2018)*

Supplementary slides

Orthogonality constraint

- Isometric mapping
  - Preserve dot-product
  - Preserve monolingual quality embeddings
  - Training more robust (no mapping collapse)

- After each training update, project the mapping to the orthogonal manifold:

$$W \leftarrow (1+\beta)W - \beta(WW^T)W$$

$$W \leftarrow W - \frac{\beta}{2}\nabla_W(\|W^TW - Id\|_F^2)$$

Cisse et al. (ICML 2017)

Unsupervised model selection

- Cannot select a model based on a supervised metric

Unsupervised model selection

- Cannot select a model based on a supervised metric
- How to select the best hyper-parameters?

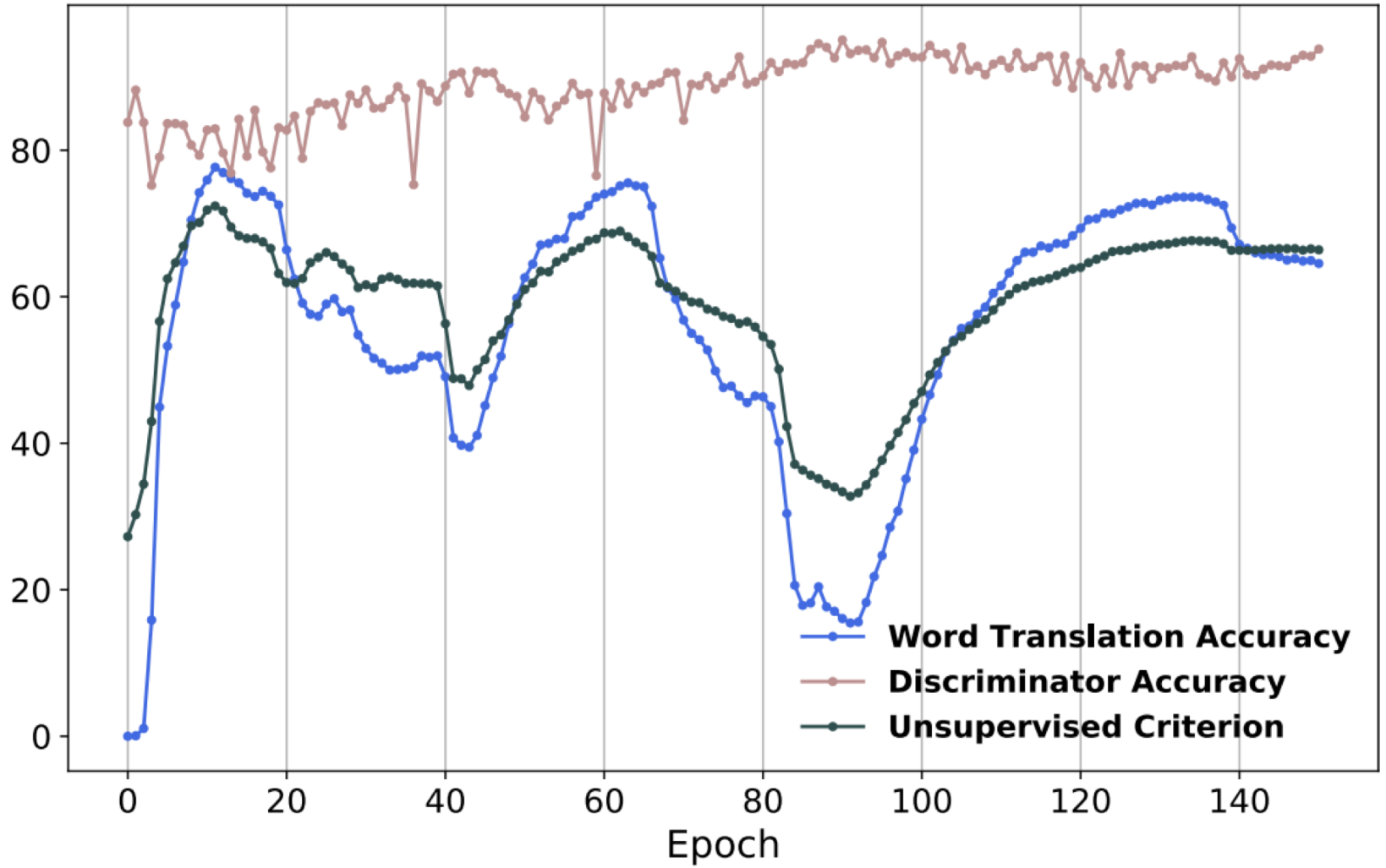
Unsupervised model selection

- Cannot select a model based on a supervised metric
- How to select the best hyper-parameters?
- Given hyper-parameters, how to select the right epoch?

Unsupervised model selection

- Cannot select a model based on a supervised metric
- How to select the best hyper-parameters?
- Given hyper-parameters, how to select the right epoch?
- Unsupervised metrics that correlates well with translation accuracy
    - Criterion: average cosine distance of 10k generated translation pairs

Unsupervised model selection



Unsupervised model selection

What do we lose compare to an oracle criterion?

	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
Unsupervised	69.8	71.3	70.4	61.9	63.1	59.6	29.1	41.5	18.5	22.3
Oracle	**72.0**	**71.7**	**71.1**	**67.5**	**64.8**	**61.3**	**34.4**	**45.1**	**20.9**	**26.3**

Unsupervised validation criterion vs Oracle criterion / Test metric

Hubness in word embedding spaces

- Hubness problem
  - Some words are the nearest neighbors of many words (hubs)

Hubness in word embedding spaces

- Hubness problem
  - Some words are the nearest neighbors of many words (hubs)
  - Some words are not the nearest neighbors of any words (anti-hubs)

Hubness in word embedding spaces

- Hubness problem
  - Some words are the nearest neighbors of many words (hubs)
  - Some words are not the nearest neighbors of any words (anti-hubs)

- Word translation retrieval
  - Many words are translated to the same word

Hubness in word embedding spaces

- Hubness problem
  - Some words are the nearest neighbors of many words (hubs)
  - Some words are not the nearest neighbors of any words (anti-hubs)

- Word translation retrieval
  - Many words are translated to the same word

Source word	un	deux	trois	quatre	cinq	six	sept	huit	neuf
Translation – NN	one	two	two	four	two	two	two	four	two

Mitigating hubness – CSLS

- Cross-Domain Similarity Local Scaling

$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t)$$

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t)$$

  - Encourage nearest neighbors reciprocity (empirical observation)
  - Increases the similarity associated with isolated word vectors

Mitigating hubness – CSLS

- Cross-Domain Similarity Local Scaling

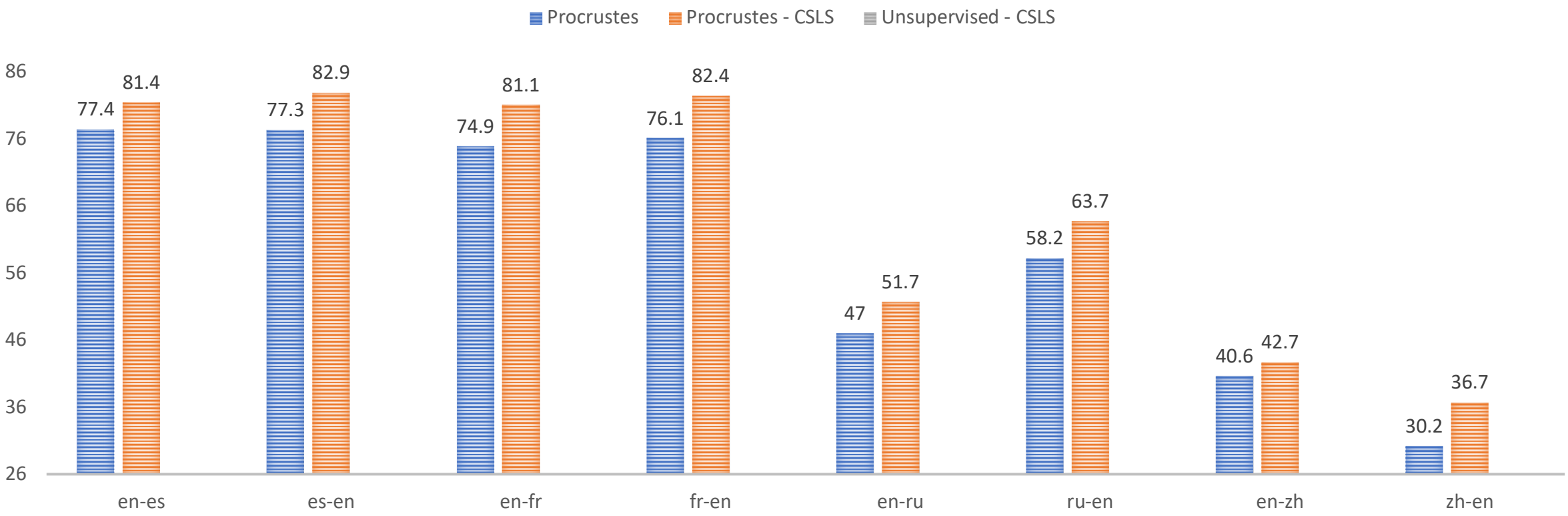
$$\mathrm{CSLS}(Wx_s, y_t) = 2\cos(Wx_s, y_t) - r_{\mathrm{T}}(Wx_s) - r_{\mathrm{S}}(y_t)$$

$$r_{\mathrm{T}}(Wx_s) = \frac{1}{K} \sum_{y_t \in \mathcal{N}_{\mathrm{T}}(Wx_s)} \cos(Wx_s, y_t)$$

  - Encourage nearest neighbors reciprocity (empirical observation)
  - Increases the similarity associated with isolated word vectors

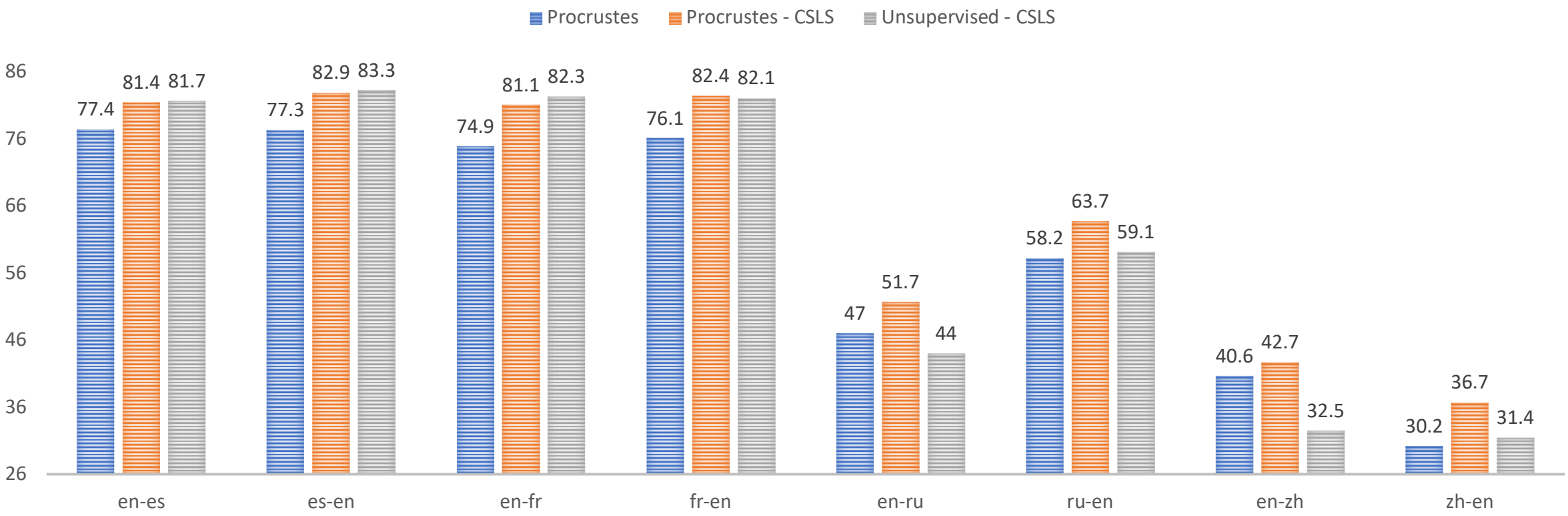
Source word	un	deux	trois	quatre	cinq	six	sept	huit	neuf
Translation – NN	one	two	two	four	two	two	two	four	two
Translation – CSLS	one	two	three	four	five	six	seven	eight	nine

Results on word translation



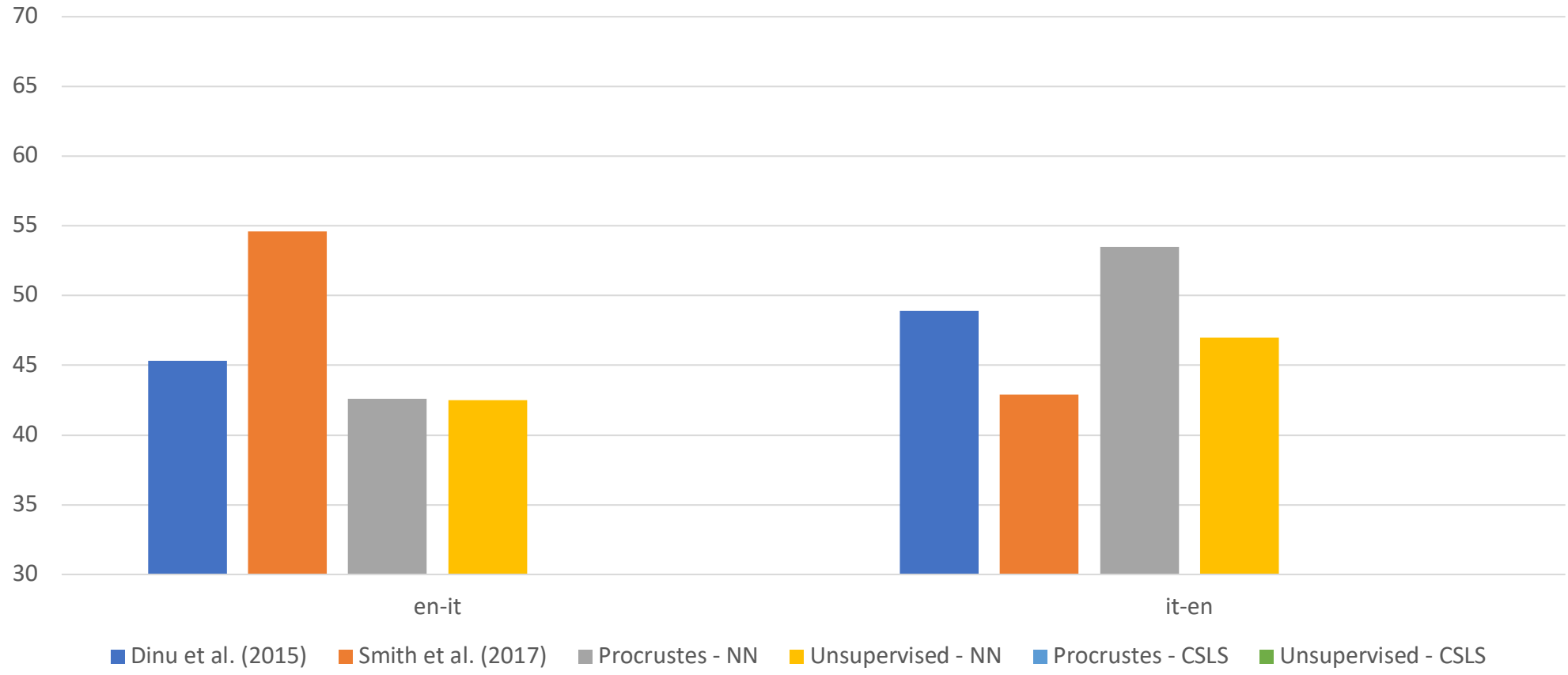
Word retrieval accuracy (P@1)

Results on word translation



Word retrieval accuracy (P@1)

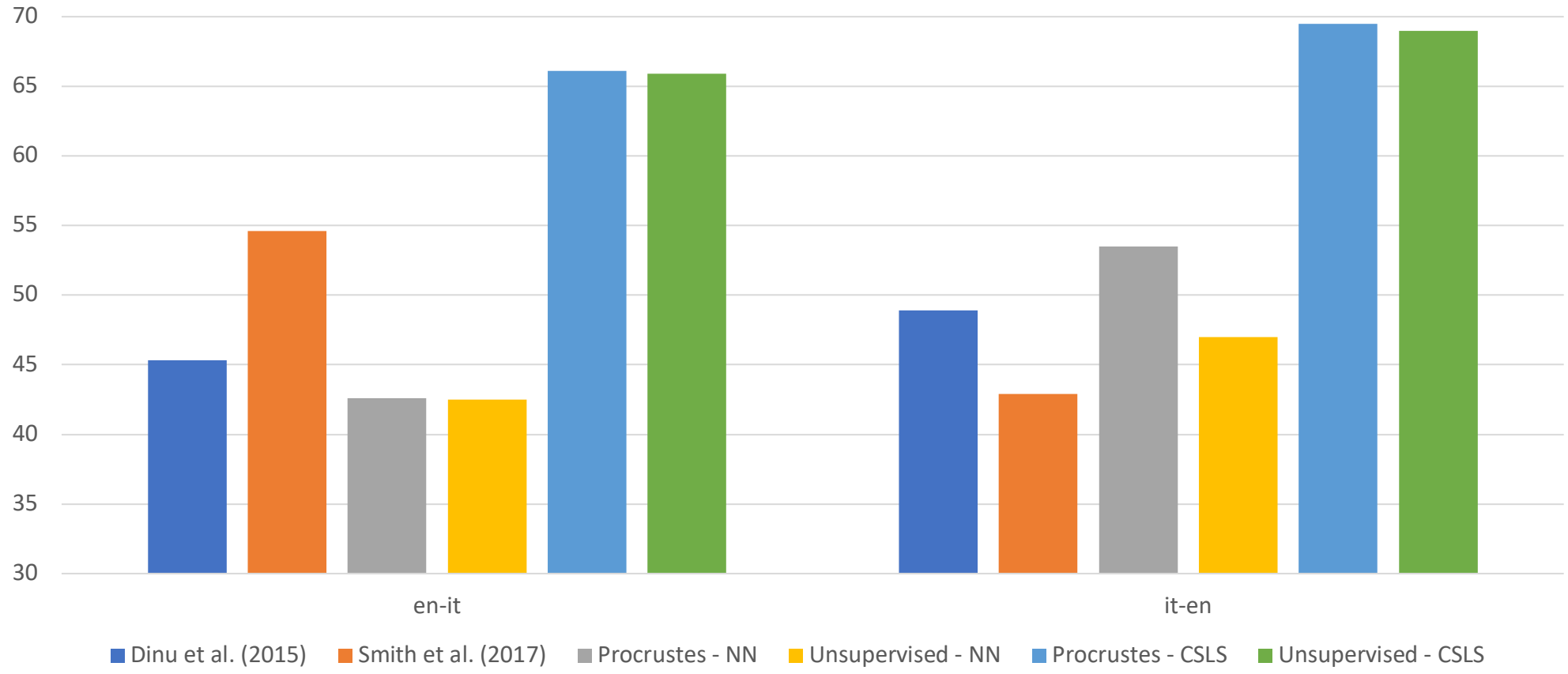
Results on sentence retrieval



Sentence retrieval accuracy (P@1)

Task: given a source sentence, find the correct translation among 200k target sentences.

Results on sentence retrieval



Sentence retrieval accuracy (P@1)

Task: given a source sentence, find the correct translation among 200k target sentences.