

The background of the slide features a light blue-grey color with a subtle, abstract pattern of interconnected dots and lines, resembling a network or a molecular structure. The dots are small and blue, while the lines are thin and light blue. The pattern is more dense in the corners and fades towards the center.

Unsupervised Learning with Text

Ludovic Denoyer

Context

- Deep Learning (on text) has shown impressive results in the very last years
- But it usually needs very large datasets with rich supervision
 - Difficult for rare languages
 - Difficult for specific application domains
 - Difficult for particular tasks
- Open Question: What can we solve with few/no supervision ?

Outline

- Unsupervised Machine Translation
- Unsupervised Text rewriting
- Unsupervised Question Answering
- Building Interpretable text classifiers

Unsupervised Machine Translation

- Machine Translation works well for high-resource language pairs
- Performance drops drastically when parallel data is scarce
 - Vietnamese, Ukrainian, Tamil, Urdu, etc.
- The creation of parallel data is difficult, and costly
- However, monolingual data is much easier to find
- Can we translate languages without any parallel data?

yes, we can...

- Unsupervised Word Translation

- Word translation without parallel data (ICLR 2018)**

- Guillaume Lample*, Alexis Conneau*, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou

- <https://github.com/facebookresearch/MUSE>

- Unsupervised Neural or Phrase-Based Translation

- Unsupervised Machine Translation Using Monolingual Corpora Only (ICLR 2018)**

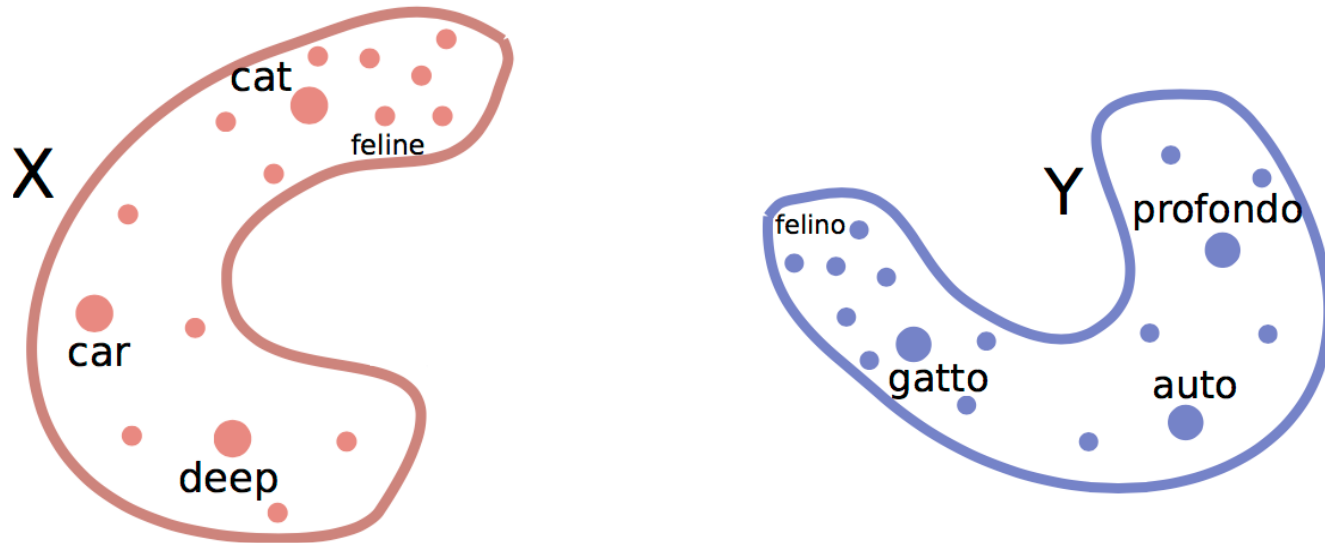
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato

- Phrase-based & Neural Unsupervised Machine Translation (EMNLP 2018 – Best paper award)**

- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato

Weakly-supervised word translation

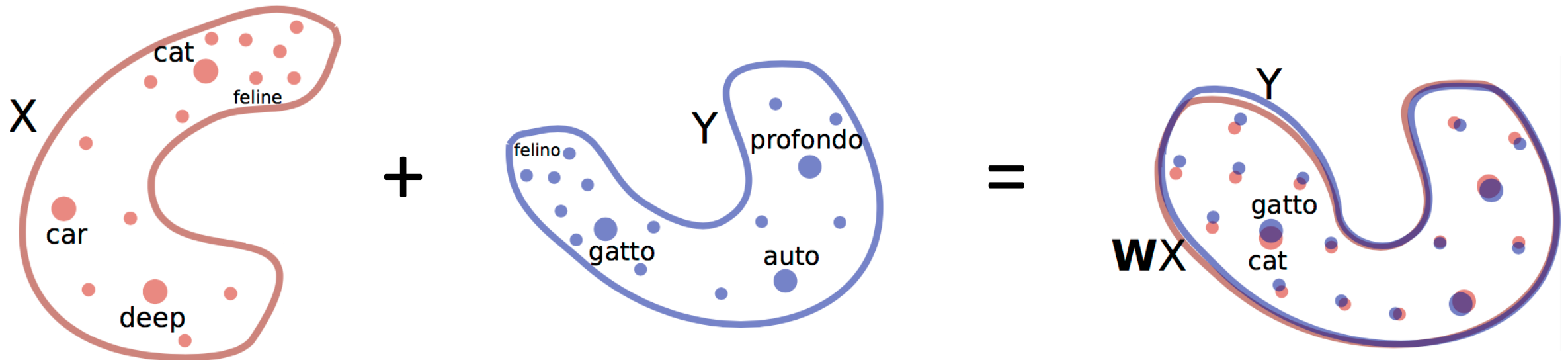
- Exploiting similarities among languages for machine translation (Mikolov et al., 2013)
 - Start from two pre-trained monolingual spaces (word2vec)



- 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



- 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)

$$W^* = \operatorname{argmin}_{W \in M_d(\mathbb{R})} \|WX - Y\|_F$$

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

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

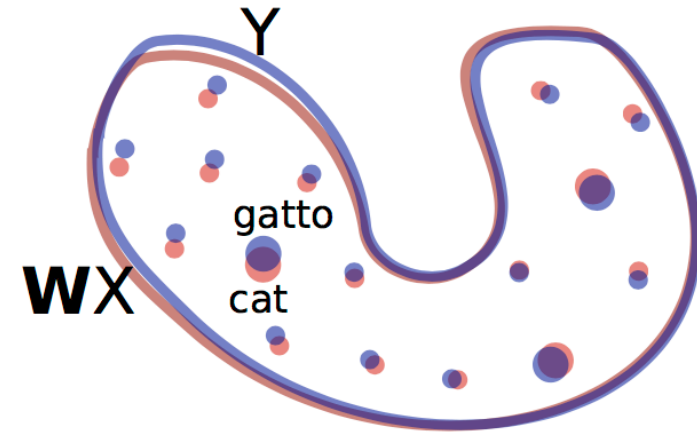
- Given a source word s , define the translation as:

$$t = \operatorname{argmax}_t \cos(Wx_s, y_t) \quad \begin{array}{l} \text{(nearest neighbor according} \\ \text{to the cosine distance)} \end{array}$$

Can we find the mapping \mathbf{W} in an unsupervised way?

Adversarial training

- If $\mathbf{W}X$ and Y are perfectly aligned, these spaces should be undistinguishable
- Train a discriminator D to discriminate elements from $\mathbf{W}X$ and Y
- Train \mathbf{W} to prevent the discriminator from making accurate predictions



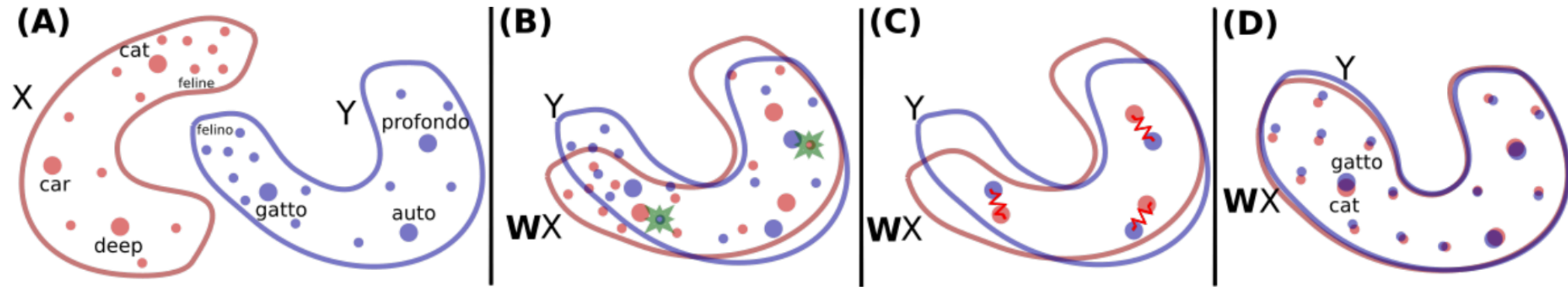
Discriminator training

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

Mapping training

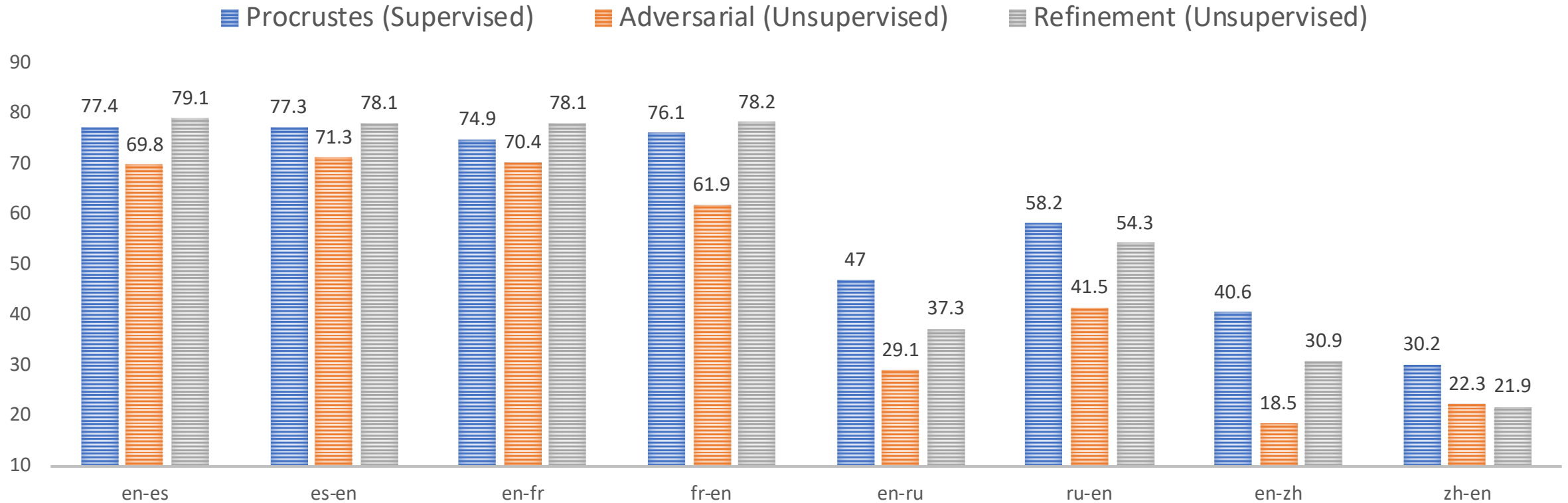
$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n} \sum_{i=1}^n \log P_{\theta_D}(\text{source} = 0|Wx_i) - \frac{1}{m} \sum_{i=1}^m \log P_{\theta_D}(\text{source} = 1|y_i)$$

Unsupervised word translation – Summary



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

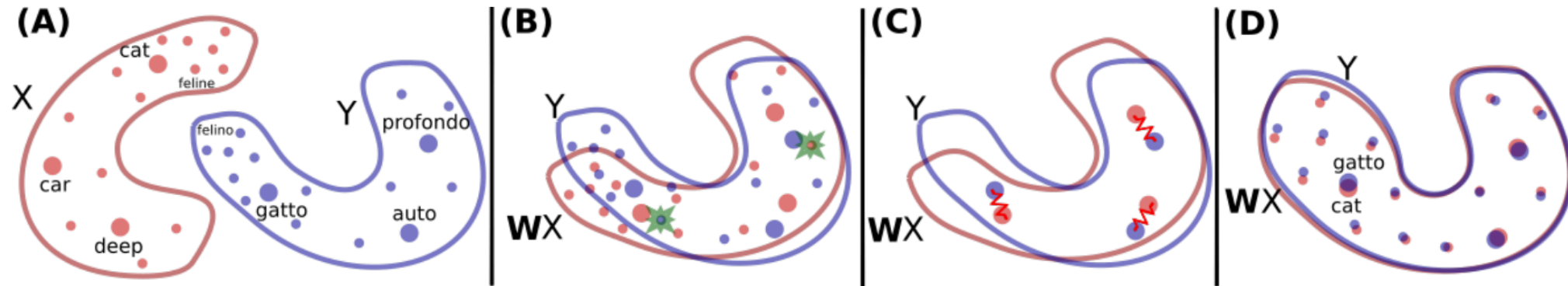
Results on word translation – Refinement



Word translation retrieval – P@1 – Adversarial + Refinement

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

Unsupervised sentence translation



- Can we apply the same 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)

Phrase-Based Statistical Machine Translation (PBSMT)

- Two main components
 - **Phrase-table (needs parallel data)**
 - Language model

Input : [il regarde] [dans] [la glace]

Reference : he looks in the mirror

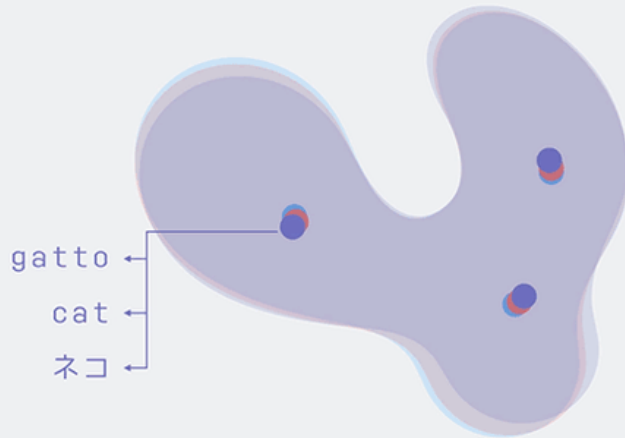
Hypotheses	Phrase scores	LM scores
[he is watching] [in] [the ice cream]	0.6 0.7 0.5	-7
[he looks at] [within] [ice]	0.1 0.1 0.3	-12
[he looks] [in] [the mirror]	0.3 0.7 0.2	-3

↙ maximizes combined phrase-table and LM scores

Source phrase	Target phrase	Score
il regarde	he is watching	0.6
	he looks	0.3
	he looks at	0.1
dans	in	0.7
	into	0.2
	within	0.1
la glace	the ice cream	0.5
	ice	0.3
	the mirror	0.2

Moses (Koehn et al., 2007)

Unsupervised ~~word~~ **phrase** translation

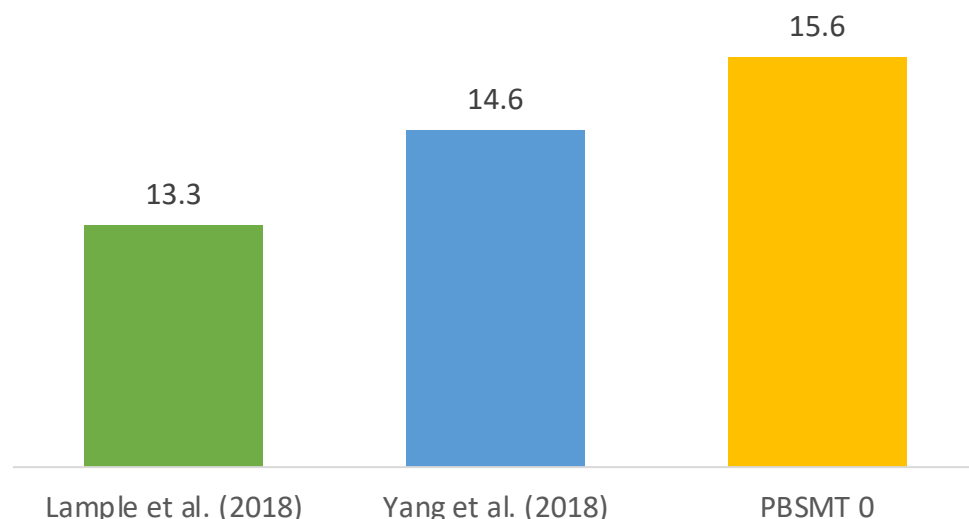


- Given independent monolingual corpora, we can generate:
 - Cross-lingual **phrase tables** dictionaries
 - Cross-lingual ~~word~~ **phrase** embeddings
- The alignment procedure is the same for words and phrases

Results with unsupervised phrase-table

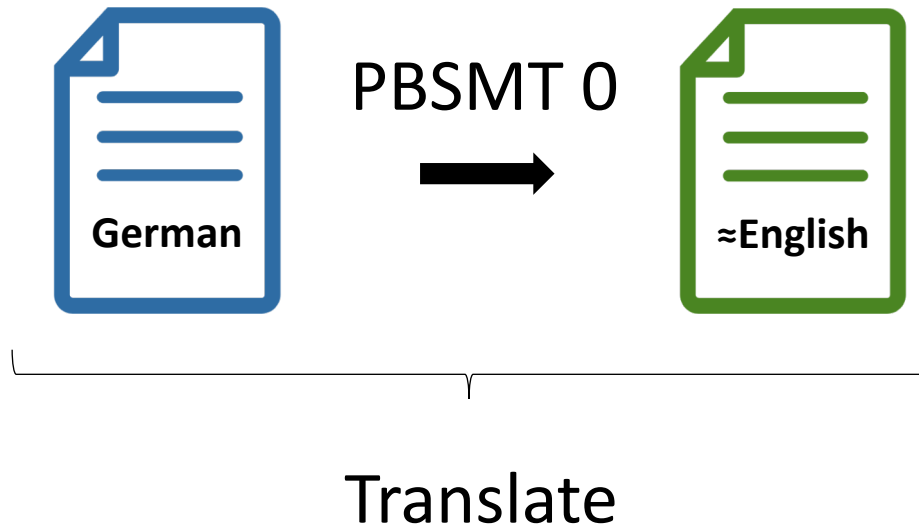
Input **[Raumsonde]** Cassini findet Ozean auf **[Saturnmond]** Enceladus
Reference Cassini space probe finds ocean on Saturn 's moon Enceladus
PBSMT 0 **[NASA spacecraft]** Cassini finds a **[Saturn moon]** Enceladus ocean

BLEU on German-English - newstest 2016



Source phrase	Target phrase	Score
Raumsonde	<u>NASA spacecraft</u>	<u>0.32</u>
	spacecraft	0.26
	moon probe	0.12
	Voyager 2	0.12
	Kepler telescope	0.09
Saturnmond	<u>Saturn moon</u>	<u>0.45</u>
	Enceladus	0.21
	liquid water	0.11
	Titan	0.10
	Earth-like planet	0.05

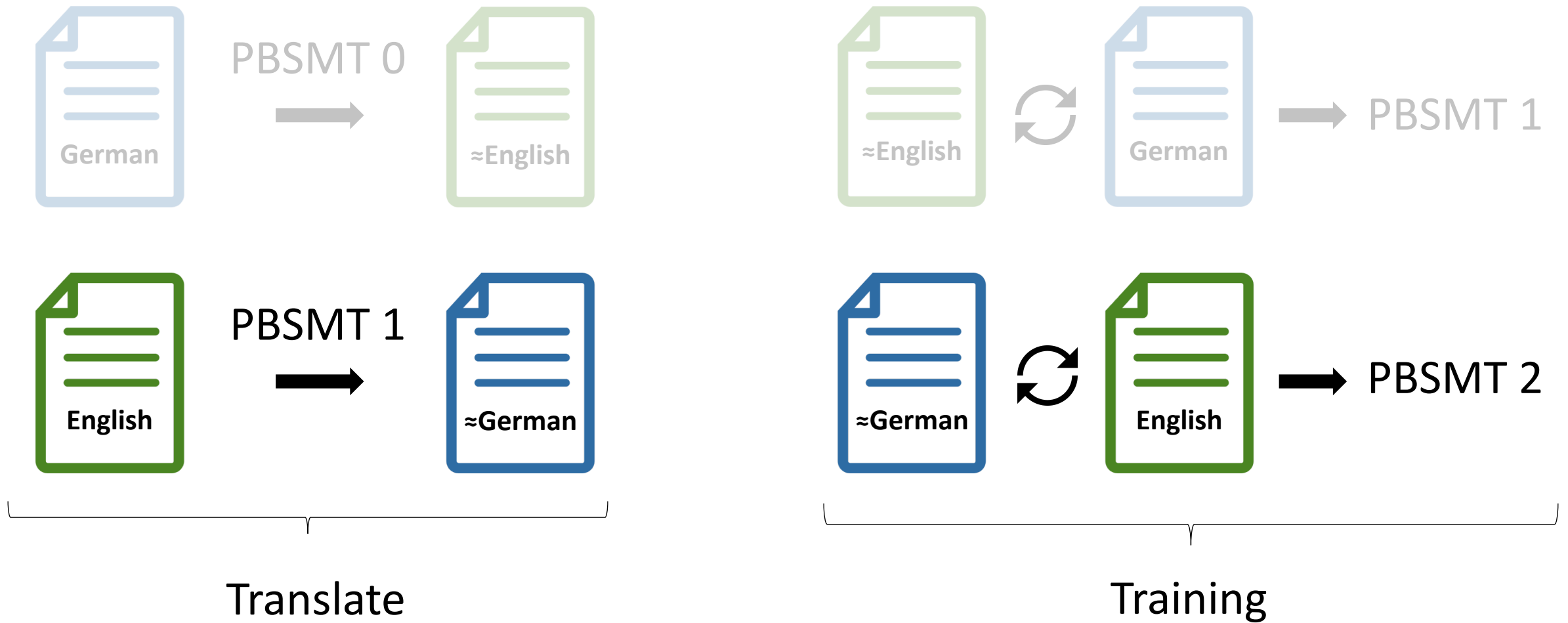
Back-translation



Improving neural machine translation models with monolingual data (Sennrich et al., 2016)

Slide courtesy of Guillaume Lample

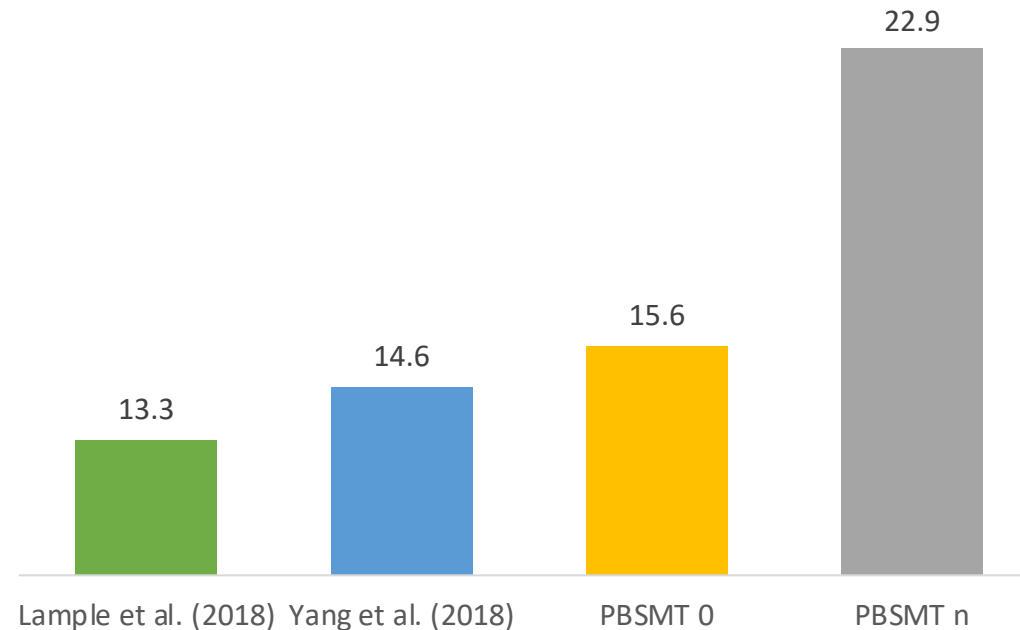
Iterative Back-translation



Results with iterative back-translation

Input	Raumsonde Cassini findet Ozean auf Saturnmond Enceladus
Reference	Cassini space probe finds ocean on Saturn 's moon Enceladus
PBSMT 0	NASA spacecraft Cassini finds a Saturn moon Enceladus ocean
PBSMT n	NASA spacecraft Cassini finds ocean on Saturn moon Enceladus

BLEU on German-English - newstest 2016

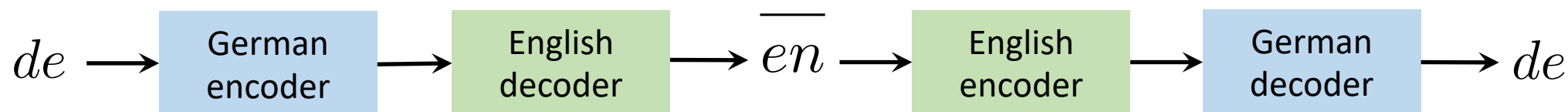


Three principles of unsupervised MT

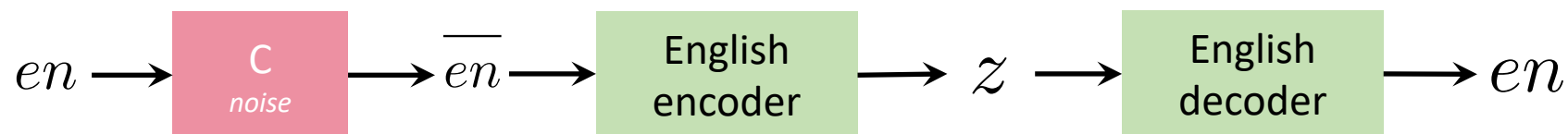
- Initialization
 - Language modeling
 - Iterative back-translation
-
- Can we apply these principles to neural architectures?

Three principles for unsupervised MT - NMT

- **Online back-translation**



- **Language modeling**

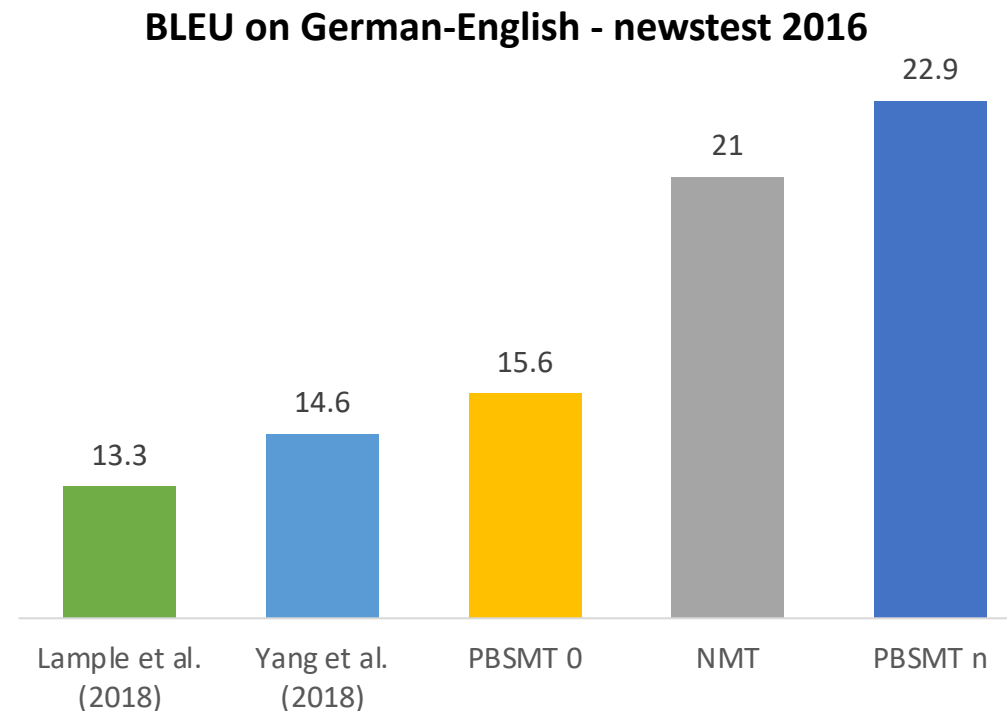


- **Initialization**

Initialized shared encoders, with shared cross-lingual BPE embeddings

Results with unsupervised NMT

Input	Raumsonde Cassini findet Ozean auf Saturnmond Enceladus
Reference	Cassini space probe finds ocean on Saturn 's moon Enceladus
PBSMT 0	NASA spacecraft Cassini finds a Saturn moon Enceladus ocean
PBSMT n	NASA spacecraft Cassini finds ocean on Saturn moon Enceladus
NMT	Cassini spacecraft takes ocean on Saturnmond Enceladus

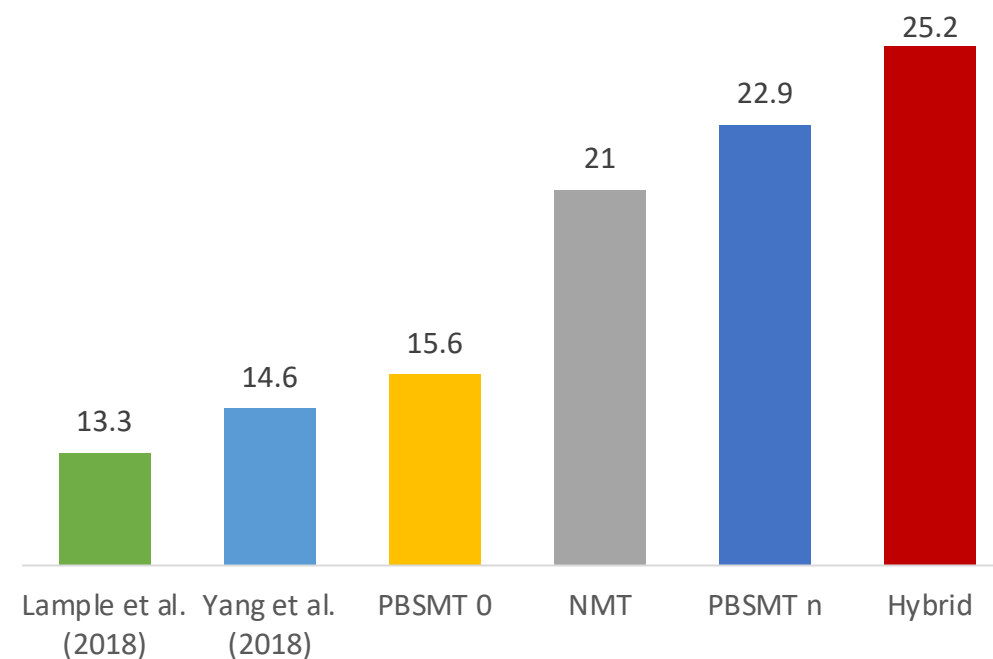


Hybrid model

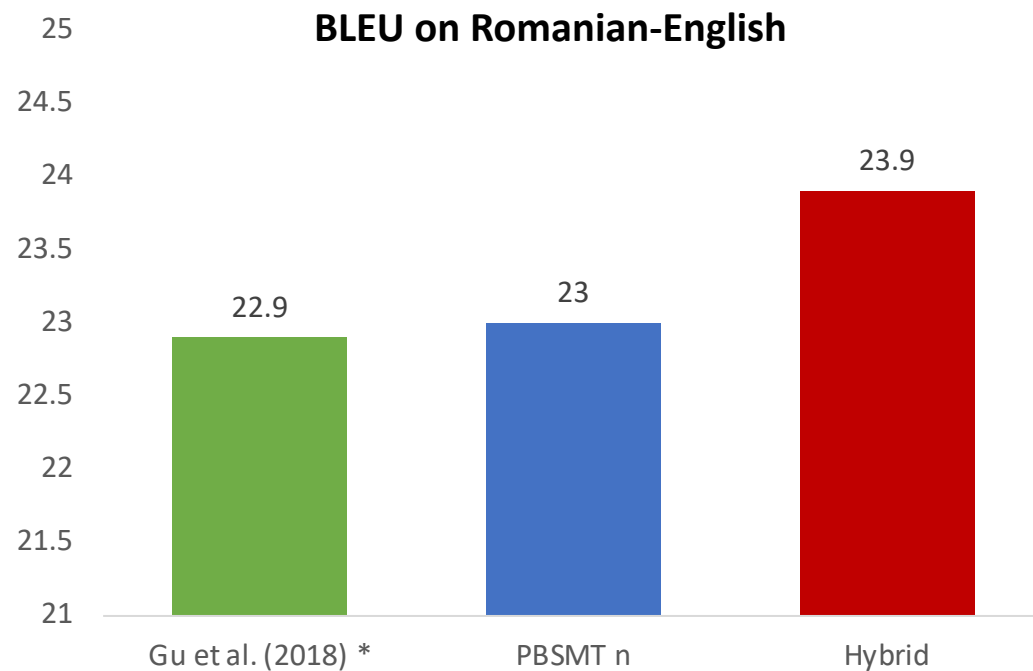
- Train unsupervised NMT and PBSMT models independently
- Fine-tune NMT model on both NMT and PBSMT back-translations

SRC	Behörden sind dabei , den Zaun schnell zu reparieren .
REF	Authorities are quickly repairing the fence .
PBSMT n	Authorities , who are on the fence to quickly repair .
NMT	Local authorities are unsure how to repair the fence .
Hybrid	Authorities are also getting the fence repaired quickly .

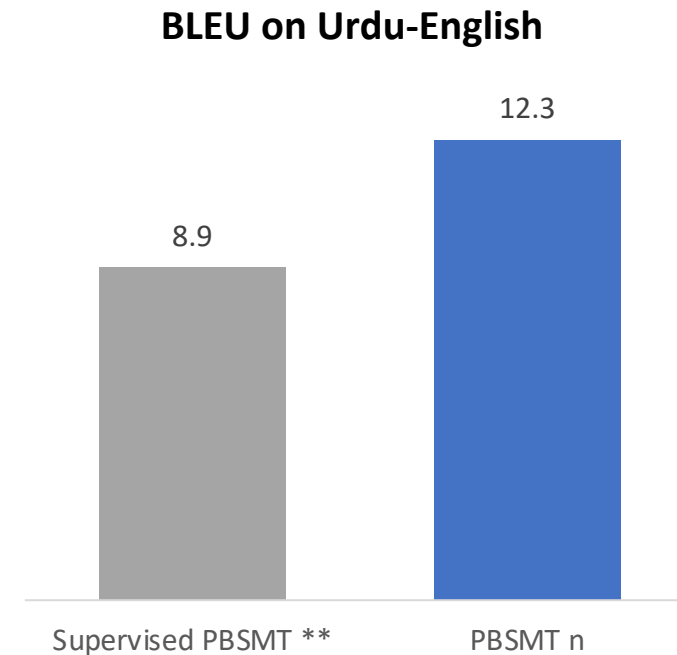
BLEU on German-English - newstest 2016



Results on low-resource language pairs



* Trained on 6,000 parallel sentences, seed dictionary, multi-NMT system with 6 languages



** Trained on 800,000 parallel sentences (Opus corpora)









Multiple-Attribute Text Rewriting

Guillaume Lample*†‡, Sandeep Subramanian*†§, Eric Smith†, Ludovic Denoyer†‡, Marc'Aurelio Ranzato†, Y-Lan Boureau†

† Facebook AI Research, § Mila, Université de Montréal, ‡ Université Pierre-et-Marie-Curie, *equal contribution



Relaxed ↔ Annoyed

Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night   
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend   
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 

1 . Text Rewriting / Style Transfer

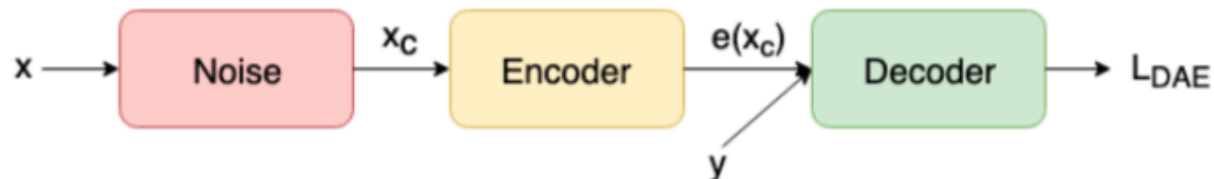
We present a simple technique for text rewriting, that

- Has the ability to control multiple attributes simultaneously
- Builds on top of techniques used in unsupervised machine translation
- Does not require any parallel data
- Does not rely on disentangling the content from the style
- Performs better than existing methods on new and challenging benchmarks

2. Model

Denoising Autoencoder

Given a sentence x with attributes y from the training set, we train the model to reconstruct x conditioned on a corrupted version of the sentence, x_c , and from the attributes y .



$$\mathcal{L}_{DAE} = \sum_{(x,y) \sim \mathcal{D}} -\log p_d(x|e(x_c), y)$$

Backtranslation

Given a sentence x and a randomly sampled configuration of attributes y^* , generate $x^* \sim p_d(x|e(x), y^*)$ using the model. Given this generation and the original y , train the model to reconstruct x .



$$\mathcal{L}_{BT} = \sum_{(x,y) \sim \mathcal{D}, y^* \sim \mathcal{Y}} -\log p_d(x|e(d(e(x), y^*)), y)$$

Male ↔ Female

Male	Gotta say that beard makes you look like a Viking...
------	--

Female	Gotta say that hair makes you look like a Mermaid...
--------	--

Female	Awww he's so gorgeous 🥰 can't wait for a cuddle. Well done 🥰 xxx
--------	--

Male	Bro he's so f***ing dope can't wait for a cuddle. Well done bro
------	---

Age 18-24 ↔ 65+

18-24	You cheated on me but now I know nothing about loyalty 😂 ok
-------	---

65+	You cheated on America but now I know nothing about patriotism. So ok.
-----	--

65+	Ah! Sweet photo of the sisters. So happy to see them together today .
-----	---

18-24	Ah 😂 Thankyou 🧡 #sisters 🧡 happy to see them together today
-------	---

Clothing ↔ Electronics (Amazon)

Clothing	got this cause it said it would help with tennis elbow and guess what my tennis elbow still bothering me
----------	--

Electronics	got this cause it said it would help with windows xp and guess what my windows xp still crashed
-------------	---

Electronics	i have no choice. this is the only black ink that works with my printer.
-------------	--

Clothing	i have no choice. this is the only black color that works with my dress.
----------	--

5 . Comparison to previous work

We achieve state-of-the-art results on existing sentiment style transfer benchmarks, measured by automatic metrics and human evaluations. We compare our model against the DeleteAndRetrieve (DAR) approach of [1].

	Fluency	Content	Sentiment
DAR [1]	3.33 (1.39)	3.16 (1.43)	64.05%
Ours	4.07 (1.12)	3.67 (1.41)	69.66%
Human [1]	4.56 (0.78)	4.01 (1.25)	81.35%
	Our Model	No Preference	DAR
DAR vs Our Fader	37.6%	32.7%	29.7%
DAR vs Ours	54.4%	24.7%	20.8%

7 . Multiple-attribute swap

- Qualitative examples of our model's ability to control multiple attributes simultaneously, on Amazon and Yelp reviews. Input sentences are in bold.

Sentiment	Category	Input / Generations
Positive	Movies	exciting new show. john malkovich is superb as always. great supporting cast. hope it survives beyond season 1
Positive	Books	exciting new book. john grisham is one of the best. great read. hope he continues to write more.
Negative	Clothing	horrible. the color is not as pictured. not what i expected. it is not a good quality.
Negative	Dessert	the bread here is crummy, half baked and stale even when "fresh." i won't be back.
Positive	Mexican	the tacos here are delicious, full of flavor and even better hot sauce. i highly recommend this place.
Positive	Dessert	the ice cream here is delicious, soft and fluffy with all the toppings you want. i highly recommend it.

Unsupervised Question Answering by Cloze Translation

Patrick Lewis, Ludovic Denoyer, Sebastian Riedel



Extractive Question Answering (EQA)

Question q: What Denver player caused two fumbles for the Panthers?

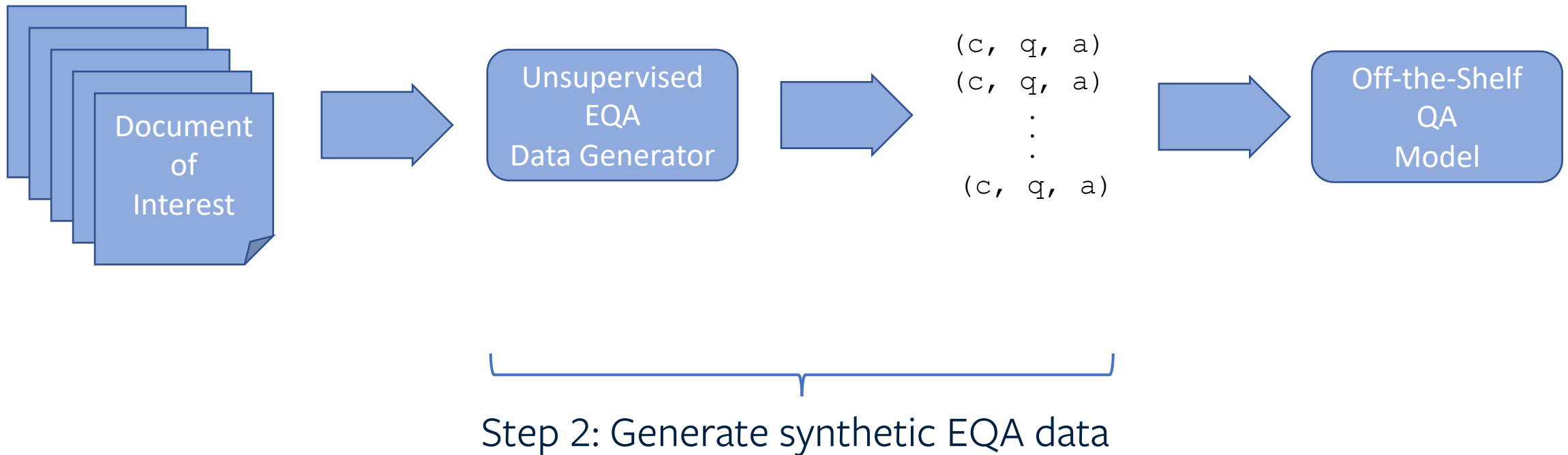
Context c: The Broncos took an early lead in Super Bowl 50 and never trailed. [...] Denver linebacker Von Miller was named Super Bowl MVP, recording five solo tackles, 2½ sacks and two forced fumbles.

Answer a: Von Miller

Extractive Question Answering (EQA) (2)

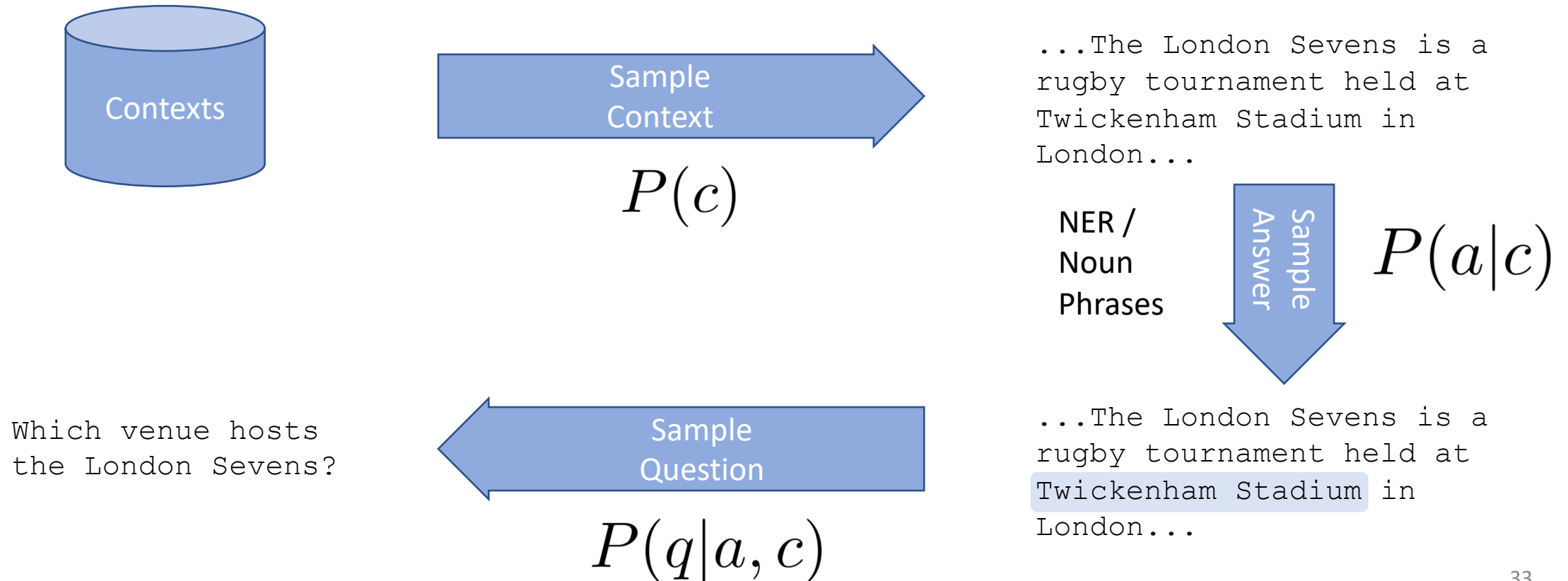
Step 1: Train Synthetic EQA data generator

Step 3: Train off-the-shelf EQA Model



Generating EQA data without Supervision

$$P(c, q, a) =$$



Experiments

- Evaluate EQA performance without explicit supervision
 - Explore impact of design decisions of data generator
-
- Context Generator: Paragraphs from English Wikipedia
 - Cloze Question boundary: Sentence or sub-clause with “S” label
 - 5M questions mined from common crawl, 5M clozes mined from Wikipedia
 - Question Answering: BiDAF + Self-attention [1] and fine-tuning BERT [2]

[1] J Devlin et al. 2019

[2] C Clark and M Gardner 2018

UNMT Translation Examples

Cloze Question

Answer

Translated Question

WALA would be sold to the Des Moines-based **ORG** for \$86 million

Meredith Corp

Who would buy the WALA Des Moines-based for \$86 million?

The **NUMERIC** on Orchard Street remained open until 2009

second

How much longer did Orchard Street remain open until 2009?

he speaks **LANGUAGE**, English, and German

Spanish

What are we , English , and German?

Form a larger Mid-Ulster District Council in **TEMPORAL**

August

When is a larger Mid-Ulster District Council?

Form a larger Mid-Ulster District Council in **TEMPORAL**

August

When will a larger Mid-Ulster District Council be formed?

Results in context

- Best results with:
 - Named entity answers
 - Unsupervised NMT questions
 - Wh* heuristic
 - Sub-clause boundaries
 - Bert-Large QA

Unsupervised Models	EM	F1
BERT-Large Unsup. QA (ens.)	47.3	56.4
BERT-Large Unsup. QA (single)	44.2	54.7
BiDAF+SA (Dhingra et al., 2018)	3.2 [†]	6.8 [†]
BiDAF+SA (Dhingra et al., 2018) [‡]	10.0*	15.0*
BERT-Large (Dhingra et al., 2018) [‡]	28.4*	35.8*
Baselines	EM	F1
Sliding window (Rajpurkar et al., 2016)	13.0	20.0
Context-only (Kaushik and Lipton, 2018)	10.9	14.8
Random (Rajpurkar et al., 2016)	1.3	4.3
Fully Supervised Models	EM	F1
BERT-Large (Devlin et al., 2018)	84.1	90.9
BiDAF+SA (Clark and Gardner, 2017)	72.1	81.1
Log. Reg. + FE (Rajpurkar et al., 2016)	40.4	51.0

Table 1: Our best performing unsupervised QA models compared to various baselines and supervised models. * indicates results on SQuAD dev set. † indicates results on non-standard test set created by Dhingra et al. (2018). ‡ indicates our re-implementation

Conclusion

- Outperform simple supervised models without explicit supervision
- Much scope for future work:
 - Questions without Answers
 - “Multi-hop” Questions
 - Other Question Answering tasks
- 4M UQA training datapoints, and code
github.com/facebookresearch/UnsupervisedQA

EDUCE: EXPLAINING MODEL DECISIONS THROUGH UNSUPERVISED CONCEPTS EXTRACTION

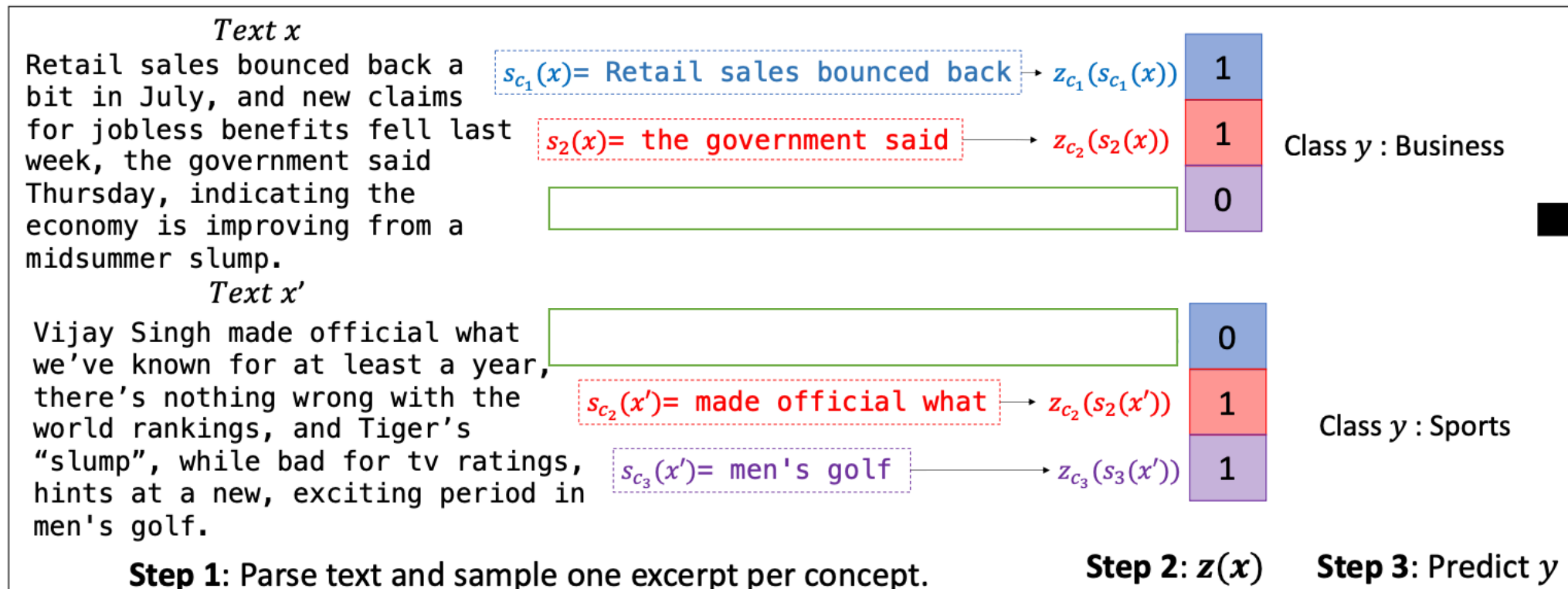
Diane Bouchacourt
Facebook AI Research
dianeb@fb.com

Ludovic Denoyer
Facebook AI Research
denoyer@fb.com

Context

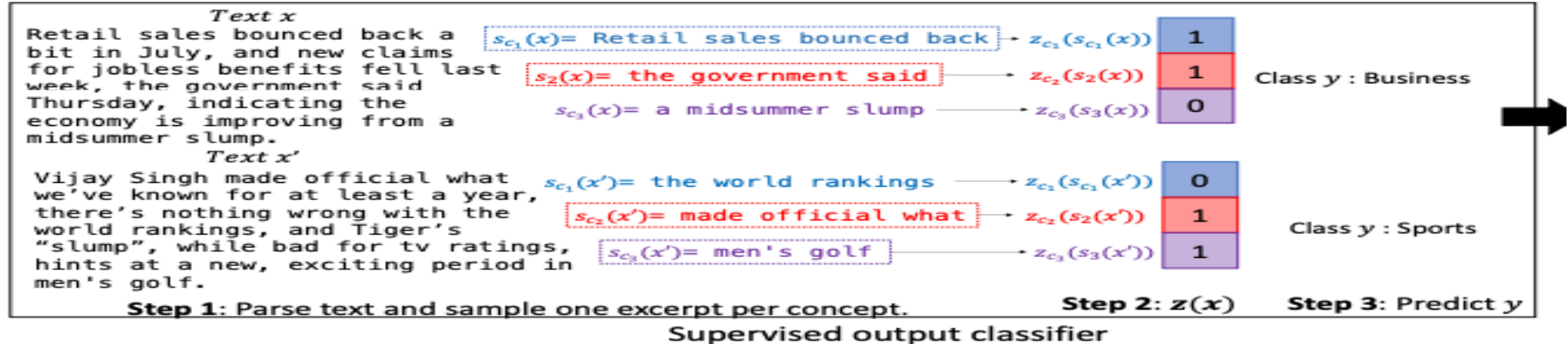
- Deep Learning Classifier are usually not interpretable
- For Text categorization, some recent efforts focus on attention models able to extract 'rationales' from text
- It usually requires human-annotated datasets
- Objective: Learn interpretable classifier without additional supervision

EDUCE: Classification through concept detection



Principles

- Step 1: For each of the C concepts, extract the excerpt that is the most relevant
- Step 2: Classify each excerpt as 'present' or not in the document
- Step 3: Classify the document based on the presence of concepts



$$\begin{aligned}\mathcal{L}^{output}(x, y, \delta, \alpha, \gamma) &= \mathbb{E}_{s(x) \sim p_\gamma} [\mathbb{E}_{z(x) \sim p_\alpha} [\mathcal{L}^{output}(y, z(x), \delta)]] \\ &= \mathbb{E}_{\forall c \ s_c(x) \sim p_\gamma(s|x, c)} [\mathbb{E}_{\forall c \ z_c \sim p_\alpha(z_c|s_c(x), c)} [-\log p_\delta(y|z(x))]].\end{aligned}$$

Ensuring concept consistency

- Without additional concept, there is no reason that EDUCE extracts meaningful concepts
- We consider that concepts will be ‘easy to understand’ if extracted excerpts are homogeneous

The concept classifier is trained by minimizing cross-entropy, where the label of each excerpt $s_c(x)$ the index of the concept for which it was extracted (i.e. c):

$$\begin{aligned}\mathcal{L}^{concept}(x, \theta, \alpha, \gamma) &= \mathbb{E}_{\mathbf{s}(x) \sim p_\gamma} [\mathbb{E}_{\mathbf{z}(x) \sim p_\alpha} [\mathcal{L}^{concept}(\mathbf{z}(x), \mathbf{s}(x), \theta)]] \\ &= \mathbb{E}_{\mathbf{s}(x) \sim p_\gamma} [\mathbb{E}_{\mathbf{z}(x) \sim p_\alpha} [\sum_c -z_c(s_c(x)) \log p_\theta(c|s_c(x))]],\end{aligned}\tag{4}$$

Results

Data	Model	Output Acc. (%)	A Posteriori Concept Acc. (%)
DBPedia	EDUCE	97.0 ± 0.1	82.4 ± 0.8
	No Concept Loss	97.4 ± 0.1	25.9 ± 0.6
	No Concept Loss+ L_1	96.5 ± 0.2	44 ± 2.6
	Baseline	98.75 ± 0.0	n/a
AGNews	EDUCE	87.5 ± 0.2	78 ± 6.5
	No Concept Loss	88.2 ± 0.1	31.0 ± 0.7
	No Concept Loss+ L_1	86.3 ± 0.7	56 ± 3.2
	Baseline	92.08 ± 0.1	n/a

Table 1: Test performance on DBPedia and AGNews (mean \pm SEM).

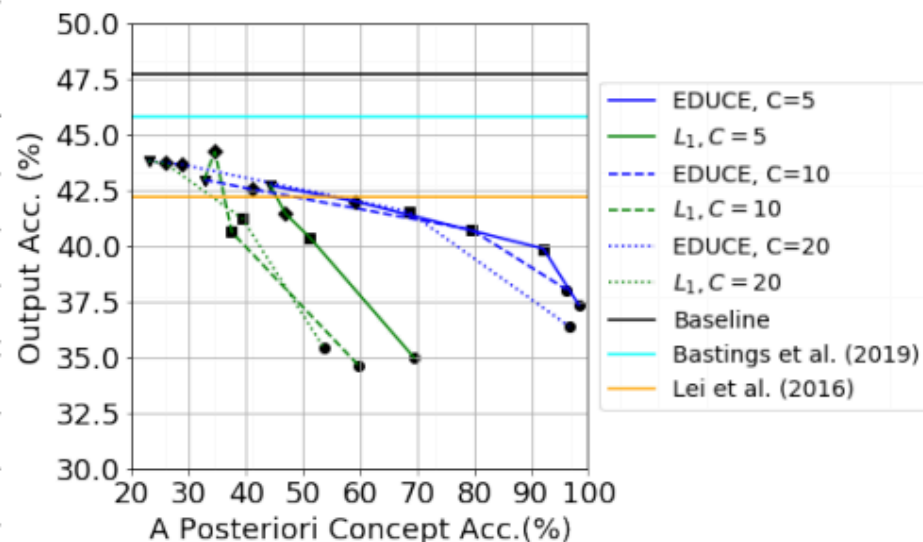


Figure 2: SST test performance, output accuracy vs a posteriori concept accuracy. A posteriori concept accuracy is not applicable for Baseline, Lei et al. (2016) and Bastings et al. (2019).

Class Business: sports retailer jjb yesterday reported a near 25 drop in profits and continuing poor sales , and ended shareholders #39 hopes of a takeover by announcing that a potential bidder had walked away .

Class World: bangkok , thailand sept . 30 , 2004 - millions of volunteers led by emergency teams fanned out across thailand on thursday in a new drive to fight bird flu after the prime minister gave officials 30 days to eradicate the epidemic .

Class Sports: the spanish government responded to diplomatic pressure from britain yesterday by starting a search for fans who racially abused england players during a quot friendly quot football match with spain .

Class Sci/Tech: los angeles (reuters) - a group of technology companies including texas instruments inc . < txn . n> , stmicroelectronics < stm . pa> and broadcom corp . < brcm . o> , on thursday said they will propose a new wireless networking standard up to 10 times the speed of the current generation .

(a) AGNews test examples correctly classified by EDUCE. Underlined set of words are excerpts extracted, one color per concept.

Concept 0	software services giant / moonwalk to home / video display chip / downloading music .
Concept 1	upcoming my prerogative video / his ever-growing swimming / launch of a video display chip / illegality of downloading
Concept 2	oil market . / oil giant sibneft / oil prices and / corp . < brcm
Concept 4	cash settlement of up to #36 50 million / six-year deal worth about \$40 million / the dollar dipped to a four-week low against the euro / five shares ,
Concept 6	olympic 100-meter freestyle / sox ' family / olympics should help / athletes were already
Concept 8	frail pope john paul / indian army major shot / unions representing workers / goverment representatives .

(b) Examples of excerpts that are extracted **accross the test set**, corresponding to the concepts detected in Figure 3a. Colors match the colors used in Figure 3a.

Concept 0	fruity esters , and/ fruits , caramelized pecans , and/ toffee and caramel accents ,/ coffee and chocolate flavors/ earthy hop resin .
Concept 2	creamy and a/ chewy and rich and drinkability/ creamy mouthfeel that/ smooth and just velvety on the/ thick , and
Concept 3	rich malt scents/ aroma is quite hoppy with big citrus/ tons of different sweet malts , toffee/ smells extremely roasty/ boom of grapefruity
Concept 7	good head and lacing/ beautiful golden-amber color/ creamy tan head/ nice deep brown color/ proud head has settled . nothing
Concept 9	carbonation is graceful/ drinkability is excellent/ mouthfeel is wonderful/ mouthfeel is exemplary/ drinkability : excellent

Table 2: Examples of excerpts extracted on the Beer test set.

Concept	Appearance	Smell	Palate	Taste
Concept 0	0.85	55.75	1.79	40.60
Concept 1	0.54	5.48	11.70	30.58
Concept 2	1.22	0.90	78.18	15.03
Concept 3	1.46	92.39	0.30	5.34
Concept 4	45.33	0.23	0.26	0.16
Concept 5	7.27	22.50	2.27	28.86
Concept 6	0.36	5.86	6.40	28.73
Concept 7	97.09	0.98	0.45	0.53
Concept 8	0.18	0.09	18.07	5.93
Concept 9	0.00	0.00	68.42	10.53

Table 3: Per-concept precision of gold rationales (in % for each aspect), trained to predict all 4 aspects and the overall score. In bold we emphasize the concepts which precision is above 50% for an aspect’s gold rationales.

Conclusion

- Learning without supervision (or with weak supervision) can provide very good results in many different problems
- It is mainly relied on the use of additional learning models as regularizers (i.e inductive bias)
- The natural extension of fully unsupervised methods is few-shot / semi-supervised learning that can greatly improve the quality of the models.