Language Generation with Continuous Outputs

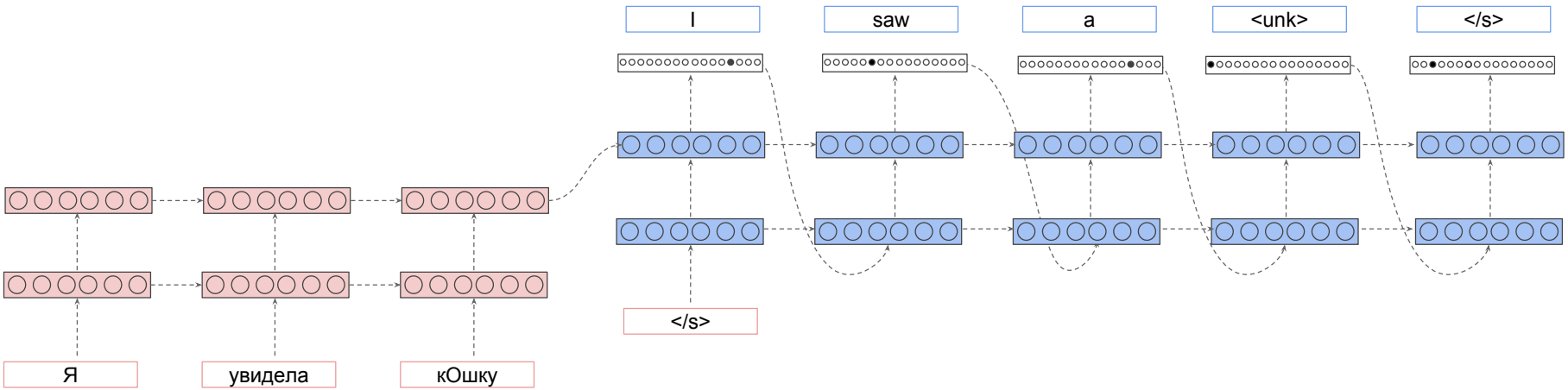
Yulia Tsvetkov

Carnegie Mellon University

August 9, 2018

Carnegie Mellon University

Encoder–Decoder Architectures

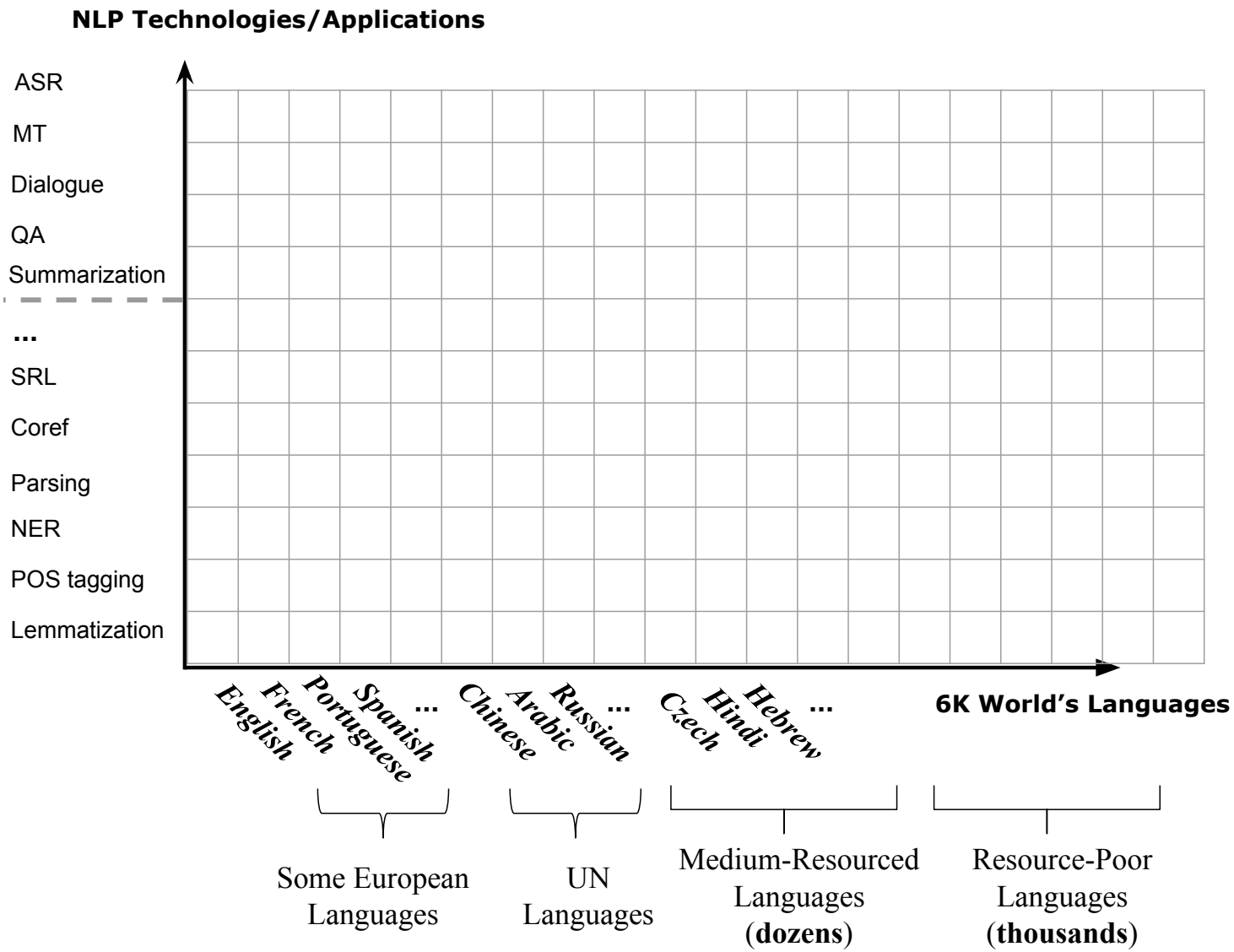


(Conditional) Language Generation

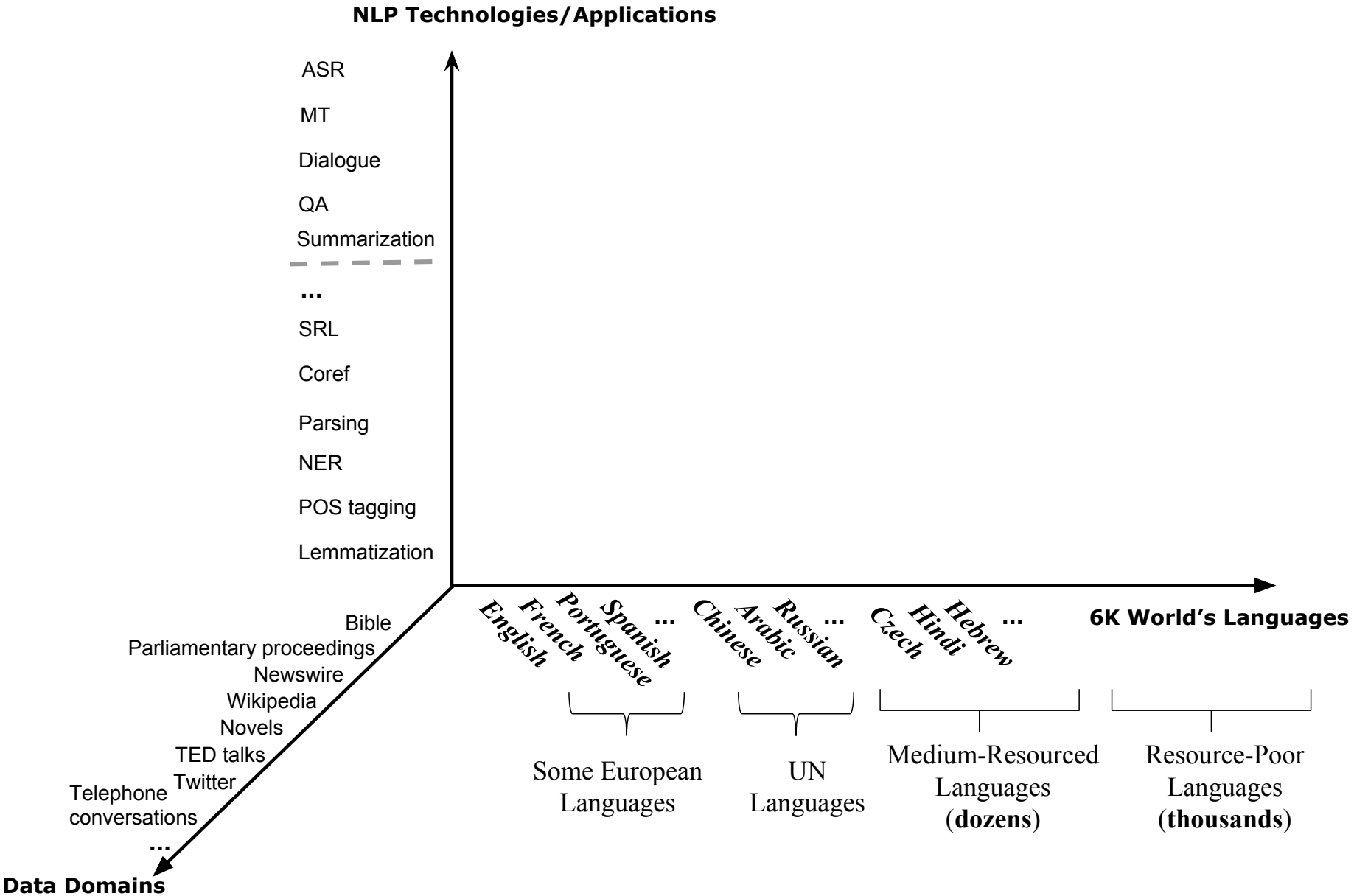
Task + Data + Language

- NLG
- Machine Translation
- Summarization
- Dialogue
- Caption Generation
- Speech Recognition
- ...

(Conditional) Language Generation – 2D



(Conditional) Language Generation – 3D

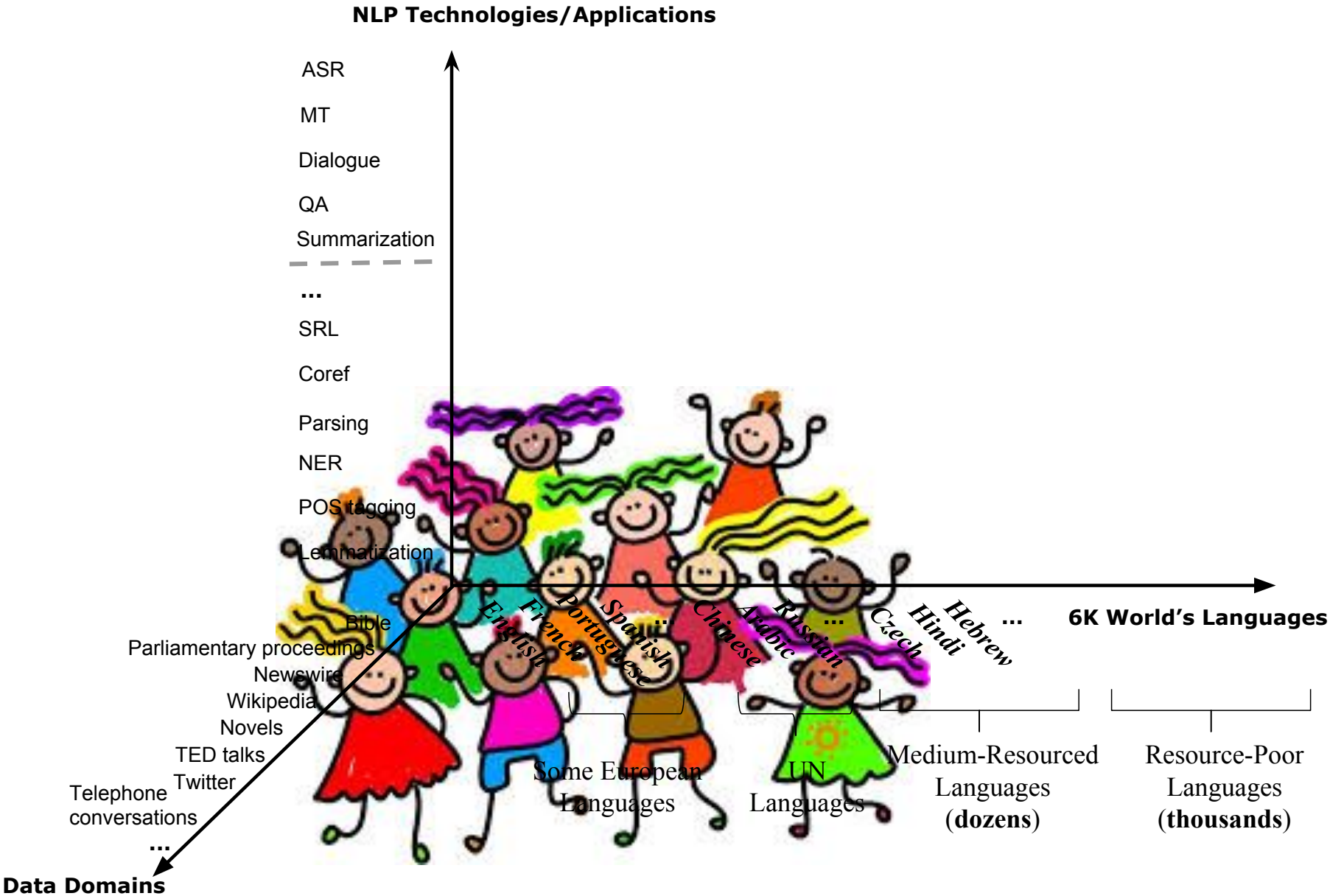


NLP ≠ Task + Data

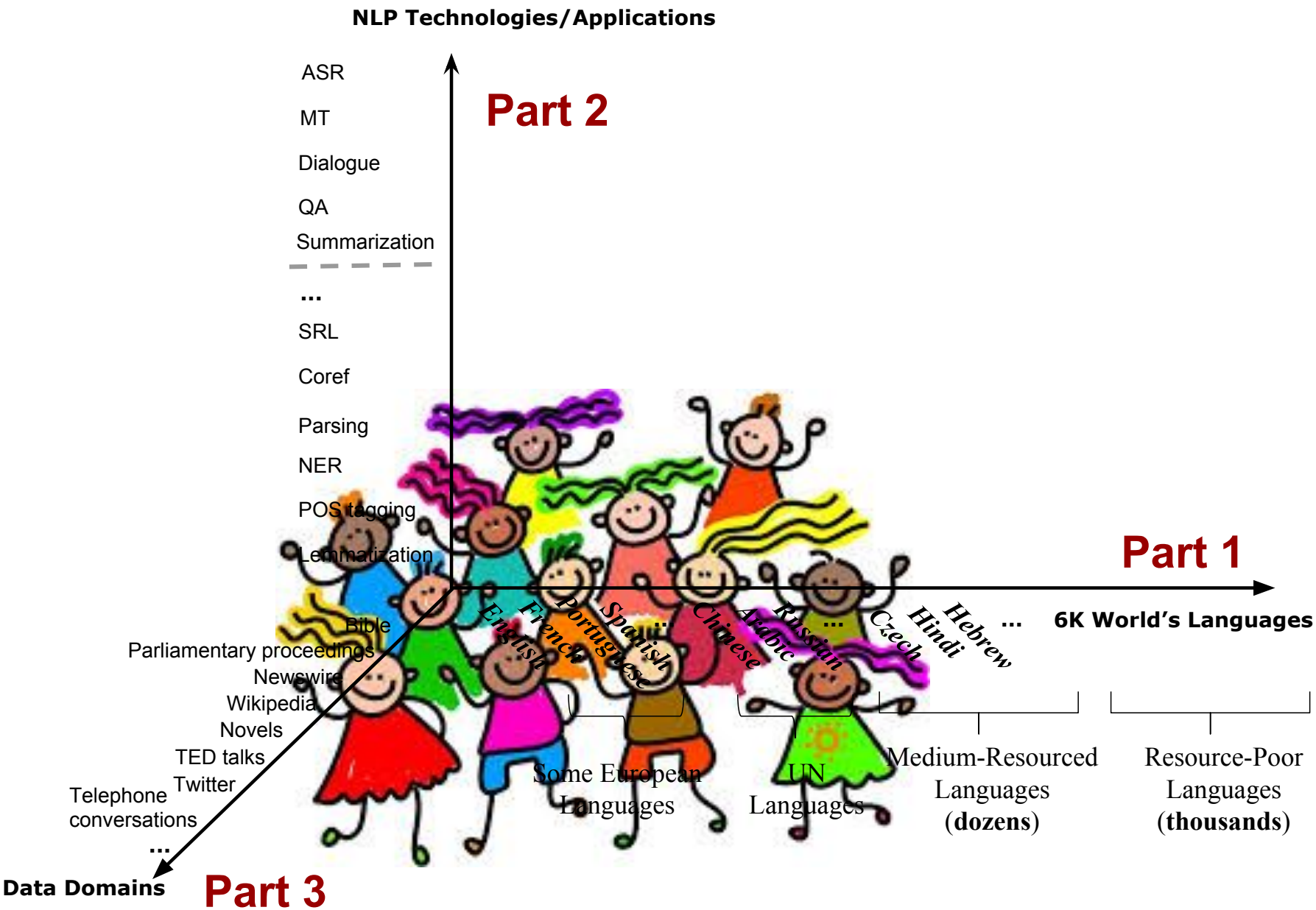
The common misconception is that language
has to do with **words** and what they mean.
It doesn't.
It has to do with **people** and what ***they*** mean.

Herbert H. Clark & Michael F. Schober, 1992
+ Dan Jurafsky's keynote at CVPR'17 and EMNLP'17

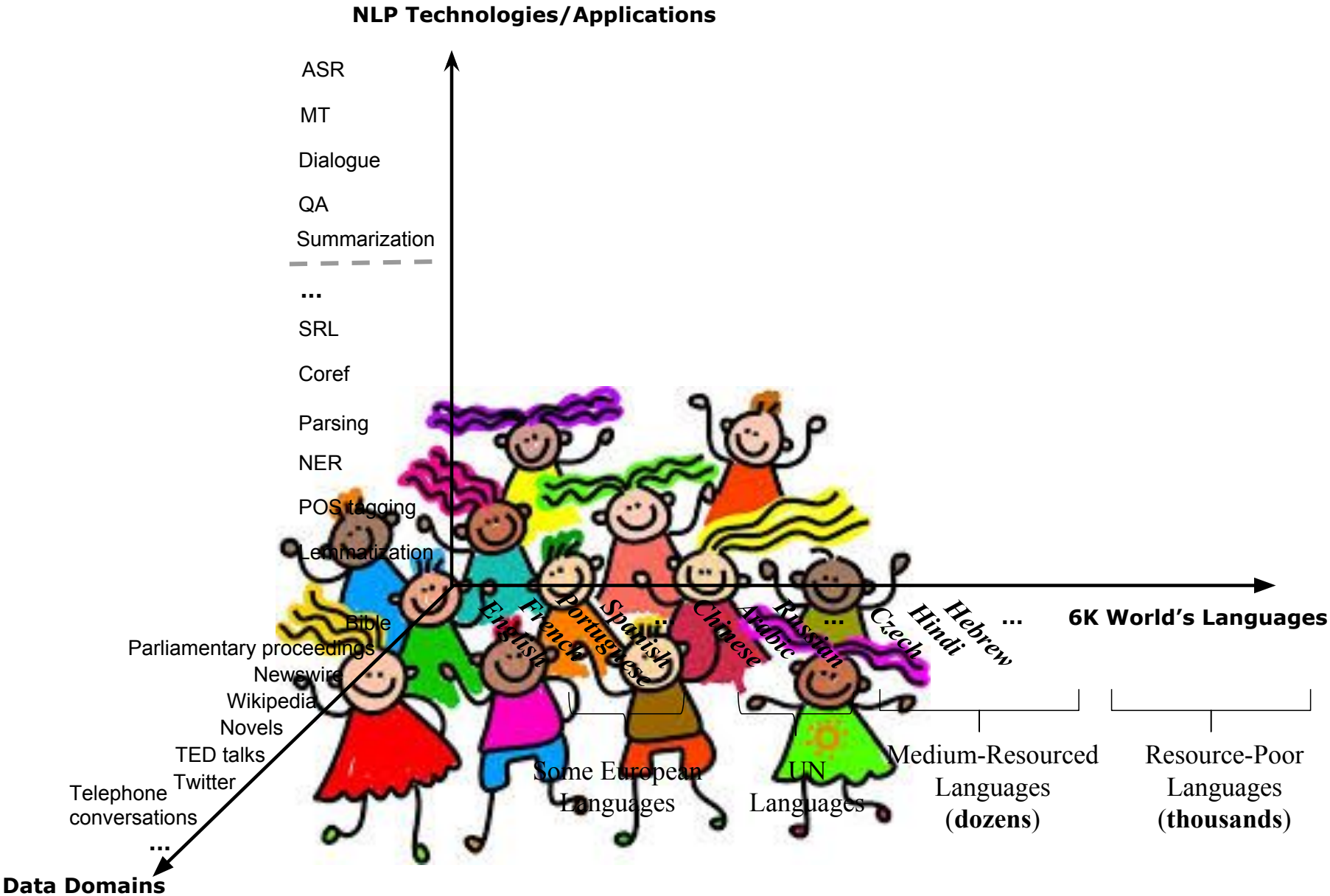
(Conditional) Language Generation – ∞D



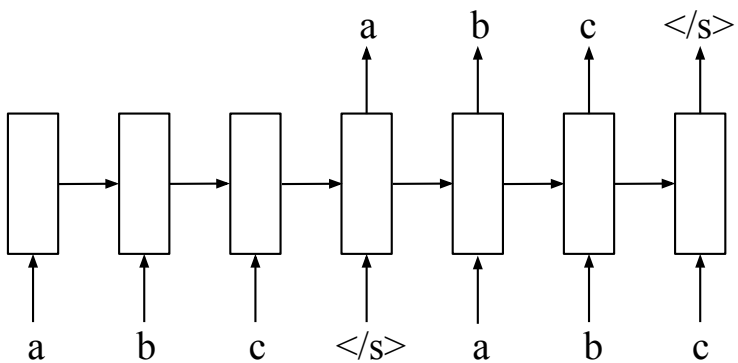
Outline

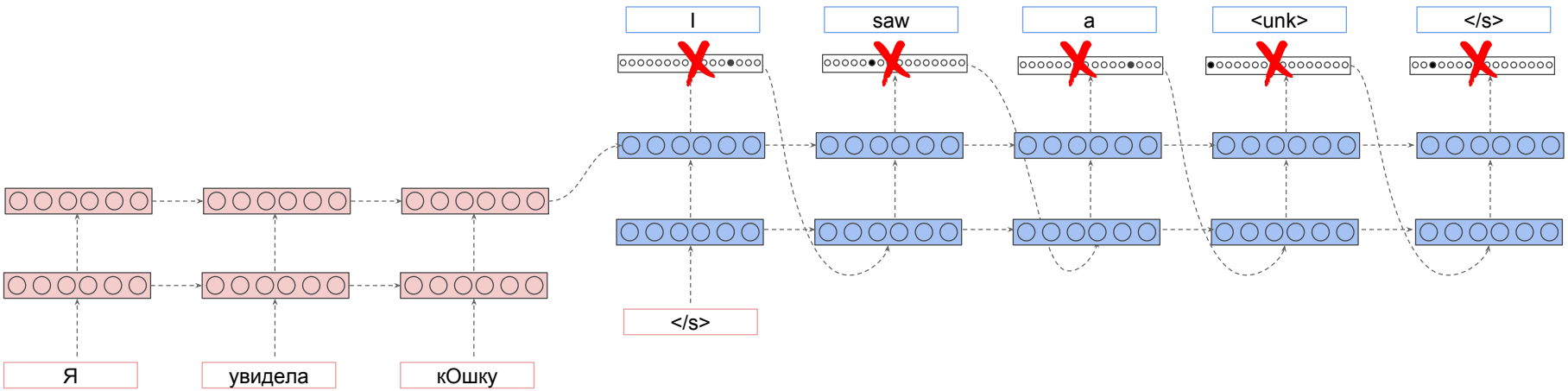

NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
Part 2
Part 1
6K World's Languages
English
French
Portuguese
Spanish
...
Chinese
Arabic
...
Russian
Czech
Hindi
Hebrew
...
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Data Domains
Part 3

(Conditional) Language Generation – ∞D

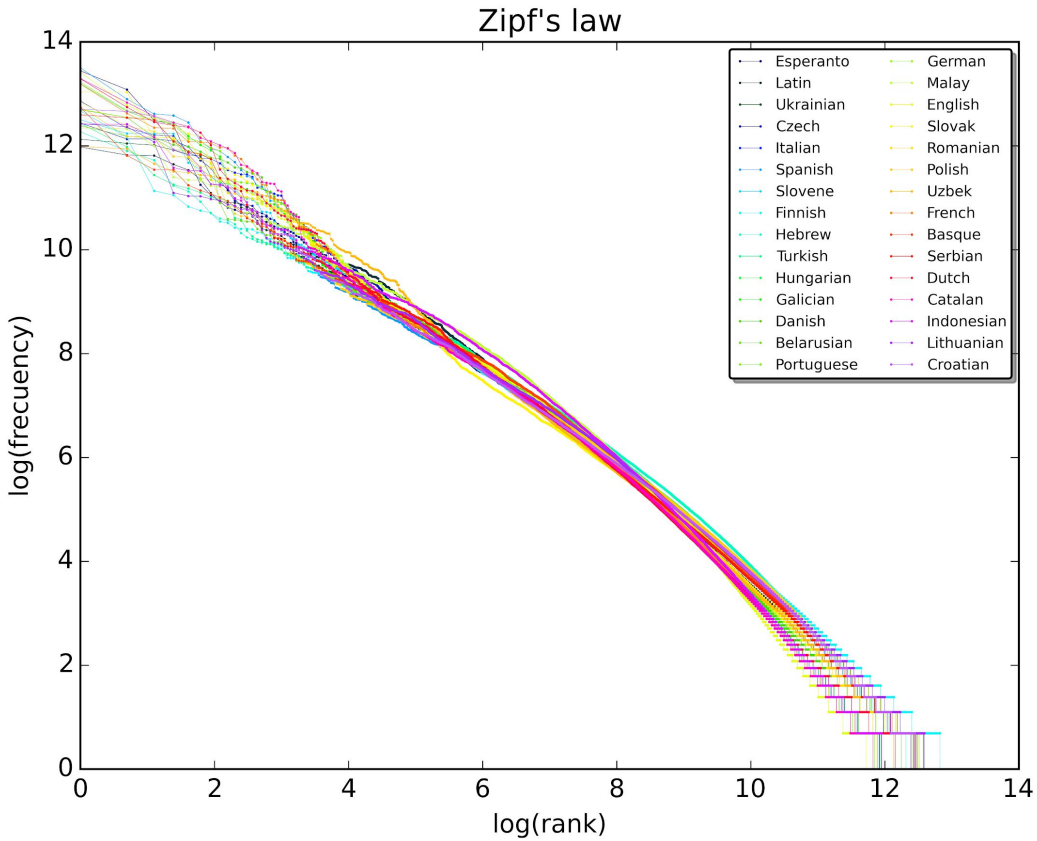


Language Generation with Continuous Outputs

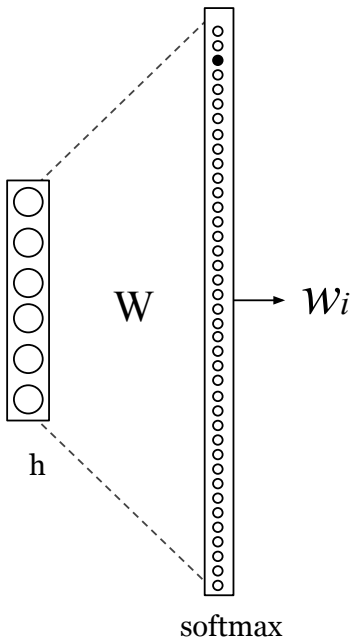



I
saw
a
<unk>
</s>
</s>
Я
увидела
кОшку

Rare Words Are Common in Language



Softmax



$$P(w_i|h) = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

- Multinomial distribution over discrete and mutually exclusive alternatives
- High computational and memory complexity
- Vocabulary size is limited to a small fraction of words plus <unk>
- Words are represented as 1-hot vectors

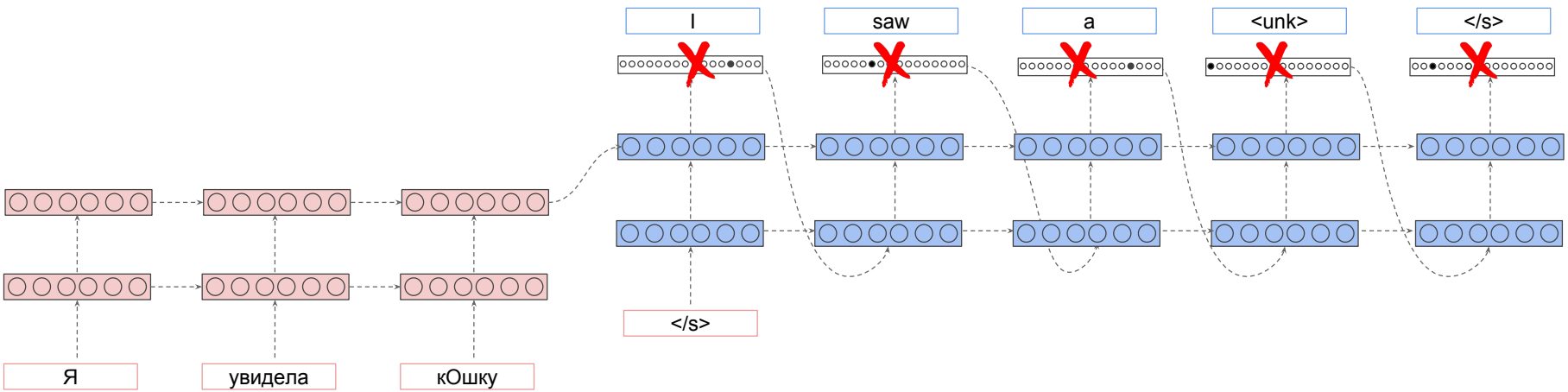
Alternatives to Softmax

- Sampling-based approximations
  - Importance Sampling: evaluate the denominator over a subset
  - Noise Contrastive Estimation: convert to a proxy binary classification problem
  - ...
- Structure-based approximations
  - Differentiated Softmax: divide the vocabulary to multiple classes; first predict a class, then predict a word of the class
  - Hierarchical Softmax: binary tree with words as leaves
  - ...
- Subword Units
  - Byte Pair Encoding (BPE) (Sennrich et al. '2016)

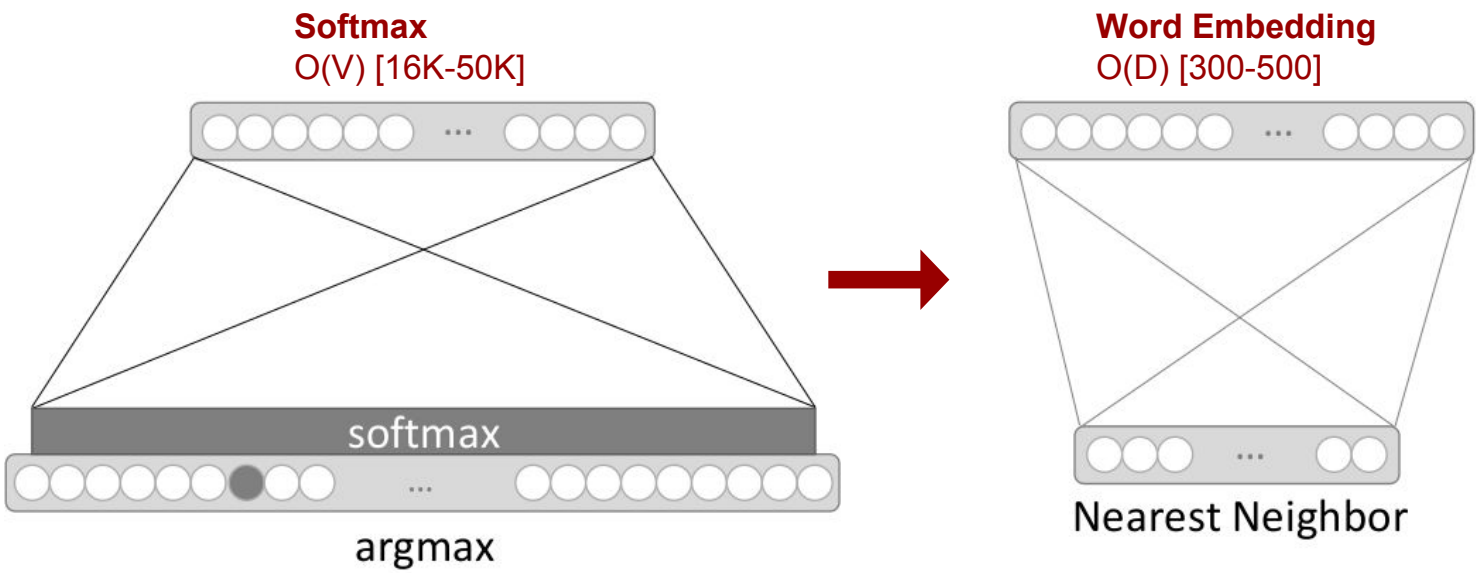
Alternatives to Softmax

	Sampling Based	Structure Based	Subword Units
Training Time	☺	☺	☺
Test Time	😐	😐	☺
Accuracy	☹	☹	☺
Memory	😐	☹	☺
Very Large Vocab	☹	☹	☺

Our Proposal: No Softmax

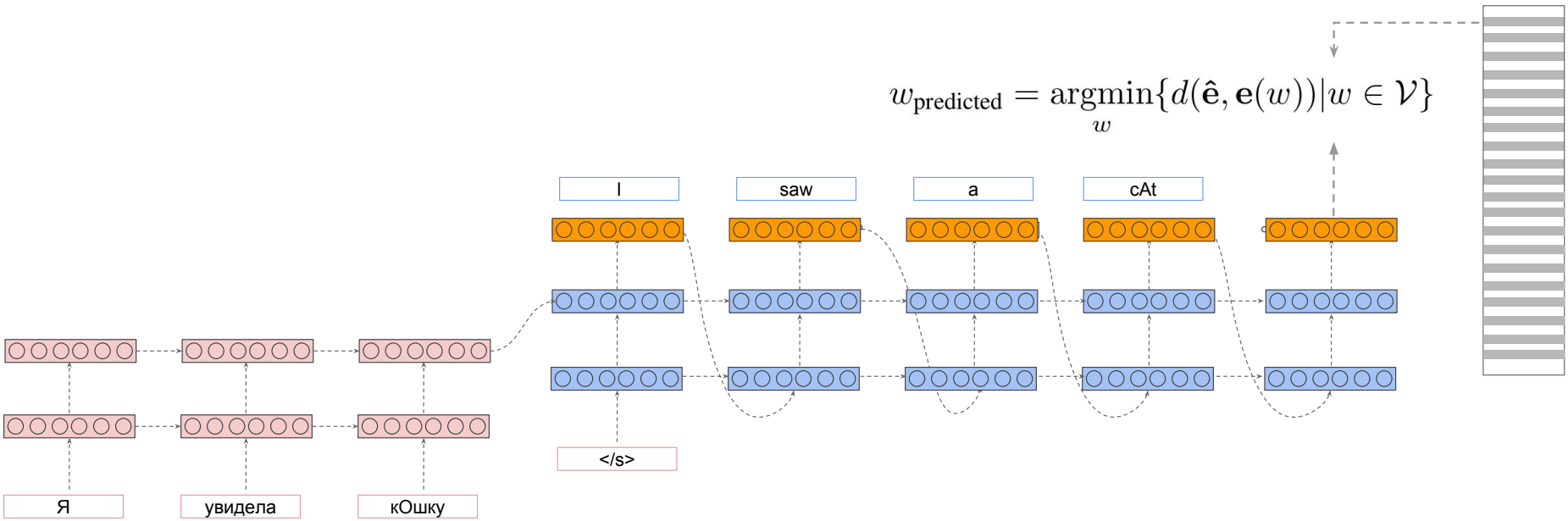


Our Proposal



- Represent each word by it's pre-trained embedding instead of a 1-hot vector

Seq2Seq with Continuous Outputs



- At each time-step t, generate the word's embedding instead of a probability distribution over the vocabulary.
- Training (next slides)
- Decoding: kNN

Training Seq2Seq with Continuous Outputs: Empirical Losses

- Euclidean Loss

$$\mathscr{L}_{\mathrm{L2}} = \|\hat{\mathbf{e}} - \mathbf{e}(w)\|^2$$

- Cosine Loss

$$\mathscr{L}_{\mathrm{cosine}} = 1 - \frac{\hat{\mathbf{e}}^T \mathbf{e}(w)}{\|\hat{\mathbf{e}}\|.\|\mathbf{e}(w)\|}$$

- Max-Margin Loss

$$\mathscr{L}_{\mathrm{mm}} = \sum_{w' \in \mathcal{V}, w' \neq w} \max\{0, \gamma + \cos(\hat{\mathbf{e}}, \mathbf{e}(w')) - \cos(\hat{\mathbf{e}}, \mathbf{e}(w))\}$$

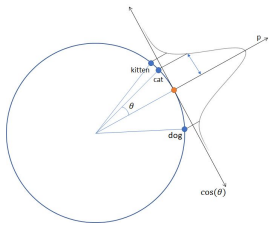
Training Seq2Seq with Continuous Outputs: Probabilistic Loss

- von Mises Fisher (vMF) Distribution

$$p(\mathbf{e}(w); \boldsymbol{\mu}, \kappa) = C_m(\kappa) e^{\kappa \boldsymbol{\mu}^T \mathbf{e}(w)}$$

We use $\kappa = \|\hat{\mathbf{e}}\|$,

$$p(\mathbf{e}(w); \hat{\mathbf{e}}) = C_m(\|\hat{\mathbf{e}}\|) e^{\hat{\mathbf{e}}^T \mathbf{e}(w)}$$



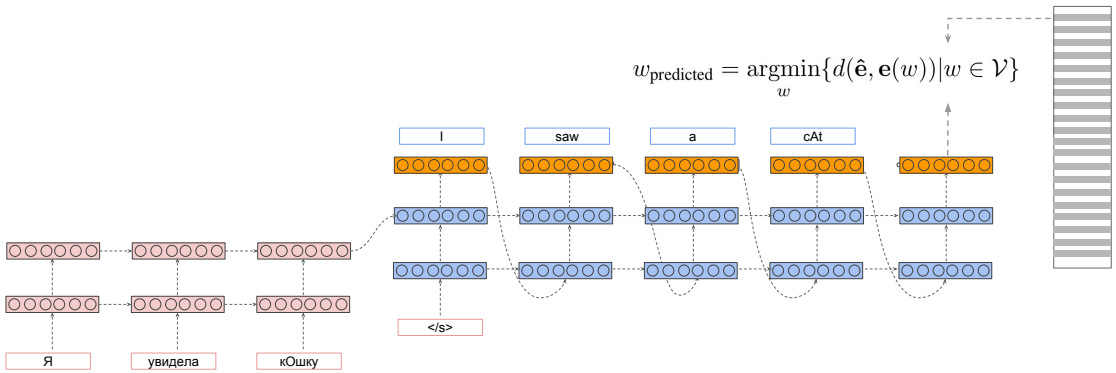
- vMF Loss

$$\mathscr{L}_{\text{NLLvMF}} = -\log(C_m \|\hat{\mathbf{e}}\|) - \hat{\mathbf{e}}^T \mathbf{e}(w)$$

+regularization

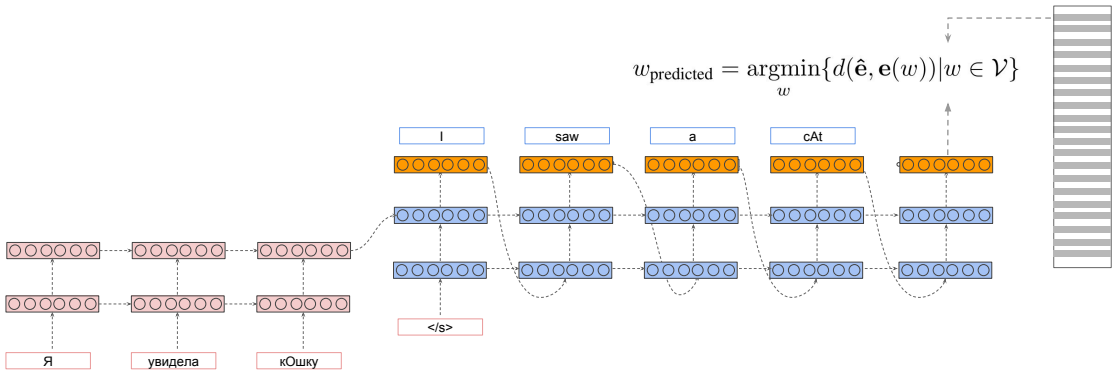
$$\mathscr{L}_{\text{NLLvMF-reg1}} = -\log C_m(\|\hat{\mathbf{e}}\|) - \hat{\mathbf{e}}^T \mathbf{e}(w) + \lambda_1 \|\hat{\mathbf{e}}\|$$
$$\mathscr{L}_{\text{NLLvMF-reg2}} = -\log C_m(\|\hat{\mathbf{e}}\|) - \lambda_2 \hat{\mathbf{e}}^T \mathbf{e}(w)$$

Training Seq2Seq with Continuous Outputs: Research Questions



- Objective function: empirical and probabilistic losses
- Embeddings: word2vec, fasttext, syntactic, morphological, ELMO, etc.
- Attention: words vs. BPE in the input
- Decoding: scheduled sampling; kNN approximations; beam search; post-processing with LMs
- OOVs: scheduled sampling; tied embeddings

Training Seq2Seq with Continuous Outputs: Research Questions



- Objective function: empirical and probabilistic losses
- Embeddings: word2vec, fasttext, syntactic, morphological, ELMO, etc.
- Attention: words vs. BPE in the input
- Decoding: scheduled sampling; kNN approximations; beam search; interpolation with LMs
- OOVs: scheduled sampling; tied embeddings

Experimental Setup

	IWSLT fr–en tst2015+tst2016
train	220K
dev	2.3K
test	2.2K

- Stronger Baselines for Trustable Results in NMT (Denkowski & Neubig '17)
- BLEU
- 50K word vocab; 16K BPE vocab
- 300-dimensional embeddings
- More setups in the paper: IWSLT de–en, IWSLT en–fr, WMT de–en

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

Translation Quality

Source Type/ Target Type	**Loss**	**BLEU fr–en**
word → word	cross-entropy	30.98
word → BPE	cross-entropy	29.06
BPE → BPE	cross-entropy	**31.44**
BPE → word2vec	L2	16.78
BPE → word2vec	cosine	26.92
word → word2vec	L2	27.16
word → word2vec	cosine	29.14
word → word2vec	max-margin	29.56
word → fasttext	max-margin	30.98
word → fasttext + tied	max-margin	**32.12**
word → fasttext	$\text{NLLvMF}_{\text{reg1+reg2}}$	30.38
word → fasttext + tied	$\text{NLLvMF}_{\text{reg1+reg2}}$	**31.63**

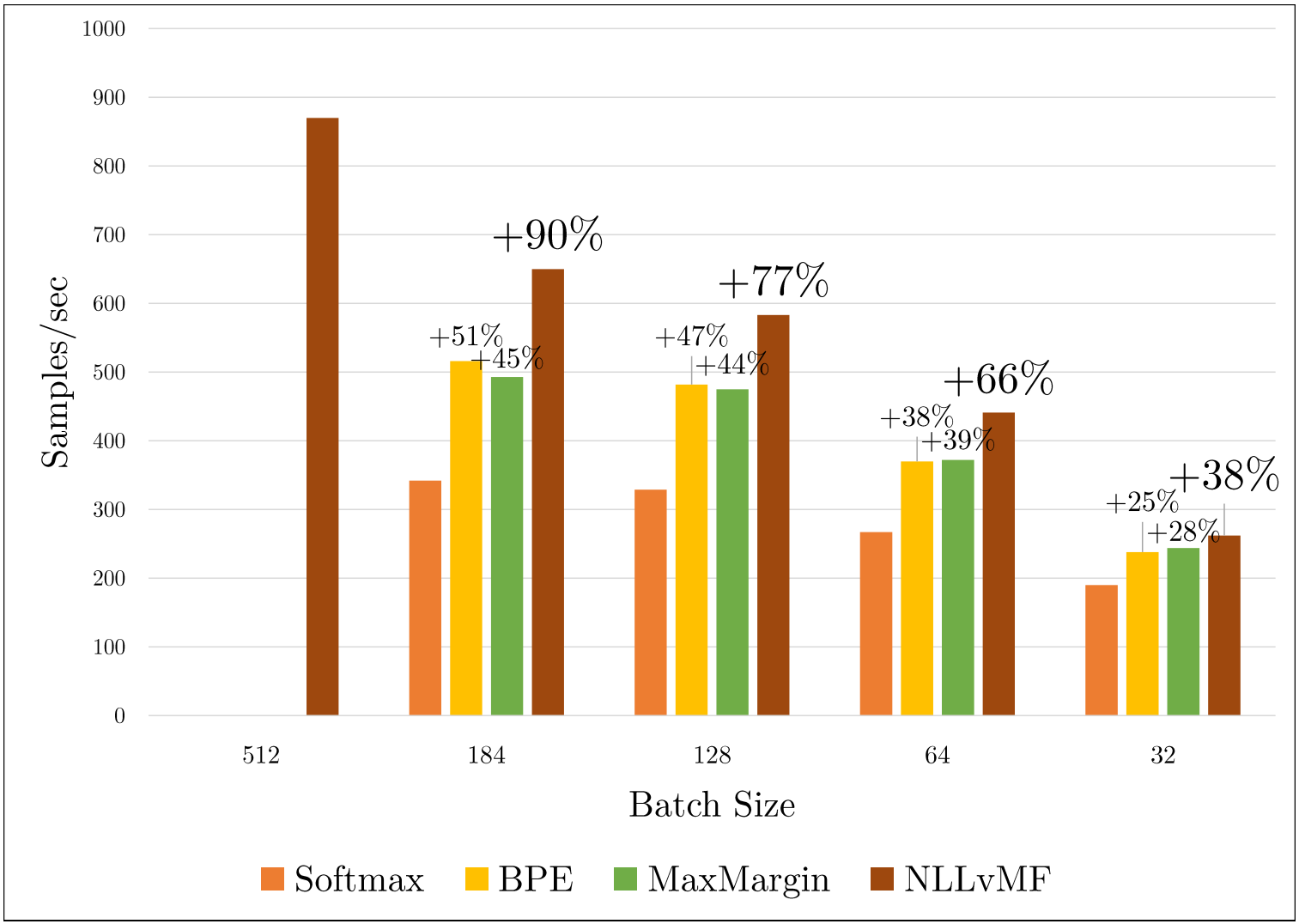
Training Time & Memory

* 1 GeForce GTX TITAN X GPU

	Softmax baseline	**BPE** baseline	**NLLvMF** best model
fr–en	4h	4.5h	**1.9h**
de–en	3h	3.5h	**1.5h**
en–fr	1.8	2.8h	**1.3h**
WMT de-en	4.3d	4.5d	**1.6d**

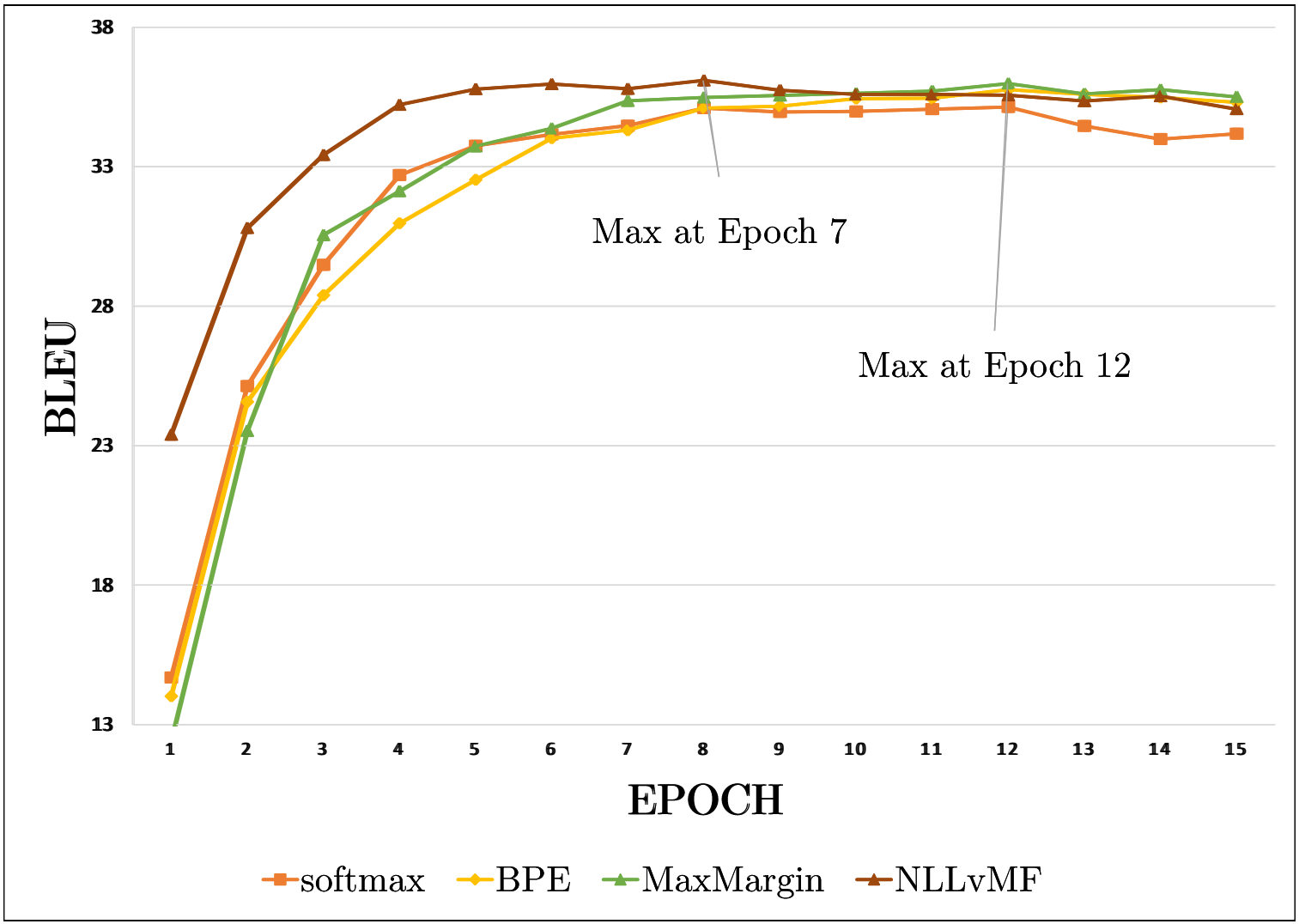
	# Parameters in the Output Layer
Softmax	51.2M (1.0x)
BPE	16.384M (0.32x)
NLLvMF	307.2K (**0.006x**)

Training Time & Memory



Encoder–Decoder with Continuous Outputs

Convergence Time



Encoder–Decoder with Continuous Outputs

Output Example

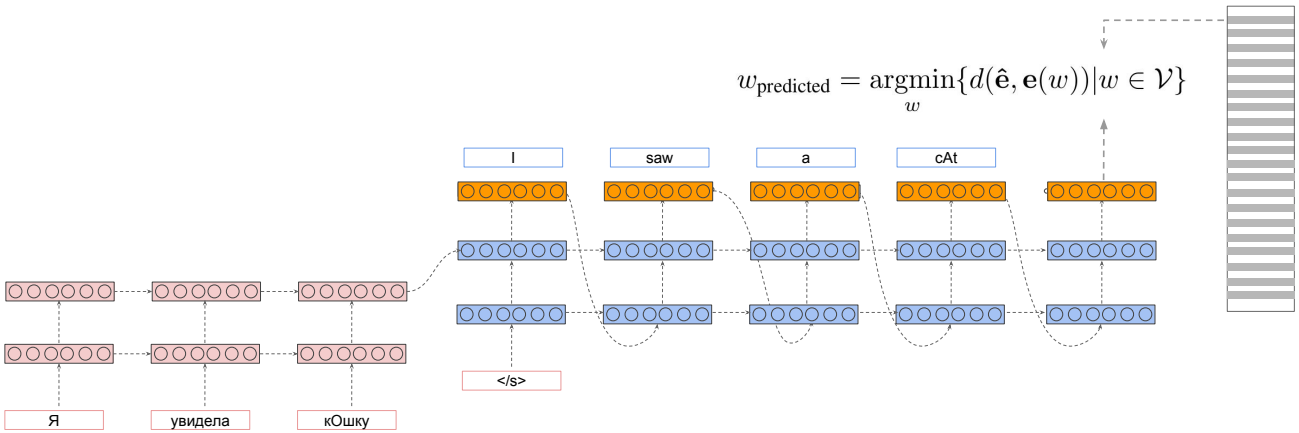
- **GOLD**
  *An education is critical , but **tackling** this problem is going to **require each and everyone of us to step up** and be **better role models** for the women and girls in our own lives .*
- **BPE2BPE**
  education is critical , but it's going to require that each of us will come in and if you do a better example for women and girls in our lives .
- **WORD2FASTTEXT**
  *education is critical , but **fixed** this problem is going to **require that all of us engage** and be **a better example** for women and girls in our lives .*

Seq2Seq with Continuous Outputs

	Sampling Based	Structure Based	Subword Units	Semfit
Training Time				
Test Time				
Accuracy				
Memory				
Handle Very Large Vocab				

Encoder–Decoder with Continuous Outputs

Future Research Questions



- Decoding
- Translation into morphologically-rich languages
- Low-resource NMT
- More generation tasks, e.g. style transfer with GANs


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
thank you
gracias
хвала
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràçie
hvala
köszönöm
akun
dankon
ačiū
djiere dieuf
tau
mochchakkeram
murakoze
enkosi
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
obrigado
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci