Language Generation with Continuous Outputs

Yulia Tsvetkov

Carnegie Mellon University

August 9, 2018
(Conditional) Language Generation

Task + Data + Language

- NLG
- Machine Translation
- Summarization
- Dialogue
- Caption Generation
- Speech Recognition
- ...
(Conditional) Language Generation – 2D

NLP Technologies/Applications

ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization

6K World’s Languages

Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
(Conditional) Language Generation – 3D

NLP Technologies/Applications
- ASR
- MT
- Dialogue
- QA
- Summarization
- SRL
- Coref
- Parsing
- NER
- POS tagging
- Lemmatization

6K World’s Languages

Medium-Resourced Languages (dozens)
- Portuguese
- Chinese
- Russian
- Czech
- Hindi
- ... (UN Languages)

Some European Languages
- English
- French
- Spanish
- ... (Resource-Poor Languages)
- Hindi
- Hebrew
- ... (thousands)

Data Domains
- Bible
- Parliamentary proceedings
- Newswire
- Wikipedia
- Novels
- TED talks
- Twitter
- Telephone conversations
- ...
NLP ≠ Task + Data

The common misconception is that language has to do with \textit{words} and what they mean. It doesn’t.
It has to do with \textit{people} and what \textit{they} mean.

Herbert H. Clark & Michael F. Schober, 1992
+ Dan Jurafsky’s keynote at CVPR’17 and EMNLP’17
(Conditional) Language Generation – ∞D

NLP Technologies/Applications

ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
...

6K World’s Languages

UN Languages

Some European Languages

Medium-Resourced Languages (dozens)

Resource-Poor Languages (thousands)

Data Domains

Parliamentary proceedings
Newswire
Wikipedia
Novels
Twitter
TED talks
Telephone conversations...
Outline

NLP Technologies/Applications

Part 1

6K World’s Languages

Part 2

Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
Twitter
TED talks
Telephone conversations

Data Domains

Part 3

ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization

Part 1

Part 2

Part 3

Resource-Poor Languages (thousands)
Medium-Resourced Languages (dozens)
UN Languages
Some European Languages
Czech
Hebrew
Hindi

6K World’s Languages

Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

Part 1

Part 2

Part 3

Data Domains

NLP Technologies/Applications

6K World’s Languages

Lemmatization

POS tagging

Parsing

NER

Coref

SRL

Summarization

...
(Conditional) Language Generation – ∞D

NLP Technologies/Applications

ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization

6K World’s Languages

Some European Languages

UN Languages

Medium-Resourced Languages (dozens)

Resource-Poor Languages (thousands)

Data Domains

Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations

Some European Languages

English
French
Spanish
Portuguese
Russian
...
Language Generation with Continuous Outputs
Я увидела кОшку
I saw a
</s>
Rare Words Are Common in Language

By SergioJimenez - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45516736
Softmax

- 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

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

<table>
<thead>
<tr>
<th></th>
<th>Sampling Based</th>
<th>Structure Based</th>
<th>Subword Units</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Time</strong></td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Test Time</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Very Large Vocab</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
</tr>
</tbody>
</table>
Our Proposal: No Softmax
Our Proposal

- Represent each word by its 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**
  \[
  \mathcal{L}_{L2} = \| \hat{\mathbf{e}} - \mathbf{e}(w) \|^2
  \]

- **Cosine Loss**
  \[
  \mathcal{L}_{\text{cosine}} = 1 - \frac{\hat{\mathbf{e}}^T \mathbf{e}(w)}{\|\hat{\mathbf{e}}\| \|\mathbf{e}(w)\|}
  \]

- **Max-Margin Loss**
  \[
  \mathcal{L}_{\text{mm}} = \sum_{w' \in \mathcal{V}, w' \neq w} \max\{0, \gamma + \cos(\hat{\mathbf{e}}, \mathbf{e}(w')) - \cos(\hat{\mathbf{e}}, \mathbf{e}(w))\}
  \]
Training Seq2Seq with Continuous Outputs: Probabilistic Loss

- von Mises Fisher (vMF) Distribution
  \[ p(e(w); \mu, \kappa) = C_m(\kappa) e^{\kappa \mu^T e(w)} \]

  We use \( \kappa = \|\hat{e}\| \),

  \[ p(e(w); \hat{e}) = C_m(\|\hat{e}\|) e^{\hat{e}^T e(w)} \]

- vMF Loss

  \[ \mathcal{L}_{\text{NLLvMF}} = -\log(C_m\|\hat{e}\|) - \hat{e}^T e(w) \]

  + regularization

  \[ \mathcal{L}_{\text{NLLvMF-reg1}} = -\log C_m(\|\hat{e}\|) - \hat{e}^T e(w) + \lambda_1 \|\hat{e}\| \]

  \[ \mathcal{L}_{\text{NLLvMF-reg2}} = -\log C_m(\|\hat{e}\|) - \lambda_2 \hat{e}^T e(w) \]
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
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

<table>
<thead>
<tr>
<th></th>
<th>IWSLT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fr–en</td>
</tr>
<tr>
<td></td>
<td>tst2015+tst2016</td>
</tr>
<tr>
<td>train</td>
<td>220K</td>
</tr>
<tr>
<td>dev</td>
<td>2.3K</td>
</tr>
<tr>
<td>test</td>
<td>2.2K</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU fr–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td><strong>31.44</strong></td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td><strong>32.12</strong></td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF$_{\text{reg1+reg2}}$</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF$_{\text{reg1+reg2}}$</td>
<td><strong>31.63</strong></td>
</tr>
</tbody>
</table>
## Translation Quality

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU fr–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td><strong>31.44</strong></td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td><strong>32.12</strong></td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF_{reg1+reg2}</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF_{reg1+reg2}</td>
<td><strong>31.63</strong></td>
</tr>
</tbody>
</table>
## Translation Quality

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU fr–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td>31.44</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td>32.12</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF_{reg1+reg2}</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF_{reg1+reg2}</td>
<td>31.63</td>
</tr>
</tbody>
</table>
## Translation Quality

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU (\text{fr–en})</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td><strong>31.44</strong></td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td><strong>32.12</strong></td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF(_{\text{reg1+reg2}})</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF(_{\text{reg1+reg2}})</td>
<td><strong>31.63</strong></td>
</tr>
</tbody>
</table>
## Translation Quality

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU fr–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td><strong>31.44</strong></td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td><strong>32.12</strong></td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF_{reg1+reg2}</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF_{reg1+reg2}</td>
<td><strong>31.63</strong></td>
</tr>
</tbody>
</table>
## Translation Quality

<table>
<thead>
<tr>
<th>Source Type/Target Type</th>
<th>Loss</th>
<th>BLEU fr–en</th>
</tr>
</thead>
<tbody>
<tr>
<td>word → word</td>
<td>cross-entropy</td>
<td>30.98</td>
</tr>
<tr>
<td>word → BPE</td>
<td>cross-entropy</td>
<td>29.06</td>
</tr>
<tr>
<td>BPE → BPE</td>
<td>cross-entropy</td>
<td><strong>31.44</strong></td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>L2</td>
<td>16.78</td>
</tr>
<tr>
<td>BPE → word2vec</td>
<td>cosine</td>
<td>26.92</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>L2</td>
<td>27.16</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>cosine</td>
<td>29.14</td>
</tr>
<tr>
<td>word → word2vec</td>
<td>max-margin</td>
<td>29.56</td>
</tr>
<tr>
<td>word → fasttext</td>
<td>max-margin</td>
<td>30.98</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>max-margin</td>
<td><strong>32.12</strong></td>
</tr>
<tr>
<td>word → fasttext</td>
<td>NLLvMF_{reg1+reg2}</td>
<td>30.38</td>
</tr>
<tr>
<td>word → fasttext + tied</td>
<td>NLLvMF_{reg1+reg2}</td>
<td><strong>31.63</strong></td>
</tr>
</tbody>
</table>
**Training Time & Memory**

* 1 GeForce GTX TITAN X GPU

<table>
<thead>
<tr>
<th></th>
<th>Softmax baseline</th>
<th>BPE baseline</th>
<th>NLLvMF best model</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr–en</td>
<td>4h</td>
<td>4.5h</td>
<td>1.9h</td>
</tr>
<tr>
<td>de–en</td>
<td>3h</td>
<td>3.5h</td>
<td>1.5h</td>
</tr>
<tr>
<td>en–fr</td>
<td>1.8</td>
<td>2.8h</td>
<td>1.3h</td>
</tr>
<tr>
<td>WMT de-en</td>
<td>4.3d</td>
<td>4.5d</td>
<td>1.6d</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Parameters in the Output Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax</td>
</tr>
<tr>
<td>BPE</td>
</tr>
<tr>
<td>NLLvMF</td>
</tr>
</tbody>
</table>
Training Time & Memory

![Graph showing the relationship between batch size and samples/sec for Softmax, BPE, MaxMargin, and NLLvMF. The graph indicates improvements in samples/sec as batch size decreases, with percentage increases marked for each batch size.]

- **Softmax**: +90% at 512, +51% at 184, +47% at 128, +38% at 64, and +25% at 32.
- **BPE**: +90% at 512, +51% at 184, +47% at 128, +38% at 64, and +25% at 32.
- **MaxMargin**: +90% at 512, +51% at 184, +47% at 128, +38% at 64, and +25% at 32.
- **NLLvMF**: +90% at 512, +51% at 184, +47% at 128, +38% at 64, and +25% at 32.
Encoder–Decoder with Continuous Outputs

Convergence Time

![Graph showing BLEU scores over epochs for different methods: softmax, BPE, MaxMargin, NLLvMF. The max score is reached at epoch 7 and 12.](image)
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

<table>
<thead>
<tr>
<th></th>
<th>Sampling Based</th>
<th>Structure Based</th>
<th>Subword Units</th>
<th>Semfit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Time</strong></td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Test Time</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td><strong>Handle Very Large Vocab</strong></td>
<td>😞</td>
<td>😞</td>
<td>😊</td>
<td>😊</td>
</tr>
</tbody>
</table>

(Note: 😊 = Positive, 😞 = Neutral, 😞 = Negative)
Future Research Questions

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

$$w_{predicted} = \arg\min_w \{d(\hat{e}, e(w)) | w \in \mathcal{V}\}$$
thank you