Deep Generative Model for Joint Alignment and Word Representation

Embedalign

Miguel Rios, Wilker Aziz, Khalil Sima’an
University of Amsterdam
Statistical Language Processing and Learning Lab

June 3, 2018
Outline

1. Introduction
2. Model
3. Evaluation
4. Conclusions and Future Work
Introduction

TL;DR

- Generative model that embeds words in their complete observed context
- Model learns from bilingual sentence-aligned corpora by marginalisation of latent lexical alignments
- Model embeds words as probability densities
- Model shows competitive results on context dependent Natural Language Processing applications
Discriminative embedding models

**word2vec**

*In the event of a chemical spill, most children know they should *evacuate* as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"
Discriminative embedding models \textbf{word2vec}

\textit{In the event of a chemical spill, most children know they should \textbf{evacuate} as advised by people in charge.}

Place words in $\mathbb{R}^d$ as to answer questions like

\textit{“Have I seen this word in this context?”}

Fit a binary classifier

- \textbf{positive} examples
- \textbf{negative} examples
In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

- Limitations
In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

- Limitations
  - Representation learning is an unsupervised problem we only observe positive/complete context
Introduction

In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

Limitations

- Representation learning is an unsupervised problem we only observe positive/complete context
- Distributional hypothesis is strong but fails when context is not discriminative
In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

- Limitations
  - Representation learning is an unsupervised problem we only observe positive/complete context
  - Distributional hypothesis is strong but fails when context is not discriminative
  - Word senses are collapsed into one vector
Embedalign

- Generative model to induce word representations
In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

- Generative model to induce word representations
- Learn from positive examples
In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

- Generative model to induce word representations
- Learn from positive examples
- Learn from richer (less ambiguous) context
  Foreign text is proxy to sense supervision (Diab and Resnik, 2002)

En caso de un derrame de productos químicos, la mayoría de los niños saben que deben abandonar el lugar según lo aconsejado por las personas a cargo.
Generative Model

quickly evacuate the area / deje el lugar rápidamente

\[ z \rightarrow x \rightarrow y \rightarrow a \]

\( \theta \)

\(|B| \)

\[
\begin{array}{c}
X \\
Z \\
A \\
Z_a \\
Y
\end{array}
\]
Generative Model

quickly evacuate the area / deje el lugar rápidamente
Generative Model

quickly evacuate the area / deje el lugar rápidamente

\[ z \rightarrow x \rightarrow y \rightarrow a \]

\[ |\mathcal{B}| \]

\[ \theta \]

\[ a_4 = 1 \]

\[ \text{rápido} \]

\[ \text{mente} \]

\[ \text{quickly}_1 \]

\[ \text{X} \]

\[ \text{Z} \]

\[ \text{Z}_1 \]

\[ \text{A} \]

\[ \text{Z}_a \]

\[ \text{Y} \]
Generative Model

quickly evacuate the area / deje el lugar rápidamente

\[ \begin{align*}
  &X \\
  &Z \\
  &A \\
  &Y \\
  &Z_a \\
  &Z_1 \\
  &Z_2 \\
  \end{align*} \]

\( a_4 = 1 \)

\(|B| \)

\( \theta \)
Generative Model

quickly evacuate the area  /  deje el lugar rápidamente
Generative Model

quickly evacuate the area / deje el lugar rápidamente
quickly evacuate the area / deje el lugar rápidamente
Generative Model

quickly evacuate the area / deje el lugar rápidamente

\[ z \rightarrow x \rightarrow y \rightarrow \theta \]

\[ a \rightarrow \text{ } m \rightarrow n \]

\[ |B| \]

\[ X \rightarrow \text{quickly}_1 \rightarrow \text{evacuate}_2 \rightarrow \text{the}_3 \rightarrow \text{area}_4 \]

\[ Z \rightarrow Z_1 \rightarrow Z_2 \rightarrow Z_3 \rightarrow Z_4 \]

\[ A \rightarrow a_1=2 \rightarrow a_2=3 \rightarrow a_4=1 \]

\[ Z_a \rightarrow Z_2 \rightarrow Z_3 \rightarrow Z_1 \]

\[ Y \rightarrow \text{deje}_1 \rightarrow \text{el}_2 \rightarrow \text{rápidamente}_4 \]
quickly evacuate the area / deje el lugar rápidamente
Learning

1. Read sentence

\begin{itemize}
  \item Read sentence
  \item Predict posterior mean $\mu_i$ and std $\sigma_i$
  \item Sample $z_i \sim N(\mu_i, \sigma_i^2)$
  \item Predict categorical distributions
  \item Generate observations
  \item Maximise a lowerbound on likelihood (Kingma and Welling, 2014)
\end{itemize}
Learning

1. Read sentence
2. Predict posterior mean $\mu_i$ and std $\sigma_i$
Learning

1. Read sentence
2. Predict posterior mean $\mu_i$ and std $\sigma_i$
3. Sample $z_i \sim N(\mu_i, \sigma_i^2)$

$\mu_1 \times \mu_2 \times \mu_3$

$\sigma_1 \times \sigma_2 \times \sigma_3$

$x_1 \times x_2 \times x_3$

$\text{evacuate}_1 \text{ the}_2 \text{ area}_3$
1. Read sentence
2. Predict posterior mean $\mu_i$ and std $\sigma_i$
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
4. Predict categorical distributions

$\text{evacuate}_1 \text{ the}_2 \text{ area}_3$
Learning

1. Read sentence
2. Predict posterior mean $\mu_i$ and std $\sigma_i$
3. Sample $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
4. Predict categorical distributions
5. Generate observations

$\text{evacuate}_1 \text{ the}_2 \text{ area}_3$ / $\text{deje}_1 \text{ el}_2 \text{ lugar}_3$
Learning

1. Read sentence
2. Predict posterior mean $\mu_i$ and std $\sigma_i$
3. Sample $z_i \sim N(\mu_i, \sigma_i^2)$

4. Predict categorical distributions
5. Generate observations
   evacuate$_1$ the$_2$ area$_3$ / deje$_1$ el$_2$ lugar$_3$

6. Maximise a lowerbound on likelihood
   (Kingma and Welling, 2014)
Outline

1. Introduction
2. Model
3. Evaluation
4. Conclusions and Future Work
## Data

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentence pairs (million)</th>
<th>Tokens (million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl EN-FR</td>
<td>1.7</td>
<td>42.5</td>
</tr>
<tr>
<td>Europarl EN-DE</td>
<td>1.7</td>
<td>43.5</td>
</tr>
</tbody>
</table>
Architecture

Diagram depicting the architecture of a generative model and an inference model. The model consists of:

- **x** and **y** as input variables.
- **f(z)** and **g(z_a)** as functions representing the generative model.
- **sample z** as an input to **f(z)** and **g(z_a)**.
- **u** and **s** as variables with dimensions 100d, connected to **h: BiRNN 100d**.
- **embedding 128d** as an output from the inference model.
- **h: BiRNN 100d** as a hidden layer between **u** and **s**.

The diagram illustrates the flow of data and the connections between different components of the model.
Word Alignment

The proposal will not now be implemented

Les propositions ne seront pas mises en application maintenant
## Word Alignment

- **Model selection on Dev set**

<table>
<thead>
<tr>
<th>Model</th>
<th>En-Fr</th>
<th>En-De</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM1</td>
<td>32.45</td>
<td>46.71</td>
</tr>
<tr>
<td>IBM2</td>
<td>22.61</td>
<td>40.11</td>
</tr>
<tr>
<td><strong>EMBALIGN</strong></td>
<td>29.43 ± 1.84</td>
<td>48.09 ± 2.12</td>
</tr>
</tbody>
</table>
Lexical Substitution

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

- event
- game
Lexical Substitution

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the event.

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the game.
The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the match.

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the event.

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the game.
## Lexical Substitution

<table>
<thead>
<tr>
<th>Model</th>
<th>GAP ↑</th>
<th>Training size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RANDOM</strong></td>
<td>30.0</td>
<td></td>
</tr>
<tr>
<td><strong>SkipGram</strong></td>
<td>44.9</td>
<td>ukWaC-2B</td>
</tr>
<tr>
<td>(Melamud et al., 2015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BSG</strong></td>
<td>46.1</td>
<td>ukWaC-2B</td>
</tr>
<tr>
<td>(Bražinskas et al., 2017)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EN</strong></td>
<td>21.31 ± 1.05</td>
<td></td>
</tr>
<tr>
<td><strong>EN-FR</strong></td>
<td>42.19 ± 0.57</td>
<td>Euro-42M</td>
</tr>
<tr>
<td><strong>EN-DE</strong></td>
<td>42.07 ± 0.47</td>
<td></td>
</tr>
</tbody>
</table>
Sentence Evaluation (SentEval)

1. Sentiment Analysis
2. Paraphrasing
3. Textual Entailment
4. Semantic Similarity

Sentence embedding → classifier
Sentence Evaluation (SentEval)

1. Sentiment Analysis
2. Paraphrasing
3. Textual Entailment
4. Semantic Similarity

Too slow for a younger crowd, too shallow for an older one.
Sentence Evaluation (SentEval)

1. Sentiment Analysis
2. Paraphrasing
3. Textual Entailment
4. Semantic Similarity

**Too slow for a younger crowd, too shallow for an older one.**
<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>CR</th>
<th>SUBJ</th>
<th>MPQA</th>
<th>SST</th>
<th>TREC</th>
<th>MRPC</th>
<th>SICK-R</th>
<th>SICK-E</th>
<th>STS14</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SkipGram</strong>&lt;sub&gt;En&lt;/sub&gt;</td>
<td><strong>70.96</strong></td>
<td><strong>76.16</strong></td>
<td><strong>87.24</strong></td>
<td><strong>86.87</strong></td>
<td><strong>73.64</strong></td>
<td><strong>65.20</strong></td>
<td><strong>70.7/80.1</strong></td>
<td><strong>0.710</strong></td>
<td><strong>76.2</strong></td>
<td><strong>0.45/0.49</strong></td>
</tr>
<tr>
<td>EN</td>
<td>57.5</td>
<td>67.1</td>
<td>72.0</td>
<td>70.8</td>
<td>57.0</td>
<td>58.0</td>
<td>70.6/80.3</td>
<td>0.648</td>
<td>74.4</td>
<td>0.59/0.59</td>
</tr>
<tr>
<td>EN-Fr</td>
<td>64.0</td>
<td>71.8</td>
<td>79.1</td>
<td>81.5</td>
<td>64.7</td>
<td>58.4</td>
<td>72.1/81.2</td>
<td>0.682</td>
<td>74.6</td>
<td>0.60/0.59</td>
</tr>
<tr>
<td>EN-De</td>
<td>62.6</td>
<td>68.0</td>
<td>77.3</td>
<td>82.0</td>
<td>65.0</td>
<td>66.8</td>
<td>70.4/79.8</td>
<td>0.681</td>
<td>75.5</td>
<td>0.58/0.58</td>
</tr>
<tr>
<td>Combo</td>
<td>66.1</td>
<td>72.4</td>
<td>82.4</td>
<td>84.4</td>
<td>69.8</td>
<td><strong>69.0</strong></td>
<td><strong>71.9/80.6</strong></td>
<td><strong>0.727</strong></td>
<td><strong>76.3</strong></td>
<td><strong>0.62/0.61</strong></td>
</tr>
</tbody>
</table>
# Sentence Evaluation (SentEval)

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>CR</th>
<th>SUBJ</th>
<th>MPQA</th>
<th>SST</th>
<th>TREC</th>
<th>MRPC</th>
<th>SICK-R</th>
<th>SICK-E</th>
<th>STS14</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkipGram&lt;sub&gt;En&lt;/sub&gt;</td>
<td>70.96</td>
<td>76.16</td>
<td>87.24</td>
<td>86.87</td>
<td>73.64</td>
<td>65.20</td>
<td>70.7/80.1</td>
<td>0.710</td>
<td>76.2</td>
<td>0.45/0.49</td>
</tr>
<tr>
<td>En</td>
<td>57.5</td>
<td>67.1</td>
<td>72.0</td>
<td>70.8</td>
<td>57.0</td>
<td>58.0</td>
<td>70.6/80.3</td>
<td>0.648</td>
<td>74.4</td>
<td>0.59/0.59</td>
</tr>
<tr>
<td>En-Fr</td>
<td>64.0</td>
<td>71.8</td>
<td>79.1</td>
<td>81.5</td>
<td>64.7</td>
<td>58.4</td>
<td>72.1/81.2</td>
<td>0.682</td>
<td>74.6</td>
<td>0.60/0.59</td>
</tr>
<tr>
<td>En-De</td>
<td>62.6</td>
<td>68.0</td>
<td>77.3</td>
<td>82.0</td>
<td>65.0</td>
<td>66.8</td>
<td>70.4/79.8</td>
<td>0.681</td>
<td>75.5</td>
<td>0.58/0.58</td>
</tr>
<tr>
<td>Combo</td>
<td>66.1</td>
<td>72.4</td>
<td>82.4</td>
<td>84.4</td>
<td>69.8</td>
<td>69.0</td>
<td>71.9/80.6</td>
<td>0.727</td>
<td>76.3</td>
<td>0.62/0.61</td>
</tr>
<tr>
<td>SkipGram</td>
<td>77.7</td>
<td>79.8</td>
<td>90.9</td>
<td>88.3</td>
<td>79.7</td>
<td>83.6</td>
<td>72.5/81.4</td>
<td>0.803</td>
<td>78.7</td>
<td>0.65/0.64</td>
</tr>
<tr>
<td>SkipGram&lt;sub&gt;(Conneau et al., 2017)&lt;/sub&gt;</td>
<td>77.7</td>
<td>79.8</td>
<td>90.9</td>
<td>88.3</td>
<td>79.7</td>
<td>83.6</td>
<td>72.5/81.4</td>
<td>0.803</td>
<td>78.7</td>
<td>0.65/0.64</td>
</tr>
<tr>
<td>NMT&lt;sub&gt;En-Fr&lt;/sub&gt;</td>
<td>64.7</td>
<td>70.1</td>
<td>84.8</td>
<td>81.5</td>
<td>-</td>
<td>82.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NMT&lt;sub&gt;(Conneau et al., 2017)&lt;/sub&gt;</td>
<td>64.7</td>
<td>70.1</td>
<td>84.8</td>
<td>81.5</td>
<td>-</td>
<td>82.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.42/0.43</td>
</tr>
</tbody>
</table>

**ACC**↑, **ACC/F1**↑, **CORR**↑
Conclusions

- Generative training
Conclusions

- Generative training
  - model learns from positive examples
Conclusions

- Generative training
  - model learns from positive examples
  - no need for context window
Conclusions

• Generative training
  • model learns form positive examples
  • no need for context window

• Translation data
Conclusions

- **Generative training**
  - model learns form positive examples
  - no need for context window

- **Translation data**
  - less ambiguous embeddings
Conclusions

- Generative training
  - model learns from positive examples
  - no need for context window

- Translation data
  - less ambiguous embeddings
  - model helps with semantic tasks e.g. **paraphrasing**
Future Work

- We modify alignment distribution
Future Work

- We modify alignment distribution
  - From IBM1 to **IBM2**
  - En-Fr 29.43 → **18.20** AER
Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings
Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings
  - Movie Reviews 66.10 $\rightarrow$ 70.55 ACC
Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    - En-Fr 29.43 → **18.20** AER

- We model word and sentence embeddings
  - Movie Reviews 66.10 → **70.55** ACC
  - Microsoft Paraphrase 71.90/80.6 → **72.93/81.27** ACC/F1

  
  **Sick R0.727 → 0.770 CORR**
Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings
  - Movie Reviews 66.10 $\rightarrow$ 70.55 ACC
  - Microsoft Paraphrase 71.90/80.6 $\rightarrow$ 72.93/81.27 ACC/F1
  - Sick R 0.727 $\rightarrow$ 0.770 CORR
Future Work

- We modify alignment distribution
  - From IBM1 to IBM2
    En-Fr 29.43 $\rightarrow$ 18.20 AER
- We model word and sentence embeddings
  - Movie Reviews 66.10 $\rightarrow$ 70.55 ACC
  - Microsoft Paraphrase 71.90/80.6 $\rightarrow$ 72.93/81.27 ACC/F1
  - Sick R 0.727 $\rightarrow$ 0.770 CORR
- We will expand the distributional context to multiple foreign languages at once
DGM4NLP research at UvA-SLPL

- Try pre-trained Europarl model on SentEval: 
  https://github.com/uva-slpl/embedalign/blob/master/notebooks/senteval_embedalign.ipynb
Try pre-trained Europarl model on SentEval: 
https://github.com/uva-slpl/Embedalign/blob/master/notebooks/senteval_embedding.ipynb

ACL-18 tutorial Variational Inference and Deep Generative Models: 
http://acl2018.org/tutorials/
## GAP

<table>
<thead>
<tr>
<th>Model</th>
<th>cos</th>
<th>KL</th>
<th>Training size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random</strong></td>
<td>30.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>SkipGram</strong></td>
<td>44.9</td>
<td>-</td>
<td>ukWaC-2B</td>
</tr>
<tr>
<td>(Melamud et al., 2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BSG</strong></td>
<td>-</td>
<td>46.1</td>
<td>ukWaC-2B</td>
</tr>
<tr>
<td>(Bražinskas et al., 2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>En</strong></td>
<td>21.31 ± 1.05</td>
<td>27.64 ± 0.40</td>
<td></td>
</tr>
<tr>
<td><strong>En-Fr</strong></td>
<td>42.19 ± 0.57</td>
<td>41.61 ± 0.55</td>
<td>Euro-2M</td>
</tr>
<tr>
<td><strong>En-De</strong></td>
<td>42.07 ± 0.47</td>
<td>40.93 ± 0.59</td>
<td></td>
</tr>
</tbody>
</table>