Training Neural Word Embeddings for Transfer Learning and Translation

Stephan Gouws

stephan@ml.sun.ac.za
Introduction

Audio Spectrogram

DENSE

Image pixels

DENSE

Word, context, or document vectors

SPARSE
Motivation

The NLP community has developed good features for several tasks, but finding

- **task-invariant** (POS tagging, NER, SRL);  **AND**
- **language-invariant** (English, Danish, Afrikaans,..)

features is non-trivial and time-consuming (been trying for 20+ years).

Learn word-level features which generalize across tasks **and** languages.
Word Embeddings

3D embedding space

dog, Berlin, Paris, January
Word Embeddings Capture Interesting, Universal Features

Male-Female

Verb tense

Country-Capital
Word Embedding Models

Models differ based on:

1. How they compute the **context** $h$
   - positional / convolutional / bag-of-words

2. How they map context to target $w_t = f(h)$
   - linear, bilinear, non-linear

3. How they measure $\text{loss}_R(w_t, f(h))$ and how they’re trained
   - **language models** (NLL, …)
   - **word embeddings**; negative sampling (CBOW/Skipgram), sampled rank loss (Collobert+Weston), squared-error (LSA)
Learning Word Embeddings: Pre-2008

Softmax classifier

Hidden layer

Projection layer

\[ \sum g(\text{embeddings}) \]

\[ \text{the cat sits on the mat} \]

\[ P_r(w_t) = \frac{\text{score}(w_t; \theta)}{\sum_{j=1}^{V} \text{score}(w_j; \theta)} \]

(Bengio et al., 2003)
Advances in Learning Word Embeddings

Not interested in language modelling (for which we need normalized probabilities), so we don’t need the expensive softmax. Can use much faster

- **hierarchical softmax** (Morin + Bengio, 2005),
- **sampled rank loss** (Collobert + Weston, 2008),
- **noise-contrastive estimation** (Mnih + Teh, 2012)
- **negative sampling** (Mikolov et al., 2013)
Learning Word Embeddings: Hierarchical Softmax

Hierarchical Softmax classifier

Hidden layer

Projection layer

\[
P(w_t|h) = \prod_{i \in L(w_t)} P(q_i|h)
\]

Significant savings since |L(w)| << V

(Morin + Bengio, 2005; Mikolov et al., 2013)
Train a non-probabilistic model to rank an observed word \( w_t \sim P_{\text{data}} \) some margin higher than \( k \) \((<<V)\) sampled noise words \( w_{\text{noise}} \sim P_{\text{noise}} \)

\[
\sum_{w_t,h \sim \text{data}} \sum_{j = 1}^{V} \max\{0, 1 - (s_\theta(w_t; h) - s_\theta(w_j; h))\}
\approx \sum_{w_t,h \sim \text{data}} \sum_{j \sim P(w)}^{k} \max\{0, 1 - (s_\theta(w_t; h) - s_\theta(w_j; h))\}
\]

(Collober + Weston, 2008)
Learning Word Embeddings: **Noise-contrastive Estimation**

Train a **probabilistic** model \( P(w|h) \) to be able to discriminate an observed nearby word \( w_t \sim P_{\text{data}} \) from sampled noise words \( w_{\text{noise}} \sim P_{\text{noise}} \)

\[
\sum_{w_t, h \sim P_{\text{data}}} P(w_t|h) + \sum_{j \sim P_{\text{noise}}} (1 - P(w_j|h))
\]

(Mnih + Teh, 2012)
Neural Word Embeddings
Why do “similar” words have similar embeddings?

Citizens of \{ France, Denmark, \ldots, Sweden \} protested today

All training objectives have the form:

$$\min_{R} \sum_{i} J(w^{(i)}_t, h^{(i)}_t)$$

$$= \min_{R} \sum_{i} \text{distance}(w^{(i)}_t, h^{(i)}_t)$$

I.e. for a fixed context, all distributionally similar words will get updated towards a common point.
Cross-lingual Word Embeddings

We want to learn an alignment between the two embedding spaces s.t. translation pairs are close.
Learning Cross-lingual Word Embeddings: Approaches

1. Align pre-trained embeddings (offline)

2. Jointly learn and align embeddings (online) using parallel-only data

3. Jointly learn and align embeddings (online) using monolingual and parallel data
Learning **Cross-lingual** Word Embeddings I

**Offline** methods: “Translation Matrix”

\[
\min_w \left\| \mathbf{R}^{en}\mathbf{W} - \mathbf{R}^{fr} \right\|^2
\]

Learn \( \mathbf{W} \) to transform the pre-trained English embeddings into a space where the distance between a word and its translation-pair is minimized

(Mikolov et al., 2013)
Learning **Cross-lingual** Word Embeddings I

**Offline** methods

Learn $W$ to transform the pre-trained English embeddings into a space where the distance between a word and its translation-pair is minimized.

$$\min_W \| \mathbf{R}^{en} W - \mathbf{R}^{fr} \|^2$$

Can also learn a separate $W$ for $en$ and $fr$ using **Multilingual CCA** (Faruqui et al, 2014)

(Mikolov et al., 2013)
Learning **Cross-lingual** Word Embeddings II

Parallel-only methods

\[
\min_R \left( \text{distance} \right)
\]

En parallel  \hspace{1cm} \text{distance} \hspace{1cm} \text{Fr parallel}

Bilingual Auto-encoders (Chandar et al., 2013)
BiCVM (Hermann et al., 2014)
Learning **Cross-lingual** Word Embeddings III

**Online** methods

\[
L_{en}(w|h) + \Omega_A(R) + L_{fr}(w|h)
\]

\(O(V_1)\) \(O(V_1V_2)\) \(O(V_2)\)

Cross-lingual regularization

en data \[\rightarrow\] \[\rightarrow\] fr data

(Klementiev et al., 2012)
Learning **Cross-lingual** Word Embeddings III

**Online** methods

\[ L_{en}(w|h) + \Omega_A(R) + L_{fr}(w|h) \]

- \( O(k) \)
- \( O(V_1V_2) \)
- \( O(k) \)

**Cross-lingual regularization**

- **en data**
- **fr data**

(Zhu et al., 2013)
## Multilingual distributed feature learning: Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix (\text{(Mikolov et al. 2013)})</td>
<td>• FAST&lt;br&gt;• Simple to implement</td>
<td>• Assumes a global, linear, <em>one-to-one mapping</em> exists between words in 2 languages. &lt;br&gt;• Requires accurate dictionaries</td>
</tr>
<tr>
<td>Multilingual CCA (\text{(Faruqui et al. 2014)})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Multilingual distributed feature learning: Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix ((Mikolov et al. 2013))</td>
<td>• FAST</td>
<td>• Assumes a global, linear, one-to-one mapping exists between words in 2 languages.</td>
</tr>
<tr>
<td></td>
<td>• Simple to implement</td>
<td>• Requires accurate dictionaries</td>
</tr>
<tr>
<td>Multilingual CCA ((Faruqui et al. 2014))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual Auto-encoders ((Chandar et al. 2014))</td>
<td>Simple to implement (?)</td>
<td>• Bag-of-words models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Learns more semantic than syntactic features</td>
</tr>
<tr>
<td>BiCVM ((Hermann et al., 2014))</td>
<td>Allows arbitrary differentiable sentence composition function</td>
<td>• Reduced training data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Big domain bias</td>
</tr>
</tbody>
</table>
# Multilingual distributed feature learning: Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix (&lt;cite&gt;Mikolov et al. 2013&lt;/cite&gt;)</td>
<td>• FAST</td>
<td>• Assumes a global, linear, one-to-one mapping exists between words in 2 languages.</td>
</tr>
<tr>
<td></td>
<td>• Simple to implement</td>
<td>• Requires accurate dictionaries</td>
</tr>
<tr>
<td>Multilingual CCA (&lt;cite&gt;Faruqui et al. 2014&lt;/cite&gt;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual Auto-encoders (&lt;cite&gt;Chandar et al. 2014&lt;/cite&gt;)</td>
<td>Simple to implement (?)</td>
<td>• Bag-of-words models</td>
</tr>
<tr>
<td>BiCVM (&lt;cite&gt;Hermann et al., 2014&lt;/cite&gt;)</td>
<td>Allows arbitrary differentiable sentence composition function</td>
<td>• Learns more semantic than syntactic features</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Reduced training data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Big domain bias</td>
</tr>
<tr>
<td>&lt;cite&gt;Klementiev et al., 2012&lt;/cite&gt;</td>
<td>Can learn fine-grained, cross-lingual syntactic/semantic features (depends on window-length)</td>
<td>• SLOW</td>
</tr>
<tr>
<td>&lt;cite&gt;Zhu et al., 2013&lt;/cite&gt;</td>
<td></td>
<td>• Requires word-alignments (GIZA++/Fastalign)</td>
</tr>
</tbody>
</table>
## Multilingual distributed feature learning: Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix (Mikolov et al. 2013)</td>
<td>• FAST • Simple to implement</td>
<td>• Assumes a global, linear, one-to-one mapping exists between words in 2 languages. • Requires accurate dictionaries</td>
</tr>
<tr>
<td>Multilingual CCA (Faruqui et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual Auto-encoders (Chandar et al. 2014)</td>
<td>Simple to implement (?)</td>
<td>• Bag-of-words models • Learns more semantic than syntactic features • Reduced training data • Big domain bias</td>
</tr>
<tr>
<td>BiCVM (Hermann et al., 2014)</td>
<td>Allows arbitrary differentiable sentence composition function</td>
<td></td>
</tr>
<tr>
<td>Klementiev et al., 2012</td>
<td>Can learn fine-grained, cross-lingual syntactic/semantic features (depends on window-length)</td>
<td>• SLOW • Requires word-alignments (GIZA++/Fastalign)</td>
</tr>
<tr>
<td>Zhu et al., 2013</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This work makes multilingual distributed feature learning more efficient for transfer learning and translation.
BilBOWA:
Fast Bilingual Bag-Of-Words Embeddings without Alignments
BilBOWA Architecture

Sampled L2 loss

∑

En BOW sentences

Fr BOW sentences

the cat sits on the mat

chat est assis sur le tapis

En monolingual

En-Fr parallel

Fr monolingual
The BilBOWA Cross-lingual Objective I

We want to learn similar embeddings for translation pairs. The **exact** cross-lingual objective to minimize is the weighted sum over all distances of word-pairs:

\[
\Omega_A(R) = \sum_{w_i \in V_{en}} \sum_{w_j \in V_{fr}} a_{i,j} \left\| R^e_{[i,:]} - R^f_{[j,:]} \right\|^2
\]

**Main contribution:** We approximate this by sampling parallel sentences.
Quick Aside: Monte Carlo integration

$$\mathbb{E}_{p(x)} f(x)$$

$$= \int_x p(x) f(x) \ (\text{Continuous})$$

$$= \sum_{x_i} Pr(x_i) f(x_i) \ (\text{Discrete})$$

$$\approx \frac{1}{N} \sum_{x \sim p(x)} f(x)$$
The BilBOWA Cross-lingual Objective II

\[ \Omega_A(R) = \sum_{i,j} a_{ij} \| R_{e[i,:]}^i - R_{f[j,:]}^j \|^2 \]

\[ = \mathbb{E}_{(i,j) \sim P(w^e, w^f)} \left[ \| R_{e[i,:]}^i - R_{f[j,:]}^j \|^2 \right] \]

\[ \approx \frac{1}{S} \sum_{s \in S} \frac{1}{mn} \sum_{(i,j) \in s} \| R_{e[i,:]}^i - R_{f[j,:]}^j \|^2 \]

Now we set \( S = 1 \) at each time step \( t \):

\[ \Omega_A^{(t)}(R) = \left\| \frac{1}{m} \sum_{i \in s_e} R_{e[i,:]}^i - \frac{1}{n} \sum_{j \in s_f} R_{f[j,:]}^j \right\|^2 \]

\( P(w^e, w^f) \) is the distribution of en-fr word alignments

Assume \( P(w^e, w^f) \) is uniform. \( m/n \) length of en/fr sentence.

mean en sentence-vector  mean fr sentence-vector
Implementation Details: Open-source

Implemented in C (part of word2vec, soon). Multi-threaded: One thread per language (monolingual), and one additional thread per language-pair (cross-lingual) (i.e. asynchronous SGD)

Runs at ~10-50K words per second on MBP.

Can process 500M words (monolingual) and 45M words (parallel, recycles) in about 2.5h on my MBP.
Cross-lingual subsampling for better results

At training step $t$, draw a random number $u \sim U[0,1]$. Then:

$$\Omega_A^{(t)}(R) = \| \frac{1}{m} \sum_{i \in s_e} \mathbb{1}_{u < f(w_i)} R_{[i,:]}^e - \frac{1}{n} \sum_{j \in s_f} \mathbb{1}_{u < f(w_j)} R_{[j,:]}^f \|^2$$

Want to estimate alignment statistics $P(e,f)$. Skewed at the sentence-level by (unconditional) unigram word frequencies.

Simple solution: 
Subsample frequent words to flatten the distribution!
Cross-lingual subsampling for better results
Qualitative Analysis: *en-fr* t-SNEs I
Qualitative Analysis: *en-fr* t-SNEs I
Qualitative Analysis: en-fr
Qualitative Analysis: en-fr
Qualitative Analysis: en-fr
Qualitative Analysis: en-fr
Experiments: *En-De* Cross-lingual Document Classification

Exact replication (obtained from the authors) of Klementiev et al.’s cross-language document classification setup:

**Goal:** Classify documents in target language using only labelled documents in source language.

**Data:** English-German RCV1 data (5K test, 100 - 10K training, 1K validation)

**4 Labels:**
- CCAT (Corporate/Industrial),
- ECAT (Economics),
- GCAT (Government/Social), and
- MCAT (Markets)
## Results: English-German Document Classification

<table>
<thead>
<tr>
<th></th>
<th>en2de</th>
<th>de2en</th>
<th>Training Size</th>
<th>Training Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority class</td>
<td>46.8</td>
<td>46.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Glossed</td>
<td>65.1</td>
<td>68.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MT</td>
<td>68.1</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Klementiev et al.</td>
<td>77.6</td>
<td>71.1</td>
<td>50M words</td>
<td>14,400 (10 days)</td>
</tr>
</tbody>
</table>
## Results: English-German Document Classification

<table>
<thead>
<tr>
<th>Method</th>
<th>en2de</th>
<th>de2en</th>
<th>Training Size</th>
<th>Training Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority class</td>
<td>46.8</td>
<td>46.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Glossed</td>
<td>65.1</td>
<td>68.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>MT</strong></td>
<td>68.1</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Klementiev et al.</td>
<td>77.6</td>
<td>71.1</td>
<td>50M words</td>
<td>14,400 (10 days)</td>
</tr>
<tr>
<td>Bilingual Autoencoders</td>
<td><strong>91.8</strong></td>
<td>72.8</td>
<td>50M words</td>
<td>4,800 (3.5 days)</td>
</tr>
<tr>
<td>BiCVM</td>
<td>83.7</td>
<td>71.4</td>
<td>50M words</td>
<td>15</td>
</tr>
<tr>
<td><strong>BilBOWA (this work)</strong></td>
<td>86.5</td>
<td><strong>75</strong></td>
<td>50M words</td>
<td><strong>6</strong></td>
</tr>
</tbody>
</table>
Experiments: WMT11 *English-Spanish* Translation

- Trained BilBOWA model on En-Es Wikipedia/Europarl data.
  - Vocabulary = 200K
  - Embedding dimension = 40,
  - Crosslingual $\lambda$-weight in \{0.1, 1.0, 10.0, 100.0\}

- Exact replica of *(Mikolov, Le, Sutskever, 2013)*:
  - Evaluated on WMT11 lexicon, translated using GTranslate
  - Top 5K-6K words as test set
## Experiments: WMT11 *English→Spanish* Translation

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimension</th>
<th>Prec@1</th>
<th>Prec@5</th>
<th>% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit distance</td>
<td>-</td>
<td>13</td>
<td>24</td>
<td>92.9</td>
</tr>
<tr>
<td>Word co-occurrence</td>
<td>-</td>
<td>19</td>
<td>30</td>
<td>92.9</td>
</tr>
<tr>
<td>Translation Matrix</td>
<td>300-800</td>
<td>33</td>
<td>51</td>
<td>92.9</td>
</tr>
<tr>
<td>BilBOWA (This work)</td>
<td>40</td>
<td><strong>39 (+6%)</strong></td>
<td>51</td>
<td>92.7</td>
</tr>
</tbody>
</table>
## Experiments: WMT11 *Spanish→English* Translation

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimension</th>
<th>Prec@1</th>
<th>Prec@5</th>
<th>% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Edit distance</em></td>
<td>-</td>
<td>18</td>
<td>27</td>
<td>92.9</td>
</tr>
<tr>
<td><em>Word co-occurrence</em></td>
<td>-</td>
<td>20</td>
<td>30</td>
<td>92.9</td>
</tr>
<tr>
<td><strong>Translation Matrix</strong></td>
<td>300-800</td>
<td>35</td>
<td>52</td>
<td>92.9</td>
</tr>
<tr>
<td><strong>BilBOWA (This work)</strong></td>
<td>40</td>
<td>44 (+11%)</td>
<td>55 (+3%)</td>
<td>92.7</td>
</tr>
</tbody>
</table>
Barista:
Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives
Motivation

Word embedding models learn to predict targets from contexts by clustering similar words into soft (distributional) equivalence classes.

For some tasks, we may easily obtain the desired equivalence classes:

- **POS**: Wiktionary
- **Super-sense (SuS) tagging**: WordNet
- **Translation**: Google Translate / dictionaries

**Barista** embeds additional task-specific semantic information by corrupting the training data according to known equivalences $C(w)$. 
Barista Algorithm

1. Shuffle $D_{en}$ & $D_{fr}$ -> $D$
2. For $w$ in $D$:
3. If $w$ in $C$ then $w'$ ~ $C(w)$ else $w' = w$
4. $D' += w'$
4. Train off-the-shelf embedding model on $D'$

For example: "we build the house":

1. we build la voiture / they run la house (POS)
2. we construire the maison / nous build la house (Translations)
Qualitative (POS)
Qualitative (Translations)

Prepositions
Cross-lingual POS tagging

<table>
<thead>
<tr>
<th>Language</th>
<th>Baseline</th>
<th>Random</th>
<th>Klmtv</th>
<th>POS-50</th>
<th>POS-300</th>
<th>Tr-50</th>
<th>Tr-300</th>
<th>DP</th>
<th>B-K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>80.6</td>
<td>81.8</td>
<td>79.8</td>
<td>82.4</td>
<td>81</td>
<td>81.6</td>
<td>82.6</td>
<td>84.2</td>
<td>80.2</td>
</tr>
<tr>
<td>German</td>
<td>80.4</td>
<td>82.7</td>
<td>82.8</td>
<td>81.8</td>
<td>84.1</td>
<td>82.6</td>
<td>84.8</td>
<td>82.8</td>
<td>81.3</td>
</tr>
<tr>
<td>Danish</td>
<td>63</td>
<td>68.9</td>
<td>-</td>
<td>68.9</td>
<td>72.4</td>
<td>71.8</td>
<td>78.4</td>
<td>83.2</td>
<td>69.1</td>
</tr>
<tr>
<td>Swedish</td>
<td>71.6</td>
<td>73.7</td>
<td>-</td>
<td>75</td>
<td>76</td>
<td>75.4</td>
<td>77.5</td>
<td>80.5</td>
<td>70.1</td>
</tr>
<tr>
<td>Italian</td>
<td>80.1</td>
<td>81.3</td>
<td>-</td>
<td>82.1</td>
<td>80.9</td>
<td>82.1</td>
<td>80.7</td>
<td>86.8</td>
<td>68.1</td>
</tr>
<tr>
<td>Dutch</td>
<td>74.5</td>
<td>77.2</td>
<td>-</td>
<td>78.3</td>
<td>77.4</td>
<td>78.7</td>
<td>80.3</td>
<td>79.5</td>
<td>65.1</td>
</tr>
<tr>
<td>Portuguese</td>
<td>76.9</td>
<td>78.1</td>
<td>-</td>
<td>77.3</td>
<td>76.1</td>
<td>80.6</td>
<td>80.5</td>
<td>87.9</td>
<td>78.4</td>
</tr>
<tr>
<td>Avg</td>
<td>75.3</td>
<td>77.7</td>
<td>-</td>
<td>78</td>
<td>78.3</td>
<td>79</td>
<td>80.7</td>
<td>83.6</td>
<td>73.2</td>
</tr>
</tbody>
</table>
Cross-lingual SuperSense (SuS) tagging

<table>
<thead>
<tr>
<th></th>
<th>Baselines</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFS</td>
<td>bl</td>
<td>Tr-300</td>
<td>WN-300</td>
</tr>
<tr>
<td>blogs</td>
<td>49.1</td>
<td>46.7</td>
<td>50.5</td>
<td>61.9</td>
</tr>
<tr>
<td>forum</td>
<td>44.5</td>
<td>41.2</td>
<td>45.0</td>
<td>53.9</td>
</tr>
<tr>
<td>magazine</td>
<td>46.5</td>
<td>45.2</td>
<td>50.4</td>
<td>51.5</td>
</tr>
<tr>
<td>newswire</td>
<td>48.4</td>
<td>45.4</td>
<td>52.7</td>
<td>60.9</td>
</tr>
<tr>
<td>reviews</td>
<td>48.4</td>
<td>44.9</td>
<td>50.4</td>
<td>55.8</td>
</tr>
<tr>
<td>speech</td>
<td>51.1</td>
<td>48.4</td>
<td>51.5</td>
<td>58.4</td>
</tr>
<tr>
<td>Avg</td>
<td>48.1</td>
<td>45.4</td>
<td>50.5</td>
<td>58.3</td>
</tr>
</tbody>
</table>
Conclusion

I presented **BilBOWA**, an efficient, bilingual word embedding model with an open-source C implementation (part of word2vec, soon) and **Barista**, a simple technique for embedding additional task-specific cross-lingual information.

*Qualitative* experiments on En-Fr & En-De show that the learned embeddings capture fine-grained cross-lingual linguistic regularities.

*Quantitative* results on Es, De, Da, Sv, It, Nl, Pt for:
- semantic transfer (document classification, xling SuS-tagging)
- lexical transfer (word-level translation, xling POS-tagging)
Thanks!

Questions / Comments?