One-Shot Learning: Language Acquisition for Machine

SS16 Computational Linguistics for Low-Resource Languages

Mayumi Ohta
July 6, 2016

Institute for Computational Linguistics
Heidelberg University
Table of contents

1. Introduction

2. Language Acquisition for Human

3. Language Acquisition for Machine
   - Zero-shot learning
   - One-shot learning
   - Application to Low-Resource Languages

4. Summary
Introduction
My Interest

**Our Focus:**
How can CL/NLP support documenting low-resource languages? (collection, transcription, translation, annotation, etc.)

**Implicit Assumption:**
Only human can produce primary language resources.
≠ Primary language resources must be produced by human only.
Our Focus:
How can CL/NLP support documenting low-resource languages? (collection, transcription, translation, annotation, etc.)

Implicit Assumption:
Only human can produce primary language resources. ≠ Primary language resources must be produced by human only.

What if a machine can learn a language?

... of course, it is still a fantasy, but ...
My Interest

Our Focus:
How can CL/NLP support documenting low-resource languages? (collection, transcription, translation, annotation, etc.)

Implicit Assumption:
Only human can produce primary language resources.
≠ Primary language resources must be produced by human only.

What if a machine can learn a language?

... of course, it is still a fantasy, but ...

Big breakthrough: Deep Learning (2010~)
→ no need for feature design
Example 1. Neural Network Language Model [Mikolov et al. 2011]

... Princess Mary was easier, fed in had oftened him. Pierre asking his soul came to the packs and drove up his father-in-law women.

generated by LSTM-RNN LM trained with Leo Tolstoy's "War and Peace"

Source: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

"Colorless green ideas sleep furiously." by Noam Chomsky
Example 1. Neural Network Language Model [Mikolov et al. 2011]

... Princess Mary was easier, fed in had oftened him. Pierre asking his soul came to the packs and drove up his father-in-law women.

generated by LSTM-RNN LM trained with Leo Tolstoy's "War and Peace"
Source: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

"Colorless green ideas sleep furiously." by Noam Chomsky

It looks as if they know "syntax". (3rd person singular, tense, etc.)
Example 2. word2vec [Mikolov et al. 2013a]

KING − MAN + WOMAN = QUEEN

Source: https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html
Example 2. **word2vec** [Mikolov et al. 2013a]

KING − MAN + WOMAN = QUEEN

Source: https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html

Intuitive characteristics of "**semantics**" are (somehow!) embedded in vector space.
Language Acquisition for Human
Vocabulary explosion

... what happened?
Helen Keller (1880 – 1968)

"w-a-t-e-r"

Image source: http://en.wikipedia.org/wiki/Helen_Keller
Language acquisition

... to simplify the problem: "Everything has a name" model

Language acquisition → Vocabulary acquisition
   → Mapping between concepts and words
     (main focus: Nouns)

↔ "water"

Image source: https://de.wikipedia.org/wiki/Wasser
Language acquisition

... to simplify the problem: "Everything has a name" model

Language acquisition $\rightarrow$ Vocabulary acquisition
$\rightarrow$ Mapping between concepts and words
(main focus: Nouns)

$\leftrightarrow$ "water"

Image source: https://de.wikipedia.org/wiki/Wasser
... to simplify the problem: "Everything has a name" model

Language acquisition → Vocabulary acquisition
    → Mapping between concepts and words
    (main focus: Nouns)

↔ "water"

Image source: https://de.wikipedia.org/wiki/Wasser
Machine vs. Human

**Machine** learns:

1. relationship between words (i.e. word2vec)
2. from manually-defined features (i.e. SVM, CRF, ...)
3. from large quantity of training examples
4. iteratively (i.e. SGD)

**Human kids** learn:

1. relationship between words and concepts
2. from raw data
3. from just one or a few examples
4. immediately (not necessarily need repetition)
Machine vs. Human

**Machine** learns:

1. relationship between words (i.e. **word2vec**)
2. from manually-defined features (i.e. **SVM**, **CRF**, ...)
3. from large quantity of training examples
4. iteratively (i.e. **SGD**)

**Human kids** learn:

1. relationship between words and concepts
2. from raw data
3. from **just one or a few examples**
4. **immediately** (not necessarily need repetition)

→ "fast mapping"
Language Acquisition for Machine
Two directions

Machine learning approach inspired from "fast mapping"?
Two directions

Machine learning approach inspired from "fast mapping"?

Zero-shot learning: unknown concept $\rightarrow$ known word
One-shot learning: unknown word $\rightarrow$ known concept

Image source: https://en.wikipedia.org/wiki/Rabbit
Zero-shot learning
Zero-shot learning: Overview

Example: Image Classification Task

Traditional supervised setting

- train a model with labeled image data

Image source: https://en.wikipedia.org/
Zero-shot learning: Overview

Example: Image Classification Task

Traditional supervised setting

- train a model with labeled image data
- classify a known label for an unseen image

Image source: https://en.wikipedia.org/
Zero-shot learning: Overview

Example: Image Classification Task

Zero-shot learning

- train a model with labeled image data

Image source: https://en.wikipedia.org/
Zero-shot learning: Overview

Example: Image Classification Task

![dog](image1)
![rabbit](image2)
![cat](image3)
![dog](image4)
![cat](image5)
![cat](image6)

Zero-shot learning

- train a model with labeled image data
- classify a **known but unseen** label for an unseen image
  → no training examples for the classes of test examples

Image source: https://en.wikipedia.org/
Zero-shot learning: Core idea

Core idea:

image features

Socher et al. 2013, modified
Zero-shot learning: Core idea

Core idea:

Socher et al. 2013, modified
Core idea:

project image features onto word embeddings

Socher et al. 2013, modified
Zero-shot learning: Core idea

Core idea:
project image features onto word embeddings

Socher et al. 2013, modified
Zero-shot learning: Formulation [Socher et al. 2013]

**Method:** Multi-layer Neural Network (Back Propagation)

**Objective function:**

$$J(\Theta) = \sum_{y \in Y} \sum_{x^{(i)} \in X} \left\| \vec{\omega}_y - \theta^{(2)} f \left( \theta^{(1)} x^{(i)} \right) \right\|^2$$

where $f(\cdot)$: non-linear activation function such as $\text{tanh}(\cdot)$  
$\theta^{(1)}$: weights for the first layer  
$\theta^{(2)}$: weights for the second layer

$\rightarrow$ update weights such that image features closes to the word embedding
One-shot learning
One-shot learning: Overview

Example: Automatic Speech Synthesis

Traditional supervised setting

- train a model with labeled audio data
  (pipelined: segment $\rightarrow$ cluster $\rightarrow$ learn transition prob.)
- generate an audio for a given concept
One-shot learning: Overview

Example: Automatic Speech Synthesis

One-shot learning

- jointly train a model with labeled audio data
- generate an audio for a given concept heard before just once
One-shot learning: Formulation [Lake et al. 2014]

Method: Hierarchical Bayesian (parametric or non-parametric)

arg max \( \Pr(\mathbf{X}_{test} | \mathbf{X}_{train}) \) = arg max \( \frac{\Pr(\mathbf{X}_{test} | \mathbf{X}_{train}) \Pr(\mathbf{X}_{train} | \mathbf{X}_{test})}{\Pr(\mathbf{X}_{train})} \) \hspace{1cm} (1)

\[
\Pr(\mathbf{X}_{test} | \mathbf{X}_{train}) \approx \sum_{i=1}^{L} \Pr(\mathbf{X}_{test} | \mathbf{Z}_{train}^{(i)}) \frac{\Pr(\mathbf{X}_{train} | \mathbf{Z}_{train}^{(i)}) \Pr(\mathbf{Z}_{train}^{(i)})}{\sum_{j=1}^{L} \Pr(\mathbf{X}_{train} | \mathbf{Z}_{train}^{(j)}) \Pr(\mathbf{Z}_{train}^{(j)})} \]

\hspace{1cm} (2)

\[
\Pr(\mathbf{X}_{train}) \approx \sum_{i=1}^{L} \Pr(\mathbf{X}_{train} | \mathbf{Z}_{train}^{(i)}) \Pr(\mathbf{Z}_{train}^{(i)}) \]

\hspace{1cm} (3)

where \( \mathbf{X}_{train}, \mathbf{X}_{test} \): sequences of features
\( \mathbf{Z}_{train} \): acoustic segments (units)
\( L \): length (number of units)
One-shot learning: Formulation [Lake et al. 2014]

Method: **Hierarchical Bayesian** (parametric or non-parametric)

\[
\text{arg max } \Pr (X_{test} | X_{train}) = \text{arg max } \frac{\Pr (X_{test} | X_{train}) \Pr (X_{train} | X_{test})}{\Pr (X_{train})} \tag{1}
\]

\[
\Pr (X_{test} | X_{train}) \approx \sum_{i=1}^{L} \Pr (X_{test} | Z_{train}^{(i)}) \frac{\Pr (X_{train} | Z_{train}^{(i)}) \Pr (Z_{train}^{(i)})}{\sum_{j=1}^{L} \Pr (X_{train} | Z_{train}^{(j)}) \Pr (Z_{train}^{(j)})} \tag{2}
\]

\[
\Pr (X_{train}) \approx \sum_{i=1}^{L} \Pr (X_{train} | Z_{train}^{(i)}) \Pr (Z_{train}^{(i)}) \tag{3}
\]

where \(X_{train}, X_{test}\): sequences of features
\(Z_{train}\): acoustic segments (units)
\(L\): length (number of units)
One-shot learning: Formulation [Lake et al. 2014]

**Method: Hierarchical Bayesian** (parametric or non-parametric)

\[
\text{arg max } \Pr(X_{\text{test}}|X_{\text{train}}) = \text{arg max } \frac{\Pr(X_{\text{test}}|X_{\text{train}})}{\Pr(X_{\text{train}})} \tag{1}
\]

\[
\Pr(X_{\text{train}}|X_{\text{test}}) \approx \sum_{i=1}^{L} \Pr(X_{\text{train}}|Z_{\text{test}}^{(i)}) \frac{\Pr(X_{\text{test}}|Z_{\text{test}}^{(i)}) \Pr(Z_{\text{test}}^{(i)})}{\sum_{j=1}^{L} \Pr(X_{\text{test}}|Z_{\text{test}}^{(j)}) \Pr(Z_{\text{test}}^{(j)})} \tag{2}
\]

\[
\Pr(X_{\text{train}}) \approx \sum_{i=1}^{L} \Pr(X_{\text{train}}|Z_{\text{train}}^{(i)}) \Pr(Z_{\text{train}}^{(i)}) \tag{3}
\]

where \(X_{\text{train}}, X_{\text{test}}\): sequences of features

\(Z_{\text{train}}\): acoustic segments (units)

\(L\): length (number of units)
One-shot learning: Formulation [Lake et al. 2014]

Method: Hierarchical Bayesian (parametric or non-parametric)

arg max  \( \Pr(X_{test}|X_{train}) \) = arg max  \( \frac{\Pr(X_{test}|X_{train}) \Pr(X_{train}|X_{test})}{\Pr(X_{train})} \)  (1)

\[ \Pr(X_{train}|X_{test}) \approx \sum_{i=1}^{L} \Pr(X_{train}|Z_{test}^{(i)}) \]  (2)

\[ \Pr(X_{train}) \approx \sum_{i=1}^{L} \Pr(X_{train}|Z_{train}^{(i)}) \Pr(Z_{train}^{(i)}) \]  (3)

where  \( X_{train}, X_{test} \): sequences of features
\( Z_{train} \): acoustic segments (units)
\( L \): length (number of units)
One-shot learning: Experiments [Lake et al. 2014]
One-shot learning: Experiments [Lake et al. 2014]

**Concept** \( \rightarrow \) **Word**

- Model 1: Human (adult English native speakers)
- Model 2: Trained with English text
- Model 3: Trained with Japanese text
- Model 4 (baseline): No-segmentation

<table>
<thead>
<tr>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>76.8%</td>
</tr>
<tr>
<td>English</td>
<td>34.1%</td>
</tr>
<tr>
<td>Japanese</td>
<td>57.6%</td>
</tr>
</tbody>
</table>
| Baseline| 17.0%   | —
One-shot learning: Experiments [Lake et al. 2014]

The concept is: has long ears, is hopping.

Task: generate a spoken Japanese word heard before just once

Model 1: Human (adult English native speakers)
Model 2: Trained with English text
Model 3: Trained with Japanese text
Model 4 (baseline): No-segmentation

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>76.8%</td>
<td>80.8%</td>
</tr>
<tr>
<td>English</td>
<td>34.1%</td>
<td>27.3%</td>
</tr>
<tr>
<td>Japanese</td>
<td>57.6%</td>
<td>60.0%</td>
</tr>
<tr>
<td>Baseline</td>
<td>17.0%</td>
<td>—</td>
</tr>
</tbody>
</table>
One-shot learning: Experiments [Lake et al. 2014]

**Task:** generate a spoken Japanese word heard before just once

- **model 1:** human (adult English native speakers)
- **model 2:** trained with English text
- **model 3:** trained with Japanese text
- **model 4** (baseline): no-segmentation
One-shot learning: Experiments [Lake et al. 2014]

**Task:** generate a spoken **Japanese** word heard before just once

<table>
<thead>
<tr>
<th>Model</th>
<th>round 1</th>
<th>round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>76.8%</td>
<td>80.8%</td>
</tr>
<tr>
<td>English</td>
<td>34.1%</td>
<td>27.3%</td>
</tr>
<tr>
<td>Japanese</td>
<td>57.6%</td>
<td>60.0%</td>
</tr>
<tr>
<td>Baseline</td>
<td>17.0%</td>
<td>—</td>
</tr>
</tbody>
</table>

---

**model 1:** human

(adolescent English native speakers)

**model 2:** trained with English text

**model 3:** trained with Japanese text

**model 4** (baseline): no-segmentation
Application to Low-Resource Languages
Potential Application to Low-Resource Languages

Zero-shot/One-shot learning as **Transfer learning**

- **modal-transfer:**
  
  image $\rightarrow$ text, text $\rightarrow$ audio, etc.

- **language-transfer:**
  
  high-resource language $\rightarrow$ low-resource language
Potential Application to Low-Resource Languages

Zero-shot/One-shot learning as Transfer learning

- **modal-transfer:**
  image $\rightarrow$ text, text $\rightarrow$ audio, etc.

- **language-transfer:**
  high-resource language $\rightarrow$ low-resource language

"learning-to-learn"
Prediction (i.e. Generative model)

\[ y = \arg \max \Pr(x, y; \theta) = \arg \max \prod_i \theta^{f(x^{(i)}, y^{(i)})} \]

Training (i.e. SGD)

\[ \theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}(\theta) \]

where

- \( x \): input (i.e. feature vector)
- \( y \): output (i.e. class label)
- \( \theta \): parameter (i.e. weights vector)
- \( f(\cdot) \): score function (i.e. probability)
- \( \mathcal{L}(\cdot) \): Loss function (i.e. squared-error)

**Training Procedure**

1. initialize \( \theta \) randomly
2. update \( \theta \)
Transfer learning: Formulation

**Prediction** (i.e. Generative model)

\[ y = \arg \max Pr(x, y; \theta) = \arg \max \prod \theta^f(x^{(i)}, y^{(i)}) \]

**Training** (i.e. SGD)

\[ \theta \leftarrow \theta - \eta \nabla \theta L(\theta) \]

where

- \( x \): input (i.e. feature vector)
- \( y \): output (i.e. class label)
- \( \theta \): parameter (i.e. weights vector)
- \( f(\cdot) \): score function (i.e. probability)
- \( L(\cdot) \): Loss function (i.e. squared-error)

**Training Procedure**

1. initialize \( \theta \) with pre-trained model
2. update \( \theta \)
Neural Machine Translation [Zoph et. al. 2016]

Attention-based Neural Machine Translation

<table>
<thead>
<tr>
<th>retrain or fix?</th>
<th>BLEU</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>0.0</td>
<td>112.6</td>
</tr>
<tr>
<td>+</td>
<td>7.7</td>
<td>24.7</td>
</tr>
<tr>
<td>+</td>
<td>11.8</td>
<td>17.0</td>
</tr>
<tr>
<td>+</td>
<td>14.2</td>
<td>14.5</td>
</tr>
<tr>
<td>+</td>
<td>15.0</td>
<td>13.9</td>
</tr>
<tr>
<td>+</td>
<td>14.7</td>
<td>13.8</td>
</tr>
<tr>
<td>+</td>
<td>13.7</td>
<td>14.4</td>
</tr>
</tbody>
</table>

BLEU: metrics
PPL: perplexity (loss)
Neural Machine Translation [Zoph et. al. 2016]

Attention-based Neural Machine Translation

(1) Pre-Train with high-resource language pair

Source RNN → Source embedding → Source word

French: la

Target RNN → Target output embedding → Target word

English: the

Target RNN → Target output embedding → Target word

English: house

BLEU: metrics
PPL: perplexity (loss)

retrain or fix?

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>0.0</td>
<td>112.6</td>
</tr>
<tr>
<td>+</td>
<td>7.7</td>
<td>24.7</td>
</tr>
<tr>
<td>+</td>
<td>11.8</td>
<td>17.0</td>
</tr>
<tr>
<td>+</td>
<td>14.2</td>
<td>14.5</td>
</tr>
<tr>
<td>+</td>
<td>15.0</td>
<td>13.9</td>
</tr>
<tr>
<td>+</td>
<td>14.7</td>
<td>13.8</td>
</tr>
<tr>
<td>+</td>
<td>13.7</td>
<td>14.4</td>
</tr>
</tbody>
</table>
Neural Machine Translation [Zoph et. al. 2016]

Attention-based Neural Machine Translation

(2) Re-Train with low-resource language pair

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>0.0</td>
<td>112.6</td>
</tr>
<tr>
<td>+</td>
<td>7.7</td>
<td>24.7</td>
</tr>
<tr>
<td>+</td>
<td>11.8</td>
<td>17.0</td>
</tr>
<tr>
<td>+</td>
<td>14.2</td>
<td>14.5</td>
</tr>
<tr>
<td>+</td>
<td>15.0</td>
<td>13.9</td>
</tr>
<tr>
<td>+</td>
<td>14.7</td>
<td>13.8</td>
</tr>
<tr>
<td>+</td>
<td>13.7</td>
<td>14.4</td>
</tr>
</tbody>
</table>

BLEU: metrics
PPL: perplexity (loss)
Experiment 1

Does Transfer model improve Non-transfer model?

**High-resource language pair**: French-English  
(\(#\text{Train}: 300m, \text{BLEU}: 26, \#\text{Epoch}: 5\) )

<table>
<thead>
<tr>
<th>Low-resource language pair</th>
<th># Train tokens</th>
<th># Test tokens</th>
<th>SBMT BLEU</th>
<th>NMT BLEU</th>
<th>Xfer BLEU</th>
<th>(\Delta_{\text{NMT BLEU}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausa-English</td>
<td>1.0m</td>
<td>11.3K</td>
<td>23.7</td>
<td>16.8</td>
<td><strong>21.3</strong></td>
<td>+4.5</td>
</tr>
<tr>
<td>Turkish-English</td>
<td>1.4m</td>
<td>11.6K</td>
<td>20.4</td>
<td>11.4</td>
<td>17.0</td>
<td>+5.6</td>
</tr>
<tr>
<td>Uzbek-English</td>
<td>1.8m</td>
<td>11.5K</td>
<td>17.9</td>
<td>10.7</td>
<td>14.4</td>
<td>+3.7</td>
</tr>
<tr>
<td>Urdu-English</td>
<td>0.2m</td>
<td>11.4K</td>
<td>17.9</td>
<td>5.2</td>
<td>13.8</td>
<td><strong>+8.6</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>+5.6</strong></td>
</tr>
</tbody>
</table>
NMT: Experiments (I) [Zoph et. al. 2016]

Experiment 1

Does Transfer model improve Non-transfer model? → Yes!

High-resource language pair: French-English
(#Train: 300m, BLEU: 26, #Epoch: 5)

<table>
<thead>
<tr>
<th>Low-resource language pair</th>
<th>#Train tokens</th>
<th>#Test tokens</th>
<th>SBMT BLEU</th>
<th>NMT BLEU</th>
<th>Xfer BLEU</th>
<th>$\Delta_{\text{NMT BLEU}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausa-English</td>
<td>1.0m</td>
<td>11.3K</td>
<td>23.7</td>
<td>16.8</td>
<td>21.3</td>
<td>+4.5</td>
</tr>
<tr>
<td>Turkish-English</td>
<td>1.4m</td>
<td>11.6K</td>
<td>20.4</td>
<td>11.4</td>
<td>17.0</td>
<td>+5.6</td>
</tr>
<tr>
<td>Uzbek-English</td>
<td>1.8m</td>
<td>11.5K</td>
<td>17.9</td>
<td>10.7</td>
<td>14.4</td>
<td>+3.7</td>
</tr>
<tr>
<td>Urdu-English</td>
<td>0.2m</td>
<td>11.4K</td>
<td>17.9</td>
<td>5.2</td>
<td>13.8</td>
<td>+8.6</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>+5.6</strong></td>
</tr>
</tbody>
</table>
NMT: Experiments (I) [Zoph et. al. 2016]

Experiment 1

Does Transfer model improve Non-transfer model? → Yes!

**High-resource language pair:** French-English

(#Train: 300m, BLEU: 26, #Epoch: 5)

Uzbek-English learning curves
NMT: Experiments (II) [Zoph et. al. 2016]

**Experiment 2**

Which high-resource language pair is better?

**Low-resource language pair:** Spanish-English

<table>
<thead>
<tr>
<th>High-resource</th>
<th>#Train</th>
<th>#Test</th>
<th>BLEU</th>
<th>$\Delta_{\text{BLEU}}$</th>
<th>PPL</th>
<th>$\Delta_{\text{PPL}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none (baseline)</td>
<td>2.5m</td>
<td>59k</td>
<td>16.4</td>
<td>—</td>
<td>15.9</td>
<td>—</td>
</tr>
<tr>
<td>French-English</td>
<td>53m</td>
<td>59k</td>
<td>31.0</td>
<td>+14.6</td>
<td>5.8</td>
<td>−10.1</td>
</tr>
<tr>
<td>German-English</td>
<td>53m</td>
<td>59k</td>
<td>29.8</td>
<td>+13.4</td>
<td>6.2</td>
<td>−9.7</td>
</tr>
</tbody>
</table>
Experiment 2

Which high-resource language pair is better? → **similar language pair**

**Low-resource language pair**: Spanish-English

<table>
<thead>
<tr>
<th>High-resource</th>
<th># Train</th>
<th># Test</th>
<th>BLEU</th>
<th>Δ_{BLEU}</th>
<th>PPL</th>
<th>Δ_{PPL}</th>
</tr>
</thead>
<tbody>
<tr>
<td>none (baseline)</td>
<td>2.5m</td>
<td>59k</td>
<td>16.4</td>
<td>—</td>
<td>15.9</td>
<td>—</td>
</tr>
<tr>
<td>French-English</td>
<td>53m</td>
<td>59k</td>
<td>31.0</td>
<td>+14.6</td>
<td>5.8</td>
<td>-10.1</td>
</tr>
<tr>
<td>German-English</td>
<td>53m</td>
<td>59k</td>
<td>29.8</td>
<td>+13.4</td>
<td>6.2</td>
<td>-9.7</td>
</tr>
</tbody>
</table>
Experiment 3

Does a look-up dictionary improve the result?

**Look-up dictionary** for source word embeddings

![Graph showing word embeddings in English and Spanish](image)

Mikolov et al. 2013b
Experiment 3

Does a look-up dictionary improve the result?

**Look-up dictionary** for source word embeddings

Uzbek-English learning curves
Experiment 3

Does a look-up dictionary improve the result? → **No!**

**Look-up dictionary** for source word embeddings

Uzbek-English learning curves
Summary
## Summary

### Take-home message

- learn from just **one or a few training examples**
  - with help of other well-known, related resources

###零-shot学习: unknown concept $\rightarrow$ known word
###一次-shot学习: unknown word $\rightarrow$ known concept

- **inspired from human kids’ "fast mapping"**
- Zero-shot learning: unknown concept $\rightarrow$ known word
- One-shot learning: unknown word $\rightarrow$ known concept
- Methods: generative models (Neural Network, Bayesian, etc.)

### Application to low-resource languages

- transfer learning: "learning-to-learn"
Summary

Take-home message

learn from just one or a few training examples
with help of other well-known, related resources

• inspired from human kids’ "fast mapping"
• **Zero-shot learning**: unknown concept → known word
• **One-shot learning**: unknown word → known concept
• Methods: generative models (**Neural Network**, **Bayesian**, etc.)

Application to low-resource languages

• transfer learning: "learning-to-learn"
  → reuse pre-trained parameters in initialization!
Future Direction (my main interest)

Promising path:

- Avoid annotation, avoid pipeline
- Use unlabeled data and a few labeled data
  - semi-supervised learning
  - weekly supervised learning
  - reinforcement learning
  - one-shot learning
Future Direction (my main interest)

Promising path:

- Avoid annotation, avoid pipeline
- Use unlabeled data and a few labeled data
  - semi-supervised learning
  - weekly supervised learning
  - reinforcement learning
  - one-shot learning

Open questions:

- How to calculate more efficiently?
Future Direction (my main interest)

Promising path:

- Avoid annotation, avoid pipeline
- Use unlabeled data and a few labeled data
  - semi-supervised learning
  - weekly supervised learning
  - reinforcement learning
  - one-shot learning

Open questions:

- How to calculate more efficiently?
- How to model what you want to say?
Future Direction (my main interest)

Promising path:

- Avoid annotation, avoid pipeline
- Use unlabeled data and a few labeled data
  - semi-supervised learning
  - weekly supervised learning
  - reinforcement learning
  - one-shot learning

Open questions:

- How to calculate more efficiently?
- How to model what you want to say?
- How to model the incentive to get to know unknown information?
"Better than a thousand days of diligent study is one day with a great teacher"
(Japanese proverb)

Thanks!
Lake, Brenden M. et al.  
**One-shot Learning of Generative Speech Concepts**  

Socher, Richard et al.  
**Zero-Shot Learning Through Cross-Modal Transfer**  

Zoph, Barret et al.  
**Transfer Learning for Low-Resource Neural Machine Translation**  
Mikolov, Tomas et al.
Distributed Representations of Words and Phrases and their Compositionality
Advances in neural information processing systems, pp. 3111-3119, 2013

Mikolov, Tomas et al.
Exploiting Similarities among Languages for Machine Translation

Mikolov, Tomas et al.
RNNLM – Recurrent Neural Network Language Modeling Toolkit
Kobayashi, Tetsuo et al.  
**Vocabulary Spurt Reexamined**  
*Baby Science* 12, pp. 40-64, 2012 (in Japanese)

Unno, Yuya.  
**PFI Seminar 2016/03/17**  
http://www.youtube.com/watch?v=DU-sNfSuCrc (in Japanese)

Tokui, Seiya.  
**PFI Seminar 2016/02/25**  
http://www.youtube.com/watch?v=uor4L7p9H0s (in Japanese)

Okanohara, Daisuke.  
**Deep Learning in real world @Deep Learning Tokyo**  
http://www.slideshare.net/pfi/deep-learning-in-real-world-deep-learning-tokyo