Machine Learning Final Project

Learning Continuous Phrase Representations for Translation Modeling

Richard Luong (2014403075)
Kerry Zhang (2014403073)

June 25, 2015

Table of Contents

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Deep learning

- Multiple processing layers
  - Layers of abstraction
- Neural networks
- State-of-the-art
  - Visual object recognition
  - Object detection
  - Speech recognition
  - Drug discoveries, genomics
- Statistical machine translation?

Problem introduction

Learning Continuous Phrase Representation for Translation Modeling - *Gao et. al. 2013*

This paper tackles the sparsity problem in estimating phrase translation probabilities by learning continuous phrase representations, whose distributed nature enables the sharing of related phrases in their representations.

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Problem introduction

Data sparsity

- Phrases as translation units
  - Not linguistic notion
  - Sequence of words
- Longer phrases
  - Better translations
  - Less likely seen
  - Data sparsity

Dataset

TED talk transcripts (English and Chinese)

Dataset

TED talk transcripts (English and Chinese)

- Source sentences (Chinese)

Dataset

TED talk transcripts (English and Chinese)

- Source sentences (Chinese)
- Target sentences (English)

Dataset

TED talk transcripts (English and Chinese)

- Source sentences (Chinese)
- Target sentences (English)
- N-best-list (N translation hypotheses per sentence)
  - Generated by Moses: open source statistical machine translation system

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

The log-linear model for SMT

Standard statistical machine translation (SMT) systems are based on linear combinations of different features:

- Phrase translation table
- Reordering model
- Language model
- Many extensions:
  - Bidirectional translation probabilities
  - Word penalty
  - Phrase penalty
  - Continuous-space Phrase Translation Model (CPTM)

N-best (translation hypothesis) list generated for each source sentence

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors
-

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors
-

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors
-

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase $\rightarrow$ Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase $\rightarrow$ Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors

Continuous-space Phrase Translation Model (CPTM)

- Addresses the data sparsity problem
- Score translation pair based on notion of similarity
- Dimensionality reduction
  - Raw phrase → Bag-of-words (word vector)
  - Neural network
    - Shared vocabulary for both languages
    - Same neural network for both languages
    - Helps discovering latent features
  - Output: lower dimension feature vector
- Similarity is measured by dot product of output feature vectors
-

Overview

Learning Continuous Phrase Representation for Translation Modeling - *Gao et. al. 2013*

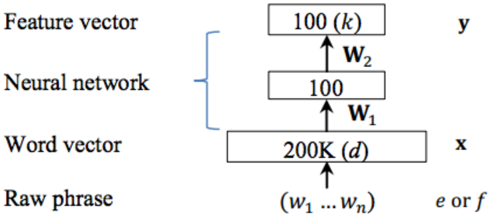


Figure 2. A neural network model for phrases giving rise to their continuous representations. The model with the same form is used for both source and target languages.

Objective function / Loss function

$$L(\theta) = -xBLEU(\theta)$$

$xBLEU$ - N-best list expected BLEU score

BLEU score

BLEU - Bilingual Evaluation Understudy

It is the most widely used automated method of determining the quality of machine translation. The BLEU metric scores a translation on a scale of 0 to 1, but is frequently displayed as a percentage value. The closer to 1, the more the translation correlates to a human translation.

Objective function / Loss function

$$L(\theta) = -xBLEU(\theta)$$

$xBLEU$ - N-best list expected BLEU score

Objective function

$$\mathrm{xBleu}(\boldsymbol{\theta}) = \sum_{E \in \mathrm{GEN}(F_i)} P(E|F_i)\,\mathrm{sBleu}(E_i, E)$$

sBLEU - sentence level BLEU

$P(E|F_i)$ - translation probabilty using *softmax*

$$P(E|F_i) = \frac{\exp(\gamma \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{h}(F_i, E, A))}{\sum_{E' \in \mathrm{GEN}(F_i)} \exp\left(\gamma \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{h}(F_i, E', A)\right)}$$

Similarity

$$\mathbf{y} \equiv \phi(\mathbf{x}) = \tanh\left(\mathbf{W_2}^{\mathrm{T}}(\tanh(\mathbf{W_1}^{\mathrm{T}}\mathbf{x}))\right) \quad (2)$$

$$\mathrm{score}(f,e) \equiv \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f,\mathbf{x}_e) = \mathbf{y}_f^{\mathrm{T}}\mathbf{y}_e$$

Parameters

Parameters: $\theta = \{W_1, W_2\}$

Parameters

Parameters: $\theta = \{W_1, W_2\}$

1. We use a baseline phrase-based SMT system to generate for each source sentence in training data an N-best list of translation hypotheses[4].
2. We set $\boldsymbol{\lambda}$ to that of the baseline system and let $\lambda_{M+1} = 1$, and optimize $\boldsymbol{\theta}$ w.r.t. a loss function on training data[5].
3. We fix $\boldsymbol{\theta}$, and optimize $\boldsymbol{\lambda}$ using MERT (Och 2003) to maximize BLEU on dev data.

Learning θ

- We learn θ by using gradient based numerical optimization algorithm (L-BFGS)
- We compute the gradient of the loss function

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{(f,e)} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)} \frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \boldsymbol{\theta}}$$

$$= \sum_{(f,e)} -\delta_{(f,e)} \frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \boldsymbol{\theta}} \qquad (7)$$

COMPUTING $\delta sim_\theta(x_f, x_e)/\delta\theta$

$$\frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \mathbf{W}_1}$$

$$= \mathbf{x}_f \left(\mathbf{W}_2 \left(\mathbf{y}_e^2 \circ \sigma'(\mathbf{z}_f^2) \right) \circ \sigma'(\mathbf{z}_f^1) \right)^{\mathrm{T}}$$

$$+ \mathbf{x}_e \left(\mathbf{W}_2 \left(\mathbf{y}_f^2 \circ \sigma'(\mathbf{z}_e^2) \right) \circ \sigma'(\mathbf{z}_e^1) \right)^{\mathrm{T}}$$

Computing $\delta sim_\theta(x_f, x_e)/\delta\theta$

$$\frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \mathbf{W}_1}$$

$$= \mathbf{x}_f \left(\mathbf{W}_2 \left(\mathbf{y}_e^2 \circ \sigma'(\mathbf{z}_f^2) \right) \circ \sigma'(\mathbf{z}_f^1) \right)^{\mathrm{T}}$$

$$+ \mathbf{x}_e \left(\mathbf{W}_2 \left(\mathbf{y}_f^2 \circ \sigma'(\mathbf{z}_e^2) \right) \circ \sigma'(\mathbf{z}_e^1) \right)^{\mathrm{T}}$$

$$\frac{\partial \mathrm{sim}_{\boldsymbol{\theta}}(\mathbf{x}_f, \mathbf{x}_e)}{\partial \mathbf{W}_2} = \frac{\partial (\mathbf{y}_f^2)^{\mathrm{T}}}{\partial \mathbf{W}_2} \mathbf{y}_e^2 + (\mathbf{y}_f^2)^{\mathrm{T}} \frac{\partial \mathbf{y}_e^2}{\partial \mathbf{W}_2}$$

$$= \mathbf{y}_f^1 \left(\mathbf{y}_e^2 \circ \sigma'(\mathbf{z}_f^2) \right)^{\mathrm{T}} + \mathbf{y}_e^1 \left(\mathbf{y}_f^2 \circ \sigma'(\mathbf{z}_e^2) \right)^{\mathrm{T}} \quad (9)$$

Computing the error term $\delta_{(f,e)}$

$$\delta_{(f,e)} = \sum_{(E,A)\in GEN(F_i)} \mathrm{U}(\boldsymbol{\theta}, E) P(E|F_i) \lambda_{M+1} N(f, e; A) \tag{14}$$

where

$$\mathrm{U}(\boldsymbol{\theta}, E) = \mathrm{sBleu}(E_i, E) - \mathrm{xBleu}(\boldsymbol{\theta}).$$

Computing the error term $\delta_{(f,e)}$

$$\delta_{(f,e)} = \sum_{(E,A)\in GEN(F_i)} \mathrm{U}(\boldsymbol{\theta}, E) P(E|F_i) \lambda_{M+1} N(f, e; A) \quad (14)$$

where

$$\mathrm{U}(\boldsymbol{\theta}, E) = \mathrm{sBleu}(E_i, E) - \mathrm{xBleu}(\boldsymbol{\theta}).$$

$N(f,e)$ - number of times (f,e) occurs in A

$$\mathrm{xBleu}(\boldsymbol{\theta}) = \sum_{E\in \mathrm{GEN}(F_i)} P(E|F_i)\mathrm{sBleu}(E_i, E)$$

Computing the error term $\delta_{(f,e)}$

$$\delta_{(f,e)} = \sum_{(E,A) \in GEN(F_i)} \mathrm{U}(\boldsymbol{\theta}, E) P(E|F_i) \lambda_{M+1} N(f,e;A)$$

where (14)

$$\mathrm{U}(\boldsymbol{\theta}, E) = \mathrm{sBleu}(E_i, E) - \mathrm{xBleu}(\boldsymbol{\theta}).$$

$N(f,e)$ - number of times (f,e) occurs in A

$$\mathrm{xBleu}(\boldsymbol{\theta}) = \sum_{E \in \mathrm{GEN}(F_i)} P(E|F_i) \mathrm{sBleu}(E_i, E)$$

$$P(E|F_i) = \frac{\exp(\gamma \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{h}(F_i, E, A))}{\sum_{E' \in \mathrm{GEN}(F_i)} \exp\left(\gamma \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{h}(F_i, E', A)\right)}$$

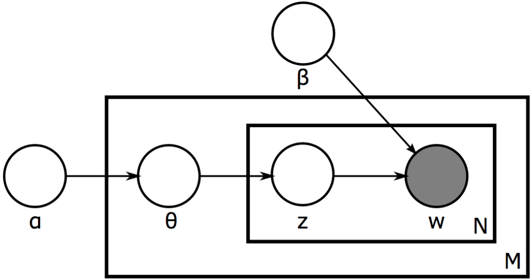
Learning θ

- The parameters are trained using **Stochastic gradient descent**
- Initialize

  W_1 bilingual topic model trained on parallell data

  W_2 identity matrix

Latent Dirichlet Allocation



Latent Dirichlet Allocation

Preprocessing:

- Stop words

Latent Dirichlet Allocation

Preprocessing:

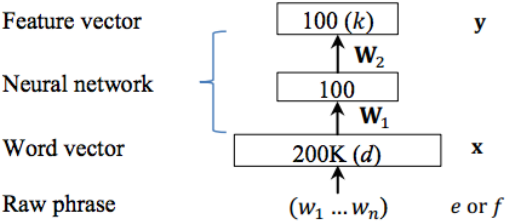
- Stop words

Output: W_1 - 100 x 50,000

Results

After multiple overnight test runs, our implementation of the CPTM algorithm conducted into satisfiable results.

Summary



$$\theta = \{W_1, W_2\}$$

Gradient descent:

$$\theta_{n+1} = \theta_n - \eta \frac{\partial L(\theta)}{\partial \theta} = \theta_n - \eta \left[\sum_{(f,e)} -\delta_{(f,e)} \right.$$

Results

In order to test the theoretical validity of our system, a plot over the loss function was made.

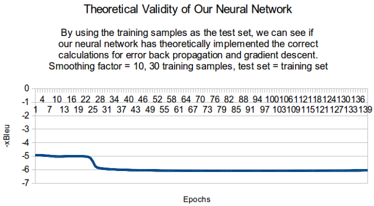


Figure: Theoretical validty of our neural network

Results

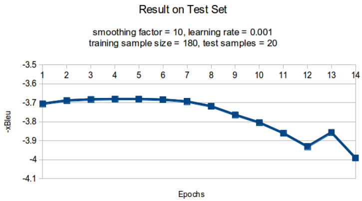


Figure: Result on test set

The results seem to suggest that the system is indeed performing a correct gradient descent.

Summary

- We have been learning continuous phrase representations for a SMT system.

Summary

- We have been learning continuous phrase representations for a SMT system.
- Using a neural network, we have projected both source and target sentences into a real-number vector space.

Summary

- We have been learning continuous phrase representations for a SMT system.
- Using a neural network, we have projected both source and target sentences into a real-number vector space.
- Loss function of the NN has been the negative expected BLEU score for a source sentence and it's possible translation.

Summary

- We have been learning continuous phrase representations for a SMT system.
- Using a neural network, we have projected both source and target sentences into a real-number vector space.
- Loss function of the NN has been the negative expected BLEU score for a source sentence and it's possible translation.
- We also tested the validity of our system, showing that the loss function decreases.

Finally

Thank you for listening! It has been an adventurous journey.