Unsupervised Learning of Neural Network Lexicon and Cross-lingual Word Embedding

Jiahui Geng

jiahui.geng@rwth-aachen.de

Master Thesis Mid-term Talk
August 3, 2018

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

Introduction

Literature

Unsupervised word-by-word translation system

- **Model**
- **Word translation**
  - **Monolingual word embedding**
  - **Linear mapping between embedding spaces**
- **Sentence Translation**
- **Experiments**

Outlook

Introduction

Motivation

- Building a machine translation system requires lots of bilingual data
- Cross-lingual word embedding offers elegant word matches between languages
- Unsupervised MT relies on back-translation which needs a long training time

Goals

- Study training details of cross-lingual word embedding
- Build a good unsupervised MT efficiently: combine with other models
- Improve the unsupervised learning algorithm for cross lingual word embedding

Literature

Unsupervised cross-lingual embedding

- [Artetxe & Labaka^{+} 17a] Learning bilingual word embeddings with (almost) no bilingual data
  - A self-learning framework combining embedding mapping and dictionary induction techniques, needs small dataset to start
- [Hoshen & Wolf 18] An Iterative Closest Point Method for Unsupervised Word Translation
  - Iterative closest point method for embedding mapping learning, without neural network but more interpretable
- [Conneau & Lample^{+} 17] Word translation without parallel data
  - Implementation of GANs: discriminator trained to distinguish between two distributions while generator fools discriminator when learning mapping

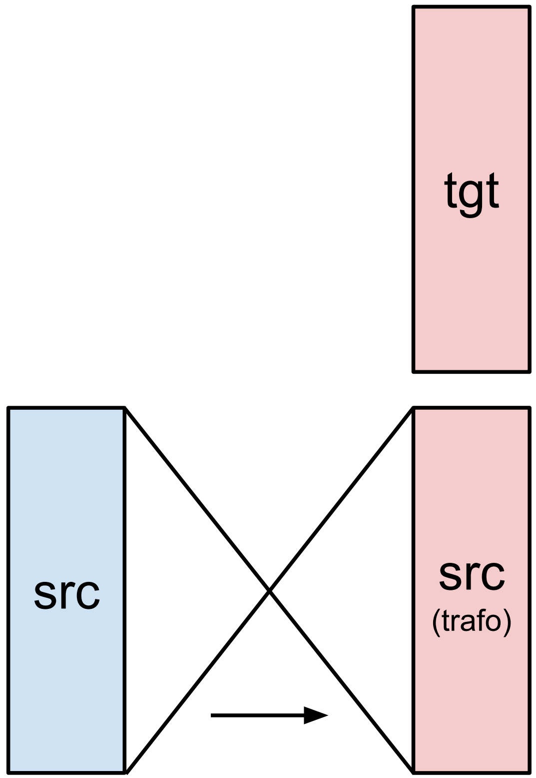
Literature

Unsupervised machine translation

- [Artetxe & Labaka+ 17b] Unsupervised Neural Machine Translation
- [Lample & Denoyer+ 17] Unsupervised Machine Translation Using Monolingual Corpora Only
  - Seq2seq model with encoder and decoder for both languages, also with denoising autoencoder and back-translation
- [Artetxe & Labaka+ 17b] Phrase-Based & Neural Unsupervised Machine Translation
  - Simplifying the architecture and loss function, still following the above mentioned principles and propose a PBSMT with back-translation

Word Translation

- Learn monolingual embedding separately
  - Skip-gram model
    `fasttext` [Joulin & Grave+ 16]
- Learn linear mapping between embedding spaces
  - Supervised learning
    - Procrustes analysis
  - Unsupervised learning
    - Adversarial learning
    - Iterative refinement
- Bidirectional dictionary induction
  - CSLS retrieval



Monolingual Embedding

Fasttext

- **Essentially an extension of skip-gram/CBOW model**
- **Treat each word as composed of character n-gram**
- **Learn the internal structure of words**

Problem

- **Not accurate for rare words (usually name entities)**

Cross-lingual Word Embedding

Definition

- Word embedding of multiple languages in a joint embedding space

Roles in unsupervised neural machine translation

- Shared latent representations
  - Shared encoder for producing a language independent representation
  - Back-translation for further improvement
- This work
  - Formulate a straightforward way to combine a language model with cross-lingual word similarities

Cross-lingual Word Embedding

Training Methods

- Mapping-based approaches (this work)
  - Train word embeddings separately then learn mapping with bilingual dictionaries
- Pseudo-multi-lingual corpora-based approaches
  - Use monolingual word embedding methods on mixed corpus of multiple languages
- Joint methods
  - Minimize the monolingual losses with the cross-lingual regularization term

Supervised Learning

Assume we have

- **Word embedding**
  trained independently for each language on monolingual corpora
- **Bilingual dictionary**
  a known dictionary with pairs of words $\{f, e\}$ size s

Learn a linear mapping $W \in \mathbb{R}^{d \times d}$ such that

$$W^* = \underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \sum_{i=1}^{s} \|W f_i - e_i\|$$

- d **: Dimension of embedding**
- $f_i, e_i \in \mathbb{R}^d$**: the embedding pair of corresponding word pair in the dictionary**

Procrustes Analysis

Constrain W to be an orthogonal matrix

- **Enforce monolingual invariance**
- **Simplify the problem as the Procrustes problem**
  - **A closed-form solution obtained from SVD**
  - **E, $F \in \mathbb{R}^{d*s}$ denotes embedding projection of word pairs $\{e, f\}$**

$$W^* = \underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \|WF - E\| = UV^T$$

$$U\Sigma V^T = \mathrm{SVD}(EF^T)$$

- **Can be efficiently computed in linear time w.r.t. seed dictionary size s**

Unsupervised Word Embedding Mapping

Problem

- Large dictionary not readily available for many language pairs

Methods

- Learn bilingual embeddings without any bilingual evidence (this work)
  - Adversarial training
- Design the seed dictionary
  - Shared words, digits and cognates
  - Design heuristics to build the seed dictionary

Adversarial Training

Model

- $\mathcal{F} = \left\{ f_1, \dots f_{V_f} \right\}$ and $\mathcal{E} = \{e_1, \dots e_{V_e}\}$: set of embeddings, not parallel
- Discriminator is trained to discriminate $W f_i$ and e_i with f_i, e_i randomly sampled from $\mathcal{F}, \mathcal{E}$
- Generator W is trained to prevent the discriminator from making accurate prediction

Discriminator loss

$$\mathcal{L}_D(\theta_D | W) = -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_D}(source = 1 | W f_i) - \frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_D}(source = 0 | e_i)$$

Generator loss

$$\mathcal{L}_D(W | \theta_D) = -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_D}(source = 0 | W f_i) - \frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_D}(source = 1 | e_i)$$

Iterative Refinement

Self-learning framework

1. **Dictionary is important to train the cross-lingual embedding**
2. **Start from a initial dictionary, use such dictionary as input to learn cross-lingual mapping**
3. **Assume the dictionary inducted from the learned mapping is better and can provide better mapping further**
4. **Design a convergence criterion, if not satisfied, keep training**

Dictionary Induction

Cross-domain Similarity Local Scaling (CSLS)

- Nearest neighbour search suffers from the hubness problem
  - Points tending to be nearest neighbors of many points in high-dimensional spaces
- Penalize the similarity score of hubs
  - $N_T(Wf)$: target neighbours for mapped source embedding
  - $r_T(Wf)$: penalty for hubness

$$r_T(Wf) = \frac{1}{K} \sum_{e \in N_T(Wf)} \cos(Wf, e)$$

$$\mathrm{CSLS}(Wf, e) = 2\cos(Wf, e) - r_T(Wf) - r_S(e)$$

Bidirection dictionary induction

- Unidirectional dictionary might lead to local optima
- Include only the mutual nearest neighbors
- Select more probable candidates as pairs

Sentence Translation

Context-aware Beam Search

- **Language model**

Denoising Autoencoder

- **Insertion**
- **Deletion**
- **Reordering**

Context-aware beam search

- Given a history h of target word before e, the score of e to be the translation of f:

$$L(e; f, h) = \lambda_{emb} q(f, e) + \lambda_{LM} p(e|h)$$

- Lexicon score $q(f, e) \in [0, 1]$ defined as:

$$q(f, e) = \frac{d(f, e) + 1}{2}$$

- $d(f, e) \in [-1, 1]$ cosine similarity between f and e
- In experiments, such lexicon score works better than others, e.g. sigmoid or softmax

Denoising

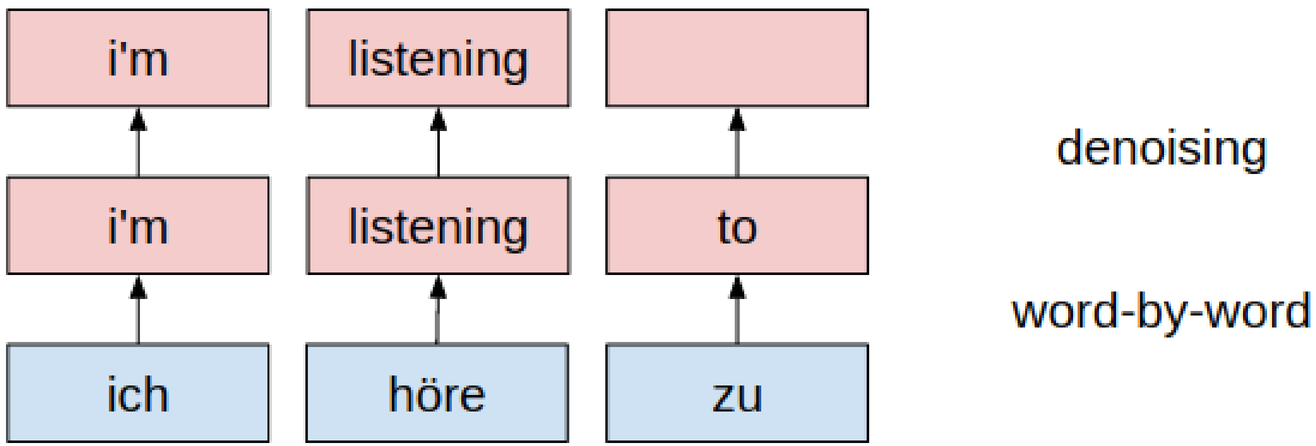
- **Model such $c(e_1^I)$ by injecting artificial noise into clean sentences e_1^I**
- **Training criterion:**

$$L = E_{e_1^I \in E}[-log(e_1^I|C(e_1^I))]$$

- **In Seq2Seq training, e_1^I as label, $c(e_1^I)$ as input**
- **Artificial noise:**
  - **insertion, deletion, reordering**

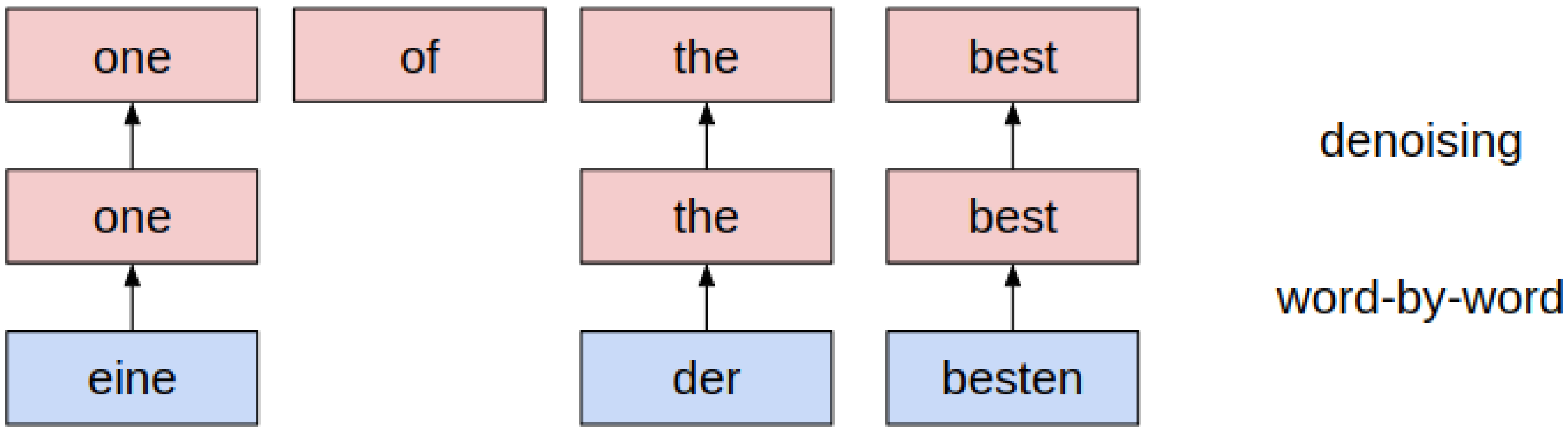
Insertion

- Motivation
  - Word-by-word translation always outputs a target word for every position
  - Some common words are considered as redundant ones
- Method
  - For each position in a sentence, insert a frequent word according from set v_{ins} to a probability distribution p_{ins}
  - Denoising network learns to delete the word when translating



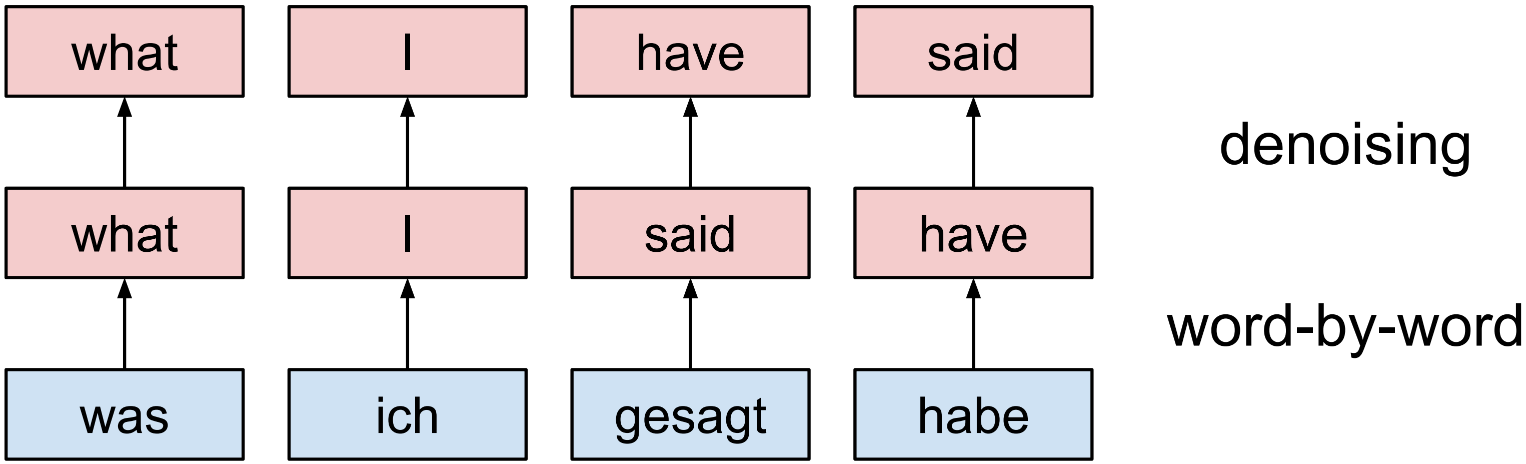
Deletion

- **Motivation**
  - **In contrary case: some words are not related to any source word**
- **Realization**
  - **For each position in a sentence, delete the word according to a probability distribution p_{del} as input**
  - **Denoising network learns to add some potential words when translating**



Reordering

- **Motivation**
  - **Generated words are not in a correct sequence of the target language**
- **Method**
  - **For each position of a sentence, swap the words within a limited distance d_{per} as input**
  - **Denoising network learns reordering information when translating**



Experiment Settings

- Word embedding and LM learned on 100M sentences from `wmt 2014-2017`
- BLEU evaluated on German↔English `newstest2016`
- Word accuracy evaluated on dictionaries released by Facebook
  - Dictionary built with internal translation tool
  - Each word has 1-4 word translation(s)
  - Top-1 accuracy: if top 1 candidate in the dictionary
- Context-aware beam search: Lexicon candidates: 100 / beam width 10

Experiments

Translation results on German↔English `newstest2016` and French↔English `newstest2014`.

System	OOV	de-en BLEU [%]	en-de BLEU [%]	fr-en BLEU [%]	en-fr BLEU [%]
Word-by-Word	None	11.1	6.7	10.6	7.8
+ LM	LM	12.9	8.9	12.7	10.0
	Copy	14.5	9.9	13.6	10.9
+ Denoising (RNN)		16.2	10.6	15.8	13.3
+ Denoising (Transformer)		17.2	11.0	16.5	13.9
[Lample & Denoyer^{+} 17]		13.3	9.6	14.3	15.1
[Artetxe & Labaka^{+} 17b]		-	-	15.6	15.1

Ablation studies

- Different sizes of training corpora
- Different vocabularies: BPE and word
- Different vocabulary sizes for cross-lingual training
- Different denoising parameters
- Phrase embedding
- Vocabulary cut-off

Different Training Corpora

Word-by-word translation from German to English

	Accuracy [%]	Bleu [%]
5M	44.9	9.7
10M	51.6	10.1
50M	59.4	10.8
100M	61.2	11.2

- Larger corpus improves the word translation accuracy
- Also improves the word-by-word translation

Different Embeddings and Traning Vocabulary Size

	Vocabulary	BLEU [%]
	Merges	
BPE	20k	10.4
	50k	12.5
	100k	13.0
	Cross-lingual training	
Word	20k	14.4
	50k	14.4
	100k	14.5
	200k	14.4

- Word-by-word translation with language model
- Word embedding performs better than BPE embedding
- Embedding trained on 20k similar to 200k $\Rightarrow$ Frequent words matter

Denoising Experiments

d_{per}	p_{del}	p_{ins}	p_{ins}	BLEU [%]
2				14.7
3				14.9
5				14.9
3	0.1			15.7
	0.3			15.1
3	0.1	0.1	10	16.8
			50	17.2
			500	16.8
			5000	16.5

- Each artificial noise improves the translation performance

Phrase Embedding

Motivation

- Many phrases have a meaning that is not a simple composition of the meaning of its individual words

Phrase detection

- Phrases formed based on the unigram and bigram counts: [Mikolov & Sutskever^{+} 13]
  - Tune a good threshold value for score

$$\text{score}(e', e) = \frac{\text{count}(e', e) - \delta}{\text{count}(e') * \text{count}(e)}$$

- Process sentences with most common phrases in training corpus
  - Count the most frequent bi-gram phrases: $\text{score}(e', e) = \text{count}(e', e)$
  - Detect phrases as top frequent phrases in the training corpus

Phrase Embeddings

Vocabulary			No LM BLEU [%]	With LM BLEU [%]	Denoising BLEU [%]
Word			11.2	14.5	17.2
[Mikolov & Sutskever+ 13]	threshold	100	11.1	13.7	15.6
		500	11.0	13.7	16.2
		2000	10.7	14.0	16.5
Top frequent	count	50k	12.0	15.7	16.8

- Phrase embeddings helps only for WBW and +LM

Souce and Target Vocanulary Cut-off

- Column: source vocabulary size/ row: target vocabulary size

Word embedding vocabulary cut-off

BLEU [%]	20k	50k	100k
50k	11.1	11.3	11.2
100k	11.2	11.2	11.1
150k	10.9	10.9	-

Phrase embedding vocabulary cut-off

BLEU [%]	50k	100k	150k
50k	11.3	-	-
100k	11.9	11.9	-
150k	12.0	11.9	11.9
200k	12.0	-	-

- Vocabulary size effects the translation performance

LM supported Cross-lingual embedding training

Basic idea:

- **Language model help to select candidates, provide better dictionary**
- **Dictionary from the sentence translation, instead of induction**
- **Training the mapping with SGD instead of Procrustes analysis**

Conclusions

Comprehensive results

- **Context-aware beam search with LM helps the lexicon choice**
- **Denoising networks aimed at insertion/deletion/reordering noise works for such problems in a small range of sentences**

Ablation studies

- **BPE embeddings performs worse than word embeddings, especially with smaller vocabulary size.**
- **Word-by-word translation with cross-lingual embedding depends highly on the frequent word mappings**
- **Phrase embedding only helps in WBW and Context-aware beam search**

Outlook

Goal: Improve the unsupervised learning for cross-lingual embedding

- Accordingly improves unsupervised MT performance
- Other applications: transfer learning for low-resource LM [Adams & Makarucha^{+}

Exchange algorithm with LM for inducing initial bilingual dictionary

- Adversarial training is not interpretable and relies on random starts
- Using LM: strong training signal and less dependence on randomness

Non-linear mapping between source and target

- Linear assumption may be too crude
- Stochastic gradient descent instead of SVD
- Also applies in supervised case

Thank you for your attention

Jiahui Geng

jgeng@cs.rwth-aachen.de

http://www.hltpr.rwth-aachen.de/

References

[Adams & Makarucha+ 17] O. Adams, A. Makarucha, G. Neubig, S. Bird, T. Cohn: Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1, pp. 937–947, 2017.

[Artetxe & Labaka+ 17a] M. Artetxe, G. Labaka, E. Agirre: Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 451–462, 2017.

[Artetxe & Labaka+ 17b] M. Artetxe, G. Labaka, E. Agirre, K. Cho: Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, Vol., 2017.

[Conneau & Lample+ 17] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou: Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, Vol., 2017.

[Hoshen & Wolf 18] Y. Hoshen, L. Wolf: An Iterative Closest Point Method for Unsupervised Word Translation. *arXiv preprint arXiv:1801.06126*, Vol., 2018.

[Joulin & Grave+ 16] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov: FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, Vol., 2016.

[Lample & Denoyer+ 17] G. Lample, L. Denoyer, M. Ranzato: Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043*, Vol., 2017.

[Mikolov & Sutskever+ 13] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean: Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.