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

Jiahui Geng

jiahui.geng@rwth-aachen.de

Master Thesis Proposal
June 22, 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
  - Lonolingual word embedding
  - Linear mapping between embedding space
- Sentence Translation
- Experiments

Outlook

Introduction

Why unsupervised learning?

- **Overcome the lack of reference translations**
- **Rich monolingual sentence resource**

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, need small dataset to start
- [Conneau & Lample+ 17] Word translation without parallel data
  - Implementation of GANs: the discriminator plays an adversarial role to a generative model and is trained to distinguish between two distributions
- [Hoshen & Wolf 18] An Iterative Closest Point Method for Unsupervised Word Translation
  - Iterative Closest Point Method for unsupervised embedding mapping, fewer hyper-parameters, more interpretable

Literature

Unsupervised machine translation

- [Artetxe & Labaka$^+$ 17b] Unsupervised neural machine translation
  - With fixed cross-lingual embeddings to train a shared encoder, train the system with de-noising and on-the-fly back-translation alternatively
- [Lample & Denoyer$^+$ 17] Unsupervised Machine Translation Using Monolingual Corpora Only
  - Seq2seq model with encoder and decoder for both language, also with denoise autoencoder and back-translation
- [Artetxe & Labaka$^+$ 17b] Phrase-Based & Neural Unsupervised Machine Translation
  - Simplifying the architecture and loss function while still following the above mentioned principles and propose a PBSMT with back-translation

Cross-lingual Word Embedding

Definition

- Word embedding of multiple languages in a joint embedding space

Roles in unsupervised machine translation

- Initilization
  - Populate the initial phrase tables using scores from cross-lingual embeddings
  - Preserve some of the original semantics
- Shared Latent representations
  - Shared encoder for producing a language independent representation
  - Enable back-translation further improvement of results

Word Translation

- Learn monolingual embedding separately
  - Fasttext
- Learn linear mapping between embedding spaces
  - Supervised learning
    - Procrustes analysis
  - Unsupervised learning
    - Adversarial learning
    - Iterative refinement
- Bidirectional dictionary induction
  - CSLS retrieval

Cross-lingual Word Embedding

Training Methods

- Mapping-based approaches
  - Train word embeddings 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 embeddings
  trained independently for each language on monolingual corpora
- Bilingual dictionary
  a known dictionary with pairs of words $\{f, e\}$

Learn a linear mapping W such that

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}}{\operatorname{argmin}} \|\boldsymbol{W}\boldsymbol{F} - \boldsymbol{E}\|$$

- d : Dimension of embeddings
- $\boldsymbol{F}$, $\boldsymbol{E}$: Aligned $d \times s$ real matrices containing the embeddings of the words in the dictionary
- s: Seed dictionary size

Procrustes Analysis

Constrain W to be an orthogonal matrix

- **Enforce monolingual invariance**
- **Simplify the problem as the the Procrustes problem which has a closed-form solution obtained from the SVD of EF^T:**

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

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

- **Can be efficiently computed in linear time $w.r.t$ number of dictionary entries**

Unsupervised Word Embedding Mapping

Problem

- Large dictionary not readily available for many language pairs

Methods

- Design the seed dictionary
  - Using document-aligned corpora to extract the training dictionary
  - Relying on shared words, digits and cognates
- Learn bilingual embeddings without any bilingual evidence
  - Adversarial training

Adversarial Training

Model

- $\mathcal{F} = \left\{ f_1, \dots f_{V_f} \right\}$ and $\mathcal{E} = \{e_1, \dots e_{V_e}\}$: Sets of word embeddings
- Discriminator is trained to discriminate between elements randomly sampled from $W\mathcal{F}$ and $\mathcal{E}$
- Generator W is trained to prevent the discriminator from making accurate prediction

Discriminator loss

$$\mathcal{L}_D = -\log D(\mathcal{E}) - \log(1 - D(W\mathcal{F}))$$

Generator loss

$$\mathcal{L}_W = -\log D(W\mathcal{F})$$

Trick: Frequency-based Vocabulary Cutoff

Problem

- **Rare word embeddings are less trained(updated)**
- **Contain noise information for alignment**

Experiment

Iterative Refinement

Algorithm 1: Self-learning framework

Input: $\mathcal{F}$ (source embeddings)
Input: $\mathcal{E}$ (target embeddings)
Input: $\mathcal{D}$ (seed dictionary)
Result: $\mathcal{W}$ (embedding mapping)

initialization;
while *not convergence criterion* **do**
  $\mathcal{W} \leftarrow learn\_mapping(\mathcal{E}, \mathcal{F}, \mathcal{D})$;
  $\mathcal{D} \leftarrow learn\_mapping(\mathcal{E}, \mathcal{F}, \mathcal{W})$;
end

Dictionay Induction

Cross-domain Similarity Local Scaling

- KNN suffers from the hubness problem
- Penalize the similarity score of hubs
  - $N_T(Wf)$: target neighbours for mapped source embedding
  - $r_T(Wf)$: penalty for hubness

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

$$CSLS(Wf, e) = 2cos(Wf, e) - r_T(Wf) - r_S(e)$$

Bidirection Dictionary Induction

- Repreated word in unidirectional dictionary might lead to local optima
- Include the dictionary in both directions

Sentence Translation

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:**

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

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

Denoising

- **Substitions, insertions, deletions, reordering viewed as noise in word-by-word translation**
- **Model such noise $c(t)$ by injecting artificial noise into clean sentences t**
- **Language modelling via denoising autoencoder can improve the translation by minimizing:**

$$L = E_{t \in T}[-log P_{t \to t}(t|C(t))]$$

- **In Seq2Seq training, t as label, $c(t)$ as input**

Results

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

	de-en	en-de	fr-en	en-fr
System	BLEU [%]	BLEU [%]	BLEU [%]	BLEU [%]
Word-by-Word	11.1	6.7	10.6	7.8
+ LM	12.9	8.9	12.7	10.0
	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

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

$$score(w_i, w_j) = \frac{count(w_i, w_j) - \delta}{count(w_i) * count(w_j)}$$

Translation performance

- Higher threshold, fewer phrases
- Topk: 500 beam size: 30

threshold	100	500	2000
BLEU	100	13.83	13.65

Outlook

Cross-lingual word embedding and word-to-word MT system

- **Develop a new training algorithm for cross-lingual embeddings**
  - **Context/domain considered, e.g. LM**
  - **Better constraints on specific (group of) words**

- **Word-to-word MT system with cross-lingual embeddings**
  - **Efficient nearest neighbour search**
  - **Combination with a language model**

- **Compare translation results with word-to-word neural lexicons**
  - **All trained/tested on intact corpora without artificial change of alignments**

Appendix: Denoising & Vocabulary

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

Translation results with different values of denoising parameters.

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

Translation results with different vocabularies.

Thank you for your attention

Jiahui Geng

`jgeng@cs.rwth-aachen.de`

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

References

[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.

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