Neural Machine Translation from Historical Japanese to Contemporary Japanese Using Diachronically Domain-Adapted Word Embeddings

Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, Kanako Komiya

Table of Contents

1. Motivation
2. Background
3. Experimental Setup
4. Results & Analysis
5. Conclusion & Discussion

SMT vs. NMT

SMT:

- analysis of existing translations (bilingual text corpora)
- rules for higher statistical probability + calibration
- no further human supervision

NMT:

- neural network
- training → adjustments based on error in output
- fine-tuning

Comparison:

- SMT: more resources + less virtual space, but dependent on source quality
- NMT: self-learning, higher quality/fluency, but dependent on clarity

1. Motivation & Background

- NMT preferable, but small parallel corpus
- → initialization with pre-trained word embeddings

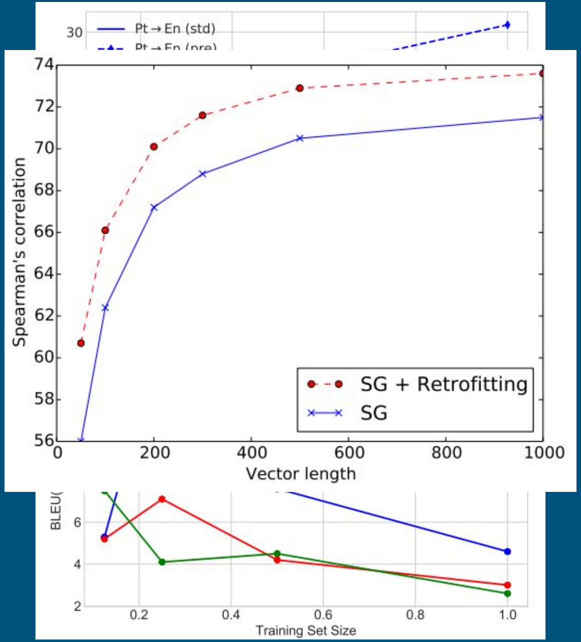
2 Problems:

- a) domains of embeddings and input differ
- b) meaning of words or words themselves change
- → gradual diachronic domain adaptation

2. Background

- Hoshino et al. (2014): sentence-based parallel corpus → SMT
- NMT self-learns word embeddings, but...

- Qi et al. (2018): pre-trained word embeddings
- Faruqui et al. (2015): retrofitting method
- Yaginuma et al. (2018): word sense disambiguation
- Kim et al. (2014): diachronic fine-tuning



3. Experimental Setup

Model:

- encoder-decoder with LSTM & attention (via OpenNMT)
- 2 unidirectional LSTM layers
- global attention
- hyper parameters based on preliminary experiments
- weights via adapted embeddings

Evaluation:

- BLEU

3. Experimental Setup

Comparison between:

- no pre-trained word embeddings (***Baseline***)
- no fine-tuning (***NWJC2vec***)
- fine-tuned with ***Entire historical corpus*** at once
- diachronic domain adaptation at one time (*+Ensemble at one time*)
- diachronic domain adaptation in order of time
  (separated and *+Ensemble in order of time*)

3. Experimental Setup

Historical Japanese	
Total Number of Sentences	86,684
Vocabulary Size	49,200
Number of Tokens	2,774,745
Contemporary Japanese	
Total Number of Sentences	86,684
Vocabulary Size	45,690
Number of Tokens	3,611,783

Data Set:

- parallel corpus by Hoshino et al. (2014)
  - → “The Complete Collection of Japanese Classical Literature” by Shogakukan
- modern (4.577), Kamakura (30.075), Heian (52.032) for total of 86.684 sentences
  - Muromachi only for fine-tuning
- training (82.591), development (2.093), test (2.093)

Period	Number of Example
Modern Test Set	123
Kamakura Test Set	739
Heian Test Set	1,231

- MeCab v0.996 as morphological analyzer (PoS)
- UniDic for Early Middle Japanese v1.3
- UniDic v2.3.0
- input limited to 100 words

3. Experimental Setup

Word Embeddings:

- NWJC2vec (2014-4Q dataset)

Number of URLs collected	83,992,556
Number of sentences (Some are overlapped)	3,885,889,575
Number of sentence (No overlapping)	1,463,142,939
Number of words (tokens)	25,836,947,421

CBOW or skip-gram	-cbow	1
Dimensionality	-size	200
Number of surrounding words	-window	8
Number of negative samples	-negative	25
Hierarchical softmax	-hs	0
Minimum sample threshold	-sample	1e-4
Number of iterations	-iter	15

- fine-tuned via Yaginuma et al. (2018)

CBOW or skip-gram	-cbow	1
Dimensionality	-unit	200
Number of surrounding words	-window	5
Number of negative samples	-negative	5
Batch size	-batchsize	1000
Number of iterations	-iter	10

4. Results & Analysis

Method	BLEU
SMT (Hoshino et al., 2014)	28.02
Baseline	19.22
NWJC2vec	19.16
Entire historical corpus	19.24
+Ensemble at one time	20.94
Diachronic domain adaptation in order of time	
(1) Modern	**19.43**
(3) Modern → Muromachi → Kamakura	19.33
(4) Modern → Muromachi → Kamakura → Heian	19.29
+Ensemble in order of time	21.59

	Modern	Kamakura	Heian
(1) Modern	5.24	25.16	19.53
(3) Modern → Muromachi → Kamakura	4.09	25.65	19.43
(4) Modern → Muromachi → Kamakura → Heian	3.59	25.76	19.40

4. Results & Analysis

(a) An example where diachronic domain adaptation until Kamakura period is better	
Input Sentence:	やまとうた (Japanese poems) の 道 、 浅き に 似 て 深く 、
Reference translation:	和歌 (Japanese poems) の 道 は 、 浅い よう で い て じつ は 深く 、
English translation:	The soul of Japanese poems seems shallow, but it is in fact profound.
Baseline:	<unk> の 道 は 、 浅い の に 似 て 深く 、
Modern:	<unk> の 道 、 浅い の と 同様 に 、 深く て 、
Modern → Muromachi → Kamakura:	和歌 (Japanese poems) の 道 は 、 浅い 時代 に 似 て 深く 、
(b) An example where ensemble method of proposed methods is better	
Input sentence:	大饗 に 劣ら ず 、 あまり 騒がしき (noisy) まで なん 集ひ たまひ ける 。
Reference translation:	大 饗 の とき に 劣ら ない ほど 、 あまり に 騒がしい (noisy) まで 大勢 お 集まり に なる の だっ た 。
English translation:	There came together as many people as people at a royal party and it was almost too noisy.
Baseline:	大 饗 に も 劣ら ず 、 あまり に も あわただしい (hasty) くらい に お 集まり に なっ た 。
Entire historical corpus:	大 饗 に 負け ず 、 あんまり 暑い (hot) まで 集まっ て お いで に なっ た 。
Diachronic domain adaptation:	大 饗 に 劣ら ず 、 あまり に 騒がしい (noisy) まで 集まっ て お いで に なっ た 。

5. Conclusion & Discussion

Cons:

- typos and redundancy (e.g. data set section)
- length of inputs limited to 100 words
- only uses BLEU as evaluation and results inconclusive, but...

Pros:

- novel idea in an under-researched field
- future work: other word embeddings, adaptation via contextual word representations

Questions:

- Why did a transformer model have both inconsistent results and worse averaged performance?
- Can diachronic domain adaptation in order of time be further improved or does the experiment setup have to be changed?