Achieving Open Vocabulary NMT with Hybrid Word-Character Models

ESC-20FALL 학술부 세션 6주차

2017122021 응용통계학과 김수연

Abstract

Achieving Open Vocabulary NMT

- previous works : used restricted vocabularies
- with subsequent method to patch <unk> tokens

with Hybrid Word-Character Models

- translate mostly at ***word*** level + consult ***character*** components for rare / unknown words

Introduction

Advantages of NMT (single dnn trained end-to-end)

- simplicity : simple decoder implementation, small memory usage
- generalization : SOTA for several language pairs

DEALING WITH UNKNOWNS!

<unk> replacement techniques (post-processing step)

- <unk> token의 occurrence를 위치 정보와 함께 기록 -> 사전에서 찾은 단어나 identity copy로 대체
- attention mechanism으로 alignment info 습득

en: The *ecotax* portico in *Pont-de-Buis* , . . . [truncated] . . . , was taken down on Thursday morning

fr: Le *portique* *écotaxe* de *Pont-de-Buis* , . . . [truncated] . . . , a été *démonté* jeudi matin

nn: Le unk de unk à unk , . . . [truncated] . . . , a été pris le jeudi matin

Introduction

Still disadvantages,,,

- crosslingually : 각 언어들은 서로 다른 알파벳들을 가지고 있는데 단어마다 모든 대응관계를 외우기 어렵

  ex. “Christopher” (English) - “Krystof” (Czech)

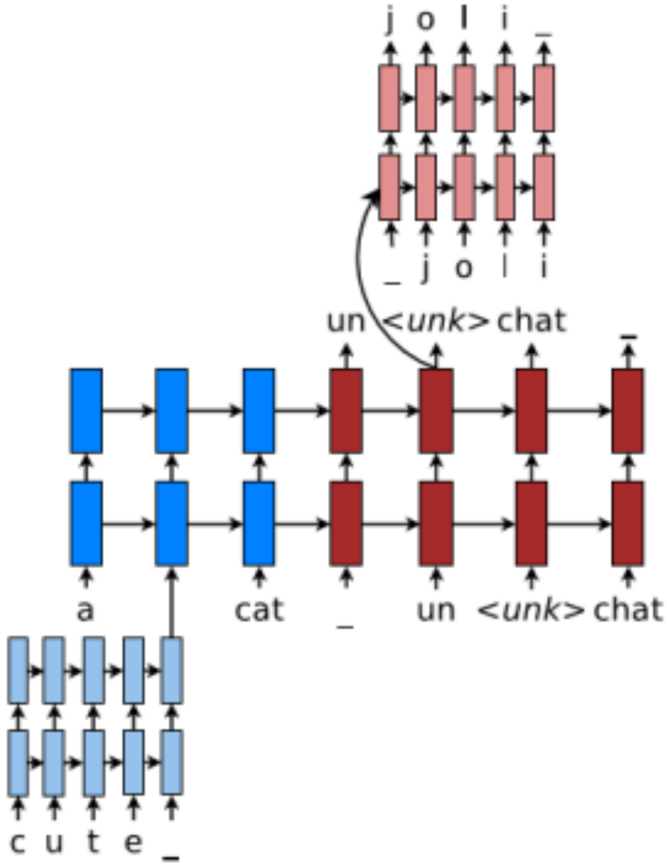
- monolingually : 단어들은 형태론적으로(morphologically) 관련되어 있는데 다른 객체로 취급해버림

  ex. “distinct” - “distinctiveness”

Hybrid Model

- compared to char-based model,,, ***fast and easier to train !***
- compared to word-based model,,, ***never produces unknown words !***
- achieve SOTA with 20.7 BLEU score in English to Czech translation task
- learn to not only generate well-formed words for target language, but also build correct representation for source language!

Introduction



* assume “cute” and “joli” is not in source, target vocabulary

Target side

- *Character level* RNN recovers <unk> tokens character-by-character

Main model

- works at *Word level*
- both components are learned jointly end-to-end

Source side

- *Character level* RNN computes representation for rare/unknown words

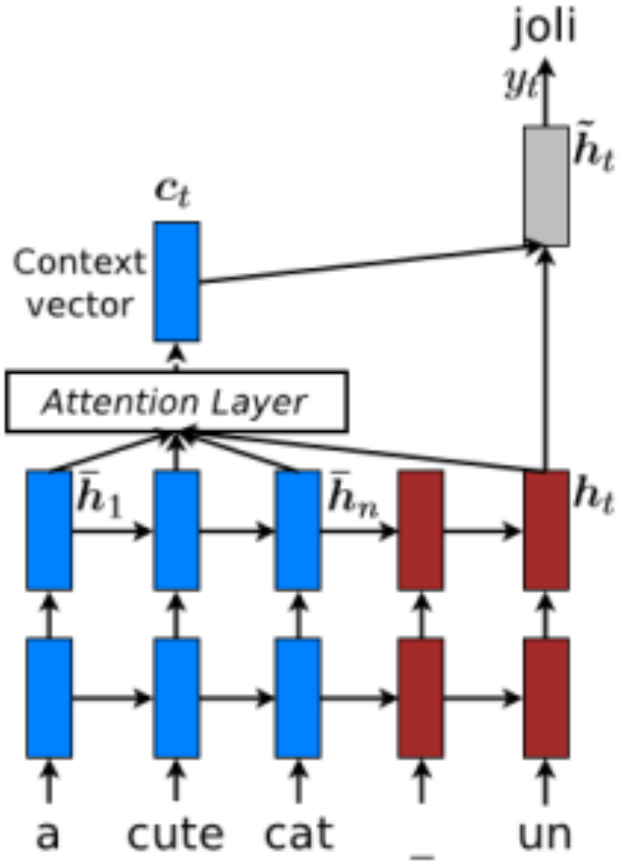
Background & Our Models

Objectives

$$J = \sum_{(x,y)\in D} -\log p(y \mid x)$$

Attention mechanism

$$p(y_t \mid y_{<t}, s) = softmax(\tilde{h}_t)$$



Hybrid Neural Machine Translation

4.1 Word-based Translation as a Backbone

- deep LSTM encoder-decoder
- vocabulary size |V| 를 임의로 조정함으로써 word based model과 char based model을 얼마나 혼용할 지 정할 수 있다. (최상의 조합 찾기 가능)

4.2 Source character-based Representation

- rare word에 일괄적으로 <unk> token을 할당해버리면 유용한 정보 누락됨 -> character based!
- character-based LSTMs are always initialized with zero states
- character-based model과 word-based LSTM의 hidden state를 연결하면 어때? (기각)

  : 너무 복잡함, rare word의 일괄적인 precomputation 불가

  : 최종적으로 pretraining 없이 end-to-end training 가능함

Hybrid Neural Machine Translation

4.3 Target Character-level Generation

- 기존엔 post-preprocessing 단계를 통해 <unk> 로 나온 결괏값들을 대체해 주었음
- Hybrid model is trained such that whenever the word-level NMT produces an <unk>, we can consult this character-level decoder to recover the correct form of the unknown target word

$$J = J_w + \alpha J_c$$

Word-Character Generation Strategy

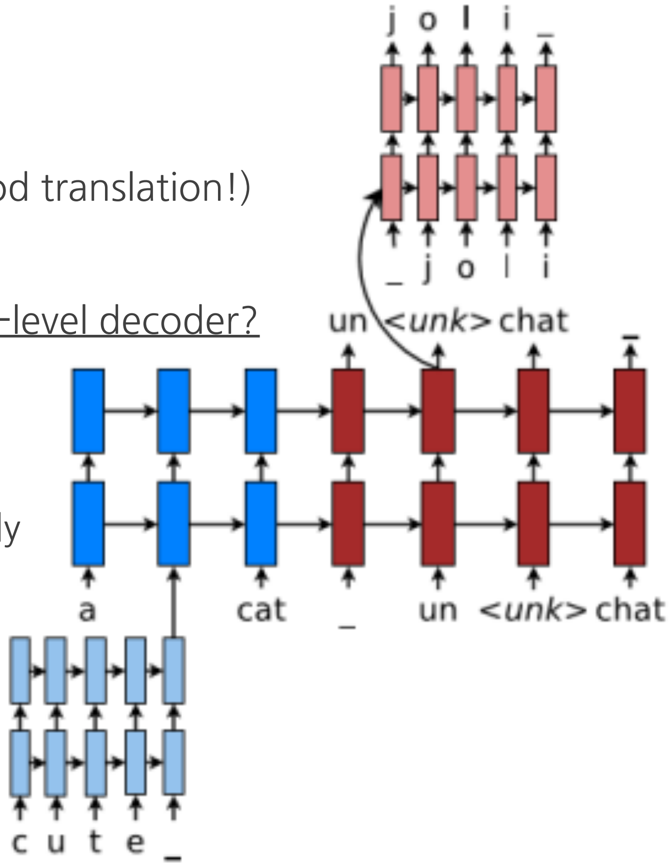
- character level decode의 마지막 hidden state를 그 단어의 representation이라고 생각하고 다음 time step으로 넘길 수 있지 않을까? -> 효용성 측면에서 기각!
- **training** : word-level NMT가 다 돌아가고 나면 char-level decoder가 실행한 모든 <unk> 객체의 실행을 잘 분리해냄 ∴ char-level decoder에서의 forward/backward pass가 batch 모드에서도 잘 호출됨
- **test** : beam search decoder @ word level -> find best translation (이 때 <unk> 포함)

  -> beam search decoder @ character level -> generate actual words for <unk>s

Hybrid Neural Machine Translation

Hidden-state Initialization

- source char-based representation : context-independent
- target char-based representation : context-dependent (for good translation!)



<u>What can best represent the current context to initialize character-level decoder?</u>

Candidate 1. *same-path* target generation approach

- final vector $\tilde{h}_t$
- all vectors $\tilde{h}_t$ might have similar value since it is directly used in softmax to predict same 〈unk〉 !

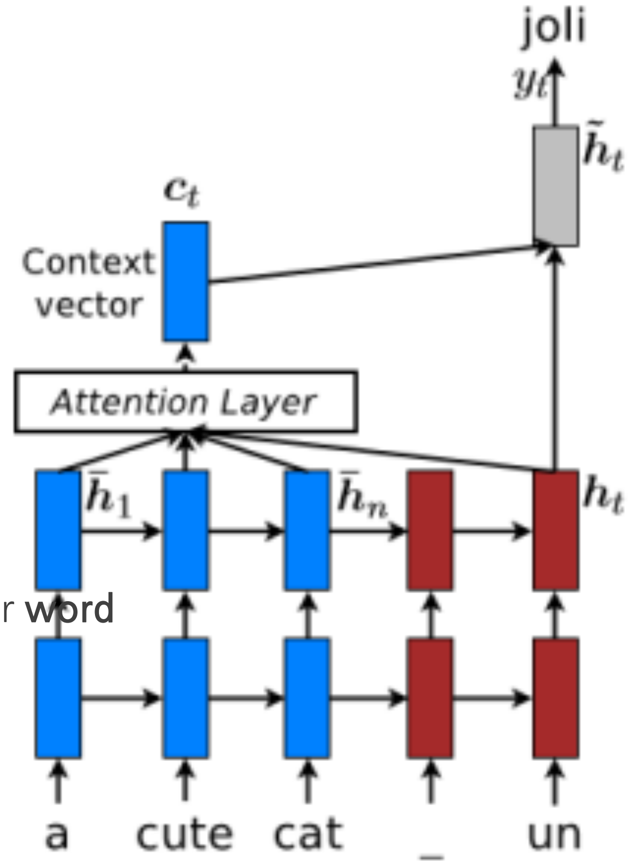
두 마리 토끼 (predicting 〈unk〉, generating character sequences)를 다 잡는 방법은 없을까??

Hybrid Neural Machine Translation

Hidden-state Initialization

Candidate 2. *separate-path* target generation approach

$$\breve{h}_t = \tanh(\breve{W}[c_t; h_t])$$



- computation in the character-level decoder is done per **word**
  not per **type** as in the source character component
  "A rose is a rose is a rose"
- context dependent nature of decoder

Experiments

English-Czech Results

- Train on WMT'15 data (12M sentence pairs)
  - newstest2015

Systems	BLEU
Winning WMT'15 (Bojar & Tamchyna, 2015)	18.8
Word-level NMT (Jean et al., 2015)	18.3
Hybrid NMT (Luong & Manning, 2016)*	**20.7**

30x data
3 systems

Large vocab
+ copy mechanism

Then SOTA!

But cf. Cherry et al. 2018: ~26 BLEU

Experiments

	English		Czech	
	word	char	word	char
# Sents	15.8M			
# Tokens	254M	1,269M	224M	1,347M
# Types	1,172K	2003	1,760K	2053
200-char	98.1%		98.8%	

Table 1: **WMT'15 English-Czech data** – shown are various statistics of our training data such as *sentence*, *token* (word and character counts), as well as *type* (sizes of the word and character vocabularies). We show in addition the amount of words in a vocabulary expressed by a list of 200 characters found in frequent words.

	System	Vocab	Perplexity w	Perplexity c	BLEU	$chrF_3$
(a)	Best WMT'15, big data (Bojar and Tamchyna, 2015)	-	-	-	*18.8*	-
	Existing NMT					
(b)	RNNsearch + unk replace (Jean et al., 2015b)	200K	-	-	15.7	-
(c)	*Ensemble* 4 models + unk replace (Jean et al., 2015b)	200K	-	-	18.3	-
	Our *word-based* NMT					
(d)	Base + attention + unk replace	50K	5.9	-	17.5	42.4
(e)	*Ensemble* 4 models + unk replace	50K	-	-	18.4	43.9
	Our *character-based* NMT					
(f)	Base-512 (600-step backprop)	200	-	2.4	3.8	25.9
(g)	Base-512 + attention (600-step backprop)	200	-	1.6	17.5	***46.6***
(h)	Base-1024 + attention (300-step backprop)	200	-	1.9	15.7	41.1
	Our *hybrid* NMT					
(i)	Base + attention + same-path	10K	4.9	1.7	14.1	37.2
(j)	Base + attention + separate-path	10K	4.9	1.7	15.6	39.6
(k)	Base + attention + separate-path + 2-layer char	10K	4.7	1.6	***17.7***	44.1
(l)	Base + attention + separate-path + 2-layer char	50K	5.7	1.6	19.6	46.5
(m)	*Ensemble* 4 models	50K	-	-	**20.7**	**47.5**

Analysis

6.1 Effects of Vocabulary Sizes

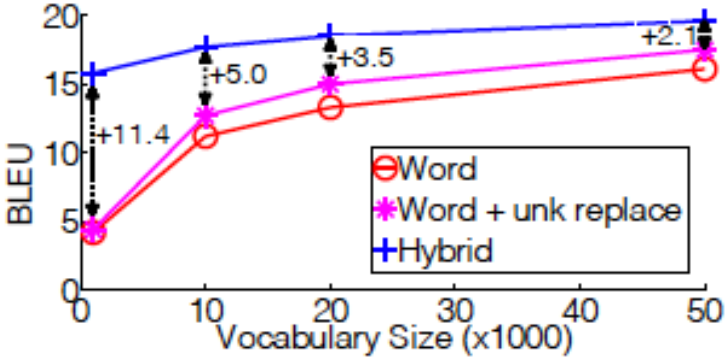


Figure 3: **Vocabulary size effect** – shown are the performances of different systems as we vary their vocabulary sizes. We highlight the improvements obtained by our hybrid models over word-based systems which already handle unknown words.

6.2 Rare Word Embeddings

System		Size	$\|V\|$	ρ
(Luong et al., 2013)		1B	138K	34.4
Glove (Pennington et al., 2014)		6B	400K	38.1
		42B	400K	**47.8**
Our NMT models				
(d)	Word-based	0.3B	50K	20.4
(k)	Hybrid	0.3B	10K	42.4
(l)	Hybrid	0.3B	50K	***47.1***

Table 3: **Word similarity task** – shown are Spearman's correlation ρ on the *Rare Word* dataset of various models (with different vocab sizes $|V|$).

Analysis

6.2 Rare Word Embeddings

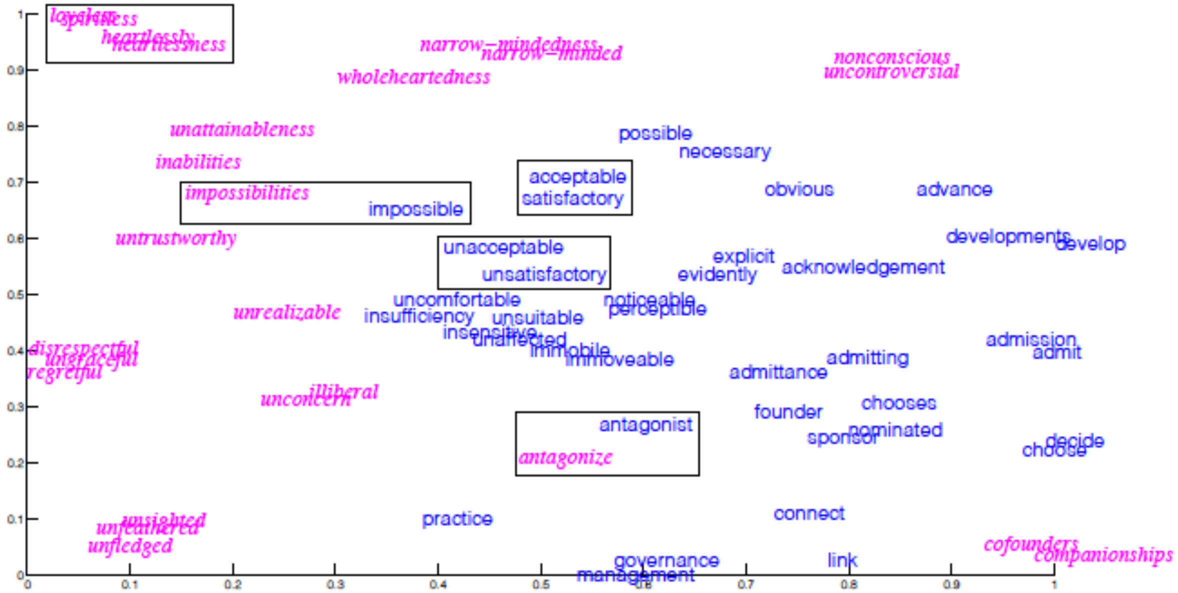


Figure 4: **Barnes-Hut-SNE visualization of source word representations** – shown are sample words from the *Rare Word* dataset. We differentiate two types of embeddings: frequent words in which encoder embeddings are looked up directly and *rare* words where we build representations from characters. Boxes highlight examples that we will discuss in the text. We use the hybrid model *(l)* in this visualization.

Analysis

6.3 Sample Translation

source	The author *Stephen Jay Gould* died 20 years after *diagnosis* .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

Analysis

6.3 Sample Translation

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její $\boxed{\textbf{jedenáctiletá}}$ dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani ,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její $\boxed{\textbf{jedenáctiletá}}$ dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

감사합니다

:-)