

Linguistics

Phonology

用音节划分

- Phonetics is the sound stream – uncontroversial “physics”
- Phonology posits a small set or sets of distinctive, categorical units: **phonemes** or distinctive features
 - A perhaps universal typology but language-particular realization
 - Best evidence of categorical perception comes from phonology
 - Within phoneme differences shrink; between phoneme magnified

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill				r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

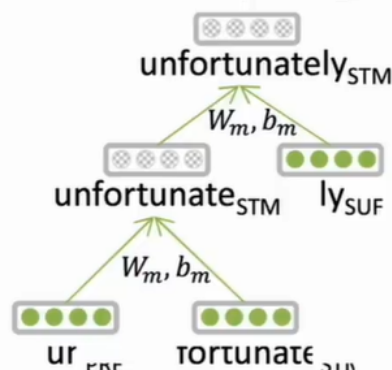
caught
cot

5

Morphology

用语素划分

- Traditionally, we have morphemes as smallest **semantic** unit
 - $[[[un \ [[fortun(e)]_{ROOT} \ ate]_{STEM}]_{STEM} \ ly]_{WORD}]$
- Deep learning: Morphology little studied; one attempt with recursive neural networks is (Luong, Socher, & Manning 2013)



A possible way of dealing with a larger vocabulary – most unseen words are new morphological forms (or numbers)

6

推广到character n-grams

Models below the word level

需要subword level的原因

- 有的word很长，可以进行拆分
- 音译
- 不常见词/人造词/缩写词，如Goooooooood, imma
-

Char-level Model

不同语言的character差别很大

Most deep learning NLP work begins with language in its written form – it's the easily processed, found data

But human language writing systems aren't one thing!

- Phonemic (maybe digraphs) jiyawu ngabulu
- Fossilized phonemic thorough failure
- Syllabic/moraic つゝくゝしゝゝ
- Ideographic (syllabic) 去年太空船二号坠毁
- Combination of the above インド洋の島

Purely character-level NMT models

English-Czech

第一个纯char-level的机器翻译，捷克语和英语，可以达到word-level的结果。

但是速度很慢，因为一个word平均有7个character。

English-Czech WMT 2015 Results

- Luong and Manning tested as a baseline a pure character-level seq2seq (LSTM) NMT system
- It worked well against word-level baseline
- But it was ssllooooww
 - 3 weeks to train ... not that fast at runtime

System	BLEU
Word-level model (single; large vocab; UNK replace)	15.7
Character-level model (single; 600-step backprop)	15.9

14

source	Her 11-year-old daughter , Shani Bart , said it felt a little bit weird
human	Její jedenáctiletá dcera Shani Bartová prozradila , že je to trochu zvláštní
char	Její jedenáctiletá dcera , Shani Bartová , říkala , že cítí trochu divně
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její 11-year-old dcera Shani , řekla , že je to trochu divné

从上图可以看出char-level很好地处理了

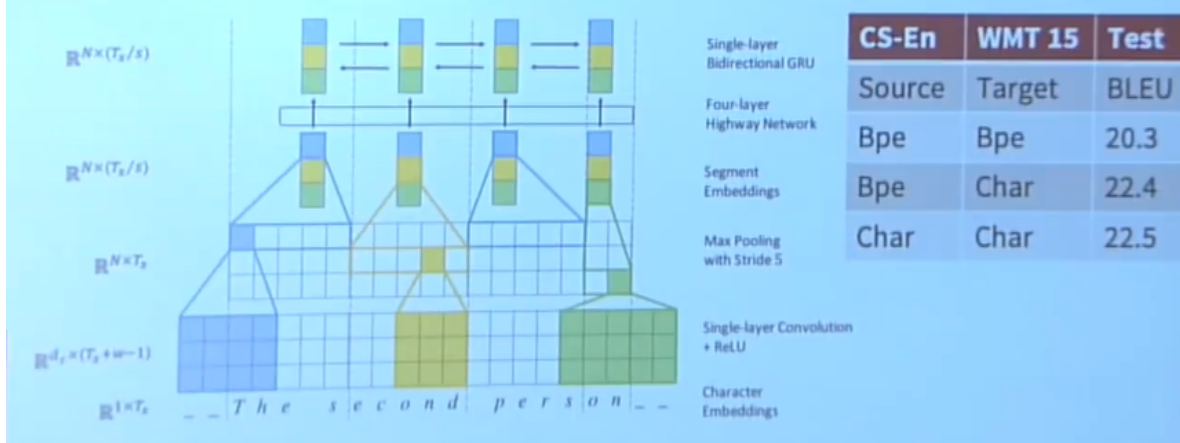
- 11-year-old 不在vocab中
- Shani Bart 需要音译

Fully char-level NMT without explicit segmentation

CNN + GRU

Fully Character-Level Neural Machine Translation without Explicit Segmentation

Jason Lee, Kyunghyun Cho, Thomas Hoffmann. 2017.
Encoder as below; decoder is a char-level GRU



Stronger char results with depth in LSTM seq2seq model

本文实验证明了简单的语言/模型适合word-level，复杂的语言/模型适合char-level。word-level时间消耗不随模型复杂度增加，char-level时间消耗随模型复杂度增加。

Word-piece model

word piece model: 和word-level结构一样，但是把word拆成各个组分

Byte Pair Encoding

BPE算法可以构造一个词典，包括所有char和常见char的n-gram，比如est/er等。

这个词典不会很大，因此速度很快。接下来就和word-level一样训练模型，把词典中的“词”当成word。

Byte Pair Encoding

- A **word segmentation** algorithm:
 - Start with a vocabulary of **characters**
 - Most frequent **ngram pairs** \mapsto a new **ngram**

Dictionary

5 lo w
2 lo w e r
6 n e w e s t
3 w i d e s t

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, e, s, e, s, t, lo

Add a pair (l, o) with freq 7

24

(Example from Sennrich)

Word-piece/Sentence-piece model

word-piece类似BPE

sentence-piece首先按n-gram words对句子进行分割

Wordpiece/Sentencepiece model

- Wordpiece model tokenizes inside words
- Sentencepiece model works from raw text
 - Whitespace is retained as special token (`_`) and grouped normally
 - You can reverse things at end by joining pieces and recoding them to spaces
- <https://github.com/google/sentencepiece>
- <https://arxiv.org/pdf/1804.10959.pdf>

27

BERT用的是word-piece model。

- BERT uses a variant of the wordpiece model
 - (Relatively) common words are in the vocabulary:
 - *at, fairfax, 1910s*
 - Other words are built from wordpieces:
 - *hypatia = h ##yp ##ati ##a*

Hybrid model

hybrid: 大部分用word-level, 少部分用char-level

Character-based LSTM

训练char-level的向量, 组成wordvec, 再输入LSTM做下游任务。

CNN + Highway Network

char-level的wordvec会在拼写上更接近, 也容易表示不在vocab中的词。

	In Vocabulary					
	<i>while</i>	<i>his</i>	<i>you</i>	<i>richard</i>	<i>trading</i>	
LSTM-Word	<i>although</i>	<i>your</i>	<i>conservatives</i>	<i>jonathan</i>	<i>advertised</i>	
	<i>letting</i>	<i>her</i>	<i>we</i>	<i>robert</i>	<i>advertising</i>	
	<i>though</i>	<i>my</i>	<i>guys</i>	<i>neil</i>	<i>turnover</i>	
	<i>minute</i>	<i>their</i>	<i>i</i>	<i>nancy</i>	<i>turnover</i>	
LSTM-Char (before highway)	<i>chile</i>	<i>this</i>	<i>your</i>	<i>hard</i>	<i>heading</i>	
	<i>whole</i>	<i>hhs</i>	<i>young</i>	<i>rich</i>	<i>training</i>	
	<i>meanwhile</i>	<i>is</i>	<i>four</i>	<i>richer</i>	<i>reading</i>	
	<i>white</i>	<i>has</i>	<i>youth</i>	<i>richter</i>	<i>leading</i>	
LSTM-Char (after highway)	<i>meanwhile</i>	<i>hhs</i>	<i>we</i>	<i>eduard</i>	<i>trade</i>	
	<i>whole</i>	<i>this</i>	<i>your</i>	<i>gerard</i>	<i>training</i>	
	<i>though</i>	<i>their</i>	<i>doug</i>	<i>edward</i>	<i>traded</i>	
	<i>nevertheless</i>	<i>your</i>	<i>i</i>	<i>carl</i>	<i>trader</i>	

Out-of-Vocabulary		
<i>computer-aided</i>	<i>misinformed</i>	<i>loooooook</i>
—	—	—
—	—	—
—	—	—
—	—	—
<i>computer-guided</i>	<i>informed</i>	<i>look</i>
<i>computerized</i>	<i>performed</i>	<i>cook</i>
<i>disk-drive</i>	<i>transformed</i>	<i>looks</i>
<i>computer</i>	<i>inform</i>	<i>shook</i>
<i>computer-guided</i>	<i>informed</i>	<i>look</i>
<i>computer-driven</i>	<i>performed</i>	<i>looks</i>
<i>computerized</i>	<i>outperformed</i>	<i>looked</i>
<i>computer</i>	<i>transformed</i>	<i>looking</i>

Hybrid NMT

Thang Luong and Chris Manning. **Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models**. ACL 2016.

核心思想是基本用word-level，在遇到不在vocab的词是用char-level。

FastText embeddings

- Represent word as char n -grams augmented with boundary symbols and as whole word:
- *where* = $\langle wh, whe, her, ere, re \rangle, \langle where \rangle$
 - Note that $\langle her \rangle$ or $\langle her$ is different from *her*
 - Prefix, suffixes and whole words are special
- Represent word as sum of these representations.
Word in context score is:
 - $s(w, c) = \sum_{g \in G(w)} \mathbf{z}_g^T \mathbf{v}_c$
 - Detail: rather than sharing representation for all n -grams, use “hashing trick” to have fixed number of vectors

55

FastText embeddings

Word similarity
dataset scores
(correlations)

		sg	cbow	sisg-	sisg
AR	WS353	51	52	54	55
	GUR350	61	62	64	70
DE	GUR65	78	78	81	81
	ZG222	35	38	41	44
EN	RW	43	43	46	47
	WS353	72	73	71	71
Es	WS353	57	58	58	59
FR	RG65	70	69	75	75
Ro	WS353	48	52	51	54
RU	HJ	59	60	60	66

56