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 g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		ⱱ		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.

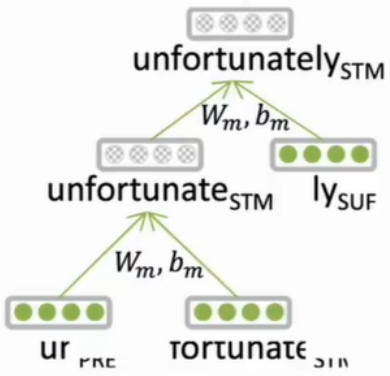
caught
cot

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

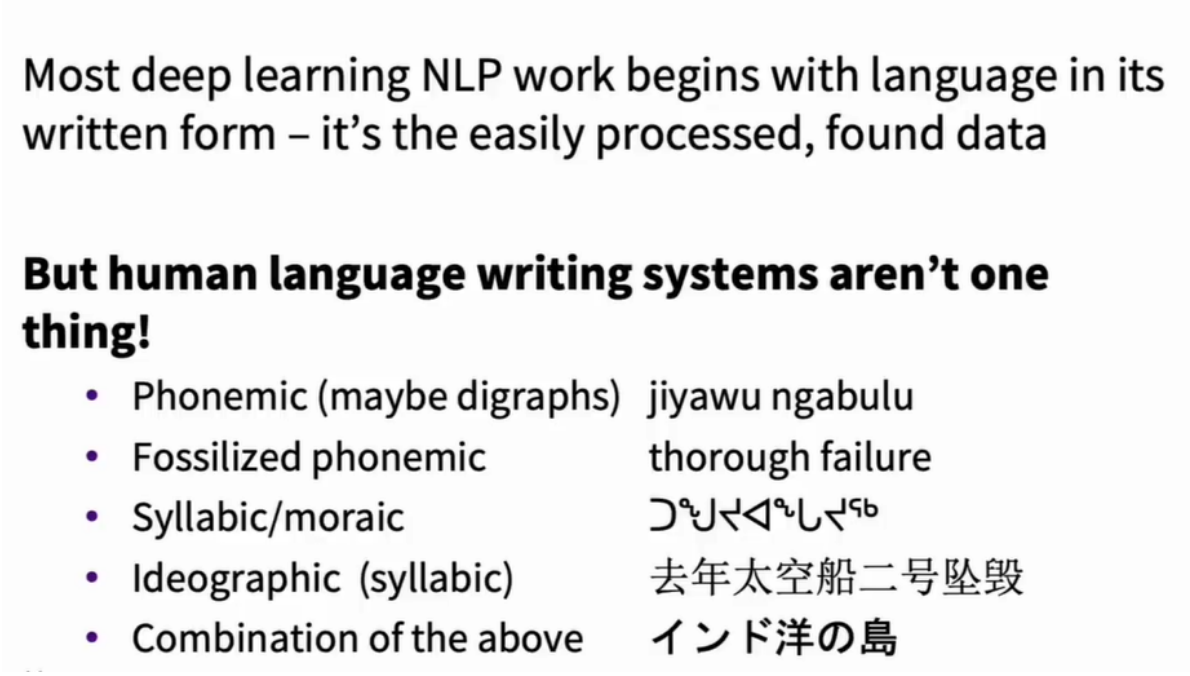
Models below the word level

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Goooooood, imma
-

Char-level Model

不同语言的character差别很大



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。

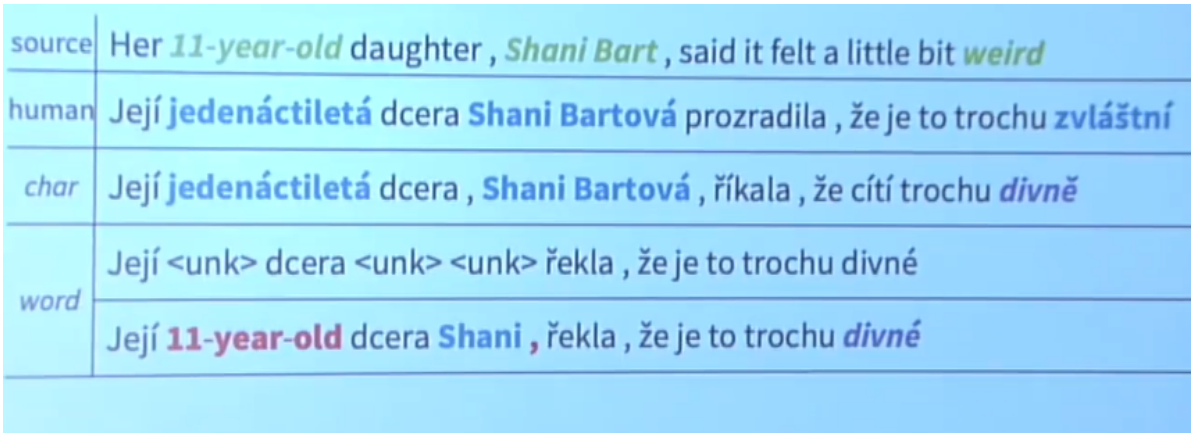
但是速度很慢，因为一个word平均有7个character。

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 sslloooooww
  - 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

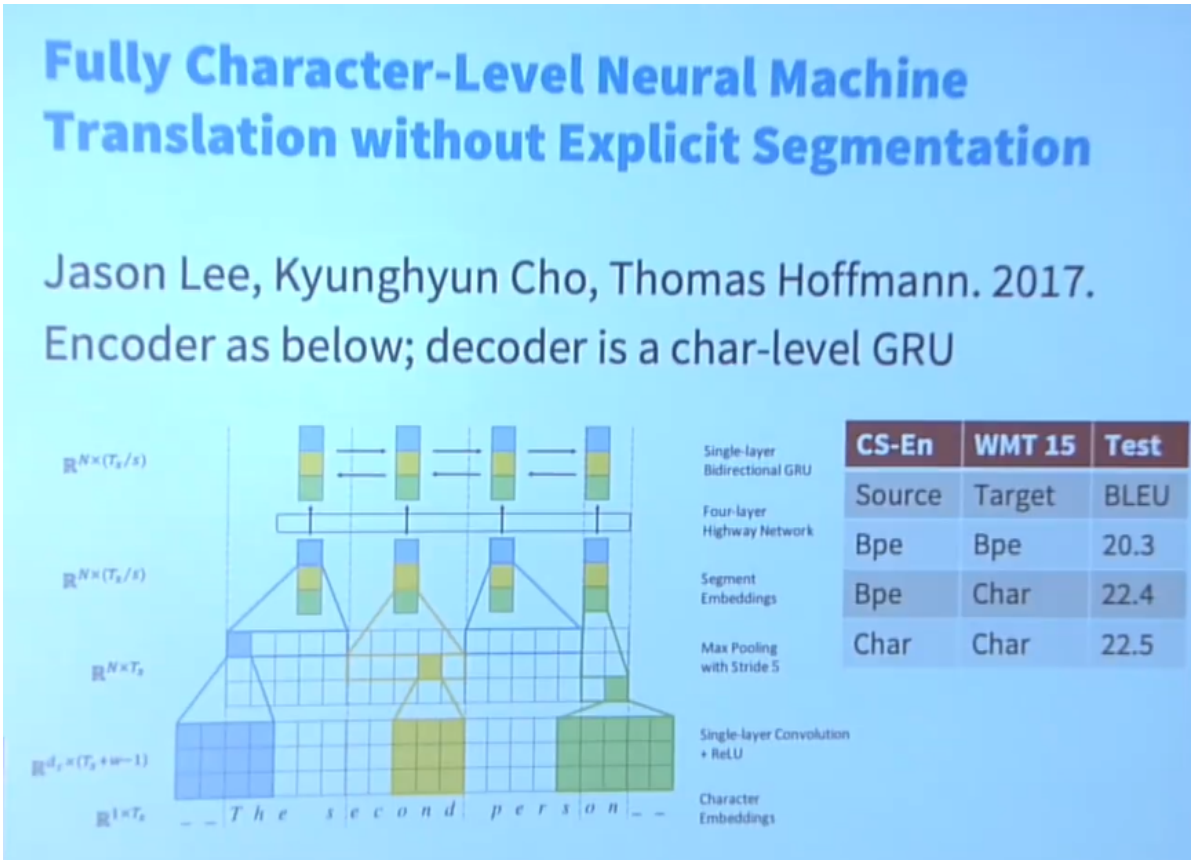


从上图可以看出char-level很好地处理了

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

Fully char-level NMT without explicit segmentation

CNN + GRU



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。

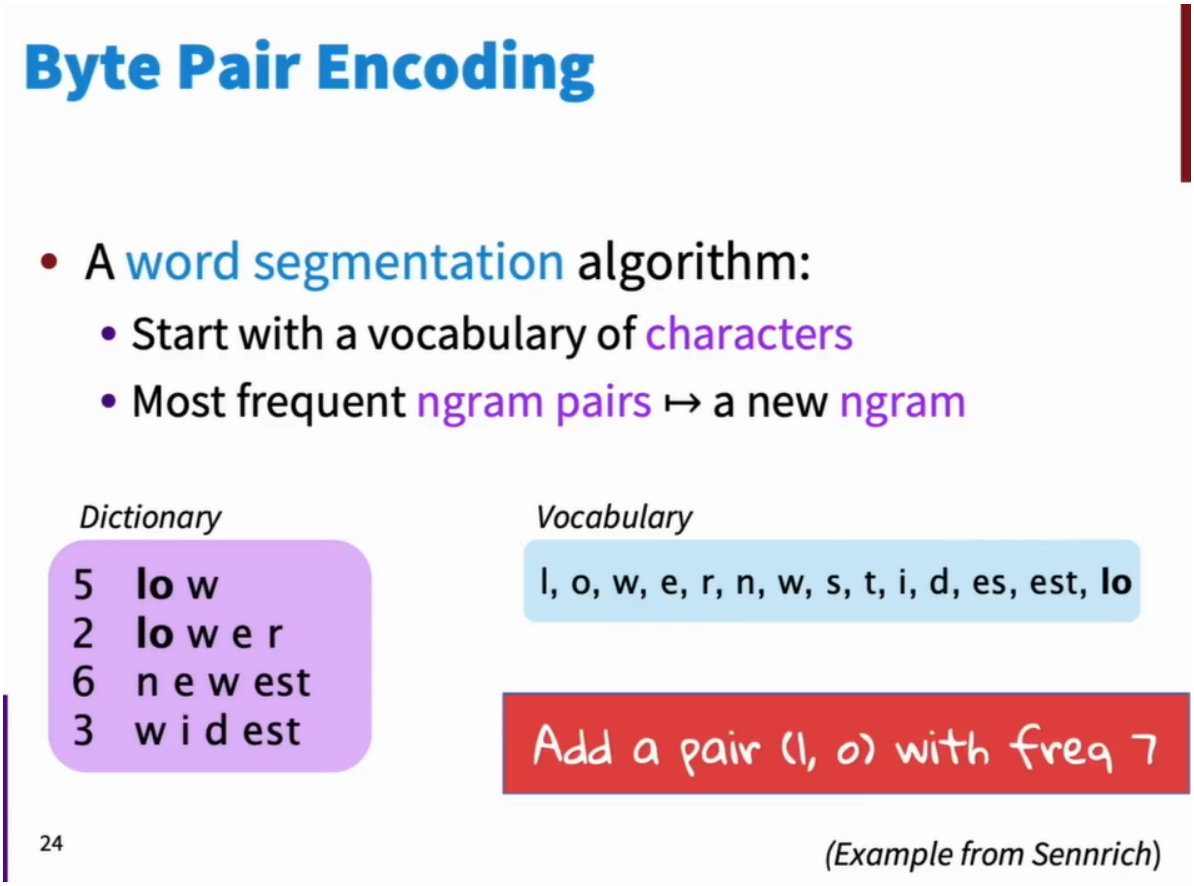
Word-piece model

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。

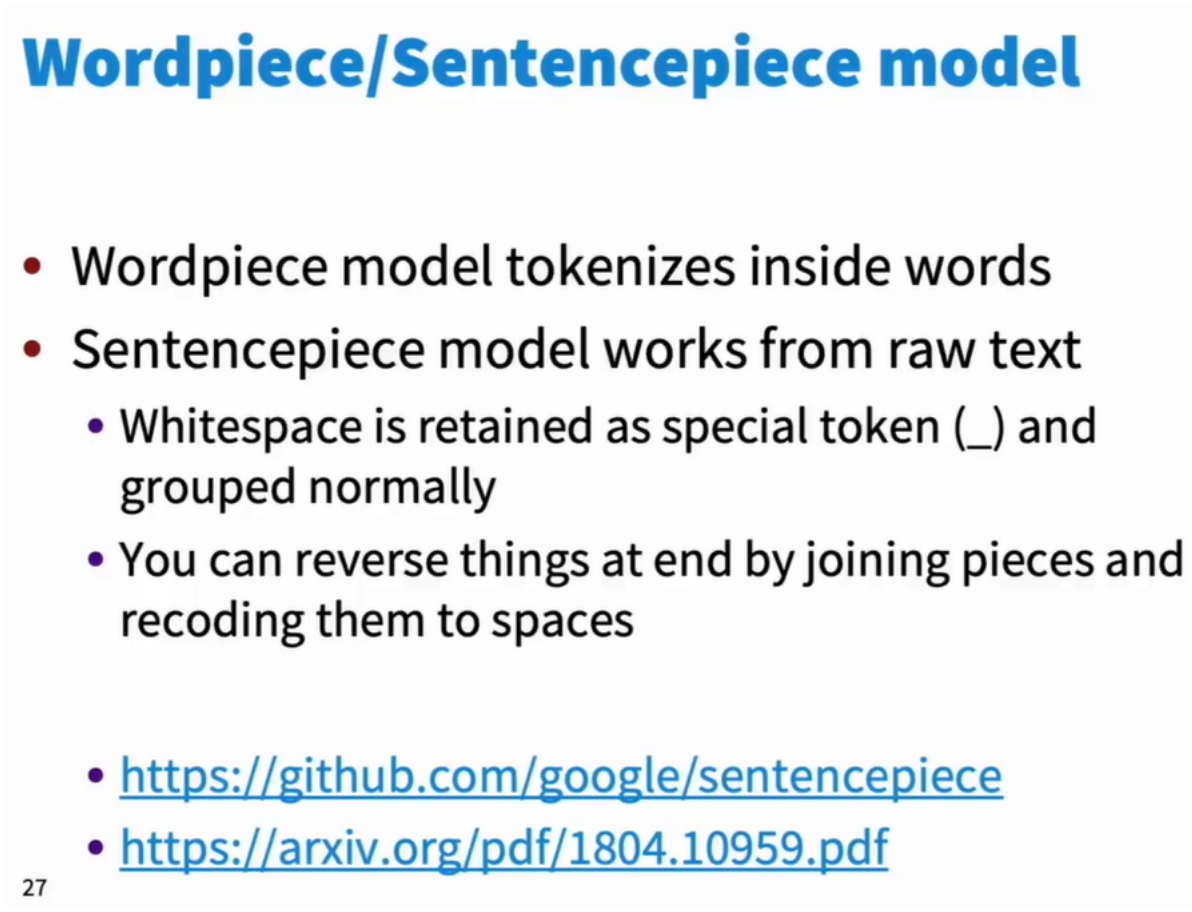
这个词典不会很大，因此速度很快。接下来就和word-level一样训练模型，把词典中的"词"当成word。



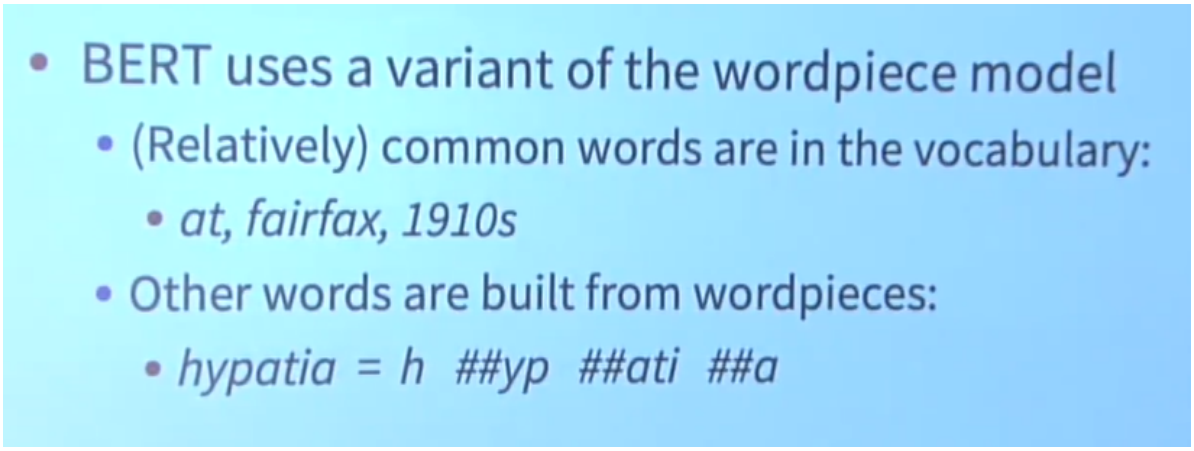
Word-piece/Sentence-piece model

word-piece类似BPE

sentence-piece首先按n-gram words对句子进行分割



BERT用的是word-piece model。



Hybrid model

hybrid: 大部分用word-level，少部分用char-level

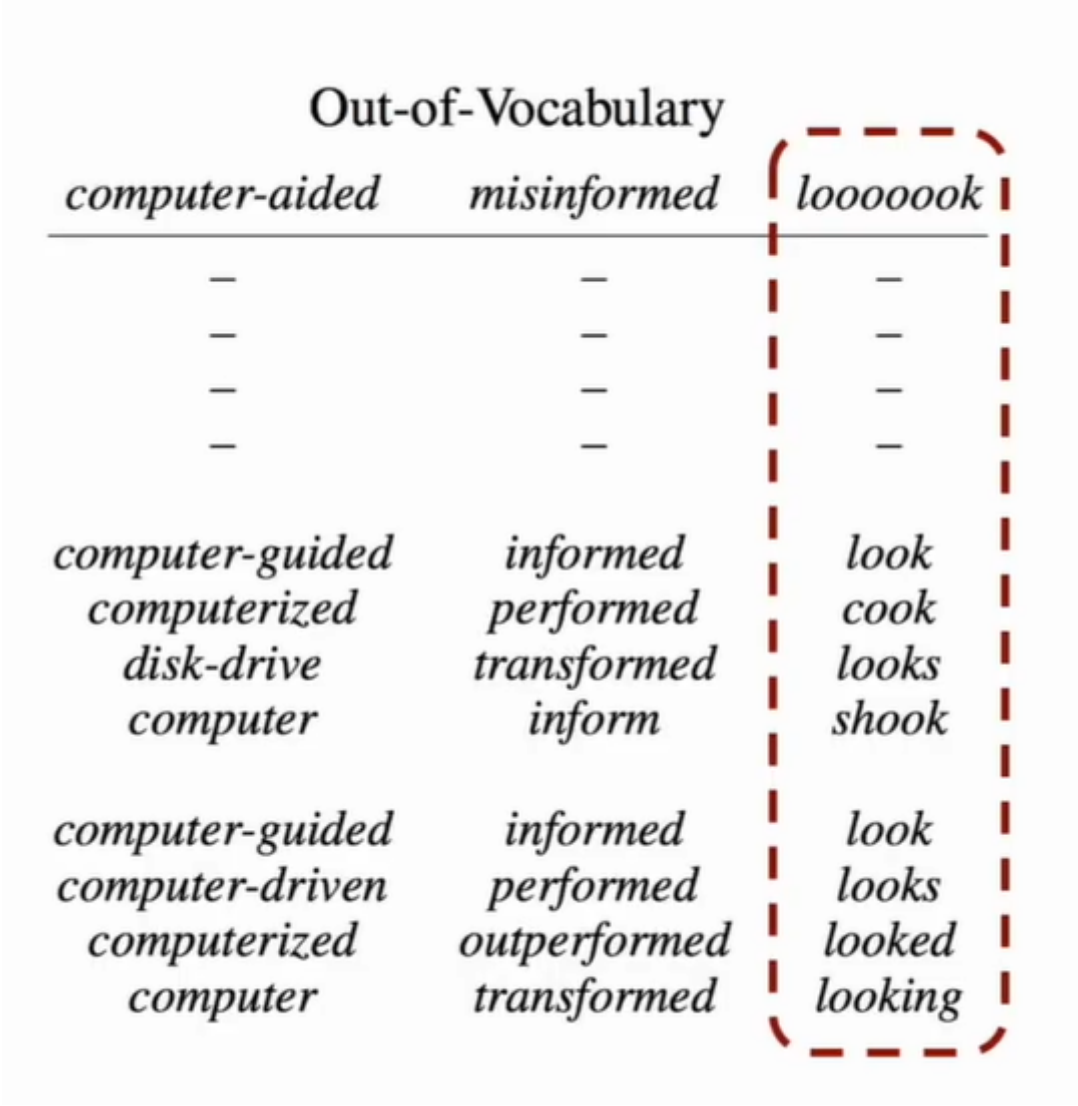
Character-based LSTM

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。

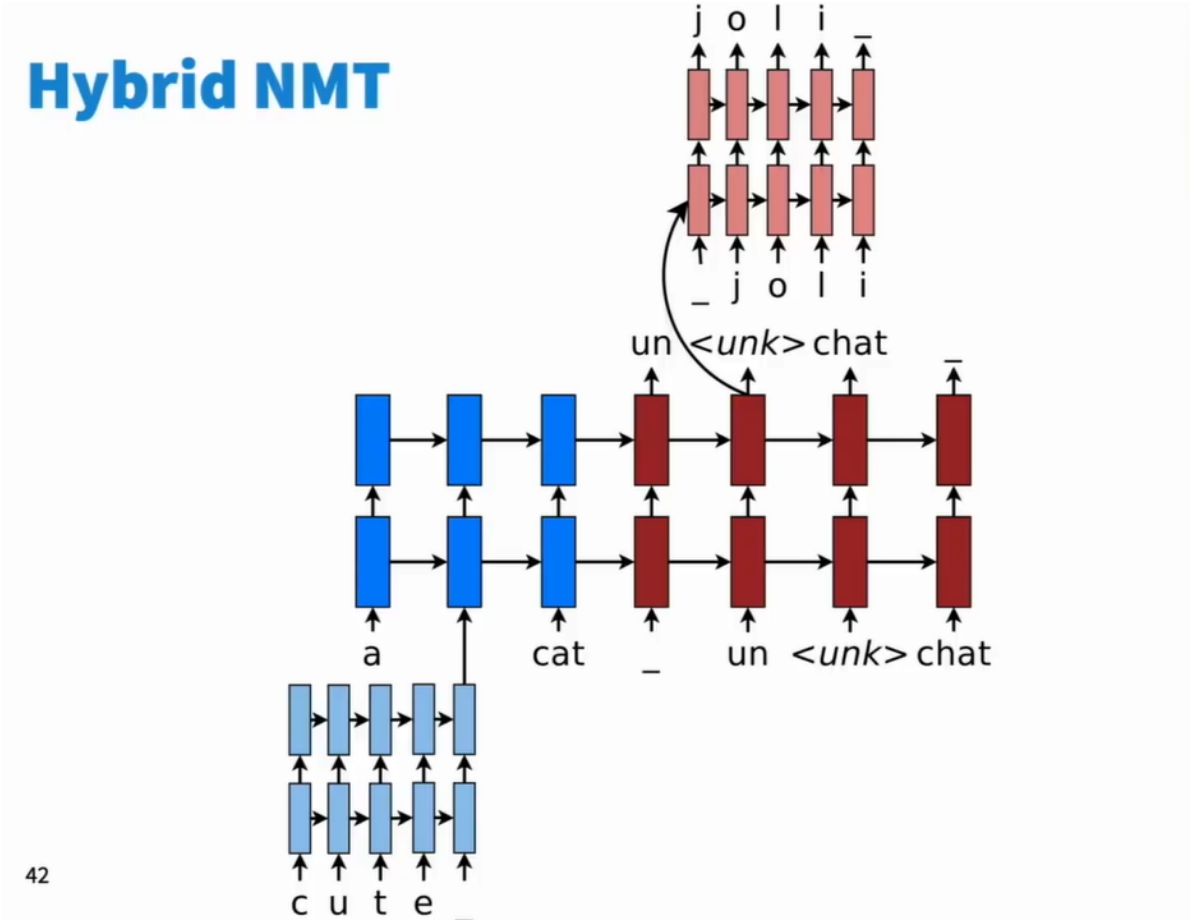
	In Vocabulary				
	while	*his*	*you*	*richard*	*trading*
LSTM-Word	*although*	*your*	*conservatives*	*jonathan*	*advertised*
	letting	*her*	*we*	*robert*	*advertising*
	though	*my*	*guys*	*neil*	*turnover*
	minute	*their*	*i*	*nancy*	*turnover*
LSTM-Char (before highway)	*chile*	*this*	*your*	*hard*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



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。



上图表示，正常走seq2seq，但是遇到预测出<unk>，就把Hidden unit输入进char的LSTM中生成char序列。

FastText embeddings

一种新的word embedding。

用传统的word2vec + char-level n-gram。

FastText embeddings

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

FastText embeddings

Word similarity dataset scores (correlations)

		sg	cbow	sisg-	sisg
AR	WS353	51	52	54	**55**
DE	GUR350	61	62	64	**70**
	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**