1314기 텍스트세미나

ToBig's 14기 장예은

Lecture 12_ subwords

Contents

잠깐 용어정리

- Phonetics: 음성학
- Phonology: 음운론
- Phonemes: 음소 (뜻 구별의 최소 단위)
- Morphology: 형태학
- Morpheme: 형태소
- Semantic: 의미론적인

Unit 01 | Purely character-level models

Character - level로 접근해야 하는 이유

1. Writing systems vary in how the represent the words

2. Need to handle large, open vocabulary

Unit 01 | Purely character-level models

1. 언어별로 특성이 상이함

- No word segmentation 安理会认可利比亚问题柏林峰会成果
- Words (mainly) segmented: *This is a sentence with words.*
  - Clitics/pronouns/agreement?
    - Separated **Je vous ai apporté** des bonbons
    - Joined ففلناها = ها +نا +قال +ف = so+said+we+it
  - Compounds?
    - Separated life insurance company employee
    - Joined Lebensversicherungsgesellschaftsangestellter

Unit 01 | Purely character-level models

2. Need to handle large, open vocabulary

- Rich morphology: nejneobhospodařovávatelnějšímu
  ("to the worst farmable one")
- Transliteration: Christopher ↦ Kryštof
- Informal spelling:

Unit 01 | Purely character-level models

Purely character – level models

- 영어 – 체코어 번역
- Pure character-level seq2seq system (2015)
- 하지만 학습시간이 너무 느렸고(3주), BLEU 15.9 의 성능에 불과함

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***

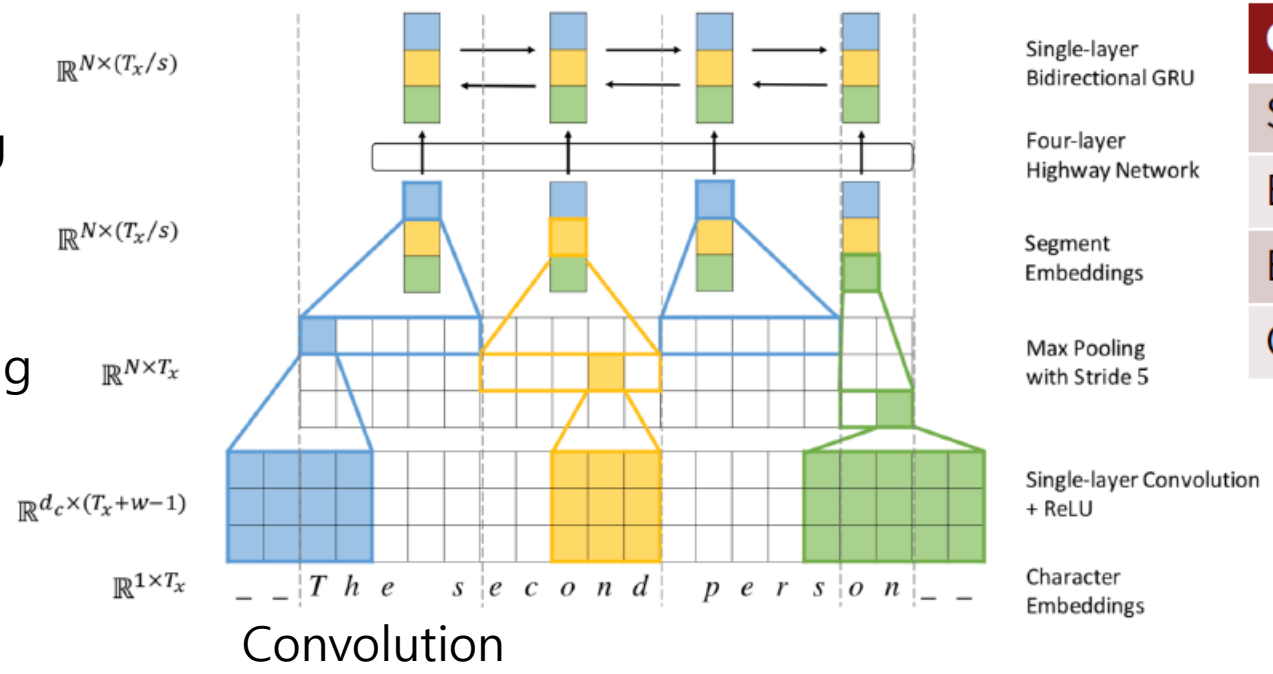
Word level 모델에 비해서 사람 이름을 잘 번역하는 경향

Unit 01 | Purely character-level models

- 체코어-영어 번역
- Fully character-level neural machine translation without explicit segmentation(2017)
- https://arxiv.org/abs/1610.03017

디코더로 single layer GRU

Max pooling



Convolution

Cs-En	WMT 15	Test
Source	Target	BLEU
Bpe	Bpe	20.3
Bpe	Char	22.4
Char	Char	22.5

앞의 모델보다 더 나아진 성능

Unit 01 | Purely character-level models

4.2 Attention and Decoder

Similarly to the attention model in (Chung et al., 2016; Firat et al., 2016a), a single-layer feedforward network computes the attention score of next target character to be generated with every source segment representation. A standard two-layer character-level decoder then takes the source context vector from the attention mechanism and predicts each target character. This decoder was described as *base decoder* by Chung et al. (2016).

Bilingual	bpe2char	char2char
Vocab size	24,440	300
Source emb.	512	128
Target emb.	512	512
Conv. filters		200-200-250-250 -300-300-300-300
Pool stride		5
Highway		4 layers
Encoder	1-layer 512 GRUs	
Decoder	2-layer 1024 GRUs	

Table 1: Bilingual model architectures. The char2char model uses 200 filters of width 1, 200 filters of width 2, ··· and 300 filters of width 8.

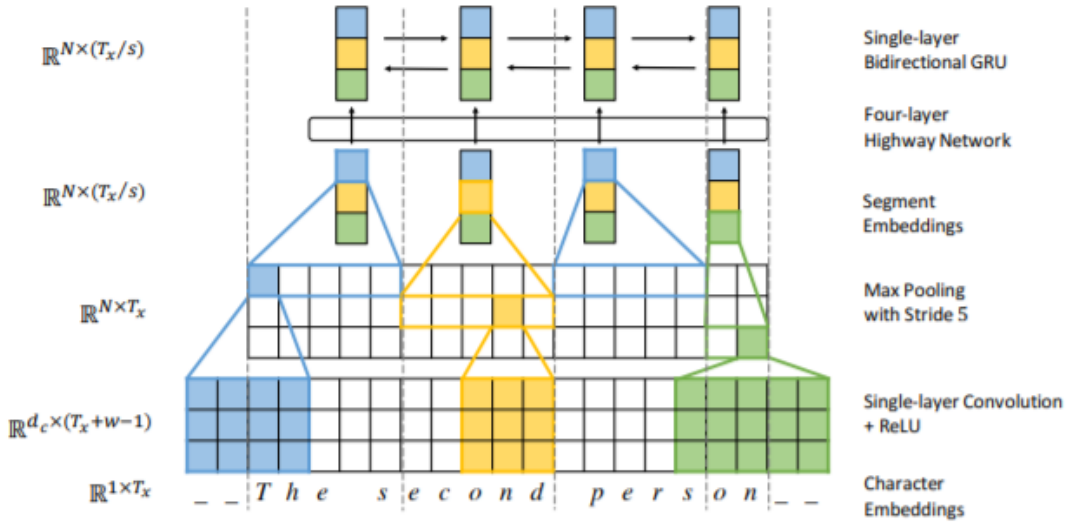
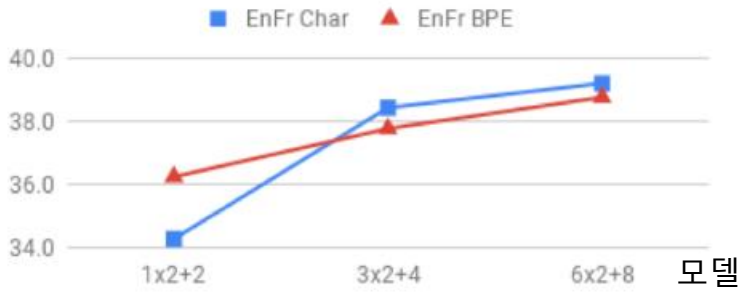


Figure 1: Encoder architecture schematics. Underscore denotes padding. A dotted vertical line delimits each segment. The stride of pooling s is 5 in the diagram.

Unit 01 | Purely character-level models

Seq2seq 모델/BPE 모델 성능 언어별 비교

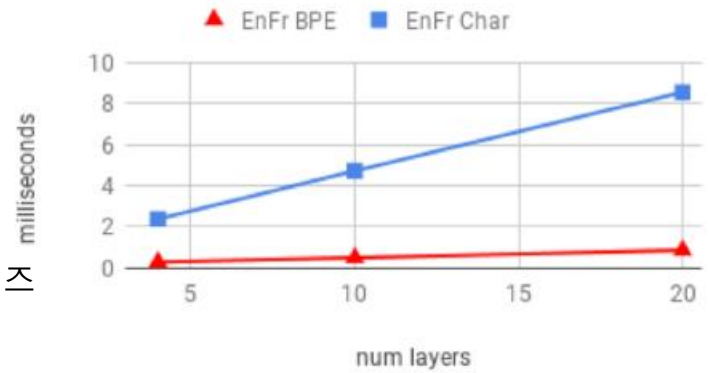
영어 -> 프랑스어



모델 사이즈 (layer)

체코어 -> 영어

BPE: Sub-word모델
(word pieces 이용)



Char 모델의
연산량이 더 많음

영어 -> 프랑스어 번역에서는 character based와 word based가 큰 차이 없지만,
체코어 -> 영어 번역에서는 char based가 훨씬 우수
(언어의 특성에 따라 효과 다름)

Unit 02 | Subword models

1. BPE

- word level 모델과 동일하지만, 더 작은 word인 word pieces를 이용함
- 딥러닝(neural network)과 전혀 무관한 간단한 아이디어
- 문서를 알집으로 압축하는것, 단축번호, 텍스트대치(?)와 유사
- 자주 나오는 byte pair(n gram)를 새로운 byte(a new gram)로 clustering

Dictionary

5 **lo** w
2 **lo** w e r
6 n e w est
3 w i d est

Vocabulary

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

- 자주 등장하는 es, est, lo를 새로운 단어로 추가(clustering)
- 새로 추가된 단어도 하나의 단어처럼 취급
- 지정한 최대 길이 넘으면 clustering 멈춤
- 성능 효과적(널리 응용됨)
- 시스템의 vocabulary를 자동적으로 결정

word piece 보충설명

참고링크 : https://wikidocs.net/22592, https://lovit.github.io/nlp/2018/04/02/wpm/

- **WordPiece Model**은 BPE의 변형 알고리즘. BPE가 빈도수에 기반하여 가장 많이 등장한 쌍을 병합하는 것과는 달리, 병합되었을 때 코퍼스의 우도(Likelihood)를 가장 높이는 쌍을 병합
- 자주 등장하는 piece는 unit으로 묶음 (BPE와 유사), 자주 등장하지 않는 것은 분리

  WPM을 수행하기 이전의 문장: Jet makers feud over seat width with big orders at stake
  WPM을 수행한 결과(wordpieces): _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

- Jet는 J와 et로 나누어졌으며, feud는 fe와 ud로 나누어짐.
- WPM은 모든 단어의 맨 앞에 _를 붙이고, 단어는 서브 워드(subword)로 통계에 기반하여 띄어쓰기로 분리함.(_는 문장 복원을 위한 장치)
- Jet → _J et : 띄어쓰기가 서브 워드(subwords)들을 구분, 기존 띄어쓰기는 _
- WPM이 수행 결과로부터 수행 전의 결과로 돌리는 방법은 현재 있는 모든 띄어쓰기를 전부 제거하고, 언더바를 띄어쓰기로 바꾸는 것

- **Sentencepiece**
- 구글에서 2018년 공개한 비지도학습 기반 형태소 분석 패키지
- 사전 토큰화 작업없이 문장 단위 input의 단어 분리 토큰화를 수행하는 단어 분리 패키지

```
I didn't at all think of it this way.
['_I', '_didn', "'", 't', '_at', '_all', '_think', '_of', '_it', '_this', '_way', '.']
[41, 623, 4950, 4926, 138, 169, 378, 30, 58, 73, 413, 4945]

I have waited a long time for someone to film
['_I', '_have', '_wa', 'ited', '_a', '_long', '_time', '_for', '_someone', '_to', '_film']
[41, 141, 1364, 1120, 4, 666, 285, 92, 1078, 33, 91]
```

Unit 02 | Subword models

다른 word piece/ sentence piece 모델

- **BERT**: vocab size가 크지만, 엄청 크지는 않음 => wordpiece 사용 필요
- 상대적으로 등장 빈도가 높은 단어들 + wordpiece를 이용
- Ex) 사전에 없는 Hypatia = h ##yp ##ati #a 라는 wordpiece로 쪼개짐,

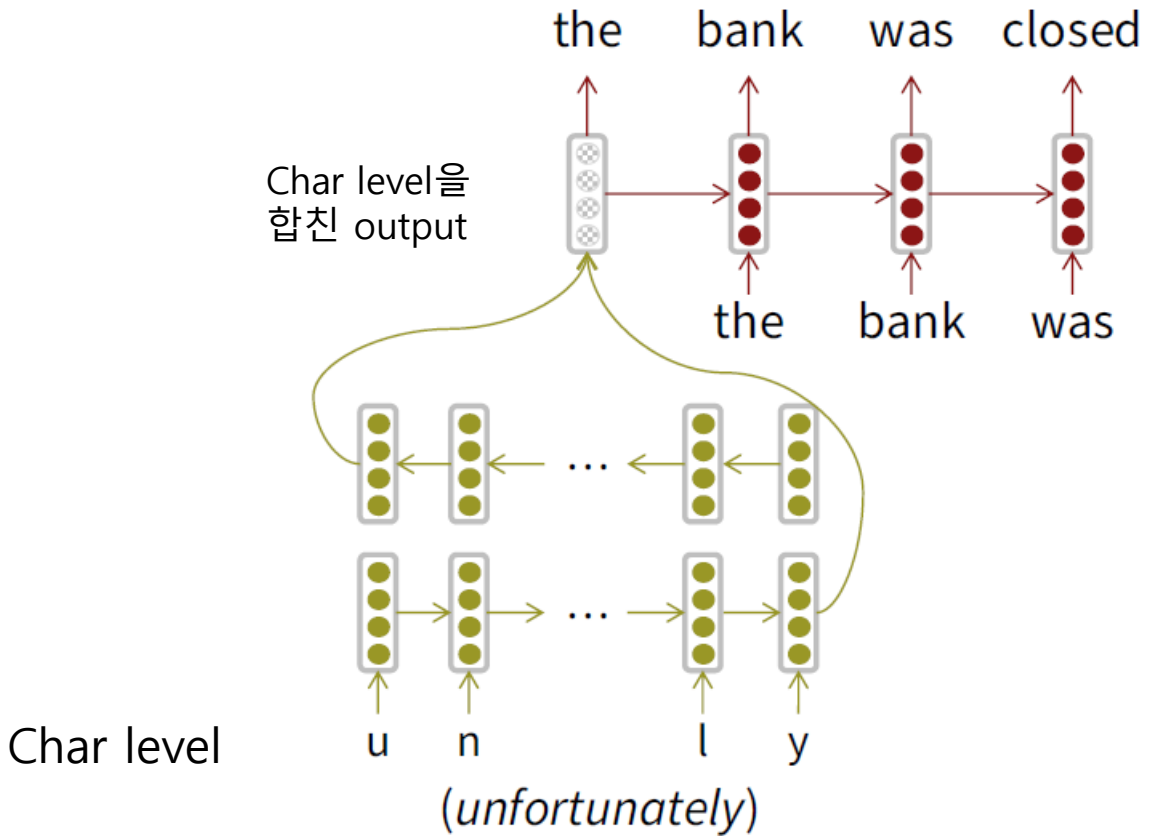
  4개의 word vector pieces

Unit 03 | Hybrid models

Hybrid models : 기본적으로 word 단위로 취급, 몇몇만 character 단위로 취급 (ex-사전에 없는 단어, 이름)

Character-based LSTM

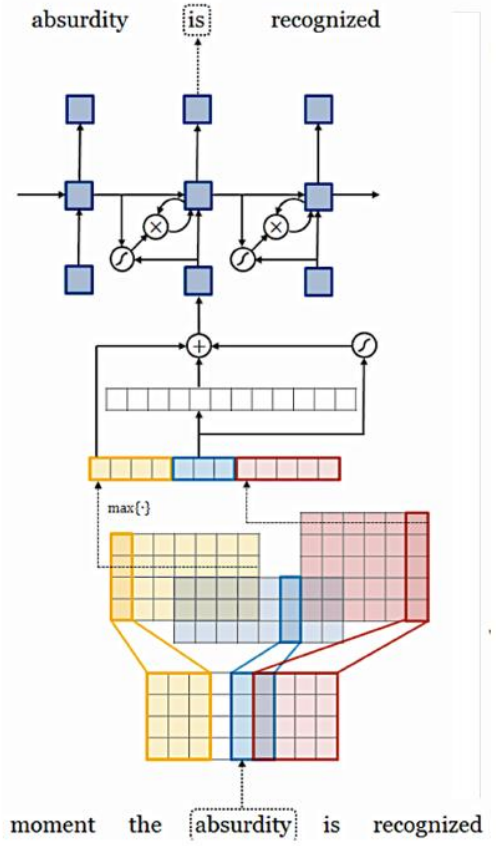
- Character level을 합친 output을 더 높은 레벨의 모델의 input으로

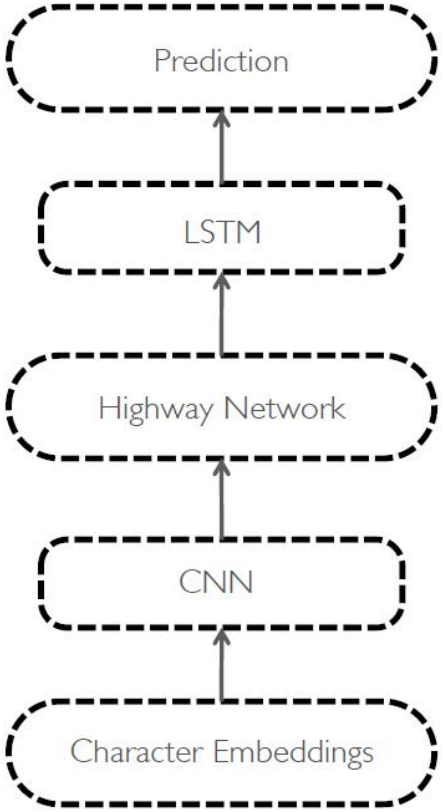


Unit 03 | Hybrid models

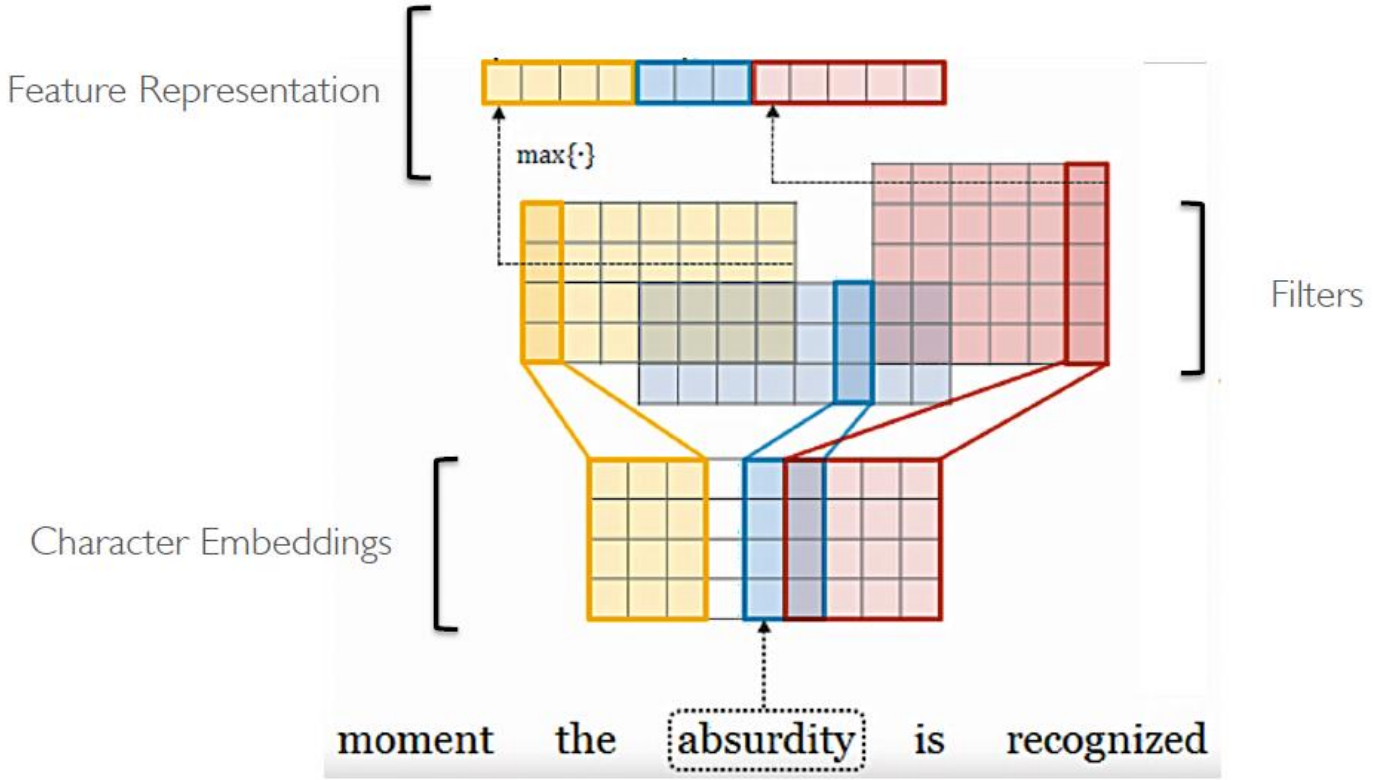
Character-Aware Neural Language Models

- Subword 관계성을 인코딩
- Ex) eventful, eventfully, uneventful
- 다른 모델이 가진 rare-word problem을 해결함
- 더 적은 파라미터 수로 비슷한 성능을 냄





Unit 03 | Hybrid models

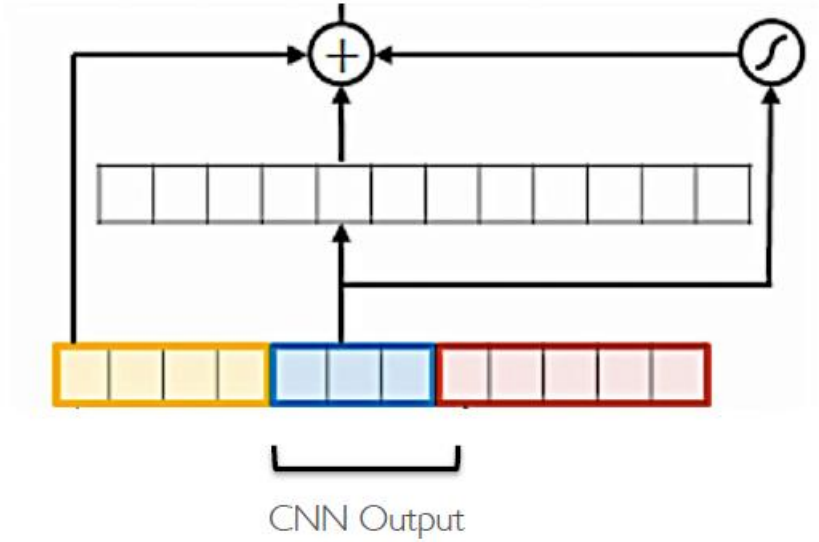


- Convolutions over character-level inputs.
- Max-over-time pooling (effectively n-gram selection).

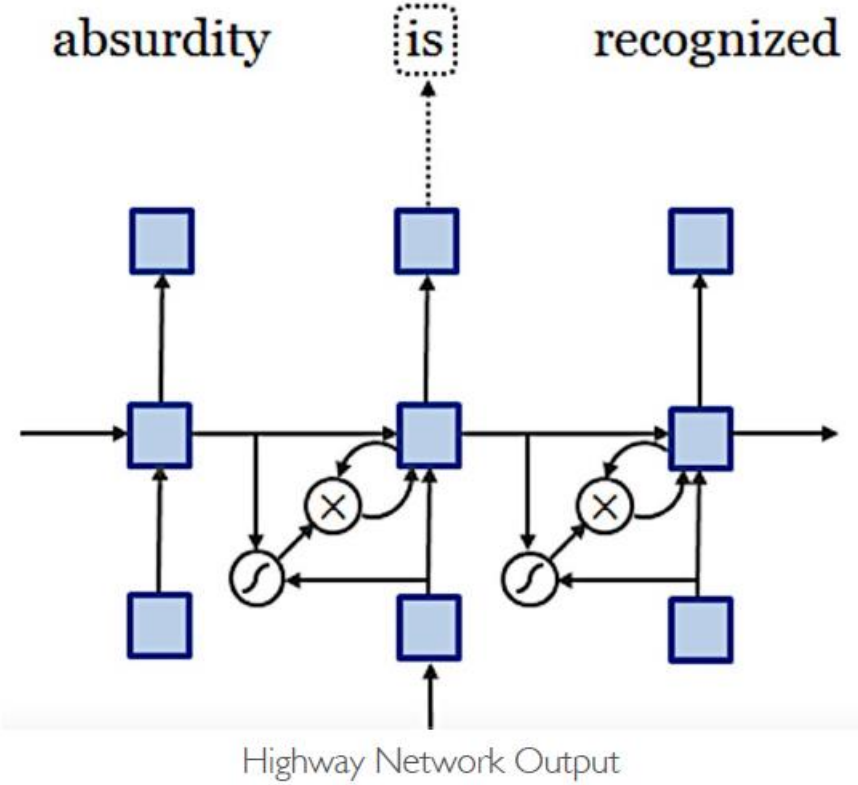
- Char 단위로 구분한 상태에서 시작
- Convolutional layer를 거쳐

  feature representation

Unit 03 | Hybrid models



- Highway Network
- LSTM과 유사한 기능



- Word-level LSTM
- 최종 출력층

Unit 03 | Hybrid models

		DATA-S					
		CS	DE	ES	FR	RU	AR
Botha	KN-4	545	366	241	274	396	323
	MLBL	465	296	200	225	304	–
Small	Word	503	305	212	229	352	216
	Morph	414	278	197	216	290	230
	Char	401	260	182	189	278	196
Large	Word	493	286	200	222	357	172
	Morph	398	263	177	196	271	**148**
	Char	**371**	**239**	**165**	**184**	**261**	**148**

		DATA-L					
		CS	DE	ES	FR	RU	EN
Botha	KN-4	862	463	219	243	390	291
	MLBL	643	404	203	227	**300**	273
Small	Word	701	347	186	202	353	236
	Morph	615	331	189	209	331	233
	Char	**578**	**305**	**169**	**190**	313	**216**

Comparable performance with fewer parameters!

	PPL	Size
LSTM-Word-Small	97.6	5 m
LSTM-Char-Small	92.3	5 m
LSTM-Word-Large	85.4	20 m
LSTM-Char-Large	78.9	19 m
KN-5 (Mikolov et al. 2012)	141.2	2 m
RNN† (Mikolov et al. 2012)	124.7	6 m
RNN-LDA† (Mikolov et al. 2012)	113.7	7 m
genCNN† (Wang et al. 2015)	116.4	8 m
FOFE-FNNLM† (Zhang et al. 2015)	108.0	6 m
Deep RNN (Pascanu et al. 2013)	107.5	6 m
Sum-Prod Net† (Cheng et al. 2014)	100.0	5 m
LSTM-1† (Zaremba et al. 2014)	82.7	20 m
LSTM-2† (Zaremba et al. 2014)	78.4	52 m

Unit 03 | Hybrid models

	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*

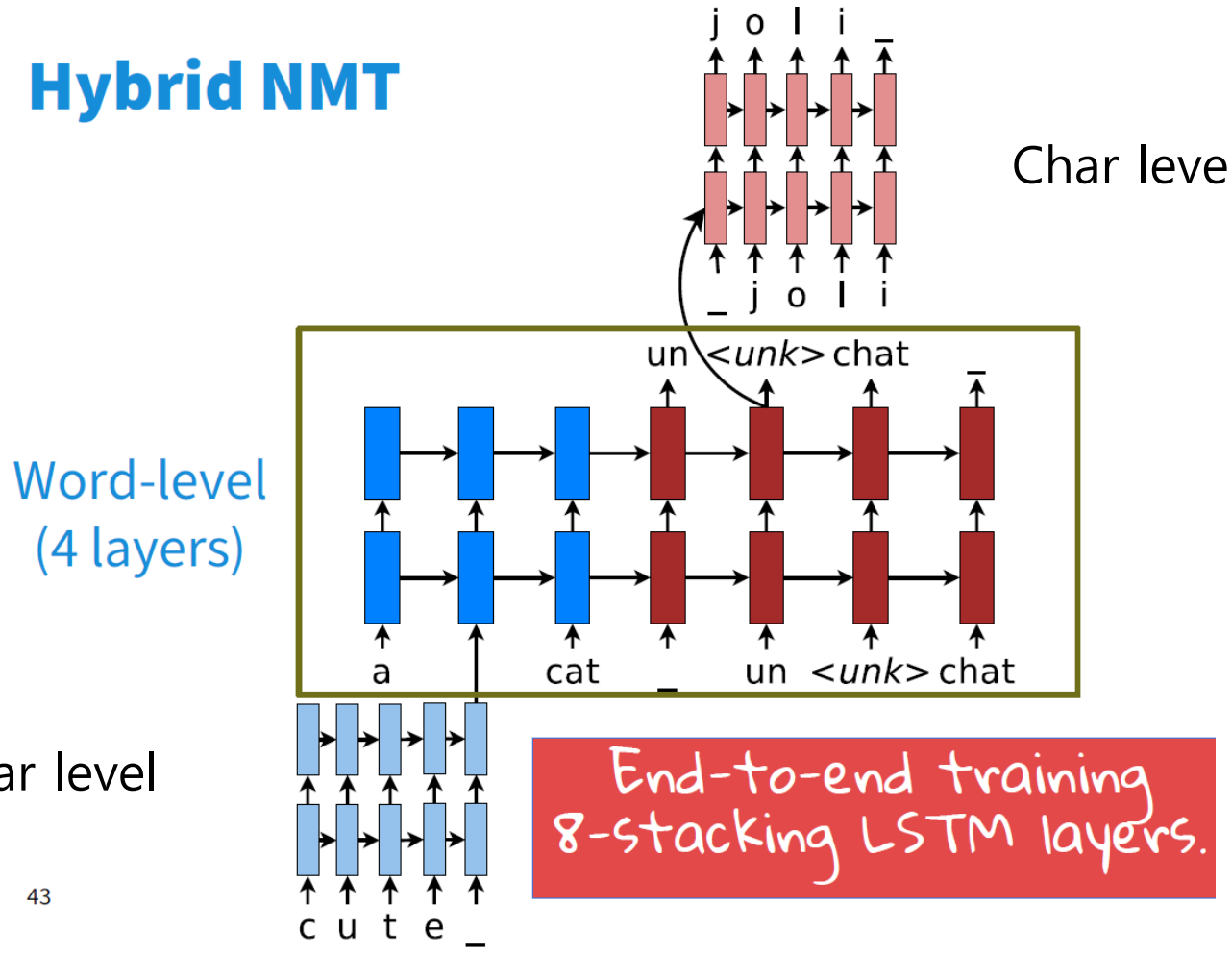
- ->가장 유사한 단어들 출력 결과

- Richard(사람 이름) 결과를 보면
- Highway block부분 이전에는
- Richard와 의미가 아닌 철자만 유사한 단어들이 가장 유사하다고 출력
- Highway block이후에는 의미를 고려하여 다른 사람 이름을 출력

Unit 03 | Hybrid models

Hybrid NMT

- 대부분 word level 사용
- 필요할때만 character level



Unit 03 | Hybrid models

Hybrid 모델의 우수성

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** .

Word 기반 모델:
Diagnosis라는 단어를
아예 잃어버림

Hybrid 모델은
완벽하게 해석

Unit 03 | Hybrid models

참고) FastText Embeddings

- 차세대 word2vec (word vector learning library
- 한 단어의 n-gram과 원래의 단어를 모두 학습에 사용

where = <*wh*, *whe*, *her*, *ere*, *re*>, <*where*>

- Note that <*her*> or <*her* is different from *her*
  - Prefix, suffixes and whole words are special

https://arxiv.org/pdf/1607.04606.pdf

들어주셔서 감사합니다.