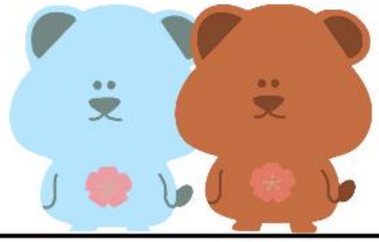
Week13_Information from parts of words : Subword Models

발표자: 임세영, 황채원

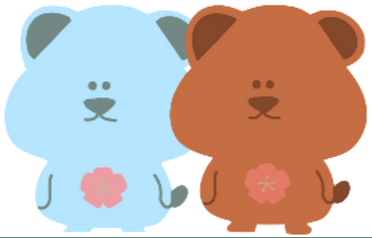
EWHA EURON

목차



EWHA EURON

Linguistic Knowledge



#01 Linguistic knowledge

Phonology : 음운론. 의미를 가지는 소리의 구조와 기능에 관한 연구

Phonetics : 음성학. 소리의 물리적 특성과 문법적 특성에 관한 연구

음운론 vs. 음성학

: 음성학에서는 모든 음성적 자질을 동등하게 중시하지만,
음운론은 의미를 변별하는 기능을 가진 변별적 자질을 중시하여
모든 음성적 자질을 동등하게 취급하는 것은 아니다.

Phonemes : 음소. 뜻 구별의 최소 단위

Morphology : 형태론. 단어의 어형 변화를 연구하는 문법의 한 분야

Morphemes : 형태소. $[[un\ [[fortun(e)\ ]_{ROOT}\ ate]_{STEM}]_{STEM}\ ly]_{WORD}$

EWHA EURON

#01 Linguistic knowledge

Words in writing systems - 언어별로 상이한 특성들

- 띄어쓰기가 없는 언어
- 띄어쓰기가 있는 언어

  -> 합성어의 띄어쓰기 문제

  -> 발음하는 대로 적는 언어, 의미대로 적는 언어

美国关岛国际机场及其办公室均接获

Je vous ai apporté des bonbons

فقلناها = ها +نا +قال +ف = so+said+we+it

life insurance company employee

Lebensversicherungsgesellschaftsangestell

#01 Linguistic knowledge

Models below the word level – 커버해야 할 단어 수가 너무 많다

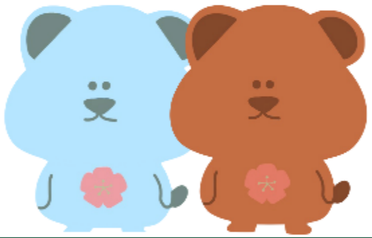
- Rich morphology

  nejneobhospodařovávatelnějšímu

  (“to the worst farmable one”)

- Informal spelling

Purely character-level models



#02 Purely character-level models

Pure character-level seq2seq system (2015)

- 영어 – 체코어 번역
- Word-level에 비해 우수한 성능 – 특히 사람 이름 번역에서 두각을 보임

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***

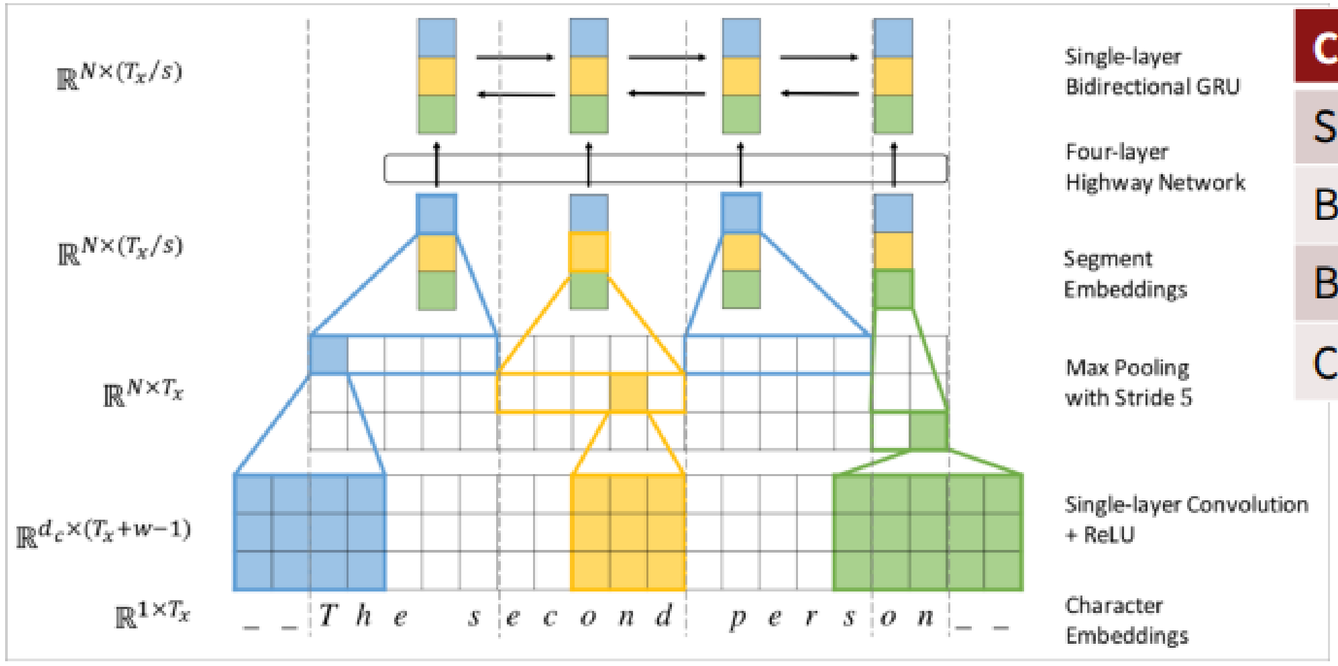
System	BLEU
Word-level model (single; large vocab; UNK replace)	15.7
Character-level model (single; 600-step backprop)	15.9

- 학습시간이 너무 느리다는 치명적인 단점 - 3주 이상 소요

#02 Purely character-level models

Fully Char-level Natural Machine Translation (2017)

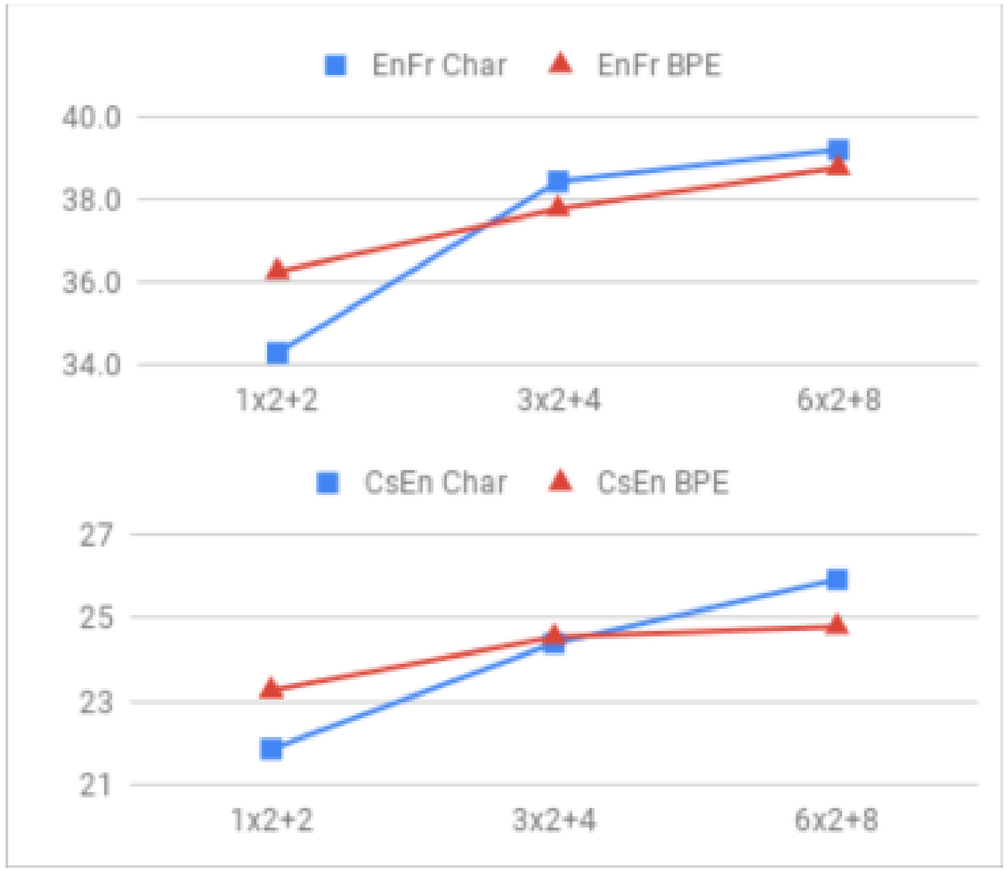
- 개선된 성능
- 인코더 : char 단위 input, convolution layer, max pooling, single layer GRU

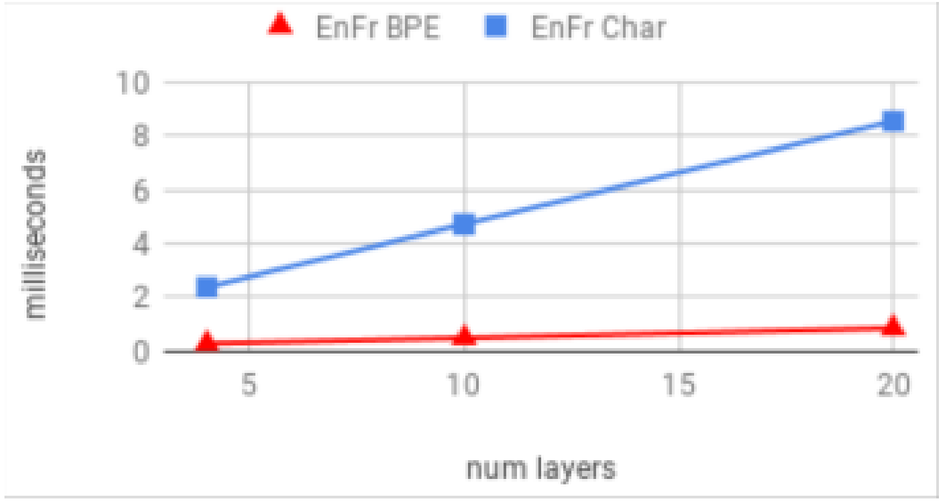


CS-En	WMT 15	Test
Source	Target	BLEU
Bpe	Bpe	20.3
Bpe	Char	22.4
Char	Char	22.5

#02 Purely character-level models

Seq2seq과 BPE 모델 성능 비교

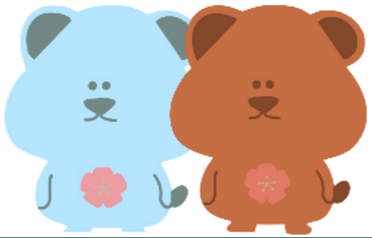




- 영어 – 프랑스어 번영에서는 큰 차이가 없지만 체코어 – 영어 번역에서 우수한 성능
- Character-level 연산량이 매우 크다

EWHA EURON

Subword models



#03 Subword models

BPE(Byte Pair Encoding)

- Word level model과 비슷하다. BPE는 더 작은 word인 word pieces를 이용한다.
- 딥러닝과는 무관한 아이디어
- Most frequent byte pair(n gram)을 새로운 byte(a new gram)으로 clustering

#03 Subword models

BPE(Byte Pair Encoding)

- Frequent 하게 등장하는 es, est, lo를 새로운 단어로 cluster
- 새로 추가된 단어도 하나의 단어처럼 취급
- Target vocab size 에 도달하면 중지
- 시스템의 vocab를 자동적으로 결정

Dictionary

5 l o w
2 l o 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

Dictionary

5 l o w
2 l o w e r
6 n e w **es** t
3 w i d **es** t

Vocabulary

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

Dictionary

5 l o w
2 l o 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**

#03 Subword models

Wordpiece model

- 단어 내에서 tokenizing을 진행
- Pre-segmentation + BPE
- Used to Transformer, EIMo, BERT, GPT-2 -> 최신 딥러닝 모델들

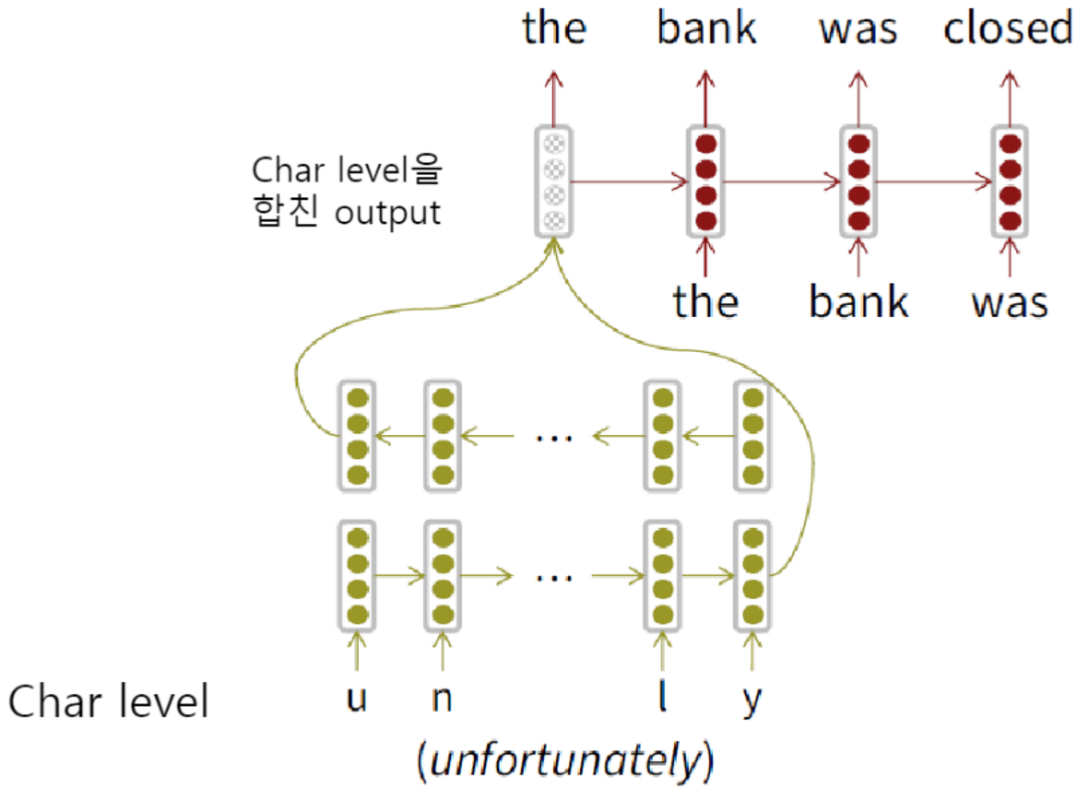
Sentencepiece model

- Raw text에서 바로 작동
- 구글에서 2018년 공개한 비지도학습 형태소 분석 패키지
- Pre-segmentation 없이 단어 분리 토큰화 진행
- bigram 각각에 대해 co-occurence 확률을 계산하고 가장 높은 값을 가지는 것을 단어장에 추가

Hybrid models

EWHA EURON

#04 Hybrid models



Hybrid models

- 기본적으로 word 단위로 취급
- 몇몇만 character 단위로 취급
  Ex) 사전에 없는 단어, 이름

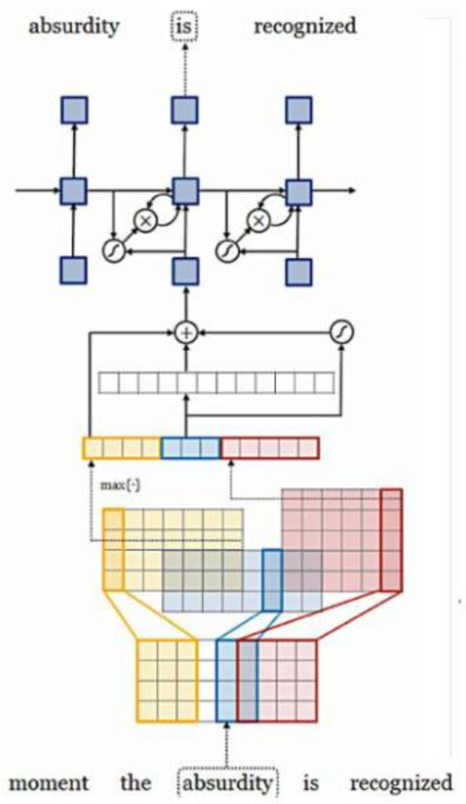
Character-based LSTM (2015)

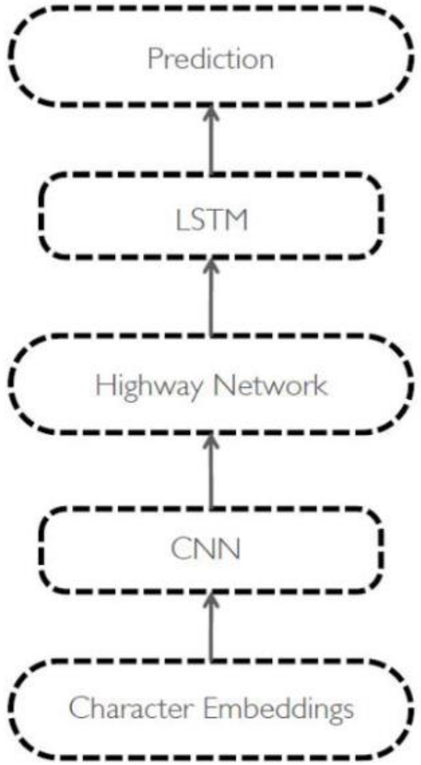
- Bi-LSTM을 통해 word embedding
- final state를 concat해서 임베딩된 단어의 벡터로 사용
- 임베딩된 단어 벡터들을 LSTM에 최종적인 task 진행
- language model, pos tagging 사용

#04 Hybrid models

Character-Aware Neural Language Models (2015)

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

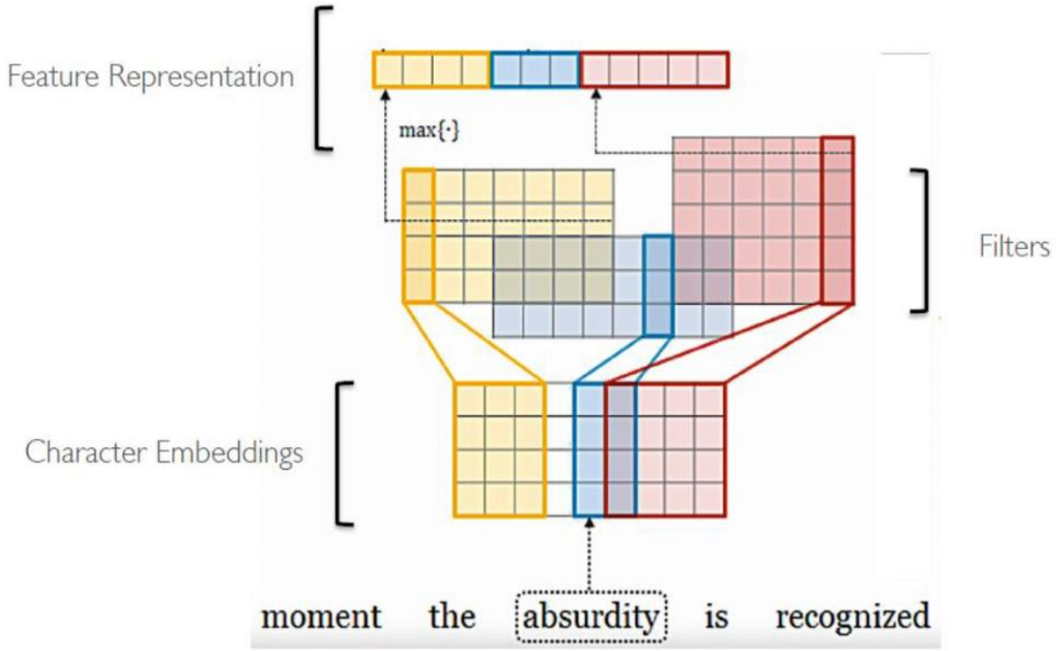




#04 Hybrid models

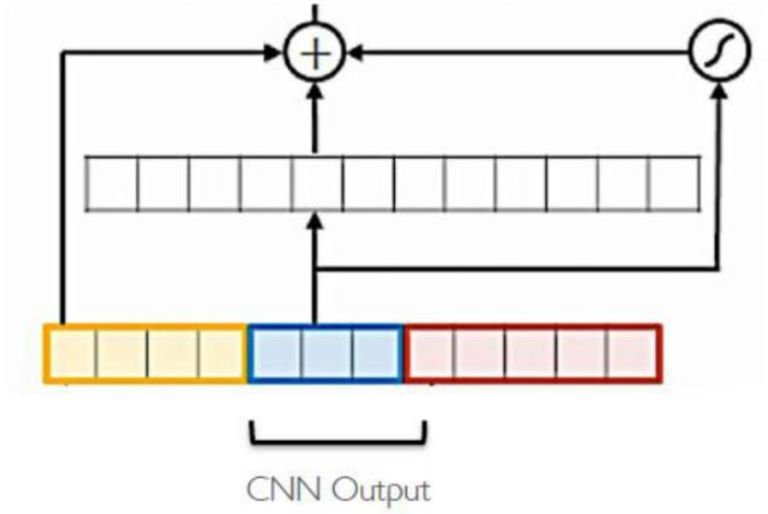
Char 단위로 구분
-> Conv layer with various filter size
-> maxpooling
(어떤 ngram이 단어의 뜻을 가장 잘 나타내는지)
-> highway network
-> Word level LSTM

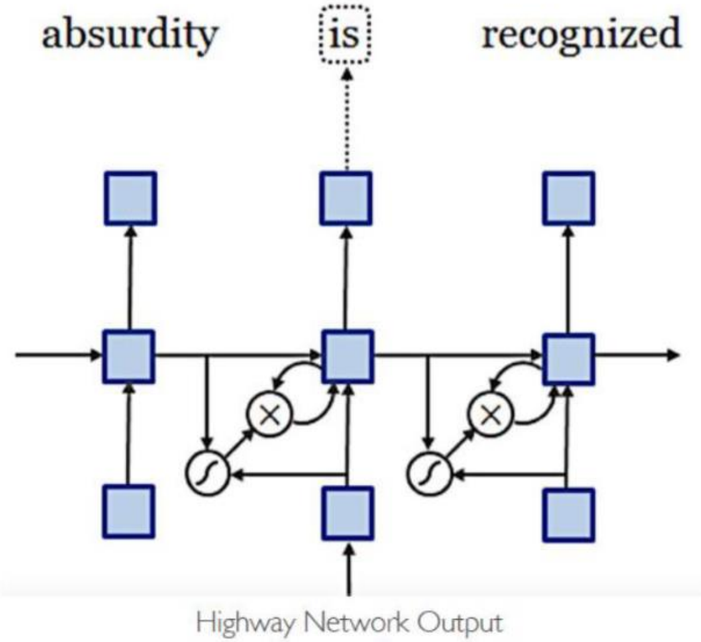
Convolutional layer를 거쳐
feature representation



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

#04 Hybrid models





Highway Network

- LSTM과 유사한 기능
- semantic을 반영하여 가장 유사한 단어를 잘 뽑아내는 결과

Word-level LSTM

- 최종 출력층

#04 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

#04 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*

Richard의 **철자**가 유사한
단어들이 가장 유사하다고 출력

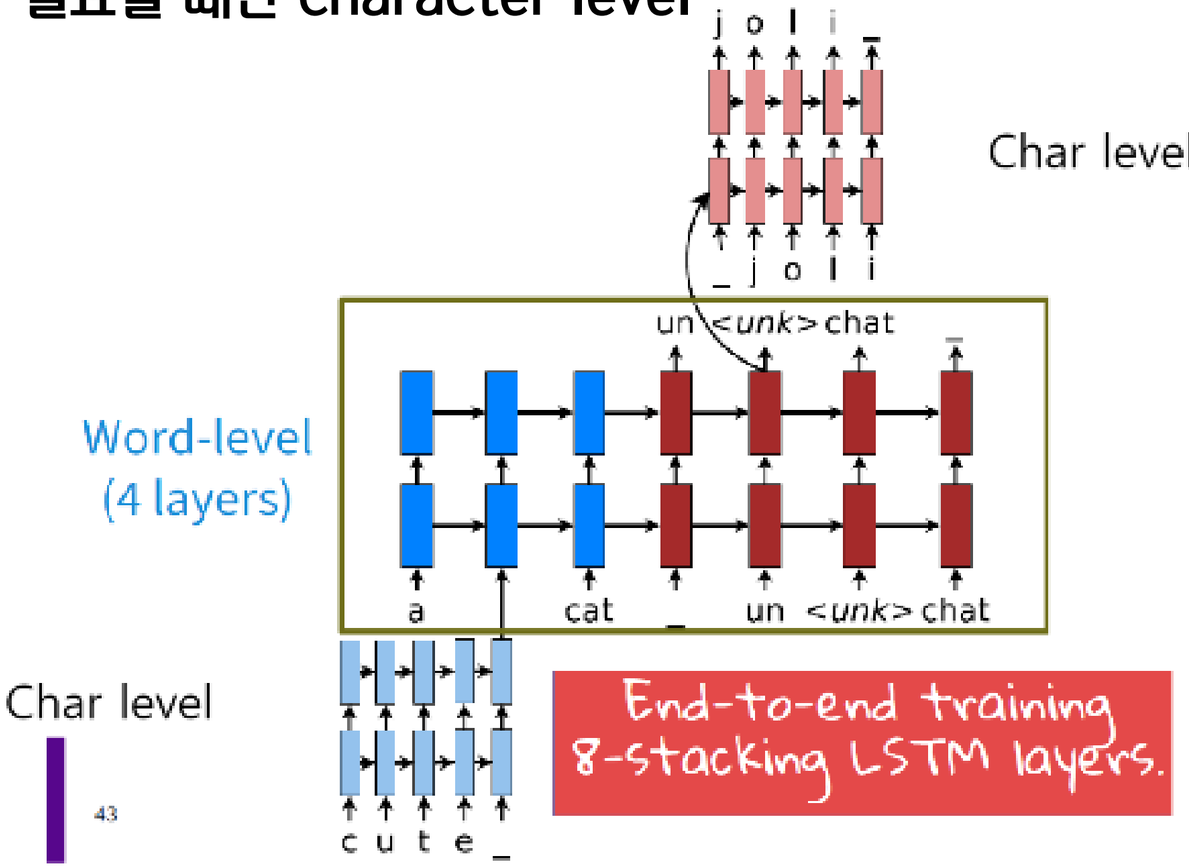
highway

의미를 고려하여
다른 사람 이름을 출력

semantic을 반영하여 더
의미 있는 단어들을 학습

#04 Hybrid models

대부분 word level 사용
필요할 때만 character level



- 16,000개 vocabulary size 이용
- seq2seq으로 word-level model 진행
- unknown word -> character-level model
- 4개의 layer 사용

#04 Hybrid models

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** .

wrong name translation

#04 Hybrid models

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**
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 **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

이름은 문제없이 옮기지만 나이 같은 경우 제대로 번역이 되지 않은 문장이 만들어짐

#04 Hybrid models

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** .

번역 결과 hybrid가 가장 우수

#04 Hybrid models

FastText Embeddings

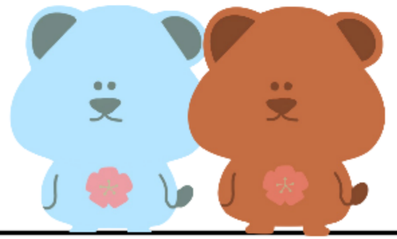
- 차세대 word2vec (word vector learning library)
- 하나의 단어에 여러 단어들이 존재하는 것으로 간주
- 한 단어의 n-gram과 원래의 단어를 모두 학습에 사용

형태소가 풍부한 언어나 희귀한 단어들을 다룰 때 성능이 더 좋음

모르는 단어에 대해서도 subword를 활용해 다른 단어와의 유사도 계산 가능

등장 빈도수가 적은 단어도 다른 단어와 n-gram을 비교해 임베딩 값 계산 가능

THANK YOU



EWHA EURON