1주차

FASTTEXT

ENRICHING WORD VECTORS WITH SUBWORD INFORMATION

NLP2 김유진 문예진 송경민 이상민 한유경

김유진 Presenter

들어가기 전에…

- fastText의 기원으로 꼽히는 논문은
  Facebook AI Research의 Enriching Word Vectors with Subword Information이며,
  완성은 Advances in Pre-Training Distributed Word Representations로 본다.

- Word2vec을 제안한 T. Mikolov가 저자로 들어있으며,
  따라서 Y. Bengio의 NNLM, T. Mikolov의 Word2Vec (CBOW , Skip-gram) 에 대한
  개념이 있다면 논문을 이해하는데 도움이 된다.

Introduction

01. 기존 연구의 문제점

- 각 단어에 독립적인 벡터를 부여하여 단어의 내재적 구조를 파악하지 못함

  예시) eat과 eating

- 희소한 단어를 임베딩하기 어려움
- 언어의 형태론적 특징(morphological feature)이 많은 언어에서는 뚜렷한 한계가 존재

  - morphologically rich languages

    예시1) 프랑스어나 스페인어의 동사는 40개 이상의 다른 형태를 가짐

    예시2) 핀란드어는 15개의 명사 형태를 가짐

02. FastText

- continuous skip-gram model을 기반으로 subword information을 추가!

Model

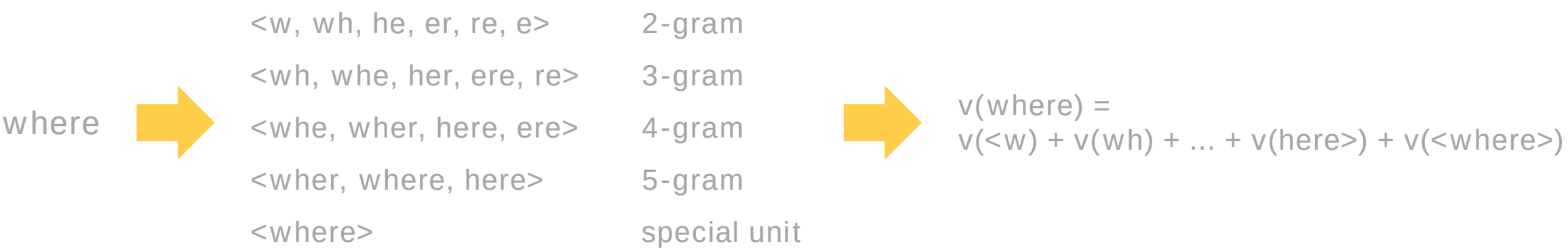
01. General Model

- Word2Vec 의 Skip-gram model을 사용

02. Subword Model

- subword units를 고려함으로써 언어의 형태론적 특징을 모델링 함
  - 단어의 시작과 끝에 boundary symbols <, >를 추가
  - 자기 자신도 special unit으로 포함
- 단어를 Bag-of-Characters로 보고, 개별 단어가 아닌 n -gram의 Charaters를 임베딩
- 최종적으로 각 단어는 Embedding된 n -gram vector들의 총합으로 표현

  예시) where $2 \leq n \leq 5$ 범위의 n -gram으로 임베딩



Model

02. Subword Model

- 새로운 Scoring Function을 정의

$$s(w,c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_c.$$

G : n -gram dictionary size
$G_w \subset \{1,\ldots, G\}$: set of n-grams appealing in word w
Z_g: vector representation to each n -gram g

03. This simple model allows ...

- OOV(out of vocaburary, 모르는 단어)에 대한 학습 능력 향상
  예시) **birthplace**라는 단어를 학습하지 않은 상태였더라도 다른 단어의 n-grams에서 **birth**와 **place**를 학습한 적이 있다면 **birthplace**의 embedding vector를 만들 수 있게 됨
- (오타와 같은) 빈도수가 낮은 단어에 대한 학습 능력 향상
  ex) **where**를 오타를 포함한 **wherre**과 character **3-grams** 기준으로 비교
  - where → <wh, whe, her, **ere**, re>
  - wherre → <wh, whe, her, **err, rre,** re>

→ **2개의 subword만 다름**

Experimental setup

01. Baseline

- Word2Vec 패키지에 있는 skipgram 모델, cbow 모델을 본 논문의 모델과 비교

02. Optimization

- SGD
- training size = T, passes over the data = P, time = t
- fixed parameter = γ_0 일 때 step size = $\gamma_0(1 - \frac{t}{TP})$

03. Implementation details (word2vec 기본 paramater 설정)

- word vectors dimension = 300
- negative samples = 5
- context window size = c, uniformly sample the size c between 1 and 5
- in order to subsample the most frequent words → rejection threshold of 10^{-4}
- keep the words that appear at least 5 times in training set
- γ_0 = 0.025 for skip-gram(baseline), 0.05 for both cbow(baseline) and our model

04. Datasets

- 위키피디아 자료 (9개 언어)

Experiment

01. Human similarity judgement

- 평가 : human judgement(사람의 유사도 평가)와 cosine similarity의 스피어만 상관계수를 계산
- n-gram vector 이용한 sisg-, sisg와 skipgram, cbow 비교
  - sisg (Subword Information Skip Gram) : 본 논문에서 제안하는 모델
  - sisg - : Training data에 없는 단어를 null vector로 대체한 모델
  - cbow 와 skipgram : Word2Vec에서 제안한 모델
- cbow와 skipgram, sisg- 모델은 Training data에 없는 단어의 경우 null vector로 처리함.

Experiment

01. Human similarity judgement

- sisg가 English WS353을 제외한 모든 데이터에 대해 우수한 성능을 보임

  → WS353 dataset은 흔한 단어(Common words)로 구성되어 있기 때문. RareWord dataset에서는 더 좋은 성능을 보임.

- 모든 Dataset에 대해 sisg가 sisg- 보다 같거나 우수한 성능을 보임

  → subword information을 제공하는 것의 장점 입증

- 영어보다는 Arabic / German / Russian data에 대해 좋은 성능을 보임.

  → 형태적으로 복잡하거나 합성어가 많은 언어

		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**

Experiment

02. Word analogy tasks

- A is to B as C is to D 형태의 문장에서

model이 D를 예측하는 문제

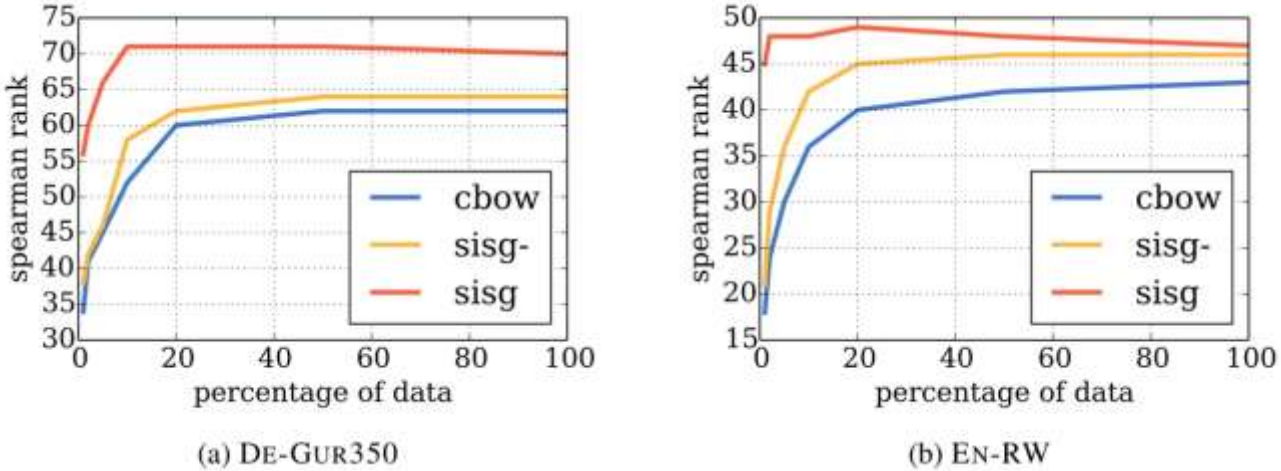
- syntactic(구조적) task에서는 baseline보다 우수한 성능
- 그러나 semantic(의미적) task에서는 좋지 않은 성적
- 심지어 German과 Italian에서는 성능이 떨어짐

  → n-grams의 size가 적절히 선택되면 개선할 수 있음
- Cs(체코어)와 De(독일어)같이 형태학적으로 풍부한 언어 (morphologically rich language)에서 성능 개선이 눈에 띄게 관찰됨

		sg	cbow	sisg
CS	Semantic	25.7	27.6	27.5
	Syntactic	52.8	55.0	77.8
DE	Semantic	66.5	66.8	62.3
	Syntactic	44.5	45.0	56.4
EN	Semantic	78.5	78.2	77.8
	Syntactic	70.1	69.9	74.9
IT	Semantic	52.3	54.7	52.3
	Syntactic	51.5	51.8	62.7

Experiment

03. Effect of the size of the training data



- 모든 경우 sisg가 baseline보다 더 나은 성능을 보임
- 데이터 세트가 축소(1~20%)될수록 어휘 부족률이 증가 → sisg- & cbow의 성능 저하
- 많은 데이터를 사용할 수록 cbow 모델의 성능은 향상되지만 sisg는 빠르게 포화
  → sisg는 더 많은 데이터를 추가한다고 결과가 개선되진 X
- sisg는 매우 작은 train dataset을 사용할 경우에도 좋은 단어 벡터를 제공

Experiment

04. Effect of the size of the training data

	2	3	4	5	6
2	57	64	67	69	69
3		65	68	70	70
4			70	70	**71**
5				69	**71**
6					70

(a) DE-GUR350

	2	3	4	5	6
2	59	55	56	59	60
3		60	58	60	62
4			62	62	63
5				64	64
6					**65**

(b) DE Semantic

	2	3	4	5	6
2	45	50	53	54	55
3		51	55	55	**56**
4			54	**56**	**56**
5				**56**	**56**
6					54

(c) DE Syntactic

	2	3	4	5	6
2	41	42	46	47	**48**
3		44	46	**48**	**48**
4			47	**48**	**48**
5				**48**	**48**
6					**48**

(d) EN-RW

	2	3	4	5	6
2	78	76	75	76	76
3		78	77	78	77
4			79	79	79
5				**80**	79
6					**80**

(e) EN Semantic

	2	3	4	5	6
2	70	71	73	74	73
3		72	74	**75**	74
4			74	**75**	**75**
5				74	74
6					72

(f) EN Syntactic

- Semantic task : 큰 n만 고려할 때 성능이 좋음 , Syntactic task : 작은 n을 고려할 때 성능이 좋음
- 5,6 column이 가장 성능이 좋은 것을 관찰할 수 있음
- 2-gram은 결합이나 축약에 해당하는 접미사를 적절하게 캡처하기에 충분하지 않아 다른 경우보다 성능이 떨어짐

Experiment

05. Language modeling

- sisg가 모든 언어에서 다른 모델(LSTM, sg)보다 좋은 성능을 보임

  → 언어 모델링 태스크에 subword 정보의 중요성을 알 수 있음

- 비교적 Vocab Size가 큰 슬라브어족 언어(Cs(체코어), Ru(러시아어))에서 perplexity가 더 작음

  → 형태학적으로 풍부한 언어에 이 논문의 모델의 벡터가 유용하다는 것을 알 수 있음

	CS	DE	ES	FR	RU
Vocab. size	46k	37k	27k	25k	63k
CLBL	465	296	200	225	304
CANLM	371	239	165	184	261
LSTM	366	222	157	173	262
sg	339	216	150	162	237
sisg	**312**	**206**	**145**	**159**	**206**

Table 5: Test perplexity on the language modeling task, for 5 different languages. We compare to two state of the art approaches: CLBL refers to the work of Botha and Blunsom (2014) and CANLM refers to the work of Kim et al. (2016).

Qualitative analysis

01. Character n-grams and morphemes

- 주어진 단어에서 가장 중요한 3개의 character n-gram을 추출한 결과

  예시) autofahrer(car driver)
  → auto(car), Fahrer(driver)

- 중요한 단어들이 n-grams에 다 표현됨

- 영어(EN)에서 복합명사와 접사가 분리되는 것을 관찰

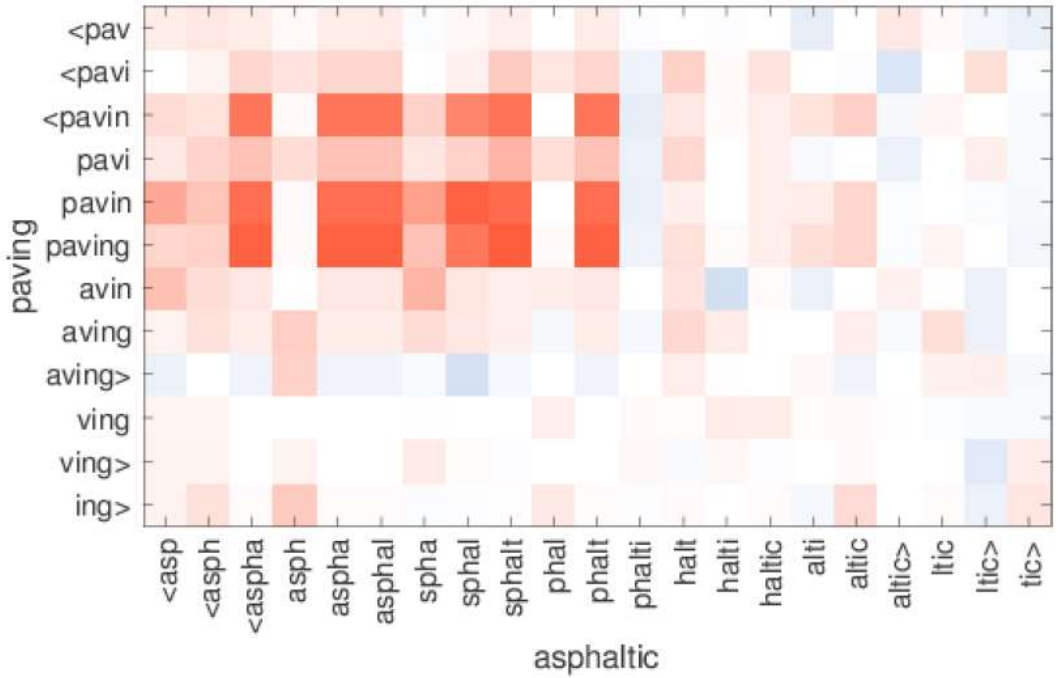
  예시) starfish → star, fish (복합명사의 분리)
  kindness → kind + ness (접사의 분리)

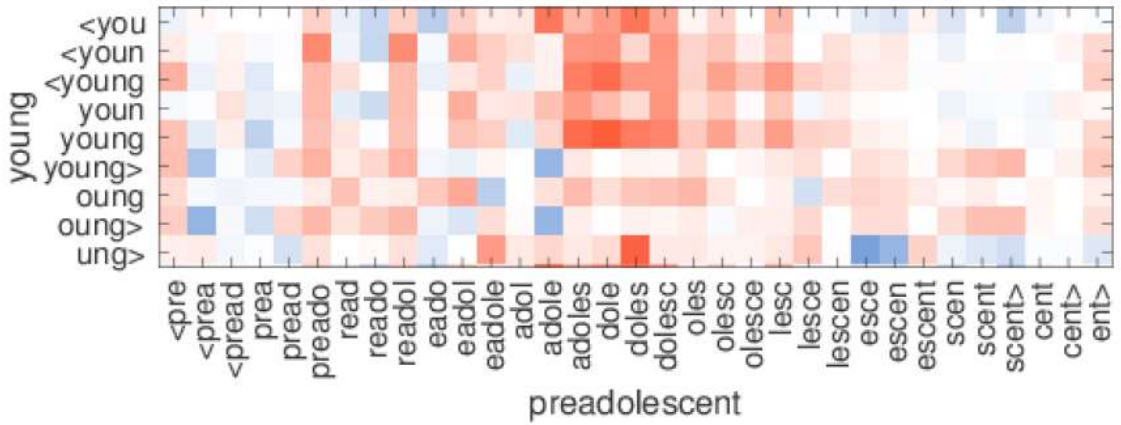
- 프랑스어(FR) 에서는 동사의 굴절(inflections) 을 관찰

	word	n-grams		
DE	autofahrer	fahr	fahrer	auto
	freundeskreis	kreis	kreis>	<freun
	grundwort	wort	wort>	grund
	sprachschule	schul	hschul	sprach
	tageslicht	licht	gesl	tages
EN	anarchy	chy	<anar	narchy
	monarchy	monarc	chy	<monar
	kindness	ness>	ness	kind
	politeness	polite	ness>	eness>
	unlucky	<un	cky>	nlucky
	lifetime	life	<life	time
	starfish	fish	fish>	star
	submarine	marine	sub	marin
	transform	trans	<trans	form
FR	finirais	ais>	nir	fini
	finissent	ent>	finiss	<finis
	finissions	ions>	finiss	sions>

Qualitative analysis

02. Word similarity for OOV words





- X축 : OOV(등장하지 않은 단어)
- Y축 : 학습 데이터 셋 내 단어

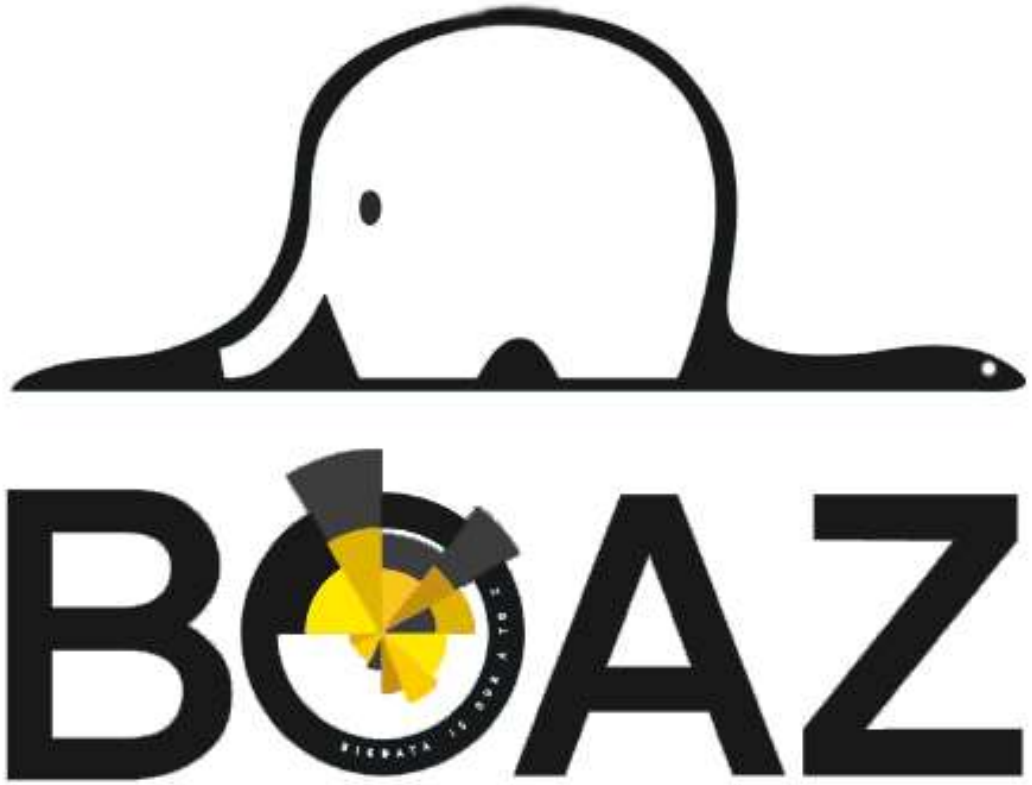
- 비슷한 뜻을 가진 단어들끼리 유사도가 높은 것을 볼 수 있다.
  예시) "pav-", "-sphal-" : (아스팔트) 포장 , "young", "adoles" : 어린(이)

Conclusion

- FastText는 skip-gram 모델을 기반으로 n-gram을 활용해 subword information을 도입
  → Word2Vec에서 반영하지 못했던 형태학적 특징까지 반영

Because of its simplicity...

- trains fast & does not require any preprocessing or supervision
- compute word representations for words that did not appear in the training data
- State-of-the-art performance !



Any Questions?