sense2vec

A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings

Under review as a conference paper at ICLR 2016

Digital Reasoning Systems, Inc.

2018.09.20 이 현 주

INTRODUCTION

- 다양한 형태의 Distributed representation 은 Part-of-Speech tagging, Named Entity Recognition, Analogy/Similarity Querying, Transliteration, Dependency Parsing 을 포함한 다양한 NLP 작업에 유용하게 사용됨.
- 대부분의 Word Embedding 기술은 각 단어의 모든 잠재적 의미를 단일 벡터로 인코딩해야 함

- 문제점 : 여러 의미가 있는 단어는 벡터가 개별적인 의미의 혼합을 취하는 벡터 공간에서 중첩이 일어날 수 있음

→ 중첩이 단어의 컨텍스트 특정 의미를 난독화하고 입력 데이터로 중첩을 사용할 경우, NLP Classifier 에 부정적인 영향을 줄 수 있음
→ 여러 단어 감각을 개별적인 퍼짐으로 구별하는 것이 이 문제와 NLP 모델에 해당되는 혼란을 완화함

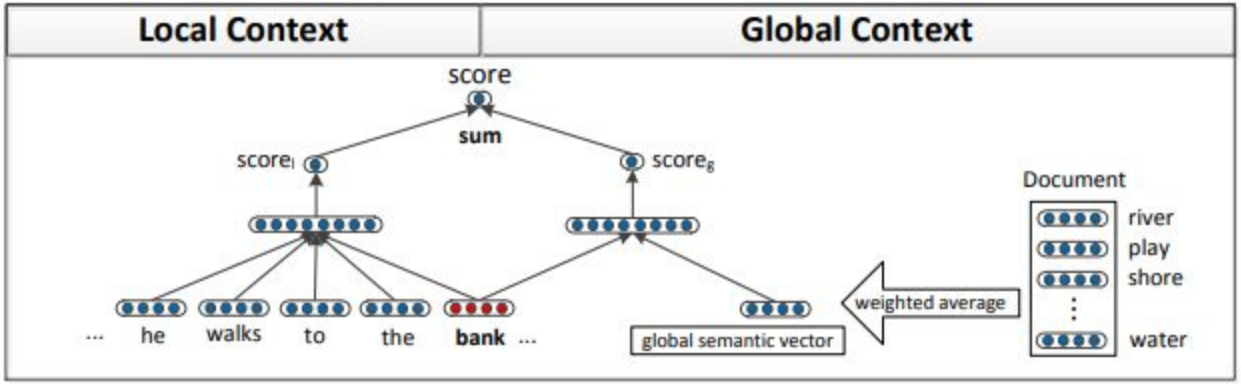
Method : The Sense2vec Model

supervised NLP label 을 활용하여 특정 단어의 인스턴스의 의미를 결정함

Huang et al. (2012)는 감독되지 않은 클러스터 사용하는 데에 반해 Sense2Vec 은
1) 임베딩을 여러 번 교육할 필요가 없음
2) 클러스터링 단계가 필요하지 않음
3) supervised classifier 가 적절한 word-sense embedding 을 사용할 수 있는 효율적인 방법을 만듬

주변 토큰이 주어진 토큰을 예측하는 대신 주변 감각(sense) 가 제공하는 단어 감각(sense) 을 예측함

* Huang et al. (2012)
Improving word representations via global context and multiple word prototypes



1) Local Context Score : 단어 순서와 구문 정보를 파악
2) Global Context Score : 문서의 의미와 주제를 파악
   a) Context word Vector 의 가중 평균을 구해 Context Representation
   b) Context Representation을 클러스터링
   c) 해당 클러스터에서 단어 표현을 학습(idf)

SUBJECTIVE EVALUATION - 1) SUBJECTIVE BASELINE

Word2vec의 Continuous Bag of Words 를 사용하여 학습(Google Word Analogy Task DataSet)

Table 1: Single-sense Baseline Cosine Similarities

bank	**1.0**	**apple**	**1.0**	**so**	**1.0**	**bad**	**1.0**	**perfect**	**1.0**
banks	.718	iphone	.687	but	.879	good	.727	perfection	.681
banking	.672	ipad	.649	it	.858	worse	.718	perfectly	.670
hsbc	.599	microsoft	.603	if	.842	lousy	.717	ideal	.644
citibank	.586	ipod	.595	even	.833	stupid	.710	flawless	.637
lender	.566	imac	.594	do	.831	horrible	.703	good	.622
lending	.559	iphones	.578	just	.808	awful	.697	always	.572

bank : proper noun, none, verb 등 3가지 방식으로 사용됨
apple : 일부 단어의 경우, 단어의 해석이 완전히 무시됨
so : vector space 에서 잘 표현되지 않는 sence(감각) 이 있는 것으로 파악됨

SUBJECTIVE EVALUATION - 2) PART-OF-SPEECH DISAMBIGUATION

Polyglot Universal Dependency 품사 Tagger를 사용하여 Dataset 에 품사 태그 표시

Table 2: Part-of-Speech Cosine Similarities for the Word: apple

apple	NOUN	1.0	apple	PROPN	1.0
apples	NOUN	.639	microsoft	PROPN	.603
pear	NOUN	.581	iphone	NOUN	.591
peach	NOUN	.579	ipad	NOUN	.586
blueberry	NOUN	.570	samsung	PROPN	.572
almond	NOUN	.541	blackberry	PROPN	.564

과일을 나타내는 명사 "apple"와
회사를 지칭하는 고유 명사 "apple"의 차이를
구별할 수 있음

Table 3: Part-of-Speech Cosine Similarities for the Word: bank

bank	NOUN	1.0	bank	PROPN	1.0	bank	VERB	1.0
banks	NOUN	.786	bank	NOUN	.570	gamble	VERB	.533
banking	NOUN	.629	hsbc	PROPN	.536	earn	VERB	.485
lender	NOUN	.619	citibank	PROPN	.523	invest	VERB	.470
bank	PROPN	.570	wachovia	PROPN	.503	reinvest	VERB	.466
ubs	PROPN	.535	grindlays	PROPN	.492	donate	VERB	.466

"bank" 이라는 단어의 세 가지 용도가
모두 각자의 품사에 의해 명확해짐

Table 4: Part-of-Speech Cosine Similarities for the Word: so

so	INTJ	1.0	so	ADV	1.0	so	ADJ	1.0
now	INTJ	.527	too	ADV	.753	poved	ADJ	.588
obviously	INTJ	.520	but	CONJ	.752	condemnable	ADJ	.584
basically	INTJ	.513	because	SCONJ	.720	disputable	ADJ	.578
okay	INTJ	.505	but	ADV	.694	disapprove	ADJ	.559
actually	INTJ	.503	really	ADV	.671	contestable	ADJ	.558

미묘한 의미의 "So"라는 단어도
품사 태그의 삽입으로 의미가 명확해짐

SUBJECTIVE EVALUATION - 3) SENTIMENT DISAMBIGUATION

라벨링된 IMDB Training Corpus 에 Polyglot POS Tagger 를 사용하여 형용사만을 추출

CBOW sense2vec 모델이 결과 데이터 집합에 대해 학습되어 품사와 감정을 명확히 구분함

Table 5: Sentiment Cosine Similarities for the Word: bad

bad	NEG	1.0	bad	POS	1.0
terrible	NEG	.905	good	POS	.753
horrible	NEG	.872	wrong	POS	.752
awful	NEG	.870	funny	POS	.720
good	NEG	.863	great	POS	.694
stupid	NEG	.845	weird	POS	.671

Table 6: Sentiment Cosine Similarities for the Word: perfect

perfect	NEG	1.0	perfect	POS	1.0
real	NEG	0.682	wonderful	POS	0.843
unfortunate	NEG	0.680	brilliant	POS	0.842
serious	NEG	0.673	incredible	POS	0.840
complete	NEG	0.673	fantastic	POS	0.839
ordinary	NEG	0.673	great	POS	0.823
typical	NEG	0.661	excellent	POS	0.822
misguided	NEG	0.650	amazing	POS	0.814

Negative "bad" vector 는 고전적인 의미의 나쁜 단어를 나타내는 단어와 가장 유사
Positive “bad” vector 는 풍자의 의미를 나타내며 "good" 등 Positive Sense 와 가장 밀접하게 관련됨

SUBJECTIVE EVALUATION - 4) NAMED ENTITY RESOLUTION

NER 에서 명확하게 할 때 embedding을 평가하기 위해 Named Entity Label 로 데이터셋을 레이블링

sense2vec이 텍스트의 multi-word sequence 와 single word sequence 사이에서 어떻게 구별하는 지를 보여줌

Table 7: Disambiguation for the word: Washington

George_Washington	PERSON_NAME	.656	Washington_D	GPE	.665
Henry_Knox	PERSON_NAME	.624	Washington_DC	GPE	.591
Philip_Schuyler	PERSON_NAME	.618	Seattle	GPE	.559
Nathanael_Greene	PERSON_NAME	.613	Warsaw_Embassy	GPE	.524
Benjamin_Lincoln	PERSON_NAME	.602	Wash	GPE	.516
William_Howe	PERSON_NAME	.591	Maryland	GPE	.507

"Wachington" 이라는 단어가 PERSON과 GPE 느낌으로 구분되어 있음

Table 8: Entity resolution for the term: Hillary Clinton

Secretary_of_State	TITLE	0.661
Senator	TITLE	0.613
Senate	ORG_NAME	0.564
Chief	TITLE	0.555
White_House	ORG_NAME	0.564
Congress	ORG_NAME	0.547

"Hillary Clinton"은 데이터 수집의 기간 내에 그녀가 가지고 있던 타이틀과 매우 유사함

SUBJECTIVE EVALUATION - 5) Neural Dependency parsing

각 임베딩을 정량적으로 평가하기 위해 6 개 언어를 사용하여 Neural Syntactic Dependency parsing task 결과 비교

불가리아어, 독일어, 영어, 프랑스어, 이탈리아어 및 스웨덴어 위키피디아 데이터셋에 sense2vec 임베딩과 wang2vec 임베딩을 학습

- Baseline Embedding : Structured skip-gram 을 사용하여 학습
- sense2vec embedding : Polyglot 품사 태거를 사용하여, Structured skip-gram 으로 학습

- 이 각각의 임베딩은 (Chen & Manning, 2014)의 Dependency parse model 을 학습하는 데 사용
- (Chen & Manning, 2014)의 파서는 gold standard POS 태그를 기반으로 인덱싱된 훈련된 품사 포함을 입력으로 사용

Table 9: Unlabeled Attachment Scores and Percent Error Reductions

	Set	Bulgarian	German	English	French	Italian	Swedish	Mean
wang	Dev	90.03	68.86	85.02	73.82	84.99	78.94	80.28
	Test*	90.17	60.25	83.61	70.10	84.99	82.47	78.60
	Test	90.39	60.54	83.88	70.53	85.45	82.51	78.88
sense	Dev	90.69	72.61	86.10	75.43	85.57	81.21	81.94
	Test*	90.41	64.17	85.48	71.66	86.13	84.44	80.38
	Test	90.86	64.43	85.93	72.16	86.18	84.60	80.69
Error Margin	Dev	7.05%	13.69%	7.76%	6.56%	3.98%	12.06%	**8.52%**
	Test	2.47%	10.95%	12.82%	5.50%	8.21%	12.71%	**8.78%**
	Abs.	5.17%	10.93%	14.54%	5.86%	5.32%	13.58%	**9.23%**
	Avg.	4.76%	12.32%	10.29%	6.03%	6.09%	12.39%	

두 가지 embedding 에 대해 학습된 parser 는 두 모델이 모두 POS 정보를 학습함
→ 단어 embedding 시 품사 정보를 추가하는 것이 두 모델의 가장 큰 차이

sense2vec를 사용한 품사 기반의 모호성 제거는 6 개 언어의 오류를 모두 줄임, 평균 감소율 8 %

CONCLUSION AND FUTURE WORK

본 연구에서는, 단어 감각의 차이를 줄이기 위해 감독된 NLP 라벨을 사용하는 새로운 모델을 제안

컨텍스트 클러스터링 형식을 활용하는 대신, 특정 단어의 컨텍스트를 분석하고 레이블을 지정할 수 있는 Supervised Label 을 사용하여 클러스터링함

→ Word Sense 모델의 연산 오버헤드를 크게 줄이고, 다른 NLP 작업에서 적절한 감지 임베딩을 선택하는 자연스러운 메커니즘을 제공함

→ 명확한 임베딩은 다양한 언어로 Syntactic dependency parsing Model의 정확성을 높일 수 있음을 보여줌