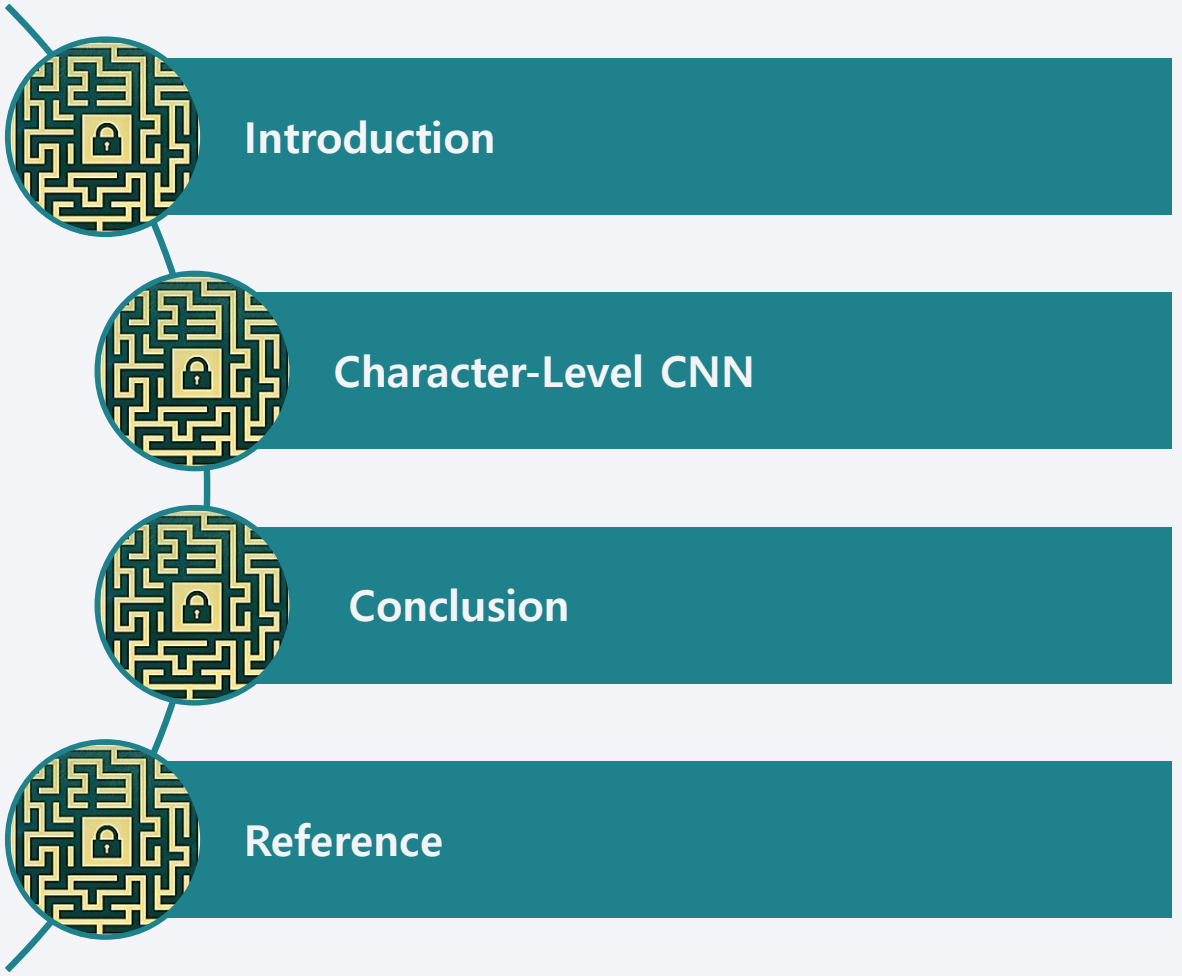
Character-Level Convolutional Neural Network for Text Classification in Korean

2016021200 모경현
2016021201 박재선
2016020324 이윤정
2017020552 장명준

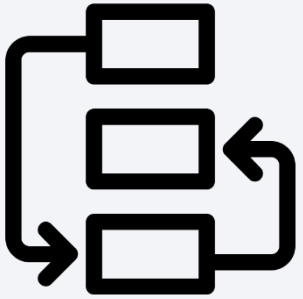
Contents

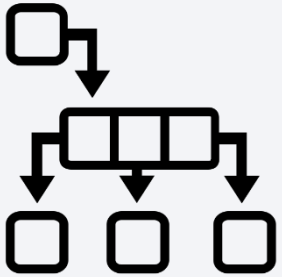


Introduction - Research Background

Traditional Method for classification

Word	Freq
Tears	2
…	…
Text	5





Documents

Classified with several categorization

Counting

Unit : word

Algorithm

TF-IDF, CNN, RNN…

Classification

Introduction - Research Background

Traditional Method for classification

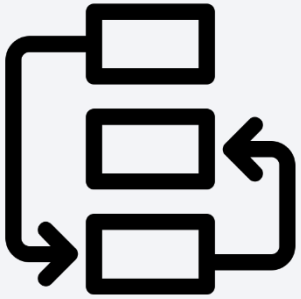
Documents

Classified with several categorization

Word	Freq
Tears	2
…	…
Text	5

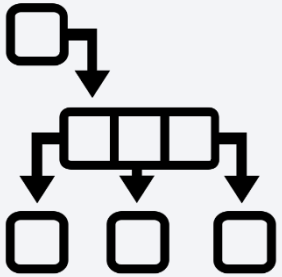
Counting

Unit : word



Algorithm

TF-IDF, CNN, RNN…



Classification

Tokenizing is the most important

Introduction - Research Background

Kinds of documents based on degree of refinement

Well-refined documents

Ex) News Articles, textbook

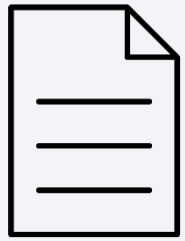
Unrefined documents

Ex) movie review, small talks on SNS

Introduction - Research Background

Preprocessing for documents

step1 **Apply rules to remove**



In Well-refined documents

1) Punctuation Marks
2) Email of journalist
3) Name of publishing company
4) Parentheses

서울과 인천에 많은 눈이 내려 퇴근길에 비상이 걸린 가운데 일부 지역에서는 황사까지 겹친 것으로 알려졌다 20일 오후 서울 등 중부지방에 많은 눈을 뿌린 눈 구름은 차차 동쪽으로 이동하고 있다 오후에는 경기 남부와 강원도 지역에 눈이 내릴 것으로 전망된다 4시 현재 서울등 중부지방에 머무르고 있는 눈 구름은 차한편 이날 제주도와 서해안 지역에는 중국발 미세먼지가 유입되면서 약한 황사 현상을 보이고 있어 황사눈 을 주의해야 한다 따라서 오늘내리는 눈은 반드시 피하거나 우산을 챙기는 것이 좋다

In unrefined documents

1) Punctuation Marks

우리들은 영화같은 삶을 살고있다고 그렇게 말해주는것 같았다. 그러니까 삶을 이어가고있는 것만으로도 우리는 가치있다고 위로해주는 것이다.

Introduction - Research Background

Preprocessing for documents

step1 **Apply rules to remove**

step2 **Apply Cohesion to tokenize**

'서울', '과', '인천', '에', '많은', '눈이', '내려', '퇴근', '길에',
'비상', '이', '걸린', '가운데', '일부', '지역', '에서', '는', '황사',
'까지', '겹친', '것', '으로', '알려', '졌다', '20', '일', '오후',
'서울', '등', '중부지방에', '많은', '눈을', '뿌린', '눈', '구',
'름은', '차차', '동쪽', '으로', '이동', '하고', '있다', '오후',
'에는', '경기', '남부', '와', '강원도', '지역', '에', '눈이', '내릴',
'것', '으로', '전망', '된다', '4시', '현재', '서울', '등',
'중부지방에', '머무르고', '있는', '눈', '구', '름은', '차',
'한편', '이날', '제주도', '와', '서해안', '지역', '에는', '중국',
'발', '미세먼지', '가', '유입되', '면서', '약한', '황사', '현상',
'을', '보이고', '있어', '황사', '눈', '을', '주의', '해야', '한다',
'따라', '서', '오늘', '내리는', '눈은', '반드시', '피하', '거나',
'우산을', '챙기', '는', '것이', '좋다'

'우리', '들은', '영화', '같은', '삶을', '살고', '있다', '고', '그',
'렇게', '말', '해주는', '것', '같', '았다', '.', '그러니까', '삶을',
'이어', '가고', '있는', '것만으로도', '우리', '는', '가치',
'있다', '고', '위로', '해주는', '것이다.'

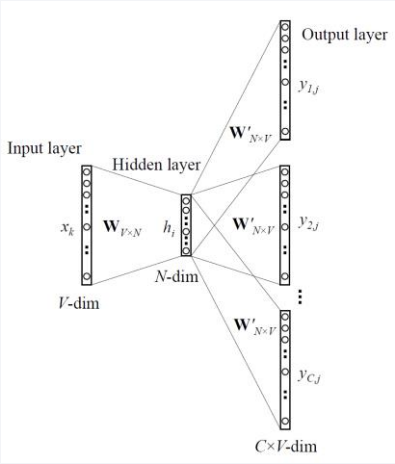
Introduction - Research Background

Preprocessing for documents

step1 **Apply rules to remove**

step2 **Apply Cohesion to tokenize**

step3 **Apply Word2Vec to corpus**



서울=	0.21	0.52	0.89	0.97	-0.34	-0.71
도쿄=	0.34	0.51	0.46	-0.37	-0.92	-0.92

Introduction – Data for this study

Explanation about Two Data sets

Well-refined documents

- News Articles published 조선, 중앙, 한겨레, 경향
- Duration for data collected : 2013.01.01 ~ 2017.03.29
- Number of data : 454,644
- Frequency Table for each category

경제	국제	사회	스포츠	연예	정치
67,078	51,498	84,022	84,022	84,022	84,022

Introduction – Data for this study

Explanation about Two Data sets

Unrefined documents

- Movie Review(Watcha)
- Duration for data collected : 2012.11.08 ~ 2016.07.08
- Number of data : 2,717,668
- Frequency Table for each category

0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
50,660	66,184	62,094	163,272	173,650	411,757	424,378	652,250	297, 327	416,096

Negative Neutral Positive

Introduction – Results Via Traditional Models

Four Algorithms we've used

1. Naïve Bayes
2. Logistic Regression
3. Convolution Neural Network
4. Recurrent Neural Network

Introduction – Results Via Traditional Models

Four Algorithms we've used

1. Naïve Bayes
2. Logistic Regression
3. Convolution Neural Network
4. Recurrent Neural Network

Results

	TF_IDF (NaiveBayes)	TF_IDF (Logistic)	CNN (word level)	RNN (word level)
Article	85.35	90.04	85.02	87.56
Watcha_2class	41.11	74.93	83.15	85.35
Watcha_3class	52.62	66.45	60.06	62.36

Character-Level CNN – Overview of this study

Problem of Word Level Analysis

1. Pre-processing에 소요되는 cost가 큼

   crawling -> pre-processing(remove punctuation…) -> tokenizing(cohesion)

   -> distributed representation(word2vec, glove…)

2. 특수 기호가 담고 있는 의미를 반영할 수 없음

   예) ^^, ㅋㅋㅋㅋㅋㅋㅋ, ㅠㅠ, :), 개쩜!!!!!!!!!!!!

Character-Level CNN – Preprocessing

Method

- **Character level embedding**

  ex) 비정형데이터분석

  ㅂ + ㅣ + ㅈ + ㅓ + ㅇ + ㅎ + ㅕ + ㅇ + ㄷ + ㅐ + ㅇ + ㅣ + ㅌ + ㅓ + ㅂ + ㅜ + ㄴ + ㅅ + ㅓ + ㄱ

  ex) 강필성 교수님

  ㄱ + ㅏ + ㅇ + ㅍ + ㅣ + ㄹ + ㅅ + ㅓ + ㅇ + + ㄱ + ㅛ + ㅅ + ㅜ + ㄴ + ㅣ + ㅁ

  ex) 21일 종강!

  2 + 1 + ㅇ + ㅣ + ㄹ + + ㅈ + ㅗ + ㅇ + ㄱ + ㅏ + ㅇ **+ !**

Character-Level CNN – Preprocessing

Method

- **Component of Character-level in Korean**

  - 초성 : 19개, 중성 21개, 종성 27개, 숫자 10개, 알파벳 26개, 특수문자

  → 193 character

ㄱ : [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

ㄴ : [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Character-Level CNN – Preprocessing

Method

- **Component of Character-level in Korean**

  - 초성 : 19개, 중성 21개, 종성 27개, 숫자 10개, 알파벳 26개, 특수문자

  → 193 character

ㄱ : [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

ㄴ : [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

193 dimensional vector

Character-Level CNN – Preprocessing

Method

- **Character level embedding**

  ex) 종강!

  ㅈ + ㅗ + ㅇ + ㄱ + ㅏ + ㅇ + !

	ㄱ	ㄴ	…	ㅇ	ㅈ	…	ㅏ	…	ㅗ	…	@	!
ㅈ	0	0	…	0	1	…	0	…	0	…	0	0
ㅗ	0	0	…	0	0	…	0	…	1	…	0	0
ㅇ	0	0	…	1	0	…	0	…	0	…	0	0
ㄱ	1	0	…	0	0	…	0	…	0	…	0	0
ㅏ	0	0	…	0	0	…	1	…	0	…	0	0
ㅇ	0	0	…	1	0	…	0	…	0	…	0	0
!	0	0	…	0	0	…	0	..	0	…	0	1

193 dimensional vector

Character-Level CNN

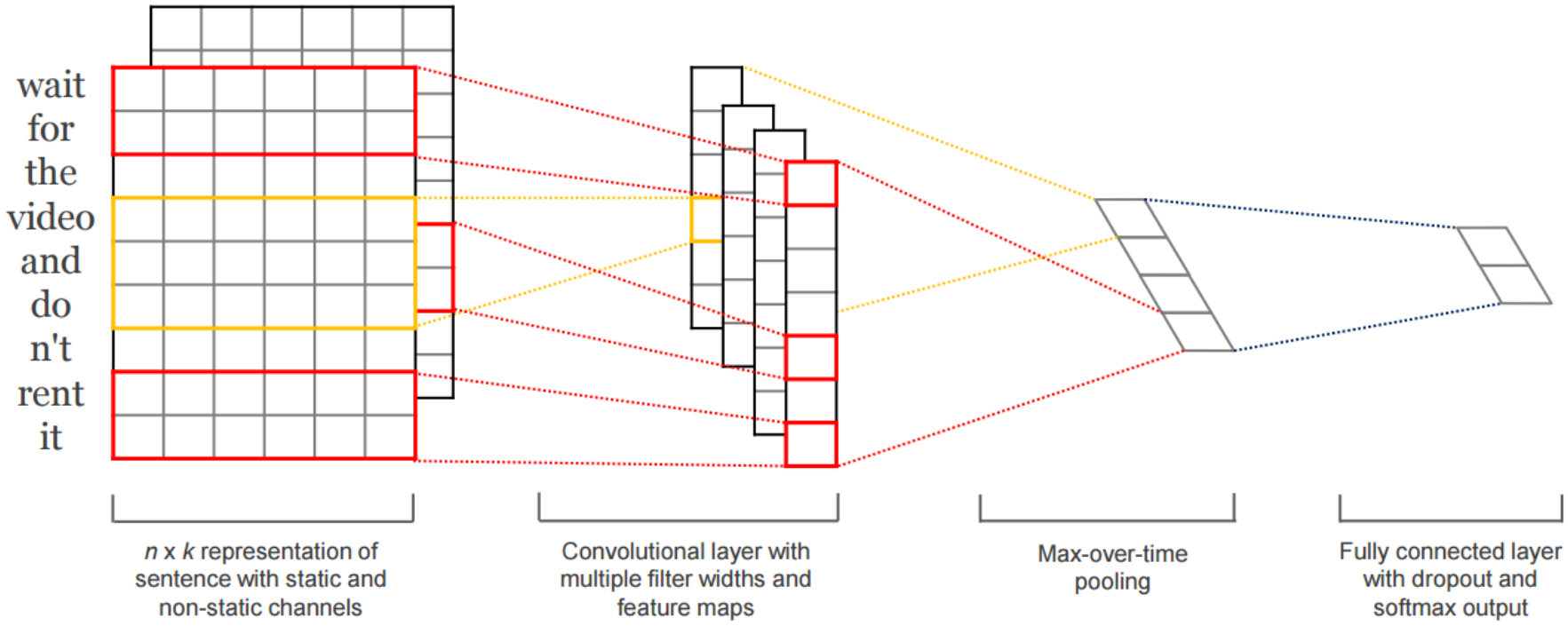


Figure 1: Model architecture with two channels for an example sentence.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882(2014).

Character-Level CNN

Article – 6class

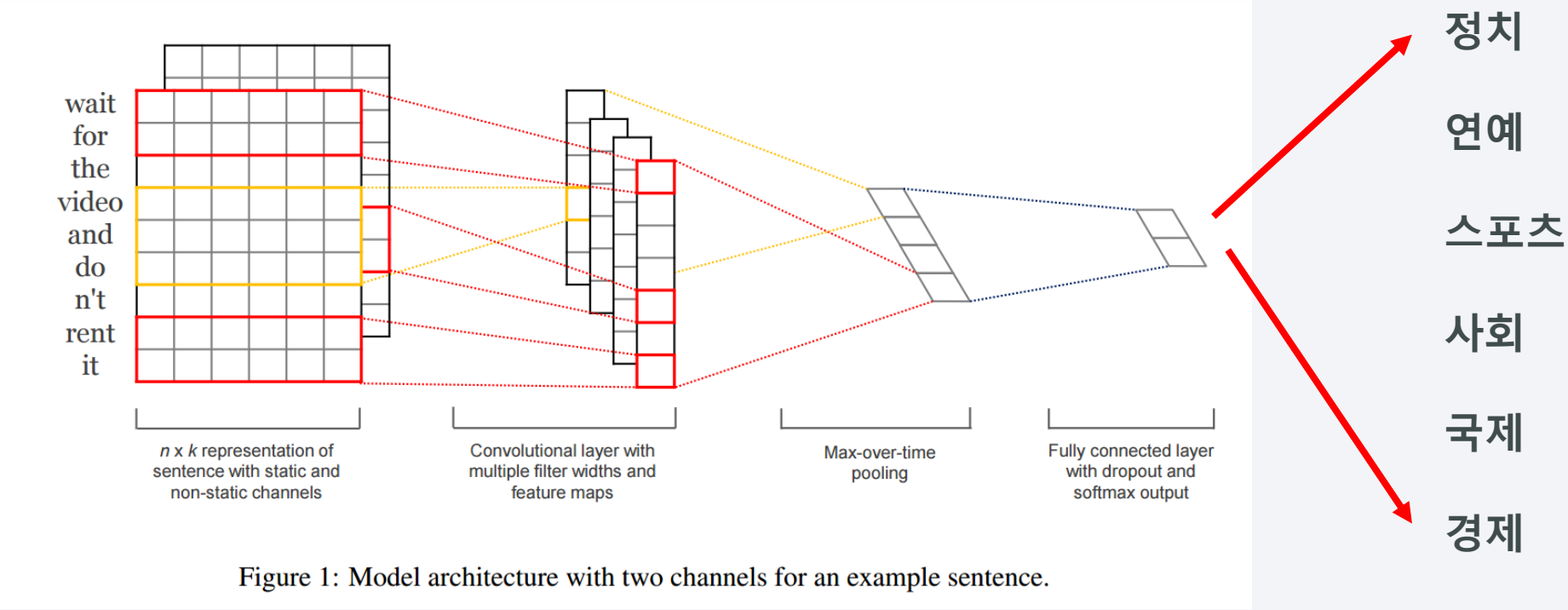


Figure 1: Model architecture with two channels for an example sentence.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882(2014).

Character-Level CNN

Movie review – 2class

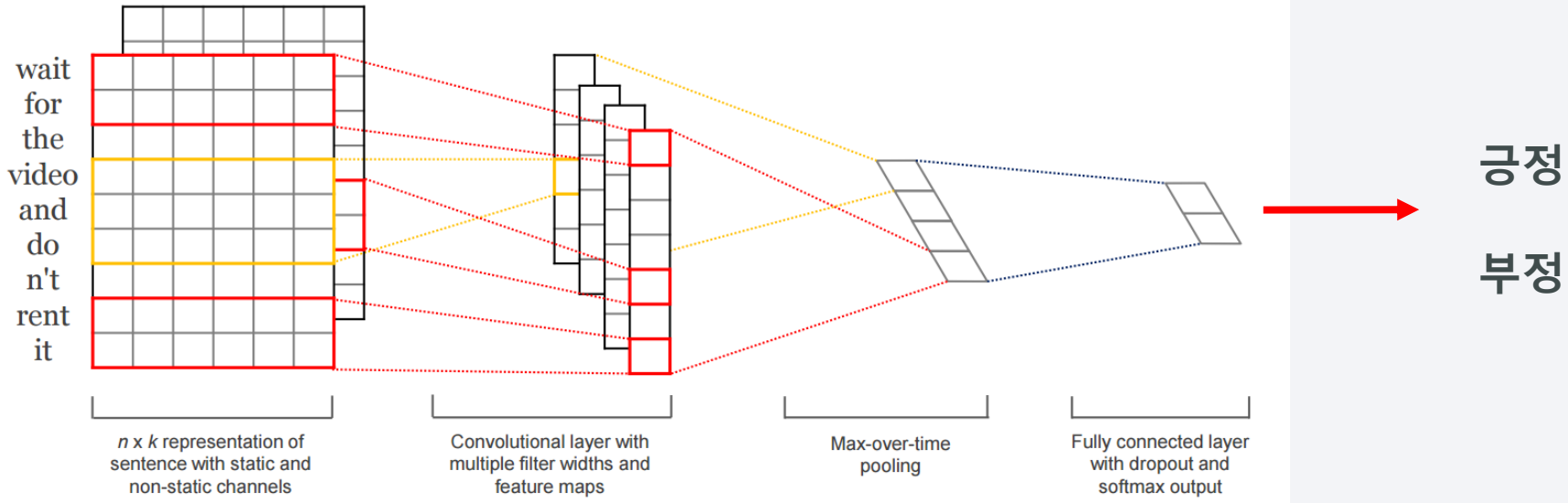


Figure 1: Model architecture with two channels for an example sentence.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882(2014).

Character-Level CNN

Movie review – 3class

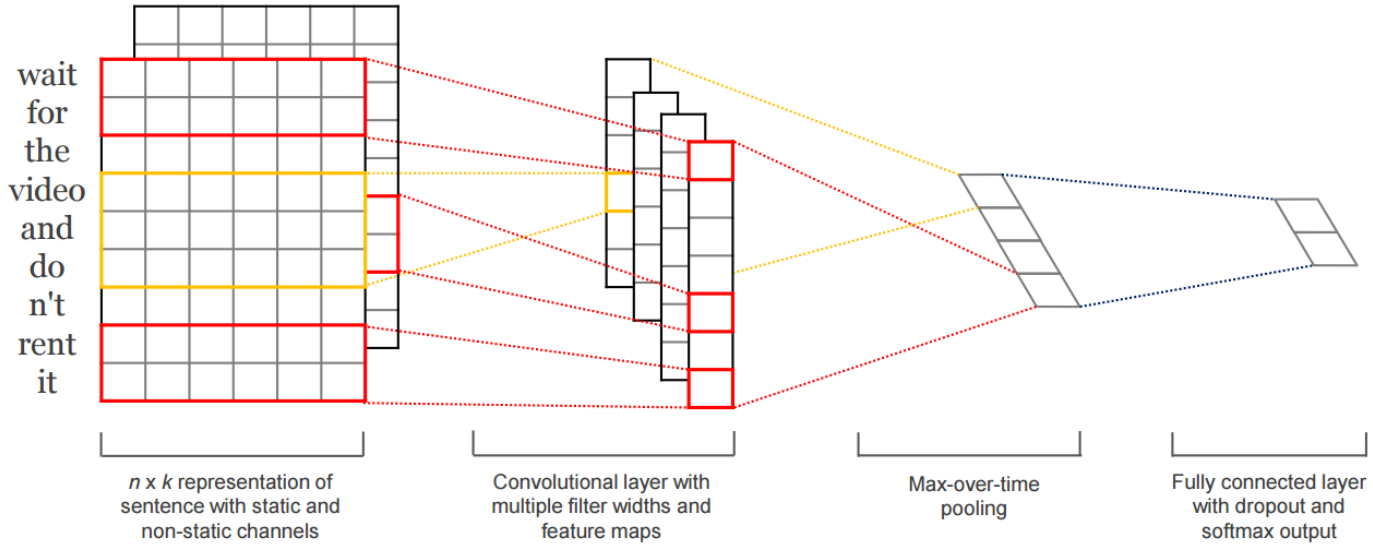


Figure 1: Model architecture with two channels for an example sentence.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882(2014).

Character-Level CNN – Result

Result

	TF_IDF (NaiveBayes)	TF_IDF (Logistic)	CNN (word level)	RNN (word level)	CNN (character level)
Article	85.35	90.04	85.02	87.56	86.27
Watcha_2class	41.11	74.93	83.15	85.35	79.87
Watcha_3class	52.62	66.45	60.06	62.36	58.21

Character-Level CNN – Result

Result

	TF_IDF (NaiveBayes)	TF_IDF (Logistic)	CNN (word level)	RNN (word level)	CNN (character level)
Article	85.35	90.04	85.02	87.56	86.27
Watcha_2class	41.11	74.93	83.15	85.35	79.87
Watcha_3class	52.62	66.45	60.06	62.36	58.21

기존 word level의 방법론과 유사한 성능을 보임

Character-Level CNN

watcha 리뷰의 특징

1. 왓챠 시스템의 특징
2. 별점과 리뷰를 통한 개인화 추천 시스템이 존재
3. 다른 영화 리뷰 사이트에 비해 비교적 정확하고 성의있는 리뷰로 구성
4. **Unrefined Document 라는 초기 가정에 적절한지 의문**

Character-Level CNN

Cohesion Tokenize

1. Character n-gram 방식
2. 연속된 글자의 연관성이 높을수록 단어일 가능성이 높음

$$\text{Cohesion(최순실)} = \{P(\text{최순}|\text{최}) * P(\text{최순실}|\text{최순})\}^{1/2}$$

3. Review data에서도 원활한 tokenizing이 수행되어 article의 word level 분석과 큰 차이점이 있는지 의문

Conclusion

1. 자소 관계를 이용한 CNN 모델이 단어 관계를 이용한 모델과 유사한 성능을 보임
2. 많은 Pre-Processing이 필요하지 않음
3. 특수기호가 담고 있는 정보도 분석이 가능함

   Ex) ^^, ㅋㅋㅋㅋㅋㅋㅋㅋ, :), ㅠㅠ
4. Unrefined 정도가 심한 Document일수록 word level과 극명한 차이를 보일 것으로 기대

Reference

- Xiang Zhang, Junbo Zhao and Yann Lecun, Character level convolutional network for text classification, NIPS 2015, Sep, 2015
- Weijie Huang, Character level convolutional network for text classification applied to Chinese corpus, Nov, 2016
- Ye Zhang and Byron Wallace, A Sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, oct, 2015
- Yoon Kim, Convolutional neural networks for sentence classification, Aug, 2014
- 조휘열, 김진화, 윤상웅, 김경민, 장병탁, 컨볼루션 신경망 기반 대용량 텍스트 데이터 분류 기술, 한국정보과학회 2015년 동계학술발표회 논문집, p. 792-794, 2015, 12
- 조휘열, 김진화, 김경민, 장정호, 엄재홍, 장병탁, 순환 신경망 기반 대용량 텍스트 데이터 분류 기술, 2016년 한국컴퓨터종합학술대회 논문집, p.968-970, 2016,6

Q & A

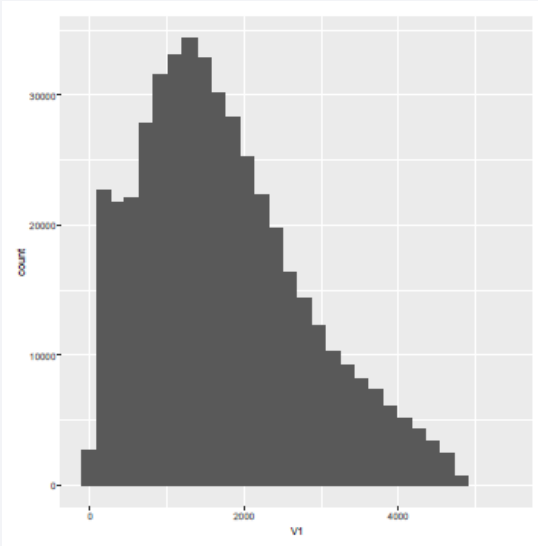
Appendix

Introduction – Data for this study

Explanation about Two Data sets

Well-refined documents

- Frequency of length of sentences



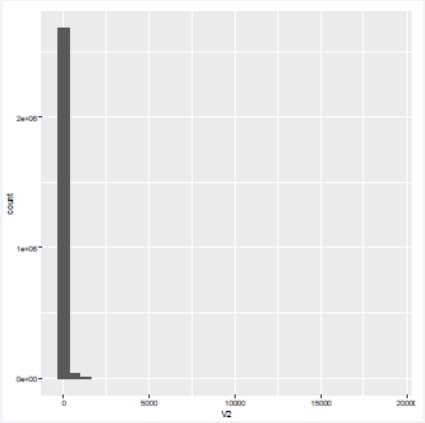
Min	Q1	Median
0	931	1,563
Mean	**Q3**	**Max**
1,726	2,365	5,364

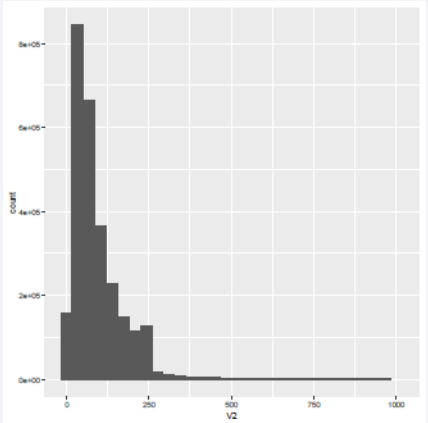
Introduction – Data for this study

Explanation about Two Data sets

Unrefined documents

- Frequency of length of sentences





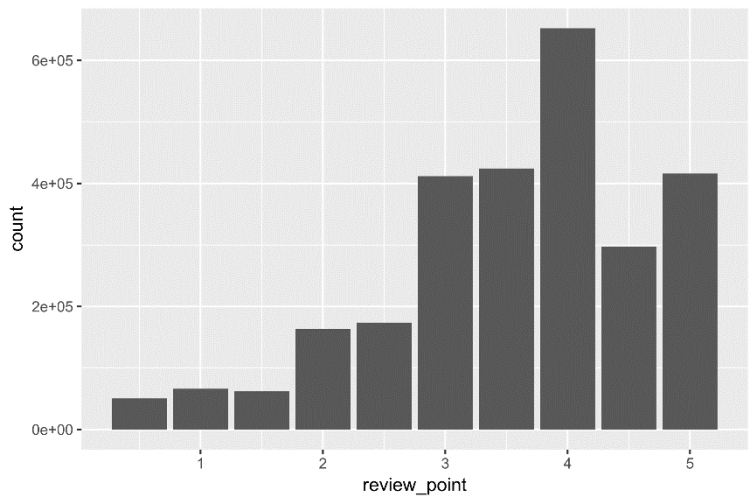
Min	Q1	Median
0	39	68

Mean	Q3	Max
95	122	18,940

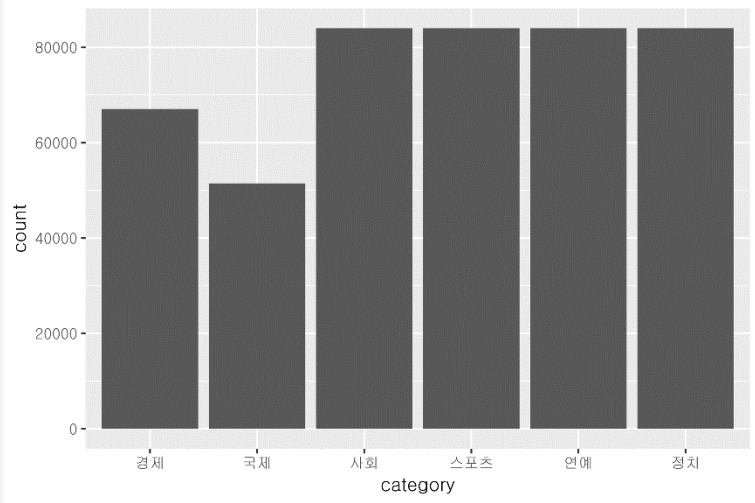
Introduction – Data for this study

Explanation about Two Data sets

0.5	1	1.5	2	2.5
50,660	66,184	62,094	163,272	173,650
3	**3.5**	**4**	**4.5**	**5**
411,757	424,378	652,250	297,327	416,096



경제	국제	사회
67,078	51,498	84,022
스포츠	**연예**	**정치**
84,022	84,022	84,022



Class 별 비율

- **Article 6class**

경제	국제	사회	스포츠	연예	정치
14%	11%	18%	18%	18%	18%

- **Movie review 2class**

긍정	부정
26%	74%

- **Movie review 3class**

긍정	중립	부정
18%	31%	51%