2020 Winter/Spring Semester URP program workshop

GRAMMATICAL AUTOCORRECTION FOR KOREAN VIA FINE-TUNING PRE-TRAINED LANGUAGE MODELS

2020.08.27 KAIST CREATIVE LEARNING B/D

20160413 Soyoung Yoon

SCHOOL OF COMPUTING

MOTIVATION & RESEARCH PROBLEM

Leaderboard

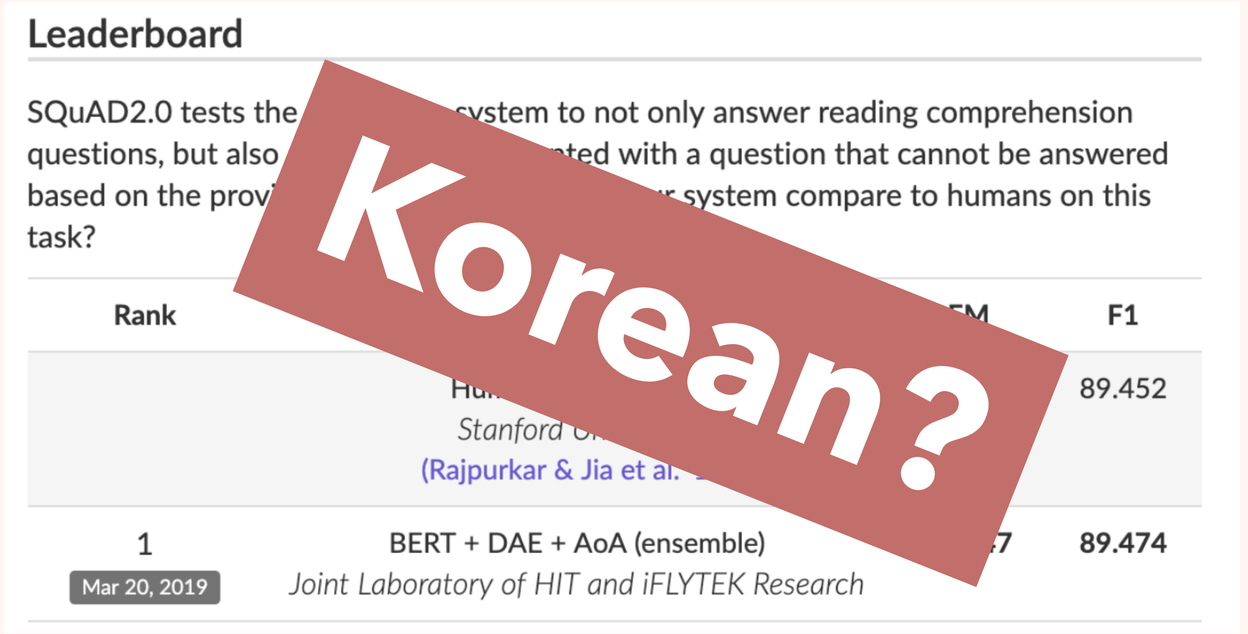
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**

NLP models exceeded the performance of human. (SQUAD 2.0)

https://medium.com/@yixing.cai/whats-aoa-a-model-that-beats-human-performance-in-squad-2-0-15422559bda2

MOTIVATION & RESEARCH PROBLEM



NLP models exceeded the performance of human. (SQUAD 2.0)

https://medium.com/@yixing.cai/whats-aoa-a-model-that-beats-human-performance-in-squad-2-0-15422559bda2

MOTIVATION & RESEARCH PROBLEM

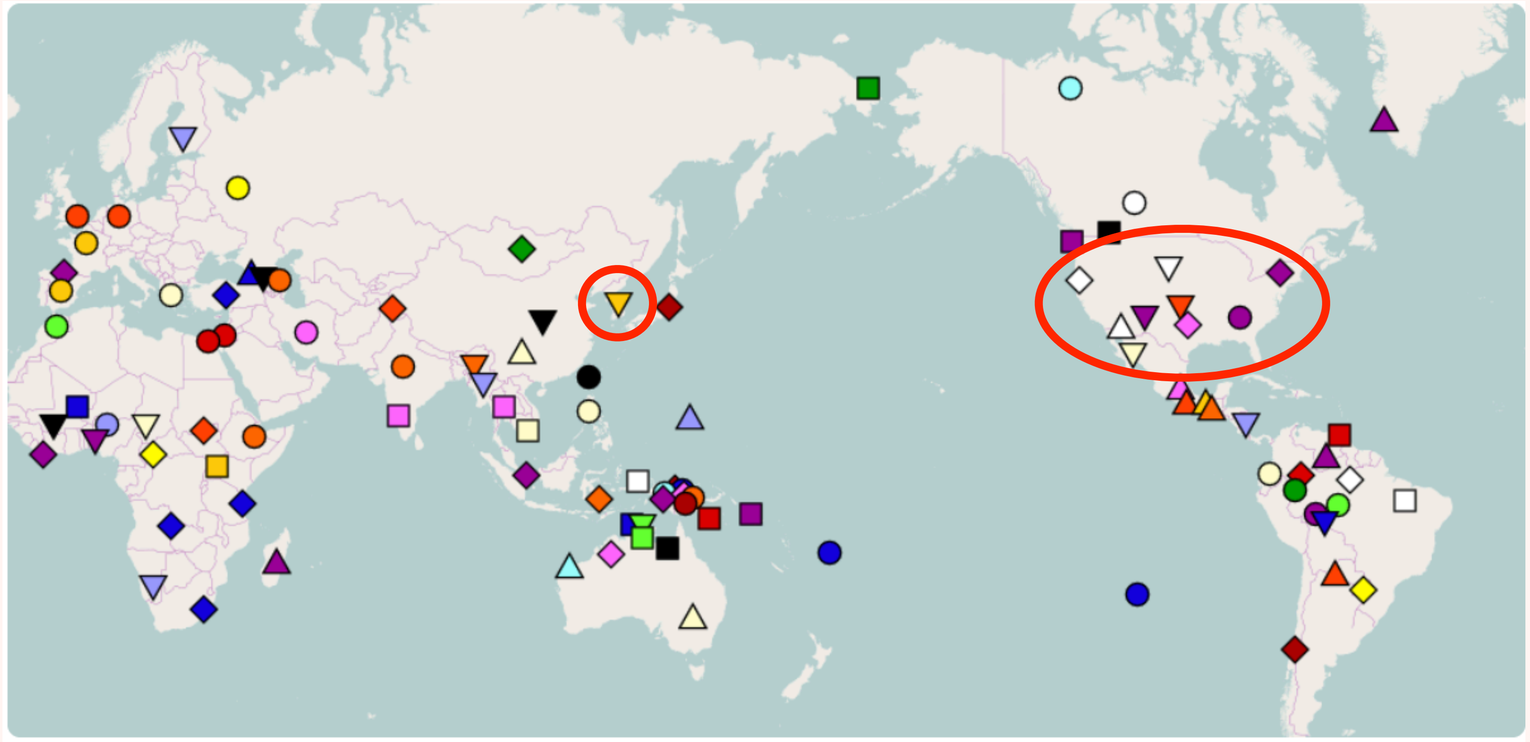
1. Fundamental difference between English and Korean

- Morphologically rich language

- Different word orders

- Even native speakers frequently make grammatical mistakes

- ....



World atlas of Language Structures (https://wals.info/)

MOTIVATION & RESEARCH PROBLEM

2. Lack of resources to train&evaluate ML model, compared with English -> we can utilize a lot more if we can pre-process raw sentences for Korean!

English clean dataset

Korean dataset
with Grammatical errors

Korean clean dataset

+ grammarly

https://news.zum.com/articles/46356026

MOTIVATION & RESEARCH PROBLEM

1. Collect Korean Corpora that can be used for Korean Grammatical Error Correction tasks

Pairs of Corrupted Sentence(나 운전 배워 안했어요) <=> Corrected Sentence(나는 운전을 안 배웠어요)

2. Provide baseline Korean GEC models and evaluate them

https://news.zum.com/articles/46356026

PREVIOUS APPROACHES

Almost no research on Korean Grammatical Error Correction

Only on narrow subset of the problem (particle error (조사) detection)

- Developing methodology for **Korean particle error detection**, Sixth Workshop on Innovative Use of NLP for Building Educational Applications, 2011.

- Detecting and Correcting Learner **Korean Particle** Omission Errors, IJNLP 2013

=> Extend to general Korean Grammatical Errors

- Improving Grammatical Error Correction via Pre-Training a **Copy-Augmented Architecture** with Unlabeled Data, NAACL 2019

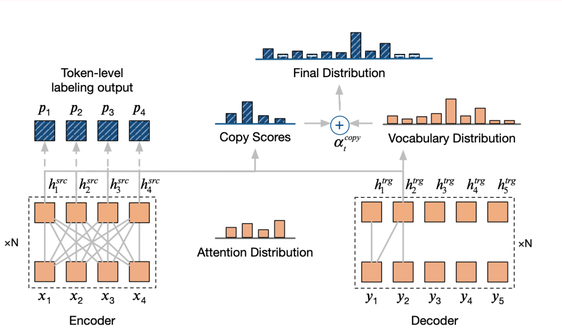


Figure 1: Copy-Augmented Architecture.

- **BART**: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, ACL 2020

=> Baseline models for GEC

OUR APPROACH

Collected Corpora for Korean GEC from various sources:

1. Kor-WikiEd(774,859), from Wikipedia

2. Kor-Lang8(100,080), from lang8

3. Kor-Native(18,198), from National Institute of Korean Language(국립국어원)

4. Kor-Learner(22,789). from 국립국어원 **-> Total of 915,926 pairs**

OUR APPROACH

Collected Corpora for Korean GEC from various sources:

1. Kor-WikiEd(774,859)

2. Kor-Lang8(100,080)

3. Kor-Native(18,198)

4. Kor-Learner(22,789).

Red: my main contribution
Blue: contributed together with Sungjoon Park

-> Total of 915,926 pairs

Then, we pre-processed the collected data, trained baseline models, and evaluated model scores for Korean GEC.

Kor-WikiEd (774,859 pairs)

Previous work: Using Wikipedia Edits in Low Resource Grammatical Error Correction, EMNLP workshop 2018.

Used Korean wikipedia edit history

Filter out by sequence length and Levenshtein distance

Used for language model pre-training

Corrupted (source)	Corrected (target)
1946년 경상남도 김해군에서 태어났다	1946년 경상남도 김해에서 태어났다
감마 함수는 다음과 같은 함수 방정식을 만족시킨다	감마 함수는 다음과 같은 방정식을 만족시킨다
선조 는 조선의 제14대 왕이다	선조는 조선의 제 14대 왕이다
백남준은 한국 태생의 미국 현대 미술가이다	백남준은 대한민국 태생의 미국 현대 미술가이다

Kor-Lang8 (100,080 pairs)

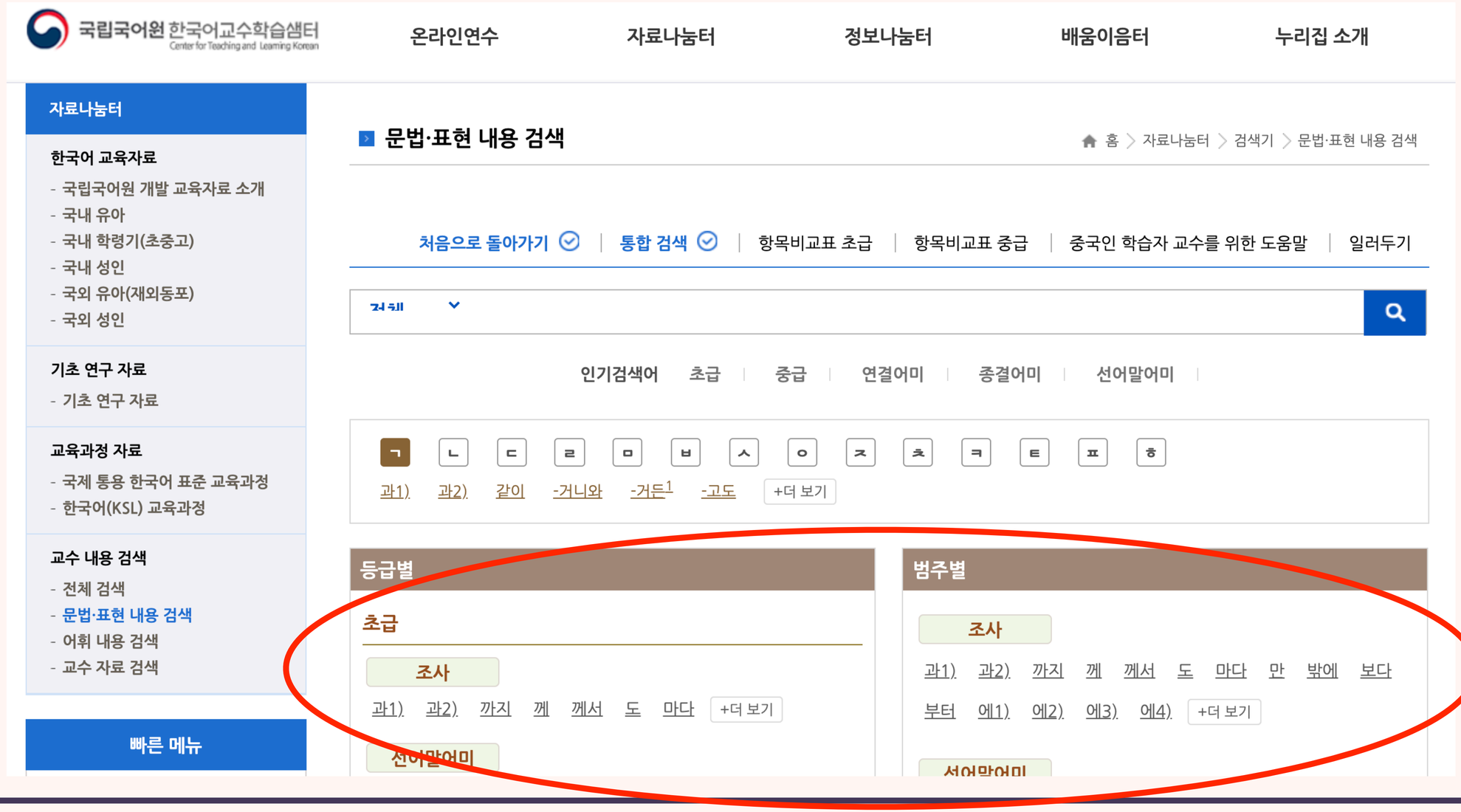
Filter out posts that are not sentence corrections (e.g. language-related questions)

Among those, filter out by the pre-post ratio & longest common subsequence

Used for language model fine-tuning

Corrupted (source)	Corrected (target)
지난날 인타넷으로 찾아냈다.	지난달 인터넷 검색으로 찾아냈다.
제가 정식으로 한국말 배워 안했어요.	제가 정식으로 한국어를 배우지 못했어요.
등산하고 바비큐도 하다.	등산도 하고 바비큐도 한다.
고맙다랑 고마워요랑 감사합니다!	고맙다! 고마워요! 감사합니다!

Kor-Native (18,198 pairs) - Generate wrong sentence FROM correct sentence!



Kor-Native (18,198 pairs)

문장 구성 정보

접기

① 과거는 '-었-'을 붙여 '-었군'으로 쓴다.

국립국어원 한국어교수학습샘터 Center for Teaching and Learning Korean

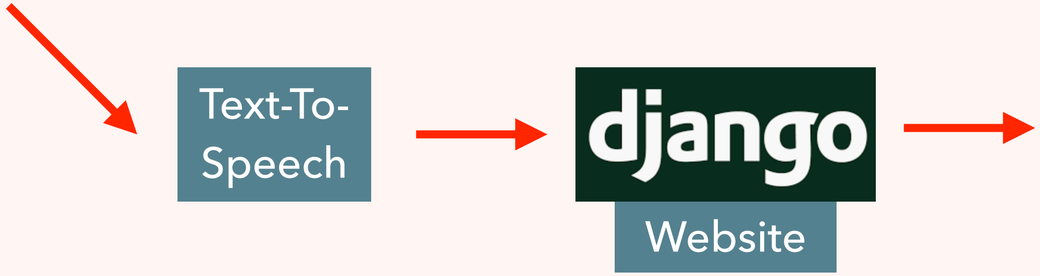
예) 작년에는 서울에 눈이 많이 **왔군요**.
어제 수지가 열심히 **공부했군요**.
줄리아 씨가 어렸을 때 정말 **귀여웠군요**.

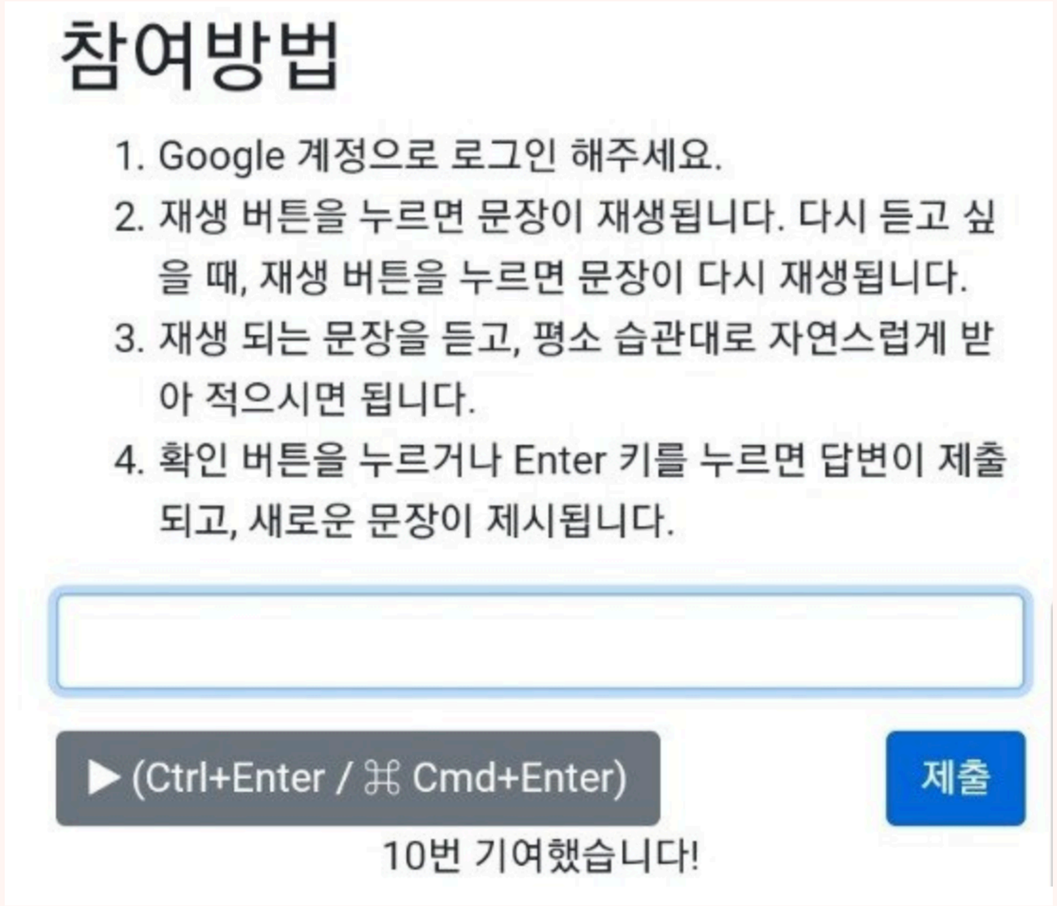
② 추측을 나타낼 때는 '-겠-'을 붙여 '-겠군'으로 쓴다.

예) 지금쯤 현우 씨는 부산에 **가겠군요**.
주말이라서 도서관에 사람이 **없겠군요**.
12월이니 한국은 날씨가 **춥겠군요**.

③ 동사의 높임은 '-시는군'을, 형용사의 높임은 '-시군'으로 쓴다.

예) 선생님께서 이제 점심을 **드시는군요**.
할아버지께서 제 방에서 **주무시는군요**.
어머니께서 어제부터 많이 **아프시군요**.
수지 씨의 아버지께서 키가 정말 **크시군요**.





참여방법

1. Google 계정으로 로그인 해주세요.
2. 재생 버튼을 누르면 문장이 재생됩니다. 다시 듣고 싶을 때, 재생 버튼을 누르면 문장이 다시 재생됩니다.
3. 재생 되는 문장을 듣고, 평소 습관대로 자연스럽게 받아 적으시면 됩니다.
4. 확인 버튼을 누르거나 Enter 키를 누르면 답변이 제출되고, 새로운 문장이 제시됩니다.

▶ (Ctrl+Enter / ⌘ Cmd+Enter)

제출

10번 기여했습니다!

Think the other way around -
It was hard to collect 'grammatically wrong' and 'corrected' pairs, so let's generate it!

Kor-Native (18,198 pairs) - advertised at many community sites & used crowdworks

카이스트 대신 전해드립니다
2019년 12월 22일 ·

[한국어 오탈자 교정을 위한 데이터 수집을 도와주세요!!]

안녕하세요, 인공지능 기반 한국어 맞춤법 검사기를 개발하
입니다.
현재 자연어처리 분야에서 사용하기 위한 영어 데이터는 많
따라서 한국어 맞춤법 검사기 개발을 위해 공개화된 한국어
발하였습니다.

간단하게 들리는 대로 평소 습관대로 작성해주시면 됩니다
혹시 한국어 corpus를 수집하는데 기여해 주실 분이 계신

시험이 끝나고 심심하신 분들이 계신다면, 부탁드리겠습니

https://korpus.junheecho.com/

감사합니다.

#잡담
https://talkyou.in/pages/KaDaejeon/posts/34430
학사과정 16학번 윤소영

KORPUS.JUNHEECHO.COM
한국어 말뭉치 모으기

정우진, 장봉준, 외 11명

dcinside.com 갤러리

갤러리 &

갤러리 + | m.갤러리 | 갤로그 | 뉴스 | 이벤트 | 만두몰 | 디시위키

언어 갤러리

최근 방문 갤러리 | 언어 | 스트리머 | 201212~

한국어 자연어처리 데이터 수집 관련

DeepNlP(115.145) | 2019.11.25 18:24:04

https://korpus.junheecho.com/

안녕하세요.

인공지능 기반 한국어 맞춤법 검사기를 개발하는 프로젝트를 위해 한국어를 모국어로 사용
자연스러운 언어 습관에 따라 (어법 오류가 포함된) 문장을 연구 및 비상업적 목적으로 수집

참가자 분들이 수집된 모든 문장은 전처리를 거쳐서, 인공지능 기반 한국어 맞춤법 검
연구자와 개발자를 포함한 누구나 사용할 수 있도록 오픈소스로 공개할 예정입니다.

crowdworks

nate 판 뉴스 스포츠 연예

홈 톡톡 판포토 판

윤소영
2019년 11월 24일 ·

제가 이번에 참여하고 있는 연구에서 한국어 말뭉치를 모으고 있습니다. 간단한 문장을 평소 적는 습관대로 받아쓰기 하시면 됩니다!! 심심하신 분들 많이 참여 부탁 드립니다... ㅎㅎ

KORPUS.JUNHEECHO.COM
한국어 말뭉치 모으기

Sungjoon Park, 이명신, 외 10명 댓글 4개

마춤뻡빌런을 위하여

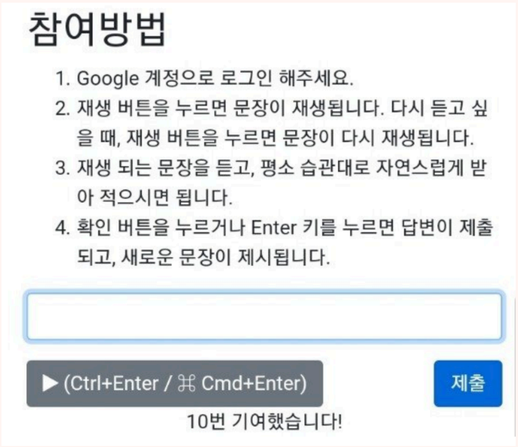
Deepnlp (판) 2020.01.04 12:57

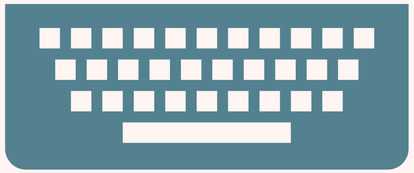
한줄요약 : 한글 자연어 처리(맞춤법) 연구 하고 있습니다. 도와주

최근 가장 많이 사랑받는 빌런은 누구일까요?
모두 같은 빌런을 떠올리진 않더라도 그 중에 분명히 '조커'가 있

Kor-Native (18,198 pairs)

작년에는 서울에 눈이 많이 왔군요.

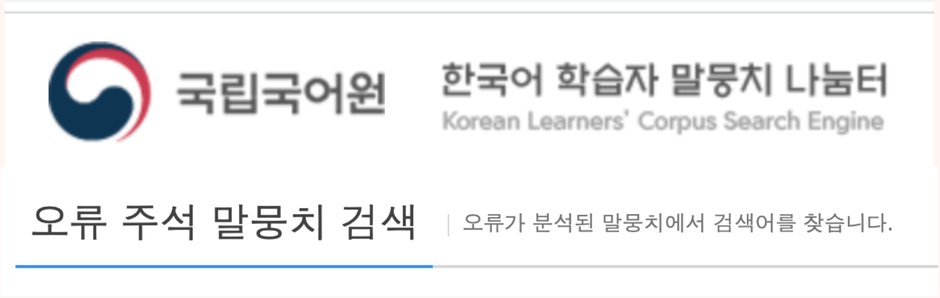




작년에는서울에 눈이 많이 왓군여.

Corrupted (source)	Corrected (target)
한국에서 뿐만 아니라 세계적으로도 유명해여.	한국에서뿐만 아니라 세계적으로도 유명해요.
이 도시는 공기도 나뿌다.	이 도시는 공기도 나쁘다.
수지씨가 예쁩니까?	수지씨가 예쁩니까?
바다에서 예쁜 조개를 주어가지고 목걸이를 만들어여.	바다에서 예쁜 조개를 주워가지고 목걸이를 만들어요.

Kor-Learner (22,789 pairs)



+ Preprocessing, filter & refine to use data for training

Corrupted (source)	Corrected (target)
우리 모두 꿈을 잘 이루어졌으면 좋겠습니다.	우리 모두의 꿈이 잘 이루어졌으면 좋겠습니다.
그 제 꿈이 교수기 도는 것입니다.	제 꿈은 교수가 되는 것입니다.
밥을 먹으로 식당에 가요.	밥을 먹으러 식당에 가요.
그래서 집 관리비가 절약할 수 있어요.	그래서 집 관리비를 절약할 수 있어요.

Tokenization - 2 ways

Vocabulary size: 10000, used sentencepiece

Char level: _이, _있, _대, _전, 이다, _보, 하는, _위, _자, _주, 였다, 으며, _사람, _남, …

***Jamo level: _ㅆㄷㅏ, ㅘ_, ㅏㅇ, ㅏㄹ_, ㅗㅇ_, ㄴ_, ㅣ_, ㅏ_, ㅡㅇ_, ㅗㄴ_, _ㄱ ㅓㅅ, …**

*Subword-level Word Vector Representations for Korean, EMNLP 2018

```
지금는 한국말 잘 못해요 => [“_지금”, “는”, “_한국”, “말”, “_잘”, “_못”, “해요 _”]
```

EVALUATION

Model	Seg.	Kor-WikiEd	Kor-Lang8	Kor-Native	Kor-Learner	
Copy-Aug. Transformer						–
BART						–
BART-pretrained						–
BART-union						–

Baseline model 1 used for English GEC

Baseline model 2 used for English GEC

Pretrained on Kor-WikiEd dataset

Fine-tuned on union of Kor-Lang8, Kor-Native, and Kor-Learner

EVALUATION

Model	Seg. Level	Kor-WikiEd		Kor-Lang8		Kor-Native		Kor-Learner				
		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU	M^2 Pre.	M^2 Rec.	$F_{0.5}$
Copy-Aug. Transformer	Char	94.7	93.9	80.8	79.0	95.9	95.4	87.9	86.8	46.17	22.09	37.90
	Jamo	93.3	92.4	76.4	74.9	91.0	90.3	83.8	83.2	44.13	19.22	35.05
BART	Char	94.5	93.7	80.9	79.2	95.3	94.5	87.2	86.1	45.15	22.01	37.31
	Jamo	93.4	92.6	76.4	75.1	93.6	92.9	83.4	82.7	43.13	19.24	34.55
BART-pretrained	Char	-	-	80.5	78.8	95.6	95.1	88.2	87.2	49.28	26.66	42.13
	Jamo	-	-	75.9	74.5	93.9	93.4	85.2	84.6	50.09	25.53	42.01
BART-union	Char	-	-	**84.2**	**82.6**	**97.5**	**97.2**	**92.8**	**92.1**	79.25	49.68	70.82
	Jamo	-	-	80.8	79.5	96.4	96.0	91.2	90.7	**82.73**	**52.08**	**74.02**

EVALUATION

Examples of the outputs:

outputs can be different from target,
but should we see them as “Wrong” or “Uncorrected”?

Corrupted Text (Source)	Corrected Text (Target)		Model Output
여러분 겨울이 좋아합니까?	여러분 겨울을 좋아합니까?	=	여러분 겨울을 좋아합니까?
200 경찰에 이상이 되었을 것이다.	경찰이 200명 이상 이었을 것이다.	=	경찰이 200명 이상 이었을 것이다.
2007년 아버지가 돌아 었습니다.	2007년에 아버지께서 돌아가셨습니다.		2007년 아버지가 돌아가셨습니다.
기회를 어떻게 찾을 수 있어요?	기회를 어떻게 찾을 수 있나요?		기회를 어떻게 찾을 수 있을까요?

CONTRIBUTION & FUTURE WORK

Submitted to EMNLP 2020

1. Collect Korean Corpora that can be used for Korean Grammatical Error Correction tasks

2. Provide baseline Korean GEC models and evaluate them

+ Better Evaluation for Korean Grammatical Error

+ Analysis on characteristics on Korean

+ Improvement of models by applying data augmentation techniques

COLLABORATORS

This work was done with the help of great people, including Alice Oh, Sungjoon Park, Gyu Tae Kim, Sumin Lim, Jae Yun Kim, Junhee Cho, Geunhoo Kim, and Gyuwan Kim.
I also thank the URP program for supporting me complete this work.