NUCLE Corpus

Patrick Jen-Yu LI
M2 Didactiques des Langues
Université Rennes 2

Supervisor: Thomas Gaillat

Outline

Introduction

Error tag set

Data challenge

Introduction

NUCLE stands for “National University of Singapore (NUS) Corpus of Learner English”

A collection of 1,414 essays (1.2 million words) written by students who are non-native English speakers

Topics include environmental pollution, healthcare, etc.

The grammatical errors in these essays have been corrected by professional instructors and saved as stand-off annotations in SGML format

Available after filling the agreement form

https://sterling8.d2.comp.nus.edu.sg/nucle_download/nucle.php

Directories and files

\NUCLE\release3.3

\bea2019

- nucle.train.gold.bea19.m2
- readme.txt

\data

- conll14st-preprocessed.conll.ann
- conll14st-preprocessed.m2
- nucle3.2.sgml

\scripts

- iparser.py
- nucle_doc.py
- nuclesgmlparser.py
- parser_feature.py
- preprocess.py
- README

nucle3.2.sgml

<DOC nid="829">

<TEXT>

<TITLE>

CREATING A HABITABLE ENVIRONMENT

</TITLE>

<P>

Humans have many basic needs and one of them is to have an environment that can sustain their lives. Our current population is 6 billion people and it is still growing exponentially. This will, if not already, caused problems as there are very limited spaces for us. The solution can be obtain by using technology to achieve a better usage of space that we have and resolve the problems in lands that inhospitable such as desserts and swamps.

</P>

...and other paragraphs...

</TEXT>

<ANNOTATION teacher_id="172">

<MISTAKE start_par="1" start_off="210" end_par="1" end_off="216">

<TYPE>Vform</TYPE>

<CORRECTION>cause</CORRECTION>

</MISTAKE>

...and other mistakes...

Error Tag Set

28 error tags and the corresponding examples are shown in the following slides

Type	Description	Example
Vt	Verb tense	Medical technology during that time [**is** → was] not advanced enough to cure him.
Vm	Verb modal	Although the problem [**would** → may] not be serious, people [would → might] still be afraid.
V0	Missing verb	However, there are also a great number of people [**who** → who are] against this technology.
Vform	Verb form	A study in 2010 [**shown** → showed] that patients recover faster when surrounded by family members.
SVA	Subject-verb agreement	The benefits of disclosing genetic risk information [**outweighs** → outweigh] the costs.
ArtOrDet	Article or determiner	It is obvious to see that [**internet** → the internet] saves people time and also connects people globally.
Nn	Noun number	A carrier may consider not having any [**child** → children] after getting married.
Npos	Noun possessive	Someone should tell the [**carriers** → carrier's] relatives about the genetic problem.
Pform	Pronoun form	A couple should run a few tests to see if [**their** → they] have any genetic diseases beforehand.

Pref	Pronoun reference	It is everyone's duty to ensure that [**he or she** → they] undergo regular health checks.
Prep	Preposition	This essay will [**discuss about** → discuss] whether a carrier should tell his relatives or not.
Wci	Wrong collocation/idiom	Early examination is [**healthy** → advisable] and will cast away unwanted doubts.
Wa	Acronyms	After [**WOWII** → World War II], the population of China decreased rapidly.
Wform	Word form	The sense of [**guilty** → guilt] can be more than expected.
Wtone	Tone (formal/informal)	[**It's** → It is] our family and relatives that bring us up.
Srun	Run-on sentences, comma splices	The issue is highly [**debatable, a** → debatable. A] genetic risk could come from either side of the family.
Smod	Dangling modifiers	[**Undeniable**, → It is undeniable that] it becomes addictive when we spend more time socializing virtually.
Spar	Parallelism	We must pay attention to this information and [**assisting** → assist] those who are at risk.
Sfrag	Sentence fragment	**However, from the ethical point of view.**

Ssub	Subordinate clause	This is an issue [**needs** → that needs] to be addressed.
WOinc	Incorrect word order	[**Someone having what kind of disease** → What kind of disease someone has] is a matter of their own privacy.
WOadv	Incorrect adjective/adverb order	In conclusion, [**personally I** → I personally] feel that it is important to tell one's family members.
Trans	Linking words/phrases	It is sometimes hard to find [**out** → out if] one has this disease.
Mec	Spelling, punctuation capitalization, etc.	This knowledge [**maybe relavant** → may be relevant] to them.
Rloc−	Redundancy	It is up to the [**patient's own choice** → patient] to disclose information.
Cit	Citation	*Poor citation practice.
Others	Other errors	*An error that does not fit into any other category but can still be corrected.
Um	Unclear meaning	Genetic disease has a close relationship with the **born** gene. (no correction possible without further clarification.)

Data Challenge

CoNLL-2014 Shared Task: Grammatical Error Correction

https://www.comp.nus.edu.sg/~nlp/conll14st.html

Building Educational Applications (BEA) 2019 Shared Task: Grammatical Error Correction (GEC)

https://www.cl.cam.ac.uk/research/nl/bea2019st/

"The aim of the shared task is to correct all types of errors in written text. This includes grammatical, lexical and orthographical errors. Participants will be provided with plain text files as input, one tokenised sentence per line, and are expected to produce equivalent corrected text files as output."

Three tracks on the Codalab competition platform:

- Restricted Track
- Unrestricted Track
- Low Resource Track

Annotated datasets

W&I+LOCNESS v2.1:

Cambridge English Write & Improve

LOCNESS

FCE v2.1:

First Certificate in English corpus is a subset of the Cambridge Learner Corpus (CLC)

Lang-8 Corpus of Learner English

NUCLE

Results

1. Restricted Track

#	User	Team Name	TP	FP	FN	P	R	$F_{0.5}$	Detailed Results
1	romang	UEDIN-MS	3127	1199	2074	72.28	60.12	**69.47**	View
2	yjchoe33	Kakao&Brain	2709	894	2510	75.19	51.91	**69.00**	View
3	goo2go	LAIX	2618	960	2671	73.17	49.50	**66.78**	View
4	HelenY	CAMB-CLED	2924	1224	2386	70.49	55.07	**66.75**	View
5	seanxu1015	Shuyao	2926	1244	2357	70.17	55.39	**66.61**	View
6	BUAA_WJP	YDGEC	2815	1205	2487	70.02	53.09	**65.83**	View
7	awasthiabhijeet05	ML@IITB	3678	1920	2340	65.70	61.12	**64.73**	View
8	fs439	CAMB-CUED	2929	1459	2502	66.75	53.93	**63.72**	View
9	tomoyamizumoto	AIP-Tohoku	1972	902	2705	68.62	42.16	**60.97**	View
10	arahusky	UFAL, Charles University, Prague	1941	942	2867	67.33	40.37	**59.39**	View
11	liuwangwang	CVTE-NLP	1739	811	2744	68.20	38.79	**59.22**	View
12	hsamswcc	BLCU	2554	1646	2432	60.81	51.22	**58.62**	View
13	yoav_kantor	IBM Research AI - HRL	1819	1044	3047	63.53	37.38	**55.74**	View
14	Masahiro	TMU	2720	2325	2546	53.91	51.65	**53.45**	View
15	qiuwenbo		1428	854	2968	62.58	32.48	**52.80**	View
16	cehinson	NLG NTU	1833	1873	2939	49.46	38.41	**46.77**	View
17	apurva.nagvenkar	CAI	2002	2168	2759	48.01	42.05	**46.69**	View
18	davidzhao	PKU	1401	1265	2955	52.55	32.16	**46.64**	View
19	SolomonLab	SolomonLab	1760	2161	2678	44.89	39.66	**43.73**	View
20	mengyang	Buffalo	604	350	3311	63.31	15.43	**39.06**	View
21	nihalnayak	Ramaiah	829	7656	3516	9.77	19.08	**10.83**	View

2. **Unrestricted Track**

#	User	Team Name	TP	FP	FN	P	R	$F_{0.5}$	Detailed Results
1	goo2go	LAIX	2618	960	2671	73.17	49.50	**66.78**	View
2	tomoyamizumoto	AIP-Tohoku	2589	1078	2484	70.60	51.03	**65.57**	View
3	arahusky	UFAL, Charles University, Prague	2812	1313	2469	68.17	53.25	**64.55**	View
4	hsamswcc	BLCU	3051	2007	2357	60.32	56.42	**59.50**	View
5	gurunathp	Aparecium	1585	1077	2787	59.54	36.25	**52.76**	View
6	mengyang	Buffalo	699	374	3265	65.14	17.63	**42.33**	View
7	nihalnayak	Ramaiah	1161	8062	3480	12.59	25.02	**13.98**	View

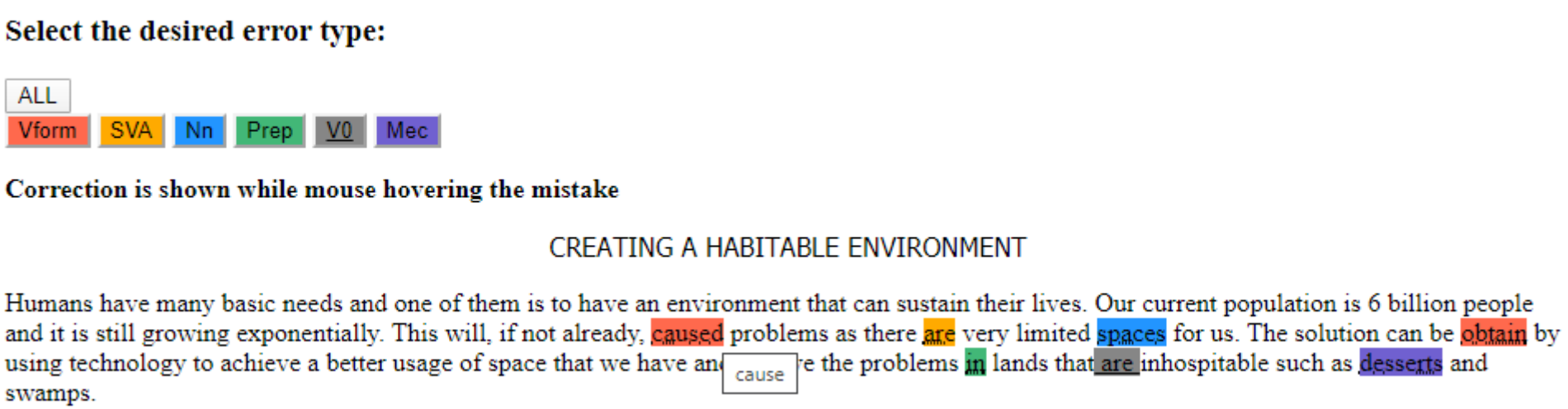
3. **Low Resource Track**

#	User	Team Name	TP	FP	FN	P	R	$F_{0.5}$	Detailed Results
1	romang	UEDIN-MS	2312	982	2506	70.19	47.99	**64.24**	View
2	JiyeonHam	Kakao&Brain	2412	1413	2797	63.06	46.30	**58.80**	View
3	goo2go	LAIX	1443	884	3175	62.01	31.25	**51.81**	View
4	fs439	CAMB-CUED	1814	1450	2956	55.58	38.03	**50.88**	View
5	arahusky	UFAL, Charles University, Prague	1245	1222	2993	50.47	29.38	**44.13**	View
6	simonHFL	Siteimprove	1299	1619	3199	44.52	28.88	**40.17**	View
7	Bohdan_Didenk	WebSpellChecker.com	2363	3719	3031	38.85	43.81	**39.75**	View
8	Satoru	TMU	1638	4314	3486	27.52	31.97	**28.31**	View
9	mengyang	Buffalo	446	1243	3556	26.41	11.14	**20.73**	View

My recent works

Visualization of the texts
Reorganization according to error tags
Wrong collocation detection

Visualization of the texts



- Use HTML, CSS, and JavaScript to implement a text browser

Reorganization according to error tags

1	TextID	ErrType	Start	End	Correction	Sentence
13	829	Wci	5	6	formation and growth	This caused problem like the appearance of slums which most of the time is not safe due to the unhealth
20	829	Wci	27	28	a greater	The only way to satisfy the increasing demands of space is by achieving a better usage of the land like d
22	829	Wci	5	6	use	It is also important to create a better material that can support the buildings despite any natural disaster l
25	829	Wci	14	15	during	It is also important to create a better material that can support the buildings despite any natural disaster l
33	829	Wci	20	27	restore the land to a livable state	Countries with a lot of inhospitable space need not only to achieve a better space usage , but also to refo
35	829	Wci	40	44	quality of the land	Countries with a lot of inhospitable space need not only to achieve a better space usage , but also to refo
37	829	Wci	10	11	transform	For example , countries with a lot of deserts can terraform their desert to increase their habitable land anc
50	829	Wci	14	15	increasing	As the number of people grows , the need of habitable environment is unquestionably essential .
52	829	Wci	12	13	environment	In this era , Engineering designs can help to provide more habitable accommodation by designing a stron
55	829	Wci	23	24	build	In this era , Engineering designs can help to provide more habitable accommodation by designing a stron
59	829	Wci	52	53	otherwise uninhabitable	In this era , Engineering designs can help to provide more habitable accommodation by designing a stron

Wrong collocation detection

Use NLTK library for collocation detection
Identify wrong collocation, firstly the Verb-Noun form

Main reference

Hwee Tou Ng *et. al.*, 2014. “The CoNLL-2014 Shared Task on Grammatical Error Correction”, Proceedings of the Eighth Conference on Computational Natural Language Learning: Shared Task, pages 1–14. Baltimore, Maryland, 26-27 July 2014