The WikEd Error Corpus

Wikipedia and Grammatical Error Correction (GEC)

Once upon a time...

Taiwanese graduate students often write...

We will discuss about the issue in ...

from https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=XugHEe/webmge?mode=basic

However...

We will <u>discuss about the issue</u> in ...

discuss

verb [T]

UK

/dɪˈskʌs/ **US**

/dɪˈskʌs/

Cambridge Dictionary

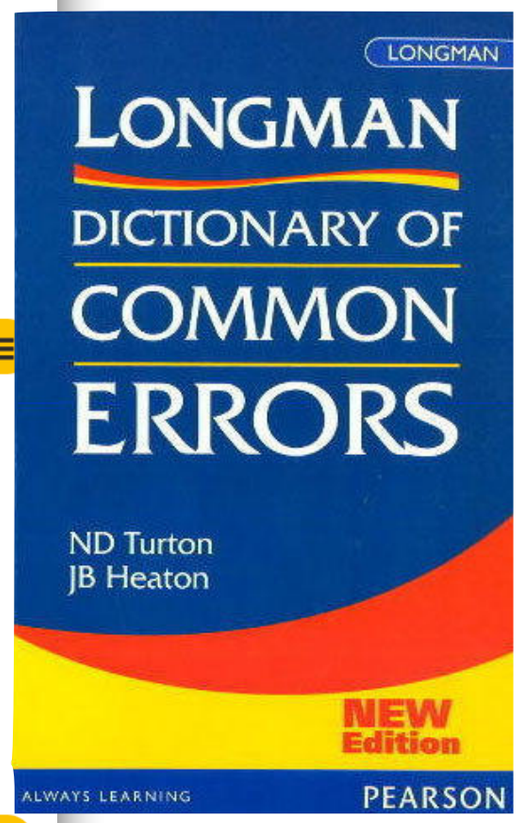
A2

to talk about a subject with someone and tell each other your ideas or opinions

討論，議論，談論

- *The police want to discuss these recent racist attacks **with** local people.*
  警方希望與當地人討論近期發生的種族主義襲擊事件。

to talk or write about a subject in detail, especially considering different ideas and



GEC is ...

We will discuss ~~about~~ the issue in ...

- Correcting errors in a sentence
- How?

  Rule-based

  Statistical approaches (language model, classifiers)

  Machine Translation approaches (state-of-the-art)

Challenges Faced in GEC

- Neural networks models typically require large data to train
- Good learner corpora are small

  FCE (~30K), NUCLE (~50K)

- Large learner corpora are noisy (Lang-8, EFCAMDAT)

  discuss about something -> discuss something
  162/852 ~= 19%

Wikipedia

- A free online encyclopedia that anyone can edit
- Very large

  Almost 6M articles (~150M sentences)

- A lot of revision history !

  Including information supplements and error correction

Revision History

− ==Rule-based vs. statistical NLP{{anchor|**Statistical_natural_language_processing_**(SNLP)}}==	+ ==Rule-based vs. statistical NLP{{anchor|**Statistical natural language processing** (SNLP)}}==
− In the early days, many language-processing systems were designed by hand-coding a set of rules,<ref name=winograd:shrdlu71>Winograd, Terry (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. http://hci.stanford.edu/winograd/shrdlu/</ref><ref name=schank77>Roger C. Schank and Robert P. Abelson (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures</ref> **e.g.** by writing grammars or devising heuristic rules for [[stemming]].	+ In the early days, many language-processing systems were designed by hand-coding a set of rules:<ref name=winograd:shrdlu71>Winograd, Terry (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. http://hci.stanford.edu/winograd/shrdlu/</ref><ref name=schank77>Roger C. Schank and Robert P. Abelson (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures</ref> **such as** by writing grammars or devising heuristic rules for [[stemming]].
− However, this is rarely robust to natural language variation.	

https://en.wikipedia.org/w/index.php?title=Natural_language_processing&type=revision&diff=919017262&oldid=913095969

The WikEd Error Corpus

- Grundkiewicz, R., & Junczys-Dowmunt, M. (2014, September). The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing* (pp. 478-490). Springer, Cham.

- Code (py2.7) & data

  https://github.com/snukky/wikiedits

Extract Edits from Wiki

- Iterate over any two adjacent revisions
- To remove vandalism, skip comments containing:
  - "reverting after (...)"
  - "remove vandalism"
  - "undo vandal's edits"
  - "delete stupid joke"

Get Hands Dirty

- https://github.com/dspp779/wikiedits (py3)

  $ pip install https://github.com/dspp779/wikiedits/archive/master.zip

- Run with the test file (tests/data/enwiki-20140102.tiny.xml)

  $ wiki_edits.py <path_to_wikixml>

Extracted edits

- You can use rsync to [-donload-] {+download+} the database .
- There [-is-] {+are+} also [-a-] two computer games based on the movie .
- These anarchists [-argue against-] {+oppose the+} regulation of corporations .
- David Zuckerman is a writer and [-producer-] {+poopface+} for television shows

```
$ cat edit_file | bin/convert_to_wdiff.py
```

WikEd vs. Lang-8

Statistics	WikEd 0.9	L8-NAIST
# sentences	12.13	2.57
# tokens (source side)	292.57	28.51
# edits	16.01	3.41
# edits per sentence	1.32/sent.	1.33/sent.
% sentences with $\geq$1 edit	91.79%	53.86%

WikEd vs. Lang-8

System	4×2-CV	ST-2013
NUCLE	22.10	27.62
+WikEd	18.21	23.63
+L8-NAIST	24.44	34.06

The WikEd Error Corpus is not an ESL error corpus

Filter Edits

- Remove sentences containing:
  - vulgarisms
  - fragments of markup (e.g., <ref>,
, [http:])
  - only changes in dates or numerical values
  - non-words tokens over the ratio (>0.5)

wikiedits/edit_filter.py

WikEd vs. Lang-8 (with select)

System	4×2-CV	ST-2013
NUCLE	22.10	27.62
+WikEd	18.21	23.63
+Select	**24.33**	**30.06**
+L8-NAIST	24.44	34.06
+Select	**26.40**	**34.15**

ERRANT

- Bryant, C., Felice, M., & Briscoe, E. J. (2017, July). Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

- ERRant ANnotaion Toolkit

- A toolkit that aligns parallel sentences (erroneous and corrected span) and determine **error types**.

  https://github.com/chrisjbryant/errant

Edit Extraction

Orig	He	only	can	look	at	the	TV	in	the	night	.
Corr	He	can	only	watch	TV	at	night	.			

- Levenshtein

Orig	He	only	can	look	at	the	TV	in	the	night	.
Corr	He	Ø	can	Ø	only	watch	TV	Ø	at	night	.

- Damerau-Levenshtein, linguistic features, merging rules

Orig	He	only	can	look	at	the	TV	in	the	night	.
Corr	He	can	only	watch		Ø	TV	at	Ø	night	.

Felice, M., Bryant, C., & Briscoe, T. (2016, December). Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 825-835).

Edit Classification

- ~50 rule based classification

Original	Corrected	Type	Rule Info
the	Ø	U:DET	POS
Ø	in	M:PREP	POS
cat	dog	R:NOUN	POS, Lemma
cat	cats	R:NOUN:NUM	POS, Lemma
eats	has eaten	R:VERB:TENSE	POS, Lemma, Parse
atack	attack	R:SPELL	Wordlist

Edit Classification

			Operation Tier		
		Type	Missing	Unnecessary	Replacement
Token Tier	Part Of Speech	Adjective	M:ADJ	U:ADJ	R:ADJ
		Adverb	M:ADV	U:ADV	R:ADV
		Conjunction	M:CONJ	U:CONJ	R:CONJ
		Determiner	M:DET	U:DET	R:DET
		Noun	M:NOUN	U:NOUN	R:NOUN
		Particle	M:PART	U:PART	R:PART
		Preposition	M:PREP	U:PREP	R:PREP
		Pronoun	M:PRON	U:PRON	R:PRON
		Punctuation	M:PUNCT	U:PUNCT	R:PUNCT
		Verb	M:VERB	U:VERB	R:VERB
	Other	Contraction	M:CONTR	U:CONTR	R:CONTR
		Morphology	—	—	R:MORPH
		Orthography	—	—	R:ORTH
		Other	M:OTHER	U:OTHER	R:OTHER
		Spelling	—	—	R:SPELL
		Word Order	—	—	R:WO
ogy Tier		Adjective Form	—	—	R:ADJ:FORM
		Noun Inflection	—	—	R:NOUN:INFL
		Noun Number	—	—	R:NOUN:NUM
		Noun Possessive	M:NOUN:POSS	U:NOUN:POSS	R:NOUN:POSS

Existing problem of WikEd

- Special html symbol remains escaped (affect tokenization)

  How to be ' Green'

- should be

  How to be ' Green '

Solution

- Unescape the escaped symbols

```
import html
text = html.unescape(text)
```

- Re-tokenize (with nltk)

```
from nltk import word_tokenize
text = ' '.join(word_tokenize(text))
```

Get Hands Dirty

- Run with the test file

  $ python parallel_to_m2.py -orig enwiki.tiny.err.tok -cor enwiki.tiny.cor.tok -out enwiki.tiny.tok.m2

- Run with WikEd 1.0 (https://github.com/snukky/wikiedits)

  $ python parallel_to_m2.py -orig wiked.tok.err -cor wiked.tok.cor -out wiked.tok.m2

Annotation Format

- We propose a new format, Diff+.

- M2
  used in NUCLE, ERRANT

- GNU wdiff[1]
  used in the WikEd corpus and "git diff --word-diff"

- Diff+
  Extend wdiff to include error type annotation

[1] https://www.gnu.org/software/wdiff/

Diff+

- All spaces in edits are substituted by full-width spaces (\u3000)
- Compactly connect *delete* and *insert* elements of replace errors
- Attach (*error_type*) at the end of the edit token

https://github.com/NTHU-NLPLAB/gec-preprocess

The benefits using Diff+

- Easy to iterate over tokens (simply **text.split(' ')**)

- Error types are explicitly indicated

  We can [-discuss-]{+talk+}(R:VERB) about it .

- Replace edits is implicitly indicated

  We can [-discuss-] {+talk+} about it . (wdiff)

- Still easy to read by human

Summary

- Introduce the WikEd error corpus and a ERRor ANnotation Toolkit
- Propose a new error annotation format
- Further clean existing learner corpora