Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Language Model Based Grammatical Error Correction Without Annotated Training Data

UNIVERSITY OF CAMBRIDGE

Christopher Bryant and Ted Briscoe

Motivation

- State of the art Grammatical Error Correction (GEC) systems rely on as much annotated training data as possible.
- Language model (LM) based approaches do not require annotated training data but still performed well in the CoNLL-2014 shared task on GEC.
- ***Question**: To what extent can a simple LM system compete with a state of the art system trained on millions of words of annotated data?*

Methodology

1. Calculate the normalised log probability of the input sentence.

Input Sentence									Prob
I	am	looking	forway	to	see	you	soon	.	**-2.71**

2. Build a confusion set for each token in that sentence.

I	am	looking	forway	to	see	you	soon	.	**-2.71**
	was	look	forward	of	seeing		sooner		
	be	looks	Norway	in	saw		soonest		-
	are	looked	foray	Ø	sees		…		

3. Rescore the sentence for each candidate correction in each confusion set.

I	am	looking	forway	to	see	you	soon	.	**-2.71**
	was -2.67	look -2.91	forward -1.80	of -2.98	seeing -3.09		sooner -3.05		
	be -3.09	looks -2.93	Norway -2.36	in -2.99	saw -3.25		soonest -3.20		-
	are -3.10	looked -2.95	foray -2.70	Ø -3.00	sees -3.39		… …		

4. Apply the single global best correction that improves the sentence probability above a threshold.

I	am	looking	forway	to	see	you	soon	.	**-2.71**
I	am	looking	**forward**	to	see	you	soon	.	**-1.80**

5. Iterate steps 1 – 4.

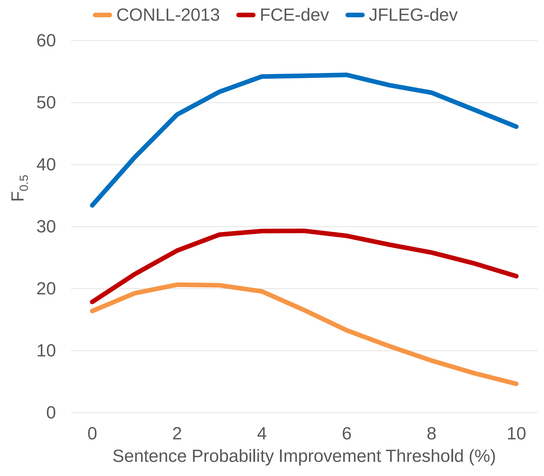
I	am	looking	forway	to	see	you	soon	.	**-2.71**
I	am	looking	**forward**	to	see	you	soon	.	**-1.80**
I	am	looking	forward	to	**seeing**	you	soon	.	**-1.65**

Confusion Set Generators

- CyHunspell
  - Spelling errors — e.g. freind → friend
  - Inflectional errors — e.g. advices → advice
- Automatically Generated Inflection Database
  - Noun number errors — e.g. cat → cats
  - Verb tense/form/agreement — e.g. eat → ate, eat → eating
  - Adjective form — e.g. bigger → biggest
- Manually defined confusion sets
  - Determiners: {Ø, the, a, an}
  - Prepositions: {Ø, about, at, by, for, from, in, of, on, to, with}

Thresholding

- Some corrections improve sentence probability more than others.
  - forway → forward — -2.71 → -1.80
  - am → was — -2.71 → -2.67
- However, smaller improvements are likely to be false positives.
  - forway → forward — -2.71 → -1.80
  - am → was — -2.71 → -2.67
- Solution: Set improvement thresholds based on a development set.



- Observation: Different datasets have different optimum thresholds even with a single tuning parameter.

Results

- We train a 5-gram LM on the 1 Billion Word Benchmark corpus with KenLM.
- We compare performance with several state of the art systems.
  - POST (2014): A LM approach that came 4th in CoNLL-2014.
  - $AMU16_{SMT}$+LSTM and $CAMB16_{SMT}$+LSTM: A hybrid combination of Statistical Machine Translation (SMT) and neural sequence labelling approaches reported in Yannakoudakis et al. (2017).
  - Sakaguchi et al. (2017): A neural reinforcement learning approach.

Test Set	System	P	R	$F_{0.5}$	GLEU
CoNLL-2014	POST 2014	34.51	21.73	30.88	59.50
	$AMU16_{SMT}$+LSTM	**58.79**	**30.63**	**49.66**	**68.26**
	$CAMB16_{SMT}$+LSTM	49.58	21.84	39.53	65.68
	Our work	40.56	20.81	34.09	59.35
FCE-test	$AMU16_{SMT}$+LSTM	40.67	17.36	32.06	63.57
	$CAMB16_{SMT}$+LSTM	**65.03**	**32.45**	**54.15**	**70.72**
	Our work	44.78	14.12	31.22	60.04
JFLEG-test	$AMU16_{SMT}$+LSTM	60.68	22.65	45.43	42.65
	$CAMB16_{SMT}$+LSTM	65.86	30.56	53.50	46.74
	Sakaguchi et al. (2017)	65.80	**40.96**	**58.68**	**53.98**
	Our work	**76.23**	28.48	57.08	48.75

Conclusions

- We improved upon the previous best LM approach by > 3 $F_{0.5}$.
- We outperformed 2 state of the art systems on JFLEG and came surprisingly close to the top system.
- State of the art systems do not seem to generalise well and probably overfit to different datasets.
- Our results are fairly competitive with data hungry systems despite
  a) requiring minimal annotated data (for tuning purposes only).
  b) only targeting ~50% of all error types.
- Our approach suggests it is possible to build a decent GEC system for any language where annotated training data may not be available.