# Robsut Wrod Reocginiton via semi-Character Recurrent Neural Network

Keisuke Sakaguchi, Kevin Duh, Matt Post, Benjamin Van Durme
Johns Hopkins University

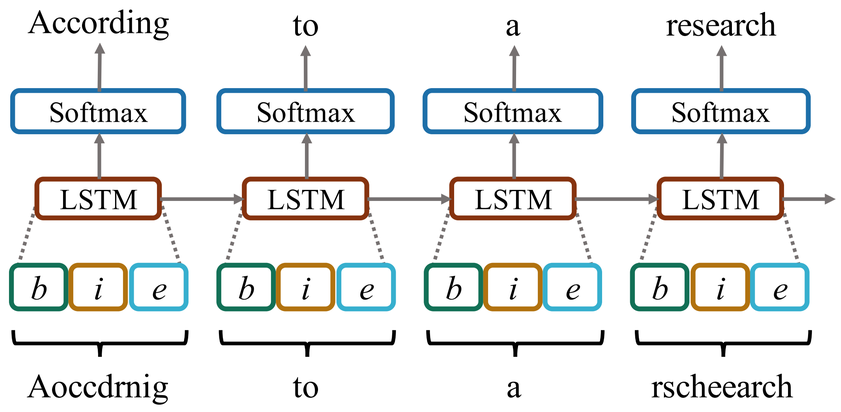
## Hey, you made spelling errors in the title!

**Good catch, but you can still read it smoothly!**
Language processing mechanism by humans is generally more robust than computers. (e.g., *Cmabrigde Uinervtisy* [*Cambridge University*] effect)

> *Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mtter in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.*

Can we build a computational model which replicates this mechanism?

## Model Overview (scRNN)




The input layer of our model consists of three sub-vectors: beginning ($b$), internal ($i$), and ending ($e$) character(s) of the input word.

## Results (Spelling Correction Experiments)

| | Jumble | Delete | Insert |
|---|---|---|---|
| scRNN (proposed) | **99.44** | **85.56** | **97.04** |
| CharCNN (Kim et al. 2016) | 16.18 | 19.76 | 35.53 |
| Enchant | 57.59 | 35.37 | 89.63 |
| Commercial A | 54.81 | 60.19 | 93.52 |
| Commercial B | 54.26 | 71.67 | 73.52 |

**Table 1:** Spelling correction accuracy (%) with different error types. (e.g., Jumble: Cmbarigde, Delete: Camridge, Insert: Cambpridge)

| Units | Acc (%) | SD | Size (KB) |
|---|---|---|---|
| 5 | 24.65 | 2.59 | 236 |
| 10 | 48.43 | 3.26 | 435 |
| 15 | 73.32 | 3.65 | 632 |
| 20 | 84.82 | 2.39 | 830 |
| 30 | 94.15 | 1.54 | 1,255 |
| 40 | 96.90 | 1.26 | 1,670 |
| 50 | 98.48 | 0.94 | 2,092 |
| 60 | 98.39 | 0.81 | 2,514 |

**Table 2:** scRNN accuracy (%), and the size of model file (KB) on jumbled word recognition with respect to the number of units of LSTM.

## Examples

| | |
|---|---|
| scRNN | According to a research at Cambridge University , it does n't matter in what order the letters in a word are , the only important thing is that the first and last letter be at the right place . The rest can be a total mess and you can still read it without problem . This is because the human mind does not read every letter by itself , but the word as a whole . |
| CharCNN | According to a research at Cambridge Minority , it deck n't mother in wait or the letters in a wood are , the tony Vermont timing is taxi the tourist and sat letter be at the fruit pile . The reset can be a total uses and you can vital rake it worthy parallel . Mips is abuse the human trim deck not rake survey latter by leftist , but the wood as a whole . |
| Enchant | Ecuadoran to a searcher at Brigade Nerviness , it does n't matter in what order the letters in a word are , the only omnipresent thing is that the freest and slat letter be at the right place . The rest can be a total mess and you can still read it outhit corbel . Tish is Ceausescu , the human mind does not read Hervey letter by leftist , but the word as a whole . |
| Comm.A | Occurring to a scholarch at Cambridge Inertias , it does n't matter in what order the letters in a word are , the only impotent thing is that the first and last letter be at the right place . The rest can be a total mess and you can still read it outhit problem . This is bcuseae the human mind does not read every letter by istle , but the word as a whole . |
| Comm.B | Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy , it does n't matter in what order the letters in a word are , the only iprmoetnt thing is that the first and last letter be at the right place . The rest can be a total mess and you can still read it wouthit problem . Tihs is bcuseae the human mind does not read every letter by itself , but the word as a whole . |

## Corroboration with psycholinguistic experiments

| Cond. | Example | # of fixations | Regression(%) | Avg. Fixation (ms) |
|---|---|---|---|---|
| Normal | The boy could not solve the problem so he asked for help. | 10.4 | 15.0 | 236 |
| Internal | The boy cuold not slove the probelm so he aksed for help. | 11.4$^*$ | 17.6$^*$ | 244$^*$ |
| Ending | The boy coudl not solev the problme so he askde for help. | 12.6$^\dagger$ | 17.5$^*$ | 246$^*$ |
| Beginning | The boy oculd not oslve the rpoblem so he saked for help. | 13.0$^\ddagger$ | 21.5$^\dagger$ | 259$^\dagger$ |

| Cond. | Example | Accuracy |
|---|---|---|
| INT | As a relust , the lnik beewetn the fureuts and sctok mretkas rpiped arapt . | 98.96 |
| END | As a rtelus , the lkni betwene the feturus and soctk msatrek rpepid atarp . | 98.68$^*$ |
| BEG | As a lesurt , the lnik bweteen the utufers and tocsk makrtes pipred arpat . | 98.12$^\dagger$ |
| ALL | As a strule , the lnik eewtneb the eftusur and okcst msretak ipdepr prtaa . | 96.79$^\ddagger$ |

**Table 3:** Example sentences and results for measures of fixation excerpt from Rayner et al., (2006) (Top) and results for spelling correction accuracy by scRNN variants depending on different jumble conditions (Bottom). Entries with $^*$ have statistically significant difference from the condition N ($p < 0.01$) and those with $^\dagger$ and $^\ddagger$ differ from $^*$ and $^\dagger$ with $p < 0.01$ respectively.

## Error Analysis

| Cond. | Examples of errors (correct/wrong) |
|---|---|
| INT | Under/under, there/three, form/from, fares/fears, trail/trial, Broad/Board |
| END | being/begin, quiet/quite, bets/best, stayed/steady, heat/hate, lost/lots +INT |
| BEG | Several/reveal, Growth/worth, host/shot, creditors/directors, views/wives + same errors in INT |
| ALL | Under/trend, center/recent, licensed/declines, stop/tops + same errors in INT, END, & BEG |

## Conclusion

- We have presented a semi-character recurrent neural network model, scRNN, which is inspired by the robust word recognition mechanism known in psycholinguistics literature as the *Cmabrigde Uinervtisy* effect.
- We also have demonstrated a similarity between scRNN and human word recognition mechanisms, by showing that scRNN replicates a psycholinguistics experiment about word recognition difficulty in terms of the position of jumbled characters.