Comparison of Interactive Knowledge Base Spelling Correction Models for Low-Resource Languages

Yiyuan Li, Antonios Anastasopoulos and Alan W Black

Carnegie Mellon University, Pittsburgh, PA

yiyuanli@andrew.cmu.edu {aanastas, awb}@cs.cmu.edu

Task and Challenges

Spelling Correction in low-resource languages.

Challenges:

- Little literacy and standardization.
- Global misspelling patterns are hard to identify.
- Gold dictionaries only preserve correct words.
- Limited data and lacking misspelled corpora. [1]

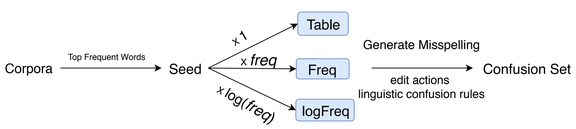
Contribution:

- Mitigating rareness of misspelling.
- A system to interact with human users.

Dataset

- **Realistic resource:** TOEFL11 (En) and Spellrueval (Ru)
- **Synthetic:** Wikipedia (En, Es, Ru, Fi, It, Tr)
- **OCR in low-resource languages:** Griko and Ainu [2]

Data Augmentation



Model

- Multi-layer LSTM (**LSTM**)
- Character-level Trigram Language Model with threshold (**CharTriLM**)

[1]Keisuke Sakaguchi, Kevin Duh et al. Robsut wrod reocginiton via semi-character recurrent neural network. In Proc. AAAI, 2016.

[2] Antonios Anastasopoulos, Marika Lekakou, et al. Part-of-speech tagging on an endangered language: a parallel griko-italian re- source. In Proc. COLING, 2018.

System Design

We build a typical system (Figure 1) and develop an user interface (Figure 2) to benefit the end users for customized languages. The decision of the words are displayed in real time and can be manually corrected by the users. Updates of the knowledge base are triggered by correction or recommendation. File upload and customized lexicon training are supported.

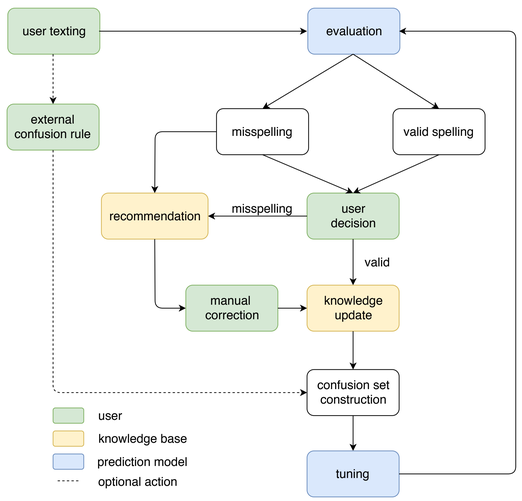


Figure 1:System Architecture

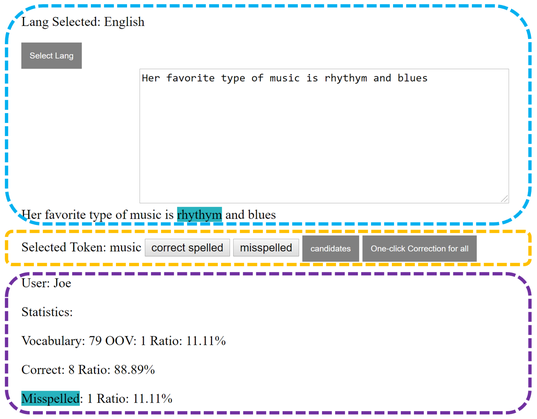
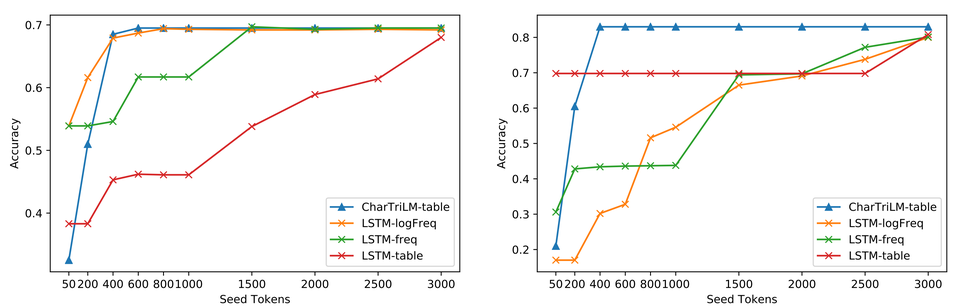


Figure 2:User Interface

Result - Realistic Resource

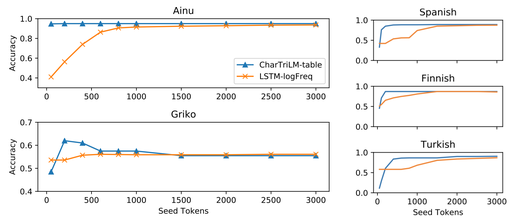


(a) Accuracy on English (b) Accuracy on Russian

- Incremental training.
- `CharTriLM` provides better convergence and LSTM models catch up when more data comes.

Demo portal: https://nativeatom.github.io/OSC/

Result - Low Resource and Synthetic Data



- Language model and LSTM models perform consistently among languages with different morphological complexities, and similar to realistic resource.
- The neural model catches up quickly, but the meeting points vary among languages.

Results - Limited Training Set

Corpora	`LSTM-logFreq`		`CharTriLM`	
	Accuracy	F1	Accuracy	F1
Real data				
English	**0.696**	0.452	0.635	**0.455**
Russian	0.301	0.217	**0.830**	**0.454**
Synthetic data				
English	0.925	**0.488**	**0.938**	0.484
Russian	0.254	0.161	**0.910**	**0.476**
Italian	**0.920**	**0.479**	0.910	0.476
Spanish	0.561	0.330	**0.880**	**0.468**
Finnish	0.729	0.435	**0.870**	**0.465**
Turkish	0.581	0.320	**0.845**	**0.458**
OCR outputs				
Griko	0.562	0.369	**0.590**	**0.467**
Ainu	0.746	0.420	**0.950**	**0.487**

- Seed of 500 most frequent words.
- `CharTriLM` performances better in most corpora than `LSTM-logFreq`.

Conclusion

- Misspelling can be identified in few training examples by character level language model.
- The difference between neural model and language model becomes minimal as the amount of data increases.