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

Data Augmentation

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

Model

- 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 - Low Resource and Synthetic Data

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

Conclusion

- Spelling correction in low-resource languages.
- Incremental training.
- CharTriLM provides better convergence and LSTM models catch up when more data comes.

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