Automatic Syllabification of English

Annika Shankwitz

The Problem

- Automatic Syllabification of English
- That is, add syllable boundaries to English grapheme forms
- E.g., …

Input: about

Output: a|bout

Why it's Interesting – Wider Applications

- Syllable boundaries can be helpful for …
  - Low-Resource Language Modeling & Translation (Oncevay et al., 2022)
    - syllable-based tokenization over sub-word or character tokenization
  - End-to-end speech recognition (Anoop & Ramakrishnan, 2023; Zhou, 2018)
    - modeling syllables over context-independent phonemes

Why it's Interesting – Previous Work

- **Dinu et al (2024)**
  - Automatic syllabification of Italian
  - Treat as a sequence labeling task
    - 0: grapheme doesn't begin a syllable
    - 1: grapheme begins a syllable

  me-da-gliò-ne → 1010100010

    - Use a GRU RNN
    - Overall Accuracy: 99.74%

Why it's Interesting – Previous Work

- Italian vs English (Seymour et al. (2003))
  - Italian:
    - shallow orthography (many 1-to-1 mappings)
    - simple syllable structure (CV structure dominant)
  - English:
    - deep orthography (few 1-to-1 mappings)
    - complex syllable structure (closed syllables, complex onsets and codas)

Why it’s Interesting – Previous Work

- Italian vs English (Seymour et al. (2003))
  - Italian:
    - shallow orthography (many 1-to-1 mappings)
    - simple syllable structure (CV structure dominant)
  - English:
    - deep orthography (few 1-to-1 mappings)
    - complex syllable structure (closed syllables, complex onsets and codas)

So, how does Dinu et al (2024)’s model perform on English?

My Approach

- Data: Corpus of 25,000 syllabified English words (e.g., a;ban;don)

- Data Preprocessing:
  - Converted data to its numeric equivalent (0100100)
    - Didn't mark with beginning of words with 1
  - Divided data randomly into train (80%), dev (10%), and test (10%)

Elman Model

- a character embedding layer, producing 96-dimensional vectors for each inputted character,
- an Elman RNN cell with an 192-dimension hidden state,
- a final linear layer which produces tag scores for each character,
- trained using Cross Entropy Loss with the Adam optimizer, a learning rate of 0.001, and 15 training epochs.

GRU Model

- a character embedding layer, producing 96-dimensional vectors for each inputted character,
- a stacked bidirectional GRU with 3 layers, a 96-dimension hidden state (96 forward & 96 backwards = total 192), and 0.2-rate dropout between GRU layers,
- 0.5-rate dropout applied to the GRU output,
- layer normalization applied to the GRU output,
- a time-distributed, fully-connected linear layer with ReLU activation, which projects each time step/inputted character onto the tag set,
- a final linear layer which produces tag scores for each character,
- trained using Cross Entropy Loss with the Adam optimizer, a learning rate of 0.001, and 15 training epochs

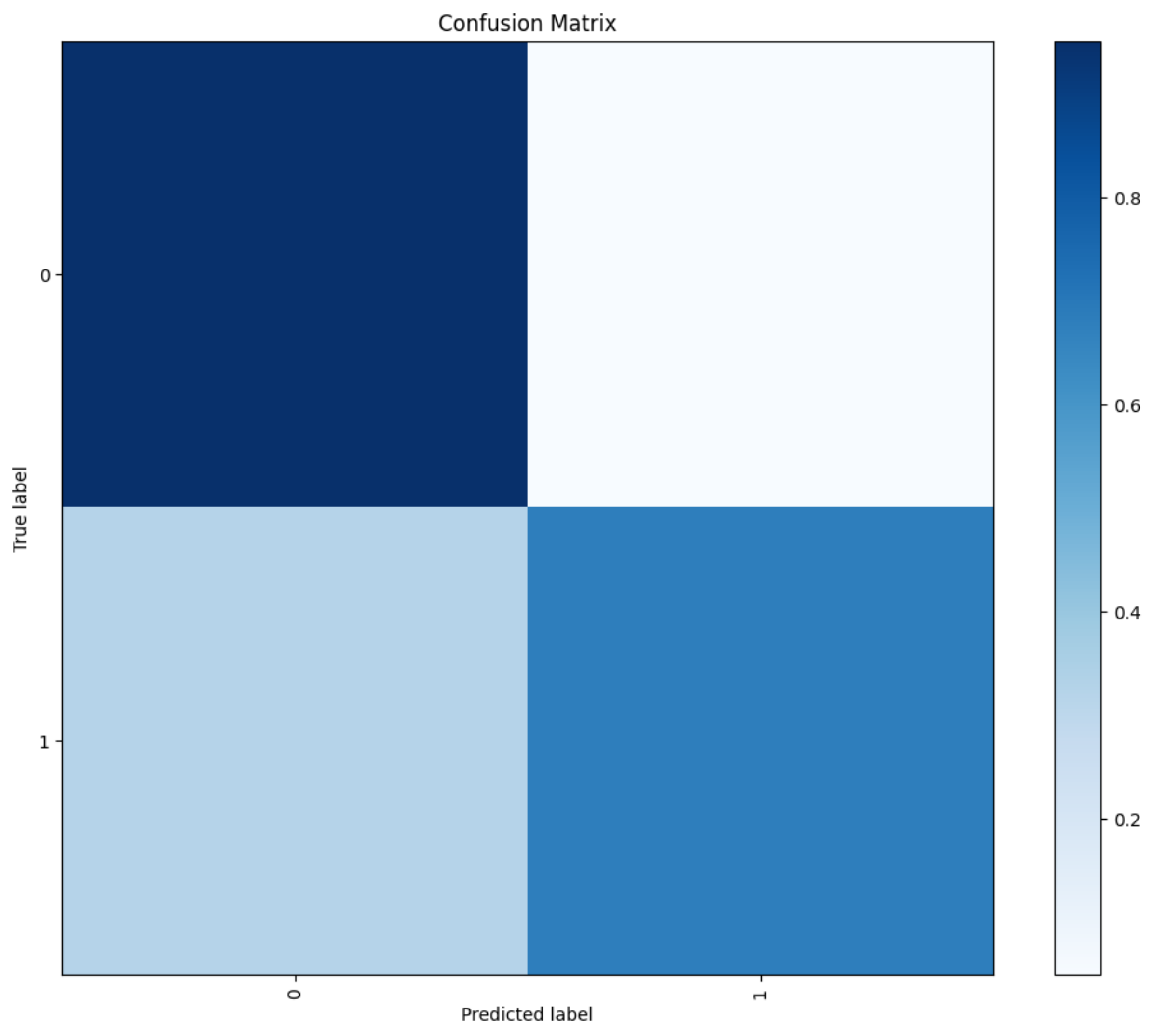
↑ Replication of Dinu et at (2024)'s model

Model Training

- Manual Grid Search Loop targeting:

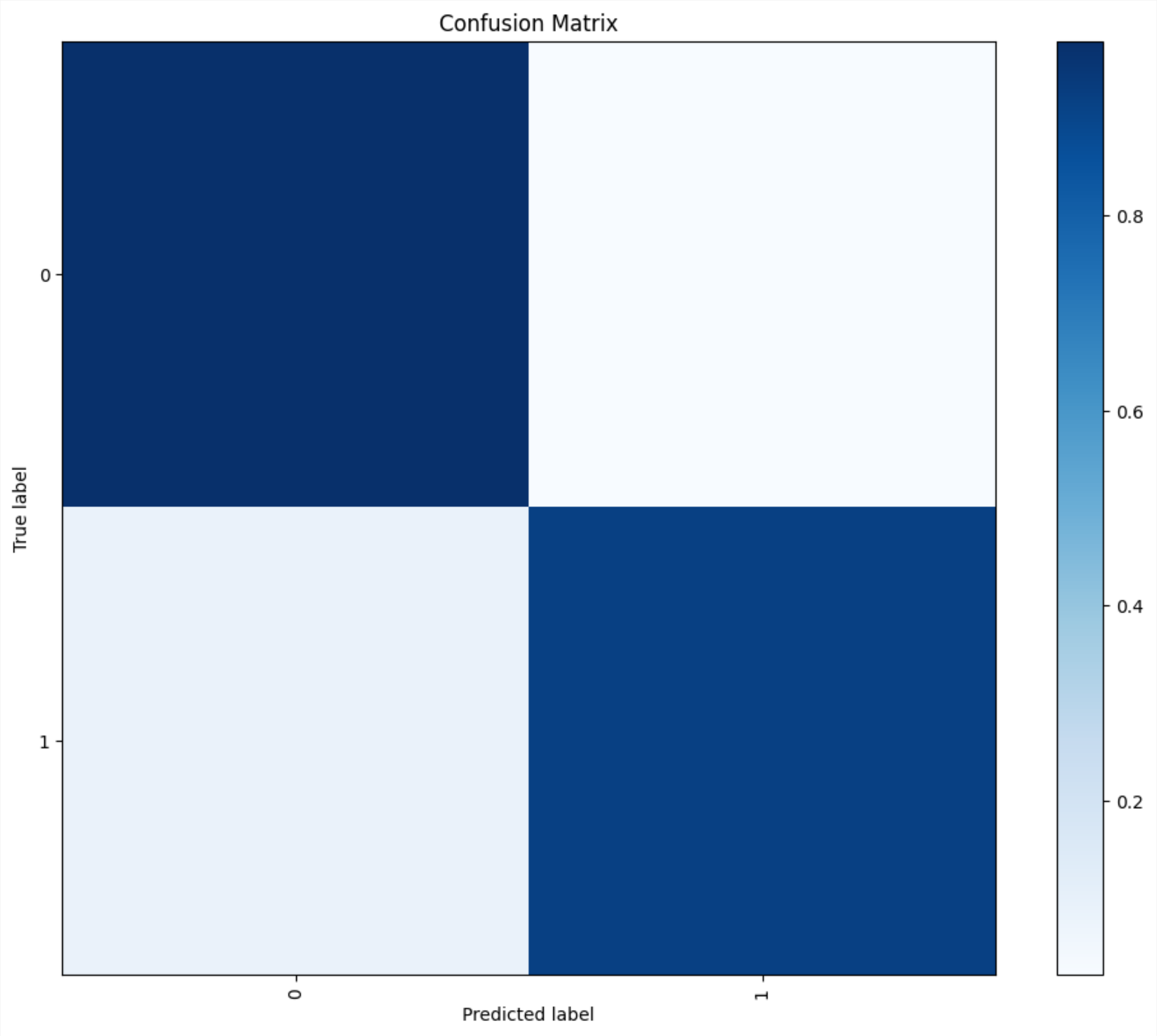
Embedding dimension	= [32, 64, 96]
Hidden dimension	= [96, 128, 192]
Learning rate	= [0.01, 0.001, 0.0001]
Number of training epochs	= [5,10,15]

Elman Model – Preliminary Results



Accuracy = 89.47%

GRU Model – Preliminary Results



Accuracy = 96.60%

TODO

- Look further into additional metrics
  - Imbalanced class problem
  - Metrics by syllable number

- Look into what type of errors the models make
  - Are the errors similar across models?
  - Are the errors different across models?

References

Anoop, Chandran Savithri, and Ramakrishnan, Angarai Ganesan. "Suitability of Syllable-Based Modeling Units for End-To-End Speech Recognition in Sanskrit and Other Indian Languages." *Expert Systems with Applications*, vol. 220, June 2023, p. 119722, https://doi.org/10.1016/j.eswa.2023.119722.

Liviu Dinu, Ioan-Bogdan Iordache, Simona Georgescu, Alina Maria Cristea, and Bianca Guita. 2024. ItGraSyll: A Computational Analysis of Graphical Syllabification and Stress Assignment in Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 316–324, Pisa, Italy. CEUR Workshop Proceedings.

Oncevay, A., Rojas, K. D. R., Sanchez, L. K. C., & Zariquiey, R. (2022). Revisiting syllables in language modelling and their application on low-resource machine translation. *arXiv preprint arXiv:2210.02509*.

Seymour, P.H.K., Aro, M., Erskine, J.M. and (2003), Foundation literacy acquisition in European orthographies. British Journal of Psychology, 94: 143-174. https://doi.org/10.1348/000712603321661859

Zhou, S., Dong, L., Xu, S., & Xu, B. (2018). Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1804.10752