Supplementary Materials: Learning to Capitalize with Character-Level Recurrent Neural Networks: An Empirical Study

Raymond Hendy Susanto[†] and **Hai Leong Chieu**[‡] and **Wei Lu**[†]
[†]Singapore University of Technology and Design
[‡]DSO National Laboratories
{raymond_susanto,luwei}@sutd.edu.sg
chaileon@dso.org.sg

Abstract

This supplementary material gives additional details to the experimental settings for the paper "Learning to Capitalize with Character-Level Recurrent Neural Networks: An Empirical Study" (Susanto et al., 2016).

Experimental Settings

The feature set in CRF-WORD follows a default property file provided in Stanford CoreNLP's code repository.[1] Table 1 shows our feature configuration. The description of each parameter in the property file is also documented online.[2] We did not include ℓ_1 regularization and only ℓ_2, since the ℓ_1 optimizer was not publicly released.

Property name	Value	Property name	Value
useClassFeature	true	useTypeySequences	true
useWord	true	useOccurrencePatterns	true
useNGrams	true	useLastRealWord	true
noMidNGrams	true	useNextRealWord	true
maxNGramLeng	6	useDisjunctive	true
usePrev	true	disjunctionWidth	true
useNext	true	wordShape	true
useLongSequences	true	usePosition	true
useSequences	true	useBeginSent	true
usePrevSequences	true	useTitle	true
useTypeSeqs	true	useObservedSequencesOnly	true
useTypeSeqs2	true		

Table 1: CRF-WORD feature configuration

[1] https://github.com/stanfordnlp/CoreNLP/blob/master/scripts/truecase/truecasing.fast.caseless.prop
[2] http://nlp.stanford.edu/nlp/javadoc/javanlp-3.6.0/edu/stanford/nlp/ie/NERFeatureFactory.html

Feature	Examples
character unigram	c[0]=j, c[1]=o, c[2]=h
character bigram	c[0]|c[1]=j|o
character trigram	c[0]|c[1]|c[2]=j|o|h
previous word	w[-1]=<s>
current word	w[0]=john
next word	w[1]=blair
word bigram	w[-1]|w[0]=<s>|john, w[0]|w[1]=john|blair
begin-of-sentence	__BOS__

Table 2: CRF-CHAR features for the character *j* in *john blair*, assuming they are the first two words in the input sentence. Similarly, we define an end-of-sentence feature for the last word in a sentence.

The feature set in CRF-CHAR is given in Table 2. We use a simpler feature set compared to CRF-WORD in order to speed up training. By working at the character level, CRF-CHAR needs to make more predictions than CRF-WORD, i.e., predictions are made for each character instead of each word. Despite the smaller feature set, CRF-CHAR outperforms CRF-WORD in 3 out of 4 data sets. We also cross-validate the ℓ_2 regularization parameter. The chosen regularization parameters are 0.1 for EN-Wiki, EN-WSJ, EN-Reuters and 0.01 for DE-ECI.

For RNN, the initial learning rate is set according to the model and data set to ensure good convergence rate. The initial learning rate is set to 0.002 for all LSTM models and GRU-SMALL and 0.001 for GRU-LARGE. We apply an exponential decay to the learning rate starting from the 10^{th} epoch, where we multiply the learning rate by a factor of 0.97 after every epoch.

References

Raymond Hendy Susanto, Hai Leong Chieu, and Wei Lu. 2016. Learning to Capitalize with Character-Level Recurrent Neural Networks: An Empirical Study. In *Proceedings of EMNLP*.