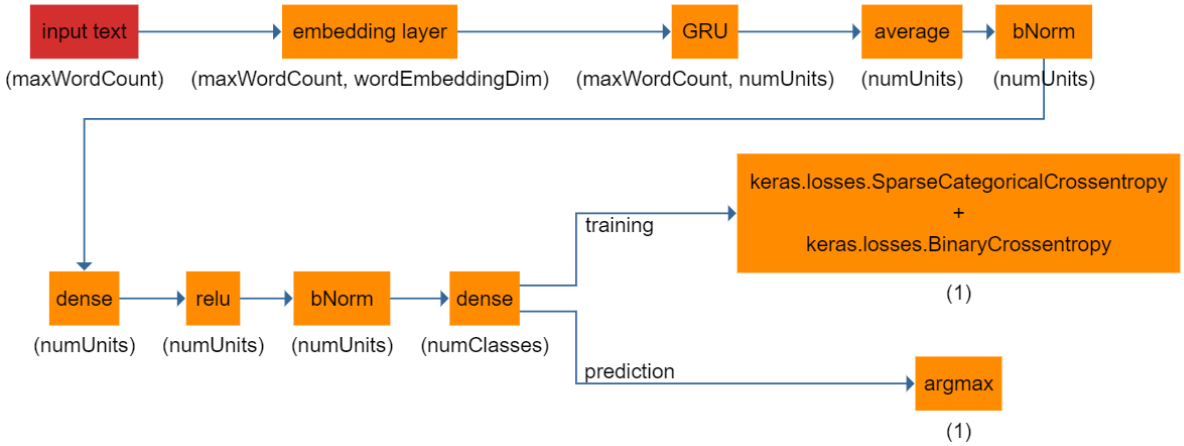
Brief summary of topic classification for MaML news corpus

To classify the topic of an article, and articles that are exclusively about foreign news topics (domestic vs non-domestic), we use the deep learning framework Keras (Srinivasa-Desikan 2018). The procedure follows the data flow illustrated in the figure below.



The input text vocabulary is limited to the 10.000 words most commonly found in the train data, and the input text is limited to the first 800 words (maxWordCount). From this input, a word embedding model is generated, using 512 dimensions (tuned). The word embedded representation of the articles is then used as input for a recurrent neural network as described in Cho et al. (2014). After passing through some intermediate neural network layers, the dense layer produces a vector of predicted probabilities for each document, for both the topic and foreign news variables. During training, these are parsed through a sparse categorical crossentropy loss function for the topic labeling, and through a binary crossentropy loss function for the foreign news labeling. Both are added together, and the foreign news loss is multiplied by a factor of 7 (tuned). For labeling, the label with the highest probability is selected.

Optimization of this model has been conducted using 90% of the manually labeled data as training data, and 10% as testing/validation data. Once the optimum model parameters have been determined, these are used in a 10-fold cross-validation procedure to determine the final performance of the model. While it is true that this 10-fold CV includes the 10% test data used to determine the optimal parameters, the stability in performance between the individual folds leads us to the conclusion that the performance estimates provided by the 10-fold CV procedure are reliable.

References:

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.