	Bengali	Hindi	Latin	Spanish
BLSTM-BLSTM	90.84/91.14	94.89/**94.90**	89.35/89.52	97.85/97.91
BGRU-BGRU	90.63/90.84	94.44/94.50	89.40/**89.59**	98.07/**98.11**
Lemming	**91.69**	91.64	88.50	93.12
Morfette	90.69	90.57	87.10	92.90

Table 2: Lemmatization accuracy (in %) without/with restricting output classes.

	Bengali	Hindi	Latin	Spanish
BLSTM-BLSTM	86.46/89.52	94.34/94.52	85.76/87.36	97.39/97.62
BGRU-BGRU	86.39/88.90	93.84/94.04	85.49/86.87	97.51/97.73

Table 3: Lemmatization accuracy (in %) without using applicable edit trees in training.

are two successive bidirectional networks - the first one for building the syntactic embedding and the next one for the edit tree classification, so basically we deal with two different models BLSTM-BLSTM and BGRU-BGRU. Table 2 shows the comparison results of these models with Lemming and Morfette. In all cases, the average accuracy over 4 fold cross validation on the datasets is reported. For an entry 'x/y' in Table 2, x denotes the accuracy without output classes restriction, i.e. taking the maximum over all edit tree classes present in the training set, whereas y refers to the accuracy when output is restricted in only the applicable edit tree classes of the input word. Except for Bengali, the proposed models outperform the baselines for the other three languages. In Hindi, BLSTM-BLSTM gives the best result (94.90%). For Latin and Spanish, the highest accuracy is achieved by BGRU-BGRU (89.59% and 98.11% respectively). In the Bengali dataset, Lemming produces the optimum result (91.69%) beating its closest performer BLSTM-BLSTM by 0.55%. It is to note that the training set size in Bengali is smallest compared to the other languages (on average, 16, 712 tokens in each of the 4 folds). Overall, BLSTM-BLSTM and BGRU-BGRU perform equally good. For Bengali and Hindi, the former model is better and for Latin and Spanish, the later yields more accuracy. Throughout the experiments, restricting the output over applicable classes improves the performance significantly. The maximum improvements we get are: 0.30% in Bengali using BLSTM-BLSTM (from 90.84% to 91.14%), 0.06% in Hindi using BGRU-BGRU (from 94.44% to 94.50%), 0.19% in Latin using BGRU-BGRU (from 89.40% to 89.59%) and 0.06% in Spanish using BLSTM-BLSTM (from 97.85% to 97.91%). To compare between the two baselines, Lemming consistently performs better than Morfette (the maximum difference between their accuracies is 1.40% in Latin).

Bengali	Hindi	Latin	Spanish
27.17	5.25	15.74	7.54

Table 4: Proportion of unknown word forms (in %) present in the test sets.

Effect of Training without Applicable Edit Trees: We also explore the impact of applicable edit trees in training. To see the effect, we train our model without giving the applicable edit trees information as input. In the model design, the equation for the final classification task is changed as follows,

$$\mathbf{l}_i = softplus(\mathbf{L}^f \mathbf{h}_i^f + \mathbf{L}^b \mathbf{h}_i^b + \mathbf{b}_l),$$

The results are presented in Table 3. Except for Spanish, BLSTM-BLSTM outperforms BGRU-BGRU in all the other languages. As compared with the results in Table 2, for every model, training without applicable edit trees degrades the lemmatization performance. In all cases, BGRU-BGRU model gets more affected than BLSTM-BLSTM. Language-wise, the drops in its accuracy are: 1.94% in Bengali (from 90.84% to 88.90%), 0.46% in Hindi (from 94.50% to 94.04%), 2.72% in Latin (from 89.59% to 86.87%) and 0.38% in Spanish (from 98.11% to 97.73%).

One important finding to note in Table 3 is that irrespective of any particular language and model used, the amount of increase in accuracy due to the output restriction on the applicable classes is much more than that observed in Table 2. For instance, in Table 2 the accuracy improvement for Bengali using BLSTM-BLSTM is 0.30% (from 90.84% to 91.14%), whereas in Table 3 the corresponding value is 3.06% (from 86.46% to 89.52%). These outcomes signify the fact that training with the ap-