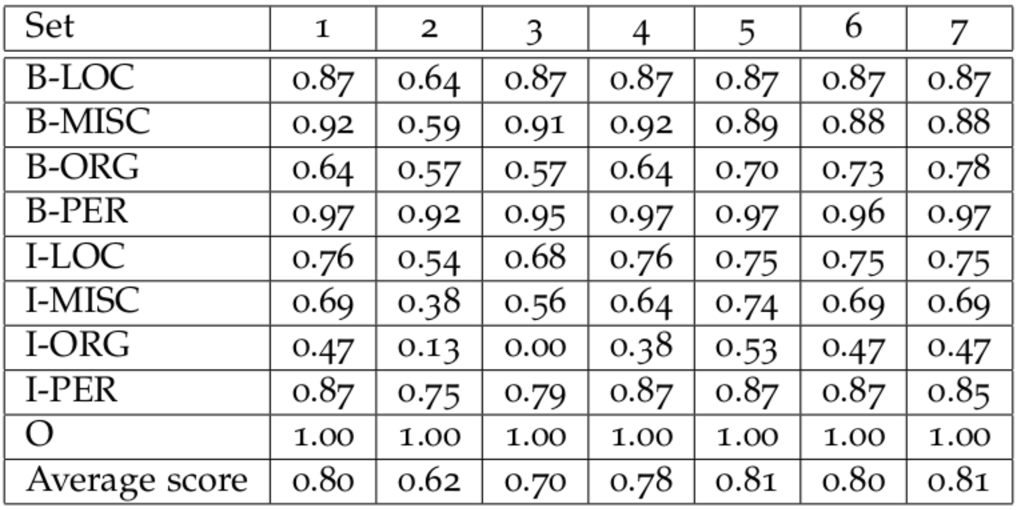
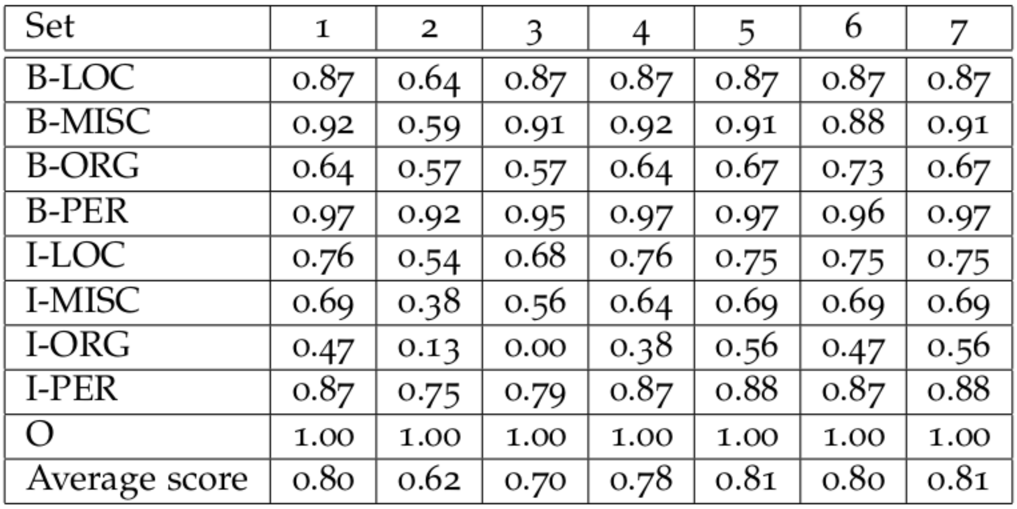


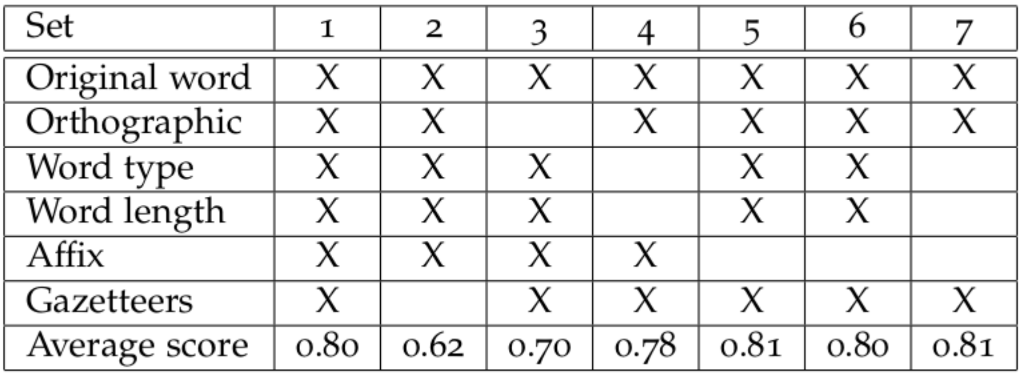
Set	1	2	3	4	5	6	7
B-LOC	0.87	0.64	0.87	0.87	0.87	0.87	0.87
B-MISC	0.92	0.59	0.91	0.92	0.89	0.88	0.88
B-ORG	0.64	0.57	0.57	0.64	0.70	0.73	0.78
B-PER	0.97	0.92	0.95	0.97	0.97	0.96	0.97
I-LOC	0.76	0.54	0.68	0.76	0.75	0.75	0.75
I-MISC	0.69	0.38	0.56	0.64	0.74	0.69	0.69
I-ORG	0.47	0.13	0.00	0.38	0.53	0.47	0.47
I-PER	0.87	0.75	0.79	0.87	0.87	0.87	0.85
O	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average score	0.80	0.62	0.70	0.78	0.81	0.80	0.81

Summary of the average F1-scores of each classifier.



Set	1	2	3	4	5	6	7
B-LOC	0.87	0.64	0.87	0.87	0.87	0.87	0.87
B-MISC	0.92	0.59	0.91	0.92	0.91	0.88	0.91
B-ORG	0.64	0.57	0.57	0.64	0.67	0.73	0.67
B-PER	0.97	0.92	0.95	0.97	0.97	0.96	0.97
I-LOC	0.76	0.54	0.68	0.76	0.75	0.75	0.75
I-MISC	0.69	0.38	0.56	0.64	0.69	0.69	0.69
I-ORG	0.47	0.13	0.00	0.38	0.56	0.47	0.56
I-PER	0.87	0.75	0.79	0.87	0.88	0.87	0.88
O	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average score	0.80	0.62	0.70	0.78	0.81	0.80	0.81

Summary of the average F1-scores of each classifier after tuning.



Set	1	2	3	4	5	6	7
Original word	X	X	X	X	X	X	X
Orthographic	X	X		X	X	X	X
Word type	X	X	X		X	X	
Word length	X	X	X		X	X	
Affix	X	X	X	X			
Gazetteers	X		X	X	X	X	X
Average score	0.80	0.62	0.70	0.78	0.81	0.80	0.81

Summary of the average F1-scores of each classifier corresponding to its feature subset.

According to the results of the experiment, the SVM model performs the best with feature set 5 which excludes affix feature and set 7 which excludes word type, word length and affix features (both F1-score = 0.81). The lowest score is from set 2 which excludes gazetteers (F1-score = 0.62) while the second lowest score is from set 3 which excludes orthographic features (F1-score = 0.70). Set 1 which includes all the features hold the second highest score as same as set 6 which excludes word type and word length features (F1-score = 0.80). Set 4 which excludes word type and word length features obtaines a score of 0.78 that is just lightly lower than set 1, 5, 6, 7. Comparing set 1, 5, 6, and 7, it is obvious that set 7 with the least features (original word, orthographic features, and gazetteers) still obtaines the highest score. This indicates that word type, word length and affix features are not meaningful for the model. Set 2 without gazetteers performes the worst, which indicates that gazetteers are the most important features for the model. Set 3 without orthographic gave the second lowest score. It means that orthographic features are also an indispensable feature for the model.

Based on the analyses above, we infer that all the features (original word, orthographic information, word type, word length, affix, and gazetteers) are potentially useful for the NER SVM model for Gronings. The second highest score of set 1 supports this point. However, using all the features is not necessary to lead to a higher performance, for example, set 1 with all the features gives the same score as set 6, and a lower score than set 7 which has the least features. Therefore, several features should be disable. Observing feature set 7, it contains the least features but obtains the highest score. This indicates that all features in set 7 are indispensable. They are original word, orthographic feature, and gazetteers. Without gazetteers, set 2 has the worst score while with gazetteer, set 4, 5, 6, and 7 has significantly higher scores. This proves that the gazetteer definitely improves the outcome of the model.

Parameters tuning do not help to improve the performance of the model because it is obvious that the average F1-score of each set is not changed. However, there are changes on the average scores of the labels. F1-score of the label I-ORG is the lowest among the other labels. The highest score for label I-ORG is from set 5 (F1-score = 0.53). Although the overall scores of each set after parameters tuning are not changed, set 5 and 7 have a better score for label I-ORG after tuning (respectively from 0.53 to 0.56 and from 0.47 to 0.56).

One of the possible reasons leading to a low generalization performance of the model on I-ORG entities is that the size the list of the non-initial words of organizations' names is not large. Moreover, there is ambiguity existing in the names, for example, “van” could be an word before a surname, or a preposition. Also, there are many names which could be both a name of a person and a name of an organization. Despite of these drawbacks of the gazetteer, its contribution to the performance of the model could not be neglected, especially under the circumstance that POS features were not available in the dataset. Comparing with the best F1-score (0.83) of the SVM NER model which was trained by POS tag in the experiment of Desmet and Hoste (2010), the score in this experiment (0.81) was not significantly lower. Therefore, the gazetteers could definitely be a potential replacement for POS tags.