Generating artificial error data for indonesian preposition error corrections

Budi Irmawati, author

Abstrak

Large-scale annotated data written by second language learners are not always available for low-resource languages such as Indonesian. To cope with data scarcity, it is important to generate ‘learner-like’ artificial error sentences when the available real learner data is insufficient and language experts cannot construct data. In this paper, we propose a new method for generating effective error-injected artificial data to proliferate training examples for preposition error correction tasks. Our method first generates a large scale of noisy artificial error data via the use of a simple error injection method. It then selectively removes the uninformative (noisy) instances from the artificial data. We assume that ‘good’ artificial preposition error data would be effective training data for error correction tasks. Therefore, to evaluate the goodness of the generated artificial data, we used the generated artificial data as training data to correct preposition errors in real learners’ sentences. The results of our study indicate that the use of our artificial data for training improves preposition error correction performance. The results also show that training on a smaller sized of good instances outperforms training on much larger-sized noisy instances as well as that on sentences written by native speakers. This method is language-independent and easy to apply to other low-resource languages because it assumes only a small size of learner error data and uses features that could be extracted automatically from linguistically annotated sentences.