Pembangunan Model Normalisasi Teks Bahasa Indonesia dengan Pendekatan Statistical Machine Translation Secara Semi-Supervised = Semi-Supervised Statistical Machine Translation Model for Indonesian Text Normalization

Tatag Aziz Prawiro, author

Abstrak

Normalisasi teks merupakan task pada NLP yang dapat digunakan untuk meningkatkan performa dari aplikasi-aplikasi NLP lain. Penelitian tentang normalisasi teks pada bahasa Indonesia masih jarang dan kebanyakan masih hanya menormalisasi pada tingkat token. Penelitian ini bertujuan untuk mengevaluasi pembangunan model normalisasi dengan menggunakan algoritma statistical machine translation (SMT). Isu dari pendekatan machine translation dalam penyelesaian task normalisasi teks adalah butuhnya data yang relative banyak. Penelitian ini juga melihat bagaimana pengaruh dari pemelajaran semi-supervised dengan cara menggunakan pseudo-data dalam pembangunan model normalisasi teks dengan algoritma statistical machine translation. Model SMT memiliki performa yang cukup baik pada data tanpa tanda baca, namun memiliki performa yang buruk pada data bertanda baca karena banyaknya noise. Pendekatan semi-supervised menurunkan performa SMT secara keseluruhan, namun, pada jenis data tidak bertanda baca penurunan relatif tidak signifikan.

.....Text normalization is a task in NLP which can be used to improve the performance of other NLP applications. Research on text normalization in Indonesian language is still rare and most only normalize at the token level. This study attempts to improve the development of the normalization model by using the statistical machine translation (SMT) algorithm. The issue in building a good performing text normalization model using the machine translation approach is the relatively large data needs. This research also looks at how using semi-supervised learning by using pseudo-data as training data in SMT approach affects text normalization performance. The SMT model has a fairly good performance on data without punctuation, but has poor performance on data with a punctuation due to the amount of noise. The semi-supervised approach reduces the overall performance of the SMT model, but the reduction in performance is relatively insignificant on data without punctuation.