Pengenalan entitas bernama pada Dokumen Wikipedia dan Berita Bahasa Indonesia dengan Pendekatan Conditional Random Field = Named-Entity Recognition On Indonesian Wikipedia and News Document Using Conditional Random Field Approach

Alif Ahsanil Satria, author

Abstrak

Pengenalan entitas bernama (named-entity recognition atau NER) adalah salah satu topik riset di bidang pemrosesan bahasa alami (natural language processing atau NLP). Pen- genalan entitas bernama merupakan langkah awal mengubah unstructured text menjadi structured text. Pengenalan entitas bernama berguna untuk mengerjakan NLP task yang lebih high-level seperti ekstraksi informasi (information extraction atau IE), Question Answering (QA), dan lain-lain. Penelitian ini memanfaatkan data berita dan wikipedia masing-masing sebanyak 200 dokumen yang digunakan untuk proses pengujian dan pelatihan. Penelitian ini mencoba mengeksplorasi entitas bernama baru yang tidak sebatas Person, Location, dan Organization. Named entity baru tersebut adalah Event, Product, Nationalities Or Religious or Political groups (NORP), Art, Time, Language, NonHuman or Fictional Character (NHFC), dan Miscellaneous. Jadi, penelitian ini menggunakan 11 entitas bernama. Dalam penelitian ini, permasalahan tersebut dipandang sebagai se- quence labelling. Penelitian ini mengusulkan penggunaan model conditional random field sebagai solusi permasalahan ini. Penelitian ini mengusulkan penggunaan fitur tambahan seperti kata sebelum, kata sesudah, kondisi huruf kapital di awal kata, dan lain-lain, serta word embedding. Penelitian ini menghasilkan performa dengan nilai F-measure terbaik sebesar 67.96% untuk data berita dan 67.09% untuk data wikipedia.

.....Named Entity Recognition or NER is one of research topics in Natural Language Pro- cessing (NLP) subject. NER is the first step to transform unstructured text to structured text. NER is used for doing more high-level NLP task such as Information Extraction (IE), Question Answering (QA), etc. This research uses news and wikipedia data with 200 documents of each, which is used for training and testing process. This research tries exploring new named entities in addition to Person, Location, and Organization. These named entities are Event, Product, Nationalities Or Religious or Political groups (NORP), Art, Time, Language, NonHuman or Fictional Character (NHFC), and Miscellaneous. Therefore, this research uses 11 named entities. This research views this problem as sequence labelling. This research proposes conditional random field model as the solution for this problem. This research proposes some features, for example additional features such as previous word, next word, initial capital letter condition, etc, and word embedding. This research results p1qerformance with the best F-Measure of 67.09% for wikipedia data and 67.96% for news data.