Tugas Pekan ke-13 Ekstraksi Informasi

Batas pengumpulan: **Jumat 18 Desember 2020, pukul 10.59 pagi**, melalui LMS

Deskripsi

Pengamatan eksperimen Named Entity Recognition (NER) berdasarkan tutorial di https://colab.research.google.com/drive/1oXVB1FaSd9VsJ75htstvXVHRW0OT-PBj?usp=sharing .

Pengamatan yang dilakukan

Pada tutorial NER tersebut, metode yang digunakan adalah CRF, di mana digunakan beberapa fitur. Lakukan eksperimen dengan menggunakan fitur yang berbeda:

1. TANPA fitur POSTag *current* kata dan POSTag kata konteks
2. TANPA fitur ortografi (terlihat dari fungsi istitle(), isupper(), dan isdigit()) *current* kata dan pada kata konteks
3. TANPA fitur akhiran pada *current* kata

Amati performansi NER dengan setting fitur yang berbeda-beda tersebut, bandingkan hasilnya dengan penggunaan keseluruhan fitur seperti pada setting awal. Berikan analisis terhadap hasil yang diperoleh, **jelaskan apa dugaan penyebabnya jika kinerja lebih baik atau lebih buruk** tanpa penggunaan fitur-fitur tersebut!

Bonus

Terapkan NER dengan metode CRF tersebut untuk dataset Bahasa Indonesia, di mana dataset dapat diperoleh di https://github.com/indobenchmark/indonlu/tree/master/dataset/nerp_ner-prosa . Gunakan data latih train_preprocess.txt untuk membangun model, dan lakukan pengujian pada data valid_preprocess.txt. Perhatikan bahwa pada dataset tersebut TIDAK terdapat informasi POSTag, sehingga fitur POSTag tidak dapat digunakan. Berikan laporan kinerja NER (precision, recall, F1) untuk tiap label.

File yang harus dikumpulkan:

1. Program: 1 file berisi modifikasi fitur yang dilakukan pada dataset Bahasa Inggris, ditambah 1 file berisi eksperimen dengan dataset Bahasa Indonesia jika mengerjakan bonus.
2. Laporan: 1 file pdf, maksimum panjang laporan adalah 5 halaman. Laporan berisi kinerja NER dengan setting fitur yang berbeda dan jawaban analisis terhadap hasil yang diperoleh .

Detail penilaian:

Program

- kelengkapan eksperimen dengan setting fitur yang berbeda [30 poin]
- kelengkapan pengukuran kinerja/evaluasi [10 poin]
- **bonus**: dapat memproses file masukan [10 poin]
- **bonus**: kelengkapan eksperimen, ekstraksi fitur dan pelatihan CRF serta evaluasi [20 poin]

Laporan:

- kelengkapan jawaban 3 setting fitur yang berbeda [30 poin]
- ketepatan analisis jawaban nomor 1 [30 poin]
- **bonus:** kelengkapan laporan kinerja/evaluasi NER Bahasa Indonesia dan analisis singkat [20 poin]

Jika ada pertanyaan, silakan disampaikan melalui *channel* pekan_13_tugas_ekstraksi_informasi di slack, https://pemrosesanbah-c5k6846.slack.com/archives/C01G5SGGUKY