BAB V

# UJI COBA

Bab kelima adalah uji coba yang dilakukan dalam tugas akhir ini. Dimulai dengan pembahasan cara hitung evaluasi dari Sequence-to-Set Network, yaitu Micro F1 Score. Kemudian uji coba terhadap metode Sequence-to-Set Network. Uji coba akan dijelaskan dalam beberapa subbab bergantung pada konfigurasi yang diubah pada Sequence-to-Set Network. Kemudian akan dijelaskan metode pembanding dengan hasil Sequence-to-Set Network.

## Evaluator Performa Hasil dari Tugas Akhir

Evaluator adalah cara penghitungan performa dari tugas akhir ini, tepatnya metode Sequence-to-Set Network dalam bahasa Indonesia. Performa prediksi sebuah metode biasanya dihitung dengan akurasi prediksi dengan target output/prediksi sebenarnya. Penghitungan ini akan menggunakan library/tool yang membantu penghitungan yang akurat dan benar. Untuk jenis evaluator yang digunakan telah dijelaskan pada subbab teori penunjang. Jenis F1 Score pada tugas akhir ini yang digunakan adalah Macro F1 Score dan Micro F1 Score. Namun untuk perbandingan F1 Score terbaik diambil dari Micro F1 Score.

Penghitungan dalam program tugas akhir ini menggunakan bantuan library yaitu *sklearn*. Fungsi yang digunakan dari sklearn adalah precision\_recall\_fscore\_support dari modul *metrics*. Fungsi tersebut memberi kemudahan dengan secara singkat program langsung mendapat nilai dari precision, recall, F1 Score, juga support tiap label. Parameter dari fungsi tersebut yang digunakan adalah y­\_true dan y\_pred diisi dengan golden entites dan entitas yang telah diprediksikan. Kemudian dua parameter lainnya adalah *labels* untuk memberikan jenis class/label yang ada, dalam tugas akhir ini label yang digunakan adalah jenis-jenis entitas. Parameter terakhir yang digunakan adalah *average*, yang digunakan adalah micro dan macro namun selain itu ada beberapa jenis average yang disediakan sklearn seperti *weighted*, *samples*, *binary*.

## Uji Coba Sequence-to-Set Network pada Nested NER

Konfigurasi dari *hyperparameter* yang akan diubah untuk uji coba ini ada empat jenis yaitu *Batch Size*, *Learning Rate*, *Gradient Norm*, *Dropout*. Penilaian akurasi menggunakan Micro F1 Score dan diambil yang terbaik dari seluruh *epoch* yang dilewatkan. Pembagian dataset adalah 90% *training* dan *development* dan 10% untuk *testing*. Total ada 42 hyperparameter yang dapat dikonfigurasikan, namun akan disebut hyperparamter yang akan diubah untuk uji coba.

Nilai *default* untuk tiap hyperparameter adalah batch size sejumlah delapan, learning rate dengan nilai 2e-5, gradient norm dengan nilai 1.0 dan nilai dropout 0.1. Catatan untuk hyperparameter dropout, terdapat tiga hyperparameter yang dapat diubah nilai dropout nya yaitu prop\_drop, lstm\_drop dan char\_lstm\_drop. Hyperparameter tersebut masing-masing yaitu, prop\_drop untuk merubah nilai probabilitas dropout dalam training model Sequence-to-Set Network. lstm\_drop untuk mengubah nilai probabilitas dropout untuk embedding akhir BiLSTM. char\_lstm\_drop untuk mengubah nilai probabilitas dropout untuk embedding character-level BiLSTM.

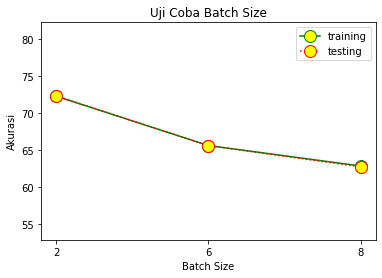
Tabel 5.1  
Hasil Uji Coba F1 Score Terbaik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Jenis Entitas** | **Precision** | **Recall** | **F1 score** | **Support** |
| TIME | 67.21 | 75.21 | 70.98 | 714 |
| LOC | 71.33 | 70.88 | 71.10 | 1583 |
| EVENT | 67.73 | 48.53 | 56.55 | 307 |
| DATE | 88.87 | 91.36 | 90.10 | 463 |
| PER | 75.62 | 81.10 | 78.26 | 3212 |
| MISC | 49.27 | 56.16 | 52.49 | 479 |
| ORG | 67.10 | 68.04 | 67.57 | 2353 |
| MICRO | 71.00 | 73.60 | 72.28 | 9111 |
| MACRO | 69.59 | 70.18 | 69.58 | 9111 |

Uji coba dilakukan dengan cara untuk hyperparameter yang sedang diuji coba akan diganti, sedangkan hyperparameter lainnya mengikuti nilai default. Secara kesimpulan dari uji coba Sequence-to-Set Network adalah F1 Score tertinggi adalah uji coba perubahan untuk batch size sebanyak dua dengan F1 Score 72.28%. Detail hasil uji coba untuk tiap jenis entitas, baik F1 Score, support, precision dan recall telah ditampilkan pada tabel 5.1.

### Pengaruh Batch Size

Uji coba pertama adalah perubahan hyperparameter jumlah batch size terhadap nilai akurasi terbaik diakhir uji coba. Nilai untuk batch size yang di uji coba adalah batch size sejumlah 2, batch size sejumlah 6, batch size sejumlah 8 (nilai default). Hyperparameter lainnya akan mengikuti nilai default selama uji coba ini dilakukan. Gambar 5.1 menunjukkan hasil dari uji coba dengan merubah jumlah batch size.

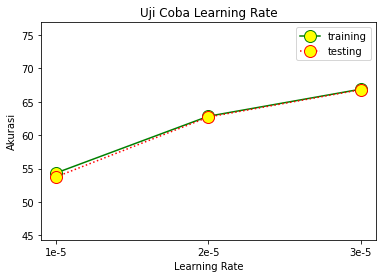


Gambar 5.1  
Hasil Uji Coba Jumlah Batch Size

Dapat dilihat perubahan terjadi cukup signifikan untuk batch size jumlah dua menuju batch size jumlah 6, sebanyak 7% (72.28 dan 65.59). Namun untuk perubahan dari batch size jumlah 6 menuju batch size jumlah 8 hanya sebanyak sekitar 3% (62.82 dan 65.59). Untuk akurasi saat proses testing, hasil akurasi tidak memiliki perbedaan jauh, perbedaan ditemukan kurang dari satu (72.31, 65.59, dan 62.7). Kesimpulan sementara untuk pengaruh jumlah batch size adalah mempengaruh tidak terlalu signifikan (kurang lebih 3%). Ada pun kesimpulan berkurangnya nilai F1 Score semakin bertambah jumlah batch size.

### Pengaruh Learning Rate

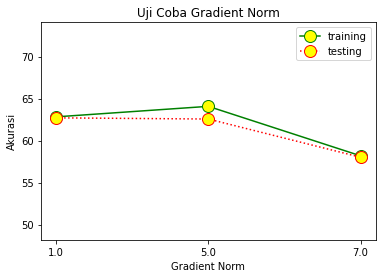
Uji coba berikutnya perubahan hyperparameter nilai learning rate. Nilai untuk learning rate yang di uji coba adalah 1e-5, 2e-5 (nilai default) dan 3e-5. Selain learning rate akan bernilai default selama uji coba ini dilakukan. Gambar 5.2 menunjukkan hasil dari uji coba dengan merubah learning rate. Perubahan yang muncul nilai learning rate 1e-5 menuju learning rate 2e-5, sebanyak 4% (54.34 dan 62.82). Jarak perubahan F1 Score dari uji coba kedua dan ketiga juga sama dengan sebelumnya yaitu 4% (62.82 dan 66.86). Akurasi proses testing tidak berbeda jauh (53.75, 62.7, dan 66.79). Kesimpulan dari uji coba ini tidak terlalu signifikan dan makin bertambah nilainya makin tinggi nilai F1 Score.



Gambar 5.2  
Hasil Uji Coba Learning Rate

### Pengaruh Gradient Norm

Uji coba berikutnya adalah perubahan nilai gradient norm. Nilai yang akan di uji coba adalah 1.0 (nilai default), 5.0, 7.0 dan hyperparameter lainnya mengikuti nilai default pada uji coba ini. Gambar 5.3 menunjukkan hasil dari uji coba dengan perubahan yang muncul adalah dari nilai 1,0 menuju 5,0, terjadi sekitar 2% (62.82 dan 64.07). Tetapi saat nilai dari gradient norm dinaikkan, akurasi mengalami penuruan sebanyak 7% sehingga dapat dikatakan perubahan jika gradient norm dinaikkan mengalami perubahan signifikan namun penurunan.



Gambar 5.3  
Hasil Uji Coba Gradient Norm

### Pengaruh Dropout

Uji coba terakhir adalah perubahan hyperparameter nilai dropout yang akan dilakukan kepada tiga variabel berbeda seperti yang dijelaskan (prop\_drop, lstm\_drop dan char\_lstm\_drop). Nilai dropout yang di uji coba adalah nilai 0.1 (nilai default), 0.3, 0.5. Hyperparameter lainnya akan mengikuti nilai default selama uji coba ini dilakukan. Gambar 5.4 menunjukkan hasil dari uji coba perubahan nilai dropout.

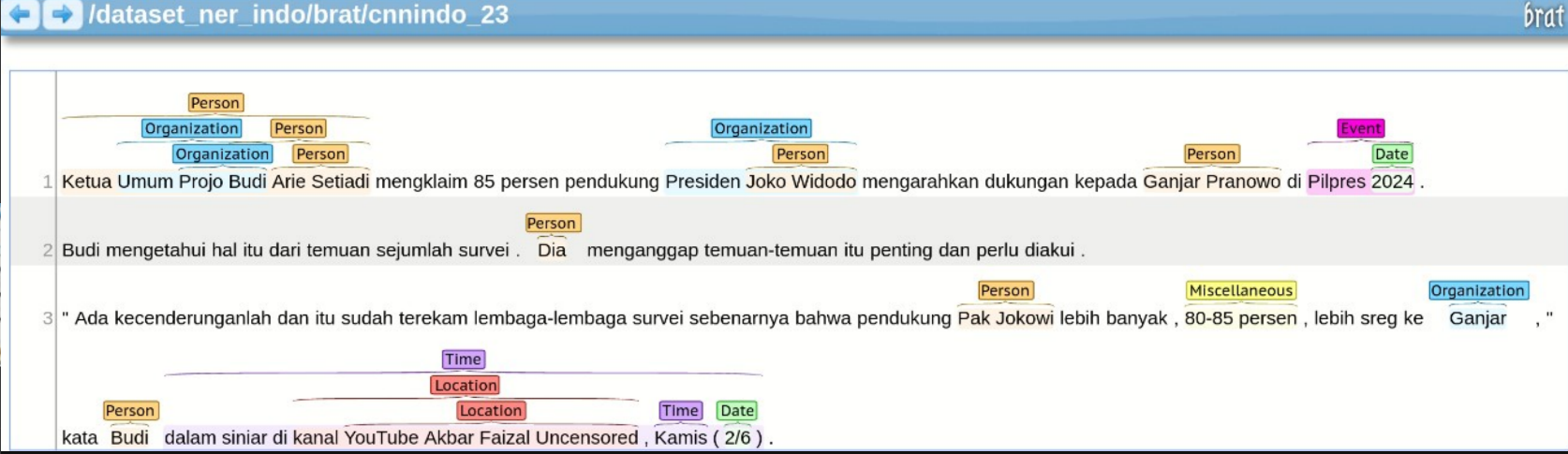


Gambar 5.4  
Hasil Uji Coba Dropout

Hasil dari perubahannya tidak signifikan dengan perubahan F1 Score yang tidak lebih dari 1% (akurasi training adalah 62.82, 62.46, 62.95). Perbandingan dengan akurasi test juga tidak beda jauh (62.7, 62.44, dan 62.6). Kesimpulan untuk pengaruh perubahan ketiga variabel dropout adalah mempengaruh tidak signifikan (kurang dari 1%).

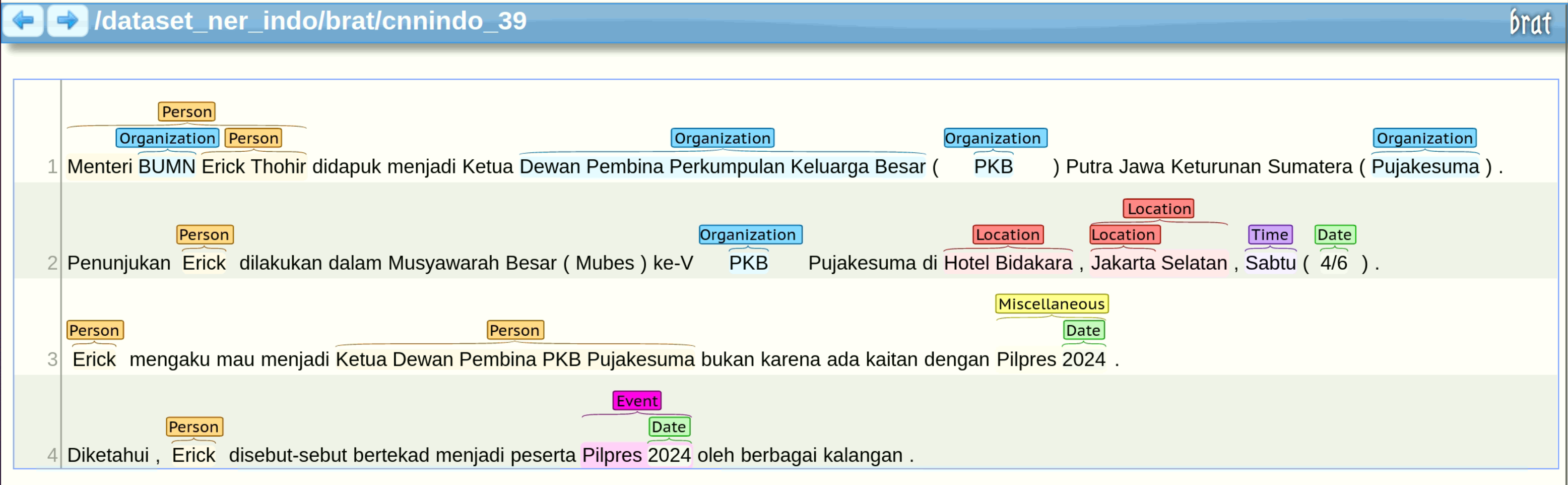
## Hasil Prediksi Sequence-to-Set Network Bahasa Indonesia

Hasil akhir dari tugas akhir ini adalah program yang dapat digunakan user untuk memberi input kalimat dan mendapat prediksi entitas dari kalimat tersebut. Program ini membutuhkan model untuk melakukan prediksi itu, dan model yang digunakan adalah hasil dari uji coba Sequence-to-Set Network dengan akurasi terbaik. Uji coba terbaik adalah uji coba perubahan nilai batch size menjadi dua dengan nilai akurasi 72.28%. Bagian dari subbab ini dilakukan analisa terhadap beberapa contoh hasil prediksi dari model terbaik ini, baik contoh dengan hasil yang baik maupun hasil yang kurang akurat. Beberapa contoh yang diberikan adalah hasil prediksi untuk domain berita politik, contoh lain diambil dari topik lain dari halaman berita CNN Indonesia seperti “Hukum dan Kriminal” atau “Peristiwa”. Adapun contoh kalimat dengan kasus tertentu seperti jika diberikan kalimat yang singkat, maka pelabelan akan kurang akurat dibandingkan dengan kalimat yang lebih banyak konteks yang diberikan.



Gambar 5.5   
Hasil Prediksi Domain Politik 1

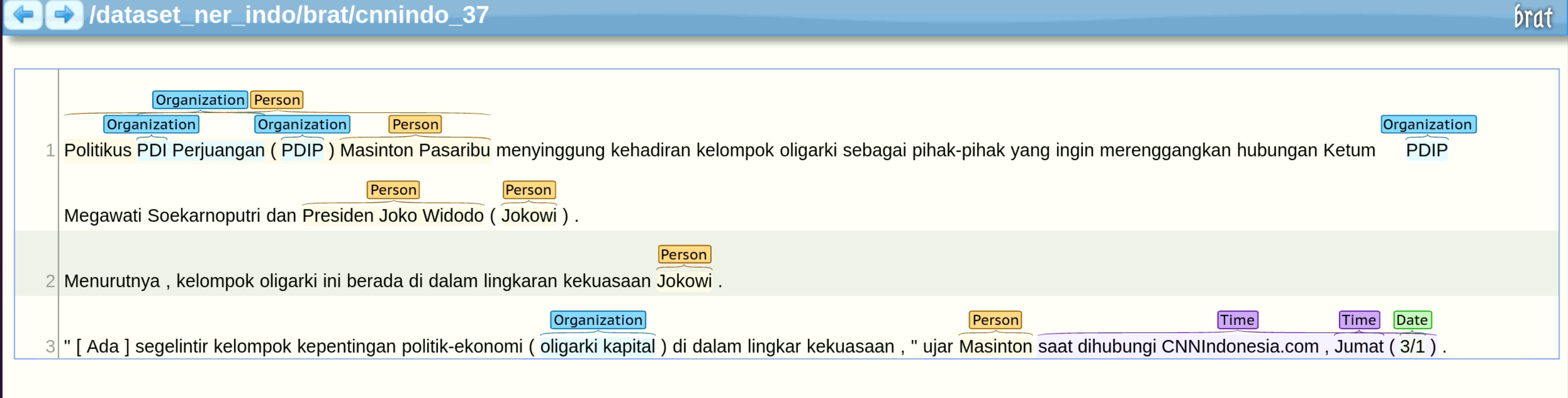
Gambar 5.5 adalah uji coba input kalimat pertama dalam domain berita politik. Sebelumnya ketentuan pelabelan tiap jenis entitas telah dijelaskan pada bab ketiga subbab Dataset dan Tagset Bahasa Indonesia. Salah satunya dijelaskan bahwa pelabelan suatu jabatan hanya akan dilabelkan jika direferensikan dengan nama seseorang atau jabatan tersebut direferensikan sebagai subyek bukan suatu jabatan saja. Pada gambar 5.5 pelabelan pada kata “Ketua Umum Projo Budi Arie Setiadi” ada kesalahan khususnya di label yang bersarang yaitu kedua label entitas organization seharusnya tidak ada. Dan label person paling dalam dengan kata-kata “Arie Setiadi” tidak perlu dilabelkan karena nama tersebut sudah dilabelkan secara lengkap dan benar dengan kata-kata “Budi Arie Setiadi”. “85 persen” seharusnya dilabelkan sebagai entitas miscellaneous, dan prediksi “Presiden Joko Widodo” sebagai entitas organization kurang tepat seharusnya kata sebelumnya yatu “pendukung” seharusnya juga dilabelkan. Dan “Presiden Joko Widodo” seharusnya dilabelkan sendiri sebagai entitas person. Kata “Budi” pada baris kedua merupakan entitas person namun tidak dilabelkan, dan kata pengganti seperti dia, mereka, dan lainnya tidak dilabelkan karena tidak termasuk sebuah entitas yang bernama. Baris ketiga tidak ditemukan kekurangan dalam pelabelan, namun lebih tepat secara konteks bahwa kedua entitas lokasi dihapuskan.



Gambar 5.6   
Hasil Prediksi Domain Politik 2

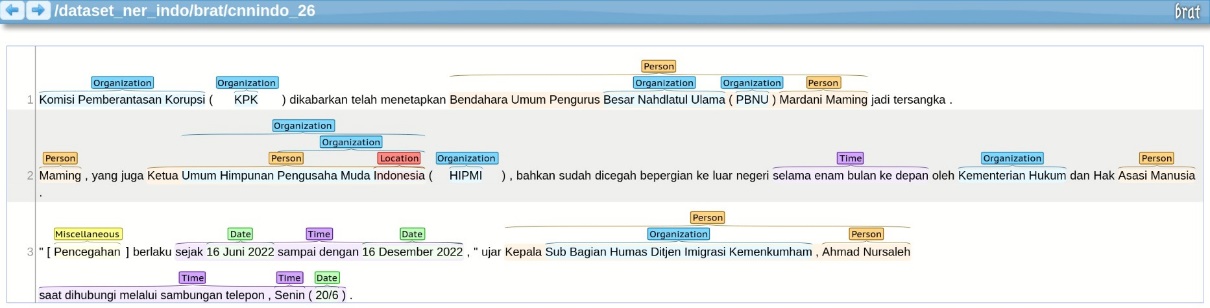
Contoh kedua dalam domain politik adalah gambar 5.6. Entitas organization tidak saja pada “Dewan Pembina Perkumpulan Keluarga Besar” namun ada satu label tambahan dengan entitas organization di atasnya yaitu pada kata-kata “Perkumpulan Keluarga Besar”. Dilewatkan juga prediksi “Putra Jawa Keturunan Sumatera” yang sebenarnya entitas organization. Tidak ada kekurangan pelabelan dari kalimat kedua kecuali “Musyawarah Besar (Mubes) ke-V PKB Pujakesuma” adalah entitas event dan yang dilabelkan organization adalah “PKB Pujakesuma” tidak hanya “PKB”.

Lanjut kepada kalimat ketiga, jabatan ketua bukanlah person karena secara konteks jabatan tersebut direferensikan hanya sebagai jabatan bukan seseorang. Label organization juga tidak ditemukan dalam kalimat ini untuk “Dewan Pembina PKB Pujakesuma”, “PKB”, dan “Pujakesuma”. Hal ini mungkin karena kurang ada nilai Transformer tidak di training ulang untuk mengenal kata-kata organization dalam bahasa Indonesia, bisa juga karena kata-kata ini tidak ditemukan dalam dataset. “Pilpres 2024” adalah entitas event, hal ini dapat terjadi karena dataset tugas akhir ini yang mungkin beberapa kata-kata yang seharusnya entitas event masih dilabelkan dengan miscellaneous. Kalimat terakhir tidak ada kekurangan.



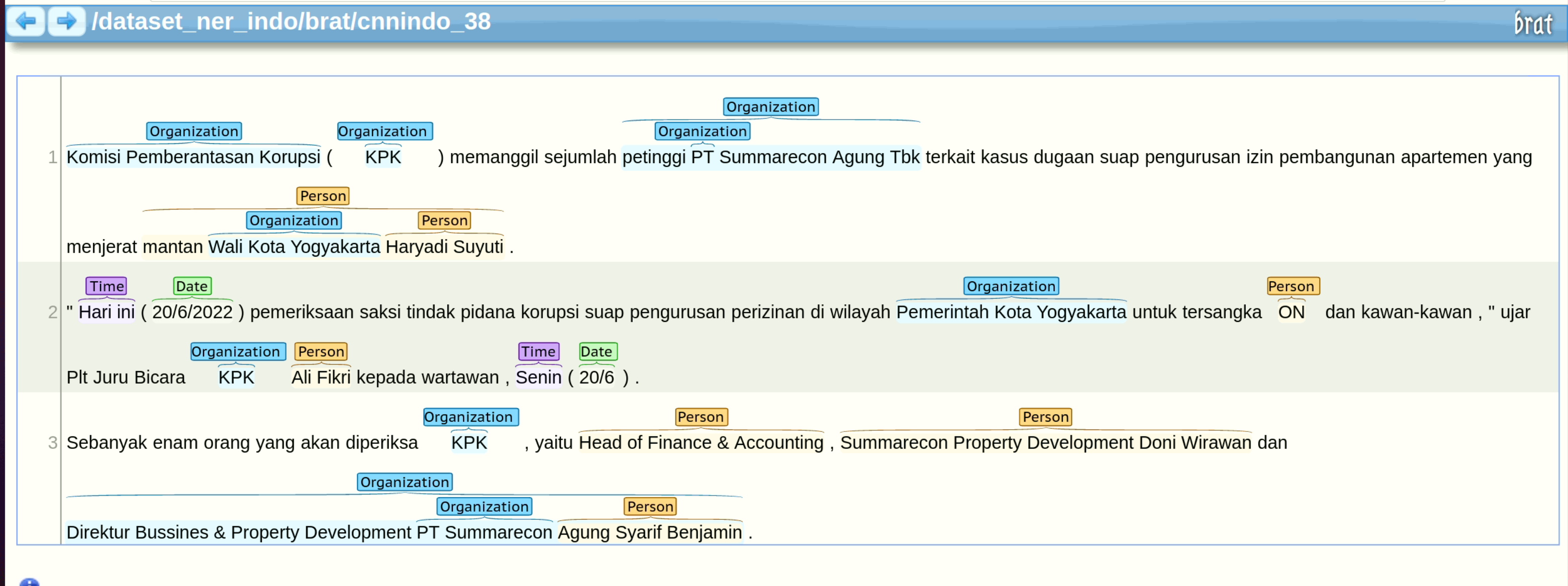
Gambar 5.7   
Hasil Prediksi Domain Politik 3

Contoh hasil prediksi domain politik terakhir adalah gambar 5.7. “PDI” tidak perlu dilabelkan sebagai organization karena kata tersebut adalah satu kesatuan nama sebagai “PDI Perjuangan”. Selain itu, “Ketum PDIP Megawati Soekarnoputri” adalah entitas person dengan entitas person juga bersarang didalamnya untuk kata-kata “Megawati Soekarnoputri”. “Joko Widodo” sendiri juga perlu dilabelkan sebagai entitas person. Untuk entitas yang masih kurang hanya pada baris ketiga kata “CNNIndonesia.com” termasuk entitas organization. Selain ini, kalimat lainnya prediksi telah dilakukan dengan baik.



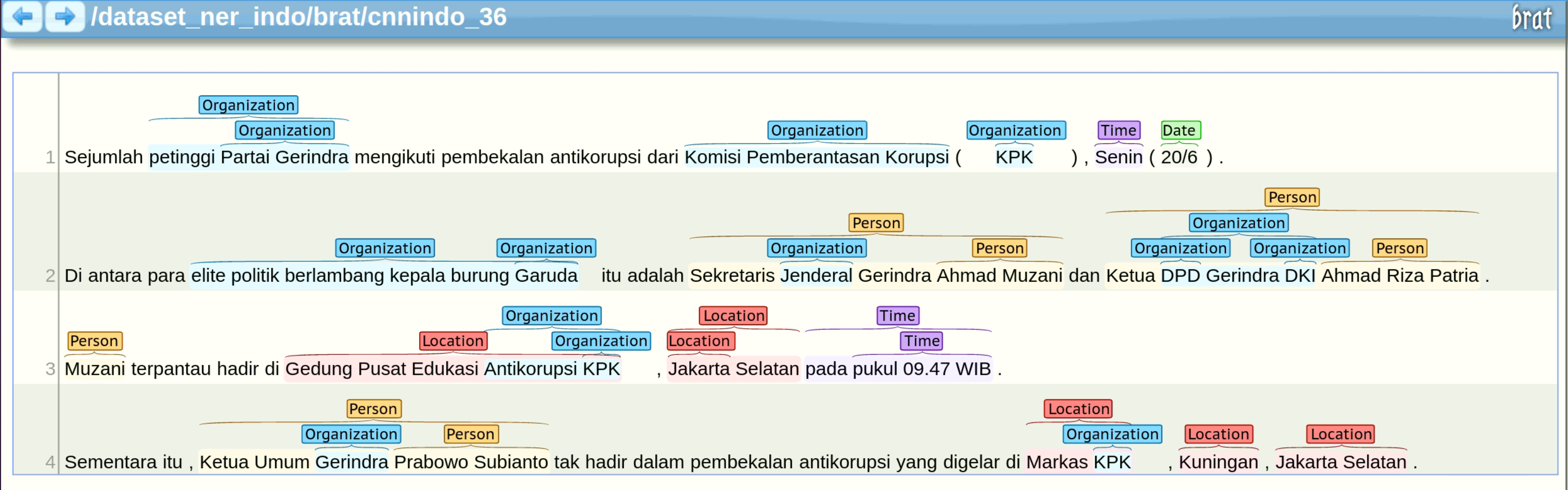
Gambar 5.8   
Hasil Prediksi Domain Hukum dan Kriminal 1

Gambar 5.8 adalah uji coba prediksi Nested NER pertama yang menggunakan contoh di luar domain politik. Untuk contoh-contoh kedepan akan diberikan kalimat-kalimat yang bukan domain politik. Kalimat pertama contoh ini tidak ada kesalahan. Untuk kalimat kedua terdapat kesalahan hanya pada span/jangkauan pelabelan organization di kata “Umum Himpunan Pengusaha Muda Indonesia” berlebihan. Span untuk entitas tersebut tidak memerlukan kata pertama sehingga seharusnya yang dilabelkan hanya “Himpunan Pengusaha Muda Indonesia”. Kesalahan berikutnya adalah “Asasi Manusia” dilabelkan sebagai entitas person perlu dihapuskan. Kesalahan pada kalimat ketiga hanya satu yaitu pelabelan miscellaneous tidak diperlukan. Dibandingkan dengan gambar 5.6, hasil prediksi pada gambar 5.8 lebih akurat. Hal ini dapat terjadi karena pada gambar 5.8, model lebih mengenal kata-kata dan konteks dari berita. Sedangkan gambar 5.6 ada kata-kata yang tidak dikenal, seperti “Musyawarah Besar (Mubes) ke-V PKB Pujakesuma” sebagai event. Untuk jenis entitas organization pada kata “Dewan Pembina Perkumpulan Keluarga Besar” juga tidak terdapat dalam dataset, namun “Himpunan Pengusaha Muda Indonesia” juga tidak ada namun model tetap bisa prediksi kata tersebut sebagai organization. Untuk “Dewan Pembina Perkumpulan Keluarga Besar” mungkin tidak dilabelkan karena model tidak mengenal adanya organization dalam dataset yang terlabel dengan “Dewan Pembina” dan disambung dengan nama organisasi. Sehingga model tidak mengenal bahwa didalam istilah “Dewan Pembina” terdapat nama organisasi yang seharusnya juga dilabelkan organisasi.



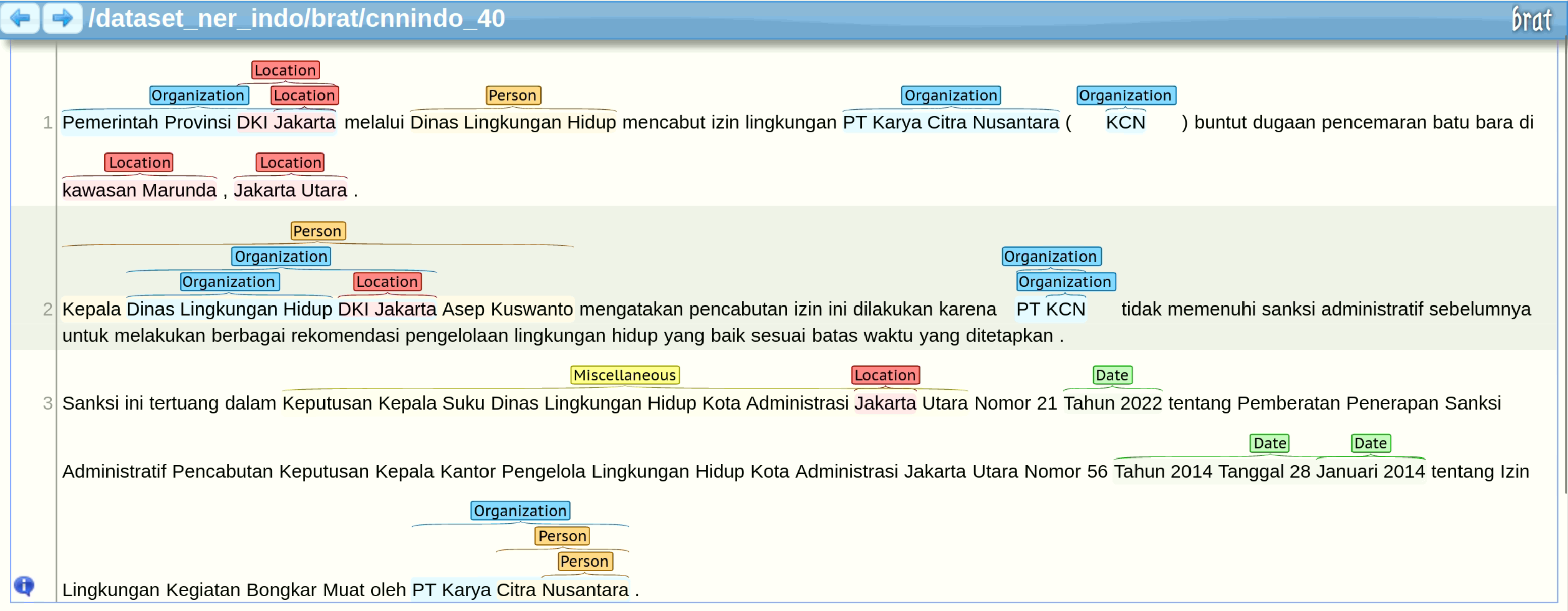
Gambar 5.9   
Hasil Prediksi Domain Hukum dan Kriminal 2

Hasil prediksi selanjutnya ditunjukkan di gambar 5.9. “Wali Kota Yogyakarta” pada kalimat pertama tidak dilabekan sebagai organization, dan kata “Yogyakarta” perlu dilabelkan sebagai location. Selain itu, jenis entitas organization dalam kata “PT” kurang lengkap, yang benar adalah “PT Summarecon Agung Tbk”. Kalimat kedua kekurangan juga dalam kata “Yogyakarta” perlu dilabelkan sebagai location. Juga “Plt Juru Bicara KPK Ali Fikri” termasuk jenis entitas person. Dan kalimat ketiga terdapat kesalahan pada pelabelan jenis person menjadi dua yaitu “Head of Finance & Accounting” dan “Summarecon Property Development Doni Wirawan” yang seharusnya menjadi satu, menjadi “Head of Finance & Accounting, Summarecon Property Development Doni Wirawan”. Telah dicoba kembali dengan kalimat ini, namun tanda koma telah dihapuskan. Jika tanda koma tersebut dihapus, maka model dapat prediksi dengan benar sesuai dengan yang baru saja dijelaskan. Membuktikan bahwa dengan kesalahan penulisan, dapat memberikan ambiguitas dan konteks yang berbeda untuk model. Perlu ditambahkan juga “Doni Wirawan” dilabelkan sebagai person. Dibuktikan juga bahwa terjadi *error propagation* dalam pelabelan dari contoh kata “Yogyakarta”. Jika dari awal model tidak mengenal kata tersebut sebagai lokasi, maka sepanjang dokumen tidak akan dianggap sebagai entitas lokasi.



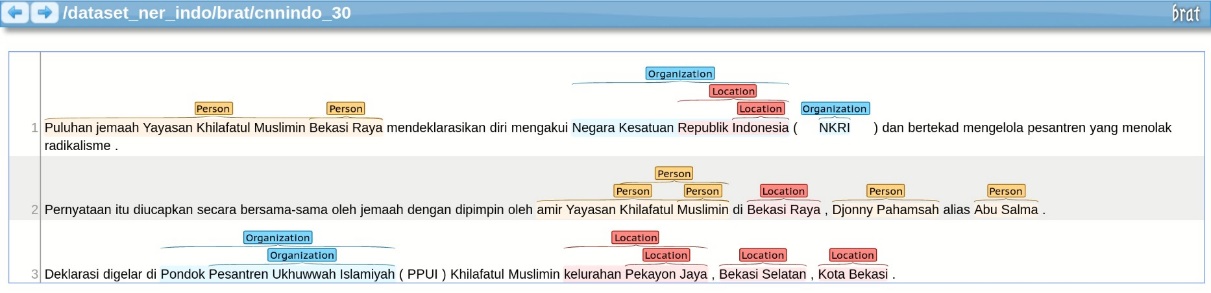
Gambar 5.10   
Hasil Prediksi Domain Hukum dan Kriminal 3

Gambar 5.10 adalah hasil prediksi dalam domain hukum dan kriminal terakhir yang akan dibahas. Kesalahan untuk dokumen ini hanya beberapa kata yang tidak perlu dilabelkan seperti “Garuda” dan “Jendral” bukan entitas organization. Dan pelabelan jenis entitas yang salah untuk “DKI” baris kedua yang benar adalah location. Kekurangan pelabelan ada pada kata “Jakarta” paling akhir dokumen sebagai entitas location, dan “Gerindra” pada baris kedua seharusnya organization.



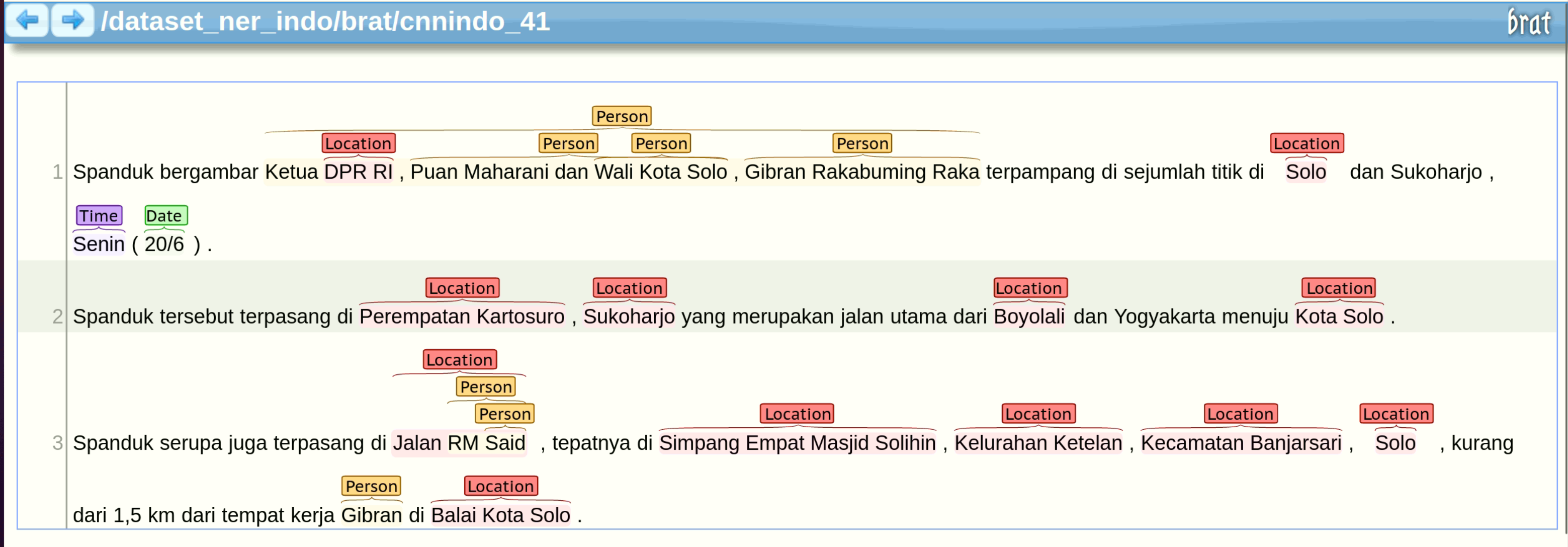
Gambar 5.11   
Hasil Prediksi Domain Peristiwa 1

Domain selanjutnya juga diambil dari CNN Indonesia yaitu berita peristiwa. Hasil prediksi yang dilampirkan pertama adalah gambar 5.11. Pada hasil prediksi ini terdapat banyak kekurangan dan kesalahan yang ditemukan. Tiga kesalahan dibaris pertama adalah “Jakarta” entitas location tidak perlu dilabelkan karena “DKI Jakarta” adalah kesatuan entitas. Prediksi jenis entitas pada kata-kata “Dinas Lingkungan Hidup” seharusnya organization bukan person. Dan kata “Jakarta” yang seharusnya dilabelkan location adalah “Jakarta” di akhir kalimat. Kalimat kedua hanya memiliki kekurangan jenis entitas person pada nama “Asep Kuswanto”. Dan kalimat ketiga memiliki kekurangan yang cukup banyak. Kalimat ketiga menyebut “Jakarta Utara” dua kali, kata-kata tersebut perlu dilabelkan sebagai location, juga untuk entitas location sebagai bersarangnya pada kata “Jakarta”. Hal ini tidak diprediksikan oleh model karena kemungkinan besar sebagian dari dataset tugas akhir ini tidak memiliki pelabelan untuk dokumen negara yang memiliki lokasi di dalamnya. Sehingga poin analisa yang muncul adalah jika ada kata yang dikenal namun secara konteks tidak direferensikan sesuai jenis entitas yang biasanya kata tersebut diberikan, maka model tidak akan memberikannya. Kemudian kesalahan terakhir adalah pelabelan person diakhir kalimat tidak diperlukan.



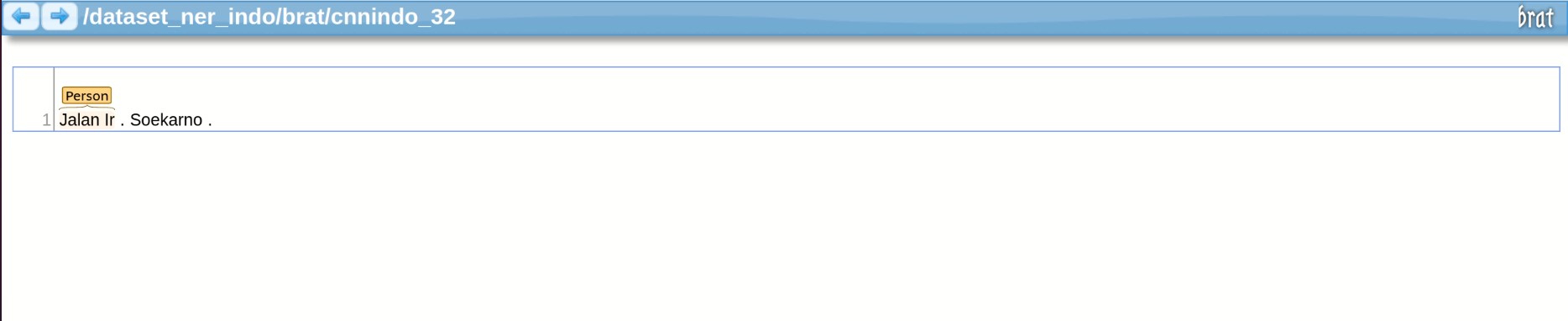
Gambar 5.12   
Hasil Prediksi Domain Peristiwa 2

Gambar 5.12 adalah hasil prediksi yang kedua dan memiliki hasil prediksi yang terdapat banyak kekurangan dapat menentukan entitas. “Yayasan Khilafatul Muslimin Bekasi Raya” adalah entitas organization, “Bekasi Raya” di baris pertama dan “Bekasi” di dalam “Bekasi Selatan” baris terakhir adalah entitas location. “Negara Kesatuan Republik Indonesia” bukan jenis entitas organization tetapi location, dan entitas bersarang di dalamnya tidak perlu dilabelkan karena sama seperti penjelasan sebelumnya, kata-kata tersebut adalah kesatuan. Pada baris kedua, kata-kata “Yayasan Khilafatul Muslimin” memang perlu dilabelkan namun jenis entitasnya adalah organization.



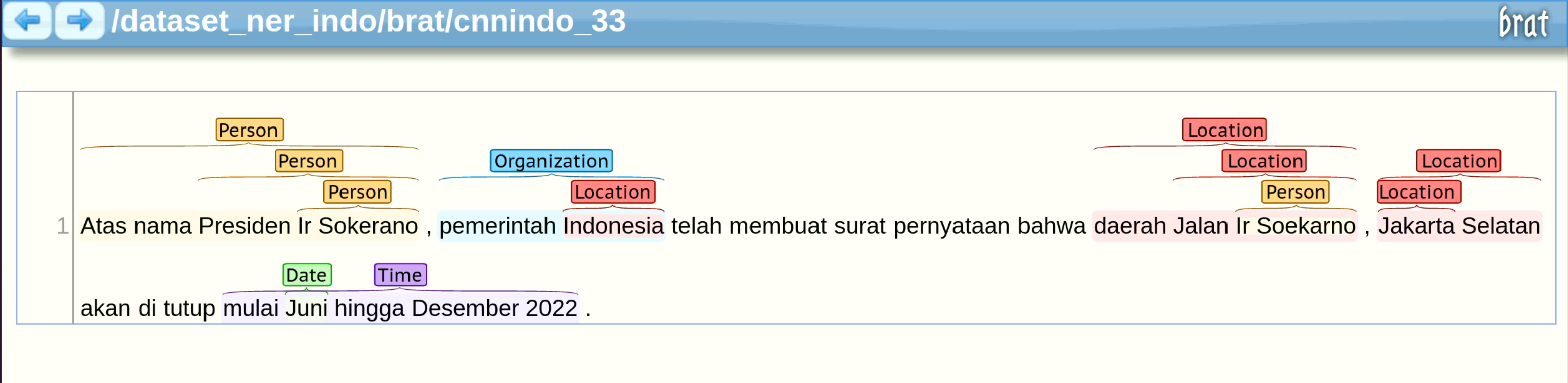
Gambar 5.13   
Hasil Prediksi Domain Peristiwa 3

Contoh hasil prediksi model terakhir dapat dilihat di gambar 5.13. Kesalahn pertama adalah penentuan kata yang perlu dilabelkan. Untuk baris pertama seluruh pelabelan entitas person salah kecuali untuk kata “Gibran Rakabuming Raka”. Dan untuk baris ketiga kata “Said” tidak perlu dilabelkan. Kata “Sukoharjo” pada baris pertama dan “Yogyakarta” pada baris kedua adalah entitas location, namun tidak dilabelkan. Dan jangkauan kata-kata yang benar untuk entitas person adalah “Ketua DPR RI, Puan Maharani”, “Puan Maharani”, “Wali Kota Solo, Gibran Rakabuming Raka”. Pelabelan “DPR RI” juga salah, jenis entitas tersebut adalah organization, juga ada pelabelan bersarang di kata “DPR” sebagai organization. Yang merupakan jenis entitas location adalah “RI”.



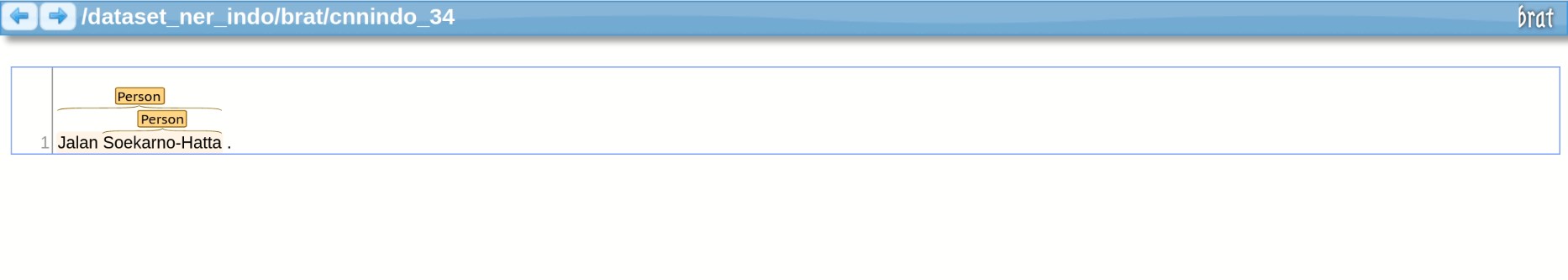
Gambar 5.14   
Hasil Prediksi Contoh Kasus 1

Beberapa contoh kasus diberikan dalam subbab ini untuk menjelaskan beberapa kasus khusus di mana input yang diterima oleh program dapat memberi dampak kepada cara model melakukan prediksi. Contohnya di gambar 5.14 adalah input user yang singkat. Pelabelan yang benar adalah “Jalan Ir. Soekarno” sebagai entitas location, “Ir. Soekarno” sebagai entitas person. Prediksi dari model yang jauh dari benar dapat disimpulkan bahwa model tidak dapat menerima kalimat yang terlalu pendek. Kalimat yang pendek tidak dapat ditemukan konteks yang lebih luas, sehingga model tidak dapat memutuskan kata tersebut sebagai entitas jenis apa. Jika dilihat pada gambar 5.15, kata “Jalan Ir. Soekarno” dan “Ir. Soekarno” dapat berhasil diprediksikan dengan benar oleh model. Hal ini dikarenakan konteks kalimat yang didapatkan oleh model.



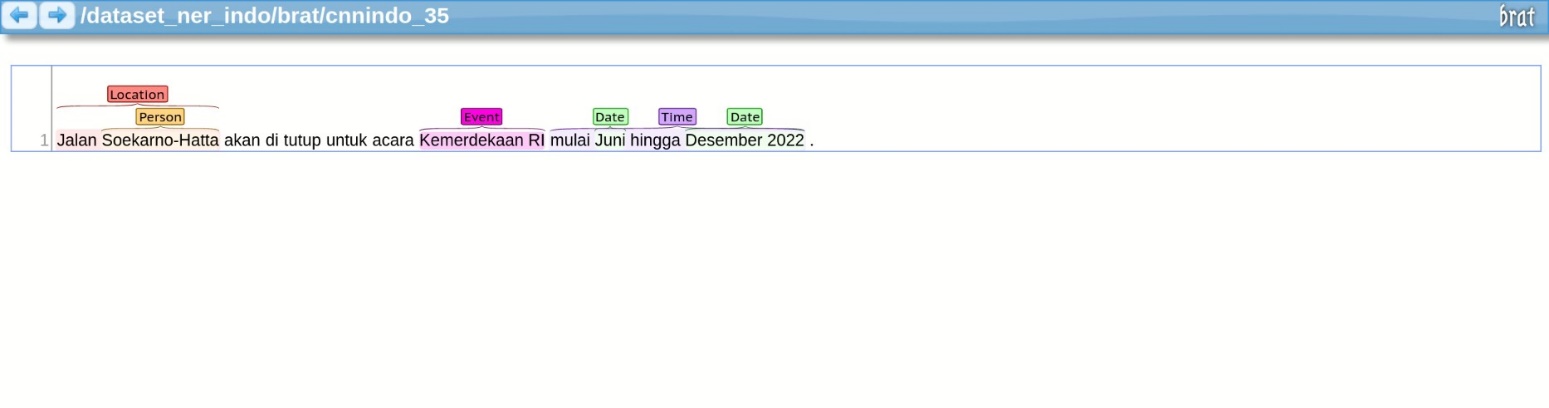
Gambar 5.15   
Hasil Prediksi Contoh Kasus 2

Contoh lainnya diambil dari gambar 5.16. Dengan struktur yang sama, input user hanya berupa nama sebuah jalan tanpa informasi lebih dari kalimat yang diberikan. Prediksi entitas yang diharapkan adalah “Jalan Soekarno-Hatta” sebagai entitas location. “Soekarno” dan “Hatta” untuk tiap kata merupakan entitas person. Tetapi prediksi yang diberikan oleh model tidak akurat.



Gambar 5.16   
Hasil Prediksi Contoh Kasus 3

Diberikan juga contoh dengan gambar 5.17. Kata-kata dari gambar 5.17 dapat ditemukan dalam gambar 5.16 tetapi dengan prediksi yang benar. Hasil-hasil prediksi ini membuktikan bahwa karena model terbiasa menerima dataset yang berkalimat panjang dengan konteks yang banyak, model tidak dapat menerima kalimat pendek dengan baik karena kekurangan informasi atau konteks.



Gambar 5.17  
Hasil Prediksi Contoh Kasus 4

## Perbandingan Metode Span-Based

Nested NER dari sebelumnya memiliki metode yang dianggap sudah cocok untuk digunakan sebagai penelitian, dan metode ini sudah berkembang banyak dalam Nested NER. Metode ini menganggap pengenalan entitas sebagai task klasifikasi dengan cara menentukan span (rentangan, suatu bagian dengan batasan kiri dan kanan). Kekurangan dari metode span-based ini, *search space* (luas pencarian) menjadi terlalu besar dan metode span-based tidak melihat nilai konteks satu entitas dengan entitas lain, yang sebenarnya dapat membantu model untuk menentukan entitas berdasarkan konteksnya. Karena itu, adanya Sequence-to-Set Network untuk memberikan solusi komputasi yang tidak seberat metode span-based tetapi memberikan akurasi yang tinggi.

Perbandingan metode akan dilakukan dengan dataset tugas akhir juga dengan dataset GENIA. Perlu diperhatikan perbedaan antar kedua dataset adalah perbandingan GENIA di ambil dari rujukkan paper tiap metode dan juga dataset GENIA memiliki keunggulan dengan proses pelabelan dilakukan oleh pakar biologis dalam domainnya sedangkan untuk dataset tugas akhir ini proses pelabelan tidak dilakukan dengan bantuan pakar domain ini.

Tabel 5.2  
Tabel Perbandingan F1 Score dengan GENIA

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Lin et al. (2019) | 75.80 | 73.90 | 74.80 |
| Luo and Zhao (2020) | 77.40 | 74.60 | 76.00 |
| Wang et al. (2020b) | 78.10 | 74.40 | 76.20 |
| Strakova et al. (2019) | - | - | 78.31 |
| Wang et al. (2020a) | 79.45 | 78.94 | 79.19 |
| Yu et al. (2020) | 81.80 | 79.30 | 80.50 |
| **Sequence-to-Set (2021)** | **82.31** | **78.66** | **80.44** |
| **Locate and Label (2021)** | **80.10** | **80.89** | **80.54** |

Metode yang akan dibandingkan adalah metode Locate and Label[[1]](#footnote-1). Locate and Label adalah metode span-based terbaru pada saat tugas akhir ini dikerjakan. Tidak hanya sebagai metode span-based/klasifikasi span, metode ini menggunakan *identifier* entitas yang memiliki dua tahap. Pertama, akan diusulkan span dengan filter dan *boundary regression* pada *seed span* untuk menemukan entitas, kemudian diberikan label kepada span yang diusulkan dan disesuaikan batas dengan kategori yang sesuai. Perbandingan F1 Score untuk metode Sequence-to-Set dengan Locate and Label dengan dataset GENIA dapat dilihat dari tabel 5.2. Dari tabel tersebut dapat dilihat F1 Score yang dimiliki keduanya hanya berbeda 0,10%. Meskipun begitu, Locate and Label menang dengan persentase 80,54% sedangkan Sequence-to-Set dengan 80,44%. Keunggulan dari Locate and Label adalah keberhasilan metode tersebut dalam metrik Recall yang menjadi 80,89% sedangkan Sequence-to-Set dengan 78,66%. Yang unggul Sequence-to-Set adalah arsitektur sistem yang lebih hemat dibandingkan dengan Locate and Label.

Tabel 5.3  
Tabel Perbandingan F1 Score dengan Dataset Tugas Akhir

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1 Score** |
| Sequence-to-Set (Terbaik) | 71.00 | 73.60 | 72.28 |
| Locate and Label | 59.21 | 78.27 | 67.42 |

Untuk perbandingan metode dalam bahasa Indonesia, konfigurasi dari metode Locate and Label telah diubah sesuai dengan kebutuhan GPU. Beberapa perubahan yang dilakukan adalah Word2Vec dan BERT akan disamakan dengan Sequence-to-Set yaitu Word2Vec WikiID dan IndoBERT. Parameter freeze\_transformer juga akan diisi sebagai True sehingga bagian transformer dari Locate and Label tidak akan dilewatkan proses training ulang. Batch size dari nilai awal 6 diubah menjadi 2 karena keterbatasan GPU, parameter konfigurasi lainnya tidak akan diubah. Tabel 5.3 merupakan perbandingan hasil performa kedua metode dengan resource dan dataset bahasa Indonesia. Kesimpulan yang dapat diambil adalah F1 Score antar metode memiliki perbedaan cukup besar yaitu sebanyak 5%. Dan untuk nilai Precision dan Recall memiliki hasil perbandingan yang sama dengan perbandingan dengan dataset GENIA. Dimana nilai Precision Sequence-to-Set akan lebih tinggi, dan Recall Locate and Label akan lebih tinggi. Dengan ini, dapat disimpulkan bahwa Sequence-to-Set lebih unggul secara prediksi Nested NER daripada Locate and Label.

## Catatan Uji Coba

Pelaksanaan uji coba tugas akhir ini memiliki beberapa kesulitan, kesulitan utamanya adalah kurangnya *resource* (fasilitas/sumber daya) secara memory untuk menjalankan arsitektur Sequence-to-Set secara keseluruhan. Penelitian pada awal dijalankan pada *environment* OS Linux (dapat dilihat dalam Lampiran B) dengan GPU GeForce Nvidia RTX 3070 8GB. Karena tidak memadai, arsitektur diupayakan di kecilkan dengan cara Autocast[[2]](#footnote-2) yaitu hidden size tidak dikurangi namun beberapa dari weight akan dibuang untuk meringankan komputasi, cara ini tidak dapat memenuhi kebutuhan dari arsitektur. Kemudian hidden size dari Transformer dan FFN dikurangi dari nilai awal 782 dan 1028 menjadi rata 504 untuk tiap dimensi. GPU dapat menjalankan metode namun mendapatkan akurasi yang kurang bagus karena pemotongan hidden size yang sekitar 50%.

Dengan ini, diputuskan membutuhkan resource yang lebih besar, sehinga penelitian menggunakan Google Colab Pro yang menyediakan GPU Nvidia Tesla T4 16GB. Tetapi GPU tersebut masih belum bisa memenuhi Transformers, sehingga penelitian dilanjutkan dengan Transformers tidak di training ulang (freeze), namun Autocast tidak digunakan, dan hidden size disesuaikan kembali 782 dan 1028. Ada pun beberapa halangan dalam Google Colab Pro karena beberapa fitur yang tidak dapat diakses. Idle runtime diberikan secara random, sehingga notebook tidak dapat ditinggalkan sepenuhnya, harus dipantau agar tidak terputus koneksi secara mendadak (unexpected error, dapat dilihat pada lampiran B). Ditengah pelaksanaan uji coba juga terjadi runtime disconnected, yang berarti Google Colab tidak mengijinkan komputasi berat tanpa aktifitas yang interaktif (terlalu idle). Karena penggunaan GPU akun tugas akhir ini terlalu berat dan tidak interaktif, maka Google Colab memberikan timeout selama 5 jam[[3]](#footnote-3).

1. Yongliang Shen, Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), (Agustus : 2021). [↑](#footnote-ref-1)
2. Automatic Mixed Precision Package, (https://pytorch.org/docs/stable/amp.html). [↑](#footnote-ref-2)
3. Google Colab FAQ, https://research.google.com/colaboratory/faq.html [↑](#footnote-ref-3)