# Laporan Tugas Akhir Natural Language Processing
# Pronoun Anaphoric Reference


Kelompok PLN

| | |
|---|---|
| Bimo Iman Smartadi | 1706039780 |
| Dafa Ramadansyah | 1706039370 |
| Helmi Alfarel | 1806235896 |

**Abstrak**
Pronominal Anaphoric Reference merupakan suatu tantangan bagi mesin untuk mengetahui hubungan antara kata yang mengacu dan kata yang diacu. Pada laporan ini, kami menggunakan metode pendekatan *rule-based* dan juga *machine learning*.

## I. Latar Belakang

Pada teks berbahasa Indonesia banyak kata maupun kalimat yang interpretasinya bergantung kepada sebuah ekspresi yang terletak sebelum ataupun setelahnya. Hal tersebut disebut dengan anafora. Anafora dalam Bahasa Indonesia memiliki perbedaan yang lumayan banyak dengan Bahasa Inggris terutama pada pengutaraan kepemilikan dengan penggunaan “nya”.

Penggunaan anafora pada setiap bahasa dapat mudah dimengerti oleh manusia. Namun untuk mengolah sebuah teks agar pendeteksian anafora dapat dilakukan merupakan hal yang menarik untuk di telusuri. Dalam bidang pengolahan bahasa manusia, pendeteksian anafora merupakan sebuah tantangan yang apabila dapat diselesaikan manfaatnya dapat banyak digunakan khususnya untuk komputer agar dapat mengolah teks dengan lebih baik dan akurat.

## II. Studi Literatur

Hal-hal yang terkait dengan pengerjaan tugas Pronominal Anaphoric Reference ini adalah sebagai berikut dengan referensi dari sumber [1] :

1. Pronominal Anaphoric Reference
2. SACR
3. Flair
4. NLTK
5. Rule based
    a. Saliency Weighting
6. Machine Learning
    a. Logistic Regression
    b. Naive Bayes
    c. Deep Learning

## III. Metode

Dalam pengerjaan tugas *anaphora resolution,* digunakan dataset yang diberikan berupa 160 teks data yang sudah diberi label anaphora dan jenisnya dalam bentuk notasi SACR.

1. Preprocessing
   Dataset yang ada dilakukan pengambilan *mention* menggunakan struktur data *stack* dengan mencari *tag* terluar berupa kurung kurawal buka ({) kemudian mengambil karakter yang ada di dalam kurung kurawal tersebut sampai ditemukan kurung kurawal tutup (}) yang akan mengeluarkan isi dari *stack* sebagai penanda untuk berhenti

mengambil karakter di luar *mention*. Kemudian dilakukan penghitungan urutan kalimat *mention* yang diambil pada teks sebagai fitur yang akan digunakan kedepannya.

2. Ekstraksi Fitur
   Fitur yang kami pilih adalah sebagai berikut:

| **Fitur** | **Deskripsi** |
|---|---|
| Id | Id dari *mention* |
| sentence | Kata / kalimat yang membuat *mention* |
| mention | Bentuk *mention* asli |
| jenis | Jenis dari *mention* |
| ref | *List* dari referensi *mention* |
| sent_id | Urutan kalimat yang mengandung *mention* pada teks |
| pronoun_1 | Berisi 1 apabila *mention* merupakan kata ganti orang pertama, 0 jika tidak |
| pronoun_2 | Berisi 1 apabila *mention* merupakan kata ganti orang kedua, 0 jika tidak |
| pronoun_3 | Berisi 1 apabila *mention* merupakan kata ganti orang ketiga, 0 jika tidak |
| role | *Grammatical role mention* pada kalimat |
| text_id | Id dari teks yang mengandung *mention* |
| filename | Nama file teks yang mengandung *mention* |

   *Mention* yang sudah diambil kemudian dilakukan penyaringan dengan menggunakan *regex* untuk mengambil beberapa fitur diatas yaitu *Id, sentence, mention, jenis, text_id, dan filename*. Untuk *sent_id* didapat dari tahap preprocessing. Untuk fitur *role* dilakukan *dependency parsing* dilanjutkan dengan POS Tagging menggunakan *library flair*.

3. Pengolahan Fitur
   Agar model *machine learning* dapat dilakukan *training*, fitur - fitur yang sudah diekstraksi harus dilakukan pengolahan lebih lanjut ke bentuk *value* yang numerik. Untuk model *rule based*, tabel fitur langsung diproses dengan modelnya.

   Bentuk tabel fitur baru adalah sebagai berikut:

| Fitur | Deskripsi |
|---|---|
| Idm | Id dari *mention* yang terkait |
| Ida | Id dari *mention* sebagai kandidat referensi dari Idm, diambil dengan mencari kandidat dengan *window* $\pm 10$ |
| distance | Jarak antar *mention* dengan kandidat referensinya |
| samesent | Bernilai 1 apabila *mention* dan kandidat referensi *mention* berada dalam 1 kalimat, 0 jika tidak |
| pronoun_1 | Bernilai 1 apabila *mention* merupakan kata ganti orang pertama, 0 jika tidak |
| pronoun_2 | Bernilai 1 apabila *mention* merupakan kata ganti orang kedua, 0 jika tidak |
| pronoun_3 | Bernilai 1 apabila *mention* merupakan kata ganti orang ketiga, 0 jika tidak |
| word_cnt | Jumlah dari kata/kalimat kandidat *mention* |
| capital | Bernilai 1 apabila awal huruf pada setiap kata di kandidat *mention* adalah huruf kapital, 0 jika tidak |
| subj_pro | Bernilai 1 jika *mention* merupakan subjek, 0 jika tidak |
| obj_pro | Bernilai 1 jika *mention* merupakan objek, 0 jika tidak |
| poss_pro | Bernilai 1 jika *mention* merupakan kepemilikan, 0 jika tidak |
| subj_ant | Bernilai 1 jika kandidat *mention* merupakan subjek, 0 jika tidak |
| obj_ant | Bernilai 1 jika kandidat *mention* merupakan objek, 0 jika tidak |
| poss_ant | Bernilai 1 jika kandidat *mention* merupakan kepemilikan, 0 jika tidak |

4. Model

   Model yang kami gunakan terdiri atas 2 tipe yaitu model *rule based* dan model machine learning. Berikut merupakan model - model yang kami gunakan pada penelitian kali ini dari kedua kategori,

   Rule-based:

   Untuk kategori *rule-based* digunakan metode *Salience weighting* dengan bobot seperti berikut:

| Karakteristik | Salience Weight |
|---|---|
| Kandidat adalah intra-sentence | +100 |
| Kandidat berupa subjek | +80 |
| Kandidat menunjukkan kepemilikan | +65 |
| Kandidat berupa objek | +45 |
| Kandidat merupakan nama orang | +60 |
| Kandidat berada di setelah kata ganti | /2 |

Machine Learning:
Untuk model Machine Learning, beberapa model yang digunakan adalah sebagai berikut:

a. Logistic Regression
   Dilakukan percobaan dengan *tolerance* 0.0001 dan iterasi maksimal 500
b. Gaussian Naive Bayes
   Dilakukan percobaan dengan *smoothing* 1e-09
c. Neural Network

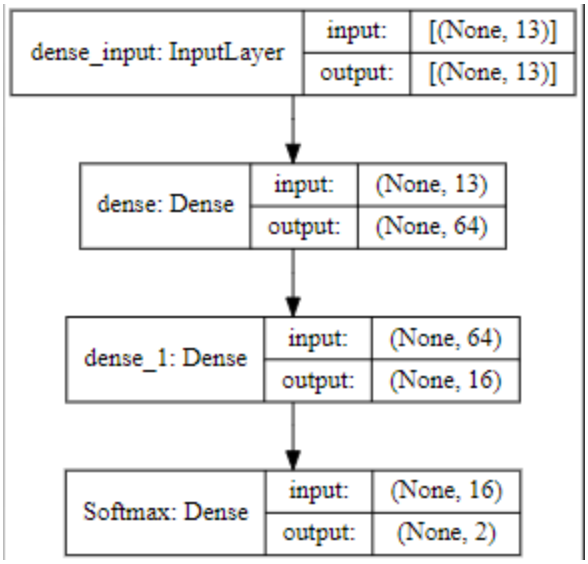
## IV. Eksperimen

**Logistic Regression & Gaussian Naive Bayes**
Kedua model *machine learning* tersebut dipilih dengan alasan sebagai tolak ukur model *machine learning* klasik. Data latih ditemukan ketidakseimbangan antar kelas positif dan negatifnya sekitar 4:1, maka dilakukan *balancing* menggunakan *downscaling* dan *upscaling*. Kedua model dicoba menggunakan masing - masing versi dari data latih. Logistic Regression dicoba dengan menggunakan *tolerance* 0.0001 dan iterasi maksimum 500 sedangkan Gaussian Naive Bayes menggunakan variabel *smoothing* 1e-09.

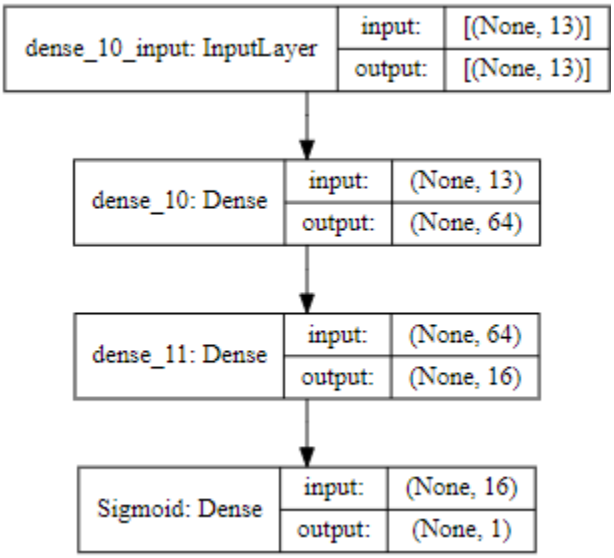
**Neural Network**

Neural network terdiri dari 1 input layer, 2 hidden layer, dan 1 output layer. Masing-masing hidden layer tidak memiliki activation function.

- Softmax sebagai activation output layer
  Struktur model




- Sigmoid sebagai activation output layer
  Struktur model




**Rule-based**

*Rule-based* dengan pendekatan metode *saliency weighting* dilakukan dengan membuat pasangan antara ID *mention* dengan *antecedent* lalu diberikan skor sesuai yang sudah didefinisikan pada penjelasan model terkait *rule-based*.

Pembuatan pasangan antara ID *mention* (idm) dan ID *antecedent* (ida) dilakukan dengan membuat data tes menjadi bentuk *Dataframe* lalu diiterasikan untuk setiap nama *file* yang berbeda dipasangkan dengan memasangkan idm dengan 12 token ida di belakang idm dan 8 token ida di depan idm. Angka ini dipilih dengan melakukan

percobaan pada text “Alien in the Attic” di mana ada token yang jarak antara idm dan ida yang sesuai menurut penulis adalah 12 token.

Lalu pasangan tersebut diberikan skor dengan memeriksa jenis dari idm dan ida serta *role* dari masing-masing ID dan diberikan skor berdasarkan acuan tersebut. Dari pasangan tersebut, dipilih yang pasangannya memiliki skor terbesar untuk setiap idm dan apabila terdapat lebih dari satu pasangan dengan skor maksimum yang sama, akan dipilih yang terdekat dari idm.

**Neural Network + Rule Based**

Menggabungkan hasil prediksi rule based dengan neural network. Neural network menghasilkan probabilitas masing-masing *mention* menjadi referen ke *mention* lain, data tersebut digunakan untuk menambahkan hasil prediksi rule based. Jika *mention* yang diprediksi oleh rule based merupakan yang mempunyai probabilitas menjadi referen paling besar, maka biarkan. Jika dia merupakan yang kedua terbesar, maka tambahkan *mention* dengan probabilitas terbesar kedalam hasil prediksi.

## V. Hasil

| Model | Macro - Exact Match | Macro - Partial Match | Micro - Exact Match | Micro - Partial Match |
|---|---|---|---|---|
| Logistic Regression | 0.1845 | 0.1973 | 0.1818 | 0.1945 |
| Logistic Regression (balanced) | 0.1718 | 0.1855 | 0.1680 | 0.1820 |
| Gaussian NB | 0.1775 | 0.1900 | 0.1705 | 0.1832 |
| Gaussian NB (balanced) | 0.1629 | 0.1765 | 0.1583 | 0.1721 |
| Neural Network (Softmax) | 0.1259 | - | - | - |
| Neural Network (Sigmoid) | 0.1512 | - | - | - |
| Neural Network + Rule Based | 0.3812 | - | - | - |
| Rule Based | 0.4975 | 0.5256 | 0.4659 | 0.4946 |

## VI. Analisis dan Kendala

**Rule-based**

Berdasarkan tabel di atas Model *Rule-based* dengan metode *saliency weighting* memiliki skor 0.4975 pada Macro Precision Exact match. Lalu, setelah dibandingkan dengan *gold label* yang diberikan, beberapa analisa yang bisa diberikan adalah:

- Terdapat satu kasus yang diminta untuk menebak referensi dari suatu ID dimana ID tersebut merupakan sebuah “named-entity”. Pada pengerjaan untuk metode ini, idm yang memiliki jenis “named-entity” akan di-skip karena hanya akan mencari yang memiliki jenis “kata ganti”
- Kasus lain yang terjadi adalah banyak anotasi jenis pada teks yang pada dataset tes yang jenisnya kosong sehingga pada percobaan awal menjadi di-*skip*. Hal ini dapat diatasi dengan memberikan daftar kata ganti, dan mengecek apabila kata tersebut berada pada daftar tersebut, jenis diberikan nilai *default* yaitu “kata ganti”. Contoh kasus yang diselesaikan dengan melakukan ini adalah pada teks “Alien in the Attic” dimana sebelum ini dilakukan banyak calon idm yang terlewat karena di-skip.
- Kekurangan dari model ini adalah, model ini hanya bisa mengeluarkan 1 hasil ida untuk setiap pasangan idm. Jadi apabila pasangan idm memiliki lebih dari satu pasangan, model ini tidak bisa mendeteksi hal tersebut.
- Kemungkinan lain yang mengurangi skor adalah *antecedent* yang terlalu jauh dari id *mention* dan role yang salah pada ida yang membuat skor yang didapat pada pasangan tersebut.

**Machine Learning**

Ditemukan hasil model *machine learning* secara umum lebih kecil dibanding *rule-based* dengan yang terbaik yaitu skor 0.3812 untuk *macro exact match* dengan menggabungkan dengan *rule-based* dan yang murni *machine learning* hanya 0.1845 untuk *macro exact match*. Beberapa analisa yang dapat diberikan adalah sebagai berikut:

- Terdapat penamaan jenis *mention* yang kurang konsisten (contoh: harusnya {M12 jenis="kata-ganti"{M10 jenis="" dia}} tetapi {M12 jenis=""{M10 jenis="kata-ganti" dia}} ) sehingga menyebabkan adanya *mention* ataupun *reference* yang tidak masuk ke data latih.
- Diduga data latih yang ada kurang representatif dengan teks yang ada, hal tersebut selain karena alasan sebelumnya juga kami asumsi dikarenakan ketidakstabilannya kelas positif dan negatif namun ketika dilakukan *balancing* hasil malah turun dibanding dengan data latih

## VII. Penutup

Task Pronominal Anaphoric Reference ini diselesaikan dengan dua kategori *approach* yaitu *rule-based* dengan metode *saliency weighting* dan *machine learning* dengan metode *deep learning, logistic regression,* serta Gaussian NB. Skor tertinggi yang didapat pada kategori *rule-based* yaitu 0.4975 dan untuk kategori *machine learning* adalah hasil modifikasi antara *rule-based* dengan deep learning dengan skor 0.3812. Fitur yang berpengaruh pada kedua kategori adalah terkait *role* yang didapat dari POS Tagging dan juga terkait jenis yang diberikan pada data anotasi latih dan tes.

## Referensi


1. Anggraito, A. (2019). *Pronominal Anaphora Resolution Pada Teks Berbahasa Indonesia Menggunakan Pendekatan Machine Learning Dan Rule-Based*.
2. Oberle, B. (n.d.). *Coreference annotation tool (SACR)*. Bruno Oberle - Coreference annotation tool (SACR). https://boberle.com/projects/coreference-annotation-with-sacr/.