POS Tagging & Named-Entity Recognition

Indonesia AI

OBJECTIVE & OUTLINE

POS Tagging & Named-Entity Recognition

Objektif: Memahami konsep dari POS Tagging dan NER dalam NLP

Outline:

1. POS(Part-of-Speech) Tagging
2. Penerapan POS Tagging
3. NER
4. Penerapan NER

POS Tagging

Indonesia AI

POS TAGGING

Proses dalam pemrosesan bahasa alami untuk **mengidentifikasi kategori** atau jenis **kata** dalam sebuah teks

POS TAGGING

Manfaat POS Tagging

- Peningkatan pemahaman teks
- Ekstraksi informasi
- Peningkatan akurasi dalam tugas NLP

POS TAGGING

Macam POS Tag:

- Tertutup
  - kata depan, kata hubung, kata ganti
- Terbuka
  - kata kerja, kata benda

POS TAGGING

Kelas Tag:

- Dionysius Thrax
  - 8 tag untuk bahasa Yunani: *noun, verb, pronoun, preposition, adverb, conjunction, particle, artiCle*
- Penn Treebank
  - 45 tag: DT, IN, JJ, JJR, NN dsb.
- Brown Corpus
  - 87 tag

POS TAGGING

No.	Tag	Description	Example
1	ADV	Adverbs. Includes adverb, modal, and auxiliary verb	sangat, hanya, justru, boleh, harus, mesti
2	CC	Coordinating conjunction. Coordinating conjunction links two or more syntactically equivalent parts of a sentence. Coordinating conjunction can link independent clauses, phrases, or words.	dan, tetapi, atau
3	DT	Determiner/article. A grammatical unit which limits the potential referent of a noun phrase, whose basic role is to mark noun phrases as either definite or indefinite.	para, sang, si, ini, itu, nya
4	FW	Foreign word. Foreign word is a word which comes from foreign language and is not yet included in Indonesian dictionary	workshop, business, e-commerce
5	IN	Preposition. A preposition links word or phrase and constituent in front of that preposition and results prepositional phrase.	dalam, dengan, di, ke
6	JJ	Adjective. Adjectives are words which describe, modify, or specify some properties of the head noun of the phrase	bersih, panjang, jauh, marah
7	NEG	Negation	tidak, belum, jangan

POS TAGGING

8	NN	Noun. Nouns are words which refer to human, animal, thing, concept, or understanding	meja, kursi, monyet, perkumpulan
9	NNP	Proper Noun. Proper noun is a specific name of a person, thing, place, event, etc.	Indonesia, Jakarta, Piala Dunia, Idul Fitri, Jokowi
10	NUM	Number. Includes cardinal and ordinal number	9876, 2019, 0,5, empat
11	PR	Pronoun. Includes personal pronoun and demonstrative pronoun	saya, kami, kita, kalian, ini, itu, nya, yang
12	RP	Particle. Particle which confirms interrogative, imperative, or declarative sentences	pun, lah, kah
13	SC	Subordinating Conjunction. Subordinating conjunction links two or more clauses and one of the clauses is a subordinate clause.	sejak, jika, seandainya, dengan, bahwa
14	SYM	Symbols and Punctuations	+,%,@

POS TAGGING

15	UH	Interjection. Interjection expresses feeling or state of mind and has no relation with other words syntactically.	ayo, nah, ah
16	VB	Verb. Includes transitive verbs, intransitive verbs, active verbs, passive verbs, and copulas.	tertidur, bekerja, membaca
17	ADJP	Adjective Phrase. A group of words headed by an adjective that describes a noun or a pronoun	sangat tinggi
18	DP	Date Phrase. Date written with whitespaces	1 Januari 2020
19	NP	Noun Phrase. A phrase that has a noun (or indefinite pronoun) as its head	Jakarta Pusat, Lionel Messi
20	NUMP	Number Phrase.	10 juta
21	VP	Verb Phrase. A syntactic unit composed of at least one verb and its dependents	tidak makan

POS TAGGING

Irina mengatakan, pendidikan usia dini dinilai ampuh menanamkan nilai positif kepada anak.

POS TAGGING

kata	Claudia	mengatakan	,	pendidikan	usia	dini	dinilai	ampuh	menanamkan
pos tag	NNP	VB	SYM	NN	NN	JJ	VB	JJ	VB
kata	nilai	positif	kepada	anak	.				
pos tag	NN	JJ	IN	NN	SYM				

POS TAGGING

Masalah pada POS Tagging

- Polisemi(Ambigu)
- OOV

Any question guys ~

Penerapan POS Tagging

Indonesia AI

PENERAPAN POS TAGGING

Penerapan POS tagging

- Rule based tagger
- Statistical/stochastic tagger

PENERAPAN POS TAGGING

Rule based tagger

- Top down
  - Pendefinisian aturan yang biasa digunakan manusia

PENERAPAN POS TAGGING

Menggunakan **aturan** atau pola yang didesain untuk **mengklasifikasikan kata** ke dalam **kategori POS tag** tertentu. Dalam pendekatan ini, aturan-aturan tersebut diterapkan secara bertahap dari **atas ke bawah** pada **struktur kalimat**.

PENERAPAN POS TAGGING

Aturan tingkat atas/top:

(1) Jika kata memiliki imbuhan [me-, mem-, ber-, per-, ter-, di, -kan, ter-kan, dan di-i], maka kata tersebut kemungkinan besar merupakan kata kerja (verb).

PENERAPAN POS TAGGING

Aturan tingkat atas/top:

(2) Jika kata diawali dengan huruf besar, maka kata tersebut kemungkinan besar merupakan kata benda (noun). (Jika bahasa inggris bisa menggunakan “The”)

PENERAPAN POS TAGGING

Aturan tingkat atas/top:

(3) Jika kata berupa angka, maka kata tersebut kemungkinan besar merupakan kata bilangan (number)

PENERAPAN POS TAGGING

Aturan tingkat bawah:

(1)Jika kata berada di dalam kalimat dan diapit oleh kata benda dan kata adverb, maka kata tersebut kemungkinan besar merupakan kata pronoun

PENERAPAN POS TAGGING

Aturan tingkat bawah:

(2) Jika kata berada di dalam kalimat dan diapit oleh kata kerja (verb), maka kata tersebut kemungkinan besar merupakan kata adverb

PENERAPAN POS TAGGING

Kuda itu sedang berlari.

PENERAPAN POS TAGGING

Penerapan rule based:

Kata pertama "Kuda" diawali dengan huruf besar.

Menggunakan aturan tingkat atas, kata "Kuda" kemungkinan besar merupakan kata benda (noun).

PENERAPAN POS TAGGING

Penerapan rule based:

Kata kedua "itu" diapit oleh kata "Kuda" dan "sedang".

Menggunakan aturan tingkat bawah, kata "itu" kemungkinan besar merupakan kata pronoun.

PENERAPAN POS TAGGING

Penerapan rule based:

Kata ketiga "sedang" diapit oleh kata "berlari".

Menggunakan aturan tingkat bawah, kata "sedang" kemungkinan besar merupakan kata kerja adverb.

PENERAPAN POS TAGGING

Penerapan rule based:

Kata keempat "berlari" diawali dengan imbuhan ber-.

Menggunakan aturan tingkat atas, kata “berlari” kemungkinan adalah kata kerja(verb)

PENERAPAN POS TAGGING

- "Kuda" diberi POS tag NN (noun).
- "itu" diberi POS tag PR (pronoun).
- "sedang" diberi POS tag ADV (adverb).
- "berlari" diberi POS tag VB (ver).

PENERAPAN POS TAGGING

Statistical/Stochastic Tagger

- Buttom up
  - Menggunakan corpus sebagai data latih di mana aturan ditetapkan secara otomatis

PENERAPAN POS TAGGING

Metode/algoritma yang populer digunakan

- HMM:hasil pelatihan berupa nilai probabilitas
- Decision tree/ Random forest: hasil pelatihan berupa pohon keputusan
- Neural network: hasil pelatihan berupa fungsi pembeda
  -

Any question guys ~

NER

Indonesia AI

NER

NER (Named Entity Recognition) adalah salah satu fitur NLP yang bertujuan untuk **mengenali** dan **mengklasifikasikan** nama **entitas** dalam teks menjadi kategori tertentu seperti orang, tempat, organisasi, dsb.

NER

Jenis entitas yang biasa digunakan:

- PER (Person)
- NOR (Nationalities or religious or political groups)
- FAC (Facility)
- ORG (Organization)
- GPE (Geo-Political Entity)
- LOC (Location)
- PRO (Product)
- EVT (Event)

NER

[Jokowi] siang ini tiba di [Denpasar].Istri [Presiden Jokowi], [Iriana] turut mendampinginya.

NER

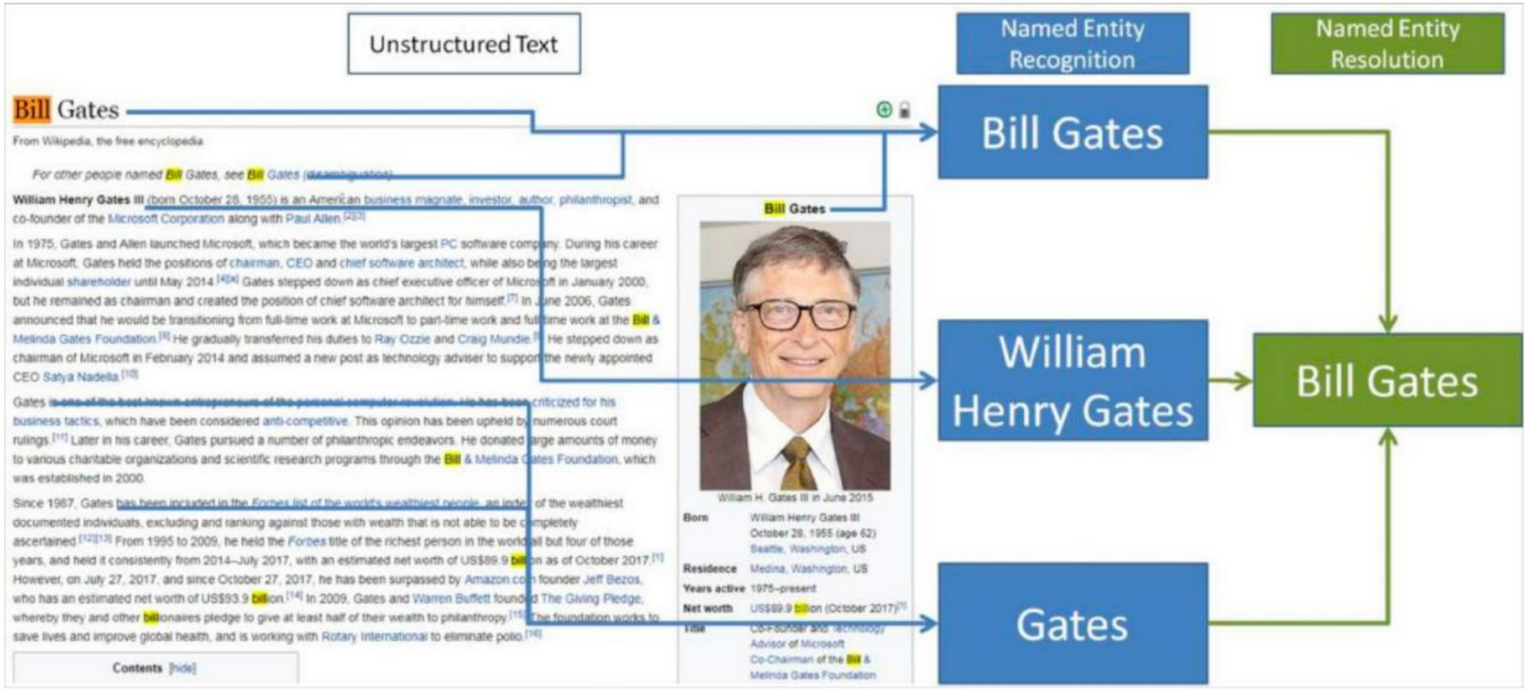
[**Jokowi**] siang ini tiba di [**Denpasar**].Istri [**Presiden Jokowi**], [**Iriana**] turut mendampinginya.

[**PER**] [**LOC**]

NER

Manfaat NER dalam tugas NLP

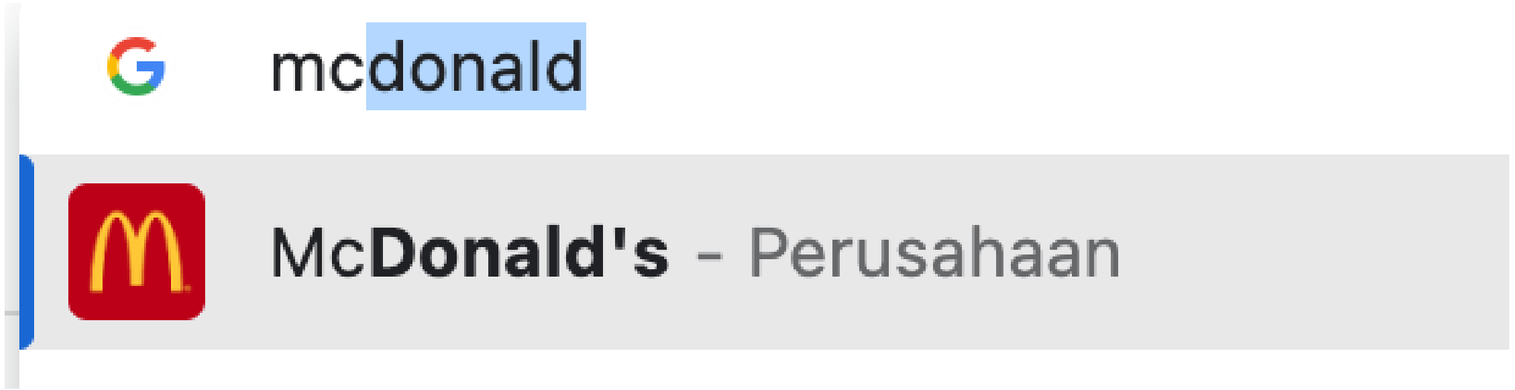
1. Entity Linking



NER

Manfaat NER dalam tugas NLP

2. Web query construction



NER

Masalah pada NER

- Polisemi(Ambigu)

Any question guys ~

Penerapan NER

Indonesia AI

PENERAPAN NER

Penerapan NER

- Rule based
- Model-based

PENERAPAN NER

Sistem **rule-based** pada NER akan **efektif** untuk **kelas entitas tertentu**. Biasanya menggunakan sistem leksikon yang mencantumkan nama, organisasi, lokasi, dll. Rule yang digunakan juga dapat dibuat menggunakan regex atau pencocokan pola lainnya.

PENERAPAN NER

Jalan <nama><nomor> <- alamat

<title><name> <- nama orang

PENERAPAN NER

Jalan <nama><nomor> <- alamat
Kejadian itu terjadi di [Jalan Merdeka III]

<title><name> <- nama orang
Pasien tersebut ditangani oleh [dr. Adli]

PENERAPAN NER

Model-based

- Machine Learning-Based
  - SVM
  - CRF
  - HMM
- Deep Learning-Base
  - RNN
  - LSTM
- Pre-trained Models-Based
  - BERT

Any question guys ~

Terima Kasih!