

Part of Speech Tagging

Teknik Informatika Unimma

Part of Speech

Part of Speech merupakan klasifikasi dari kata-kata yang dikategorikan dari peran dan fungsinya dalam struktur kalimat sebuah bahasa atau kategori sintaktik.

Sebuah kategori yang diberikan ke sebuah kata sesuai fungsi sintaktiknya.

Contoh Part of Speech

- *noun* (kata benda) = people, animal, things
- *pronoun* (kata ganti) = I, you, he, she, some
- *verb* (kata kerja) = run, study, have, do, like, work, sing, can
- *adjective* (kata sifat) = soft, diligently, good, big, red, well, interesting
- *adverb* (kata keterangan) = quickly, silently, well, badly, very, really
- *preposition* (kata depan) = to, at, after, on, but
- *conjunction* (kata hubung) = and, but, when
- *interjection* (kata seru) = oh!, ouch!, hi!, well

Kategori POS

- **Closed class:** relatively fixed set
 - Composed of a small, fixed set of grammatical function words for a given language, e.g. **Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions**
 - Prepositions: **of, in, by, ...**
 - Auxiliaries: **may, can, will, had, been, ...**
 - Pronouns: **I, you, she, mine, his, them, ...**
 - Usually **function words** (short common words which play a role in grammar)
- **Open class:** productive
 - Open class categories have large number of words and new ones are easily invented

Open vs Close?

- Bakso
- Pedoman
- Menyakitkan
- Tentang
- Menyanyi
- Bawah
- Melalui

POS TAGGING

Part of Speech (POS) Tagging merupakan salah satu implementasi dari Natural Language Processing. POS Tagging merupakan proses pemberian label kelas kata pada kalimat.

- **Tagging** adalah proses menentukan kelas kata / part-of-speech tag untuk setiap kata dalam sebuah teks.
- Input: rangkaian kata + tagset.
- Output: tag yang paling tepat untuk setiap kata.
- \approx tokenization untuk bahasa pemrograman (di mana bedanya?) \rightarrow POS tagging bisa ambiguous!
- Contoh:
Book that flight .
VB DT NN .
Book bisa juga Nn (malahan lebih sering?).

POS TAGGING

- POS Tagging atau *part-of-speech tagging* merupakan salah satu tahapan yang dilakukan dalam proses pengolahan teks bahasa (NLP), yaitu memberikan POS Tag pada setiap kata dalam satu atau lebih kalimat dengan penanda part-of-speech.

Contoh: The/**AT** ball/**NN** is/**VB** green/**JJ**

- POS Tagger merupakan sebuah aplikasi yang mampu melakukan proses anotasi part-of-speech tag untuk setiap kata di dalam dokumen secara otomatis.
- Akurasi dari POS Tagger sudah mencapai lebih dari 96% (untuk Bahasa Inggris)

Kenapa menggunakan Pos tagging

- **speech synthesis**
 - Bagaimana cara pembacaan
 - DIScount or disCOUNT, CONten or conTENT
- **Parshing**

Mengetahui apa saja kata-kata yang ada pada kalimat
- **Information extraction**

Menemukan nama, lokasi dll
- **Machine translation**

-I like you , I like you , Where like

Contoh Tagset Bahasa Inggris

- Brown corpus tagset: 87 tag (Francis and Kucera, 1982)
- Penn Treebank tagset: 45 tag (Marcus et al., 1993)
- C5 CLAWS BNC tagset: 61 tag (Garside et al., 1997)
- C7 tagset: 146 tag (Leech et al., 1994)

Perbedaannya?

- Penn Treebank menggabungkan beberapa tag menjadi satu, karena ada informasi parse tree.
- Tergantung kebutuhan!

Pentreebank tagset

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	les
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables

Pentreebank tagset

Contoh kalimat yang di-tag:

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN
of/IN other/JJ topics/NNS ./.

Universal Part of Speech tagset

Tag	Meaning	English Examples
ADJ	adjective	<i>new, good, high, special, big, local</i>
ADP	adposition	<i>on, of, at, with, by, into, under</i>
ADV	adverb	<i>really, already, still, early, now</i>
CONJ	conjunction	<i>and, or, but, if, while, although</i>
DET	determiner, article	<i>the, a, some, most, every, no, which</i>
NOUN	noun	<i>year, home, costs, time, Africa</i>
NUM	numeral	<i>twenty-four, fourth, 1991, 14:24</i>
PRT	particle	<i>at, on, out, over per, that, up, with</i>
PRON	pronoun	<i>he, their, her, its, my, I, us</i>
VERB	verb	<i>is, say, told, given, playing, would</i>
.	punctuation marks	<i>. , ; !</i>
X	other	<i>ersatz, esprit, dunno, gr8, univeristy</i>

TABLE 11. Final proposed Indonesian tagset

Tag	Description	Example
CC	Coordinating conjunction	dan, tetapi, atau
CD	Cardinal number	dua, juta, enam, 7916, sepertiga, 0,025, 0,525, banyak, kedua, ribuan, 2007, 25
OD	Ordinal number	ketiga, ke-4, pertama
DT	Determiner / article	Para, Sang, Si
FW	Foreign word	climate change, terms and conditions
IN	Preposition	dalam, dengan, di, ke, oleh, pada, untuk
JJ	Adjective	bersih, panjang, hitam, lama, jauh, marah, suram, nasional, bulat
MD	Modal and auxiliary verb	boleh, harus, sudah, mesti, perlu
NEG	Negation	tidak, belum, jangan
NN	Noun	monyet, bawah, sekarang, rupiah
NNP	Proper noun	Boediono, Laut Jawa, Indonesia, India, Malaysia, Bank Mandiri, BBKP, Januari, Senin, Idul Fitri, Piala Dunia, Liga Primer, Lord of the Rings: The Return of the King
NND	Classifier, partitive, and measurement noun	orang, ton, helai, lembar
PR	Demonstrative pronoun	ini, itu, sini, situ
PRP	Personal pronoun	saya, kami, kita, kamu, kalian, dia, mereka
RB	Adverb	sangat, hanya, justru, niscaya, segera
RP	Particle	pun, -lah, -kah
SC	Subordinating conjunction	sejak, jika, seandainya, supaya, meski, seolah-olah, sebab, maka, tanpa, dengan, bahwa, yang, lebih ... daripada ..., semoga
SYM	Symbol	IDR, +, %, @
UH	Interjection	brengsek, oh, ooh, aduh, ayo, mari, hai
VB	Verb	merancang, mengatur, pergi, bekerja, tertidur
WH	Question	siapa, apa, mana, kenapa, kapan, di mana, bagaimana, berapa
X	Unknown	statemen
Z	Punctuation	"...", "?", "

Chalange

- **Ambiguity**

- glass of water (noun) vs water(verb) the plans
- lie (verb) down vs tell a lie (noun)
- wind(verb) down vs a mighty wind (noun)

- **Sparse data**

- Data yang banyak kita belum pernah menemui sebuah kata-kata sebelumnya
- belum adanya kata yang memiliki tag

Pemmasalahan dalam POS TAGGING

- Sebetulnya, seberapa sulitkah POS tagging ini?
- DeRose (1988): Hanya 11.5% kata Inggris unik (**type**) dari Brown corpus adalah rancu.
- Namun demikian, 40% dari (**token**) di Brown corpus rancu!
- Type vs. token = class vs. instance.
“the boy saw the man”: 5 token, 4 type
“pria itu berdiri di tengah pria-pria lainnya”: 7 token, 7 type
- (Perhatikan ini adalah beda tag, bukan makna/sense.)

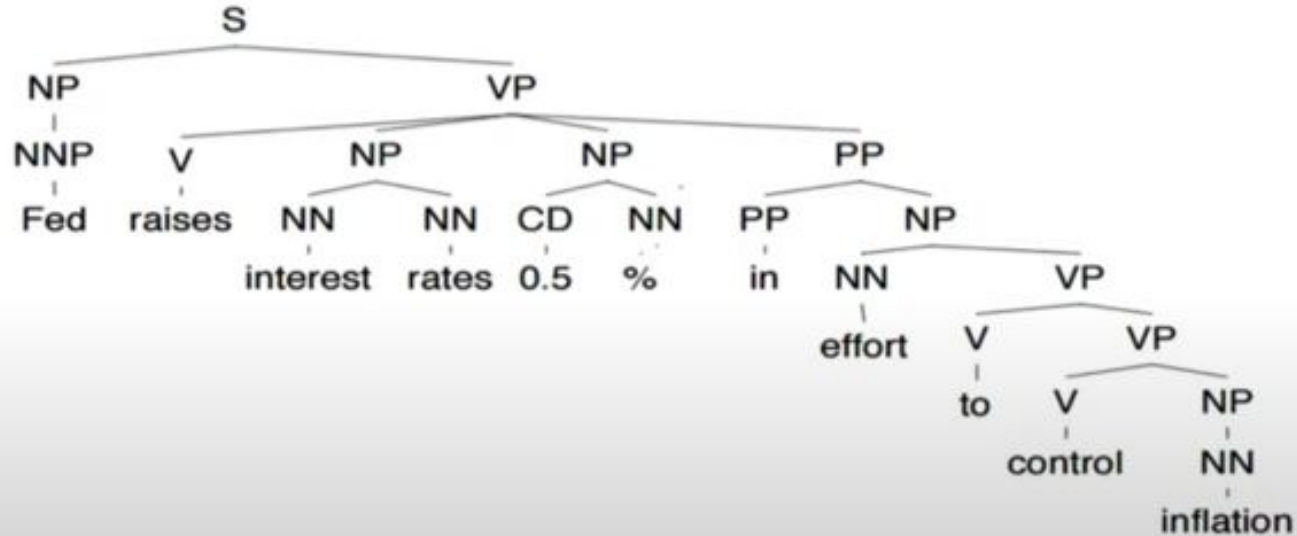
Unambiguous (1 tag)	35340
Ambiguous (2-7 tags)	4100
2 tags	3760
3 tags	264
4 tags	61
5 tags	12
6 tags	2
7 tags	1 (“still”)

Teknik POS TAGGING

- **Rule-based tagging.** Cara top-down – konsultasi ahli linguistik; definisikan aturan-aturan yang biasa digunakan manusia.
- **Learning-based tagging.** Cara bottom-up – gunakan corpus sebagai training data, seperti Penn Treebank
 - **Stochastic tagger.** untuk menentukan secara probabilistik tag yang terbaik untuk sebuah kata (dalam sebuah konteks): HMM, Maximum Entropy Markov Model, Conditional Random Field (CRF)
 - **Transformation-based tagger.** Semacam gabungan teknik di atas. Tetap belajar dari corpus, tapi knowledge yang dipelajari dinyatakan sebagai rule.

Pada umumnya, pendekatan berbasis pembelajaran terbukti lebih efektif secara keseluruhan, dengan mempertimbangkan jumlah total keahlian dan upaya manusia yang terlibat

Pemanfaatan informasi POS Tag : Parsing



Pemodelan statistika

- Pada intinya, kita ingin memilih tag yang memaksimalkan rumus berikut:
 $P(kata|tag) \times P(tag|n \text{ tag sebelumnya})$
- Sebagai aproksimasi, sebuah *bigram* tagger memilih tag untuk kata ke- i (t_i) berdasarkan tag sebelumnya (t_{i-1}) dan kata ke- i tersebut (w_i)

$$t_i = \operatorname{argmax}_j P(t_j|t_{i-1}, w_i)$$

Melalui beberapa asumsi Markovian, diperoleh:

$$t_i = \operatorname{argmax}_j P(t_j|t_{i-1}) \times P(w_i|t_j)$$

(Pada kenyataannya, kita ingin melakukan *tagging* pada seluruh kalimat sekaligus, bukan hanya satu kata!)

Rule base

I have to go there.
verb

I had a go at it.
noun

- If the previous word is “to”, then it’s a verb.
- If the previous word is “a”, then it’s a noun.
- If the next word is ...

:

➡ Writing rules manually is impossible

Learning base

The involvement of ion channels in B and T lymphocyte activation is
DT NN IN NN NNS IN NN CC NN NN NN VBZ
supported by many reports of changes in ion fluxes and membrane
VBN IN JJ NNS IN NNS IN NN NNS CC NN
.....
.....

Unseen text

We demonstrate
that ...



Machine Learning
Algorithm



We demonstrate
PRP VBP
that ...
IN



training