

## Pemrosesan Bahasa Alami

# Pengantar Pemrosesan Bahasa Alami

Putra Pandu Adikara, S.Kom, M.Kom  
Budi Darma Setiawan, S.Kom, M.Cs  
Sigit Adinugroho, S.Kom, M.Sc  
**Universitas Brawijaya**

# Outline

---

- Apakah itu Pemrosesan Bahasa Alami?
- Topik-Topik dalam Pemrosesan Bahasa Alami
- Sejarah Pemrosesan Bahasa Alami

# Pengertian Pemrosesan Bahasa Alami

---

- Bahasa alami adalah bahasa yang digunakan oleh manusia secara lisan maupun tulisan
- Tapi, bahasa manusia tidak dimengerti oleh komputer
- Oleh karena itu, agar komputer mengerti bahasa manusia maka kita harus memberikan pengetahuan kepada komputer tentang bahasa manusia
- Sistem Pemrosesan Bahasa Alami/*Natural Language Processing* (NLP) adalah perangkat lunak yang mengolah (naskah) bahasa manusia

# Pengertian Pemrosesan Bahasa Alami


---

- Komponen dalam sistem NLP
  - Naskah: tulisan (*written*) vs. lisan (*speech*)
  - Pengolahan: memahami (*understanding*), menghasilkan (*generation*), keduanya (*dialogue, question and answer/QA*)
  - Bahasa manusia: Indonesia, Inggris, Jepang, Perancis, Jawa, dll.
  - Domain: undang-undang, iklan, micro text (SMS, *tweet*), dll

# Contoh Sistem dan Aplikasi NLP

- Question Answering

- Wolfram Alpha, Ask.com → knowledge engine, answer dan search engine untuk mencari jawaban/situs relevan dari pertanyaan
- Intelligence Personal Assistant: Alexa, Assistant & Now, Cortana, Siri



The screenshot shows the Wolfram Alpha interface. At the top, the search bar contains the text "first president of indonesia". Below the search bar, the input is interpreted as "Indonesia President 1<sup>st</sup>". The results section displays "Sukarno (from 17-08-1945 to 12-03-1967)". Below this, a table provides basic information about Sukarno's presidency.

WolframAlpha<sup>®</sup> computational knowledge engine.

first president of indonesia

Web Apps Examples Random

Input interpretation:

Indonesia President 1<sup>st</sup>

Enlarge Data Customize PlainText Interactive

result:

Sukarno (from 17-08-1945 to 12-03-1967)

Basic information:

official position	President
country	Indonesia
start date	17-08-1945 (71 years 6 months 2 days ago)
end date	12-03-1967 (49 years 11 months 7 days ago)
duration of leadership	21 years 6 months 26 days

# Contoh Sistem dan Aplikasi NLP

- Ekstraksi Informasi:

Penguji Skripsi

Yth. Bapak/Ibu John Doe

**Event:** Penguji Skripsi  
**Name:** Andi  
**Subject:** Sistem temu kembali untuk xxx  
**Date:** Jan 27, 2017  
**Start:** 08:45  
**End:** 10:00  
**Where:** FILKOM UB / R. C1.4

Sesuai dengan penugasan dari Ketua Program Studi maka Bapak/Ibu dimohon hadir sebagai penguji pada sidang skripsi berikut:

Nama : Andi Judul : Sistem temu kembali untuk xxx Tgl/Jam  
Ujian : Jan 27, 2017 / 08:45 s/d 10:00 Ruang : FILKOM UB /  
R. C1.4

Create new Calendar entry

# Contoh Sistem dan Aplikasi NLP

---

- Machine Translation
  - Google Translate ([translate.google.com](https://translate.google.com)) adalah sistem/mesin penerjemah yang dapat menerjemahkan antar bahasa
- Speech Recognition
  - Sphinx ([cmusphinx.sf.net](http://cmusphinx.sf.net)) adalah aplikasi yang dapat mengubah sinyal lisan (speech) menjadi tulisan (text) atau *speech-to-text* dan *text-to-speech* (TTS)
- Spelling & grammar checker/correction
  - Microsoft Word dapat mendeteksi kesalahan penulisan dan memberikan usulan perbaikan

# Contoh Sistem dan Aplikasi NLP

- *Query* basis data dalam bahasa manusia
- Peringkasan dokumen: menghasilkan ringkasan dokumen (abstrak)
- Intelligent Tutoring System: berdialog dengan siswa
- Automated Essay Scoring/Grading: aplikasi untuk menilai suatu essay
- Plagiarism checker, dan masih banyak lagi

Keberhasilan teknik ditentukan oleh pemahaman tentang **bahasa** itu sendiri dan **teknik mengolah bahasa**.



# Contoh Sistem dan Aplikasi NLP

banyak diselesaikan

## Spam detection

Let's go to Agra!



Buy V1AGRA ...



## Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

## Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Berkembang dengan baik

## Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



## Coreference resolution

Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



## Parsing

I can see Alcatraz from the window!

## Machine translation (MT)

第13届上海国际电影节开幕...



The 13<sup>th</sup> Shanghai International Film Festival...

## Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

masih sulit tapi berkembang

## Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

## Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

## Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

## Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?



- Sumber: Dan Jurafsky, NLP

# Mengapa belajar NLP?

---

- Agar mesin dapat mengolah bahasa alami dengan baik.
- Bahasa menyatakan pengetahuan manusia (*knowledge*). Jika mesin bisa paham, maka bisa mengolah dan melakukan banyak hal

# Bahasa Manusia

- Kita harus mempelajari dan memahami bahasa manusia terlebih dahulu.
- Bidang-bidang ini membagi bahasa dalam tingkatan/detail representasinya:

## Fonetik (Phonetics)

Mengubah sinyal suara menjadi fonem dan sebaliknya (**sounds** → **phonem or words**)

## Morfologi (Morphology)

Bentuk dan makna kata (**morphemes** → **words**): *ajar, belajar, mengajari, diajari*

# Bahasa Manusia

## Sintaksis (Syntax)

Membentuk urutan kata dan perubahan kata menjadi kalimat yang sah bahkan baik (**word sequence** → **sentence structure**)

John loves Jane.                      vs. John love Jane.                      → sah

Ani merangkai bunga.                vs. Ani dirangkai bunga. → ???

## Semantik (Semantics)

Memahami makna kata dari teks/kalimat (**sentence structure + word meaning** → **sentence meaning**)

- Mis. “hapus semua file” → `del *.*`

## Pragmatik (Pragmatics)

Memahami interpretasi kata di dalam konteks (wacana, domain, dll) (**sentence meaning + context** → **deeper meaning**)

- Mis. “Pindahkan file-file *ini* ke dalam folder yang *tadi*.”

# Bahasa Manusia

## Discourse & World Knowledge

Memahami arti kata secara umum dan khusus dalam kalimat dan antar kalimat sebelum dan sesudahnya/percakapan. (**connecting sentence + background knowledge → utterances**).

## Semiotik (Semiotics)

Memahami makna suatu tanda, simbol termasuk analogi, metafora, dan simbolisme

# Teknik Mengolah Bahasa

---

- **Pendekatan Linguist (Top-down)**
  - Mengimplementasikan algoritma dan struktur data berdasarkan teori dan model linguistik
- **Pendekatan Empiricist (Bottom-up)**
  - Menggunakan model “black-box” berdasarkan statistik atau *machine learning*
- Kedua pendekatan membutuhkan sumber informasi tentang bahasa:
  - *linguistic resources*, mis: kamus (**lexicon**), aturan tata bahasa (**grammar**), kumpulan dokumen (**corpus**)

# Contoh pengolahan

- Menginterpretasikan sinyal lisan:
  1. “I scream is delicious.”
  2. “Ice cream is delicious.”
- Dapat dilakukan dengan dua pendekatan sebelumnya:

## Model linguistik

Kalimat (1) dinyatakan tidak valid, sedangkan kalimat (2) valid

## Model empiris

“Ice cream is” lebih sering muncul/dijumpai daripada “I scream is” dalam koleksi dokumen

# Mengapa mempelajari & memodelkan bahasa itu sulit?

- Kerancuan (*ambiguity*) pada banyak tingkat:
  - “Bisa ular bisa mematikan.” → homofon
  - “Ani beli apel sebelum apel pagi.” → homograf
  - “Anto makan mie dengan sumpit. Budi makan bakso dengan Ani.”
  - “The boy saw the man with the telescope.”
  - “Anto memukul Budi. Dia meraung kesakitan.”
- Ada aturan (rule), tapi banyak pengecualian (*exception*)
- Bahasa senantiasa berubah terutama dalam percakapan
  - “Dia sama Ani pergi.” (sama → bersama)
  - “Kamu makan sama apa? Makan ayam” (makan ayam atau sama ayam?)
  - Kata tidak baku (ngapain, cius miapah, dll)



# Contoh aturan & pengecualian

- Aturan morfologis: be + VK → beR + VK
  - beR+uban, beR+ujung, dll
- kata dasar: ajar
  - be+ajar → be**l**ajar (kenapa bukan berajar?)
- Pada kata dengan awalan konsonan huruf **k**
- Aturan morfologis: me + KV → meNG + (K luluh) V
  - me+kurung → mengurung, me+koordinasi → mengoordinasi
- Kata dasar: kaji
  - me**ng**aji atau me**ngk**aji?

Keduanya baku di KBBI, sama-sama dari kata dasar kaji

# Mengapa mempelajari & memodelkan bahasa itu sulit?

- Kita tidak mengerti dengan jelas bagaimana manusia mengolah bahasa.
- Kata *empiricist*: abaikan manusia, pelajari (banyak) data!

## **Tapi, harus hati-hati!**

Data yang banyak tapi tidak benar akan memberikan hasil yang salah!

Misal: kata tidak baku yang sering salah (***silahkan* atau *silakan?***), hoax/fake news

# Sejarah NLP

---

- **1940-1950an**
  - Pembentukan teori dasar
    - Teori bahasa formal (Chomsky)
    - Noisy channel model, information theory (Shannon & Weaver)
  - Optimisme naif tentang machine translation
    - The spirit is willing but the flesh is weak.
    - The vodka is strong but the meat is rotten.

# Sejarah NLP

---

- **Mid 1950-1970**

- Mulai terbentuk komunitas simbolik (FIB) vs. statistik (FT)
- Model simbolik *berdasarkan context-free grammar* dan *transformational grammar*-nya Chomsky.
- Program NLU/dialogue sederhana berdasarkan pattern-matching, mis: ELIZA (Weizenbaum)
- Metode statistik digunakan untuk OCR dan penentuan pengarang.

# Sejarah NLP

---

- **1970-1983**

- SHRDLU-nya Winograd (bimbingan Minsky)
- Penggunaan *grammar* dan *parser* yang semakin canggih.
- Pendekatan *logic-based untuk syntax & semantics*  
→ PROLOG (Colmerauer)
- *Going beyond the sentence: discourse modeling*  
(Grosz & Sidner)
- **Hidden Markov Models** (HMM) untuk *speech recognition*

# Sejarah NLP

---

- **1983-1993**

- Bangkit kembali: *finite-state model*, terutama untuk *morphology*
- Bangkit kembali: *probabilistic model* (*speech recognition*-nya lab IBM): *part-of-speech tagging*, statistical parsing, dll.
- Riset ke dalam *Natural Language Generation* (NLG)

# Sejarah NLP

---

- **1994-sekarang**
  - Makin maraknya penggunaan model probabilistik dan empiris, dengan bantuan teori linguistik.
  - Ilmu semakin matang, metodologi evaluasi yang jelas.
  - **Meledak!** WWW, Google, Facebook, Twitter, data, hardware, uang!
- Tren sekarang Big Data untuk NLP → Intelligence Personal Assistant, fake news/hoax

# Portal Paper Jurnal & Conference

---

- **Association for Computational Linguistics (ACL) Web**
  - <https://www.aclweb.org/>
- **IEEE (proxy UB)**
  - IEEE Computer Society: <https://www.computer.org>
  - IEEE Xplore: <http://ieeexplore.ieee.org>
- **Association for Computing Machinery (ACM)**
  - <https://www.acm.org>
- **ScienceDirect (proxy UB)**
  - <http://www.sciencedirect.com/>
- **CiteseerX**
  - <http://citeseerx.ist.psu.edu/>
- **Google Scholar**
  - <http://scholar.google.com/>



# Tugas

---

- Bahasa pemrograman yang utama digunakan dalam kuliah ini adalah Python!
  - Cross platform, default di Unix/Linux
- Instalasi Python 2.7
  - Bisa melalui distribusi: Continuum Anaconda, Python(x,y), WinPython
  - Disarankan Anaconda
- Pelajari Python: sintaksis dan struktur, input-output, seleksi kondisi, iterasi
- Pertemuan berikutnya: Pemrosesan Teks dengan Python