

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 anwer/QA)
 - Bahasa manusia: Indonesia, Inggris, Jepang, Perancis, Jawa, dll.
 - Domain: undang-undang, iklan, micro text (SMS, tweet), dll



- 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
 WolframAlpha computational Moderation of Moder



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



Machine Translation

 Google Translate (translate.google.com) adalah sistem/mesin penerjemah yang dapat menerjemahkan antar bahasa

Speech Recognition

- Sphinx (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

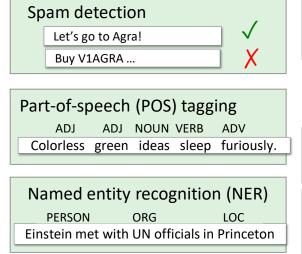


- Query basis data dalam bahasa manusia
- Peringkasan dokumen: menghasilkan ringkasan dokumen (abstrak)
- Intelligent Tutoring System: berdialog dengan siswa
- Automated Essay Scoring/Grading: aplikasi untuk menilai suatu essay
- Plagiarism checker, dan masih banyak lagi

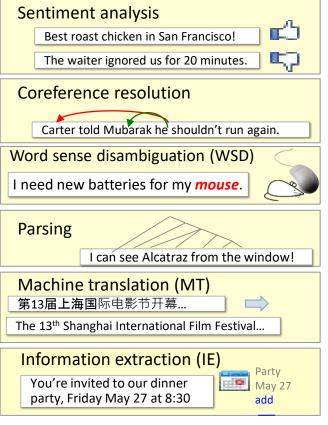
Keberhasilan teknik ditentukan oleh pemahaman tentang bahasa itu sendiri dan teknik mengolah bahasa.



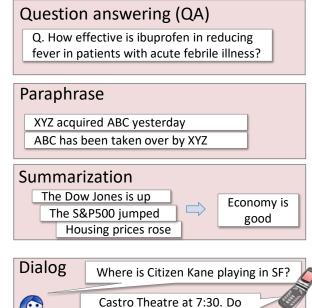
banyak diselesaikan



Berkembang dengan baik



masih sulit tapi berkembang



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 \rightarrow 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. \rightarrow sah Ani merangkai bunga. vs. Ani dirangkai bunga. \rightarrow ???

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
 → belajar (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
 - mengaji atau mengkaji?

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



• 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.



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.



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



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)



• 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, inputoutput, seleksi kondisi, iterasi
- Pertemuan berikutnya: Pemrosesan Teks dengan Python

