
Natural Language Processing Introduction

Ahmad Rio Adriansyah, M.Si.

STT Terpadu Nurul Fikri

*diadaptasi dari slide Raymond J. Mooney (Univ. Texas)
dan slide Dan Jurafsky (Univ. Stanford)

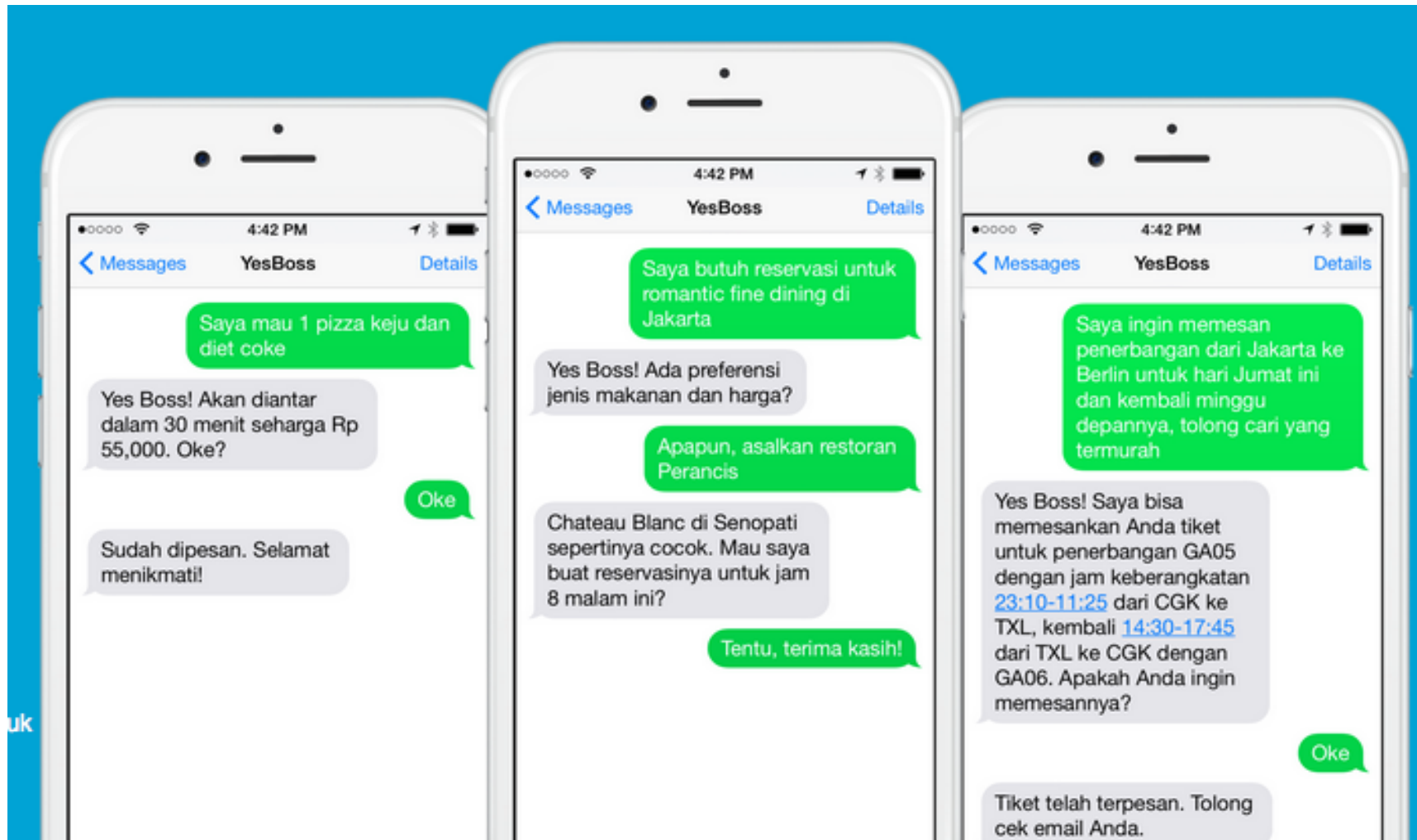
Natural Language Processing

- Natural Language Processing (NLP) adalah cabang ilmu komputer yang fokus ke pengembangan sistem yang membuat komputer dapat berkomunikasi dengan manusia menggunakan bahasa sehari hari.
- Natural Language Processing \neq Neuro Linguistic Programme
- Disebut juga **Computational Linguistics**
 - Membahas juga tentang metode komputasi untuk memahami bahasa manusia

Area Terkait

- Kecerdasan Buatan
- Teori Bahasa Formal dan Automata
- Pembelajaran Mesin (Machine Learning)
- Interaksi Manusia Computer (HCI)
- Bahasa (Linguistics)
- Psikologi Bahasa (Psycholinguistics)
- Cognitive Science
- Filosofi Bahasa (Philosophy of Language)

Penerapan : Question Answering



Penerapan : Ekstraksi Informasi

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30. ▼

-Chris

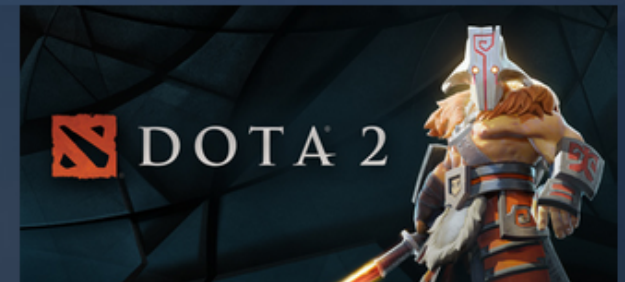
Create new Calendar entry

Penerapan : Sentimen Analisis

All Games > Free to Play Games > Dota 2

Dota 2

Community Hub



Every day, millions of players worldwide enter battle as one of over a hundred Dota heroes. And no matter if it's their 10th hour of play or 1,000th, there's always something new to discover. With regular

81% of the 14,858 user reviews in the last 30 days are positive.

RECENT REVIEWS: **Very Positive** (14,858)

ALL REVIEWS: **Very Positive** (861,774)

RELEASE DATE: 9 Jul, 2013

DEVELOPER: Valve

PUBLISHER: Valve

Popular user-defined tags for this product:

Free to Play MOBA Strategy Multiplayer PvP +

Penerapan : Machine Translation



Translate

French Japanese Indonesian Detect language ▼



Indonesian Japanese English ▼

Translate

苦あれば 楽あり



Ada sukacita jika ada kepahitan



8/5000



Ku areba raku ari

Penerapan : Summarization



#NoLimitUpdate

18 - 24 Februari 2019

TOP TRENDING
PEOPLE
WEEK #08

Ahmad Dhani - 2.655 Talk

Pemberitaan terkait kunjungan rekan dan keluarga Ahmad Dhani di rutan Medaeng Sidoarjo

Syahrini - 2.491 Talk

Perberitaan terkait rencana pernikahan Syahrini dengan Reino Barack

Reino Barack - 1.721 Talk

Pemberitaan terkait rencana pernikahan Syahrini dengan Reino Barack

Jusuf Kalla - 1.685 Talk

Pemberitaan terkait ucapan JK bahwa lahan milik Prabowo sesuai dengan Undang-Undang

Indra Sjafri - 1.602 Talk

Pemberitaan terkait Keberhasilan Indra Sjafri yang membawa Timnas U-22 ke final Piala AFF



#NoLimitUpdate

18 - 24 Februari 2019

TOP TRENDING
LOCATION
WEEK #08

Mall Taman Anggrek - 1.124 Talk

Pemberitaan terkait ledakan di Mall Taman Anggrek

Rutan Medaeng - 668 Talk

Pemberitaan terkait kunjungan rekan dan keluarga Ahmad Dhani di rutan Medaeng Sidoarjo

Stadion Kanjuruhan - 498 Talk

Pemberitaan terkait pertandingan leg kedua babak 16 besar, antara Arema FC dengan Persib Bandung

Stadion Si Jalak Harupat - 450 Talk

Pemberitaan terkait pertandingan leg pertama babak 16 besar, antara Persib Bandung dengan Arema FC

Pelabuhan Muara Baru - 431 Talk

Pemberitaan terkait kebakaran sejumlah kapal nelayan di pelabuhan Muara Baru



Simpan nomor
081-211-211-851
dan kirim

#NOLIMITUPDATE

onm.nolimit.id



Simpan nomor
081-211-211-851
dan kirim

#NOLIMITUPDATE

onm.nolimit.id

@nolimitid

@nolimitid

Penerapan : Speech Recognition / Captioning



Lecture 1 | Natural Language Processing with Deep Learning

215,812 views

1K 22 SHARE ...

NLP Bahasa Inggris

Dan Jurafsky



Language Technology

making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy VIAGRA ... ✗

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

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 13th Shanghai International Film Festival...

Information extraction (IE)

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

Party
May 27
add

still really hard

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?

Communication

- Tujuan dari **produksi** dan **pemahaman** bahasa alami adalah komunikasi.



Communication (cont)

- Dari sisi pembicara (komunikator):
 - **Intention**: Menentukan kapan dan informasi apa yang akan dikirim (nama lain : *content selection, strategic generation*). Membutuhkan perencanaan dan pemahaman dari tujuan dan anggapan.
 - **Generation**: Mengubah informasi yang akan dikomunikasikan (dalam representasi logika atau “bahasa pikiran”) ke dalam bentuk string dalam bahasa tertentu (nama lain : *surface realization, tactical generation*).
 - **Synthesis**: Mengeluarkan output dalam bentuk modality yang dikehendaki, teks atau oral

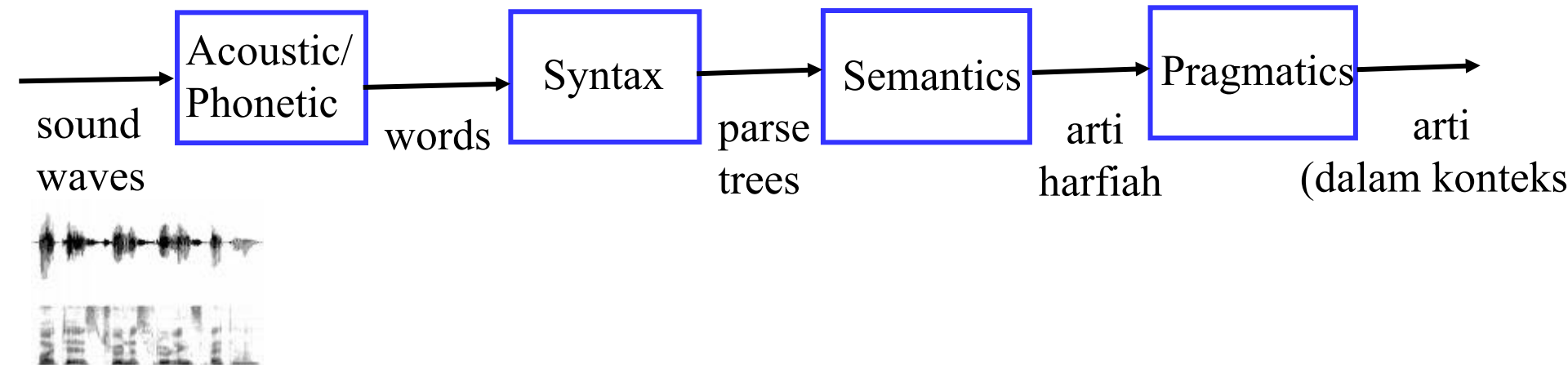
Communication (cont)

- Dari sisi pendengar (komunikasikan):
 - **Perception**: Memetakan input modality ke dalam rangkaian kata (contoh : *optical character recognition* (OCR) atau *speech recognition*(ASR)).
 - **Analysis**: Menentukan konten informasi dari string / rangkaian kata
 - **Syntactic interpretation (parsing)**: Menentukan parse tree yang tepat, yang menunjukkan struktur dari string tersebut.
 - **Semantic Interpretation**: Menyarikan arti (harfiah) dari string (*logical form*).
 - **Pragmatic Interpretation**: Mempertimbangkan efek konteks keseluruhan yang dapat mengubah arti dari sebuah kalimat.
 - **Incorporation**: Menentukan apakah akan mempercayai sebuah konten informasi atau tidak dan menambahkannya ke basis pengetahuan (knowledge base).

Syntax, Semantic, Pragmatics

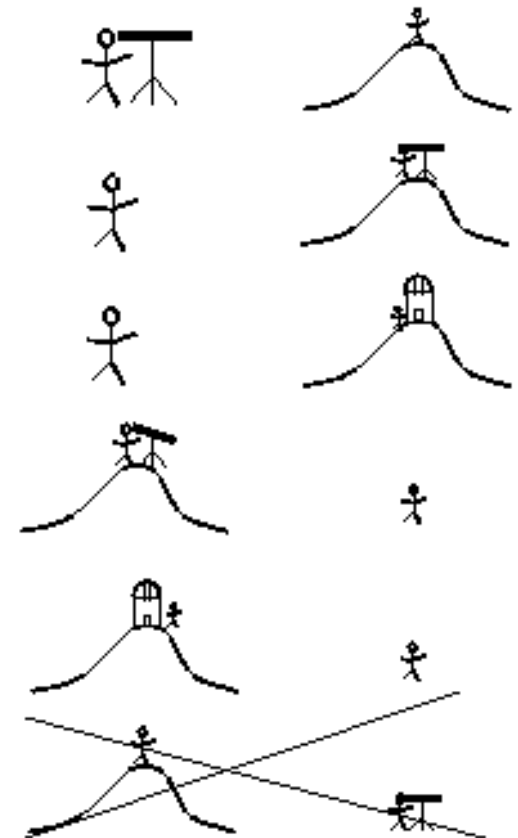
- Syntax (sintaksis) berkaitan dengan urutan baku dari kata dan pengaruhnya ke arti.
 - Anjing itu menggigit anak laki-laki.
 - Anak laki-laki menggigit anjing itu.
 - Anak anjing itu menggigit laki-laki.
 - Anjing laki-laki anak itu menggigit.
- Semantics (semantik) berkaitan dengan arti (harfiah) dari kata, frasa, dan kalimat.
 - “buku” sebagai lembaran kertas yang berjilid
 - “buku” sebagai tempat pertemuan dua ruas
- Pragmatics (pragmatik) berkaitan dengan keseluruhan konteks sosial dan konteks komunikasi dan pengaruhnya ke interpretasi.
 - Doktor itu bertangan dingin
 - Karena sering mabuk, wanita itu memukul anaknya

Modular Comprehension



Ambiguity

- Bahasa alami **sangat** ambigu dan harus didisambiguasikan (dibuat tidak bermakna ganda).
 - Saya melihat orang di gunung dengan teleskop.
 - *I made her duck.*



Ambiguity is Ubiquitous

- Speech Recognition
 - “recognize speech” vs. “wreck a nice beach”
 - “youth in Asia” vs. “euthanasia”
- Syntactic Analysis
 - “Saya makan mi pakai garpu” vs. “Saya makan mi pakai bakso.”
- Semantic Analysis
 - “The dog is in the **pen**.” vs. “The ink is in the **pen**.”
 - “I put the **plant** in the window” vs. “Ford put the **plant** in Mexico”
- Pragmatic Analysis
 - **From “The Pink Panther Strikes Again”:**
 - **Clouseau:** Does your dog bite?
Hotel Clerk: No.
Clouseau: [*bowing down to pet the dog*] Nice doggie.
[*Dog barks and bites Clouseau in the hand*]
Clouseau: I thought you said your dog did not bite!
Hotel Clerk: That is not my dog.

Ambiguity is Explosive

- Kalimat yang ambigu bisa membuat sangat banyak interpretasi yang mungkin.
- Dalam bahasa Inggris, kalimat dengan n buah frasa preposisi memiliki lebih dari 2^n interpretasi sintaksis (cf. Catalan numbers).
 - “I saw the man with the telescope”: 2 parses
 - “I saw the man on the hill with the telescope.”: 5 parses
 - “I saw the man on the hill in Texas with the telescope”: 14 parses
 - “I saw the man on the hill in Texas with the telescope at noon.”: 42 parses
 - “I saw the man on the hill in Texas with the telescope at noon on Monday” 132 parses

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
 - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
 - She criticized my apartment, so I knocked her flat.
 - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.
 - Policeman to little boy: “We are looking for a thief with a bicycle.” Little boy: “Wouldn't you be better using your eyes.”
 - Why is the teacher wearing sun-glasses. Because the class is so bright.

Why is Language Ambiguous?

- Ekspresi bahasa yang unik untuk setiap kemungkinan konsep bisa membuat bahasa menjadi sangat kompleks dan kalimatnya menjadi sangat panjang.
- Membiarkan ambiguitas yang mudah dibedakan membuat ekspresi bahasa lebih singkat (kompresi data).
- Bahasa bergantung pada kemampuan manusia untuk menggunakan pengetahuan dan kemampuan penyimpulan menangani ambiguitas secara tepat.
- Disambiguasi kadang gagal (lossy compression)

Natural Languages vs. Computer Languages

- Ambiguitas adalah pembeda utama antara bahasa alami dan bahasa komputer.
- Bahasa pemrograman formal didesain untuk tidak ambigu. Bisa dijelaskan dengan grammar yang menghasilkan uraian (parse) yang unik untuk setiap kalimat dalam bahasa tersebut.
- Bahasa pemrograman juga didesain untuk penguraian yang efisien (deterministik CFL)
 - Kalimat dalam DCFL dengan panjang string n bisa diurai dalam waktu $O(n)$.

Natural Language Tasks

- Mengolah bahasa alami melibatkan banyak pekerjaan sintaksis, semantik, dan pragmatik.
- Pengolahan bahasa alami tidak terbatas pada hal tersebut saja.

Why Natural Language Processing is Hard?

- Penggunaan bahasa tidak standard
- Masalah segmentasi
- Idioms
- Kata-kata baru
- Pengetahuan umum yang trivial bagi manusia tapi sulit bagi komputer
- Name Entity yang rancu
- dll

Syntactic Tasks

Word Segmentation

- Membagi string dari karakter (graphemes) ke dalam rangkaian kata.
- Beberapa bahasa tulis (misal Mandarin) kata tidak dipisahkan menggunakan spasi.
- Dalam bahasa Indonesia pun, karakter lain selain spasi juga dapat digunakan untuk memisahkan kata [misal , ; . - : ()]
- Contoh:
 - bukalapak.com \Rightarrow buka lapak .com
 - thetabledownthere
 - \Rightarrow the table down there
 - \Rightarrow theta bled own there

Morphological Analysis

- ***Morphology (Morfologi)*** adalah bidang linguistik yang mempelajari struktur internal dari kata
- ***Morpheme (Morfem)*** adalah unit terkecil dari kata yang memiliki arti semantik
 - contoh : “nir-”, “pra-”, “-isasi”, “ber-”, “-an”, “-logi”
- Analisis morfologi adalah pekerjaan mengurai kata ke dalam morfemnya :
 - nirkabel \Rightarrow nir + kabel (tidak / tanpa)
 - standardisasi \Rightarrow standard + isasi (proses)
 - kelima \Rightarrow ke + lima
 - Googlers \Rightarrow (Google + er) + s (plural)
 - unlockable \Rightarrow un + (lock + able) ?
 \Rightarrow (un + lock) + able ?

Part Of Speech (POS) Tagging

- Menganotasi tiap kata dalam kalimat dengan sebuah **part-of-speech**.

I ate the spaghetti with meatballs.

Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

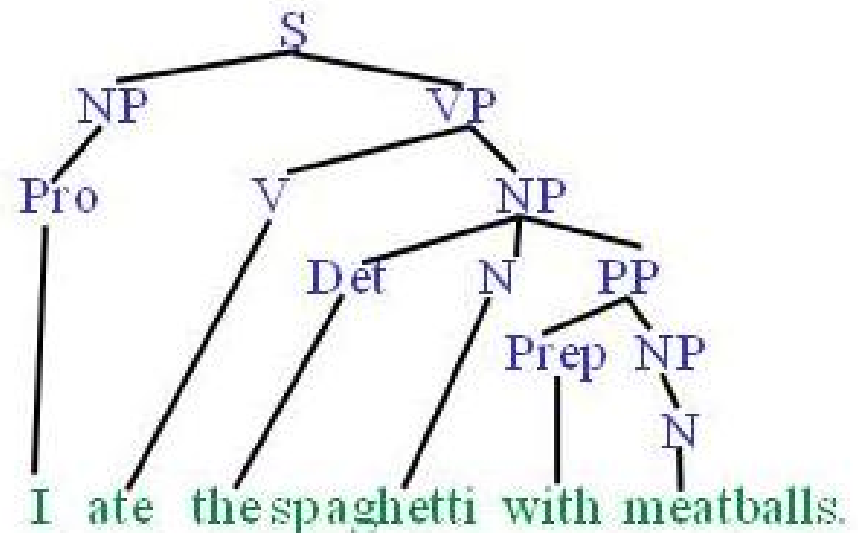
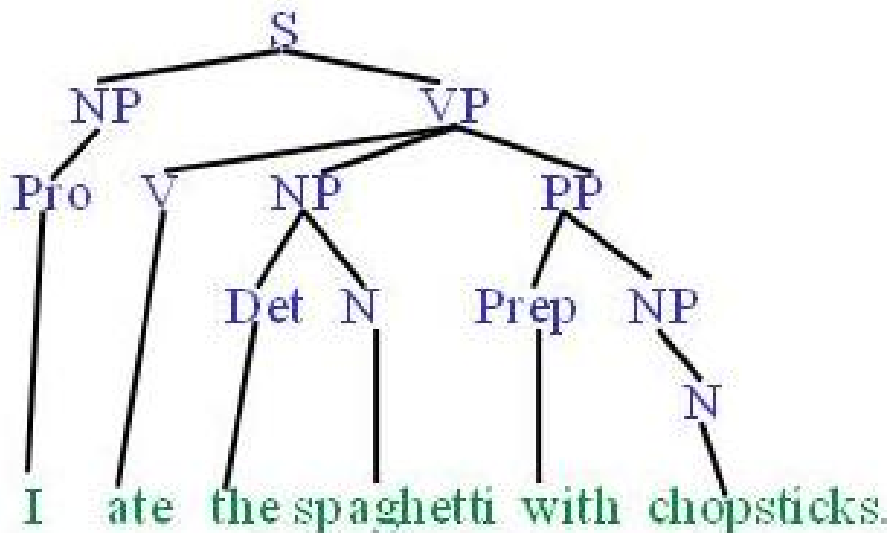
- Berguna untuk proses berikutnya, yaitu **syntactic parsing** dan **word sense disambiguation**.

Phrase Chunking

- Menemukan semua noun phrases (NPs) dan verb phrases (VPs) yang nonrekursif dalam sebuah kalimat.
 - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
 - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Menghasilkan pohon urai sintaksis (syntactic parse tree) yang tepat dari sebuah kalimat.



Semantic Tasks

Word Sense Disambiguation (WSD)

- Sebuah kata dalam bahasa alami biasanya bisa memiliki beberapa arti yang berbeda.
 - Orang tua Ellen bekerja di bidang komputasi linguistik.
 - Ellen melihat ada orang tua tersesat di jalan.
- Untuk banyak pekerjaan (question answering, translation), harus ditentukan arti kata yang tepat untuk tiap kata yang ambigu dalam kalimat.

Semantic Role Labeling (SRL)

- Untuk tiap klausa, menentukan peran semantik (semantic role) dari masing masing noun phrase yang merupakan argumen dari verba.

agent patient source destination instrument

— John drove Mary from Austin to Dallas in his Toyota Prius.

— The hammer broke the window.

- Nama lainnya “case role analysis,” “thematic analysis,” dan “shallow semantic parsing”

Semantic Parsing

- *Semantic parser* memetakan kalimat dalam bahasa alami ke dalam representasi semantik (*logical form*) yang lengkap dan detail.
- Untuk beberapa aplikasi output yang diharapkan adalah yang bisa dieksekusi langsung oleh program lainnya.
- Contoh: Memetakan kalimat untuk database query

How many cities are there in the US?

```
answer(A, count(B, (city(B), loc(B, C),  
                                const(C, countryid(USA))),  
A))
```

Textual Entailment



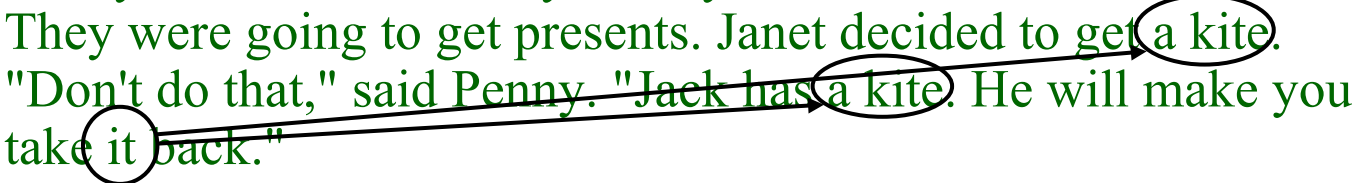
- Menentukan apakah sebuah kalimat dalam bahasa alami menyiratkan kalimat lain dalam interpretasi biasa

Textual Entailment Problems from PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
<i>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</i>	<i>Yahoo bought Overture.</i>	TRUE
<i>Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</i>	<i>Microsoft bought Star Office.</i>	FALSE
<i>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</i>	<i>Israel was established in May 1971.</i>	FALSE

Pragmatics/Discourse Tasks

Anaphora Resolution/ Co-Reference

- Menentukan frasa mana saja dalam dokumen yang mengacu ke hal yang sama.
 - John put the carrot on the plate and ate it.
 - Bush started the war in Iraq. But the president needed the consent of Congress.
- Dalam beberapa kasus, penalarannya sulit.
 - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Kata dan frasa yang berulang kadang dihilangkan dari kalimat jika konteksnya cukup jelas dan kata tersebut bisa disimpulkan dari kalimat yang ada.

"Wise men talk because they have something to say; fools, because they have to say something." (Plato)

"Wise men talk because they have something to say; fools **talk** because they have to say something." (Plato)

Other Tasks

Information Extraction (IE)

- Mengidentifikasi frasa dalam bahasa yang mengacu ke tipe entitas dan relasi tertentu pada teks.
- **Named entity recognition** adalah pekerjaan untuk mengidentifikasi nama dari orang, tempat, organisasi, dll pada teks.

people organizations places

– Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

- **Relation extraction** mengidentifikasi hubungan tertentu antar entitas.

– Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.



The diagram illustrates relation extraction by drawing arcs between entities in the sentence "Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas." An orange arc connects "Michael Dell" to "CEO", and another orange arc connects "Dell Computer Corporation" to "CEO". A brown arc connects "Dell Computer Corporation" to "lives in", and another brown arc connects "Austin Texas" to "lives in".

Question Answering

- Menjawab pertanyaan dalam bahasa alami berdasarkan informasi yang ada dalam corpora dokumen teks (misalnya web).
 - When was Barack Obama born? (*factoid*)
 - August 4, 1961
 - Who was president when Barack Obama was born?
 - John F. Kennedy
 - How many presidents have there been since Barack Obama was born?
 - 9

Text Summarization

- Membuat ringkasan dari dokumen atau artikel yang panjang.
 - **Article:** With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee....
 - **Summary:** Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Machine Translation (MT)

- Menerjemahkan kalimat dari satu bahasa alami ke bahasa alami lainnya.
 - Hasta la vista, bebé \Rightarrow
Until we see each other again, baby.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
 - “John plays the guitar.” → “John toca la guitarra.”
 - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
 - “The spirit is willing but the flesh is weak.” ⇒ “The liquor is good but the meat is spoiled.”
 - “Out of sight, out of mind.” ⇒ “Invisible idiot.”

Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires knowledge of:
 - Syntax
 - An agent is typically the subject of the verb
 - Semantics
 - Michael and Ellen are names of people
 - Austin is the name of a city (and of a person)
 - Toyota is a car company and Prius is a brand of car
 - Pragmatics
 - World knowledge
 - Credit cards require users to pay financial interest
 - Agents must be animate and a hammer is not animate

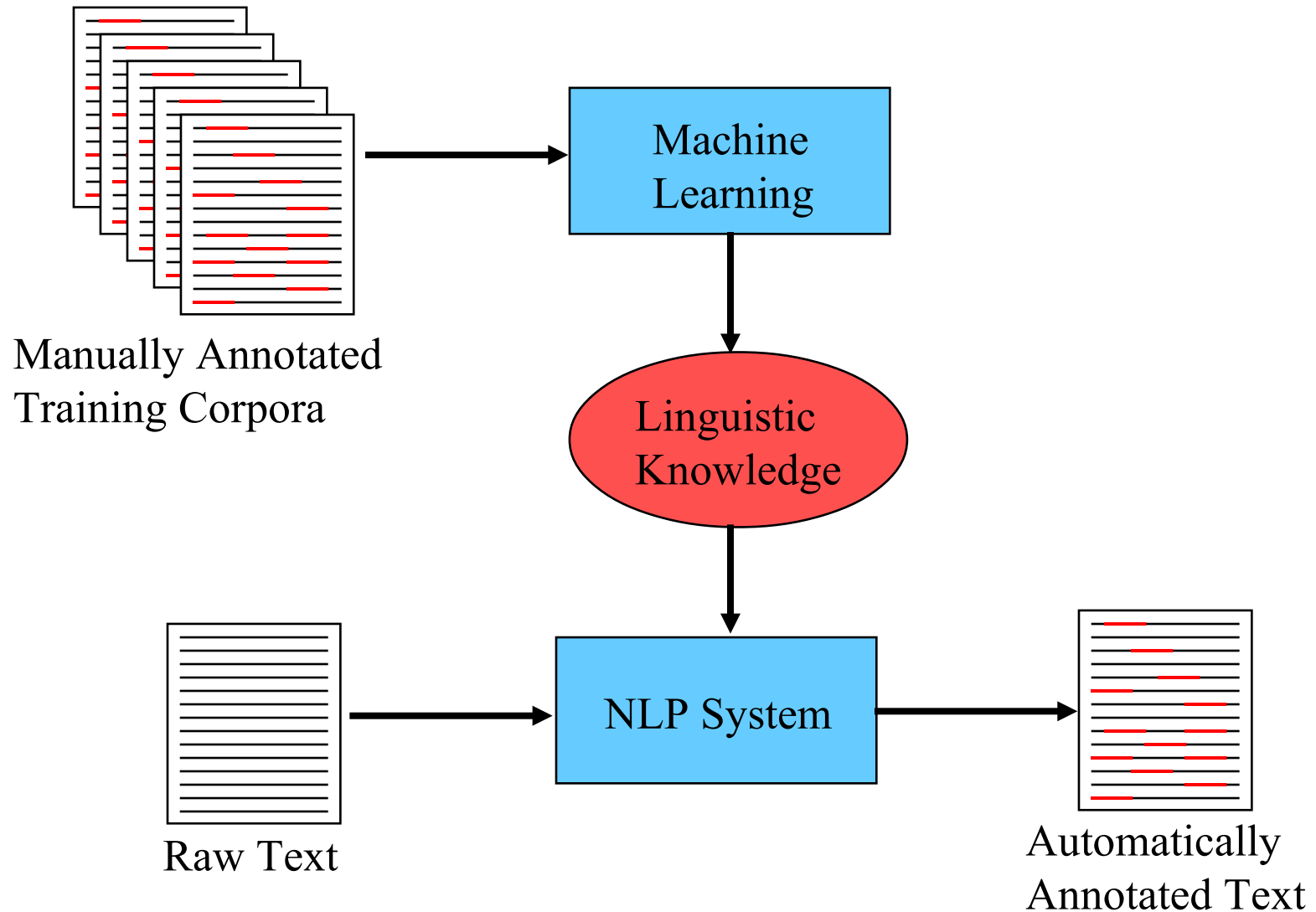
Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
 - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Various referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

Learning Approach



Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
 - “The a are of I” is a valid English noun phrase (Abney, 1996)
 - “a” is an adjective for the letter A
 - “are” is a noun for an area of land (as in hectare)
 - “I” is a noun for the letter I
 - “Time flies like an arrow” has 4 parses, including those meaning:
 - Insects of a variety called “time flies” are fond of a particular arrow.
 - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
 - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Human Language Acquisition

- Human children obviously learn languages from experience.
- However, it is controversial to what extent prior knowledge of “universal grammar” (Chomsky, 1957) facilitates this acquisition process.
- Computational studies of language learning may help us to understand human language learning, and to elucidate to what extent language learning must rely on prior grammatical knowledge due to the “poverty of the stimulus.”
- Existing empirical results indicate that a great deal of linguistic knowledge can be effectively acquired from reasonable amounts of real linguistic data without specific knowledge of a “universal grammar.”

Pipelining Problem

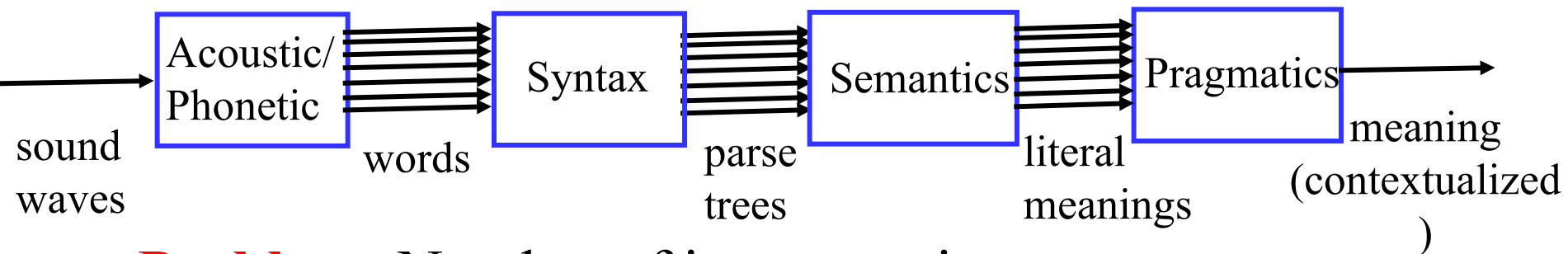
- Assuming separate independent components for speech recognition, syntax, semantics, pragmatics, etc. allows for more convenient modular software development.
- However, frequently constraints from “higher level” processes are needed to disambiguate “lower level” processes.
 - Example of syntactic disambiguation relying on semantic disambiguation:
 - At the zoo, several men were showing a group of students various types of flying animals. Suddenly, one of the students hit the man **with** a **bat**.

Pipelining Problem (cont.)

- If a hard decision is made at each stage, cannot backtrack when a later stage indicates it is incorrect.
 - If attach “with a bat” to the verb “hit” during syntactic analysis, then cannot reattach it to “man” after “bat” is disambiguated during later semantic or pragmatic processing.

Increasing Module Bandwidth

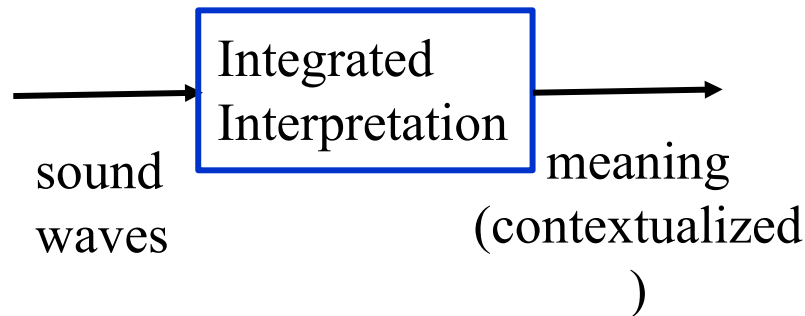
- If each component produces multiple scored interpretations, then later components can rerank these interpretations.



- **Problem:** Number of interpretations grows combinatorially.
- **Solution:** Efficiently encode combinations of interpretations.
 - Word lattices
 - Compact parse forests

Global Integration/ Joint Inference

- Integrated interpretation that combines phonetic/syntactic/semantic/pragmatic constraints.



- Difficult to design and implement.
- Potentially computationally complex.

Early History: 1950's

- Shannon (the father of information theory) explored probabilistic models of natural language (1951).
- Chomsky (the extremely influential linguist) developed formal models of syntax, i.e. finite state and context-free grammars (1956).
- First computational parser developed at U Penn as a cascade of finite-state transducers (Joshi, 1961; Harris, 1962).
- Bayesian methods developed for *optical character recognition* (OCR) (Bledsoe & Browning, 1959).

History: 1960's

- Work at MIT AI lab on question answering (BASEBALL) and dialog (ELIZA).
- Semantic network models of language for question answering (Simmons, 1965).
- First electronic corpus collected, Brown corpus, 1 million words (Kucera and Francis, 1967).
- Bayesian methods used to identify document authorship (*The Federalist* papers) (Mosteller & Wallace, 1964).

History: 1970's

- “Natural language understanding” systems developed that tried to support deeper semantic interpretation.
 - SHRDLU (Winograd, 1972) performs tasks in the “blocks world” based on NL instruction.
 - Schank *et al.* (1972, 1977) developed systems for conceptual representation of language and for understanding short stories using hand-coded knowledge of scripts, plans, and goals.
- Prolog programming language developed to support logic-based parsing (Colmerauer, 1975).
- Initial development of hidden Markov models (HMMs) for statistical speech recognition (Baker, 1975; Jelinek, 1976).

History: 1980's

- Development of more complex (mildly context sensitive) grammatical formalisms, e.g. unification grammar, HPSG, tree-adjoining grammar.
- Symbolic work on discourse processing and NL generation.
- Initial use of statistical (HMM) methods for syntactic analysis (POS tagging) (Church, 1988).

History: 1990's

- Rise of statistical methods and empirical evaluation causes a “scientific revolution” in the field.
- Initial annotated corpora developed for training and testing systems for POS tagging, parsing, WSD, information extraction, MT, etc.
- First statistical machine translation systems developed at IBM for Canadian Hansards corpus (Brown *et al.*, 1990).
- First robust statistical parsers developed (Magerman, 1995; Collins, 1996; Charniak, 1997).
- First systems for robust information extraction developed (e.g. MUC competitions).

History: 2000's

- Increased use of a variety of ML methods, SVMs, logistic regression (i.e. max-ent), CRF's, etc.
- Continued developed of corpora and competitions on shared data.
 - TREC Q/A
 - SENSEVAL/SEMEVAL
 - CONLL Shared Tasks (NER, SRL...)
- Increased emphasis on unsupervised, semi-supervised, and active learning as alternatives to purely supervised learning.
- Shifting focus to semantic tasks such as WSD, SRL, and semantic parsing.

History: 2010's

- Grounded Language: Connecting language to perception and action.
 - Image and video description
 - Visual question answering (VQA)
 - Human-Robot Interaction (HRI) in NL
- Deep Learning: Neural network learning with many layers or recurrence.
 - Long Short Term Memory (LSTM) recurrent neural networks using encoder/decoder sequence-to-sequence mapping.
 - Neural Machine Translation (NMT)
 - Spreading to syntactic/semantic parsing and most other NLP tasks.

Relevant Scientific Conferences

- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- International Conference on Computational Linguistics (COLING)
- Empirical Methods in Natural Language Processing (EMNLP)
- Conference on Computational Natural Language Learning (CoNLL)
- International Association for Machine Translation (IMTA)

Skema Perkuliahan

14 pertemuan (bahasa Indonesia/Inggris):

- model bahasa ngram
- pos tagging
- hidden markov model
- neural network
- sintaktik parsing
- semantik role labeling, semantik parsing
- information extraction / machine translation / text categorization

Skema Penilaian

Kehadiran : 5%

Tugas : 50%

UTS / UAS : 45%

Lab (optional)

Referensi

Bird, Steven et.al. Natural Language Processing with Python. 2009

Indurkha, Nitin et.al. Handbook of Natural Language Processing. 2010

Papers