



Natural Language Processing

Aryanto, M.Si



Outline

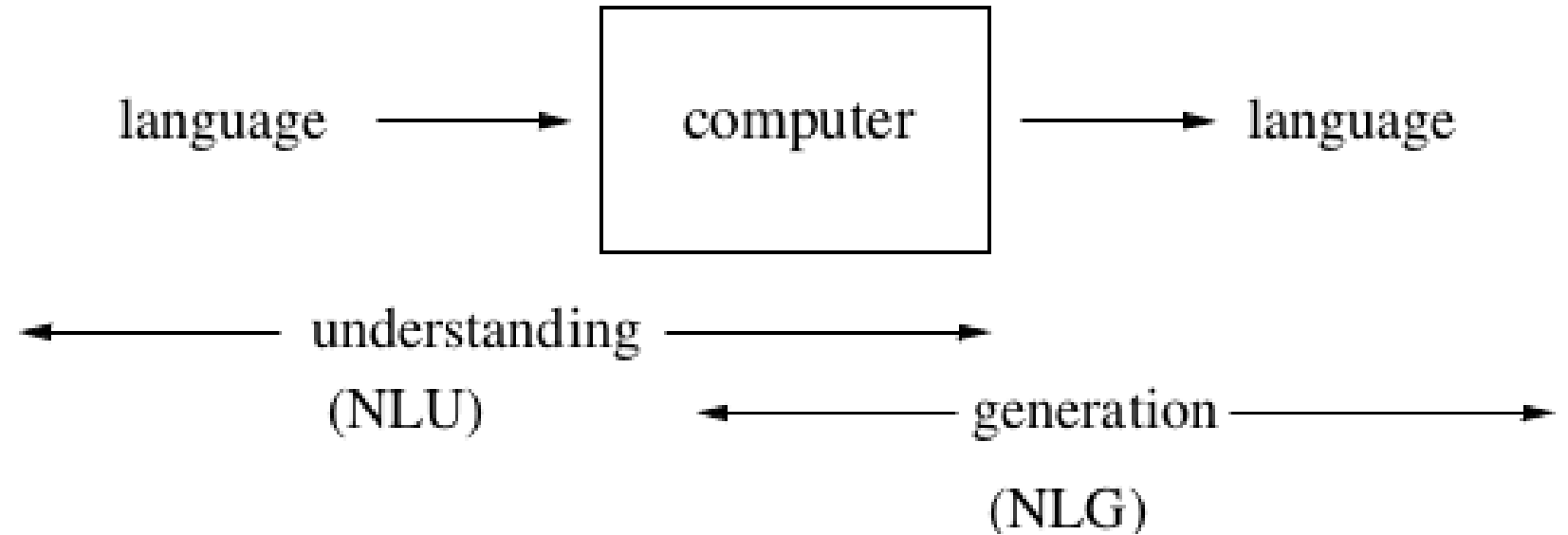
1. What is NLP?
2. Language Processing Class
3. NLP Task and Applications
4. Why NLP is Hard?



Topic one

What is Natural Language Processing?

What is Natural Language Processing?

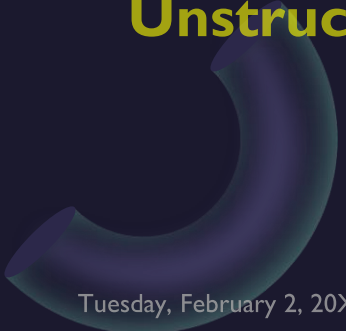


- computers using natural language as input and/or output

Motivation for NLP



IT CAN AS :

- **Understand language analysis & generation**
 - **Communication**
 - **Language is a window to the mind**
 - **Data is in linguistic form**
 - **Data can be in Structured (table form), Semi structured (XML form), Unstructured (sentence form).**
- 



Two Contrasting Views of Language

- Language as a phenomenon
- Language as a data

Topic Two

Language Processing Class

Language Processing

- *Level 1* – Speech sound (*Phonetics & Phonology*)
- *Level 2* – Words & their forms (*Morphology, Lexicon*)
- *Level 3* – Structure of sentences (*Syntax, Parsing*)
- *Level 4* – Meaning of sentences (*Semantics*)
- *Level 5* – Meaning in context & for a purpose (*Pragmatics*)
- *Level 6* – Connected sentence processing in a larger body of text (*Discourse*)

Language Processing – More Simple

1. Morphological Analysis/ Lexical Analysis

2. Syntax Analysis

3. Semantic Analysis

4. Discourse

5. Pragmatics



Morphological/ Lexical Analysis

Analisis Morfologis atau Leksikal berurusan dengan teks pada tingkat kata individu. Itu mencari morfem, unit terkecil dari sebuah kata. Misalnya, nonfungsi dapat dipecah menjadi non (awalan), miliknya dan -nya (akhiran). Analisis Leksikal menemukan hubungan antara morfem ini dan mengubah kata menjadi bentuk akarnya. Penganalisis leksikal juga menetapkan kemungkinan Part-Of-Speech (POS) ke kata tersebut. Ini mempertimbangkan kamus bahasa. Misalnya, kata "karakter" dapat digunakan sebagai kata benda atau kata kerja.



Syntax Analysis

Analisis Sintaks memastikan bahwa potongan teks yang diberikan adalah struktur yang benar. Analisis ini mencoba mengurai kalimat untuk memeriksa tata bahasa yang benar di tingkat kalimat. Mengingat kemungkinan POS yang dihasilkan dari langkah sebelumnya, penganalisis sintaks menetapkan tag POS berdasarkan struktur kalimat. Misalnya:

Sintaks yang Benar: Matahari terbit di timur.

Sintaks yang Salah: Terbit matahari di timur.



Semantic Analysis

Perhatikan kalimat: "Apel memakan pisang". Meskipun kalimatnya benar secara sintaksis, namun tidak masuk akal karena apel tidak bisa memakan pisang. Analisis semantik mencari makna dalam kalimat yang diberikan. Ini juga berkaitan dengan menggabungkan kata-kata menjadi frase. Misalnya, "apel merah" memberikan informasi mengenai satu objek; karenanya kita memperlakukannya sebagai frasa tunggal. Demikian pula, kita dapat mengelompokkan nama yang mengacu pada kategori, orang, objek, atau organisasi yang sama. "Aryanto Jack" mengacu pada orang yang sama dan bukan dua nama yang terpisah - "Aryanto" dan "Jack".



Discourse (Percakapan)

Percakapan berurusan dengan efek kalimat sebelumnya pada kalimat dalam pertimbangan. Dalam teks, “Jack adalah mahasiswa yang cerdas. Dia menghabiskan sebagian besar waktunya di perpustakaan.” Di sini, wacana menugaskan "dia" untuk merujuk pada "Jack".



Pragmatics

Tahap akhir dari NLP, Pragmatik menginterpretasikan teks yang diberikan menggunakan informasi dari langkah sebelumnya. Diberikan kalimat, “Matikan lampu itu” adalah perintah atau permintaan untuk mematikan lampu.

Topic Three

NLP Tasks and Applications

NLP tasks

Machine Translation (MT)

Information Extraction (IE)

Text Summarization

Dialogue Systems

Tagging (POS, NER)

Speech Recognition

Machine Translation (MT)

- Sub-bidang linguistik komputasi yang menyelidiki penggunaan perangkat lunak untuk menerjemahkan teks atau ucapan dari satu bahasa ke bahasa lain. (Wikipedia)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)



Parallel corpus

41. Chapter 3, Stenmarck (SV)	fr	nl
context That is true as long as account is taken of the 20 per cent of the total postal services market where , in practice , there is still a monopoly , that is where the state is the only player .	C' est exact si l' on considère la question en tenant compte des 20 pour cent du marché total des services postaux où le monopole s' est maintenu dans la pratique , c'est-à-dire là où l' État est le seul acteur .	Dat klopt als men alleen kijkt naar 20% van de totale postmarkt , waar de staat in de praktijk nog steeds het monopolie heeft .
42. Chapter 3, MacCormick	fr	nl
context The Commission should not , for example , take a stepwise jump from 350 grammes to , as some have suggested , as low as 50 grammes .	Par exemple , la Commission devrait éviter de passer de 350g à 50g , comme l' ont suggéré certains .	De Commissie moet bijvoorbeeld niet helemaal van 350 gram naar 50 gram gaan zakken , zoals sommigen hebben geopperd .

Information Extraction (IE)

- Ekstraksi informasi (IE) adalah tugas mengekstraksi informasi terstruktur secara otomatis dari dokumen yang tidak terstruktur (Wikipedia).



Information Extraction Example

10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automotive account. Experience in online marketing, automotive and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables ... Compensation: \$50,000-\$80,000 Hiring Organization: 10TH DEGREE



INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Irvine, CA
COMPANY	10TH DEGREE
SALARY	\$50,000-\$80,000

Information Extraction

Goal: Map a document collection to structured database

Motivation:

- Pencarian yang kompleks (“Temukan saya semua pekerjaan di bidang statistika yang membayar setidaknya Rp 10Jt di Indonesia”)
- Pertanyaan statistik (“Bagaimana jumlah pekerjaan di bidang statistika berubah selama bertahun-tahun?”)

Text Summarization

Document



Summary



Dialogue Systems

User: I need a flight from Boston to Washington, arriving by 10 pm.

System: What day are you flying on?

User: Tomorrow

System: Returns a list of flights

Google

Siapa penemu listrik



Berita

Gambar

Brainly

Video

Statis

Shopping

Bohlam

Generator

Telegraf

Semua filter ▾

Alat

SafeSea

Listrik / Penemu



William Greener



Schuyler Wheeler



Henry M. Leland



Ebenezer Kinnersley



ruangguru.com

<https://www.ruangguru.com/sejarah-penemuan-listrik>

Siapa Penemu Listrik? Michael Faraday atau Benjamin ...

20 Nov 2021 — Sejarah Perjalanan Penemuan **Listrik** · Sang Jagoan **Listrik** Benjamin Franklin · Michael Faraday Si Bapak **Listrik** Dunia.



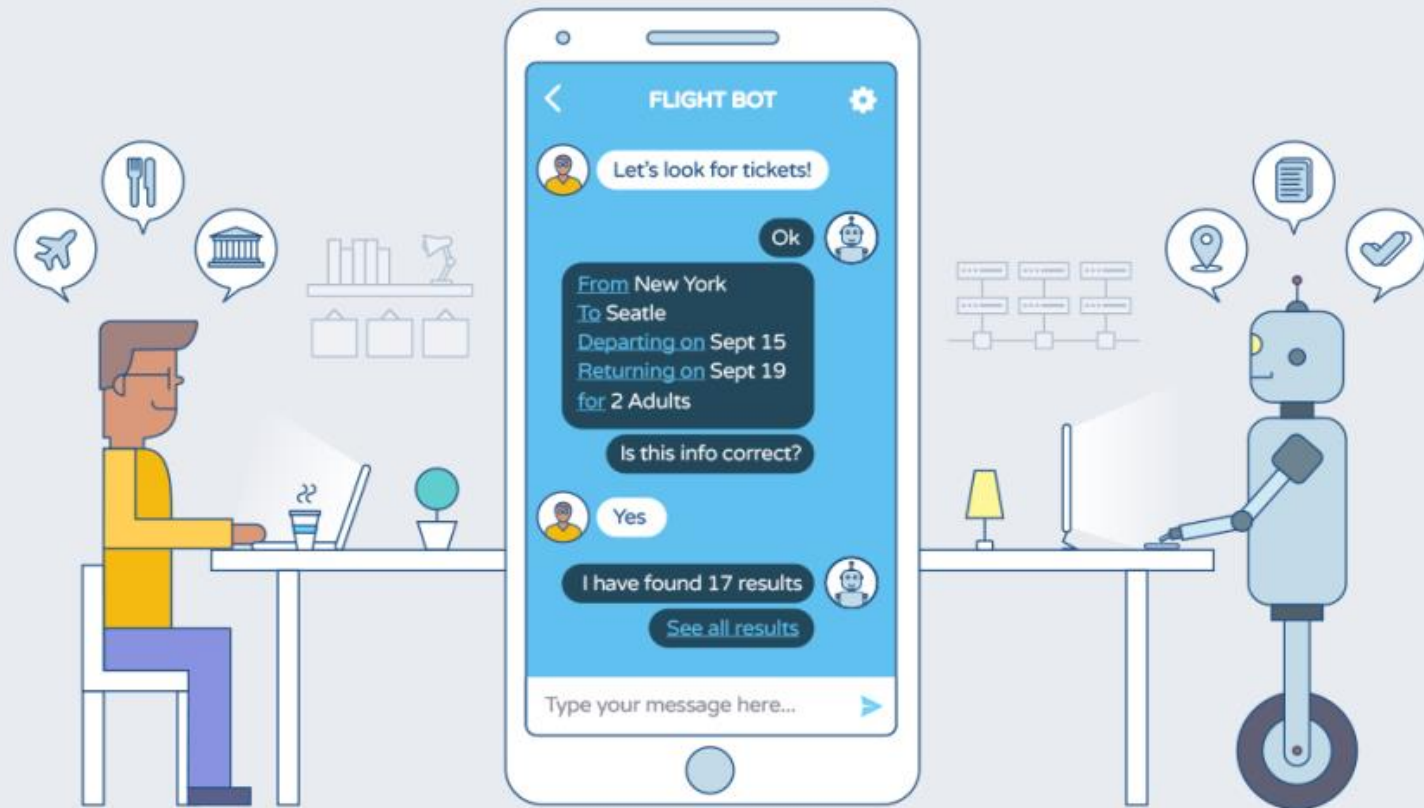
kompas.com

Listrik



Chatbots (text understanding and generation)

is a computer program which conducts a conversation via auditory or textual methods



NLU / NLG

Seq2seq

LSTM

GRU

Tagging

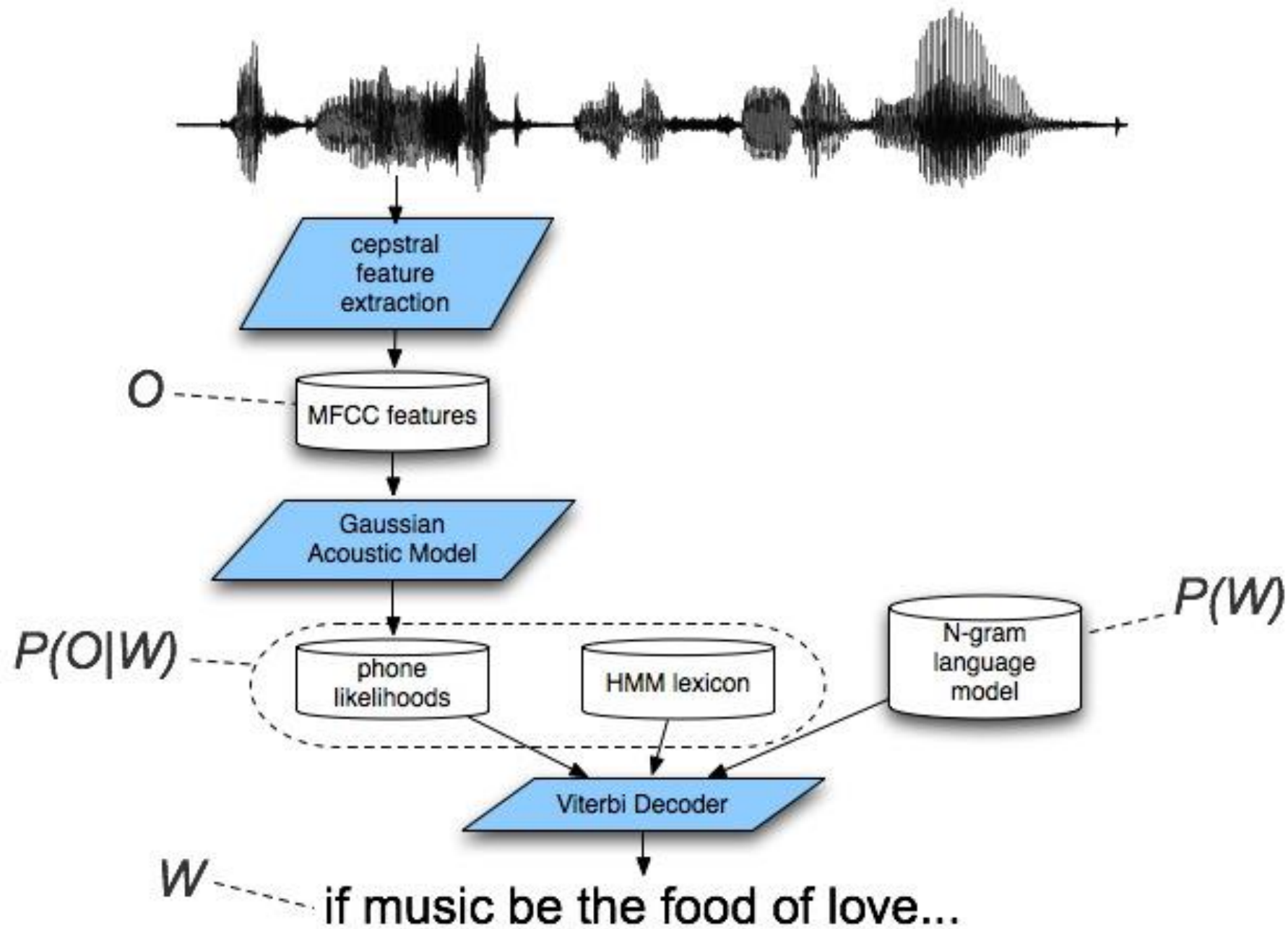
Example 1: Part-of-speech tagging

- Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ./.

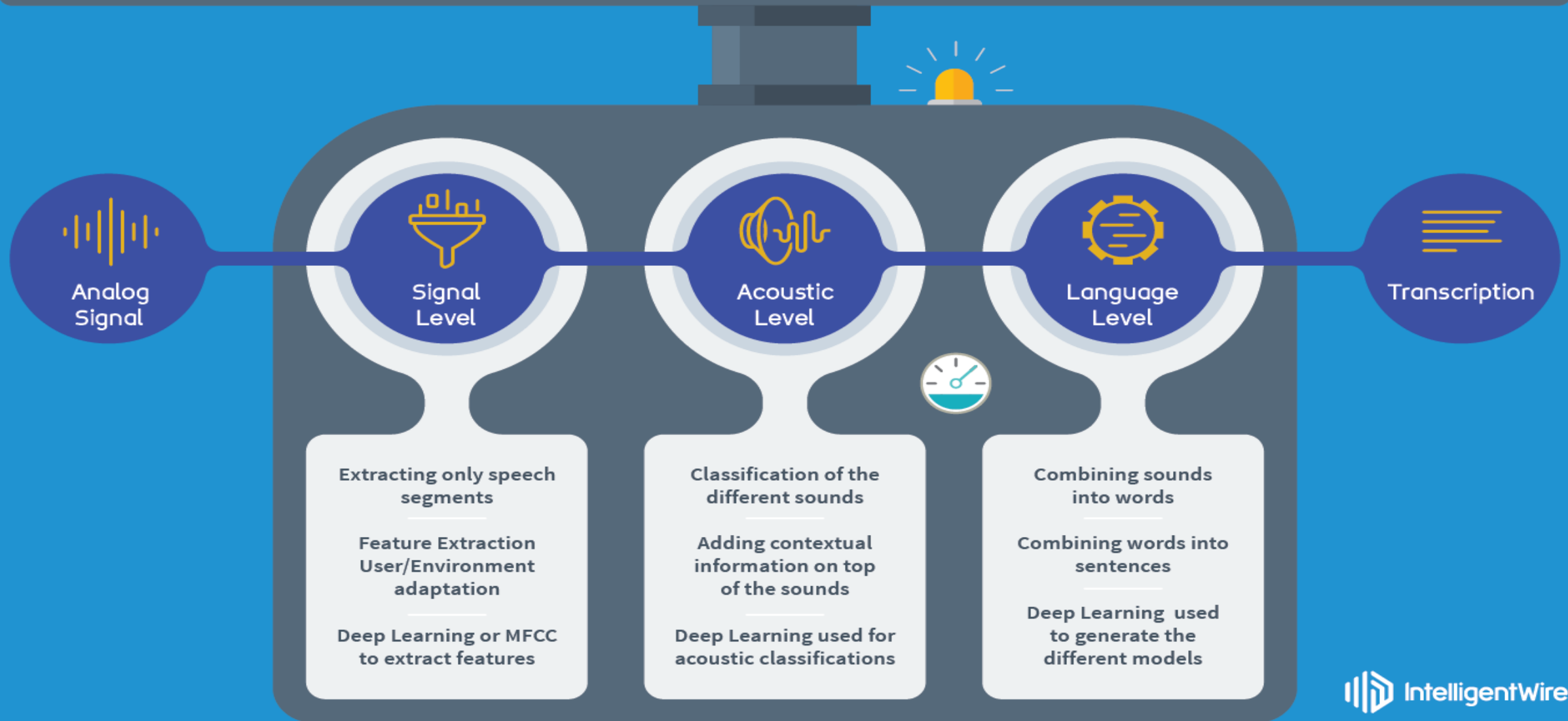
Example 2: Named Entity Recognition

- Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ./.

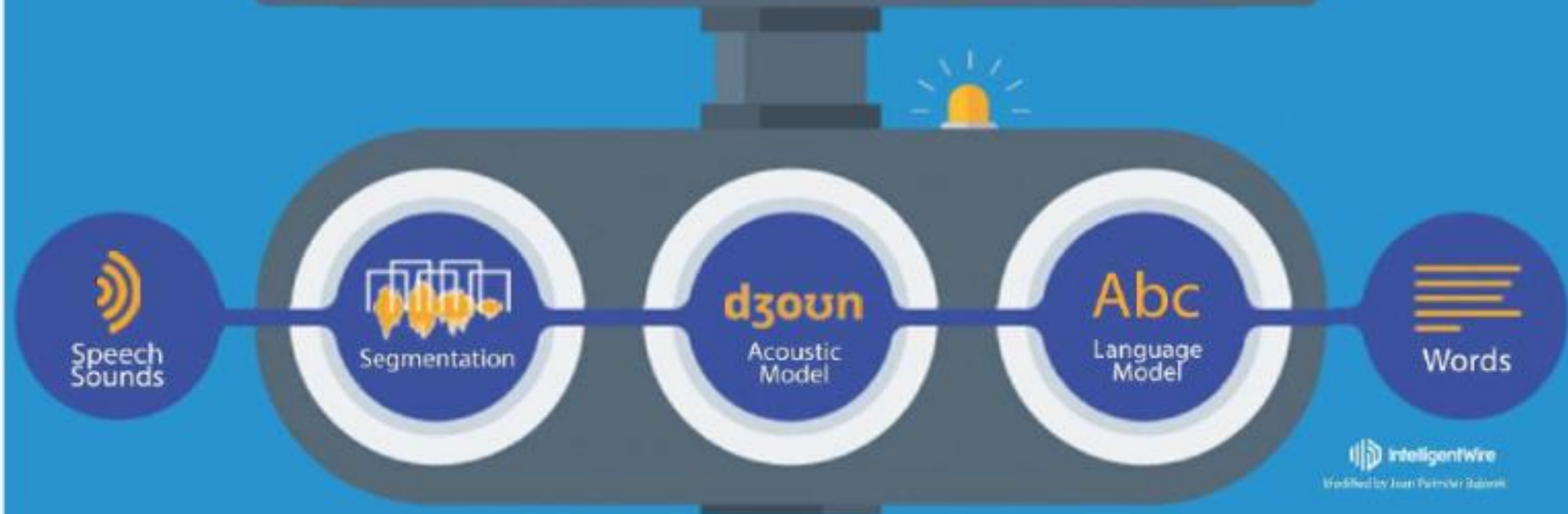
Speech Recognition



Speech Recognition System



Speech Recognition System



Topic Four

Why is Natural Language Processing so Hard?

Ambiguity



“At last, a computer that understands you like your mother”



1. (*) It understands you as well as your mother understands you



2. It understands (that) you like your mother



3. It understands you as well as it understands your mother



1 and 3: Does this mean well, or poorly?

Ambiguity At
the acoustic
level (speech
recognition)

I'm eight or duck

Eye maid; her duck

I maid her duck

I'm aid her duck

I'm ate her duck

I'm ate or duck

Ambiguity at the semantic (meaning) level



Two definitions of “mother”



I a woman who has given birth to
a child



I a stringy slimy substance consisting
of yeast cells and bacteria; is added
to cider or wine to produce vinegar



This is an instance of word sense
ambiguity

More Word
Sense
Ambiguity
semantic
(meaning)
level




I They put money in the
bank = = buried in mud?



I saw her duck with
a telescope

Book

- Jurafsky and Martin:
 - Speech and Language Processing
 - <https://web.stanford.edu/~jurafsky/slp3/>



NLP Libraries in Python

NLTK (Classical NLP)

Spacy (Classical NLP)

Textblob (Classical NLP)

TensorFlow (Deep learning, supported by Google)

Chainer (Deep learning, is led by Japanese venture company in partnership with IBM, Intel, Microsoft, and Nvidia)

Google tools <https://ai.google/tools/>

Facebook tools <https://ai.facebook.com/tools/>

Thank You

Let's jump into Codes

