MOOC Kapita Selekta Sistem Cerdas
Semester Genap 2021

The World with Language Technologies

Meisyarah Dwiastuti
NLP engineer @ Prosa.ai

March 1st, 2021

The Speaker

Meisyarah Dwiastuti

meisyarah.dwiastuti@gmail.com

Education background:

2012 - 2017	Computer Science at Universitas Gadjah Mada, Indonesia
2017 - 2019	Computational Linguistics at Saarland University, Germany
	Computer science/computational linguistics at Charles University, Czech Rep.

Research & Professional

2018	Research Assistant @ Dept. of Language Science & Technology, Saarland Uni.
2019 - 2020	Research Assistant @ German Research Center for AI (DFKI)
2020 - now	NLP Engineer @ Prosa.ai

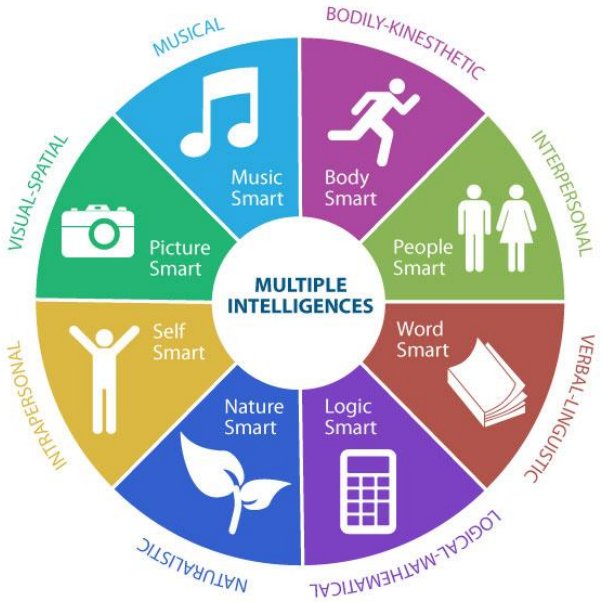
Research interest: natural language processing, machine learning
Also likes: running, traveling, books, and coffee

Gaining linguistic capabilities with technologies

- What are the capabilities?
- What are the technologies?
- Where are we now?
- What are the challenges?
- What can be in the future?

Towards artificial intelligence systems

- ***Artificial intelligence*** is an imitation of human intelligence demonstrated by machines to solve some problem
- One of the intelligences we would like to achieve is linguistic intelligence
  - how human use languages (natural languages)
  - both in written and spoken settings.

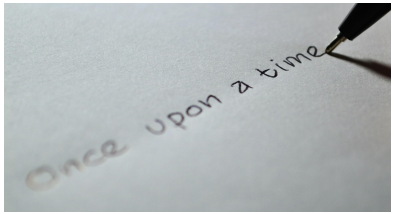


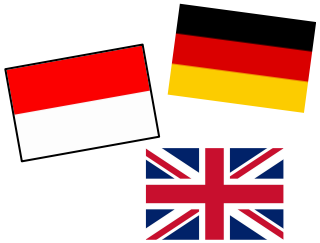
Human intelligence category according to Howard Gardner

What are the capabilities?

Some examples are:

- Producing a speech
- Writing a story
- Responding to a question
- Speaking more than one language
- Giving some recommendation
- Understanding analogy
- Having conversation
- etc

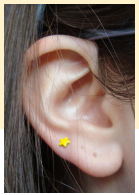




Human communication (spoken)

Pemrosesan bahasa

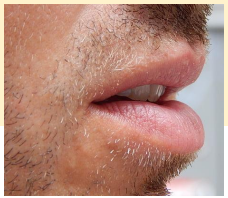
menerima sinyal suara



menyampaikan informasi

memproses informasi

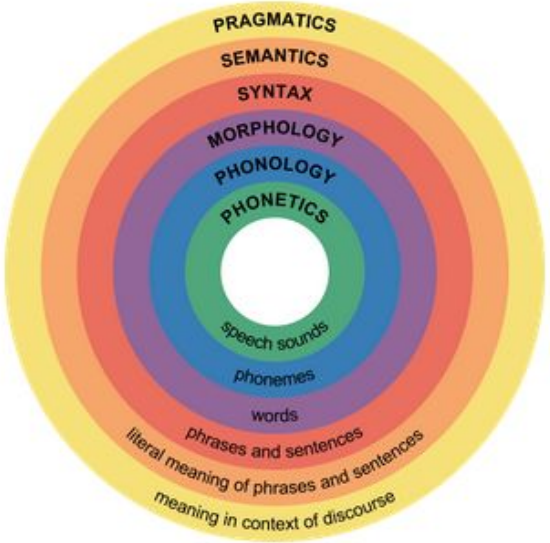
menyampaikan informasi



mengutarakan dalam sinyal suara

Language and Linguistics

- **Language** is the ability to produce and comprehend spoken and written words
- **Linguistics** is the study of language
- Linguistic level:



Source: lumenlearning.com

Can we make technologies to have such capabilities?

Actually, we have used a lot of language technologies.

Language Technologies

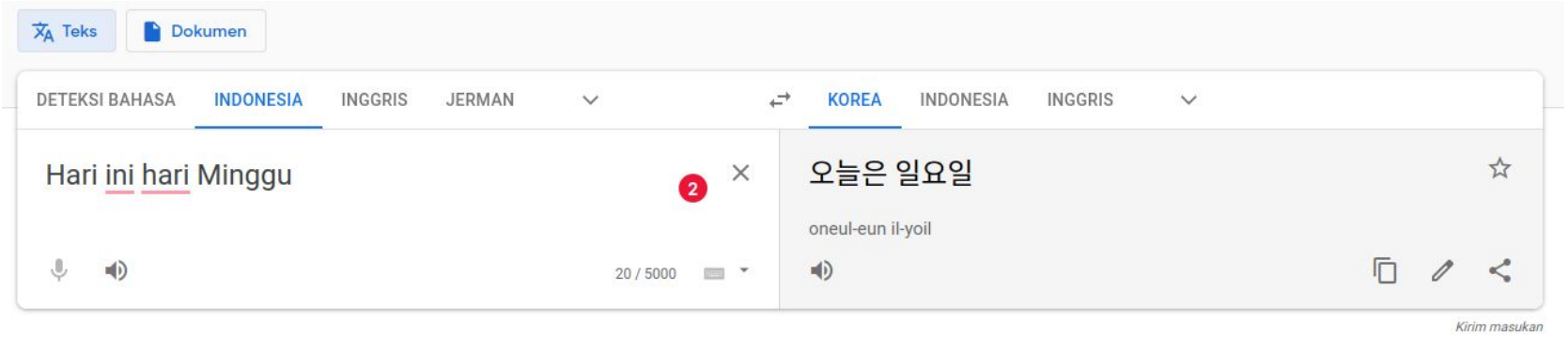
- Spam detection

Language Technologies

- Keyword suggestion

Language Technologies

- Machine Translation



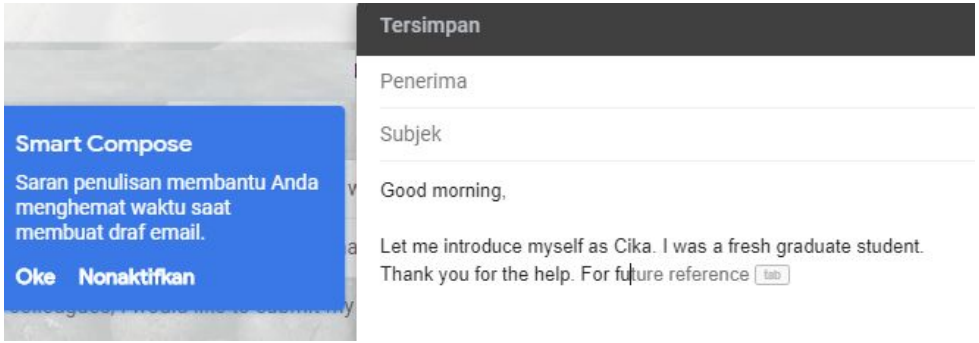
Language Technologies

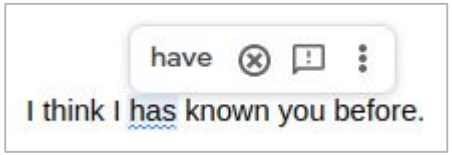
- Dialogue system

Chatbot Prosa.ai

Language Technologies

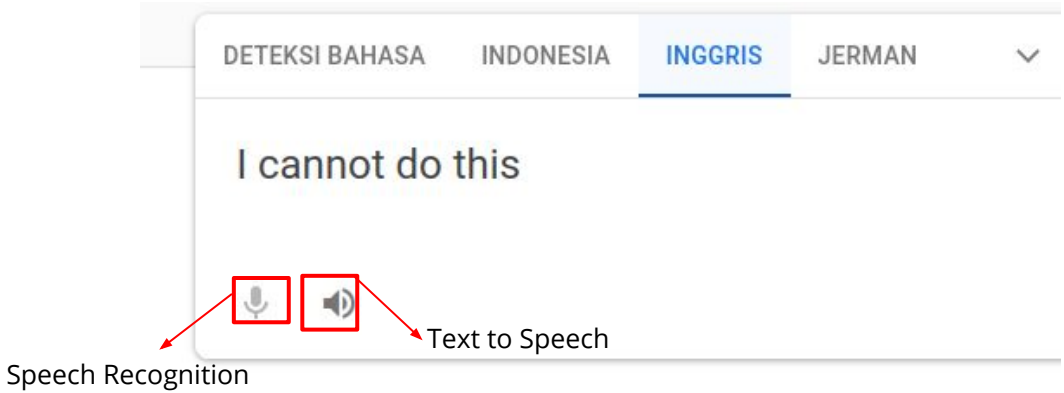
- Auto-complete, auto-correct, and spell-checker





Language Technologies

- Speech recognition (ASR) and Text-to-Speech (TTS)



In order to imitate the language capabilities, the technologies must run some *processing* on the language as input (both written and spoken).

Hence, the discipline learning how machine process natural languages is called **natural language processing (NLP)**.

NLP Tasks

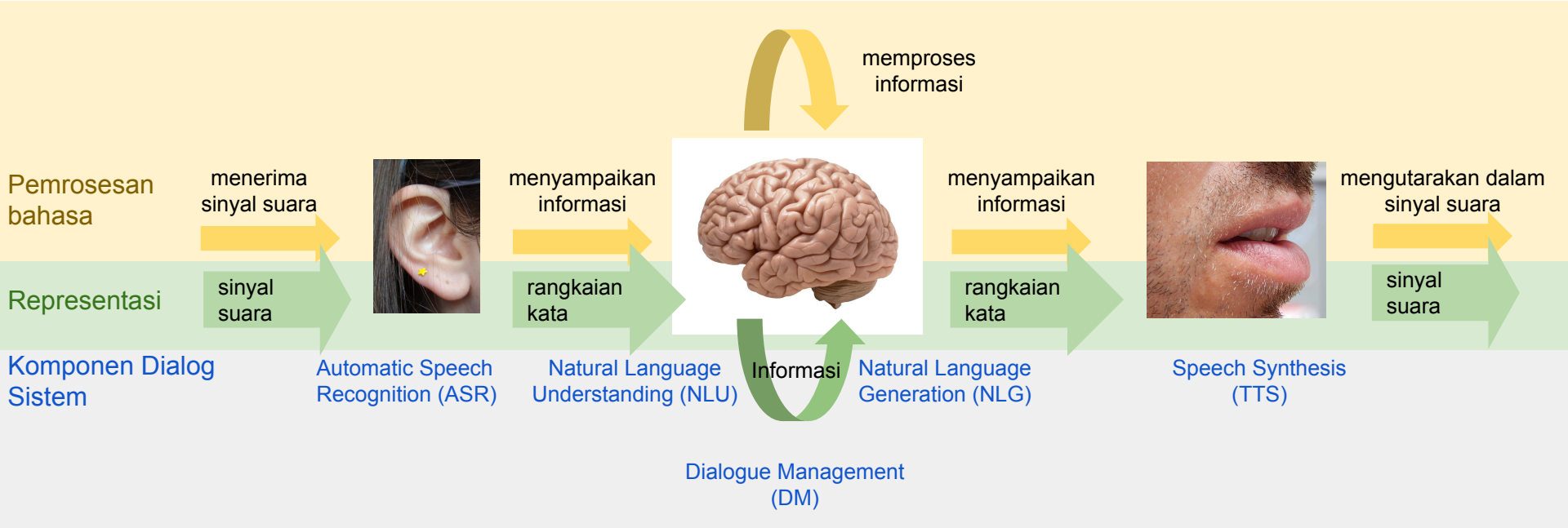
For the sake of simplicity, NLP tasks can be categorized into two types:

- Natural Language Understanding (NLU)
- Natural Language Generation (NLG)

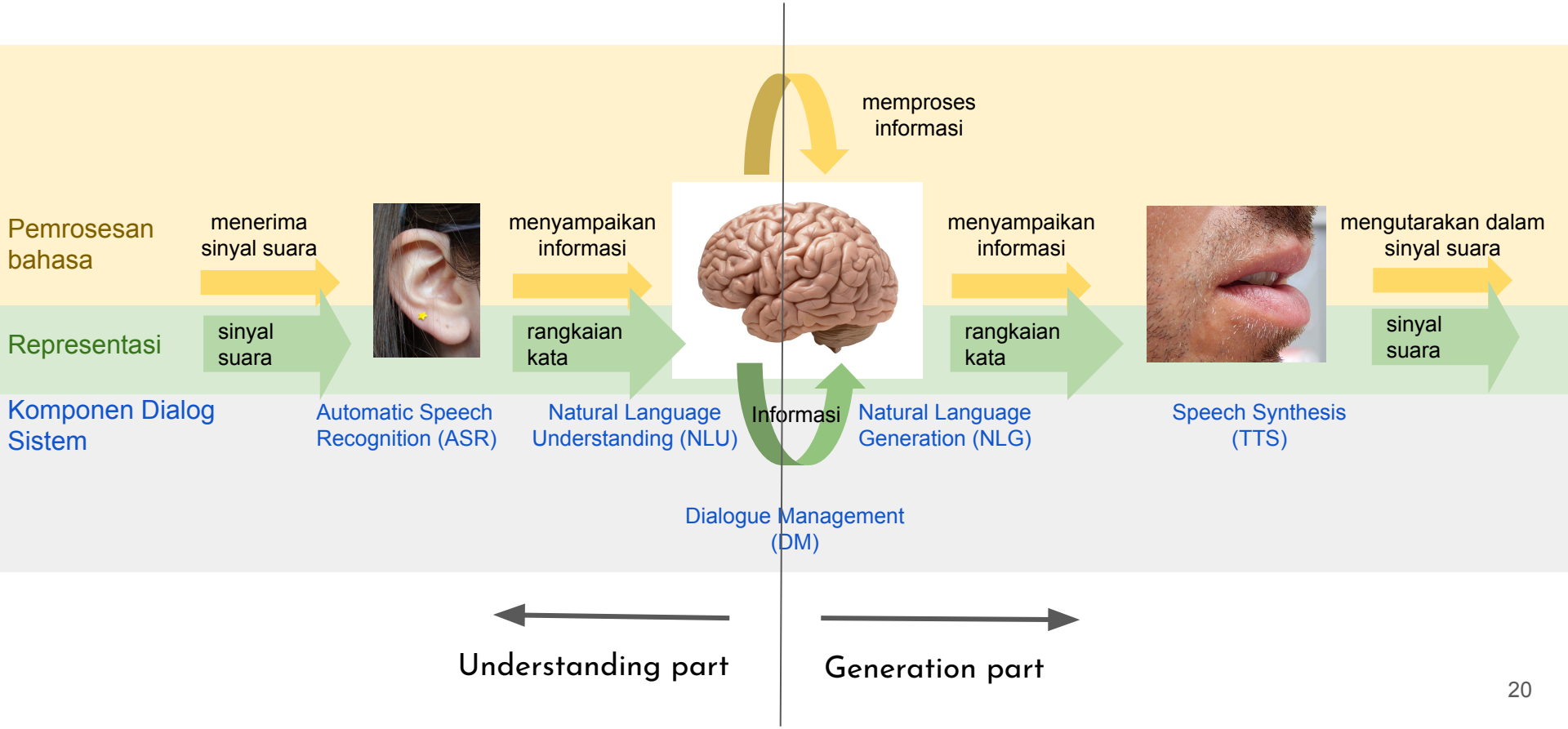
NLP Tasks

- We can be dealing with different **linguistic level**:
  - Phonetics/phonology → sound
  - Morphology → morphemes
  - Syntax → sentence structure
  - Semantics/Pragmatics → meaning
- or **input types**:
  - sound signal
  - text (our main focus)
  - others: picture, table, ontology, etc.

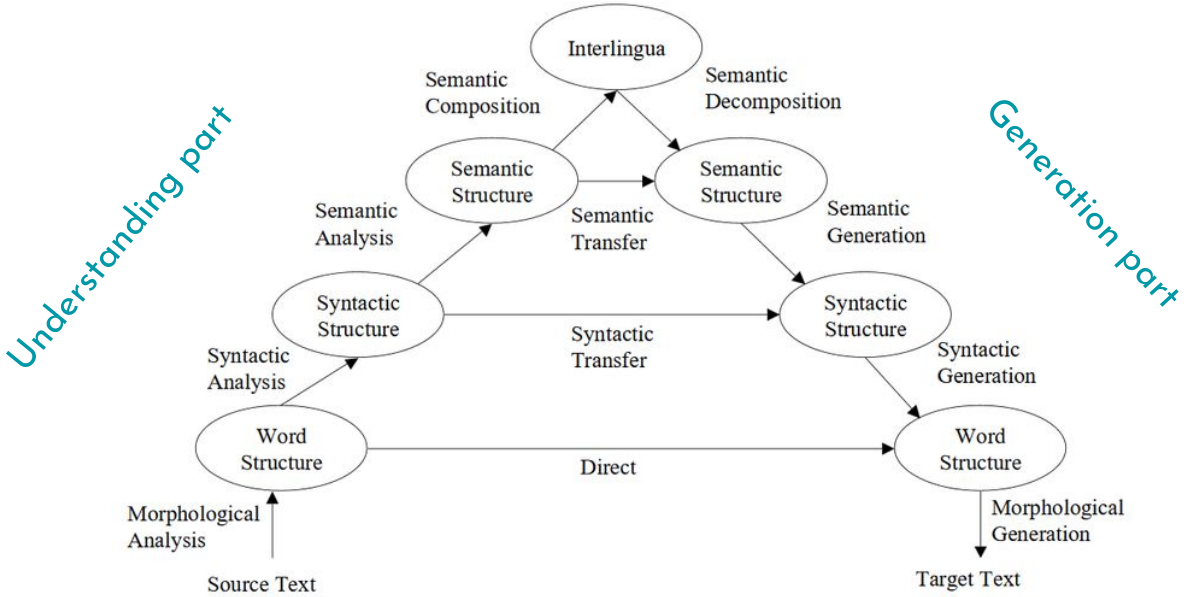
Example 1: Dialogue system



Example 1: Dialogue system



Example 2: Machine Translation



Vauquois triangle for rule-based machine translation systems

Example 3: Text Classification (Understanding)

- To assign some text to some fixed category
- We should have labeled text
- Examples:
  - News Classification: Sport, Health, Finance, Politics, Technology
  - Hate Speech Detection: Yes, No
  - Plagiarism Detection: Yes, No
  - Hoax News Detection: Yes, No
  - Sentiment Classification: Positive, Negative, Neutral

Example 4: Information Extraction (Understanding)

- To extract some structured information from a non/semi-structured information, e.g. texts
- Examples:
  - Template filling
  - Named entity recognition

Figure 1: An example of NER application on an example text

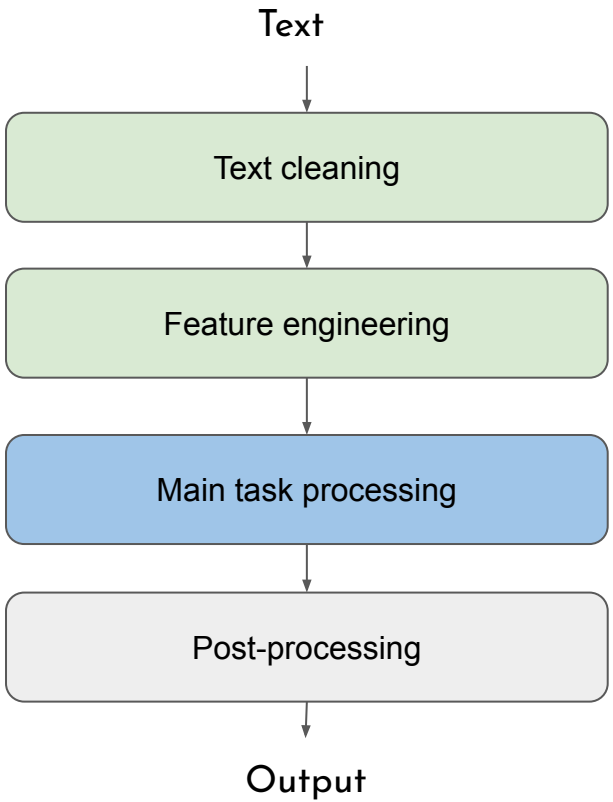
Source: inspiratron.org

Example 5: Text generation

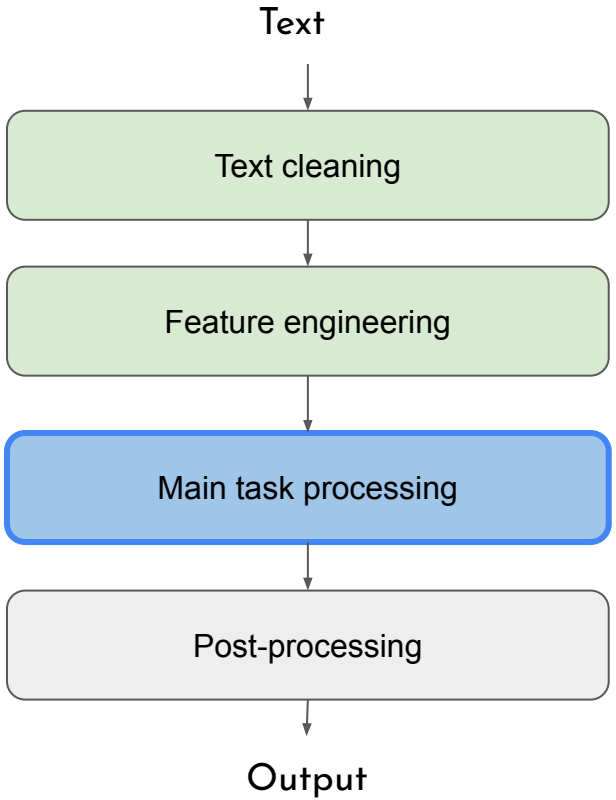
- To generate text given some input:
  - Text seed, table, image, etc..
- Examples:
  - Answer generation in QA system
  - Story/article generation:
    - e.g. http://ai-writer.com, https://notrealnews.net/

How do machines process the languages?

Text processing pipeline



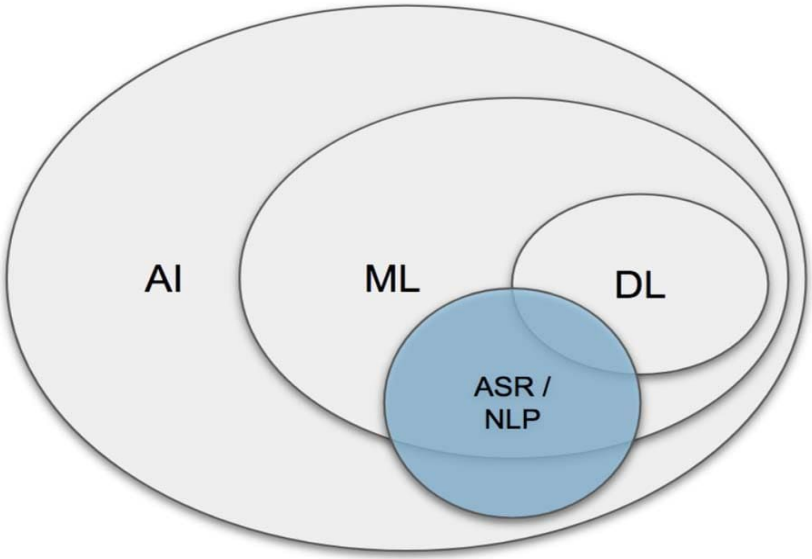
Text processing pipeline: main task processing



Common approaches:

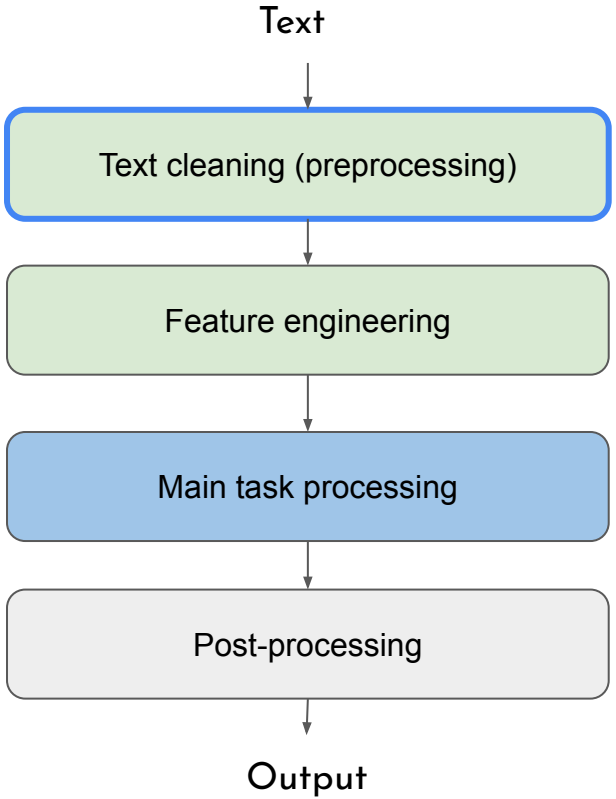
- Rule-based
  - Using handcrafted rules
- Statistical/machine learning (ML)
  - Data-driven
- Neural/deep learning (DL)
  - Also data-driven and various neural network architectures

Text processing pipeline: main task processing



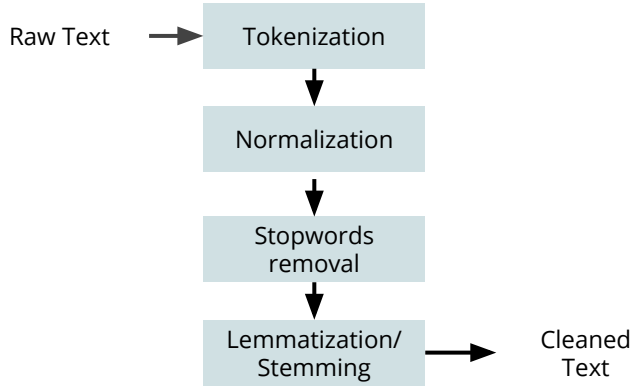
source: sonix.ai

Text processing pipeline: text cleaning

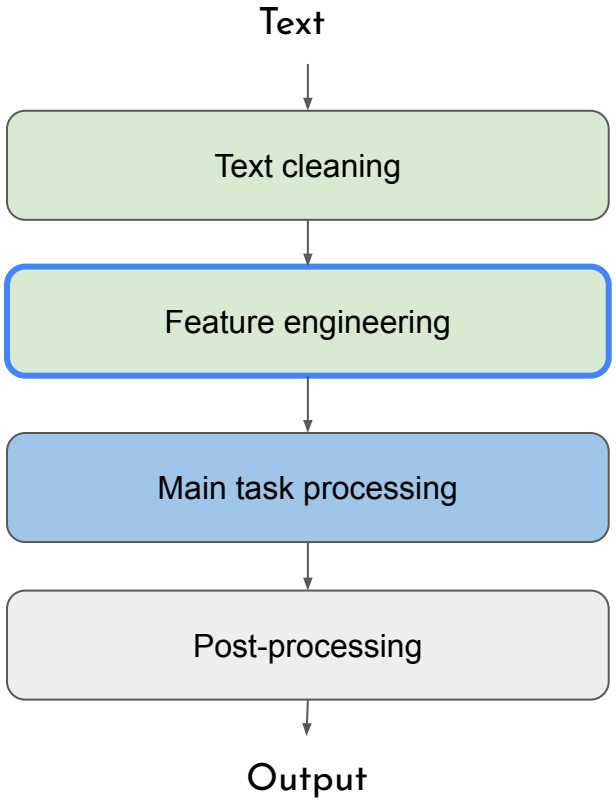


To clean the raw text, so it becomes a kind of text that is useful for our system.

Example pipeline:



Text processing pipeline: feature engineering

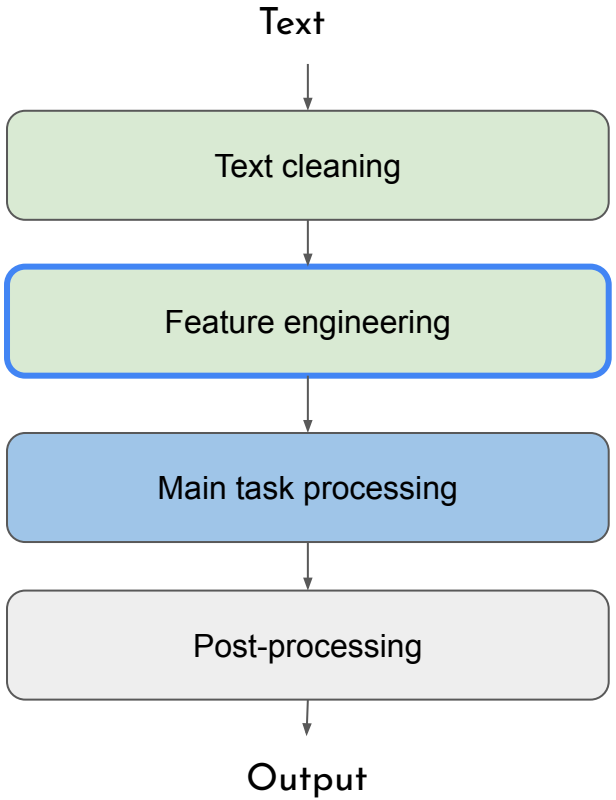


- **Feature** is a property of an object being observed.
- **Objective**: to gain extra information from the existing data that can helps our system solve the problem.
- It requires **understanding of the problem and data** to select appropriate features.

Examples:

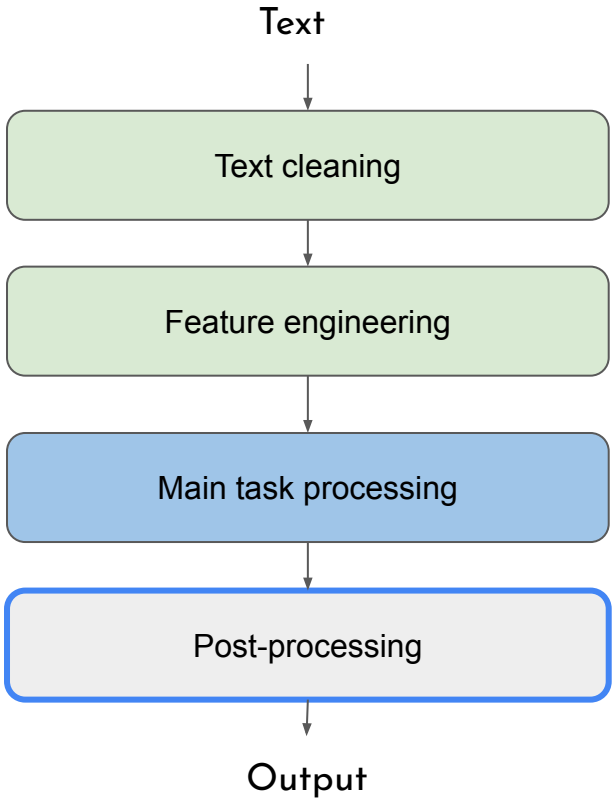
- Part-of-Speech (POS) tags
- Word position
- Word-before, word-after
- N-gram
- Lemmas
- Text length
- etc

Text processing pipeline: feature engineering



- We need to transform the features into numbers because machines can only read numbers
- Mostly used representation: vectors
- Approaches:
  - Label encoder
  - Bag-of-words
  - Term Frequency-Inverse Document Frequency (TF-IDF)
  - (word) embeddings
  - etc

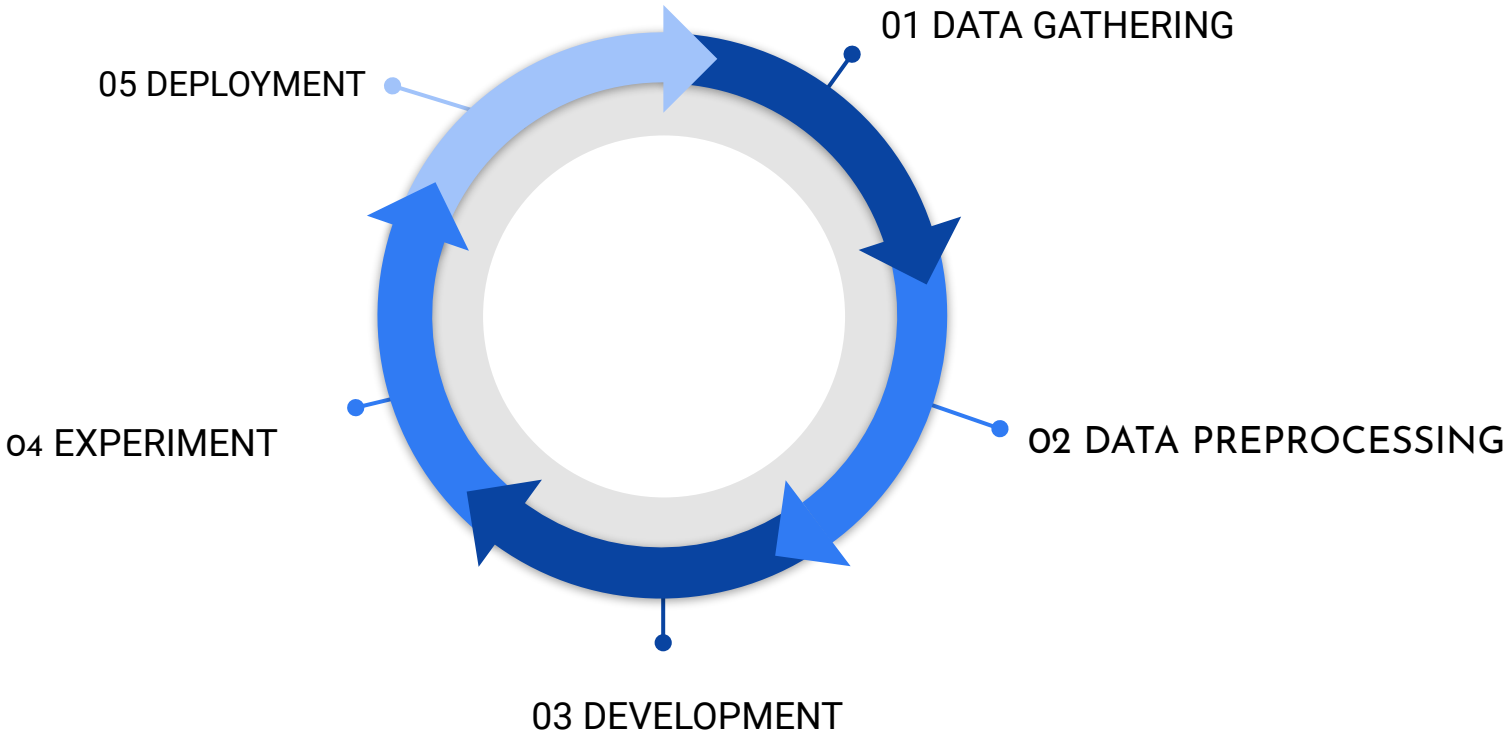
Text processing pipeline: post–processing



The main process may not be “perfect”

Optionally, we can post-process the output of our main approach to get results with better quality

It's not a one-time process



Evaluation

- Research:
  - Performance on test dataset
  - vs baseline system or human participants
- Application:
  - Performance on real users
    - benchmark may vary, depending on the application goal
  - e.g. A/B Testing

At which state are we now?

AI in Sci–fi: what people expect

AI in reality: what people encounter

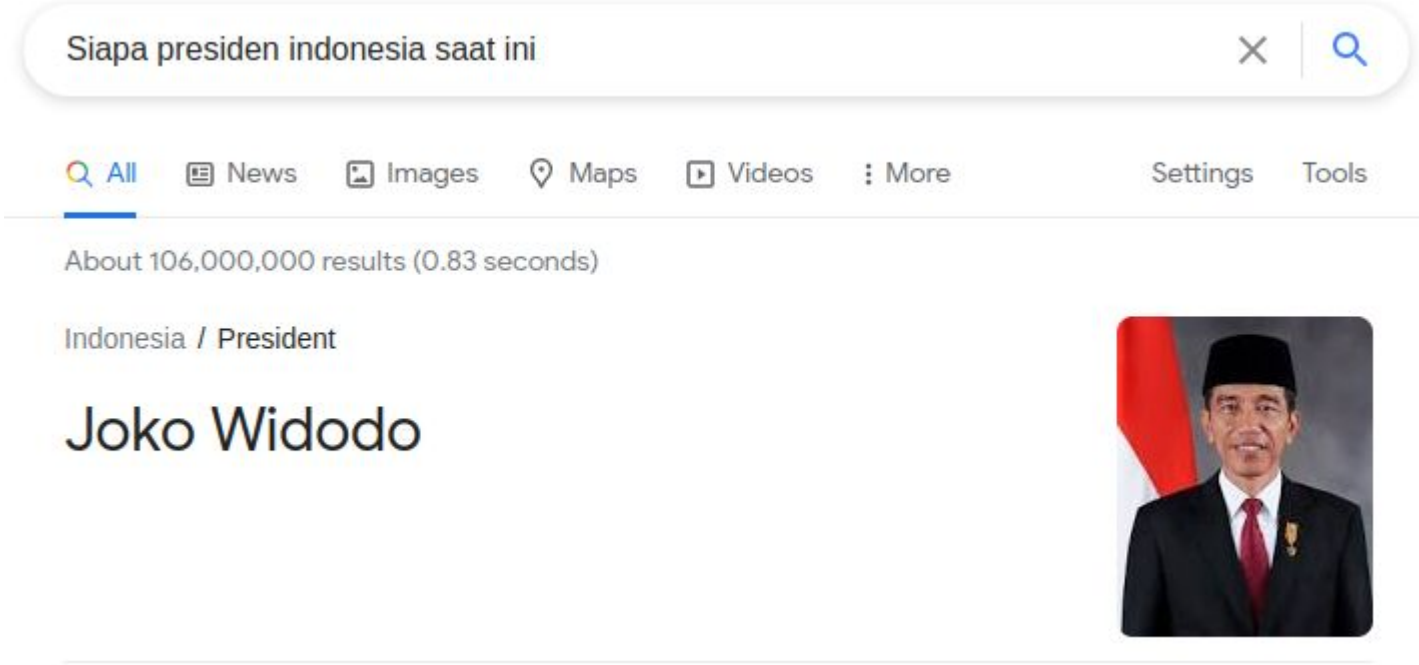
Capturing “meaning” is not easy

Language Technology Progress

Nice summary for various tasks: https://nlpprogress.com/

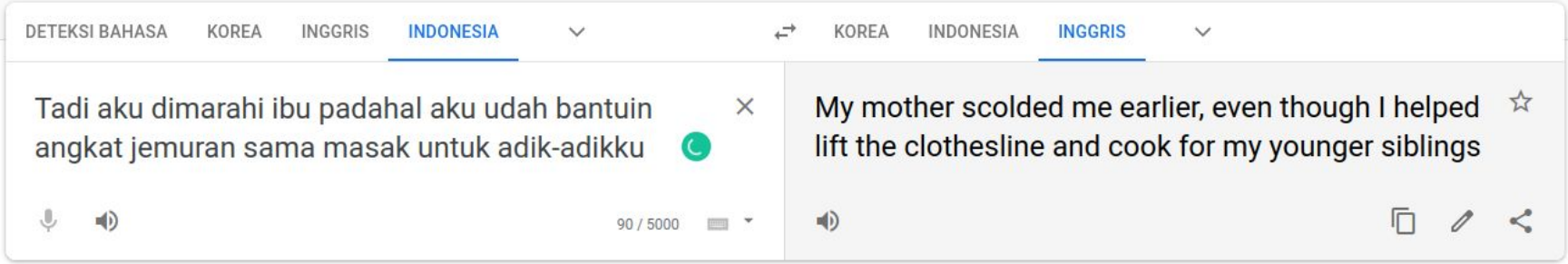
- Different **task** may have different performance
- Different **language** may have different performance
  - English is still the most heavily researched language
- Different **domain** may have different performance
  - E.g. News, sport, legal, health, formal language

Example: Question Answering and Information Retrieval



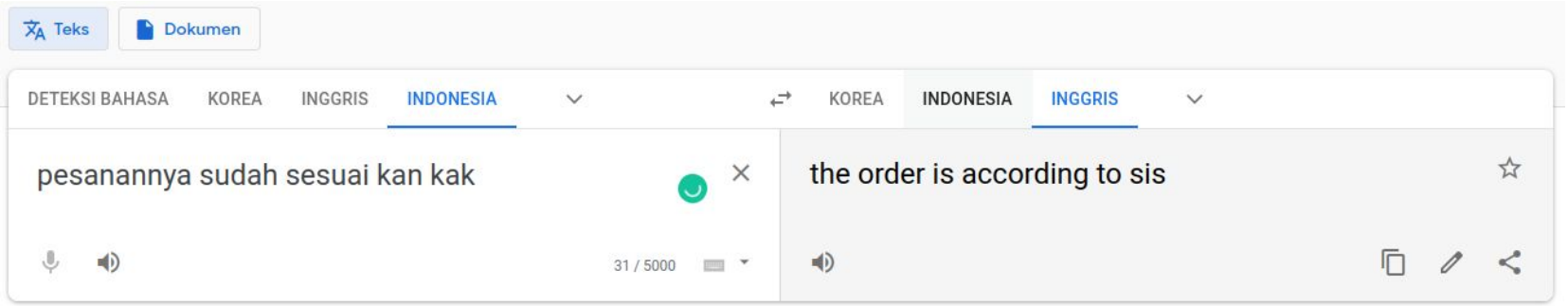
Accessed on March 1st, 2021

Example: Machine translation



Accessed on March 1st, 2021

Example: Machine translation



Accessed on March 1st, 2021

There are many applications

Only for **dialogue systems**:

- Smart Speakers
  - Siri, Alexa, Google Assistant
- Telephone
- Computer games
- Chatbots
- Assistive technologies
  - Therapy, elderly care, psychology consultation
- Built-in car dialogue system
- Research systems

Can we develop a new one?

There are still rooms for progress

Challenges in building language technologies

Ambiguity (at all level)

- Lexical (word)

  Kakak datang untuk memberi **tahu**.
- Syntactic

  Saya membaca **buku sejarah musik baru**.
- Semantic

  Budi berlibur dengan istrinya, **begitu juga Arif**.
- Pragmatic
  - **Jam berapa sekarang?**

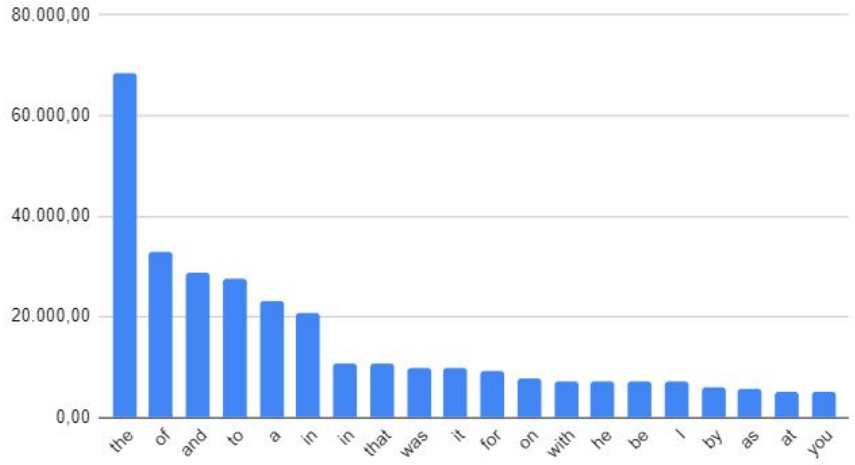
    Bisa memiliki dua makna:

    - menanyakan jam (arti sebenarnya)

    - menyindir/marah karena orang lain terlambat

Language is creative and infinite

- There are so many words
  - Zipf law: The frequency of a word occurrence in inversely proportional to its rank



Top-20 words in English (from a corpus)

Language is creative and infinite

- New words inclusion
  - e.g. in KBBI: *daring*, *warganet*, *luring*, *pramusiwi*
- Slang words
  - e.g. mager, woles, baper, bucin
- Code-switching
  - ***Seriously,*** harganya mahal banget
  - Nanti ***nek misale*** Bapak udah nyampe, kamu kabari aku ya
- One may produce a sentence that no one has ever produced
  - Factors: length, choose of words, etc

Data-driven methods NEED data

- Quality vs quantity
  - Should achieve both
  - Internet is a good resource but there is so much noise
- This is why systems using low-resource language struggles

Ethical use of language technologies

- Data collection
- Data privacy in delivery
- The result of language generation:
  - Hate speech, false news, racism, harassment
  - Plagiarism
  - Killing creative industry?

Language Technologies in the future

The research is still going on ...

- Inclusion of other languages (i.e. not just English)
- More application in different domains:
  - The use in healthcare is getting more attention
  - Open-domain understanding system is still unsolved
- Multimodality: combining multiple “intelligence”
  - Text + sound + images

Thank You.