AIFE: UNIT-3
MS. SHWETA TIWARI
Published: DECEMBER, 2021

"Artificial Intelligence for Engineering/Engineers (KMC-201)"_

UNIT-3: NATURAL LANGUAGE PROCESSING

FALL SEMESTER, YEAR (2ND, 1ST)

FALL SESSION (2021-22)
(AIFE)
MS. SHWETA TIWARI
Published: DECEMBER, 2021

Rajkiya Engineering College | Ambedkar Nagar, UP, India

Faculty Name: Ms. Shweta Tiwari, Subject: Artificial Intelligence for Engineering (KMC201) Year- 1st Year, Semester- 2nd Sem, Branch- CE, Session- Even Semester (2021-22)

PREPARED FOR
Engineering Students
All Engineering College

PREPARED BY
SHWETA TIWARI
Guest Faculty

December, 2021

Notes Part-1
MS. SHWETA TIWARI

UNIT-3

“Artificial Intelligence for Engineering/Engineers (KMC-201)”

By SHWETA TIWARI

Artificial Intelligence for Engineering/Engineers (KMC-201)”

UNIT-3: NATURAL LANGUAGE PROCESSING

December, 2021

Notes Part-1

MS. SHWETA TIWARI

UNIT-III

UNIT 3: NATURAL LANGUAGE PROCESSING

December, 2021

Notes Part-1

MS. SHWETA TIWARI

Natural Language Processing

Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English. Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to hear decision from a dialogue based clinical expert system, etc.

The field of NLP involves making computers to perform useful tasks with the natural languages humans use. The input and output of an NLP system can be –

- Speech
- Written Text

Components of NLP

There are two components of NLP as given

Natural Language Understanding (NLU) Understanding involves the following tasks –

- Mapping the given input in natural language into useful representations.
- Analyzing different aspects of the language.

Natural Language Generation (NLG)

It is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation.

It involves –

- **Text planning** – It includes retrieving the relevant content from knowledge base.
- **Sentence planning** – It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
- **Text Realization** – It is mapping sentence plan into sentence structure.

The NLU is harder than NLG.

Difficulties in NLU

NL has an extremely rich form and structure.

It is very ambiguous. There can be different levels of ambiguity –

- **Lexical ambiguity** – It is at very primitive level such as word-level.
- For example, treating the word "board" as noun or verb?
- **Syntax Level ambiguity** – A sentence can be parsed in different ways.
- For example, "He lifted the beetle with red cap." – Did he use cap to lift the beetle or he lifted a beetle that had red cap?
- **Referential ambiguity** – Referring to something using pronouns. For example, Rima went to Gauri. She said, "I am tired." – Exactly who is tired?
- One input can mean different meanings.
- Many inputs can mean the same thing.

NLP Terminology

- **Phonology** – It is study of organizing sound systematically.
- **Morphology** – It is a study of construction of words from primitive meaningful units.
- **Morpheme** – It is primitive unit of meaning in a language.
- **Syntax** – It refers to arranging words to make a sentence. It also involves determining the structural role of words in the sentence and in phrases.
- **Semantics** – It is concerned with the meaning of words and how to combine words into

meaningful phrases and sentences.

- **Pragmatics** – It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.
- **Discourse** – It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.
- **World Knowledge** – It includes the general knowledge about the world.

Steps in NLP

There are general five steps –

- **Lexical Analysis** – It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

- **Syntactic Analysis (Parsing)** – It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.

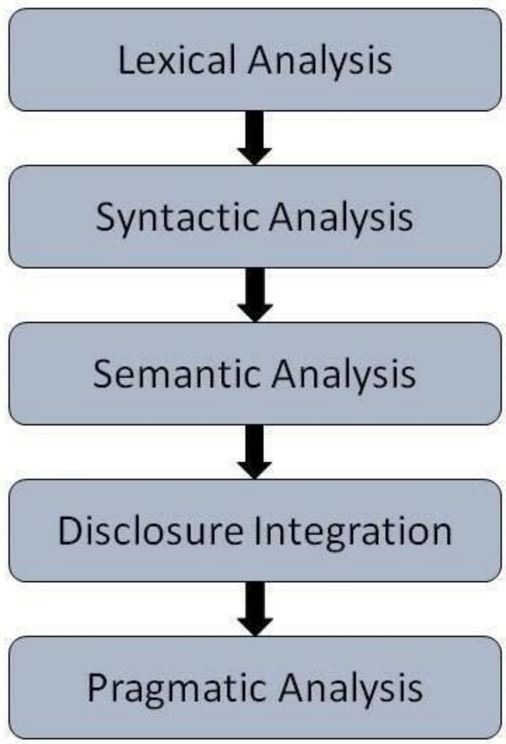


Figure 3.1: NLP Steps

- **Semantic Analysis** – It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as "hot ice-cream".
- **Discourse Integration** – The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

- **Pragmatic Analysis** − During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

3.1Speech Recognition

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, is a capability which enables a program to process human speech into a written format. While it’s commonly confused with voice recognition, speech recognition focuses on the translation of speech from a verbal format to a text one whereas voice recognition just seeks to identify an individual user's voice.

Many speech recognition applications and devices are available, but the more advanced solutions use AI and machine learning. They integrate grammar, syntax, structure, and composition of audio and voice signals to understand and process human speech. Ideally, they learn as they go — evolving responses with each interaction.

The best kind of systems also allow organizations to customize and adapt the technology to their specific requirements — everything from language and nuances of speech to brand recognition. For example:

- **Language weighting:** Improve precision by weighting specific words that are spoken frequently (such as product names or industry jargon), beyond terms already in the base vocabulary.
- **Speaker labeling:** Output a transcription that cites or tags each speaker's contributions to a multi-participant conversation.
- **Acoustics training:** Attend to the acoustical side of the business. Train the system to adapt to an acoustic environment (like the ambient noise in a call center) and speaker styles (like voice pitch, volume and pace).
- **Profanity filtering:** Use filters to identify certain words or phrases and sanitize speech output.

Meanwhile, speech recognition continues to advance. Companies, like IBM, are making inroads in several areas, the better to improve human and machine interaction.

Speech recognition algorithms

The vagaries of human speech have made development challenging. It's considered to be one of the most complex areas of computer science – involving linguistics, mathematics and statistics. Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models to determine the appropriate output.

Speech recognition technology is evaluated on its accuracy rate, i.e. word error rate (WER), and speed. A number of factors can impact word error rate, such as pronunciation, accent, pitch, volume, and background noise. Reaching human parity – meaning an error rate on par with that of two humans speaking – has long been the goal of speech recognition systems. Research from Lippmann (link resides outside IBM) estimates the word error rate to be around 4 percent, but it's been difficult to replicate the results from this paper.

Read more on how IBM has made strides in this respect, achieving industry records in the field of speech recognition.

Various algorithms and computation techniques are used to recognize speech into text and improve the accuracy of transcription. Below are brief explanations of some of the most commonly used methods:

- **Natural language processing (NLP):** While NLP isn't necessarily a specific algorithm used in speech recognition, it is the area of artificial intelligence which focuses on the interaction between humans and machines through language through speech and text. Many mobile devices incorporate speech recognition into their systems to conduct voice search—

  e.g. Siri—or provide more accessibility around texting.
- **Hidden markov models (HMM):** Hidden Markov Models build on the Markov chain model, which stipulates that the probability of a given state hinges on the current state, not its prior states. While a Markov chain model is useful for observable events, such as text inputs, hidden markov models allow us to incorporate hidden events, such as

part-of-speech tags, into a probabilistic model. They are utilized as sequence models within speech recognition, assigning labels to each unit—i.e. words, syllables, sentences, etc.—in the sequence. These labels create a mapping with the provided input, allowing it to determine the most appropriate label sequence.

- **N-grams:** This is the simplest type of language model (LM), which assigns probabilities to sentences or phrases. An N-gram is sequence of N-words. For example, “order the pizza” is a trigram or 3-gram and “please order the pizza” is a 4-gram. Grammar and the probability of certain word sequences are used to improve recognition and accuracy.
- **Neural networks:** Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold) and an output. If

that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. While neural networks tend to be more accurate and can accept more data, this comes at a performance efficiency cost as they tend to be slower to train compared to traditional language models.

- **Speaker Diarization (SD):** Speaker diarization algorithms identify and segment speech by speaker identity. This helps programs better distinguish individuals in a conversation and is frequently applied at call centers distinguishing customers and sales agents.

3.2. Natural language understanding

Natural language understanding (NLU) is a sub-topic of natural language processing, which involves breaking down the human language into a machine-readable format. Interesting applications include text categorization, machine translation, and question answering.

NLU uses grammatical rules and common syntax to understand the overall context and meaning of "natural language," beyond literal definitions. Its goal is to understand written or spoken language the way a human would.

NLU is used in natural language processing (NLP) tasks like topic classification, language detection, and sentiment analysis:

- Sentiment analysis automatically interprets emotions within a text and categorizes them as positive, negative, or neutral. By quickly understanding, processing, and analyzing thousands of online conversations, sentiment analysis tools can deliver valuable insights about how customers view your brand and products.

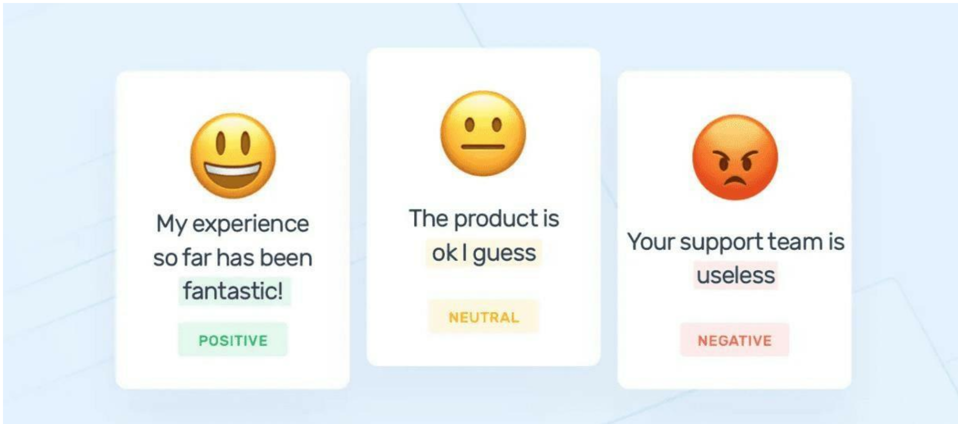


Figure 3.2: Natural language understanding

- Language detection automatically understands the language of written text. An essential tool to help businesses route tickets to the correct local teams, avoid wasting time passing tickets from one customer agent to the next, and respond to customer issues faster.
- Topic classification is able to understand natural language to automatically sort texts into predefined groups or topics. Software company Atlassian, for example, uses the tags Reliability, Usability, _and_ Functionality_ to sort incoming customer support tickets, enabling them to deal with customer issues efficiently.

NLP Vs NLU: What's The Difference?

Natural language understanding is a subfield of natural language processing.

While both NLP and NLU aim to make sense of unstructured data, but they are not the same thing. NLP is concerned with how computers are programmed to process language and facilitate "natural" back-and-forth communication between computers and humans. Natural language understanding, on the other hand, focuses on a machine's ability to understand the human language. NLU refers to how unstructured data is rearranged so that machines may "understand" and analyze it.

Look at it this way. Before a computer can **process** unstructured text into a machine-readable format, first machines need to **understand** the peculiarities of the human language.

Natural Language Understanding Examples

Machine Translation (MT)

Accurately translating text or speech from one language to another is one of the toughest challenges of natural language processing and natural language understanding.

Using complex algorithms that rely on linguistic rules and AI machine training, Google Translate, Microsoft Translator, and Facebook Translation have become leaders in the field of “generic” language translation.

You can type text or upload whole documents and receive translations in dozens of languages using machine translation tools. Google Translate even includes optical character recognition (OCR) software, which allows machines to extract text from images, read and translate it.

Automated Reasoning

Automated reasoning is a subfield of cognitive science that is used to automatically prove mathematical theorems or make logical inferences about a medical diagnosis. It gives machines a form of reasoning or logic, and allows them to infer new facts by deduction.

Simply put, using previously gathered and analyzed information, computer programs are able to generate conclusions. For example, in medicine, machines can infer a diagnosis based on previous diagnoses using IF-THEN deduction rules.

Automatic Routing of Tickets

A useful business example of NLU is customer service automation. With text analysis solutions like MonkeyLearn, machines can understand the content of customer support tickets and route them to the correct departments without employees having to open every single ticket. Not only does this save customer support teams hundreds of hours, but it also helps them prioritize urgent tickets.

According to Zendesk, tech companies receive more than 2,600 customer support inquiries per month. Using NLU technology, you can sort unstructured data (email, social media, live chat, etc.) by topic, sentiment, and urgency (among others). These tickets can then be routed directly to the relevant agent and prioritized.

Question Answering

Question answering is a subfield of NLP and speech recognition that uses NLU to help computers automatically understand natural language questions. For example, here's a common question you might ask Google Assitant: "What's the weather like tomorrow?" NLP tools can split this question into topic (weather) and date (tomorrow), understand it and gather the most appropriate answer from unstructured collections of "natural language documents": online news reports, collected web pages, reference texts, etc

By default, virtual assistants tell you the weather for your current location, unless you specify a particular city. The goal of question answering is to give the user response in their natural language, rather than a list of text answers.

3.3. Natural language generation

Natural-language generation (NLG) is a software process that transforms structured data into natural language. It can be used to produce long form content for organizations to automate custom reports, as well as produce custom content for a web or mobile application. It can also be used to generate short blurbs of text in interactive conversations (a chatbot) which might even be read out by a text-to-speech system.

Automated NLG can be compared to the process humans use when they turn ideas into writing or speech. Psycholinguists prefer the term language production for this process, which can also be described in mathematical terms, or modeled in a computer for psychological research. NLG systems can also be compared to translators of artificial computer languages, such as decompilers or transpilers, which also produce human-readable code generated from an intermediate representation. Human languages tend to be considerably more complex and allow for much more ambiguity and variety of expression than programming languages, which makes NLG more challenging.

NLG may be viewed as the opposite of natural-language understanding (NLU): whereas in natural- language understanding, the system needs to disambiguate the input sentence to produce the machine representation language, in NLG the system needs to make decisions about how to put a concept into words. The practical considerations in building NLU vs. NLG systems are not symmetrical. NLU needs to deal with ambiguous or erroneous user input, whereas the ideas the system wants to express through NLG are generally known precisely. NLG needs to choose a specific, self-consistent textual representation from many potential representations, whereas NLU generally tries to produce a single, normalized representation of the idea expressed.

3.4. Chatbots

A chatbot is an artificial intelligence (AI) software that can simulate a conversation (or a chat) with a user in natural language through messaging applications, websites, mobile apps or through the telephone.

Why are chatbots important? A chatbot is often described as one of the most advanced and promising expressions of interaction between humans and machines. However, from a technological point of view, a chatbot only represents the natural evolution of a Question

Answering system leveraging Natural Language Processing (NLP). Formulating responses to questions in natural language is one of the most typical Examples of Natural Language Processing applied in various enterprises' end-use applications.

Behind the scenes: How a chatbot works

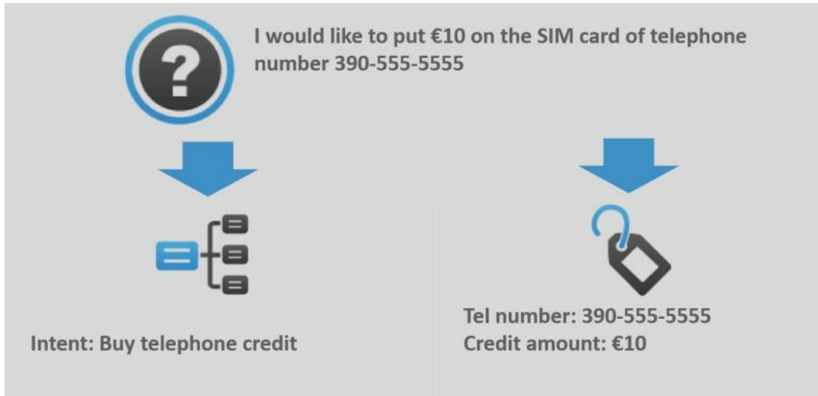
There are two different tasks at the core of a chatbot:

1) User request analysis

2) Returning the response

1) User request analysis: this is the first task that a chatbot performs. It analyzes the user's request to identify the user intent and to extract relevant entities.

The ability to identify the user's intent and extract data and relevant entities contained in the user's request is the first condition and the most relevant step at the core of a chatbot: If you are not able to correctly understand the user's request, you won't be able to provide the correct answer.



2) Returning the response: once the user's intent has been identified, the chatbot must provide the most appropriate response for the user's request. The answer may be:

- a generic and predefined text
- a text retrieved from a knowledge base that contains different answers
- a contextualized piece of information based on data the user has provided
- data stored in enterprise systems

- the result of an action that the chatbot performed by interacting with one or more backend application
- a disambiguating question that helps the chatbot to correctly understand the user's request

Why chatbots are important

Chatbot applications streamline interactions between people and services, enhancing customer experience. At the same time, they offer companies new opportunities to improve the customers engagement process and operational efficiency by reducing the typical cost of customer service.

To be successful, a chatbot solution should be able to effectively perform both of these tasks. Human support plays a key role here: Regardless of the kind of approach and the platform, human intervention is crucial in configuring, training and optimizing the chatbot system.

3.5. Machine Translation

Machine translation (MT) refers to fully automated software that can translate source content into target languages. Humans may use MT to help them render text and speech into another language, or the MT software may operate without human intervention.

MT tools are often used to translate vast amounts of information involving millions of words that could not possibly be translated the traditional way. The quality of MT output can vary considerably; MT systems require "training" in the desired domain and language pair to increase quality.

Translation companies use MT to augment productivity of their translators, cut costs, and provide post-editing services to clients. MT use by language service providers is growing quickly. In 2016, SDL—one of the largest translation companies in the world—announced it translates 20 times more content with MT than with human teams.

Customizable MT refers to MT software that has a basic component and can be trained to improve terminology accuracy in a chosen domain (medical, legal, IP, or a company's own preferred terminology). For example, WIPO's specialist MT engine translates patents more accurately than generalist MT engines, and eBay's solution can understand and render into other languages hundreds of abbreviations used in electronic commerce.

Adaptive MT offers suggestions to translators as they type in their CAT-tool, and learns from their input continuously in real time. Introduced by Lilt in 2016 and by SDL in 2017, adaptive MT is believed to improve translator productivity significantly and can challenge translation memory technology in the future.

There are over 100 providers of MT technologies. Some of them are strictly MT developers, others are translation firms and IT giants.

Examples of MT Providers		
Google Translate	Microsoft Translator / Bing	SDL BeGlobal
Yandex Translate	Amazon Web Services translator	Naver
IBM - Watson Language Translator	Automatic Trans	BABYLON
CCID TransTech Co.	CSLi	East Linden
Eleka Ingeniaritza Linguistikoa	GrammarSoft ApS	Iconic Translation Machines
K2E-PAT	KantanMT	Kodensha
Language Engineering Company	Lighthouse IP Group	Lingenio
Lingosail Technology Co.	LionBridge	Lucy Software / ULG
MorphoLogic / Globalese	Multilizer	NICT
Omniscien	Pangeanic	Precision Translation Tools (Slate)
Prompsit Language Engineering	PROMT	Raytheon

Reverso Softissimo	SkyCode	Smart Communications
Sovee	SyNTHEMA	SYSTRAN
tauyou	Tilde	Trident Software

Table 3.1:MT providers

MT Approaches

There are three main approaches to machine translation:

First-generation rule-based (RbMT) systems rely on countless algorithms based on the grammar, syntax, and phraseology of a language.

Statistical systems (SMT) arrived with search and big data. With lots of parallel texts becoming available, SMT developers learned to pattern-match reference texts to find translations that are statistically most likely to be suitable. These systems train faster than RbMT, provided there is enough existing language material to reference.

Neural MT (NMT) uses machine learning technology to teach software how to produce the best result. This process consumes large amounts of processing power, and that is why it's often run on graphics units of CPUs. NMT started gaining visibility in 2016. Many MT providers are now switching to this technology.

A combination of two different MT methods is called Hybrid MT.

Questions

1. What is Natural Language Processing? Discuss with some applications.
2. List any two real-life applications of Natural Language Processing.
3. What is Speech recognition
4. Explain the Natural language understanding and Natural language generation
5. Show the working of chatbots.
6. Analyze how statistical methods can be used in machine translation
7. Describe the different components of a typical conversational agent