Natural Language Processing

Chapter 1 An Overview of Natural Language Processing

Dr Raymond Lee
Associate Professor, DST
BNU-HKBU United International College

Natural Language Processing
Prologue

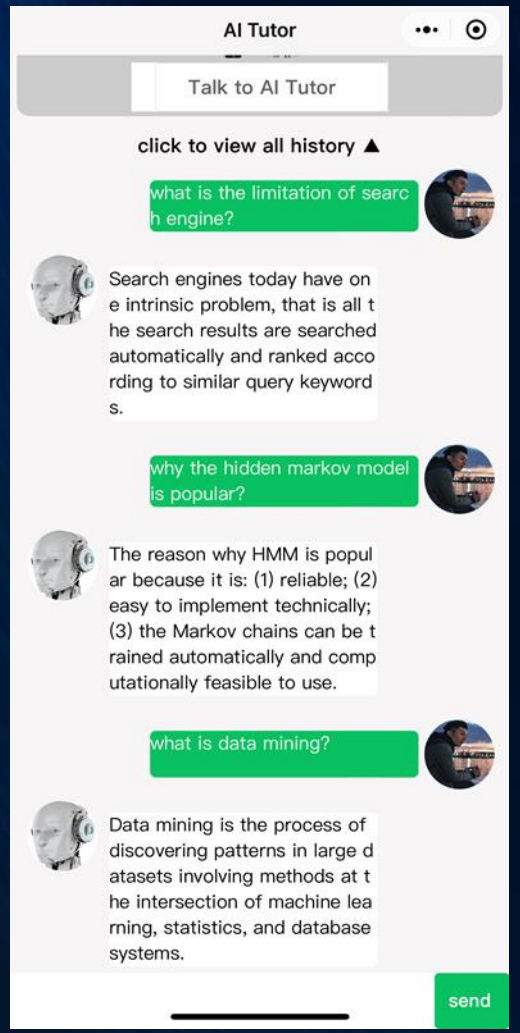

Fig. 1.1 A snapshot of AI Tutor Chatbot

(Consider this scenario: Late in the evening, John start a mobile apps and talk with AI Tutor Max)

Max: *Good evening, John. How are you?*

John: *Busy day. Just finished dinner after whole afternoon AI lecture.*

Max: *I see. Anything I can help?*

John: *Yes. Today Dr Lee has talked about data mining and search engine. Some concepts are not so clear, might need you help.*

Max: *Sure thing. Which part you are so clear?*

John: *What is the limitation of search engine?*

Max: *Search engines today have one intrinsic problem, that is all the search results are searched automatically and ranked according to similar query keywords.*

John: *Why Hidden Markov Model is popular?*

Max: *The reason why HMM is popular because it is: 1) reliable; 2) easy to implement technically ; and 3) the Marko chains can be trained automatically and computationally feasible to use*

John: *What is Data Mining?*

Max: *Data mining is the process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics and database systems.*

John: *I see. Thank you very much. Have a nice evening.*

Max: *You too, bye!*

- Is Max a human tutor? Maybe or maybe not. But in term of John, of course "he" is a very help tutor.
- Actually, these conversation is come from our latest research project so-called AI-Tutor 2.0. AI- Tutor 2.0 (Lee 2020a) is an AI chatbot that have learnt all the knowledge inside my latest book "AI in Daily Life" (Lee 2020b) and can use these knowledge to communicate with students in both Chinese and English conversations.
- You may wonder: how can we do that?
- In this chapter, we will introduce this fascinating technology by discussing the main components of NLP.
- You will soon find out that NLP technology is closely related different disciplines include linguistics, statistical engineering, machine learning, data mining, human voice processing, etc.
- You will also be surprised by the genius and efforts of these AI scientists and NLP engineers in the past 20 years to turn this important research topic into commercial products that can be used and assist us in many applications of our daily life.
- So, let's start our journey of Natural Language Processing with "Human Language and Intelligence".

Human Language and Intelligence

Fig. 1.2 The Turing Test

- "What you behave define who you are". I think that clause is very truth. Since we cannot (and should) "see" what people, the only cue to judge or evaluate a person whether s/he is good or bad, genius or dumb can only be achieved by observing his/her behaviours. And of course, the most direct way is what s/he said. That's why Alan Turing, the father of AI devised the famous Turing Test in 1950's as a way to judge whether a machine is intelligence or not, as shown in Fig. 1.2.
- In terms of AI perspective, the core technology of Turing Test in fact is NLP (Natural Language Processing), the technology to recognize and understand questions in human language, and how to response back to the judge also by human language – a ultimate challenge of NLP.
- In fact, "Human Language" is a crucial component in human civilization and one of the most fundamental aspect of our behavior.
- In general sense, it can be categorized into two main aspects: written and oral languages.
- For written-language, the main function is to store and pass our knowledge to other people of this generation and the future generations.
- For oral-language (spoken-language), the core function is to act as a medium for communication behind one person to other person in daily live and activities.
- In fact, language study is so important that different disciplines have their own focus and interpretation, and each discipline comes with its own set of language related problems to tackle and a set of solution to address those problems.
- Table 6.1 shows a summary of these language related disciplines and how can they solve the related problems.

Discipline	Problems to tackle with	Solutions & tools
Philosophy	What is meaning and knowledge? How do words and sentences acquire meaning? How can we relate ideas and concept into words and meanings?	Ontology and epistemology. Natural language argumentation using intuition. Mathematical models such as logic theory and model theory.
Psychology	How can we identify the structure of sentences? How the meaning of words can be identified? When does understanding take place?	Psychological experiments to measure the performance. Statistical analysis of observations.
Linguistics	How to form phrases and sentences with words? How can we represent the meaning of a sentence?	Mathematical model of language structure. Logical model for the representation of language structure and patterns.
Computational linguists & NLP	How to model different type of human languages? How to model knowledge and meaning? How to use human language for human-machine direct communication?	Agent ontology and ontological tree modelling. NLP techniques discussed in this chapter.

Levels of Linguistic in Human Language

- Levels of linguistic refers the functional analysis of any human language with including both the written and spoken languages.
- In terms of linguistic analysis, there are totally six levels of linguistics, which can be categorized into three main categories.
- Basic linguistic levels of "Sound"
  - Which include two levels of "Sound": Phonetics and Phonology sound levels.
  - These two levels related to the sound of the spoken language.
  - Phonetics is about the physical aspect of sounds, it studies the production and the perception of sounds, called phones. Phonetics deals with the production of speech sounds by humans, often without prior knowledge of the language being spoken.
  - Phonology is about the abstract aspect of sounds and it studies the phonemes. Phonology is about establishing what are the phonemes in a given language, i.e. those sounds that can bring a difference in meaning between two words.
  - For instance, the vowels in the English words "cool", "whose" and "moon" are all similar but slightly different. The different variants are dependent on the different contexts in which they occur.
- Intermediate linguistic levels of "Structure" – which refer to the basic structure of the language
  - Which include two levels of structure: Morphology and Syntax Levels of language structure.
  - Morphology is the level of forms and words. It is what one normally understands by grammar (along with syntax). The term morphology refers to the analysis of minimal forms in language which are, however, themselves comprised of sounds and which are used to construct words which have either a grammatical or a lexical function. Lexicology is concerned with the study of the lexicon from a formal point of view and is thus closely linked to (derivational) morphology.
  - Syntax is the level of clauses and sentences. It is concerned with the meanings of words in combination with each other to form phrases or sentences. In particular, it involves differences in meaning arrived at by changes in word order, the addition or subtraction of words from sentences or changes in the form of sentences. It furthermore deals with the relatedness of different sentence types and with the analysis of ambiguous sentences.
- Advanced linguistic levels of "Meaning" – which refer to the actual meaning of the language
  - Which include two levels of meaning: Semantics and Pragmatics Levels of language meaning.
  - Semantics is the area of meaning. It might be thought that semantics is covered by the areas of morphology and syntax, but it is quickly seen that this level needs to be studied on its own to have a proper perspective on meaning in language. Here one touches, however, on practically every other level of language as well as there exists lexical, grammatical, sentence and utterance meaning.
  - Pragmatics is with the use of language in specific situations. The meaning of sentences need not be the same in an abstract form and in practical use. In the latter case one speaks of utterance meaning. The area of pragmatics relies strongly for its analyses on the notion of speech act which is concerned with the actual performance of language. This involves the notion of proposition – roughly the content of a sentence – and the intent and effect of an utterance.

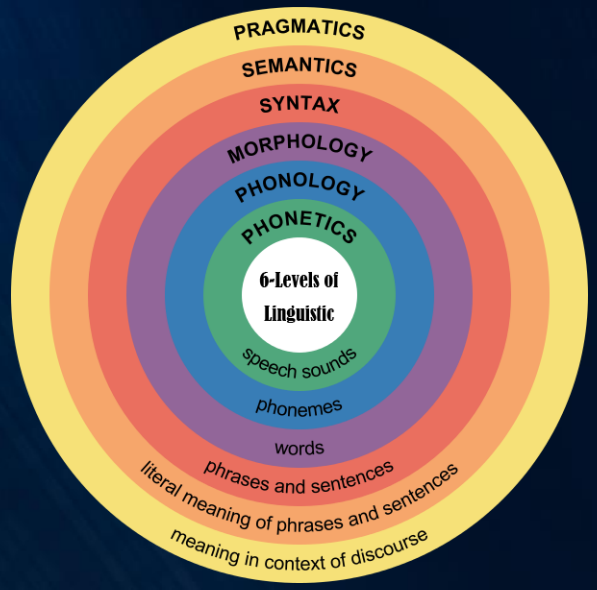

Fig. 1.3 Levels of Linguistic in Human Languages

Ambiguity in Human Language

Fig. 1.4 What do you meant by "bank"

- Ambiguity and Uncertainty in Language Ambiguity, generally used in natural language processing, can be referred as the ability of being understood in more than one way.
- In simple terms, we can say that ambiguity is the capability of being understood in more than one way. Natural language is very ambiguous. NLP has the following types of ambiguities:
  - Lexical ambiguity
    - The ambiguity of a single word is called lexical ambiguity.
    - For example, treating the word "silver" as a noun of metal, an adjective of silver colored, or a verb of process of silvering.
  - Syntactic ambiguity
    - This kind of ambiguity occurs when a sentence is parsed in different ways.
    - For example, the sentence "The man saw the girl with the telescope".
    - It is ambiguous whether the man saw the girl carrying a telescope or he saw her through his telescope.
  - Semantic ambiguity
    - This kind of ambiguity occurs when the meaning of the words themselves can be misinterpreted.
    - In other words, semantic ambiguity happens when a sentence contains an ambiguous word or phrase.
    - For example, the sentence "The van hits the boar while it is moving" is having semantic ambiguity because the interpretations can be "The van, while moving, hit the boar" and "The van hits the boar while the boar is moving".
  - Pragmatic ambiguity
    - Such kind of ambiguity refers to the situation where the context of a phrase gives it multiple interpretations.
    - In simple words, we can say that pragmatic ambiguity arises when the statement is not specific.
    - For example, the sentence "I like you too" can have multiple interpretations like I like you (just like you like me), I like you (just like someone else dose).
- Figure 6.3 shows a typical example of ambiguity in language in high-level of pragmatic meaning.
- When one says "I go to river bank this morning", the meaning of "river bank" one refers to should probably "riverbank" that close to a river, right? But if she says "I go to river bank this morning to take some cash". In that case, the second part of the sentence "to take some cash" gives us more cue about what "river bank" she refer to, maybe a bank called "river bank".

A Brief History of NLP

First Stage - Machine Translation on NLP (Before 1960's)

- The history of Natural Language Processing (NLP) can be traced back to 17th century when philosophers such as Leibniz and Descartes who proposed to use codes to relate words between different languages. Although such proposals only remained theoretical back then, they laid the ground for the development of language translation machine.
- The first invention patent related to translation machine was applied in the mid-1930s by Georges Artsrouni's proposal.
- However, the history of natural language processing (NLP) "officially" started from 1950's with Alan Turing for the publication of his famous article "Computing Machinery and Intelligence" and the propose of Turing Test to explore the intelligence of machine using NLP as judging criteria.
- At the time, NLP was mainly focused on the R&D of intelligent machine on language translation – Machine Translation.
- The first international conference on Machine Translation (MT) was held in 1952 and second was held in 1956, NLP still only focused on machine translation, mainly used simple rule-based methods and statistical techniques.
- In 1954, Georgetown-IBM experiment involved fully automatic translation of more than sixty Russian sentences into English. The inventors at that time were too optimistic to claim that the whole machine translation problem will be totally solved within 3 – 5 years. However, real progress was much slower than expectation.
- In 1957, The Noam Chomsky's Syntactic Structures helped revolutionized Linguistics with universal grammar.
- In 1961, the work presented in Teddington International Conference on Machine Translation of Languages and Applied Language analysis was the climax of this phase.
- However, after the release of ALPAC report in 1966, which revealed that ten-year-long research had failed to fulfill the original expectation of machine translation, funding all related research and projects was dramatically reduced.

Second Stage - Early AI on NLP (Late 1960's to 1970's)

- With the popularity of AI over this period of time, major development of NLP focused on how AI can be applied be applied to the exploration of knowledge, so-called "Ontology", and its role on the construction and manipulation of meaning representations.
- Typical example includes the development of BASEBALL system in late 1960's, a question-answering AI-based expert system was also developed. However, the input to this system was restricted and the language processing involved was a simple one.
- A more advanced NLP system was proposed by Minsky 1968 – a major founder of AI. As compared with the BASEBALL Q&A system, this NLP system employed AI based inference engine and knowledgebase for the interpretation of the Q&As.
- 1970: William A. Woods introduced the augmented transition network (ATN) to represent natural language input. Many programmers also began writing conceptual ontologies which structured real-world information into computer-understandable data in the 1970s.
- But because of these high expectation of AI and Expert Systems are not truly realized, both the US Government and Commercial started to withdraw further funding on AI & expert systems research which leads to the FIRST WINTER OF AI (1974 – 1980).

A Brief History of NLP

Third Stage – Grammatico-logical on NLP (1970's – 1980's)

- This phase can be described as the grammatico-logical phase. Due to the failure of practical system building in last phase, the researchers moved towards the use of logic for knowledge representation and reasoning in AI.
- The grammatico-logical approach, towards the end of decade, helped us with powerful general-purpose sentence processors like SRI's Core Language Engine and Discourse Representation Theory, which offered a means of tackling more extended discourse.
- In this phase we got some practical resources & tools like parsers, e.g. Alvey Natural Language Tools along with more operational and commercial systems, e.g. for database query.
- Although the R&D of the NLP developed in this period still bounded by the computational capacity of the computer system, the work on lexicon in 1980s provided a solid foundation in the direction of grammatico-logical approach of NLP for future development.

Fourth Stage – AI & Machine Learning (1980's – 2000's)

- Up to the 1980s, most natural language processing systems were based on complex sets of hand-written rules until 1980.
- With the re-birth of AI in 1980's coined by the success of J. Hopfield for his ground-breaking Hopfield Network in machine learning, a revolution in NLP began in the 1980s with the introduction of machine learning algorithms for language processing.
- Thanks for the acceleration for the advance of computer technology in terms of computational capacity and memory storage, together with the dominance of Chomskyan theories of linguistics, whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing.
- This stage can also be known as the stage of "Lexical & Corpus" on NLP. The phase had a lexicalized approach to grammar that appeared in late 1980s and became an increasing influence.
- In 2006, Watson, a question-answering computer system capable of answering questions posed in Natural language was developed in IMB's DeepQA project by a research team which was led by Principal Investigator David Ferrucci.

A Brief History of NLP

Fifth Stage – Rise of BERT, Transformer, ChatGPT, and LLMs (2000s – present)

- The advancement of NLP since the early 2000s has been largely driven by innovations in neural networks and deep learning architectures.
- Recurrent Neural Networks (RNNs), introduced in the early 2000s, addressed sequential data but struggled with long-term dependencies.
- Long Short-Term Memory (LSTM) networks improved on RNNs by incorporating a memory mechanism for better performance in tasks needing long-range context.
- The introduction of the Transformer architecture in 2017 marked a major breakthrough, utilizing a self-attention mechanism for improved speed and accuracy in NLP tasks.
- BERT (2018) utilized a bidirectional approach for context understanding, achieving state-of-the-art results in various NLP tasks.
- OpenAI's Generative Pre-trained Transformers (GPT) began with GPT-1 in 2018, which showcased significant unsupervised learning capabilities with 117 million parameters.
- GPT-2 (2019) expanded to 1.5 billion parameters, demonstrating the ability to generate coherent text but raised concerns about potential misuse.
- GPT-3 (2020) had 175 billion parameters, excelling in few-shot and zero-shot learning, capable of tasks that mimic human reasoning.
- ChatGPT, based on GPT-4 and released in 2023, enhanced coherence and reasoning, becoming a versatile tool for conversational AI and problem-solving.
- The growth of LLMs from GPT-1 to GPT-4 reflects exponential increases in parameters and computational power, raising ethical questions regarding AI's impact on human-like capabilities.

Natural Language Processing and AI

Fig 1.5 NLP and AI

What is Natural Language Processing (NLP)?

- Natural language processing (NLP) can be defined as the automatic (or semi-automatic) processing of human language.
- In some sense, the term `NLP' is sometimes used rather more narrowly than that, often excluding information retrieval and sometimes even excluding machine translation.
- Many computer scientists consider NLP as "computational linguistics". In some sense, it is rather true as in terms of computer science, NLP can be considered as a kind of "computer modelling" or "computerization" of linguistics, just like the term "computational finance" to refer to the computational modeling of finance theory.
- As said, NLP is a multidiscipline topics which involve extensive knowledge and basic concepts on linguistics and logic theory in theoretical mathematics.
- Nowadays NLP research also covers cognitive science, psychology and even philosophy in terms of epistemology and ontology.

NLP and AI

- In terms of AI, NLP is a field of AI in which computers analyze, understand, and derive meaning from human language in a smart and useful way.
- By utilizing NLP, AI developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.
- NLP is used to analyze text, allowing machines to understand how human's speak and response.
- This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more.
- NLP is commonly used for text mining, machine translation, and automated question answering.
- With the rapid growing of AI and computer technology, current research and implementation of NLP also involve various AI-based machine learning, data-mining, deep learning and agent ontology.

Main Components of NLP

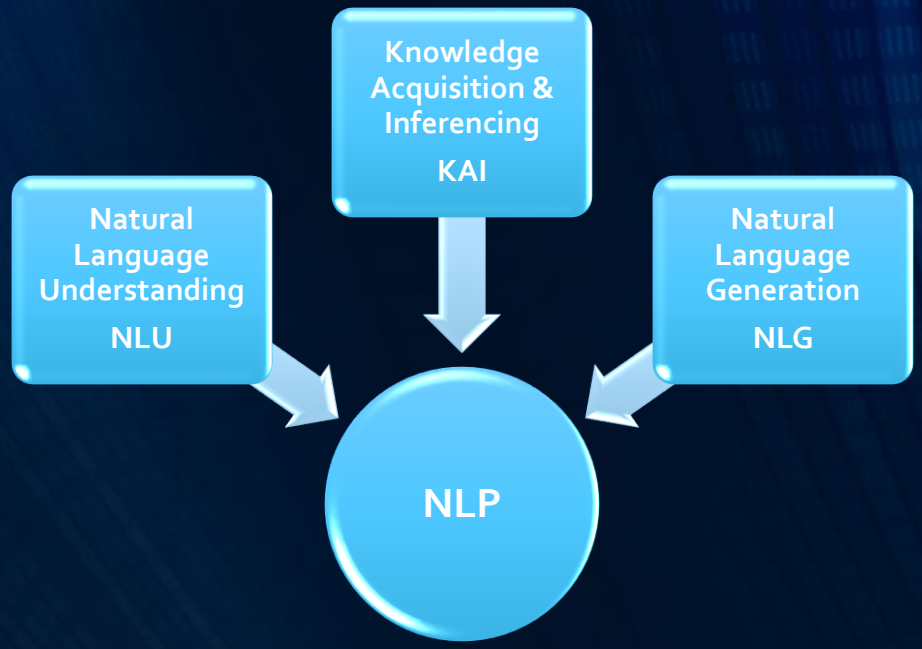

Fig. 1.6 Main components of NLP

- The main components of NLP consists of THREE main components (Figure 6.3):
  - Natural Language Understanding (NLU)

    NLU corresponds to all basic functions and operations of NLP from the voice recognition of human language to the 3 levels analysis of understanding the meaning of the spoken language: syntax, semantic and pragmatic analysis. Traditional NLP with restrict domain of application is only focus of this component.
  - Knowledge Acquisition and Inferencing (KAI)

    Once the spoken language is "clearly understood" from the NLU, KAI focus on the generation of the proper response and answer. In terms of machine learning and AI, is a kind of knowledge acquisition and inferencing problem. Traditionally, such task is achieved by rule-based system. That's the "If-then" kind of question and response, which is commonly used in many expert system. However, with the complexity of natural language and conversation, most rule-based system are failed to applied successfully. To solve this intrinsic problem, most KAI system will try to restrict the knowledge domain to certain area such as customer service knowledge on a particular industry (e.g. insurance, IT). With the advance of AI technology, new technology on agent ontology have been implement with certain success. We will discuss it in detail in the next chapter on Ontological-based Search Engine.
  - Natural Language Generation (NLG)

    NLG involve the generation of reply, response and feedback in the human-machine conversation. This consists of the process of: 1) formulation of the response into texts and sentences with the target language; 3) text-to-voice synthesis based on the target language to produce near-human voice response.

Natural Language Understanding (NLU)

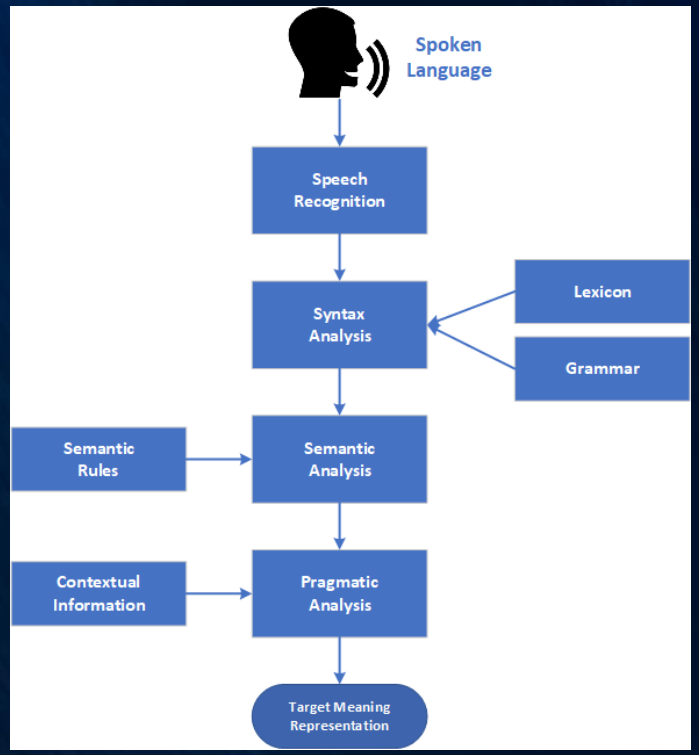

Fig. 1.7 Systematic diagram of NLU

Basically, NLU focus on the "recognition and understanding" of the spoken language.
It consists of 4 main processes: speech recognition, syntax analysis, semantic analysis and pragmatic analysis.
Figure 6.6 shows the systematic diagram of NLU.

Speech Recognition

- It is the first phase of NLP, which corresponds to the implementation of the phonetics, phonology and morphological processing of the spoken language mentioned in the linguistic model.
- The main purpose of this phase is to break chunks of spoken language input into sets of "tokens" which correspond to the paragraphs, sentences and words.
- Current speech recognition basically applies frequency spectrogram technology to extract different frequency of the spoken sounds for speech recognition.
- For example, a word like "uncertain" can be broken into two sub-word tokens as "un-certain".

Syntax Analysis

- It is the second phase of NLP, which directly corresponds to the first level for the analysis of the structural meaning of spoken sentence(s).
- The purpose of this phase is two folds: 1) to check that a sentence is well formed or not and 2) to break the spoken sentence(s) into a syntactic structure that can reflect the syntactic relationships between different words.
- For example, the sentence "The apple goes to the girl" would be rejected by syntax analyzer or parser.

Semantic Analysis

- It is the third phase of NLP, which directly corresponds to the second level for the analysis of the semantic meaning fot he spoken sentence(s).
- The purpose of this phase is to extract exact meaning, or one can say the meaning defined by the dictionary extracted from the text. In order words, the extracted text is checked for its meaningfulness (or rejected with meaningless).
- For example, semantic analyzer would reject a sentence like "Hot snowflakes".

Pragmatic Analysis

- It is the fourth phase of NLP and also the most difficult level of meaning analysis of the spoken sentence(s).
- Pragmatic analysis deals with outside word knowledge, which means knowledge that is external to the spoken sentence(s).
- Pragmatics analysis that focuses on what was described is reinterpreted by what it actually meant, deriving the various aspects of language that require real world knowledge.
- For example, the sentence "Will you crack open the door? I am getting hot."
- Semantically, the word "crack" would mean to break, but pragmatically we know that the speaker means to open the door just a little to let in some air.

Speech Recognition
The Basics

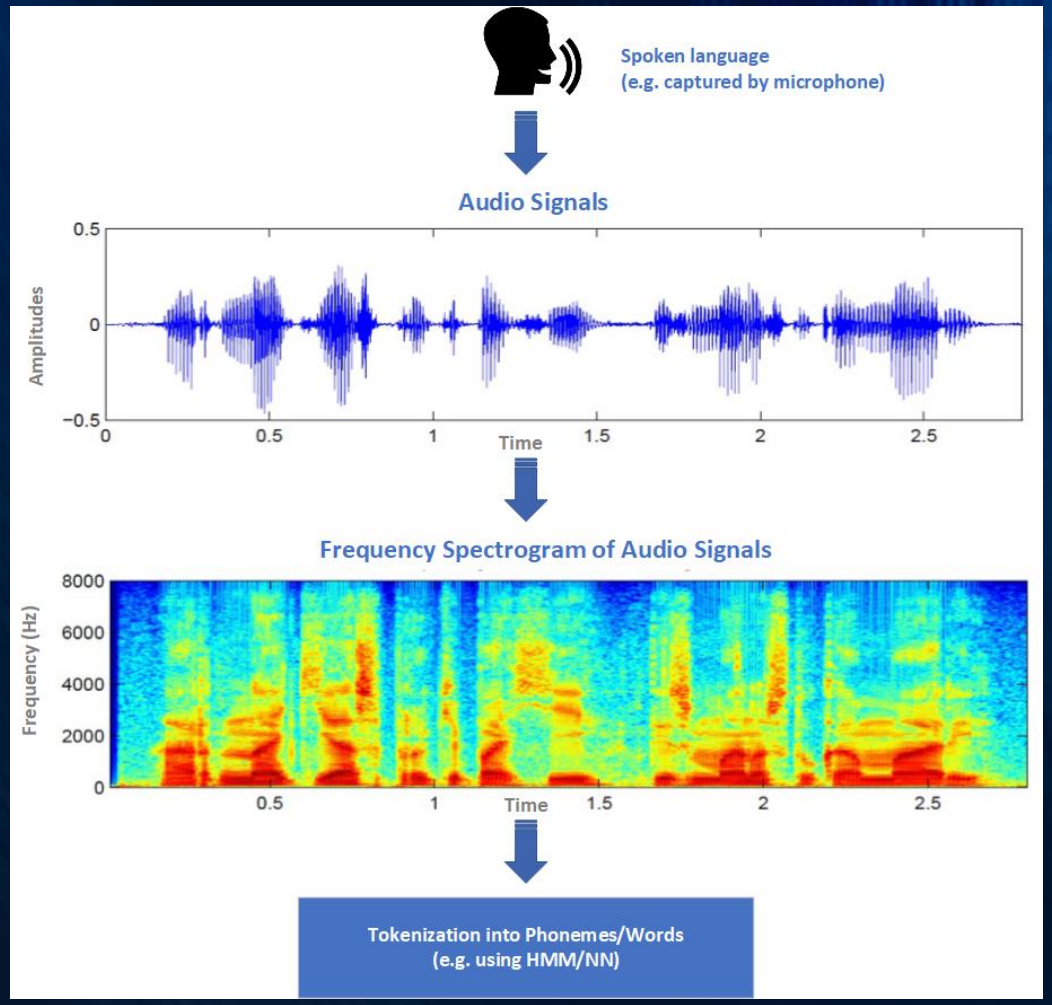

Fig. 1.8 Basic operation of speech recognition

- Speech recognition – technically known as "voice-to-text" process is a complex process which consists of 4 basics steps: 1) Voice capture / recording step; 2) Analog-to-digital conversion (ADC) step; 3) Frequency spectrogram generation step and 4) Phonemes / words generation step, as shown in Figure 6.7.
- In voice capture /recording step, spoken voice is captured or recorded voice communication devices such as microphones of PC or the user's mobile phone, via internet or mobile network to the NLP system such as the technical support of the IT computer.
- In analog-to-digital converter (ADC) step, the system translates these analog sound wave into digital data so that it can be processed by the speech recognition system. To achieve this, the system samples or digitizes the sound wave by taking measurements of the sound waves at different time intervals.
- In frequency spectrogram generation step the system filters the background noises to improve the sound quality, and then separate the sound waves into different bands of frequency to generate the so-called Frequency Spectrogram. It also normalizes the sound or adjusts it to a constant volume level. As people don't always speak at the same speed, so the sound must be adjusted to match the speed of the template sound samples already stored in the system's memory.
- In phonemes / words tokenization step, the processed signal is divided into small segments as short as a few hundredths of a second, or even thousandths in the case of plosive consonant sounds -- consonant stops produced by obstructing airflow in the vocal tract -- like "p" or "t." The system then matches these segments to known phonemes in the appropriate language. A phoneme is the smallest element of a language -- a representation of the sounds we make and put together to form meaningful expressions. There are roughly 40 phonemes in the English language (different linguists have different opinions on the exact number), while other languages have more or fewer phonemes. The most widely used voice-to-text tokenization technology is Hidden Markov Model (HMM), nowadays Deep Neural Networks (DNN) are also be used. Next, take a look on how HMM works.

Speech Recognition
Hidden Markov Model (HMM)

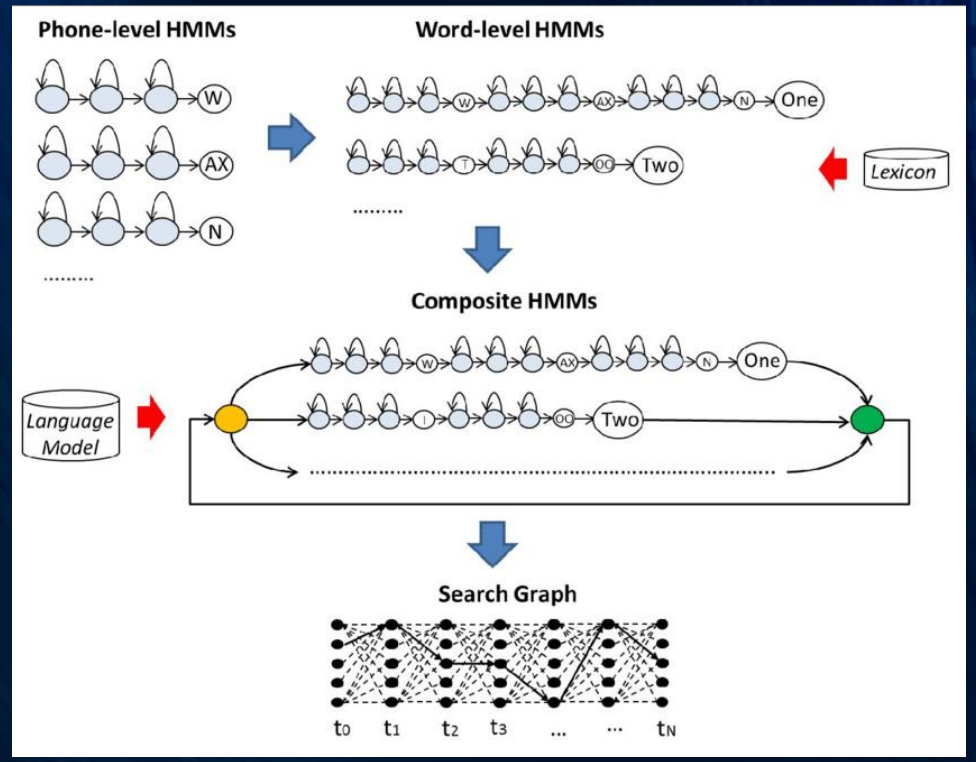

Fig. 1.9 Hidden Markov Model

- Frankly speaking, most speech recognition systems we are now using commercially are based on Hidden Markov Models (HMM). HMM is a very powerful statistical models based on basic concept of Markov process. HMM is a statistical models that output a sequence of symbols or quantities. HMMs are used for the tokenization of voice signal because it can be viewed as a piecewise stationary signal. Basically, within a short time-scale such as 5-10 msec, human voice can be modeled as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes.
- Another reason why HMMs are popular is because: 1) HMM is highly reliable; 2) It is technically easy to implement; 3) the Markov chains can be trained automatically and computationally feasible to use.
- An HMM is a system where a variable can switch (with varying probabilities) between several states, generating one of several possible output symbols with each switch (also with varying probabilities). The sets of possible states and unique symbols may be large, but finite and known.
- The basic process in HMM in speech recognition are:
  ① Inference: given a particular sequence of output symbols, compute the probabilities of one or more candidate state switch sequences.
  ② Pattern matching: find the state-switch sequence most likely to have generated a particular output-symbol sequence.
  ③ Training: given examples of output-symbol sequence (training) data, compute the state-switch/output probabilities (ie, system internals) that fit this data best.
- Figure 6.8 shows a typical HMM models on the tokenization of spoken voice in phonemes and words level.
- As shown in this figure, to accomplish the speech recognition process, a comprehensive lexicon for the targeted language is needed. We will talk about that in the coming section.

Syntactic Analysis
Parsing

- Syntax analysis is also known as parsing.
- Parsing is the process of determining whether a string of tokens can be generated by a grammar.
- It is performed by syntax analyzer which can also be termed as parser.
- Figure 6.9 shows the basic mechanism of parsing in syntactic analysis.
- As shown in Figure 6.9, the tokenized input texts (words or phonemes) generated by the speech recognition system are fed into the Lexical Analyzer and cross-checked with the lexicon database such as WordNet in order to check for correct syntax and grammar.
- After that it passes into the Parser to establish a data structure generally in the form of a Parse Tree or other syntax structures.
- The main roles of the parse include:
  - To check for any syntax or grammatic error.
  - To recover from commonly occurring error so that the processing of the remainder of program can be continued.
  - To construct and modify the parse tree.
  - To construct and modify the symbol table.
  - To produce intermediate representations for information retrieval.
- It may be defined as the software component designed for taking input data (text) and giving structural representation of the input after checking for correct syntax as per formal grammar. It also builds a data structure generally in the form of parse tree or abstract syntax tree or other hierarchical structure.

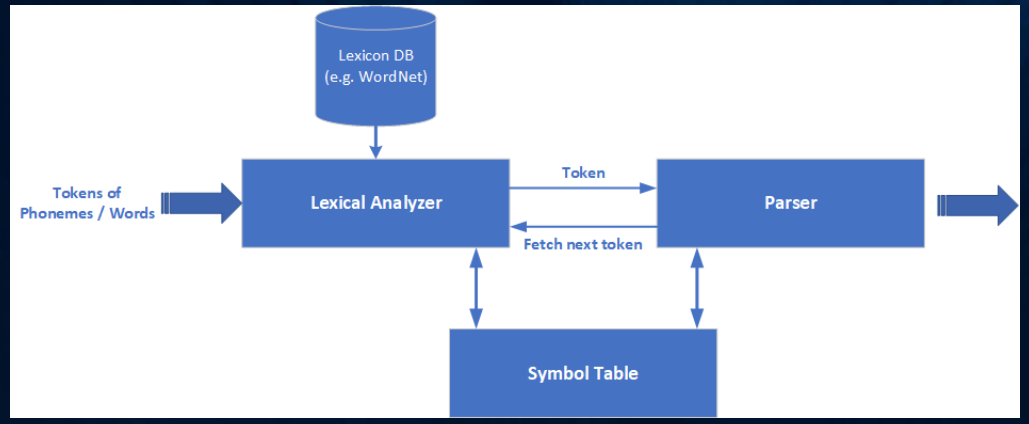

Fig. 1.10 Basic mechanism of Parsing in Syntactic Analysis

Syntactic Analysis
Parse Tree

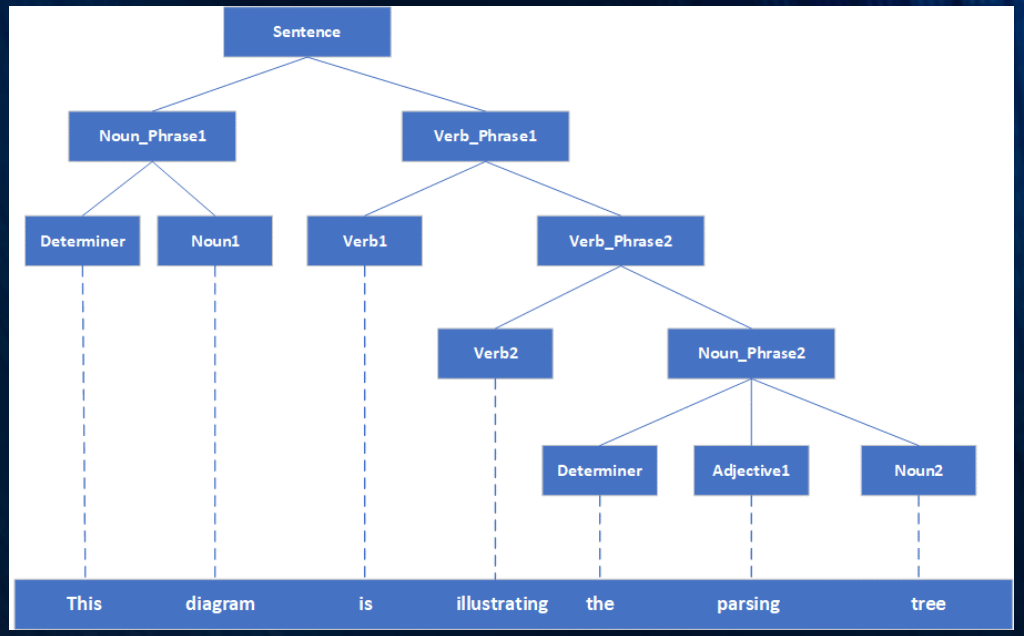

Fig. 1.11 An example of Parse Tree for the sentence "This diagram is illustrating the parsing tree"

Grammar in Parser

- Grammar is very essential and important to describe the syntactic structure of well-formed languages (and even programming languages).
- In the literary sense, it defines the syntactical rules for conversation for different languages.
- Linguistics have attempted to define grammars since the inception of natural languages like English, Chinese, Japanese, Hindi, etc.
- The theory of formal languages is also applicable in the fields of Computer Science mainly in programming languages and data structure.
- Even for a computer language such as Java language, the precise grammar rules state how functions are made from lists and statements.

What is a Parse Tree?

- Parse Tree is defined as the graphical depiction of a derivation.
- The start symbol of derivation serves as the root of the parse tree – the "sentence (s)" root node.
- The other nodes include all the constitutional components in a sentences include: Noun_Phrase (NP), Verb_Phrase (VP), Determiner (D), Verb (V), Noun (N), Proper_Noun (PN), Adjective (AJ), Adverb (AV), etc.
- For each parse tree, the leaf nodes are terminals and interior nodes are non-terminals.
- A property of parse tree is that in-order traversal will produce the original input text sentence.
- So, the main function of parsing is: Given a sentence with a grammar, the parser will check the sentence whether it is correct according with the grammar and if so, returns a Parse Tree representing the structure of the sentence.
- Figure 6.10 shows an example of parse tree for the sentence "This diagram is illustrating the parsing tree".

Semantic Analysis

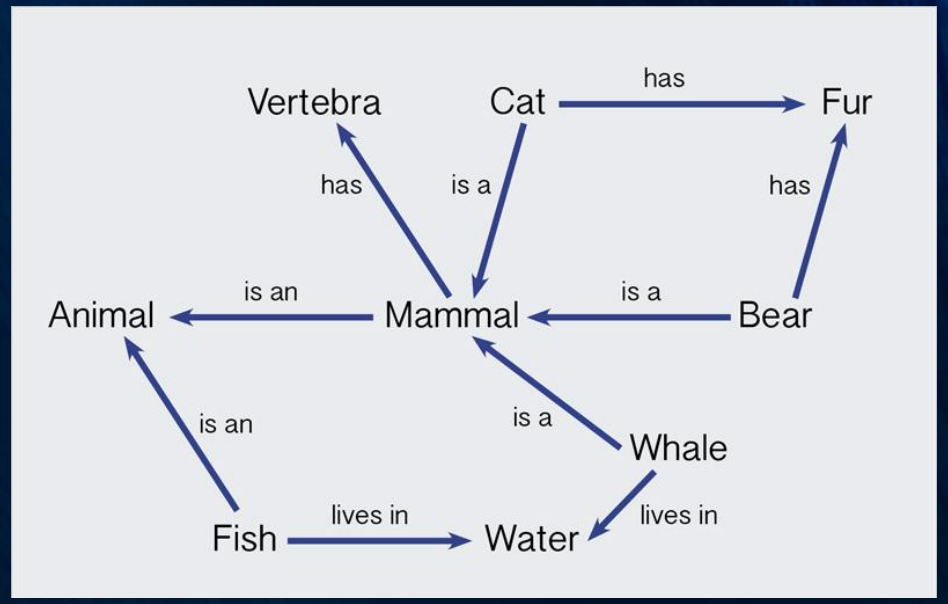

Fig. 1.12 An example of Semantic Network for all the meanings and concepts related to "Mammal"

What is Semantic Analysis?

- In short, semantic analysis is to extract the actual meaning of the text, or one may says the "dictionary meaning". Anyway, the main task of semantic analysis to check for the "meaningfulness" of the text.
- One may wonder: Lexical analysis in syntactic analysis in some sense already check for the meanings of the words, so what is the different between lexical analysis and semantic analysis?
- The truth is: lexical analysis only check for the meaning of the individual words from the lexicon, while semantic analysis extract the "overall meaning" of the text which most likely involves the combination of more than one or several token of words, in order to extract the actual meaning of the whole text message.
- For example, the sentence "Einstein is great scientist" means "Albert Einstein, the one who proposed General Relativity is a great scientist; or another scientist called "Einstein" is a great scientist? To solve this problem, we need to know more about the sentence to give us more cues which "Einstein" refers to, in order investigate the actual meaning of the sentence. That is the reason why we need semantic analysis.

Semantic Network

- A semantic network is a knowledgebase that represents semantic relations between concepts in a network.
- This is often used as a form of knowledge representation.
- In short, a semantic network is a directed or undirected graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between concepts, mapping or connecting semantic fields.
- In general, most semantic networks are cognitively based. They also consist of arcs and nodes which can be organized into a taxonomic hierarchy.
- Figure 6.11 shows a typical example of semantic network of concept "Mammal". As one can see, even from such simple semantic network, we can extract several useful knowledge.
- To use the semantic network, simple start with the node corresponds to "word" you are interested with, forward the arrow of the graph to find a path. Each path will correspond to a particular knowledge related to the concept you are interested with. For instance:
  - Mammal is an animal.
  - Bear is a mammal which has vertebra.
  - Whale is a mammal, lives in water.
  - Cat is a mammal has fur.

Pragmatic Analysis

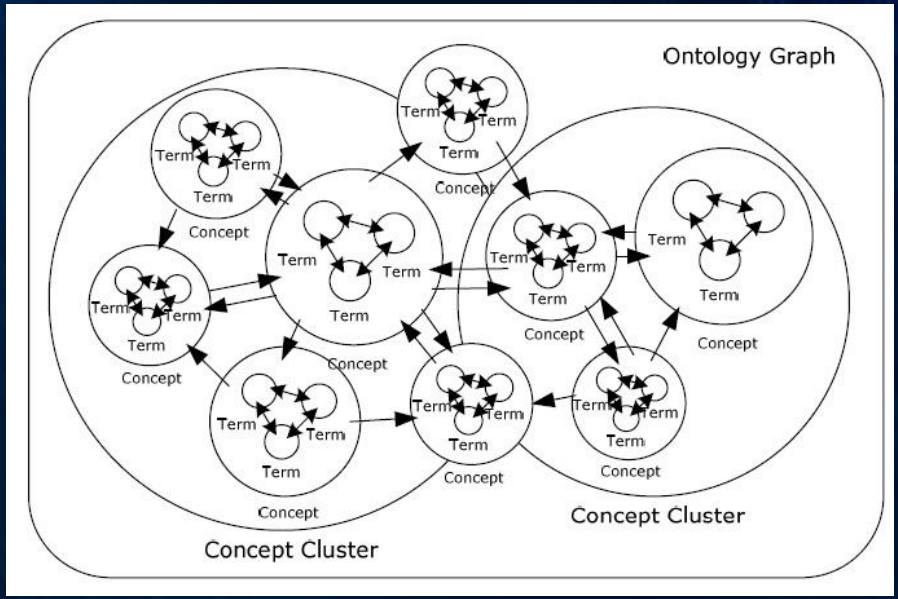

Fig. 1.13 Conceptual diagram of Ontology Graph

- Pragmatic Analysis is the last phase of linguistic analysis.
- It is part of the process of extracting information from text.
- Specifically, it's the portion that focuses on taking a structures set of text and figuring out what the actual meaning is.
- Unlike semantics, which examines meaning that is conventional or "coded" in a given language, pragmatic analysis studies how the transmission of meaning depends not only on structural and linguistic knowledge (such as grammar and lexicon) of the speaker but also on the context of the utterance, which might involves any pre-existing knowledge and the inferred intent of the speaker.
- The most famous example is the clause: "Raining cats and dogs". We all know it means raining heavily. But how can we draw this knowledge. As such concept of knowledge is totally unrelated to the syntax, semantic or even the semantic networks of either: cat, dog or rain – this the task for Pragmatic Analysis.
- In other words, the main purpose of pragmatic analysis is tried to extract the "true knowledge (meaning)" of the spoken speech, that maybe or may not be directly reflected by its semantic meaning.
- Due to the highly complexity and ambiguity to extract the "embedded meaning" of the spoken language, which not only related to the spoken message, but also other related knowledge and concepts outside the topic context and knowledge domain.
- Pragmatic analysis is believed to be one of the most difficult topic in linguistic and AI in terms of the implementation of NLP and knowledgebase and search engine.
- Different from voice recognition, syntactic and semantic analysis which are technically mature enough will widely adopted methods and technology, pragmatic analysis is still in the R&D stage without any dominant technology and solution.
- Latest research of pragmatic analysis include the R&D of an emerging AI technology – Agent Ontology, which focus on the ultimate problem of how human knowledge is generated, stored and retrieval. And more importantly, how difference concepts and ideas are related together in order to retrieve the actual and embedded meaning, which will be discussed in detail in the next Chapter – Ontological Search Engine.
- Figure 6.12 shows a snapshot of an example ontology graph.

Human Voice Synthesis
Text-to-Speech Synthesis

- Speech synthesis is artificial simulation of human speech with by a computer or other device.
- As the counterpart of the voice recognition, speech synthesis is mostly used for translating text information into audio information and also known as TTS (Text-To-Speech) technology.
- Apart from this, it is also used in assistive technology for helping vision-impaired individuals in reading text content.
- Similar to speech recognition technology, speech synthesis technology can be considered as a mature technology and are now widely used in many daily life related applications such as voice-enabled services and mobile applications.
- A typical speech synthesis system consists of three main modules: Text Analyzer; Linguistic Analyzer and Wave Form Generator.
- The main function of the Text Analyzer is to convert the response text which generated (says) from the ontology-based knowledgebase back into tokens of words – tokenization process. After that, it passes to the Linguistic Analyzer for further processing.
- In the Linguistic Analyzer, it assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation. The processed phonemes then passes to the Wave Form Generator.
- The Wave Form Generator, commonly known as "speech synthesizer" converts the symbolic linguistic representation into sound. In most systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech in the form of human voices.

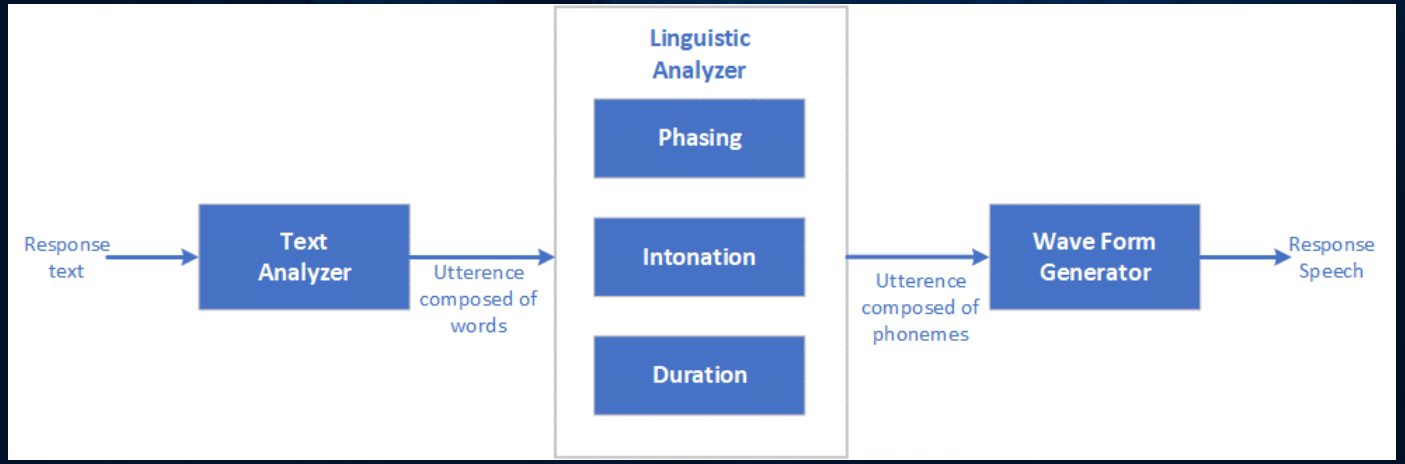

Fig. 1.14 Systematic diagram of Speech Synthesis

Linguistic Resources

Corpus

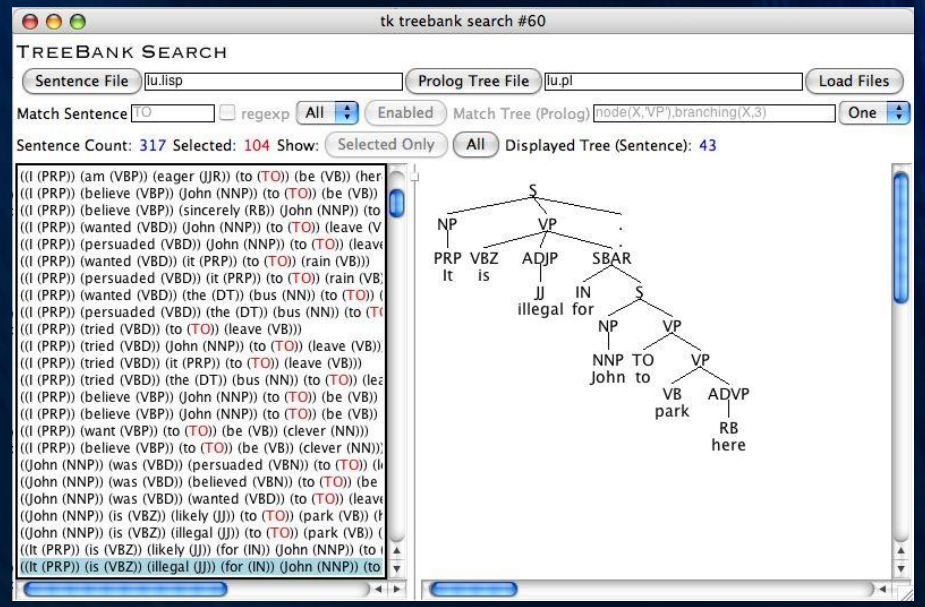

Fig. 1.15 Snapshot of Parse Tree using TreeBank Corpus

What is Corpus?

- In Linguistics, corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. Technically speaking, corpus can be derived in different ways like text that was originally electronic, transcripts of spoken language and optical character recognition, etc.
- As language is infinite but a corpus has to be finite in size. For the corpus to be finite in size, we need to sample and proportionally include a wide range of text types to ensure a good corpus design.
- Another important element of corpus design is its size. How large the corpus should be? There is no specific answer to this question. The size of the corpus depends upon the purpose for which it is intended as well as on some practical considerations as follows:
  - Kind of query anticipated from the user.
  - The methodology used by the users to study the data.
  - Availability of the source of data.
- With the advancement in technology, the corpus size also increases. For example, the size of the "Brown and LOB" corpus is around 1 million words, while the commonly used nowadays "Bank of English" corpus is over 650 million words in total.

TreeBank and ProBank Corpus

- TreeBank Corpus is linguistically parsed text corpus that annotates syntactic or semantic sentence structure. Geoffrey Leech coined the term 'treebank', which represents that the most common way of representing the grammatical analysis is by means of a tree structure. Semantic and Syntactic Treebanks are the two most common types of Treebanks in linguistics.
- ProBank Corpus also known as "Proposition Bank" is a corpus, which is annotated with verbal propositions and their arguments. The corpus is a verb-oriented resource; the annotations here are more closely related to the syntactic level. Martha Palmer et al., Department of Linguistic, University of Colorado Boulder developed it. In NLP, the PropBank project has played a very significant role and helps in semantic role labeling.
- Figure 6.14 shows a sample snapshot of the parse tree construction using TreeBank Corpus.

Linguistic Resources
WordNet and VerbNet

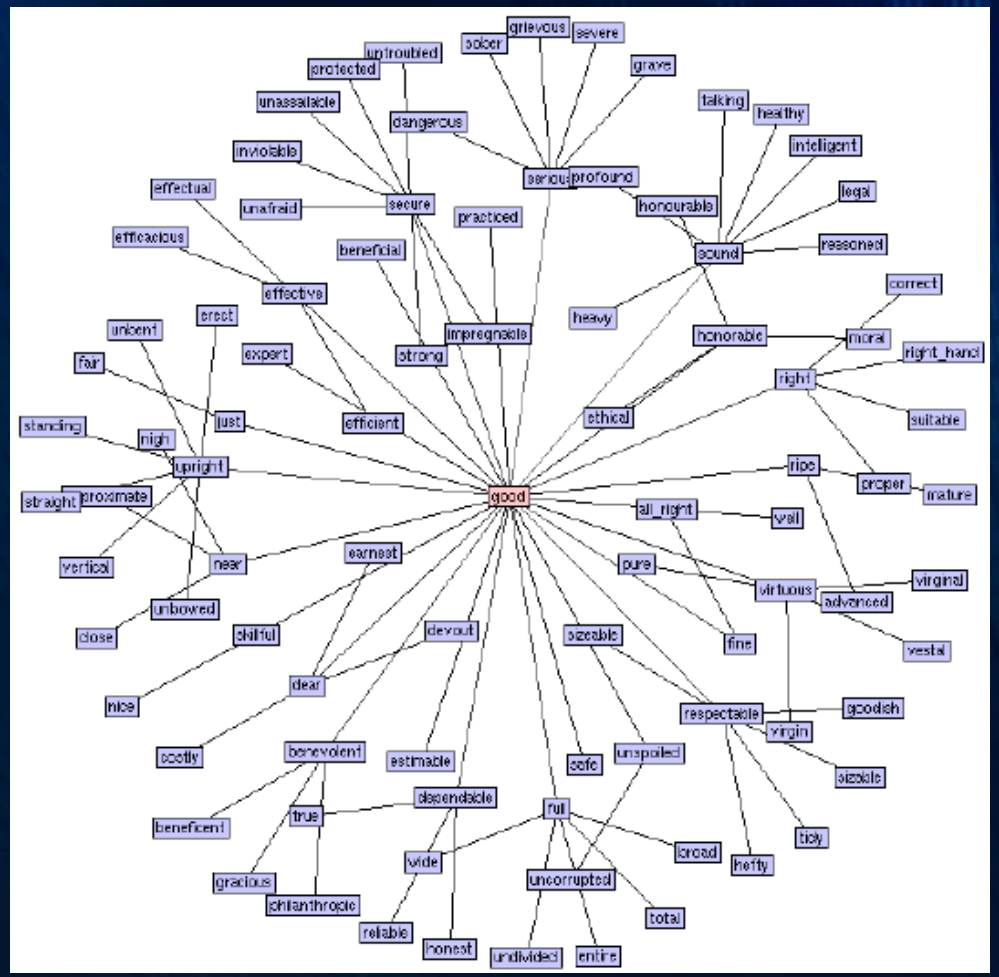

Fig. 1.16 Snapshot of WordNet database from perspective of adjective 'good'

- WordNet and VerbNet are the two most commonly used corpus in NLP.
- VerbNet (VN)
  - VerbNet (VN) is the hierarchical domain-independent and largest lexical resource present in English that incorporates both semantic as well as syntactic information about its contents.
  - VN is a broad-coverage verb lexicon having mappings to other lexical resources such as WordNet, Xtag and FrameNet.
  - It is organized into verb classes extending Levin classes by refinement and addition of subclasses for achieving syntactic and semantic coherence among class members.
- WordNet (WN)
  - WordNet, created by Princeton is a lexical database for English language.
  - WordNet is a lexical database of semantic relations between words in more than 200 languages and the most commonly used corpus in the world, both for commercial and academic research in NLP, knowledgebase and agent ontology R&D,
  - It is the part of the NLTK corpus. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called Synsets. All the synsets are linked with the help of conceptual-semantic and lexical relations.
  - In terms of information systems, WordNet is used for various purposes like word-sense disambiguation, information retrieval, automatic text classification and machine translation. One of the most important uses of WordNet is to find out the similarity among words.
  - Due to its popularity, various functions and algorithms on syntactic and semantic analysis have been implemented in various NLP related development platform as functional libraries or packages include C, C++, Perl, ADW in Java and NLTK in Python.
  - Figure 6.15 shows a snapshot WordNet database from perspective of adjective 'good'

Applications of NLP in Daily Life

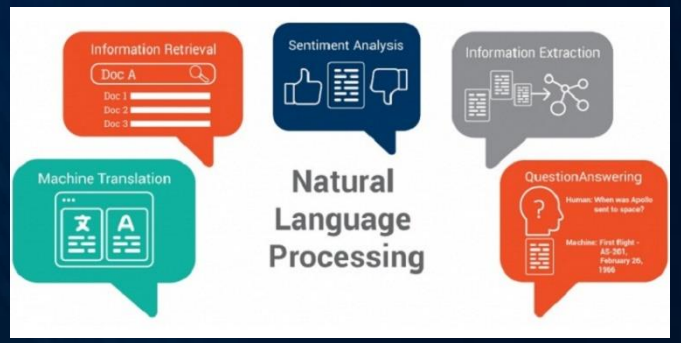

Fig. 1.17 Applications of NLP to daily life

- After over 20 years of R&D and actual implementation, NLP technology is now being using in many applications that are closely related to our daily life.
- They include: Machine Translation (MT), Information Extraction (IE), Information Retrieval (IR), Sentiment Analysis, QA Chatbot systems.
- As shown in Figure 6.15.
- Machine Translation (MT)
  - Machine translation is the oldest and one of the most important application of NLP. It is also one of the most well-studied, earliest applications of NLP.
  - One major challenge in MT nowadays is two folds: 1) the "naturalness (or fluency) - machine translation that is natural in the target language while preserving the exact meaning expressed by the input; 2) the "adequacy" – the degree of MT to which the output reflects the meaning of the source.
  - These two are often in conflict, especially when the source and the target languages are not very similar (e.g. translation between Chinese and English).
  - Experienced human translators address this trade-off in a creative way.
  - The goal of nowadays MT is to apply various AI technology such as Deep Learning to learn from experts in order to achieve human-quality translations.
- Information Extraction (IE)
  - IE is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources.
  - In most of the cases this activity concerns processing human language texts by means of NLP.
  - Recent activities in multimedia document processing like automatic annotation and content extraction out of images, audio, video, and text documents that could be seen as information extraction.
  - Due to the difficulty of the problem, many commercial IE applications are domain specific such as the focus of a particular discipline (e.g. Law, environments) or topic interest.

Applications of NLP in Daily Life

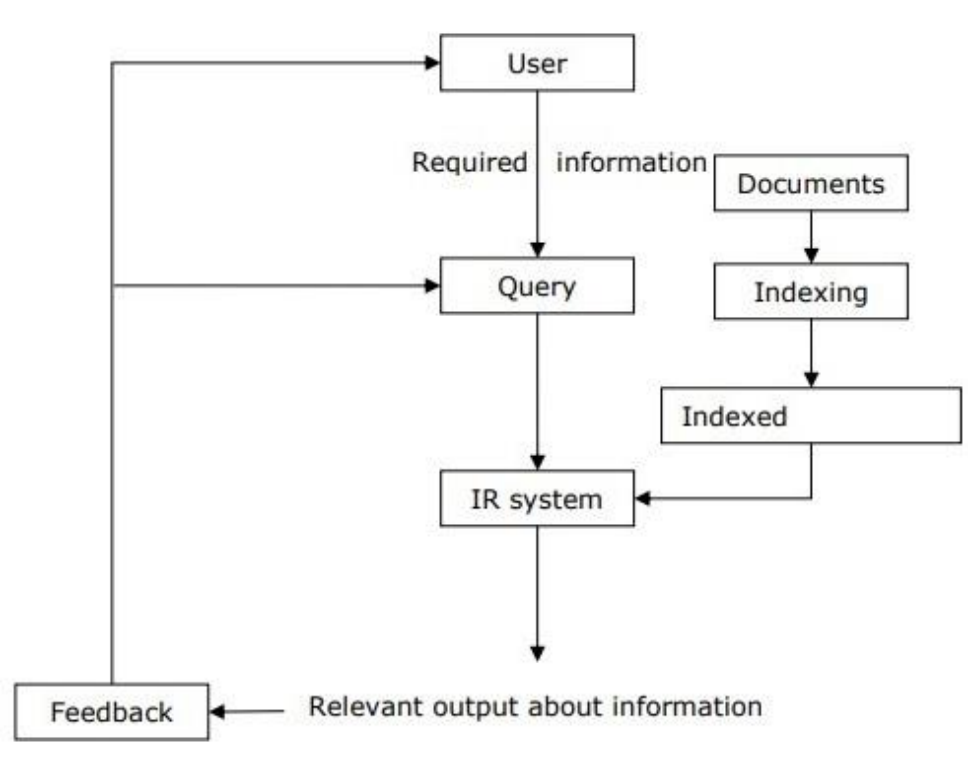

Fig. 1.18 Information Retrieval System

- Information Retrieval (IR)
  - IR is software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information.
  - The system assists users in finding the information they require but it does not explicitly return the answers of the questions.
  - It informs the existence and location of documents that might consist of the required information.
  - The documents that satisfy user's requirement are called relevant documents.
  - A user who needs information will have to formulate a request in the form of query using NLP.
  - Then the IR system will respond by retrieving the relevant output, in the form of documents, about the required information.
  - In fact, the main objective of IR system is to develop a model for retrieving information from the repositories of documents.
  - A typical example of IR system is so-called "ad-hoc retrieval problem".
  - Figure 6.16 shows the process flowchart a typical information retrieval system using ad-hoc retrieval method.
  - In ad-hoc retrieval, the user must enter a query in natural language that describes the required information.
  - Then the IR system will return the required documents related to the desired information.
  - For example, suppose we are searching something on the Internet and it gives some exact pages that are relevant as per our requirement but there can be some non-relevant pages too. This is due to the ad-hoc retrieval problem.

Applications of NLP in Daily Life

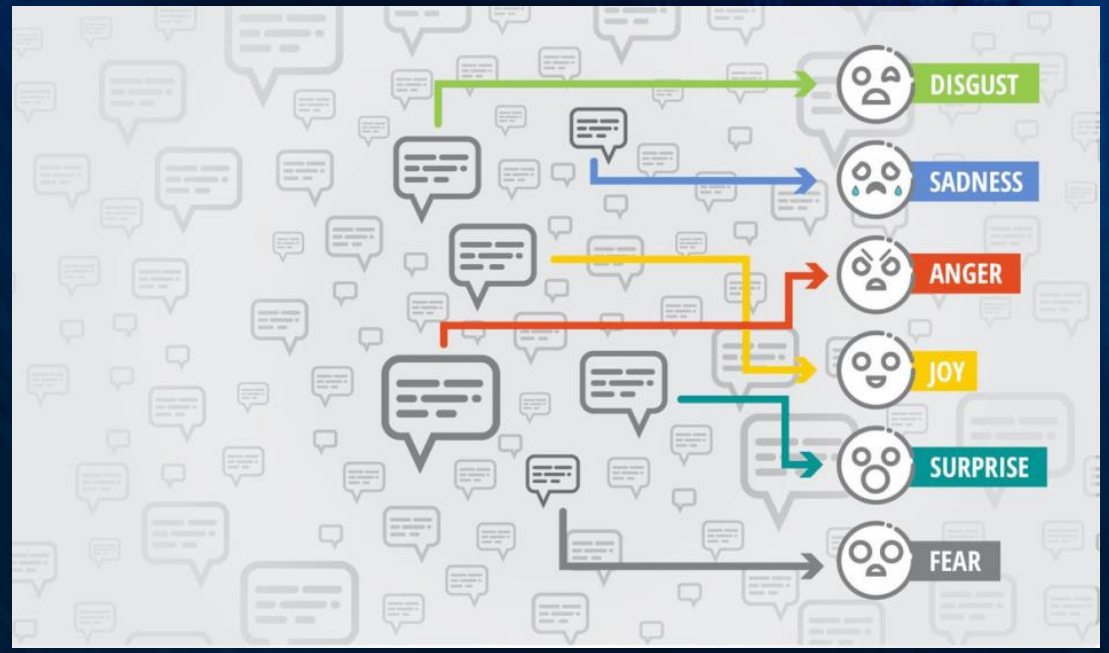

Fig. 1.19 NLP on Sentiment Analysis

- Sentiment Analysis
  - Sentiment analysis is a type of data mining that measures the inclination of people's opinions through NLP, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources.
  - The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information.
  - Customer services use sentiment analysis, an application of NLP to identify the user opinion and sentiment.
  - It will help companies to understand what their customers think about the products and services.
  - Companies can also judge their overall reputation from customer posts with the help of sentiment analysis.
  - In this way, we can say that beyond determining simple polarity, sentiment analysis understands sentiments in context to help us better understand what is behind the expressed opinion.
  - Figure 6.17 shows a typical scenario of sentiment analysis using NLP technology.
  - As shown in Figure 6.10, with the integration of data mining technology and NLP, user responses and comments on various topic of interest can be converted into machine understandable concepts and ideas that can be effectively classified into different degree of emotions and response, which can be used by the companies to better understand and analyze the customer needs.
  - For news agency and public forum, sentiment analysis together with NLP technology can be used to data-mine the public opinions and comments more effectively and objectively.
  - In fact, NLP-based sentiment analysis is widely used to major social media and forum such as Facebook and Twitter to user opinion and real-time response to some ah-hoc events and incidences.

Applications of NLP in Daily Life

Fig. 1.20 Customer service robots using NLP technology

- Question & Answering (Q&A) Robots (Systems)
  - Another main application of natural language processing (NLP) is question-answering robots (or systems).
  - In general, Q&A systems is the "ultimate challenge" of NLP and AI which is the main theme of Turing Test.
  - It concerns with the building of AI system that and automatically answer questions raised by humans using our own languages.
  - In other words, the Q&A not only need to recognize and understand human language, but also need to know the "actual" meaning from syntax, semantic up to pragmatic levels.
  - Besides, it need to response with human voice, which also involve high-level knowledge-based and inferencing, together with human voice generation system.
  - With the fast growing in AI and NLP technology, Q&A robots and systems are widely used in many industries includes:-
    - Technical support robots e.g. IT support using technical support robot to provide basic technical support via internet and traditional hotline support.
    - Customer service robots e.g. For product promotion and after-sale support services.
    - Language learning tutor e.g. English language robot to teach and train student in language center.
  - Figure 6.18 shows a typical scenario of Q&A customer service robots.

Case Study
Language Learning Chatbot using NLP

Fig. 1.21 An English Language Learning Chatbot

- AI Chatbot is not new thing in AI world, date back to the 80's, Japan's industry had already developed several famous companion robots which can interact with human and provide limited NLP capability.
- With the advance AI, computing and robotic technology, nowadays companion robots are capable to provide more powerful service, such as Foreign Language Learning Robots (LLR) to teach students how to speak foreign languages, such as English LLR to teach Asian students to learn English, or Spanish LLR to teach English-language oriented students to learn Spanish in their daily life and activity.
- Suppose you are the English LLR system designer. Based on the NLP technology learnt in this chapter, together with various machine learning and data mining technique learnt in the previous chapters:
  ① What are the THREE basic machine learning techniques in AI?
  ② Discuss and explain how these machine learning techniques can be applied to English learning?
  ③ Discuss and explain how to integrate these machine learning techniques, together with the NLP technology learnt in this chapter to implement an English Learning Robot?
  ④ Many learning system have different level of challenge. Suppose you have to design this English LLR with 3 levels of challenge. What are these 3 level of challenge?
  ⑤ What kind of AI and NLP methods you have to use to implement these 3 levels of functions?

Natural Language Processing
Summary

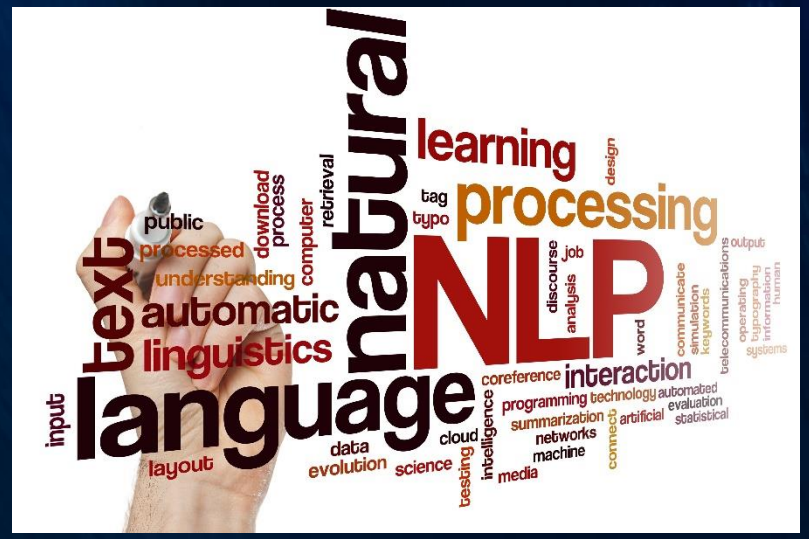

Fig. 1.22 NLP Framework

- In this chapter, we have discussed a very challenging and important AI technology – Natural Language Processing (NLP).
- As said at the beginning of the chapter, NLP is not a new topic.
- In fact, the Turing Test proposed by Alan Turing in the 1950's is focused on NLP performance of the machine to determine the degree of intelligence.
- Besides, it is also one of the core components of Generalized AI (GAI) and the design of robots during 1960 – 80's.
- However, owing to over-expectation of AI and limitation of computational capability, NLP technology developed slowly and mainly focused on statistical-based Machine Learning (ML) applications.
- With the advance of computational speed and the popularity of AI, Machine Learning, Deep Network, Big Data and Data Mining, NLP technology and related applications are developed rapidly in the past 20 years.
- Nowadays, various NLP-related technology such as human voice synthesis systems for car navigation, information retrieval (IR), information extraction (IE), NLP-based customer service robot, sentiment analysis in social media, machine translation apps and systems and Q&A chatbots become part of our daily life.
- In fact, NLP is not just a human voice related technology, it is closely related to how we acquire, learn, store and manipulate our knowledge. The so-called "ultimate challenge" of AI – knowledgebase and agent ontology problem.
- In the next chapter, we will discuss this challenging topic – Ontological-based Search Engine.

Next

NLP Workshop #1

NLP Basics with NLTK