Natural Language Processing: Introduction

Dr. Muskan Garg

Thapar Institute of Engineering & Technology

muskan@thapar.edu

February 25, 2022

Overview

1. Introduction

2. NLP Definitions

3. Knowledge in NLP

4. Approaches to Natural Language Processing
   - Rationalist Approach
   - Empiricist Approach

5. Applications of NLP

Introduction

- Language is the primary means of communication used by humans.
- It is the tool with which we express our ideas and emotions.
- Language shapes our thought, has a structure, and carries meaning.
- Representing ideas and thoughts is so natural that we hardly realize that how we process knowledge. There must be some kind of representation (model) of content of language in our mind that helps to represent it.

Summary

Natural language processing is concerned with the development of computational models of aspects of human language processing.

Introduction (Contd.)

- Building computational models with human language processing abilities require a knowledge of how humans acquire, store and process knowledge.
- It also requires a knowledge of the world and of language.
- These computational models are useful:
  1. In developing automated tools for language processing.
  2. To gain better understanding of human communication.

Naming Conventions

Natural Language processing is an interdisciplinary field with many names such as *Speech and Language Processing*, *human language technology*, *natural language processing*, *computational linguistics*, *speech recognition and synthesis*.

NLP: Formal definitions

- Natural language processing (NLP) is a field of computer science, artificial intelligence concerned with the interactions between computers and human (natural) languages.
- Natural Language Processing is a field to compose the computational models which understand, generate and manipulate the human language.
- NLP consists of Natural Language Generation (NLG) and Natural Language Understanding (NLU).

Types of NLP

- Converting language of human beings (NLU) to *machine readable form* so that it can be manipulated and processed to *produce the natural language* (NLG) (1950s till late 1990)[1]
- Due to certain problems, we discarded this classification and we moved to ML/ DL in early 2000's[2]
- NLP is divided into Rule based methods (Traditional NLP) and statistical methods (ML based or Advanced NLP)

[1] Traditional NLP

[2] ML based NLP

Rule based NLP

- The input form of data for rule based NLP can be text or speech.
- To understand the the structure and meaning of sentences, we use some linguistic rules.
- These linguistic rules have many problems like ambiguity (recognizing phonology and phonetics), requirement of linguistic resources (dictionaries/ lexicons of different languages: WordNet, SentiNet, sentence structure: context free grammar). Available for English language but problems with low resource languages. Thus, Rule based NLP fails.
- Formerly, many language-processing tasks typically involved the direct hand coding of rules, which is not in general robust to natural language variation.

Statistical NLP/ Computational Linguistics

- The machine-learning paradigm calls instead for using statistical inference to automatically learn such rules through the analysis of large corpora of typical real-world examples (a corpus (plural, "corpora") is a set of documents, possibly with human or computer annotations.
- Statistical NLP focus on corpus-driven methods that make use of supervised and unsupervised machine learning approaches and algorithms.
- Systems based on machine-learning algorithms have many advantages over hand-produced rules:
  - No Language Dependency and Expertise.
  - Automatically focus on the most common cases
  - Robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Can be made more accurate simply by supplying more input data

Knowledge in Natural Language Processing

- NLP applications like other conventional programs or data processing systems require knowledge (*Knowledge of language*).
- For instance, Consider a UNIX wc program, which counts the total number of bytes, words, and lines in the text.
    1. When it counts the number of bytes it is a simple data processing application.
    2. But when it is used to count the number of words in a file, it requires knowledge about what it means to be a word in a language and thus become a language processing system.

Summary

Similarly, the advanced applications like conversational agent (such as Siri), machine translation, question answering agent require different levels of knowledge.

Phonetics and Phonology

- Any conversational agent or speech recognition and synthesis system must be able to recognize words from an audio signal and to generate words from an audio signal.
- These applications require knowledge about phonetics and phonology- how words are pronounced in terms of sequence of sounds and how each of these sounds is recognized acoustically.
- **Phonetics** deals with the organs of sound production.
- **Phonology** deals with the sounds and their changes due to various factors such as climatic change, race, influence of other languages and the like.

Knowledge: Morphology and syntax knowledge

- The systems must be able to produce and understand variations like I'm , can't, or other variations like pluralization, verb forms, etc.
- Producing and recognizing these variation of individual words require knowledge about *morphology*
- **Morphology** is the the study of words, their structure and parts of words such as stems, root words, prefixes, and suffixes.
- Moving beyond words, these applications require structural knowledge to group together the words which constitute a response.
- The knowledge required to group the words in correct order is called **syntax knowledge**.

Knowledge: Semantics

- To answer the questions or to understand the text, NLP questions need to have the knowledge of *semantics i.e. meanings*.
- Consider the question to be answered by a question answering system:

Question

Question: How much Chinese silk was exported to Western Europe by the end of the 18th century?
Now, to answer this question, we need to know about **lexical semantics**- meaning of all words (export, silk) and *compositional linguistics* (what constitutes *Western Europe* as opposed to Eastern or Southern Europe, what does *end* mean in context of 18th century).

Knowledge: Pragmatics or dialogue knowledge

- *Pragmatic or dialogue knowledge*- knowledge of the relationships of meaning to the goals and intentions of speaker.
- Consider the question put to conversational agent:

Question

Question: *Hey Siri, is the John's door open?* Or *Hey Siri, open the John's door*?
Despite the bad behavior, HAL knows to be polite to speaker by not simply replying as *No* or *No, I won't open the door*, it response with the phrase *I am sorry. I can't*. This knowledge of kindness of actions that speakers intend by their use of sentences is **pragmatic or dialogue knowledge**.

Knowledge: Co-reference resolution or discourse

- *Discourse Analysis*- 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.
- Discourse knowledge is the knowledge about linguistic units larger than a sentence.
- Consider the question

Question

Question: *How many states were in the US that year?*
To answer this question, it must be able to interpret words like *that year*. This task of **conference resolution** makes use of knowledge about words like *that* or pronouns like *it* or *she* refers to the previous parts of **discourse**.

Knowledge: Summary

To summarize, complex language processing applications require various kinds of knowledge:

- **Phonetics and Phonology**- knowledge about linguistic sounds
- **Morphology**- knowledge of representation, structure and parts of words.
- **Syntax**- knowledge of structural knowledge between words.
- **Semantics**- knowledge of meaning
- **Pragmatic**- knowledge of the relationships of meaning to the goals and intentions of speaker
- **Discourse**- knowledge about linguistic units larger then sentence.

Approaches to Natural Language Processing

There have been two major approaches to natural language processing:

- Rationalist approach
- Empirical approach.

Between about 1960 and 1985, most of linguistics, psychology, artificial intelligence, and natural language processing was completely dominated by a rationalist approach.

Rationalist Approach

- A rationalist approach is characterized by the belief that a significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance.
- Within linguistics, this rationalist position has come to dominate the field due to the widespread acceptance of arguments by *Noam Chomsky* for an *innate language hypothesis* (natural language processing capability).
- Chomsky suggested that it is difficult to see how children can learn something as complex as a natural language from the limited input (of variable quality and interpretability) that they hear during their early years.
- Within artificial intelligence, rationalist beliefs can be seen as supporting the attempt to create intelligent systems by hand coding into them a lot of starting knowledge and reasoning mechanisms, so as to duplicate what the human brain begins with.

Empiricist Approach

- An *empiricist approach* also begins by postulating some cognitive abilities as present in the brain.
- The empiricist approach assumed that a baby's brain begins with general operations for *association, pattern recognition, and generalization* and that these can be applied to the rich sensory input available to the child to learn the detailed structure of natural language.
- An empiricist approach to NLP suggests that we can learn the complicated and extensive structure of language by specifying an appropriate general language model, and then inducing the values of parameters by applying statistical, pattern recognition, and machine learning methods to a large amount of language use.

Application: Pre-processing

- A **spell checker (or spell check)** is an application program that flags words in a document that may not be spelled correctly. Spell checkers may be stand-alone, capable of operating on a block of text, or as part of a larger application, such as a word processor, email client, electronic dictionary, or search engine.
- Text preprocessing is an important part of Natural Language Processing (NLP), and **normalization of text** is one step of preprocessing. The goal of normalizing text is to group related tokens together, where tokens are usually the words in the text.

Application: POS Tagging

- **Part-of-speech tagging (POS tagging):** also called **grammatical tagging** or **word-category disambiguation**, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

Use of POS tagging

The **POS tagger** is used as a *preprocessor*. *Text indexing and retrieval* uses POS information. Speech processing uses POS tags to decide the *pronunciation*. POS tagger is used for making *tagged corpora*.

Application: WSD and NER

- **Word Sense Disambiguation (WSD):** WSD is identifying the sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings.
- **Named-entity recognition (NER)** (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Use of WSD and NER

The **WSD** and **NER** systems are also used as a preprocessor in a number of NLP applications such as machine translation, information retrieval and extraction, etc.

Application: Summarisation

- **Keywords and keyphrases extraction** as they are very useful in analyzing large amount of textual material quickly and efficiently search over the internet besides being useful for many other purposes.
- **Automatic text summarization** is the process of shortening a text document with software, in order to create a summary with the major points of the original document.

Use of Summarization

- Text summarization systems are used in applications like entity timelines, storylines of events, sentence compression, summarization of user generated content.

Application: Machine Translation

- **Machine translation** is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.
- Machine Translation systems are used in search engines, cross-lingual information retrieval, social networking, military applications, mobile applications, etc.

Use of Summarization

- Machine translation is used in many multimodal and multilingual downstream NLP tasks like Code switching conversation summarization, multi-lingual models, classifying code switched texts.

Application: Document Classification

- **Document classification** or document categorization is to assign a document to one or more classes or categories.

Use of Document Classification

It is used in number of applications like e-mail filtering, mail routing, spam filtering, news monitoring, selective dissemination of information to information consumers, automated indexing of scientific articles, automated population of hierarchical catalogues of Web resources, identification of document genre, authorship attribution, survey coding and so on.

Application: Speech Synthesis/ Text-to-speech Synthesis

- **Speech Synthesis/ Text-to-speech Synthesis:** Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a *speech computer or speech synthesizer*.

Use of Speech Synthesis

A *text-to-speech (TTS) system* converts normal language text into speech. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written words on a home computer.

Application: Opinion Mining/ Sentiment Analysis

- **Opinion Mining/ Sentiment Analysis:** Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event.

Use of Opinion Mining

Sentiment analysis is widely applied to: voice of the customer materials such as reviews and survey responses, online and social media, healthcare.

Application: Optical character recognition

- **Optical character recognition** is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).

Use of Optical character recognition

It is widely used as a form of information entry from printed paper data records like passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation.

Application: Question Answering Systems

- Question answering Systems: is concerned with building systems that automatically answer questions posed by humans in a natural language.
- QA research attempts to deal with a wide range of question types including: fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions, open ended, closed ended.

Application: Textual entailment (TE)

- **Textual entailment (TE)** in NLP is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text.

Use of Textual Entailment

Many natural language processing applications, like Question Answering (QA), Information Extraction (IE), (multi-document) summarization and machine translation (MT) evaluation, need to recognize that a particular target meaning can be inferred from different text variants.

Application: Topic segmentation and recognition

- **Topic segmentation and recognition:** Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

Use of Topic segmentation and recognition

It can improve information retrieval or speech recognition significantly (by indexing/recognizing documents more precisely or by giving the specific part of a document corresponding to the query as a result). It is also needed in topic detection and tracking systems and text summarizing problems.

Application: Relationship extraction

- **Relationship extraction:** Given a chunk of text, identify the relationships among named entities (e.g. who is married to whom).

Use of Relationship extraction

Application domains where relationship extraction is useful include gene-disease relationships, protein-protein interaction, etc.

Application: Paraphrase Identification

Paraphrase Identification: Paraphrase detection is the task of examining two text entities (ex. sentence) and determining whether they have the same meanings.

- **Lexical level**: Example - *solve* and *resolve*
- **Phrase level**: Example - *look after* and *take care of*
- **Sentence level:** Example - *The table was set up in the carriage shed* and *The table was laid under the cart-shed*
- **Pattern level**: Example - *[X] considers [Y]* and *[X] takes [Y] into consideration*.
- **Collocation level**: Example - *(turn on, OBJ light)* and *(switch on, OBJ light)*.

Application: Miscellaneous-I

- **Automated essay scoring (AES)**: The use of specialized computer programs to assign grades to essays written in an educational setting. It is a method of educational assessment and an application of NLP. It can be considered a problem of statistical classification.
- **Automatic image annotation**: A process by which a computer system automatically assigns textual metadata in the form of captioning or keywords to a digital image. The annotations are used in image retrieval systems to organize and locate images of interest from a database.
- **Automatic taxonomy induction**: Automated construction of tree structures from a corpus. This may be applied to building taxonomical classification systems for reading by end users, such as web directories or subject outlines.

Application: Miscellaneous-II

- **Code-Switching**: Code-switching is the communication phenomenon where the speakers switch between different languages during a conversation. With the widespread adoption of conversational agents and chat platforms, code-switching has become an integral part of written conversations in many multi-lingual communities worldwide.
- **Causal analysis and Discourse relations**: The discourse relation is a prominent cohesive agent which identified the reason behind the intent of a user out of the textual information. Some cause expressions are general, others relate more specifically to result, reason or purpose. For instance: so, then, therefore, in consequence, on account of this, for that purpose, etc.

The End