Natural Language Processing

By Omar AbedelKader

Contents

Processing natural languages such as English has always been one of the central research issues of artificial intelligence, both because of the key role language plays in human intelligence and because of the wealth of potential applications. Many of the knowledge representation and inference techniques that have been applied successfully in knowledge-based systems were originally developed for processing natural language, but the language-processing applications themselves have always seemed far from being realized. The special series on natural-language processing is an attempt to bring language processing and its applications into focus/spl minus/to demonstrate techniques that have recently been applied to real-world problems, to identify research ripe for practical exploitation, and to illustrate some promising combinations of natural-language processing with other emerging technologies.

Chapter 1

Introduction

1.1 What is Natural Language Proccessing

- NLP is a subfield of linguistics, computer science, and AI concerned with the interactions between computers and human language, and how to program computers to process and analyze large amounts of natural language data.

- Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can.

- Natural Language Processing is different Neuro-Linguistic Programming. The most example used in NLP is e-mail filtering, when an e-mail is received he enters in a classifier is based on NLP to classify the e-mail spam from the email which is not spam.

- Steps: Data Collection, Pre-processing, Feature extraction and selection, Learning and Classification, Recognizing activity.

1.2 History of NLP

Natural language processing has historical roots:

- Philosophers such as Leibniz and Descartes made proposals for symbols that would connect words between languages. All of these proposals remained theory, neither of which resulted in the creation of a real machine.

- Since the 1930s, there have been ideas and designs for a translation machine, but none with real efficiency.

- In 1950 Alan Turing (Who is considered one of the founding fathers of artificial intelligence) published his famous article ”Computing Machines and Intelligence” which introduced what is now called the Turing Test, as a standard for artificial intelligence. This criterion is based on the ability of a computer program to impersonate a human in written conversation with a human judgment at a time. Real, well enough that the judge is unable to reliably distinguish – based on the content of the conversation alone - between the program and a real human.

- The so-called ”Georgetown Experiment” was devised in cooperation with IBM in 1954 with a fully automatic translation of more than sixty Russian sentences into English. The authors of the program claimed that within three or five years the automatic translation problem would be solved.

- In 1968. The SHRDLU program was invented by the American engineer ”Terry Winograd” at MIT, who succeeded in establishing a kind of tactful dialogue with the computer.

- In 1991, a model of an ”intelligent psychotherapist” was made, based on the idea of a chatbot and it was working on DOS, and it begins with a sentence.

- In 2006, the first version of Watson, the mother of IBM's business, was released, which succeeded in outperforming humans in answering questions in the famous American program: Jeopardy . Personal Assistant launched: Siri in 2011, Alexa and Cortana in 2014, Google Assistant in 2016, and Samsung Bixby in 2017.

1.3 Challenges

- Ceiling Analysis: Ceiling analysis means that before you work, the models of machine learning or deep learning measure the efficiency of humans in it, and no model in machine learning or deep learning can increase the efficiency of humans and exceed humans in speed in quantity, but in intelligence, It is impossible to understand If a person does not understand a certain word, Character Recognition cannot understand this word.

- After the knowledge of NLP is one of the most important and complex sciences of Deep Learning as it deals with texts, which are often misunderstood by humans themselves because it contradicts the principle of Ceiling Analysis and the problem of computers is that it facilitates them to deal with numbers simply, and do all the arithmetic operations On it, but it is not easy for it to deal with texts, especially since it is unstructured data (Structured Data: Table, Unstructured Data: Videos, Audios, Image...) and it comes in different shapes, forms, and languages, but one of the most difficult challenges in NLP is the so-called (crashed blossoms).

- What are Crashed Blossoms? They refer to phrases that are written in a certain way and are misunderstood by the reader. *Examples: "A woman, without her man, is nothing." "A woman: without her, man is nothing".*

- Also among the challenges facing the other NLP:
  - Great difference in dialects, even in the same language.
  - Unofficial language is used in many conversations (you are, your, you're).
  - Incorrect intersections in metaphorical sentences.

- New terms: Format it, ring it.
- The ironic uses of the words.
- Words of multiple uses and meanings.
- Historical facts.
- The difficulty of defining special terms.

1.4 Applications

- First application: Feature extraction: Here we conclude that there may be a meeting from 10:00 to 11:30 that's why he proposed to create a new calendar entry
- Second application: Sentiment analysis: the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. (Twitter)
- Third Application: Machine Translation: Machine translation, sometimes referred to by the abbreviation MT, is a subfield of computational linguistics that investigates the use of software to translate text or speech from one language to another.
- Fourth Application: Information Retrieval: Information retrieval in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing.
  EX: Google Search is the most famous example of information retrieval.
- Fifth Application: Information Extraction: Information extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. EX: emails and Web pages to reports, presentations, legal documents, and scientific papers.
- Sixth Application: Question Answering: Most example uses: Alexa, Siri, Fake news detection, Classify emails, Predicting disease, Error detection(Cybersecurity), IVR application (Receive voice calls).

- NLP applications can be divided into three types, depending on how difficult each of them :
  - Mostly Solved: Spam detection, Part of speech tagging, Named entity recognition
  - Making good process: Sentiment Analysis, Coreference resolution, Word sense disambiguation (WSD), Parsing, Machine Translation, Information extraction (IE)
  - Still Really Hard: Question Answering (QA), Paraphrase, Summarization, Dialog And one of the most important applications expected to be achieved in the coming years is Chatbot Either writing or voice.

Chapter 2

Basics Of Natural Language Processing

2.1 Libraries

2.1.1 Natural Language Toolkit (NLTK)

- Libraries that are used for NLP: nltk, spacy, re, genism, fast text, NumPy, pandas, sklearn, matplotlib.

- Library is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.

2.1.2 spaCy

- spaCy is a modern Python library for industrial-strength Natural Language Processing. In this free and interactive online course, you'll learn how to use spaCy to build advanced natural language understanding systems, using both rule-based and machine learning approaches.

- There are several important stages in spacy dealing with language processing:
  - Loading and calling the library
  - Building the path pipeline
  - Using tokens
  - Selecting specific words tagging
  - Understanding token attributes

- This graphic shows the path the spaCy pipeline takes (Figure) :

- First, the text is entered, then all the NLP operations are done such as tokenizer, tagger, then an array, followed by the process of identifying objects called Named Entity Recognition (NER), as well as the parser operations Stemming, POS, Lemmatization, and others.

2.1.3 Regular Expression (Re)

A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing)

2.2 Reading Text

2.2.1 Text File (txt)

A TXT file is a standard text document that contains plain text. It can be opened and edited in any text-editing or word-processing program. TXT files are most often created by Microsoft Notepad and Apple TextEdit, which are basic text editors that come bundled with Windows and macOS, respectively.

2.2.2 Comma-separated values (csv)

A Comma Separated Values (CSV) file is a plain text file that contains a list of data. These files are often used for exchanging data between different applications. For example, databases and contact managers often support CSV files.

2.2.3 Tab-separated values (tsv)

A TSV file is a tab-separated values file commonly used by spreadsheet applications to exchange data between databases. It stores a data table in which each record in the table is on a separate line, and data columns are separated by tabs.

2.2.4 Excel (xlsx)

XLSX is a zipped, XML-based file format. Microsoft Excel 2007 and later uses XLSX as the default file format when creating a new spreadsheet. Support for loading and saving legacy XLS files is also included. XLS is the default format used with Office 97-2003.

2.3 Handling PDF

PyPDF2 1.26.0 PDF toolkit Project description A Pure-Python library built as a PDF toolkit. It is capable of: extracting document information (title, author, ...) splitting documents page by page merging documents page by page cropping pages merging multiple pages into a single page encrypting and decrypting PDF files.

- pip install PyPDF2: for installing PyPDF2
- import PyPDF2: for import library PyPDF2
- There are three important functions in the library, read, combine, and write : PdfFileMerger, PdfFileReader, PdfFileWriter

2.3.1 PdfFileReader

dfFileReader in Python offers functions that help in reading viewing the pdf file. It offers various functions using which you can filter the pdf on the basis of the page number, content, page mode, etc.

2.3.2 PdfFileWriter

As for writing a pdf file, the current method is to read pages from a specific file and rewrite them in another file, and thus a file can be made with assemblies of many pages

2.3.3 PdfFileMerger

Is used to merge multiple pdf files together.

2.4 Search in Text

Chapter 3

Natural Language Processing Tools

3.1 Tokenization

- Tokenization is the process of dividing a sentence into several parts (words) called (token) to deal with it and know its type, etc. It is based on separating words from sentences so that each word is alone, and there are two types of it Word Tokenizer i.e. separating words second is Sentence Tokenizer which means separating the sentences, each sentence separately

- The idea of Word Tokenizer is close to the split command, but with some differences. The split command is based on the presence of the space or the symbol through which the decoding will be done, but tokenize is based on the meaning of the word and the difference often appears in the attached words.

- In Japanese and Chinese, there are no spaces between words. So you use the max-match method: which means starting from the beginning of the sentence and specifying the maximum length that can lead to a specific meaning. We start from the next letter and repeat the process if we reach it.

Example: thetabledownthere ; the table down there ; theta bled own there

3.2 Sentence Segmentation

- Sentence Segmentation: It is the division of whole speech into sentences according to its beginning and context Divide sentences into parts.It is a very important process, and it may be done in one of two ways Either with a clear sign that it is the end of a sentence (? ! ,), and sometimes it is done through the use of an indistinct mark such as the dot (.), Which may have other uses such as the end of the abbreviation Dr or decimal numbers and so on.

- End Of Statements (EOS):Therefore, ML models are sometimes used to find out if this is not the case (.) for the end of the sentence or another use? And the ML algorithm can be done manually to find out whether this is the end of the sentence or not

- Also, the length of the word after it, just as looking at the case of the letters before and after the dot indicates whether it is EOS.
  - Case of word with ".": Upper, Lower, Cap, Number
  - Case of the word after ".": Upper, Lower, Cap, Number

- Numeric features
  - Length of word with "."
  - Probability (a word with "." occurs at the end-of-s)
  - Probability (word after "." occurs at beginning-of-s)

3.3 Part Of Speech (POS)

- POS, which determines the type of a word grammatically, is it a verb, a noun, or an adjective, based on the context of the word, the sentence in it, and not the word itself.

- It is based on the fact that the meaning of any word is not in itself but it's content and context and according to the words surrounding it, and therefore it makes an interpretation of each word according to its content, content, and context, and categorizes it among many sections.

3.4 Stemming Lemmatization

- It is a tool that allows stripping any word of all the addition it contains and returning to its source. It is based on the idea of returning the word to its origin and deleting all additions to it, whether at the beginning of the end. Words such as play all belong to the word (playing, player, plays, played).

- It is very useful in getting to know the meaning of words and joining all words of one origin to the same word. In many forms, the total number of it must be calculated after returning it to its original win. The words are repeated, all revolving in the orbit of the word.

- Words such as (book, books, library, writer, scribes, written, kotaib), are all from the original (books), and the same is true in most languages. Also, the sentences (I was riding in the car.) (I was taking a ride in the car.) have the same meaning even if the words are different.

- But sometimes the stemming tool fails to find the root of the word, and it deletes a vowel at the end, while it is an original letter since=¿ sinc. And the Spacey library does not support stemming because it supports a similar feature, which is lemmatization, so let's deal with it from the NLTK library.

- And lemmatization is similar to stemming in the idea, but it is more powerful and effective, as it is not satisfied with removing the extra letters in the words, but by searching for their meaning and basis, for the words were, was, its root is (be) and so on. It also takes into account the meaning of the sentence. The word "meeting" may have its origin "meet" if it is a present verb, and its origin may be the same as "meeting" if it is a noun and not a verb (meaning a meeting).

- The tool returns each word to its origin, with the deletion of any additions to it, whether previous or subsequent, which is what is done in my step stemming lemmatization. So we start with the normalization step, which includes deleting the extra points to turn a word like the U.S.A to US A.

- And also taking into account the letter s plural, capital, and small in the research.

3.5 Named-Entity Recognition (NER)

- It is the abbreviation of Named Entity Recognition, which is for identifying and categorizing important words, such as the names of people or institutions, the names of countries or cities, time, money, percentages, as well as identifying the names of the media from the names (People, companies, cities, currencies, and so on..)

- And from NER. applications
  - The extracted names can be linked with links to them (Wikipedia)
  - Products are tied to specific companies, civic cars are tied to Honda
  - Certain questions are linked to answers elsewhere

- How to Train a Model to Succeed at NER Words, This is done through several steps:
  - Gathering a sufficiently large amount of data
  - Manually make a selection for the type of each word
  - Create an appropriate mechanism to extract the required features from the words
  - Create an ML algorithm to train it on features and identify categories
  - Then the test was conducted, to calculate the extent of its efficiency in determining the categories

- IO Encoding IOB Encoding :
  - IO Encoding: is an abbreviation for Inside Outside Encoding: In which we define the category of each word

  - IOB Encoding: it is an abbreviation for Inside Outside Begin Encoding: We determine whether this name is the beginning of a long name. *Ex: Fred showed Sue Mengqui Huang‘s new painting*

- It was determined that there are words that are names of Per and other words of O. Names are peculiar, as some names are one name for a specific person and some names are two consecutive names for the same person. It is important to determine the features that will be used in training.

- They are:

  - current word
  - The previous or next word, or both
  - POS value of the word

3.6 Stopwords

They are words that spread a lot in writing and we do not want them to be crowned because of their spread and lack of influence on the meaning, and they are often used repeatedly and permanently and do not have a significant impact on the meaning, so they can be deleted, bearing in mind that some words A task has a difference in meaning like the difference between or, not, and.

3.7 Matchers

It is a tool that allows words to be linked together, to make nlp realize that we mean they have the same meaning or in words And it is very important when we are looking for a specific word that is written in more than one way (Hadi, Hady, abdelkader, abd al kader), indifferent letters, but we want to unite them in a specific word.

3.8 Syntactic Structure

- Here we discuss the structure of words, Which draws the relationship between the words in the sentence based on the rules of language and grammatical foundations, and it shows, How dependent is each on the other?

- There are two models of this structure.

  - The first is the constituency circular model : Which depends on dividing the sentence into small parts, including linking each word with the words related to it, then making a structural link between them and between each other.
    It is clear that the sentence is divided as follows:
    John talked to the children about drugs.
    Each part of it is considered an independent sentence, and if you make a rearrangement like this:
    John talked about drugs to the children
    about drugs, John talked to the children
    The meaning will remain clear, even if the sentence is not tactful

- The second type is the reliability structure Which is based on dealing with the most important word in the sentence, then we start by defining what it depends on, and we complete the process as a sentence:
  The boy put the tortoise on the rug
  The most important word here is put, so we make an arrow from the outside to this word And this verb, we see here 3 words that depend on it, which are the subject, the object of it, and the place of the verb, so there are three arrows going from them like this:

3.9 Text Visualization

It is a function that refers to the data and the results of the results tool from the same, transfer, transfer, ESE, transfer, a specific sentence, and we ask to make a visual position for it showing the relationships between words and dependencies or important words, calling a class from the spacy library, specifying the type of the pattern, Is it ent, dep and in the type of relationships the arrow can be circular or straight lines, and the drawing can be displayed on Jupiter only with a submission order or attached to a page on a site with the service order

Chapter 4

Simple Text Processing

4.1 Word Meaning

Here we discuss important terms in the meanings of the words, and their synonyms to get to know the different terms in this field:

- Lemma Form : It is the word after it has been returned to its roots: book, run, sing.
- word form body : It is the word after the additions have been added to it: bookings, runner, sung.
- Meaning type: These are the different meanings that the same word can have:
  - . . . a bank1 can hold the investments in a custodial account. . .
  - ‘’. . . as agriculture burgeons on the east bank2 the river will shrink even more ‘’.
- Words Homographs :
  - A word that has the same letters, but more than one meaning: serve, book, like, bank.
  - The idea of the test is to find two sentences with different meanings:
Ahmed runs 2 km daily.
Ahmed runs the restaurant.
The meaning is ” Ahmad runs 2 km daily at the restaurant .”
- Homophones lyrics : Words that have the same pronunciation, but with different letters, and different meanings: peace: piece, right: write.
- Synonyms words : These are the words that replace each other in most places, such as: Big/large , car/automobile , water/H2O.

- Antonyms : And they are opposites, i.e. words that have the opposite meaning of another word, in a certain perspective. Dark/light, short/long.

- Hyponym Hypernym concept It is a concept of Hypernym assets and branches Hyponym, The car is considered a hyponym for vehicles, Strawberry is a hyponym for fruits. The concept of Hypernym is the opposite of the original, That is, the vehicles are Hypernym for the car, Fruits are Hypernym for strawberries.

4.2 Word Embedding

- It is one of the most important terms in Neuro-Linguistic Programming, which is meant by the matrix of words that each word receives to determine its meaning and to know the extent of convergence or distance. we have to imagine an example.
  Example :We have five different words (patience, man, apple, dog, book).

- We want to make mathematical relationships between them, and equality by asking types of questions :
  Is this thing alive? ; Are you able to speak? ; Is he male? ; Is it tangible? ; Can it be eaten? ; Can it be sold or bought?

- What about questions that don't have a yes or no value, but rather a certain percentage, for example, a question: Is it important for a person?

- The inclusion of words, based on this basis, but on a larger level, the average libraries give us a value of 300 numbers based on Describe each word perfectly.

- These numbers are used to identify the approximate meaning of the word in circulation, and also to compare the words, and know-how close. The word apple comes from the word orange, and how far away are both of them from the word patience. And the creation of the embedding word is not done in this old way but depends on training the model to know the approximate meaning of each.

4.3 Text Vectors

- Account information, Based on another number of words...

- Simple example: If we have the capital of the United States and we want to deduce the capital of Russia, how is this done?

- It is done by drawing countries and capitals in a certain graph. If we imagine that the values of WE are only two values, like this drawing, and the locations, the important countries and cities here, and a subtraction process was applied between America and its capital, which will be (5, -1), then with this difference with the values of Russia, which will be a certain value (10,4)=¿moscow.

- It is also possible to add or subtract the famous sentence, for example, we can say that (x = King – Man + Woman). The value of x here is to find the difference between the king and the man, which will be the attributes of kings, and if we add them to the woman, it will be the value of the queen.

- The idea of subtracting words from each other, or equalizing the differences with each other, means that the difference between the first and second words is a vector, equal In value and direction, from the vector that represents the difference between the third and fourth words.

- how to calculate how close or far away two words are from each other?

  The first approach is the Euclidean distance method:

  The second appraoch is the similarity cosine:

4.4 Word To Vec (Word2Vec)

- word2vec is a two-layer neural network that handles text processing. The input to it is the corpus text, and the output is a set of arrays for the features of the text.

- It is simply a neural network, which is trained on the basis of including words, and its goal is to calculate the importance and value of each word in the sentence.And then, we infer the remainder of the word, The main task of the word2vec tool is to do the grouping of matrices of similar and related words together, which is done through the mathematical similarities of each word, And these similarities and analogies are like (man - boy) == (woman - girl).

- There are two main types of it:

  - First-CBOW Technique : The idea of BOW depends on using a number of words in the text and making numbers 1 and 0 according to the presence of each word, and the idea of NGram depends on the use of one or more of the previous words to infer the next word, and thus by using CBOW. It can be defined as a neural network, but it is used to predict what is the missing word in a given sentence (often the last word) One of the practical applications of CBOW is what is called “article spinning”.Which is intended to make a change in some of the words of the articles, while preserving the meaning, in order to make a kind of quotation from Some sites and put them on other sites, without Google reducing the classification of the second site as a content thief.
  - Second-SkipGram Technique : If the CBOW deduces an incomplete word by eating the words of the sentence, then the gram-skip is the opposite, deducing a

sentence by word. Its idea is based on calculating the relationship between words and each other. In the first time, the relationship was calculated between the first word and the second (is, this) and between the first and the third (a, this). And in the second time, when the focus was on the word is, the relationship between the second and the first, and the second and third, and the second and the fourth. Thus, noting that a maximum limit for the relationship of words with each o ther is specified, here the maximum is 4. The idea originally comes from the idea of the bigram, in which a word can be used to infer a word following it, i.e. from the word jumps.Deduce the word over, but as the idea expands, skipping can be done, i.e. skipping a number of words so that we can deduce the words, The third, fourth or previous, so it is called skipgram. But what is the mechanism by which the words will be chosen, how can it be possible through word processing to deduce the next word for it? There are two methods, a less successful method, which is hierarchical selection, and a more successful method: the negative sample

- Hierarchical Selection: So the idea of hierarchical selection is based on dividing all the words we have in the form of a complete hierarchical division, so that there is Levels in words, so that there are words at higher levels, and others at lower levels, and so on
- Sampling negative: This idea comes to avoid the disadvantages of the previous method. The idea is that instead of training the model so that there is a multi-classification between the correct word and hundreds of thousands of wrong words, we first select the entry word, which is jumps, and then select the correct words, which are (fox, brown,

the, over, (and instead of training it on the rest of the other words in the English language, we say by choosing a number of other words at random, for example, four other incorrect words are chosen randomly, let it be (apple .) Tokyo, boat, orange, (, then train the model by calculating the probability of finding the correct words and increasing them, then Calculate and reduce the probability of misspellings.

4.5 Bag Of Words (BOW)

4.6 Term Frequency-Inverse Document Freque
IDF)

4.7 Text Similarity

4.8 Distributional Similarity

Chapter 5

Advanced Text Processing

5.1 Text Classification

5.2 Text Clustering

5.3 Latent Dirichlet Allocation (LDA)

5.4 Non-Negative Matrix Factorization

5.5 Word for Vectors Global Representation (GloVe)

5.6 NGrams

5.7 Language Modelling

5.8 Text Generation

5.9 Sentiment Analysis

5.10 Naives Bayes

5.11 Vader

5.12 Auto Correct

5.13 Questions Answering

5.14 Summarization Snippet

5.15 Distributional Similarity

Chapter 6

Data Collection

6.1 Tweet Collecting

6.2 Data Scrapping

6.3 Information Retrieval

6.4 Relative Extraction

6.5 Search Engine

Bibliography

[1] Books of Shrii Shrii Anandamurti (Prabhat Ranjan Sarkar):
http://shop.anandamarga.org/

[2] Avtk. Ananda Mitra Ac., *The Spiritual Philosophy of Shrii Shrii Anandamurti: A Commentary on Ananda Sutram*, Ananda Marga Publications (1991)
ISBN: 81-7252-119-7