NLP AND APPLICATIONS

1. WHAT IS NATURAL LANGUAGE PROCESSING

1.1. Natural Language Processing vs Computational Linguistics

Computational study of human language includes:

- Natural language processing (NLP)
- Computational Linguistics (CL)

The differences are in different ways: Level of abstraction, methods used

1) Level of abstraction

A key difference between NLP and Computational Linguistics is the level of abstraction.

NLP

=> practical implementation of algorithms and models that can process natural language

=> to develop methods for solving **practical problems** involving language, such as:

- information extraction
- automatic speech recognition
- machine translation
- sentiment analysis
- question answering
- summarization

- etc.

CL

=> theoretical and mathematical foundations of language processing.

=> employs:

- computational methods **to understand properties** of human language
- How do we understand language?
- How do we produce language?
- How do we learn languages?
- What relationships do languages have with one another ?

2) The methods used

NLP typically relies on machine learning techniques:

- deep learning
- statistical modeling

⇨ to analyze and generate natural language.

CL (Computational Linguistics), on the other hand, relies on a combination of:

- Linguistics
- mathematics,
- computer science

⇨ to develop algorithms and models that can analyze and generate language.

In literature, it is common to see a crossover of methods and researchers, from CL to NLP and vice versa.

Lessons from CL about language can be used in NLP, and

Statistical and machine learning methods from NLP can be applied to answer questions CL seeks to answer.

In this communication, we concern ourselves with only NLP

1.2. The evolution of natural language processing

NLP draws from a variety of disciplines, including computer science and computational linguistics developments dating back to the mid-20th century. Its evolution included the following major milestones:

- **1950s.**

  Alan Turing developed the Turing Test to determine whether or not a computer is truly intelligent

  The test involves automated interpretation and the generation of natural language as criterion of intelligence

- **1950s-1990s.**
  - NLP was largely rules-based
  - Rules developed by linguists to determine how computers would process language.

- **1990s-2000s.**
  - More statistical approach
  - Data-driven natural language processing became mainstream during this decade
  - Shift from a linguist-based approach to an engineer-based approach

- **2000s-present.**
  - Natural language processing saw dramatic growth in popularity as a term.
  - Numerous real-world applications
  - A combination of classical linguistics and statistical methods.

1.3. Why is natural language processing important?

- Businesses use massive quantities of unstructured, text-heavy data and need a way to efficiently process it.
- A lot of the information created online and stored in databases is natural human language, and until recently, businesses could not effectively analyze this data.
- NLP plays a vital part in technology and the way humans interact with it.
- It is used in many real-world applications: chatbots, cybersecurity, search engines, and big data analytics

1.4. How does natural language processing work?

There are two main phases to natural language processing: data preprocessing, algorithm development.

3.4.1. Data preprocessing:

Purposes:

preparing and "cleaning" text data for machines to be able to analyze it.

Ways to do:

- **Tokenization**

  Text is broken down into smaller units to work with.

- **Stop word removal**

  Common words are removed from text so unique words that offer the most information about the text remain.

- **Lemmatization and stemming**

  Words are reduced to their root forms to process

- **Part-of-speech tagging**

Words are marked based on the part-of speech: nouns, verbs and adjectives, etc.

Once the data has been preprocessed, an algorithm is developed to process it.

3.4.2. Algorithm development

Many different natural language processing algorithms: two main types are commonly used:

- **Rules-based system.**

  Carefully designed linguistic rules.

- **Machine learning-based system.**

  They learn to perform tasks based on training data

  Exemple: neural networks

1.5. Techniques and methods of natural language processing

Two main techniques: Syntax and semantic analysis

1.5.1. Syntax

Syntax is about arrangement of words in a sentence to make grammatical sense.

Syntax techniques include:

- **Parsing.**

  Grammatical analysis of a sentence.

  Example: “The dog barked."

  ⇨ breaking this sentence into parts of speech:

dog = noun,

barked = verb.

- **Word segmentation**

  Deriving word forms from a string of text

  Example:

  A person scans a handwritten document into a computer. The algorithm would be able to analyze the text and recognize that the words are divided by white spaces.

- **Sentence breaking**

  This places sentence boundaries in large texts.

  Example:

  "The dog barked. I woke up."

  The algorithm can recognize the period that splits up the sentences using sentence breaking.

- **Morphological segmentation**

  This divides words into smaller parts called morphemes.

  Example:

  The word *untestably* => [[un[[test]able]]ly]

  "un," "test," "able" and "ly" are morphemes.

  This is especially useful in machine translation and speech recognition.

- **Stemming**

  This divides words with inflection in them to root forms.

  Example:

  "The dog barked,"

  The root of the word "barked" is "bark."

This would be useful if a user was analyzing a text for all instances of the word bark, as well as all of its conjugations. The algorithm can see that they are essentially the same word even though the letters are different.

1.5.2. **Semantics**

Semantics is about understanding the meaning and structure of sentences.

Semantics techniques include:

- **Word sense disambiguation.**

  This derives the meaning of a word based on context.

  Example:

  Consider the sentence, "The pig is in the pen."

  The word *pen* has different meanings.

  An algorithm using this method can understand that the use of the word *pen* here refers to a fenced-in area, not a writing implement.

- **Named entity recognition**

  This determines words that can be categorized into groups.

  Example:

  "Daniel McDonald's son went to McDonald's and ordered a hamburger,"

  The algorithm could recognize the two instances of "McDonald's" as two separate entities: one a restaurant and one a person.

- **Natural language generation**

  This uses a database to determine semantics behind words and generate new text.

  Example:

  Automatically generating news articles or tweets based on a certain body of text used for training.

1.6. What is natural language processing used for?

Some of the main functions that natural language processing algorithms perform are:

- **Text classification**
  - Assigning tags to texts to put them in categories
  - Useful for sentiment analysis
  - To help the natural language processing algorithm determine the sentiment, or emotion behind a text.

- **Text extraction**

  Automatically summarizing text and finding important pieces of data.

  One example of this is keyword extraction, which pulls the most important words from the text, which can be useful for search engine optimization.

- **Machine translation**

  This is the process by which a computer translates text from one language, such as English, to another language, such as French, without human intervention.

- **Natural language generation**

  This involves using natural language processing algorithms to analyze unstructured data and automatically produce content based on that data.

  Example: language models such as GPT3

The functions listed above are used in a variety of real-world applications, including:

- customer feedback analysis

  where AI analyzes social media reviews;

- customer service automation

where voice assistants on the other end of a customer service phone line are able to use speech recognition to understand what the customer is saying, so that it can direct the call correctly;

- automatic translation

  using tools such as Google Translate, Bing Translator and Translate Me;

- academic research and analysis

  where AI is able to analyze huge amounts of academic material and research papers not just based on the metadata of the text, but the text itself;

- analysis and categorization of medical records

  where AI uses insights to predict, and ideally prevent, disease;

- word processors used for plagiarism and proofreading

  using tools such as Grammarly and Microsoft Word;

- stock forecasting and insights into financial trading

  using AI to analyze market history and documents, which contain comprehensive summaries about a company's financial performance;

- talent recruitment in human resources

- automation of routine litigation tasks

  one example is the artificially intelligent attorney.

Sentiment analysis is another primary use case for NLP. Using sentiment analysis, data scientists can assess comments on social media to see how their business's brand is performing, or review notes from customer service teams to identify areas where people want the business to perform better.

1.7. Benefits of natural language processing

- Improves the way humans and computers communicate with each other

- improved accuracy and efficiency of documentation;
- ability to automatically make a readable summary of a larger, more complex original text;
- useful for personal assistants such as Alexa, by enabling it to understand spoken word;
- enables an organization to use chatbots for customer support;
- easier to perform sentiment analysis;
- provides advanced insights from analytics that were previously unreachable due to data volume.

1.8. Challenges of natural language processing

- **Precision.**
  - human speech is not always precise
  - it is often ambiguous
  - the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.
- **Tone of voice and inflection**
  - Natural language processing has not yet been perfected.

  For example, semantic analysis can still be a challenge.
  - Other difficulties include the fact that the abstract use of language is typically tricky for programs to understand.

  For instance, natural language processing does not pick up sarcasm easily.

  These topics usually require understanding the words being used and their context in a conversation.

  As another example, a sentence can change meaning depending on which word or syllable the speaker puts stress on.
  - The tone and inflection of speech may vary between different accents, which can be challenging for an algorithm to parse.

- **Evolving use of language**
  - Language and the way people use it is continually changing.
  - Hard computational rules that work now may become obsolete as the characteristics of real-world language change over time.

2. TOOLS FOR NATURAL LANGUAGE PROCESSING

2.1. Programming languages for NLP

2.1.1. LISP

Lisp has long been a favored language in the realm of artificial intelligence.

Lisp, with its symbolic processing capabilities, is well-suited for various Natural Language Processing (NLP) tasks.

Basic NLP tasks with Lisp includes:

- Tokenization
- Part-Of-Speech Tagging
- Stemming
- Named Entity Recognition
- Parsing

More complex NLP applications in Lisp:

- Semantic Analysis
- Coreference Resolution
- Dependency Parsing
- Topic Modeling
- Sentiment Analysis

2.1.2. PROLOG

Prolog is used in NLP because linguists can write natural language grammars almost directly as PROLOG programs. This allows fast-prototyping of NLP systems and facilitates analysis of NLP theories.

To implement natural language processing (NLP) capabilities in a Prolog program, one can use the following steps:

1) Define the facts and rules that govern the domain of knowledge that the NLP algorithm will be able to understand
2) Implement the main entry point for the NLP algorithm
3) Implement the predicates that handle the natural language input
4) Test the NLP algorithm to ensure that it behaves as expected

2.1.3. PYTHON

Nowadays, Python is one of the best choices for natural language processing projects:

- Python has transparent semantics and syntax
- Python developers can enjoy solid support for integration with other languages and tools to build machine learning models.
- Python offers a collection of NLP tools and libraries that enable developers to handle different NLP tasks, including:
    - sentiment analysis
    - POS tagging
    - document classification
    - topic modeling
    - word vectors
    - and more.

2.2. The most frequently used libraries in Python.

NLTK(www.nltk.org/)

TextBlob (http://textblob.readthedocs.io/en/dev/index.html)

SpaCy (https://spacy.io/)

Gensim (https://pypi.python.org/pypi/gensim)

Pattern (https://pypi.python.org/pypi/Pattern)

Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/)

2.2.1. NLTK (www.nltk.org)

NLTK (Natural Language Tool Kit) is the most common Python package for working with corpora, categorizing text, analyzing linguistic structure, and more.

The recommended way of installing the NLT K package:

```
pip install nltk.
```

Tokenizing a given sentence into individual words:

```
import nltk
# Tokenization
sent_ = "I am almost dead this time"
tokens_ = nltk.word_tokenize(sent_)
tokens_
>> ['I', 'am', 'almost', 'dead', 'this', 'time']
```

Getting a synonym of a word

```
# Make sure to install wordnet, if not done already so
# import nltk
# nltk.download('wordnet')
# Synonyms
from nltk.corpus import wordnet
word_ = wordnet.synsets("spectacular")
print(word_)
>> [Synset('spectacular.n.01'), Synset('dramatic.s.02'),
Synset('spectacular.s.02'), Synset('outstanding.s.02')]
# Printing the meaning along of each of the synonyms
print(word_[0].definition())
print(word_[1].definition())
print(word_[2].definition())
print(word_[3].definition())
>> a lavishly produced performance
>> sensational in appearance or thrilling in effect
>> characteristic of spectacles or drama
```

```
>> having a quality that thrusts itself into attention
```

Stemming and lemmatizing words

```
# Stemming
from nltk.stem import PorterStemmer
stemmer = PorterStemmer() # Create the stemmer object
print(stemmer.stem("decreases"))
#The result may not be a real word
>> decreas
#Lemmatization
from nltk.stem import WordNetLemmatizer
# Create the Lemmatizer object
lemmatizer = WordNetLemmatizer()
print(lemmatizer.lemmatize("decreases"))
#The result may is a real word
>> decrease
```

2.2.2. TextBlob

TextBlob (http://textblob.readthedocs.io/en/dev/index.html) is a Python library for processing textual data. It provides a simple API for common NLP tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and much more

TextBlob for sentiment analysis

Sentiment refers to a feeling hidden in the sentence.

Polarity defines negativity or positivity in the sentence

Subjectivity implies whether the sentence discusses something vaguely or with complete surety.

Example:

```
from textblob import TextBlob
# Taking a statement as input
statement = TextBlob("My home is far away from my school.")
# Calculating the sentiment attached with the statement
statement.sentiment
Sentiment(polarity=0.1, subjectivity=1.0)
```

TextBlob to deal with spelling errors

Example

```
sample_ = TextBlob("I thinkk the model needs to be
trained more!")
print(sample_.correct())
>> I think the model needs to be trained more!
```

TextBlob for language translation

Example

```
# Language Translation
lang_ = TextBlob(u"Voulez-vous apprendre le français?")
lang_.translate(from_lang='fr', to='en')
>> TextBlob("Do you want to learn French?")
```

3. ASSOCIATIONS AND RESOURCES FOR NLP

3.1. Associations

3.1.1. ISCA

International Speech Communication Association

Web site:

https://www.isca-speech.org/iscaweb/index.php

Conference

INTERSPEECH

ISCA organizes an annual conference, INTERSPEECH, which integrates two previous series of biennial international conferences: EUROSPEECH, the European Conference on Speech Communication and Technology, and ICSLP, the International Conference on Spoken Language Processing.

Publications

1) *Speech Communication*

http://www.elsevier.com/locate/specom

Speech Communication is a publication of the European Association for Signal Processing (EURASIP), and the International Speech Communication Association (ISCA)

2) *Computer Speech and Language*

https://www.sciencedirect.com/journal/computer-speech-and-language

An official publication of the International Speech Communication Association (ISCA)

3.1.2. ACL

Association for Computer Linguistics

Web site:

https://www.aclweb.org/portal/

Conferences

- Annual Meetings of the Association for Computational Linguistics
- EACL: Annual Meeting of the European chapter of the Association for Computational Linguistics
- EMNLP: Empirical Methods in Natural Language Processing
- NACL: The North American Chapter of the Association for Computational Linguistics
- IJCNLP : International Joint Conference on Natural Language Processing

Journals

- Computational Linguistics (MIT Press)
- Computational Linguistics (MacQuarie University, Australia)

Antology

https://aclanthology.org/

The ACL Anthology currently hosts 88660 papers on the study of computational linguistics and natural language processing.

3.1.3. ELRA

European Language Resource Association

http://www.elra.info/en

A non-profit organization whose main mission is to make Language Resources for Human Language Technologies available to the community at large.

3.2. Language Resource

The term Language Resource refers to:

- a set of speech or language data and descriptions
- in machine readable form
- used for building, improving or evaluating natural language and speech algorithms or systems
- or, as core resources for the software localization and language services industries, for language studies, ...

Where to Find NLP Resources in Languages ?

- **Hugging Face's model hub**

https://huggingface.co
Currently has more than 22,000 models that one can filter by language and application.

- **Kaggle**

https://www.kaggle.com

A repository of community-published models, data and code

3.3. Journals

- **Natural Language Processing Journal, Elsevier**

https://www.sciencedirect.com/journal/natural-language-processing-journal

- **Applied Linguistics, Oxford University Press**

https://academic.oup.com/applij

- **Natural Language Engineering, Cambridge**

https://www.cambridge.org/core/journals/natural-language-engineering

- **IEEE/ACM Transactions on Audio Speech and Language Processing, IEEE**

https://signalprocessingsociety.org/publications-resources/ieeeacm-transactions-audio-speech-and-language-processing

- **ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), ACM**

https://dl.acm.org/journal/tallip

- **Applied Corpus Linguistics, Elsevier**

https://www.sciencedirect.com/journal/applied-corpus-linguistics

- **International Journal of Speech Technology, Springer**

https://www.springer.com/journal/10772

REFERENCES

1. TAWEH BEYSOLOW II, *Applied Natural Language Processing with Python*, Apress Media LLC 2018

2. PALASH GOYAL, SUMIT PANDEY, KARAN JAIN, *Deep Learning for Natural Language Processing*, Apress Media LLC 2018

3. DAN JURAFSKY, JAMES H. MARTIN, *Speech and Language Processing (3rd ed. Draft 2023),* https://web.stanford.edu/~jurafsky/slp3

4. R. KIBBLE, *Introduction to natural language processing*, University of London International Programmes Publications Office, Department of Computing, Goldsmiths 2013

5. HOBSON LANE, HANNES HAPKE, COLE HOWARD, *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python,* 1st Edition, Manning 2019

6. BEN LUTKEVIC, *Natural Language processing*, https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

7. DELIP RAO, BRIAN MCMAHAN, *Natural Language Processing with PyTorch,* O'Reilly Media, Inc., 2019

8. TUTORIALSPOINT,*NLTK*, , https://www.tutorialspoint.com/natural_language_toolkit/