Natural Language Processing

Unit 1: Introduction and Concepts

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

January 2020

Who am I?

Mario Alejandro Campos Soberanis (mario.campos@upy.edu.mx)

Master in Computer Science (UADY)
Research Engineer at SoldAI
Experience as Webmaster, Chief technology officer and Research engineer at SoldAI
Interest in conversational systems, Automatic reasoning and Biologically inspired algorithms

Course Syllabus

- Unit 1: Classical approaches: 31/01/2020
  - Introduction and concepts
  - Preprocessing
  - Lexical analysis
  - Sintactic analysis
  - Semantic analysis
  - Part of speech tagging

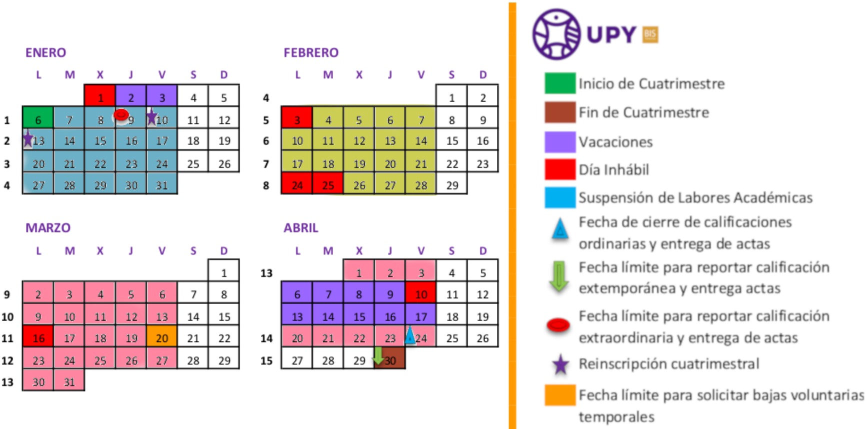
Course Syllabus

- Unit 2: Statistical approaches: 28/02/2020
  - Text Corpora
  - Bag of Words
  - Naive Bayes
  - Classification
  - Evaluation

Course Syllabus

- Unit 3: Deep learning: 21/04/2020
  - Multi layer perceptron
  - Activation functions
  - Loss functions
  - Optimization methods

Calendario



ENERO

	L	M	X	J	V	S	D
			1	2	3	4	5
1	6	7	8	●9	★10	11	12
2	★13	14	15	16	17	18	19
3	20	21	22	23	24	25	26
4	27	28	29	30	31		

FEBRERO

	L	M	X	J	V	S	D
4						1	2
5	3	4	5	6	7	8	9
6	10	11	12	13	14	15	16
7	17	18	19	20	21	22	23
8	24	25	26	27	28	29	

MARZO

	L	M	X	J	V	S	D
							1
9	2	3	4	5	6	7	8
10	9	10	11	12	13	14	15
11	16	17	18	19	20	21	22
12	23	24	25	26	27	28	29
13	30	31					

ABRIL

	L	M	X	J	V	S	D
13			1	2	3	4	5
	6	7	8	9	10	11	12
	13	14	15	16	17	18	19
14	20	21	22	23	▲24	25	26
15	27	28	29	↓30			

UPY

- 🟩 Inicio de Cuatrimestre
- 🟫 Fin de Cuatrimestre
- 🟪 Vacaciones
- 🟥 Día Inhábil
- 🟦 Suspensión de Labores Académicas
- ▲ Fecha de cierre de calificaciones ordinarias y entrega de actas
- ↓ Fecha límite para reportar calificación extemporánea y entrega actas
- ● Fecha límite para reportar calificación extraordinaria y entrega de actas
- ★ Reinscripción cuatrimestral
- 🟧 Fecha límite para solicitar bajas voluntarias temporales

About the course

- Homework
  - Individual
  - Teams (2 persons)

About the course

- Homework
  - Individual
  - Teams (2 persons)
- Evaluation
  - Participation
  - Assignments (40 %)
  - Exam and projects (60 %)

About assignments

- Deadline weekly (if an assignment is requested on monday the deadline is next monday before 23:59:59 email/schoology time)

About assignments

- Deadline weekly (if an assignment is requested on monday the deadline is next monday before 23:59:59 email/schoology time)
- Format
  - Reports/Essays/Presentations: PDF
  - Programming assignments: Jupyter Notebok (.ipynb)
  - Projects: Python code (.py)

About assignments

- Deadline weekly (if an assignment is requested on monday the deadline is next monday before 23:59:59 email/schoology time)

- Format
  - Reports/Essays/Presentations: PDF
  - Programming assignments: Jupyter Notebok (.ipynb)
  - Projects: Python code (.py)

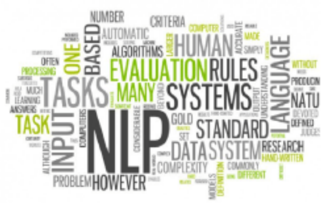
- Naming Individual:
  NLP_{*homework_no*}_{*last_name*}_{*first_name*}.{*file_extension*}
  Team:
  NLP_{*homework_no*}_{*team*}_{*last_names*}.{*file_extension*}
  examples: NLP_01_Campos_Mario.pdf,
  NLP_03_TeamA_Campos_Soberanis_Perez.pdf

Enrole the course

Enrole the schoology course:

CGG9-MGWT-TQDJB

What is Natural Language Processing?



What NLP is all about?

Natural Language Processing

“Natural Language Processing (NLP) is the interdisciplinary field of study between artificial intelligence, linguistics and computer science whose goal is to make computers perform useful tasks that involve human language”

What is NLP used for?

- Allow communication between humans and machines (conversational agents)

What is NLP used for?

- Allow communication between humans and machines (conversational agents)
- Enhance communication between humans automatic translation)

What is NLP used for?

- Allow communication between humans and machines (conversational agents)
- Enhance communication between humans automatic translation)
- Make useful processing of text and speech (ortographic correction)

Why is NLP important?

Why is relevant?

- Increment of enterprise, commercial and industrial applications using natural language (search, publicity, translation, automatic speech recognition, automatic support).

Why is relevant?

- Increment of enterprise, commercial and industrial applications using natural language (search, publicity, translation, automatic speech recognition, automatic support).
- Human language as universal communication paradigm (Siri, Google Assistant, Cortana, Messenger, Alexa).

Why is relevant?

- Increment of enterprise, commercial and industrial applications using natural language (search, publicity, translation, automatic speech recognition, automatic support).
- Human language as universal communication paradigm (Siri, Google Assistant, Cortana, Messenger, Alexa).
- Tool to obtain knowledge of a bunch of unestructured data.

Why is hard?

NLP
TASKS
EVALUATION
RULES
SYSTEMS
LANGUAGE
HUMAN
STANDARD
DATA
SYSTEM
INPUT
TASK
MANY
ONE
BASED
ALGORITHMS
AUTOMATIC
NUMBER
CRITERIA
COMPUTER
GOLD
RESEARCH
COMPLEXITY
HOWEVER
PROBLEM
PROCESSING
DIFFERENT
NATU

Why is hard?

Human language is an efficient communication system, most things are not sayed but implicitly understood

Why is hard?

Human language is an efficient communication system, most things are not sayed but implicitly understood

- Computational representation complexity
  - People use discourse, computers data and commands (NLP tries to close that gap)

Why is hard?

Human language is an efficient communication system, most things are not sayed but implicitly understood

- Computational representation complexity
  - People use discourse, computers data and commands (NLP tries to close that gap)
- Human lnaguage is inheretibily ambiguous
  - Evolution, slang, regionalisms
  - Humor and sarcasm
  - Writing and grammatical errors

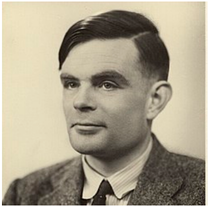
Why is hard?

Human language is an efficient communication system, most things are not sayed but implicitly understood

- Computational representation complexity
  - People use discourse, computers data and commands (NLP tries to close that gap)
- Human lnaguage is inheretibily ambiguous
  - Evolution, slang, regionalisms
  - Humor and sarcasm
  - Writing and grammatical errors

The perfect understanding of the human language is an AI-complete problem.

Turing Test



"A computer can be considered intelligent if it's able to hold a conversation with a human being without realizing to be talking with a machine"

— Alan Turing

Ambiguity

- I saw the mountains flying to New York
- After the death, the miners refuse to work
- In Mexico a woman gives birth every 15 minutes
- The officer shot the man with the knife

Lost in translation

"The spirit is willing, but the flesh is weak"
Translates to:

Lost in translation

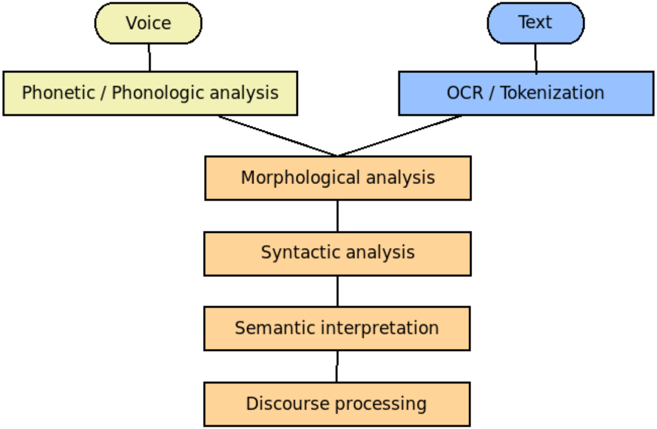
"The spirit is willing, but the flesh is weak"
Translates to:

"The vodka is agreeable, but the meat is rotten"

Combinatorial space for words

- A highschool student knows around 60,000 words
- Almost each sentence produced by a person is a combination generated for the first time in it's life.

NLP Levels


Voice
Text
Phonetic / Phonologic analysis
OCR / Tokenization
Morphological analysis
Syntactic analysis
Semantic interpretation
Discourse processing

Pre-processing

- Cleaning
  - Deletion of empty meaning words (stopwords)
  - Capitalization
  - Processing of characters and symbols

Pre-processing

- Cleaning
  - Deletion of empty meaning words (stopwords)
  - Capitalization
  - Processing of characters and symbols
- Annotation
  - Grammar labeling (Part Of Speech tagging)
  - Structural marking

Pre-processing

- Cleaning
  - Deletion of empty meaning words (stopwords)
  - Capitalization
  - Processing of characters and symbols
- Annotation
  - Grammar labeling (Part Of Speech tagging)
  - Structural marking
- Normalization
  - Stemming
  - Lematizing

Pre-processing

- Cleaning
  - Deletion of empty meaning words (stopwords)
  - Capitalization
  - Processing of characters and symbols
- Annotation
  - Grammar labeling (Part Of Speech tagging)
  - Structural marking
- Normalization
  - Stemming
  - Lematizing
- Others
  - Tokenizing / Segmentation
  - Counting and grouping

Main approaches

- Rule based methods
  - Regular expressions
  - Free context Grammars
  - First order logic

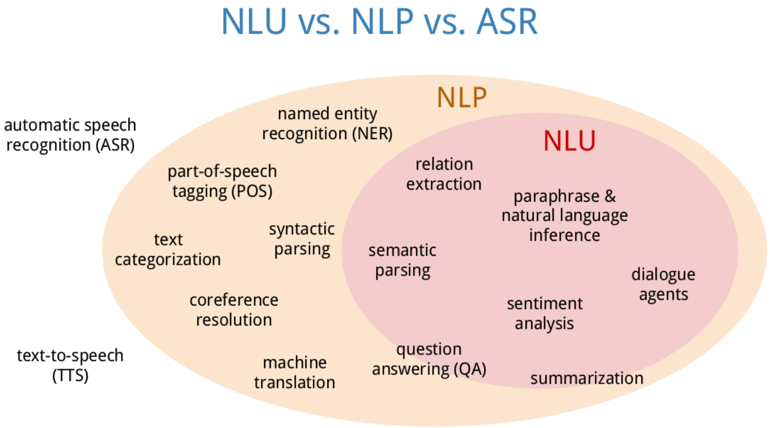
Main approaches

- Rule based methods
  - Regular expressions
  - Free context Grammars
  - First order logic
- Probabilistic models and Machine Learning
  - Maximum likelihood
  - Linear classification
  - Markov hidden models

Main approaches

- Rule based methods
  - Regular expressions
  - Free context Grammars
  - First order logic
- Probabilistic models and Machine Learning
  - Maximum likelihood
  - Linear classification
  - Markov hidden models
- Deep learning
  - Representation Learning
  - Embeddings
  - Convolutional, Recursive, Long Short Term Memory and Recurrent Neural Networks

Task terminology



NLU vs. NLP vs. ASR

NLP

NLU

automatic speech recognition (ASR)

named entity recognition (NER)

part-of-speech tagging (POS)

relation extraction

paraphrase & natural language inference

text categorization

syntactic parsing

semantic parsing

dialogue agents

coreference resolution

sentiment analysis

text-to-speech (TTS)

machine translation

question answering (QA)

summarization

Some interesting applications

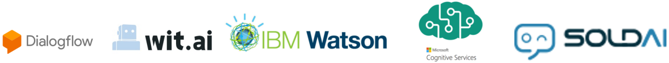
- Sentiment analysis
- Ortographic correction
- Search engines
- Information extraction
- Document classification
- Automatic translation
- Dialog systems and digital assistans
- Automatic question answering
- Natural language database interfaces
- Automatic summary

Tools

- Libraries
  - NLTK (Natural Language Toolkit)
  - Stanford CoreNLP
  - Apache OpenNLP
  - Spacy
  - Keras
  - Tensorflow
  - Pytorch
  - fastai

Tools

- Platforms



Assignments

Assignment 1: Write a report about two of the following NLP tasks:

- Automatic speech recognition
- Dialogue agents
- Sentiment analysis
- Question answering

The report must include:

- Applications
- Approaches to solve the task
- Commercial products using it
- References

Let’s code

References

[1] Jurafsky, D., Martin, J.: Speech and Language Processing 2nd. ed. (2009).

[2] Bird S., Klein E., Loper, E.: Natural Language Processing with Python, (2009). ISBN: 978-0-596-51649-9

[3] Indurkhya N., Damerau F. Handbook of Natural Language Processing, Second Edition (2010). ISBN: 978-1-4200-8593-8

[4] Kao, A., Poteet S. Natural Language Processing and Text Mining (2007). ISBN: 78-1-84628-175-4

[5] Mikolov, T., Corrado, G., Chen, K. y Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: CoRR, (2013). https://dblp.org/rec/bib/journals/corr/abs-1301-3781

[6] Pinker, S.: The Stuff of Thought - Language as a window into human nature. https://www.youtube.com/watch?v=5S1d3cNge24