MINICURS:
INTRODUCCIÓ A LA INTEL·LIGÈNCIA ARTIFICIAL
APLICADA ALS SISTEMES COMPLEXOS

APRENETATGE AUTOMÀТИC
PER AL PROCESSAMENT DEL
LLENGUATGE NATURAL

Professora: Dra. Mireia Farrús Cabeceran
Facultat de Filologia i Comunicació, UB

http://ubics.ub.edu
http://ubics.ub.edu/AI_course
Natural Language Processing

Process / analysis of natural language data
Advanced interaction with computers
- in natural language
Linguistics, Computer Science, Artificial Intelligence

Turing test: Are we talking to a machine or to a human?

ELIZA

Virtual assistants:

LUNA, Google Duplex Assistant AI
Language technologies

Natural Language Processing

Signal / Speech Processing

Applications in NLP
Text

Machine translation
Information extraction
  - summarisation
  - document classification
Information retrieval: question answering
Language identification
Dialogue systems
Sentiment analysis and Opinion mining
Speech

Automatic speech recognition
Speaker recognition
Speaker diarisation
Speech synthesis (text-to-speech)
Speech-to-speech translation
Language/dialect/sociolect identification

Medical applications - Accessibility
Tasks

Written Language Dialogue

Syntactic Parsing → Semantic Analysis

Semantic Analysis → Dialogue Management

Dialogue Management → Natural Lang. Generation

Semantic Analysis:
- Reasoning: What is the concern of the user?
- Reasoning: How shall the system react?
- How shall the system present its reaction?

Syntactic Parsing:
- Syntactic str.

Written Language:
- Input
- Output
Tasks

Spoken Language Dialogue

- Automat. Speech recognition (ASR)
- Syntactic Analysis
  - Synt. str
  - Reasoning: What is the concern of the user?
  - Reasoning: How shall the system react?
  - How shall the system present its reaction?
- Semantic Analysis
  - Semantic str.
- Dialogue Management
  - Semantic str.
- Natural Lang. Generation
  - Text
- Speech synthesis (TTS)
  - Text
  - Speech
Tasks

Multimodal Dialogue
Architecture: dialog system
Natural language: Challenges

Language is **discrete, compositional, and sparse**

Characters $\rightarrow$ words (objects, concepts, events, actions, ideas)

Letters form words, words form phrases and sentences

Words can be combined in infinite ways to form meanings

Speech: + prosody!
Natural language: Challenges

NLP includes:

1. algorithms that take human-produced text as input (analysis).
2. algorithms that produce natural looking text as outputs (generation).
   – The need for such algorithms is ever increasing.

NLP is challenging: human language is inherently ambiguous, ever changing, and not well defined.

Natural Language is:

1. **symbolic** in nature → based on logic, rules and ontologies.
2. also **highly ambiguous/variable** → statistical approach
Natural language: Challenges

Ambiguity: all levels of analysis

(1) Les cases.
   (les DET cases NOM ) SN   ((les PRO) SN, OD cases V)SV)F
(2) Van veure l’avió volant a Nova York.
(3) Van atracar el port / Van atracar al port.
(4) Va donar un pastís de xocolata als nens de P3.
(5) (Els seus amics)i Van comprar [unes maduixes]j, Øi lesj
   van posar a la nevera i Øi se lesj van menjar.
Levels of a linguistic structure

<table>
<thead>
<tr>
<th>Unit</th>
<th>Structure</th>
<th>Meaning</th>
<th>Applications in NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound</td>
<td>Phonetics</td>
<td>Phonetic structure of words &amp; utterances</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prosody</td>
<td>Meaning of sounds and sound contours</td>
</tr>
<tr>
<td>Word</td>
<td>Morphology</td>
<td>Structure of words</td>
<td>Lexical semantics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Meaning of words</td>
</tr>
<tr>
<td>Sequence of words</td>
<td>Syntax</td>
<td>Structure of sentences</td>
<td>Communicative semantics</td>
</tr>
<tr>
<td>Sentence</td>
<td>Syntax</td>
<td>Structure of sentences</td>
<td>Semantics</td>
</tr>
<tr>
<td>Text</td>
<td>Discourse</td>
<td>Text structure</td>
<td>Discourse semantics</td>
</tr>
</tbody>
</table>

Applications in NLP
Deep Learning

• Branch of machine learning
• Inspired by the way computation works in our brain, and which can be characterised as learning of parameterised differentiable mathematical functions.
  – Many layers of these differentiable functions are often chained together

<table>
<thead>
<tr>
<th>Traditional ML</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning to make predictions based on past observations</td>
<td>Learning not only to predict but also to correctly represent the data, such that they are suitable for prediction</td>
</tr>
</tbody>
</table>
Simple Decisions

Speaker Verification / Identification

Is she Clara?

unknown user

impostor

DATABASE

Who is he?

unknown user

user 1

user 2

... 

user N

DATABASE

He is user i

closed set identification

He is user i

or

Unidentified user

open set identification

Applications in NLP
Complex Decisions

Machine Learning in NLP

• Machine Translation
• Automatic Speech Recognition
• Speech Synthesis
“[...] MT has to deal with almost every imaginable problem in computational linguistics, in at least two languages. Because of the necessity for this broad coverage, many of the problems (and indeed solutions) that we will mention in the following may also be found in other areas of NLP. MT has the special position of having to bring them all together.” (H. Somers 2000)
Rule-based MT

Linguistic information about source/target languages
Monolingual/Bilingual/Multilingual...
  - dictionaries
  - grammars

Morphological/Syntactic/Semantic analysis of the languages involved

e.g. SYSTRAN, Apertium
## Rule-based MT

### Monolingual lexicons

<table>
<thead>
<tr>
<th>Source language</th>
<th>Target language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LEMA</strong> = LOVE</td>
<td><strong>LEMA</strong> = LOVE</td>
</tr>
<tr>
<td><strong>ARREL</strong> = LOVE</td>
<td><strong>ARREL</strong> = LOVE</td>
</tr>
<tr>
<td><strong>CAT</strong> = VERB</td>
<td><strong>CAT</strong> = VERB</td>
</tr>
<tr>
<td><strong>C_FLEXIVA</strong> = RG</td>
<td><strong>C_FLEXIVA</strong> = AR</td>
</tr>
<tr>
<td><strong>SUBCAT</strong> = SUBJ +Nom + Humà OBJ +Nom +/- Humà</td>
<td><strong>SUBCAT</strong> = SUBJ +Nom - Humà OI PREP=A +Humà</td>
</tr>
<tr>
<td><strong>SEM</strong> = PSICOLÒGIC</td>
<td><strong>SEM</strong> = PSICOLÒGIC</td>
</tr>
</tbody>
</table>

**Example:**

- **LEMÀ** = AGRADAR
  - **ARREL** = AGRAD
  - **CAT** = VERB
  - **C_FLEXIVA** = AR
  - **SUBCAT** = SUBJ +Nom - Humà OI PREP=A +Humà
  - **SEM** = PSICOLÒGIC
Rule-based MT

Bilingual or transfer lexicons

**LEMA** = LOVE
**ARREL** = LOVE
**CAT** = VERB
**C_FLEXIVA** = RG
**SUBCAT** = SUBJ +Nom + Humà
**OBJ** +Nom +/- Humà
**SEM** = PSICOLÒGIC

Condicions de selecció: ?**OBJ** = +Humà
Accions associades: !**ESTIMAR**

**LEMA** = AGRAD
**ARREL** = AGRAD
**CAT** = VERB
**C_FLEXIVA** = AR
**SUBCAT** = SUBJ +Nom - Humà
**OBJ** +Nom - Humà
**OI PREP**=A +Humà
**SEM** = PSICOLÒGIC

Condicions de selecció: ?**OBJ**= -Humà
Accions associades: !**AGRADAR**

SUBJ +Nom + Humà  →  **OI Prep** + Humà
OBJ +Nom –Humà  →  SUBJ +Nom -Humà
Statistical-based MT

CATALAN/ENGLISH bilingual (parallel) text

statistical analysis

source text

possible translations

quim fred que tinc

have I that cold

ENGLISH text

statistical analysis

most likely translation

I am so cold
cold I am so

...
Statistical-based MT

CATALAN/ENGLISH bilingual (parallel) text → statistical analysis → possible translations → most likely translation

ENGLISH text → statistical analysis → possible translations → most likely translation

Source text → Translation model

Language model

Decoding algorithm: \( \text{argmax } p(e) \cdot p(c|e) \)
Statistical-based MT

Translation model

Given a pair of strings \(<c, e>\), computes \(P(c|e)\)

\[\text{High } P(c|e) \rightarrow \text{good translation}\]

\[\text{Low } P(c|e) \rightarrow \text{poor translation}\]

Language model

Given an English string \(e\), computes \(P(e)\)

\[\text{High } P(e) \rightarrow \text{good string}\]

\[\text{Low } P(e) \rightarrow \text{bad string}\]

Decoding algorithm

Given a language model, a translation model and a new sentence \(c\), find translation \(e\) maximizing \(P(c|e) \cdot P(e)\) (Bayes’ rule)

\[P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad P(e|c) = \frac{P(c|e)P(e)}{P(c)}\]
Statistical-based MT

• Word-based translation

• Phrase-based N-grams

\[ P(W) = P(w_1, w_2, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) ... P(w_n|w_1, w_2, ..., w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \]

**Approximation:** Markov model of order N-1 (preceding N-1 words):

\[ P(W) \approx \prod_{i=1}^{n} P(w_i|w_{i-N+1}, w_{i-N+2}, ..., w_{i-1}) \]

**Unigram model**

\[ P(\text{cats sleep a lot}) = P(\text{cats}) \cdot P(\text{sleep}) \cdot P(\text{a}) \cdot P(\text{lot}) \]

**Bigram model**

\[ P(\text{cats sleep a lot}) = P(\text{cats|<START>}) \cdot P(\text{sleep|cat}) \cdot P(\text{a|sleep}) \cdot P(\text{lot|a}) \cdot P(\text{<END>|lot}) \]

• Syntax-based

Applications in NLP
Neural MT

- Word embeddings
- End-to-end architectures
- Attention mechanism

Word embeddings
Vectorial representations of words reflecting their similar contextual usages.

https://medium.com/@hari4om/word-embedding-d816f643140
Neural MT

One-hot versus dense vector representation

Embedding layer: mapping of discrete symbols to continuous vectors in a relatively low dimensional space.

<table>
<thead>
<tr>
<th>Words</th>
<th>Embedded words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolated distinct symbols</td>
<td>Mathematical objects that can be operated on</td>
</tr>
</tbody>
</table>

This representation of words as vectors is learned by the network as part of the training process.

- The network also learns to combine word vectors in a way that is useful for prediction.
- This capability alleviates discreteness and data-sparsity problems.
Neural MT

End-to-end architectures
- No separate modules for language model, translation model, etc.
- Single sequence model predicting one word at a time.
- RNN1: \textit{encoder} $\rightarrow$ source sentence representation
  RNN2: \textit{decoder} $\rightarrow$ word prediction in the target language
- Later: CNN
Types of NN

1. Feed-forward networks
   - Multi-layer perceptrons (MLPs)
     Allow to work with fixed sized inputs, or with variable length inputs in which we can disregard the order of the elements
   - Convolutional networks (CNNs)
     Extract meaningful local patterns that are sensitive to word order, regardless of where they appear in the input
     (e.g. Identifying indicative phrases or idioms in long sentences or documents)

2. Recurrent/recursive networks (RNNs)
   - Recurrent NN
     Specialised models for sequential data
     Rarely used as a standalone component. Trainable components that can be fed into other network components
   - Recursive NN
     Generalisation of Recurrent NN that can handle trees.
**Types of NN**

1. **Feed-forward networks**
2. **Recursive/recurrent networks**


Applications in NLP
Neural MT

Attention mechanism

RNNs have a memory problem in long-range dependencies (words they saw a long time ago that are somehow related to the next word)

- **Attention** fixes that

Looks over all the information contained in the original sentence, and generates the proper word taking the context into account.

Puts *the same sentence* along the columns and rows (matrix) to understand how some parts of the sentence relate to others. e.g. where are my pronouns’ antecedents? (“self-attention”).

https://www.tensorflow.org/tutorials/text/nmt_with_attention
Neural MT

Transformer (2017)
- Handle sequential data (as RNNs)
- But...
  - do not require the sequential data to be processed in order
  - allows for parallelization
  - enables training on larger datasets
    - pre-trained systems, e.g. BERT (Bidirectional Encoder Representations for Transformers)
      - fine-tuned to specific language tasks

Model of choice for many NLP tasks
Automatic Speech Recognition

History...

1922 Radio Rex (Elmwood Button Company) : dog’s name
1952 First true speech recogniser (Bell Laboratories): digits
1959 Denes & Fry, bigram language model: 4 vowels & 9 consonants, words
The Problem – The Solution

Text

$X = (x_1, x_2, \ldots, x_t)$  

Acoustic evidence

$W = (w_1, w_2, \ldots, w_t)$  

Word or sequence of words

Template framework  1980s  Statistical framework

Dynamic Time Warping (DTW)  
Vector Quantization (VQ)

$P(W|X)$  

Probability that the words $W$ were spoken given that the acoustic evidence $X$ was observed

Bayes’ rule:  

$P(W|X) = \frac{P(W)P(X|W)}{P(X)}$

$\widehat{W} = \arg\max_{W \in \omega} P(W)P(X|W)$

Applications in NLP
Statistical Framework

Four subproblems (steps) to be addressed:
1. Feature Extraction (Signal Processing):
   \[ P(X|W) \]
2. Acoustic Modeling:
3. Language Modeling:
   \[ P(W) \]
4. Decoding:
   \[ \tilde{W} = \arg\max_{W} P(W)P(X|W) \]
Step 1: Feature Extraction

**Sounds into bits:** turning sound waves into numbers

Waveform: representation of amplitude over time. At every moment in time, sound waves have a single value based on the height of the wave.

We sample the wave, recording the height of the wave at equally-spaced points. For speech recognition, a sampling rate of 16kHz (16,000 samples per second) is enough to cover the frequency range of human speech.

We group our sampled data into e.g. 20-millisecond-long chunks. (In the picture, the first chunk).

Feature Extraction

Plotting the numbers to obtain a rough approximation of the original sound wave for this period of time.

Breaking the complex sound wave into its component parts (low-F0 parts, next lowest-F0 parts, etc.) by using the Fourier transform, and then adding up how much energy is contained in each one.

**Result:** score of how important each frequency range is, from low to high F0.

**Example:** each number represents how much energy is in each 50hz band of our 20 ms audio.
Spectrogram

Why are **spectrograms** so important?

**How to read a spectrogram**

1. Fourier transform of a signal excerpt.
2. Map powers of the spectrum obtained onto the **mel scale** (approximates the human auditory system’s response)
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers.
5. The **Mel Frequency Cepstrum Coefficients (MFCC)** are the amplitudes of the resulting spectrum. (representation of the short-term power spectrum of a sound).

Fayek, 2016
Step 2: Acoustic Modeling

Markov Chains

Markov property: the probability of being in any particular state only depends on the previous state it was at before.

\[ P[q_t=S_j|q_{t-1}=S_i, q_{t-2}=S_k, ...] = P[q_t=S_j|q_{t-1}=S_i] \]

At every state, a given observable event is produced (knowing the state we are at, we know the event)

Hidden Markov Models (HMM) (1967)
Step 3: Language Modeling

This new display can recognize speech
This nudist play can wreck a nice beach

\[ P(W) = P(w_1, w_2, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) ... P(w_n|w_1, w_2, ..., w_{n-1}) \]

\[ = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \]

**N-gram Language Models:**

**Aproximation:** Markov model of order N-1 (preceding N-1 words):

- **Unigram model**
  \[ P(\text{cats sleep a lot}) = P(\text{cats}) P(\text{sleep}) P(\text{a}) P(\text{lot}) \]

- **Bigram model**
  \[ P(\text{cats sleep a lot}) = P(\text{cats} | \text{\langle START\rangle}) P(\text{sleep} | \text{cat}) P(\text{a} | \text{sleep}) P(\text{lot} | \text{a}) P(\text{\langle END\rangle} | \text{lot}) \]
Step 4: Decoding

**Objective:** Search for the most likely word sequence $W$ given some observed acoustic data $X$:

$$\hat{W} = \arg\max_{W \in \omega} P(W)P(X|W)$$

**Solution:** The Viterbi algorithm

Dynamic programming algorithm for finding the most likely sequence of hidden states—called the Viterbi path—that results in a sequence of observed events. (Andrew Viterbi, 1967)
Deep Learning ASR

Towards *end-to-end* models

1. Connectionist Temporal Classification
2. Sequence-to-Sequence
3. Online Sequence-to-Sequence

Some pictures taken from https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a and https://heartbeat.fritz.ai/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380
Deep Learning ASR

Connectionist Temporal Classification (CTC)
Maps from acoustic frames to character sequences

- Mapping of each audio slice to the most likely spoken letters:

Softmax over vocabulary \( \{a, b, c, d, e, f, \ldots, z, ?, !, \ldots\} \) and extra token \(<b>\).

Softmax at step, \( t \), gives a score \( s(k,t) = \log Pr(k, t|X) \) to category \( k \) in the output at time \( t \).
Deep Learning ASR

Sequence-to-sequence (seq2seq)

Makes next-step predictions

Listen, Attend, and Spell (LAS) model

- Needs an attention mechanism!


Applications in NLP
Speech Synthesis (TTS)

**Articulatory:** Models the movements of articulators and acoustics of vocal tract.

**Formant:** Starts with acoustics, creates rules & filters to create each formant.

**Concatenative:** Uses databases of stored speech to assemble new utterances.
  - Diphone / Unit selection

**Statistical:** Trains parameters on databases of speech (HMM)

**Deep learning:** Neural network architectures
TTS Systems

1990’s – present: concatenative, statistical, deep learning

- Diphone synthesis
- Unit selection synthesis
- HMM synthesis
- NN synthesis
TTS Architecture

- He stole $100 million from the Banks
- It’s 13 St. Andrews St.
- The homepage is http://www.upf.edu
- Yes, see you the following tues, that’s 11/12/2016

<table>
<thead>
<tr>
<th>ARPABET</th>
<th>TIMT</th>
<th>CMU</th>
<th>WSJ</th>
<th>SWB</th>
<th>ICSI</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>as</td>
<td>lOck</td>
<td>lOck</td>
<td>lOck</td>
<td>lOck</td>
<td>sOck</td>
<td>eC</td>
</tr>
<tr>
<td>oc</td>
<td>bAt</td>
<td>bAt</td>
<td>bAt</td>
<td>bAt</td>
<td>bAt</td>
<td>eT</td>
</tr>
<tr>
<td>ah</td>
<td>bUt</td>
<td>bUt</td>
<td>bUt</td>
<td>bUt</td>
<td>bUt</td>
<td>eT</td>
</tr>
<tr>
<td>ao</td>
<td>bOught</td>
<td>bOught</td>
<td>/</td>
<td>bOught</td>
<td>bOught</td>
<td>o</td>
</tr>
<tr>
<td>ew</td>
<td>cOWn</td>
<td>cOW</td>
<td>cOW</td>
<td>cOW</td>
<td>tOWn</td>
<td>eu</td>
</tr>
<tr>
<td>swh</td>
<td>/</td>
<td>/</td>
<td>cOW</td>
<td>/</td>
<td>/</td>
<td>eu</td>
</tr>
<tr>
<td>ox</td>
<td>About</td>
<td>/</td>
<td>/</td>
<td>About</td>
<td>About</td>
<td>ə</td>
</tr>
</tbody>
</table>

Table 1: PHONEME MAPPINGS

row text in

Text analysis
- Text Normalization
- POS tagging
- Homograph disambiguation

Phonetic analysis
- Dictionary lookup
- Grapheme-to-phoneme

Prosodical analysis
- Boundary placement
- Pitch and accent assignment
- Duration computation

Waveform synthesis

Speech out

Applications in NLP
Deep Learning TTS

**WaveNet** (2016) “changes the other paradigms [unit selection and statistical] by directly modelling the raw waveform of the audio signal, one sample at a time.”

**SampleRNN** (2017)

**Tacotron/Tacotron2** (2017)

---

**VOCODERS**

- Griffin-Lim (2017)
- Waveglow (2018)
- MelGAN (2019)
Q1. What types of food are a good source of vitamins?

Q2. Are legumes a source of vitamins?

Q3. I’ve heard that legumes are healthy, but what are they a good source of?

Where did he come from?
We do not know.
From far off.

Paragraph structure
Discourse structure
Narrative structure

Applications in NLP
### Data

- **Switchboard**
- **LibriSpeech** (ASR)
- **LibriTTS**
- **LJ Speech Dataset**
- **CommonVoice**

### My own system

#### Machine Translation
- **OpenNMT**
- **ModernNMT**

#### ASR
- **Wav2letter**
- **Fairseq**
- **ASR-pytorch, (colab)**

#### TTS
- **Tacotron (colab)**

#### CollectivaT
- **Workshops**
- **Catotron (demo)**
Primary Literature