Introduction to AI

NATURAL LANGUAGE PROCESSING

Pedro Meseguer
Institut d'Investigació en Intel·ligència Artificial (IIIA)
Consejo Superior de Investigaciones Científicas (CSIC)

1

Contents

1. Introduction
2. Technical limitations: Markov chain, Bayes'th. (cond. prob.)
3. Language models
4. 3-gram character model: lenguaje identification
5. 3-gram word model: text classification (spam detection)
6. Word Embeddings
7. Machine Translation
8. LLMs: ChatGPT
9. Wrap-up

2

1.Introduction: some dates on evolution

		Equivalence
Age of the Earth	4,600 x 10^6 years	*365 days*
Life starts	3,500 x 10^6 years	*March 29*
Dinosaurs extintion	66 x 10^6 years	*Dec 26*
Homo sapiens	200 x 10^3 years	*Dec 31, 23:37*
Spoken language	100 x 10^3 years	*Dec 31, 23:48:30*
Written language	5 x 10^3 years	*Dec 31, 23:59:26*

Language
- Very recent (in evolution terms)
- Key aspect in the development of intelligence

3

1.Introduction

- Natural Language Processing (NLP): why?

- Natural language, whether spoken, written, or typed, is the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce. As computers play a larger role in the preparation, acquisition, transmission, monitoring, storage, analysis, and transformation of information, endowing them with the ability to understand and generate information expressed in natural languages becomes more and more necessary.

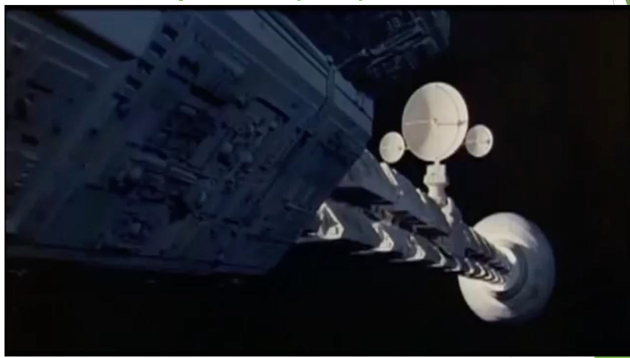
4

1.Introduction (II)

- Specific reasons:
  1. For humans to communicate with computers: commands, instructions…
  2. To learn from humans: documents, Wikipedia…
  3. To help humans
     1. Automatic translation
     2. Information extraction
     3. Question answering

5

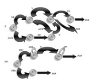
1.2001, a Space Odyssey

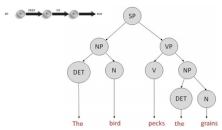


6

1.Introduction (III)

- An old AI goal, many techniques:
  - Grammar
  - Special (ATN) grammars
  - Parsing, Semantics, Pragmatics
  - Statistical methods
  - Neural (deep learning) approach
- Why so elusive?
  - Ambiguous
    - *el banco cerca del río*
    - *the bank/bench by the river*
  - Exceptions / Messy





7

1.Introduction (IV)

Recent significant advances in NLP:

- **Web-scale information retrieval**
- **Question answering:** IBM's Watson
- **Interactive Dialogue Systems** (Siri / Cortana / Google Now)
- **Spam classification**

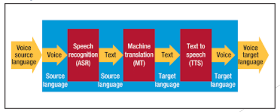
8

1.Introduction (V)

Recent significant advances in NLP:

- Automatic Machine Translation

- Speech-to-Text; Text-to-Speech; Translation



9

1.Introduction (V)

Recent significant advances in NLP:

- Large Language Models (LLMs): ChatGPT

10

1.Introduction (VI)

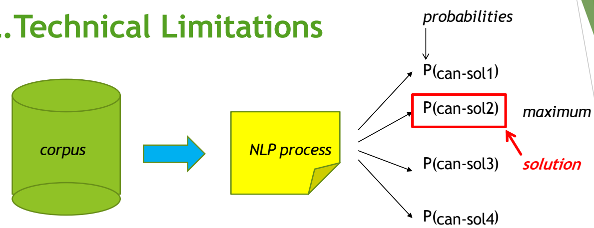
- Welcome to the realm of

probability

- In NLP, anwers are selected by their probability
- Many odd anwsers are candidates with very low probability; but not zero, perhaps
  - it is an exception
  - a non-frequent word (language is open)
  - a new word (language is dynamic)

11

2.Technical Limitations



- To technically advance, elementary notions of:
  - Markov processes: chain of states, the current state only depends on a fixed number of previous states
  - Probability theory (conditioned probabilities, Bayes' theorem)
- **We will develop a qualitative approach, with acts of faith**

12

2. Basics on probability

- probability(sentence) = product of probabilities of its words?
- Independent vs dependent events (13.4)
  - A: You are at the UAB B: you dice outputs 6
  - A: You are at the UAB B: the AI degree is your course
- And words in a sentence are not independent:
  - Language is not a bag of words
- Consider **a sentence as a Markov chain:**
- probability(sentence) = product of conditioned probabilities of its words

13

13

2.History: Andrei Markov

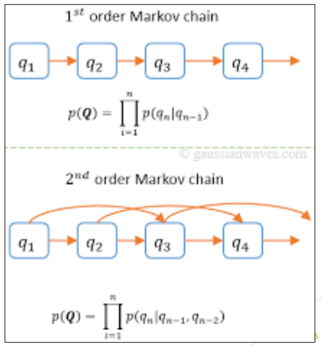
- Russian mathematician
- 1856-1922
- Number theory
- Probability theory

14

14

2.Markov processes

- 15.1.2
- Markov assumption: the current state depends on only a finite fixed number of previous states.
- Markov chain:
  - First order: depends on the previous state
  - Second order: depends on the two previous states



15

2 Example

- Corpus: *Dábale arroz a la zorra el abad*
- Characters: sentence of 25+6 characters, formed by 9 different ch.

	frequency
d	2
a	7
b	2
l	3
e	2
r	4
o	2
z	2
space	6

16

16

2 Example: tri-gram model

- Corpus: *Dábale arroz a la zorra el abad*
- Compute the probability of
  - each character in isolation: pr(x)
  - conditioned probability of each character to any previous two: tri-gram model

17

17

2 Example

- Corpus: *Dábale arroz a la zorra el abad*

<table>
<tr><td>d|dd</td><td>d|ad</td><td>d|bd</td><td>d|ld</td><td>d|ed</td><td>d|rd</td><td>d|od</td><td>d|zd</td><td>d|_d</td></tr>
<tr><td>d|da</td><td>d|aa</td><td>d|ba</td><td>d|la</td><td>d|ea</td><td>d|ra</td><td>d|oa</td><td>d|za</td><td>d|_a</td></tr>
<tr><td>d|db</td><td>d|ab</td><td>d|bb</td><td>d|lb</td><td>d|eb</td><td>d|rb</td><td>d|ob</td><td>d|zb</td><td>d|_b</td></tr>
<tr><td>d|dl</td><td>d|al</td><td>d|bl</td><td>d|ll</td><td>d|el</td><td>d|rl</td><td>d|ol</td><td>d|zl</td><td>d|_l</td></tr>
<tr><td>d|de</td><td>d|ae</td><td>d|be</td><td>d|le</td><td>d|ee</td><td>d|re</td><td>d|oe</td><td>d|ze</td><td>d|_e</td></tr>
<tr><td>d|dr</td><td>d|ar</td><td>d|br</td><td>d|lr</td><td>d|er</td><td>d|rr</td><td>d|or</td><td>d|zr</td><td>d|_r</td></tr>
<tr><td>d|do</td><td>d|ao</td><td>d|bo</td><td>d|lo</td><td>d|eo</td><td>d|ro</td><td>d|oo</td><td>d|zo</td><td>d|_o</td></tr>
<tr><td>d|dz</td><td>d|az</td><td>d|bz</td><td>d|lz</td><td>d|ez</td><td>d|rz</td><td>d|oz</td><td>d|zz</td><td>d|_z</td></tr>
<tr><td>d|d_</td><td>d|a_</td><td>d|b_</td><td>d|l_</td><td>d|e_</td><td>d|r_</td><td>d|o_</td><td>d|z_</td><td>d|__</td></tr>
</table>

18

18

2 Smoothing n-gram models

- Issues:
  - Characters or words unseen, not present in the corpus but existing in the language:
    - The model will assign pr 0 → bad, the whole sentence will have pr 0
    - pr small but greater that 0
    - Different ways to assing this small pr
    - Should follow the probability axioms
  - At the beginning of a sentence

19

19

2 Probabilities of Independent Events

- Rely on your previous knowledge about probability
- If A and B independent, pr(A and B) = pr(A,B) = pr(A) pr(B)
- Example: dice games
  - A: 1st dice=6
  - B: 2nd dice=6
  - Two sixs, one after the other,

    pr(A,B) = pr(A) pr(B) = 1/36

20

20

2 Probabilities of Dependent Events

- If A and B dependent, pr(A and B)= pr(B|A) pr(A)
- pr(B|A): conditioned probability
  - How probable is B, in the cases where A holds (A and B are dependent)?
- Dependency:
  - Causal:

    A: making love   B:get pregnant
  - Co-occurence

    A: meningitis   B: stiff neck

NLP: The characters or the words of a sentence are dependent of the previous characters or words.

21

21

2. History: Thomas Bayes

- 1702 – 1761
- Conditional probability
- Famous theorem

22

22

2.Conditional Probability: Bayes' theorem

pr(A and B) = pr(A,B) = pr(A|B) pr(B)

pr(A and B) = pr(A,B) = pr(B|A) pr(A)

pr(A|B) = pr(B|A) pr(A) / pr(B) — Bayes's theorem

23

23

2. Example on Bayes' theorem

- 13.5.1
- pr(cause|effect) = pr(effect|cause) pr(cause)/pr(effect)
- pr(cause|effect): diagnosis direction
- pr(effect|cause): causal direction
- Example: *meningitis and stiff neck*

pr(s|m) = 0,7

pr(m) = 1/50.000

pr(s)=0,01

pr(m|s) = pr(s|m) pr(m)/pr(s) = (0,7*1/50.000)/0,01 = 0,0014

24

24

3.Language Models

- Formal languages (e.g. Python, Logic)
  - Grammar (generative)
  - Semantics
- But natural languages (e.g. English, German…)
  - Grammaticality are less clear

    Not to be invited is sad / To be not invited is sad
  - -Ambiguity at many levels (syntax, semantics, …)

    He saw her duck

25

3.Note on N-gram models

- Often, computing Bayes' rule we are interested in argmax_{Ai}

  pr(B|Ai) pr(Ai)/pr(B)
- Since the denominator is constant (equal for all Ai), we ignore it

  pr(B|Ai) pr(Ai)

26

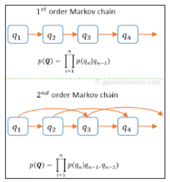
3.N-gram character models

- P(c1:n) is the probability of a sequence of N *characters* c1 through cN
  - Typically corpus-based (uses a body of text)
  - P("the") = .03
  - P("zgq") = .000000000002
- Application: language identification
  - Corpus: P(Text|Language) (trigrams)
  - Language Identification -use BayesRule!

27

4.3-gram character Model

- Probability that a sentence is from a language? `my tailor is rich`
  1. Multiply the probability of its characters → they are not independent, multiply the conditioned probabilities (Bayes' rule)
     [in fact, the question is a conditioned probability $P(l|c_{1:N})$]
  2. Assume the probability of a character depends on the previous two:
     $P(c_i|c_{i-1},c_{i-2})$ [Markov chain]



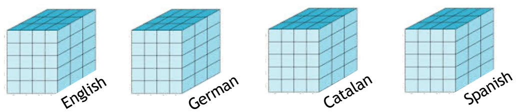
  3. Compute them in a language corpus (vocabulary 100 char, → array of 1.000.000 entries, corpus of 10.000.000 char)

28

4.3-gram character Model

- Probability of a sentence in a language? `my tailor is rich`
  4. L = argmax$_l$ P(l|c$_{1:N}$) = argmax$_l$ P(c$_{1:N}$|l) P(l) [*by Bayes' theorem*]
     [observe that you don't know P(l|c$_{1:N}$), but you know P(c$_{1:N}$|l)]

  You try this for different languages (you have computed arrays for English, German, Catalan, Spanish…)



5. The language that you want is the one with highest probability
6. Useful to automatically identify languages in translators

29

5.N-gram *word* models

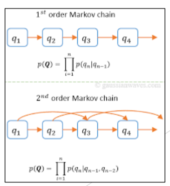
- Much larger "vocabulary" of units
- Since units are open, out of vocabulary becomes a problem
- "Word" needs to be defined precisely

30

30

5.3-gram word Model

- Words instead of characters: P(class|mess) (spam detector)
- Vocabulary:
  - Long list of words
  - Always consider the unknown word
- Multiply the probability of its words → they are not independent, multiply the conditioned probabilities
  Assume the probability of a word depends on the previous two:
  P(w$_i$|w$_{i-1}$,w$_{i-2}$) [Markov chain]



- Corpus: several thousand sentences of two classes: spam and ham (= ¬spam)

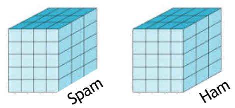
31

31

5.3-gram word Model

- Probability of a message in a class? `Spam detector`
  1. C = argmax$_{class}$ P(class|mess) = argmax$_{class}$ P(mess|class)P(class)
     [observe that you don't know P(class|mess), but you know P(mess|class)]

  P(mess|class): you will work with the arrays for the clases *spam* and *ham*

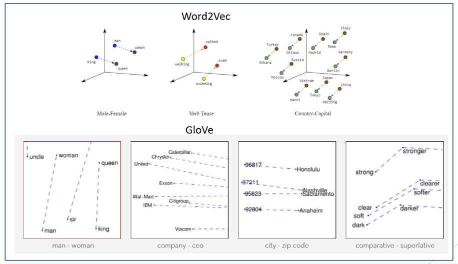


2. Then, you classify the incoming *mess* in the class with higher probability

32

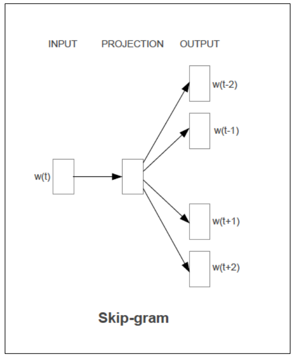
6.Word Embeddings

- Transform a word into a vector of many dimensions
- Somehow capture semantic meaning, since similar words are mapped into close regions of that multidimensional space.



33

6.Word Embeddings: How?



- Neural network
- Input: a word W
- Output: words forming the context of W
- The output vector → embedding
- Training:
  - Thousands of words
  - Millions of sentences

34

6.Word Embeddings (II)

- Very useful in machine translation:
  - Operate words:
    - king - man + woman = queen
  - “A word is characterized by the company it keeps”

35

6.Word Embeddings (III)

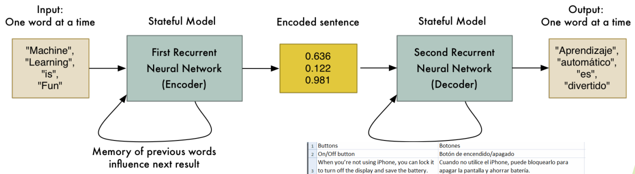
A	B	C	D=C+(B-A)	Relatrionship
Athens	Greece	Oslo	Norway	Capital
Astana	Kazahastan	Harare	Zimbabwe	Capital
Angola	kwanza	Iran	Rial	Currency
Copper	Cu	Gold	Au	At. Symbol
Microsoft	Windows	Google	Android	Op. System
New York	New York Times	Baltimore	Baltimore Sun	Newspaper
Berlusconi	Silvio	Obama	Barak	First name
Switzerland	Swiss	Cambodia	Cambodian	Nationality
Einstein	Scientist	Picasso	painter	Occupation

A word embedding model can sometimes answer the question “A is to B as C is to [what]?” with vector arithmetic: given the word embedding vectors for the words A, B, and C, compute the vector D=C+(B−A) and look up the word that is closest to D. (The answers in column D were computed automatically by the model. The descriptions in the “Relationship” column were added by hand.) Adapted from Mikolov et al. (2013, 2014).

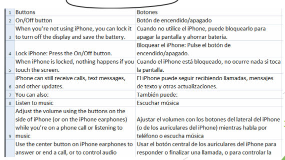
36

7.Automatic Translation

- State-of-the-art: RNN (recurrent neural networks, deep learning)
- Sentence embedding + attention mechanism



- Trained with a long parallel corpus



1	Buttons	Botones
2	On/Off button	Botón de encendido/apagado
3	When you're not using iPhone, you can lock it to turn off the display and save the battery.	Cuando no utilice el iPhone, puede bloquearlo para apagar la pantalla y ahorrar batería.
4	Lock iPhone: Press the On/Off button.	Bloquear el iPhone: Pulse el botón de encendido/apagado.
5	When iPhone is locked, nothing happens if you touch the screen.	Cuando el iPhone está bloqueado, no ocurre nada si toca la pantalla.
6	iPhone can still receive calls, text messages, and other updates.	El iPhone puede seguir recibiendo llamadas, mensajes de texto y otras actualizaciones.
7	You can also:	También puede:
8	Listen to music	Escuchar música
9	Adjust the volume using the buttons on the side of iPhone (or on the iPhone earphones) while you're on a phone call or listening to music	Ajustar el volumen con los botones del lateral del iPhone (o de los auriculares del iPhone) mientras habla por teléfono o escucha música
	Use the center button on iPhone earphones to answer or end a call, or to control audio	Usar el botón central de los auriculares del iPhone para responder o finalizar una llamada, o para controlar la

37

37

7.Automatic Translation (II)

- Applications are available in Internet:
  - Google Translate: 109 languages
  - DeepL: 24 laguages
  - Limits for the free versions
  - Revolution in the manual translation field: now, in many cases, the starting point is a text automatically pretranslated.

38

38

8.LLMs: ChatGPT

- ChatGPT:
  - Huge deep neural network
    - 175 x 10^9 parameters (weighths)
    - Trained with millions of texts
- Compute a "reasonable continuation": given the text so far, what should the next word be?
- It works in a space of "meanings" ← *radically new !!*

39

39

8. ChatGPT: Tokens

From words to tokens

40

40

8. ChatGPT: text predictor

- Predict the next word (token)
- Associate each candidate with a probability
- Selecting the most probable does not generate good text
- Random selection among the most probable
- Assessed by a parameter called Temperature

41

41

Wrap-up

1. Introduction
2. Markov chains
3. Conditioned probabilities
4. 3-gram character model (language identification)
5. 3-gram word model (spam classification)
6. Word embeddings
7. Machine translation
8. LLMs: ChatGPT

42

42

Further reading

- Russel & Norvig 3rd ed:
  22.1.1, 22.2, 22.3, 22.4,
  Bibliographical Notes chapter 22

- For the word embeddings & machine translation: Russel & Norvig 4rd ed
  25.1, 25.2, 25.3

43

43