Tecnologie del Linguaggio Naturale 20/21

Parte Prima

Esercizi d'esame

per la parte prima

26-03-2021

Scegliere un esercizio tra 1 o 2

1. PoS Tagger per lingue morte

2. CKY per lingue immaginarie

3. Bonus Track

1 PoS Tagger

Costruire un PoS tagger statistico basato su HMM per il Greco antico e il Latino

A. Implementare Learning (contare) e Decoding (Viterbi)
B. Addestrare il sistema su Greco e Latino (separatamente) usando 2 Treebank del progetto UD
C. Valutare il sistema, usando diverse strategie di smoothing
D. Valutare rispetto ad una baseline facile e ad una difficile

1.A algoritmo per il learning

- Elenchi di parole e di PoS TAG
- Probabilità PoS->PoS: **$P(t_i|t_{i-1})$** $$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$
- Probabilità PoS->Word: **$P(w_i|t_i)$** $$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$
- Viterbi
- **HINT**: usare i logaritmi per le probabilità!

1.A algoritmo per il learning

- Elenchi di parole e di PoS TAG

```
# sent_id = test-s1
# text = + In Dei omnipotentis nomine, regnante domno nostro Karolus divina faventem clementia imperatore augu
quinta.
# reference = document_id='85:47'-span='146'
1       +       +       PUNCT   SYM     _       5       punct   _       _
2       In      in      ADP     r|r|-|-|-|-|-|-|-       _       5       case    _       _
3       Dei     Deus    PROPN   Propn|n|-|s|-|-|-|m|g|- Case=Gen|Gender=Masc|Number=Sing        5       nmod
4       omnipotentis    omnipotens      ADJ     a|a|-|s|-|-|-|m|g|-     Case=Gen|Gender=Masc|Number=Sing
5       nomine  nomen   NOUN    n|n|-|s|-|-|-|n|b|-     Case=Abl|Gender=Neut|Number=Sing        7       obl
...       ...             ...
```

```
# newdoc id = tlg0003.tlg001.perseus-grc1.1.tb.xml
# sent_id = tlg0003.tlg001.perseus-grc1.1.tb.xml@3
# text = τὰ γὰρ πρὸ αὐτῶν καὶ τὰ ἔτι παλαίτερα σαφῶς μὲν εὑρεῖν διὰ χρόνου πλῆθος ἀδύνατα ἦν, ἐκ δὲ τεκμηρίων ὧν ἐπὶ
νομίζω γενέσθαι οὔτε κατὰ τοὺς πολέμους οὔτε ἐς τὰ ἄλλα.
1       τὰ      ὁ       DET     l-p---na-       Case=Acc|Gender=Neut|Number=Plur        11      det     _       _
2       γὰρ     γάρ     ADV     d--------       _       15      advmod  _       _
3       πρὸ     πρό     ADP     r--------       _       4       case    _       _
4       αὐτῶν   αὐτός   PRON    p-p---ng-       Case=Gen|Gender=Neut|Number=Plur        1       nmod    _       _
5       καὶ     καί     CCONJ   c--------       _       1       cc      _       _
6       τὰ      ὁ       DET     l-p---na-       Case=Acc|Gender=Neut|Number=Plur        8       det     _       _
7       ἔτι     ἔτι     ADV     d--------       _       8       advmod  _       _
...       ...             ...
```

1.B data: training-dev-test

- Greco antico -> Perseus
  - INFO: https://universaldependencies.org/treebanks/grc_perseus/index.html
  - DATA: https://github.com/UniversalDependencies/UD_Ancient_Greek-Perseus
- Latino -> LLCT
  - INFO: https://universaldependencies.org/treebanks/la_llct/index.html
  - DATA: https://github.com/UniversalDependencies/UD_Latin-LLCT

1.C smoothing

Ipotesi di smoothing per le parole sconosciute:

- Sempre nomi P(unk|NOUN) = 1
- Sempre nomi/verbi P(unk|NOUN)=P(unk|VERB)=0.5
- $P(unk|t_i) = 1/\#(PoS\_TAGs)$
- Statistica PoS sul development set: parole che compaiono 1 sola volta
- Approcci Syntax-based? Suffissi? (opzionale)

1.D Valutare

Calcolare la precisione sul test set.

Implementare 2 baselines:

- Facile: assegnare il tag più frequente se c'è nel training, altrimenti nome.
- Difficile (opzionale): MEMM https://github.com/Michael-Tu/ML-DL-NLP/tree/master/MEMM-POS-Tagger

Quali sono gli errori più comuni?

2. CKY

- Implementare CKY
- Provare l'implementazione su una grammatica scritta a mano per la lingua Dothraki
  - https://wiki.dothraki.org
  - https://docs.dothraki.org/Dothraki.pdf
  - *Hollywood language* ->The Art of Language Invention, David J. Peterson

2.A-B CKY

A. Implementare CKY from scratch

   - Struttura dati
   - Algoritmo-> triplo ciclo

B. Provare l'algoritmo sulla grammatica L1 di Jurafsky (Lez04-Syntax02: slides 39-40), sulle frasi:

   - *Book the flight through Houston*
   - *Does she prefer a morning flight*

2.B CKY: L1

Grammar	Lexicon
$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

2.B CKY: L1 in normal-form

$S \rightarrow NP\ VP$	$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$	$S \rightarrow X1\ VP$
	$X1 \rightarrow Aux\ NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb\ NP$
	$S \rightarrow X2\ PP$
	$S \rightarrow Verb\ PP$
	$S \rightarrow VP\ PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det\ Nominal$	$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal\ Noun$	$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$	$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb\ NP$	$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$	$VP \rightarrow X2\ PP$
	$X2 \rightarrow Verb\ NP$
$VP \rightarrow Verb\ PP$	$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$	$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$	$PP \rightarrow Preposition\ NP$

2.C CKY

C. Costruire una CF (senza semantica) per la lingua Dothraki, e parsificare le seguenti 3 frasi usando l'algoritmo implementato:

- **Hash yer astoe ki Dothraki?** (Do you speak Dothraki?)

  frase interrogativa

- **Anha zhilak yera** (I love you)

  frase dichiarativa

- Anha gavork (I'm hungry)

  frase copulative

- Fare riferimento alla sintassi del Dothraki: https://wiki.dothraki.org/Syntax
- HINT: scivere la grammatica direttamente in Chomsky normal form

3. Bonus Track

Costruire una CF (**con semantica!**) per la lingua Dothraki ispirandosi alla grammatica simple-sem.fcfg, e parsificare le seguenti 3 frasi usando l'interprete seman tico di NLTK visto a lezione:

- **Hash yer astoe ki Dothraki?** (Do you speak Dothraki?)

  frase interrogativa

- **Anha zhilak yera** (I love you)

  frase dichiarativa

- Anha gavork (I'm hungry)

  frase copulative

Consegna

Bisogna consegnare il codice e una breve relazione (5-10 pagine) almeno due giorni prima della data dell'esame dell'orale concordata.

Attenzione: gli esercizi si possono fare in gruppi formati da un massimo di 2 persone