Tecnologie del Linguaggio Naturale

2021/2022

Esercizi d'esame

per la parte prima

01-04-2022

Scegliere un esercizio tra 1 o 2

1. La magia NER nascosta

2. DS per apprendisti Streghe o Stregoni

1 La magia NER nascosta

Implementare un NER con HMM:

A. Implementare Learning (contare) e Decoding (Viterbi)

B. Addestrare il sistema su Wikipedia ENG e ITA

   - https://github.com/Babelscape/wikineural/tree/master/data/wikineural/en
   - https://github.com/Babelscape/wikineural/tree/master/data/wikineural/it

C. Valutare il sistema, usando diverse strategie di smoothing per ENG e ITA

D. Valutare rispetto ad una baseline facile e ad una difficile

1.A algoritmo per il learning

- Elenchi di parole e di NER TAG
- Probabilità TAG->TAG: **$P(t_i|t_{i-1})$** $P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$
- Probabilità TAG->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 NER TAG:

- PER persona
- ORG organization
- LOC location
- MISC miscellanea

0	La	O
1	pellicola	O
2	è	O
3	stata	O
4	presentata	O
5	in	O
6	concorso	O
7	alla	O
8	61ª	B-MISC
9	Mostra	I-MISC
10	internazionale	I-MISC
11	d’	I-MISC
12	arte	I-MISC
13	cinematografica	I-MISC
14	di	I-MISC
15	Venezia	I-MISC
16	.	O

1.C smoothing

Ipotesi di smoothing per le parole sconosciute:

- Sempre O: P(unk|O) = 1
- Sempre O o MISC: P(unk|O)=P(unk|B-MISC)=0.5
- Uniforme: $P(unk|t_i) = 1/\#(NER\_TAGs)$
- Statistica TAG sul development set: parole che compaiono 1 sola volta
- Altro? (opzionale)

1.D Valutare

Calcolare sul test set:

1. Accuracy generale (come per il PoS Tagging)
2. Precision e recall sulle entità

Implementare 2 baselines:

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

Quali sono gli errori più comuni?

1.D Valutare

Provare il vostro sistema sulle frasi:

- *La vera casa di Harry Potter è il castello di Hogwards.*
- *Harry le raccontò del loro incontro a Diagon Alley.*
- *Mr Dursley era direttore di una ditta di nome Grunnings, che fabbricava trapani.*

2 DS per apprendisti Streghe o Stregoni

Specifiche:

1. Il DS (ITA o ENG) deve impersonare il personaggio Severus Piton. Il DS è **task-based**: deve interrogare l'utente sulla composizione di 3 pozioni magiche a scelta tra:
   https://www.potterpedia.it/?speciale=elenco&categoria=Pozioni
2. Algoritmo: ANALISI-DM-GENERAZIONE

2.A: ANALISI

2 possibili approcci:

- Come Eliza: espressioni regolari su stringhe
- Con le dipendenze: si usi un parser a dipendenza (es. Spacy, Stanza, Tint) si cerchino le regolarità nell'albero. Es: *Un ingrediente è l'acqua di luna.*
  - *ingrediente –nsubj-> acqua*
  - *è –cop-> acqua*
  - *acqua –nmod-> luna*

2.B: DM

- **L'iniziativa è del sistema** che deve interrogare
- **Frame-Based:** ogni pozione viene rappresentata come un frame da riempire i cui slot sono gli ingredienti -> **Common-ground**
- Il DM deve interrogare sugli ingredienti ancora mancanti, eventualmente proponendo risposte vere o false
- Dovrebbe dare un voto e un commento (sagace) alla fine dell'interazione
- Backup-strategy
- Memory

2.C: Generazione

- Definire una struttura per il Text-Plan e una per il Sentence-Plan
- Usare Simple-NLG o SimpleNLG-it
  - https://github.com/simplenlg/simplenlg
  - https://pypi.org/project/simplenlg/
  - https://github.com/alexmazzei/SimpleNLG-IT

2.D Valutare

- Analizzare almeno 3 dialoghi (-> relazione)
- Quali sono gli errori più comuni?
- Quali fenomeni linguistici si riescono a gestire?

  Trindi Tick List

2.E: Bonus Tracks

1. SpeechRecognition e Text2Speech: cosa cambia negli errori?
2. Approccio alternativo all'analisi basato su logica: costruire un CFG con semantica con la libreria NLTK

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 2/3 persone