Tecnologie del Linguaggio Naturale

Parte Prima

A.A. 2023/2024

> Esercizi d'esame

15 Marzo, 2024

Scegliere un esercizio tra 1, 2 o 3

1. NER Multilingua

2. Dependency parser proiettivo per l'italiano

3. DS: Prof. Danny

1 NER Multilingua

Implementare un NER con HMM:

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

B. Addestrare il sistema su Wikipedia EN, ES, IT

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

C. Valutare il sistema, usando diverse strategie di smoothing per le tre lingue

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.B: 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.C: Valutare

Calcolare sul test set:

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

ATTENZIONE: conviene rappresentare le entità, sia nel gold che nel system, come insiemi di quadruple (PER, Sent-24,3,4) (ORG, Sent-12,9,10), etc.

Quali sono gli errori più comuni?

1.C: Valutare

Implementare 2 baselines:

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

2 Dependency parser proiettivo per l'italiano

- Realizzare un transition-based parser statistico proiettivo (stile “MALT”)
- Un tutorial: https://spacy.io/blog/parsing-english-in-python
- Usare una libreria di ML (anche neurale) per costruire un classificatore che decide se fare **LEFT-RIGHT-SHIFT** sullo stato attuale. Oppure implementare un algoritmo di classificazione statistica
- **ATTENZIONE**
  - RELAZIONI H-D NON TIPATE
  - NON BISOGNA IMPLEMENTARE IL POS TAGGER

2.A: Corpus

- https://github.com/UniversalDependencies/UD_Italian-ISDT/tree/master
  - it_isdt-ud-train.conllu
  - it_isdt-ud-dev.conllu
  - it_isdt-ud-test.conllu

2.B: Preprocessing

1	Inconsueto	inconsueto	ADJ	A	Gender=Masc|Number=Sing	2	amod	_	_
2	allarme	allarme	NOUN	S	Gender=Masc|Number=Sing	0	root	_	_
3-4	alla	_	_	_	_	_	_	_	_
3	a	a	ADP	E	_	5	case	_	_
4	la	il	DET	RD	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	5	det	_	_
5	Tate	Tate	PROPN	SP	_	2	nmod	_	_
6	Gallery	Gallery	PROPN	SP	_	5	name	_	_
7	:	:	PUNCT	FC	_	2	punct	_	_

1	Inconsueto	inconsueto	ADJ	A	Gender=Masc|Number=Sing	2	amod	_	_
2	allarme	allarme	NOUN	S	Gender=Masc|Number=Sing	0	root	_	_
3	a	a	ADP	E	_	5	case	_	_
4	la	il	DET	RD	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	5	det	_	_
5	Tate	Tate	PROPN	SP	_	2	nmod	_	_
6	Gallery	Gallery	PROPN	SP	_	5	name	_	_
7	:	:	PUNCT	FC	_	2	punct	_	_

2.C: Feature Modelling

- Le feature considerano le parole e i PoS delle parole sullo stack e sulla lista
- Usare i feature template standard riportati nel libro di testo nel Capitolo 18

2.D: Valutazione del Parser

- Valutare i risultati usando il programma eval07.pl:

  eval07.pl -q -g it_isdt-ud-test.conllu -s out.conllu

- Considerare solo la misura **unlabelled attachment score**

2.E: Bonus Tracks

- Pseudo-projective parsing [Nivre and Nilsson 2005]
- Preprocess training data, post-process parser output
- http://www.maltparser.org/optiondesc.html

-> proj+deproj

3. Prof. Danny

Il DS (ITA o ENG) deve impersonare il personaggio di Danny, un prof. di Linguistica Computazionale.

Il DS è **task-based**: deve interrogare l'utente sulla conoscenza degli argomenti trattati nella parte 1 di TLN2324.

3. Prof. Danny

Usare la base di conoscenza: **DomandeTLN2324.txt** (almeno 8 domande).

Esempi:

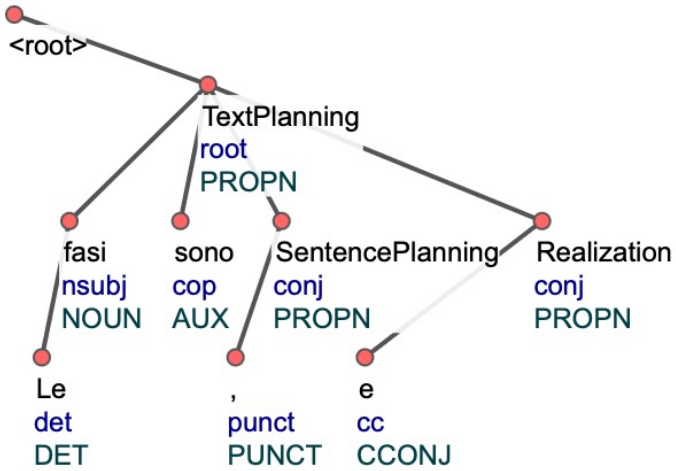
- HMM è un modello generativo o discriminativo?
  - *Generativo*
- Quali sono le fasi della NLG simbolica?
  - *Le fasi sono TextPlanning , SentencePlanning e Realization*
- Cosa è una Probabilistic CFG?
  - *È una grammatica CFG in cui ogni regola di produzione ha una probabilità associata*

3.A ANALISI

Algoritmo: ANALISI-DM-GENERAZIONE

Due possibili approcci:

- Come Eliza: espressioni regolari per controllare le risposte
- Con le dipendenze: si usi un parser a dipendenza (es. udpipe, Spacy, Stanza, Tint) si cerchino le regolarità nell'albero.
  - ***fasi –nsubj-> TextPlanning***
  - ***TextPlanning –conj-> SentencePlanning***
  - ***TextPlanning -conj-> Realization***



3.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 le risposte corrette -> **Common-ground**
- Il DM segue una GUS-Policy: prova a riempire il frame. Deve interrogare anche proponendo, eventualmente, risposte vere o false
- Dovrebbe dare un giudizio, un commento (sagace?) e decidere se l’utente l'esame alla fine dell’interazione
- Backup-strategy
- Memory?

3.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

3.D Valutare

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

3.E Bonus Tracks

- SpeechRecognition e Text2Speech: cosa cambia negli errori?
- 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 1, 2 o 3 persone (dello stesso A.A.)

TLN - Fine Parte 1