Ilaria Salogni
Matricola: 640091

Relazione

Linguistica Computazionale II

Anno Accademico 2021/2022

Questo progetto *hands-on* consiste nella revisione dell'annotazione semi-automatica di un corpus testuale, effettuato con la pipeline addestrabile UDPipe, utilizzando un modello data-driven, con la finalità di creare dopo la correzione manuale di due annotatori umani, un golden corpus.

I testi sono stati appositamente scelti in generi e epoche differenti da quello del corpus di addestramento del modello, per osservare le difficoltà create dalla domain-adaptation, e allo stesso scopo il golden corpus è stato usato per valutare altri due modelli di annotazione automatica dell'italiano.

Durante il lavoro è risultato più comodo riferirsi ai testi con un nome univoco, che riportiamo qui affiancato a quello con cui ci sono stati assegnati:

Cucina-1	cucina-sample_lazzariturco_1947.txt
Cucina-2	cucina-sample_centosessantamaniere_1907.txt
Cucina-3	cucina-sample_boni_1927.txt
Giornali-1	giornali-sample_lastampa_1977.txt
Giornali-2	giornali-sample_corriere_1996.txt

Sentence Splitting

1. **Due punti, punto e virgola**
   Generalmente abbiamo osservato l'inconsistenza del *parser* nella segmentazione in presenza dei due punti e del punto e virgola, anche in contesti apparentemente molto simili. Il nostro approccio come annotatori è stato il più segmentante possibile: tranne in tre casi che costituiscono eccezione, sia i due punti che il punto e virgola sono stati posti come limite della frase.
   La prima eccezione è quella in cui abbiamo trovato **due punti seguiti da elenchi**[1] , ritenendo che sarebbe stato complesso analizzare l'elenco isolato dal resto della frase.
   Per coerenza con questa linea guida abbiamo poi deciso anche di non isolare gli elenchi che contenessero un verbo principale, come i casi nell'esempio[2]. Non abbiamo separato la frase anche quando dopo i **due punti** fossero **seguiti da una congiunzione coordinante**[3], e quando i due punti si trovassero **all'interno di un discorso diretto** racchiuso fra virgolette, costruzione che abbiamo sempre voluto considerare come unica frase.
2. **Abbreviazioni**
   La performance dell'annotatore automatico in presenza di forme abbreviate è stata piuttosto deludente. Questa situazione è particolarmente evidente nel testo Cucina-1, dove si può vedere come le forme di abbreviazione più diffuse ("gr. "e "N. 1") sono state correttamente analizzate, mentre la punteggiatura nelle forme come "chilog.", "chilogr." o l'elenco numerato ("5. Polenta…") è stata sistematicamente confusa per punto fermo.

[1] Esempio da Cucina-3: *Secondo (…) variano le proporzioni dei componenti: burro, farina e latte.*
[2] Esempio da Giornale-2: *Abbiamo saputo che Silvio Berlusconi, partito da zero, ha fatto tutto da solo: Fininvest, Standa, Mondadori, assicurazioni, mentre Prodi ha risanato l'iri, che però c'era già.*
[2] Esempio da Giornale-2: *Il solo brivido è stato l'elenco dei soprannomi dei contendenti: ricordo Berluskaz, Nano pelato, (…).*
[3] Esempio da Giornale-2: *Non c'è nessuno che possa mettere qualche brivido a questa compagnia elettorale: **e** Lucia Annunziata, (…).*

Tokenizzazione

L'utilità di un approccio incrementale è stata particolarmente evidente in questa fase: senza la corretta partizione di clitico e verbo si sarebbero avuti risultati disastrosi nel *POS tagging*.

1. **Clitici**
   In tutti i testi il riconoscimento delle **particelle pronominali enclitiche** è stato problematico. I testi di cucina in particolare si prestano all'approfondimento della questione per densità di occorrenza.

	Cucina-1	Cucina-2	Cucina-3
Clitici totali:	21	35	20
Riconosciuti:	14	15	4

Tabella 1

   L'annotatore ha dato esiti diversi anche in contesti simili (si veda *immagine 1*), e non si può dire che abbia avuto una performance migliore su verbi d'uso più comune.

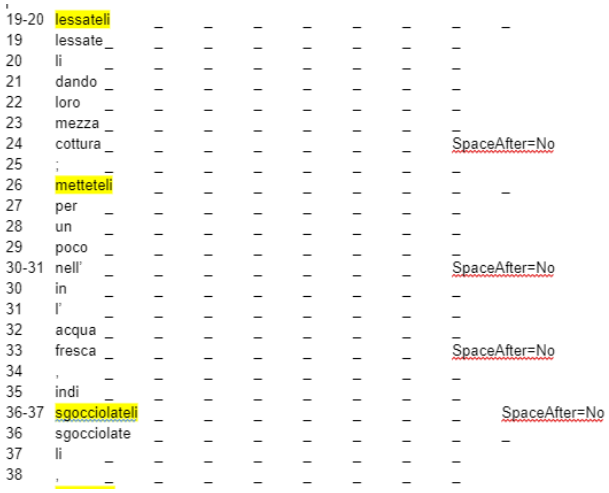

Immagine 1

   Inoltre c'è stato un risultato diverso anche su flessioni dello stesso verbo: la forma "fateli" è stata correttamente separata, ma non la forma "fatelo" oppure "fatela".
2. Le **preposizioni articolate** a differenza dei primi sono sempre state correttamente trattate, tranne in due casi che attribuibili al fattore diacronico, cioè "colla" e "col".[4]

[4] "*Accuracies also drop markedly when there are differences in topic, epoch, or writing style between the training and operational data*". Manning, C. D. (2011, February). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *International conference on intelligent text processing and computational linguistics* (pp. 171-189). Springer, Berlin, Heidelberg.

3. L'univerbazione "Rai-tv" è stata trattata senza consistenza dal modello, e abbiamo convenuto per la separazione in tre token, ritenendo fosse l'atteggiamento più riproducibile. Abbiamo fatto lo stesso nel caso di numeri che indicano un intervallo come "1-1/2".

Pos Tagging

1. Gli **imperativi** sono stati riconosciuti con molta difficoltà, dato che la disambiguazione è affidata solo al contesto: solo per fare un esempio, nella frase "passate da casa" il verbo può essere un participio, un presente indicativo o un imperativo.

   Questo si vede molto bene nell'*immagine 2*, in cui l'arcaismo "poscia", scorrettamente interpretato come sostantivo, ha fatto sì che "bagnate" fosse etichettato come articolo indeterminativo.

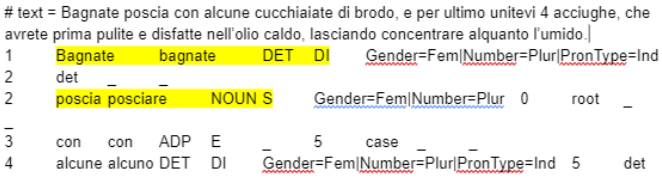

immagine 2

"*While taggers are quite good, they regularly make egregious errors*"[5]: questo aspetto è particolarmente evidente qui, in cui la maggior parte degli imperativi comunque sono correttamente riconosciuti come verbi, ma questo non significa che manchino dei casi tremendamente sbagliati, come mostrano appunto *immagine 2* e *immagine 3.*

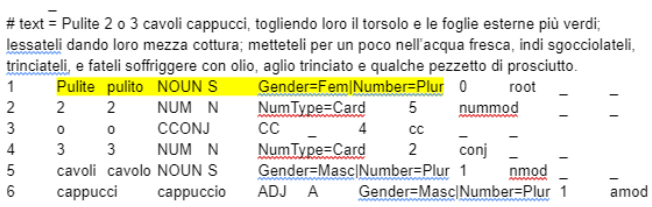

immagine 3

in ogni caso la situazione più frequente è stata quella *di immagine 4,* in cui è stata riconosciuta la categoria sintattica, ma non il corretto tempo verbale.

[5] Manning, C. D. (2011, February). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *International conference on intelligent text processing and computational linguistics* (pp. 171-189). Springer, Berlin, Heidelberg.

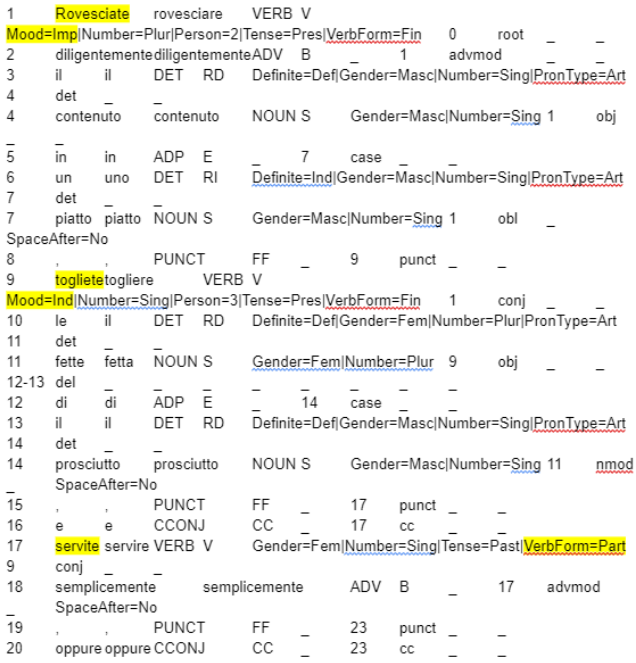

immagine 4

2. Similarmente all'imperativo, anche il riconoscimento del ruolo delle forme **participiali** nella frase in funzione aggettivale o verbale, è affidata al contesto. Nonostante ciò, mi sarei aspettata una performance migliore in quanto il participo passato ha funzione verbale quando è unito all'ausiliare del verbo essere o avere, e quindi ha un contesto caratterizzato.

   È stato invece più complesso analizzare il participio che introduce la subordinata, ad esempio in "pronta che sia".

3. Anche quando i clitici sono manualmente separati in fase di tokenizzazione, possono esser scambiati per avverbi o per articoli determinativi.
   Questo errore si riscontra anche per i **pronomi diretti** con ruolo di oggetto, come in "la verserete".
4. Errori sistematici da parte del sistema nel riconoscimento di **nomi posposti** rispetto all'aggettivo, incontrata specialmente in ambito giornalistico.

Lemmatizzazione

Gli arcaismi e le parole rare sono stati i più complessi per il task di lemmatizzazione, come potevamo già aspettarci.

Tuttavia è interessante osservare come lemmi non presenti nel dizionario del modello ma ricorrenti in un dato testo, siano stati ricondotti spesso a **radici differenti**: è il caso di "polenta", per la quale gli esiti della lemmatizzazione sono stati "polento", "polentire", "polentare".

Sporadicamente, alcuni tempi meno comuni dei verbi come il futuro “getterete” sono stati scorrettamente lemmatizzati (in “gettereto”), benché tutto il resto dell’analisi fosse corretta.

Parsing

In questa sezione abbiamo sperimentato la difficoltà di essere coerenti con la propria annotazione: non è stato facile decidere se un *giardino curato* fosse da annotare allo stesso modo del *brodo ottenuto.*

1. Si è notato che nel caso di lunghe coordinazioni fra elementi connessi da una congiunzione coordinante, indicate con CONJ, l’accumulazione riparte spesso da un nuovo antecedente, generando una catena, benché le linee guida richiedano di ricollegare al primo elemento anche gli elementi successivi.

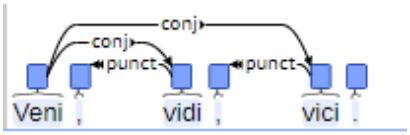

immagine 5 – linee guida

2. La presenza di un inciso può costituire un elemento di disturbo notevole, e portare ad una analisi non corretta, come mostrato in *immagine 6.* Questo esempio introduce quanto sarà meglio discusso successivamente, cioè come il criterio di prossimità sia determinante per l’assegnazione di una analisi corretta alla frase.

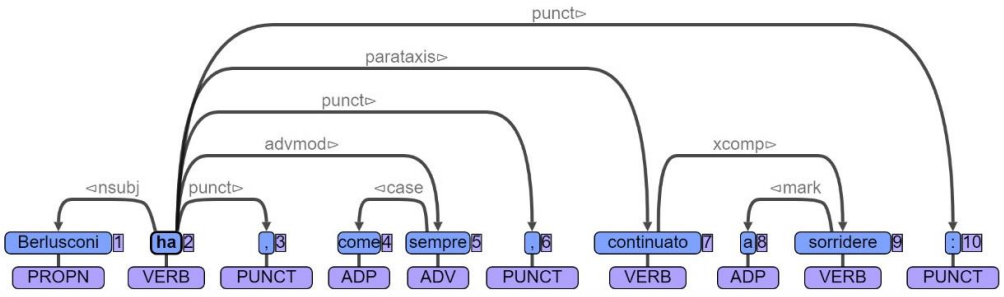

immagine 6

3. Per sua costruzione, UDpipe non collega il pronome al referente. Nonostante sia riconosciuta la frase relativa, con una buona affidabilità, il pronome relativo è interpretato nel suo ruolo sintattico della subordinata relativa, ma non è connesso all'antecedente. Questo rende più solido il modello perché anche con un pronome scorretto (nel senso di non concordante in genere e numero) è comunque prodotta una analisi del periodo corretta. Se la frase da analizzare fosse "la cucina, del cui si intravede il portone", l'analisi restituita resterebbe la medesima dell'*immagine 7*.

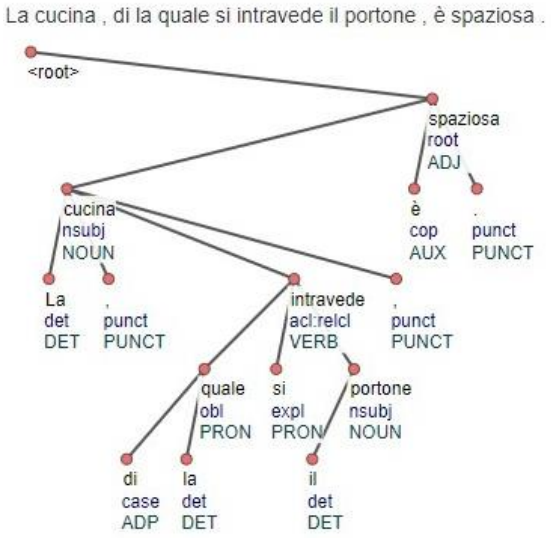

immagine 7

4. I casi (piuttosto rari) in cui la relativa non è stata riconosciuta, si possono ricondurre alla generale tendenza di cercare la testa nelle immediate vicinanze del dipendente. Questo è quanto intendo per criterio di prossimità. Il parlante umano invece affida la disambiguazione proprio all'accordo in genere e numero, e quindi possiamo notare un approccio totalmente diverso del modello rispetto a quello naturale. L'esempio dell'*immagine 8* è stato costruito ad hoc per indagare questo aspetto.

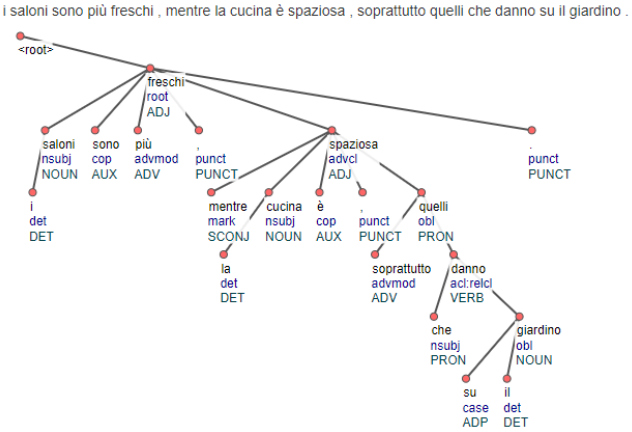

immagine 8

In questa frase esempio ci sono due soggetti, uno maschile e plurale (i saloni) e uno femminile e singolare (la cucina). Il pronome relativo è maschile e plurale, ma la clausola relativa è assegnata alla frase immediatamente precedente, che ha soggetto femminile e singolare. Questa cosa è data da un limite se vogliamo così dire tecnico della macchina, dato che il modello estrae feature contestuali solo per limitate parole precedenti e successive.

Inter-Annotator Agreement

Di seguito sono riportati i valori dei due indicatori *Average Observed Agreement* e *K di Cohen*, sia per quanto riguarda le categorie morfosintattiche che per le relazioni di dipendenza sintattica:

	Cucina1 lazzariturco	Cucina 2 Centosessanta maniere	Cucina 3 Boni	Giornale 1 La Stampa	Giornale 2 Corriere
K Cohen	0.957	0.983	0.981	0.988	0.975
Average Observed Agreement	0.962	0.985	0.983	0.989	0.977

tabella 2 – annotazione delle categorie morfosintattiche

	Cucina-1 lazzariturco	Cucina-2 Centosessanta maniere	Cucina-3 Boni	Giornale-1 La Stampa	Giornale-2 Corriere
K Cohen	0.747	0.857	0.753	0.881	0.902
Average Observed Agreement	0.749	0.855	0.751	0.882	0.903

tabella 3 – annotazione delle relazioni di dipendenza sintattica

Operando un'analisi quantitativa delle disuguaglianze fra le analisi dei due annotatori, si arriva ai dati contenuti in *tabella 4*, che mostrano le tre principali discordanze fra annotatori per quanto riguarda il *POS-tagging,* e le loro ricorrenze fra parentesi.

Cucina-1	adj-verb (3)	adj-noun (2)	det-pron(2)
Cucina-2	adj-verb(4)	adp-adv (1)	det-num(1)
Cucina-3	noun-verb(2)	det-pron(1)	aux-verb(1)
Giornale-1	adj-noun(2)	adj-propn(2)	noun-propn(1)
Giornale-2	adj-noun (4)	noun-verb(1)	pron-sconj(1)

tabella 4

In *tabella 5* sono invece mostrate le discordanze nell'assegnazione delle dipendenze sintattiche: tutte appaiono motivabili e collegate alla superficiale conoscenza del tagset, o in alcuni casi alla peculiarità del testo.

Cucina-1	nmod-obl (5)	obj-obl (4)	case-fixed(2)
Cucina-2	nmod-obl (5)	acl-amod (4)	case-fixed(2)
Cucina-3	expl-expl:impers(3)	appos-nmod(2)	obj-obl (2)
Giornale-1	nmod-obl (10)	acl-advcl(2)	acl-advcl(2)
Giornale-2	appos-obj(3)	nsubj-root(3)	flat-flat:name(2)

tabella 5

Golden Corpus

Per la costituzione del Golden Corpus, oltre al chiarimento delle discordanze mostrate nelle tabelle sopra, è stato necessario adottare linee guida più rigide per attribuire la testa della dipendenza agli elementi di punteggiatura. Corrette queste, si è potuto procedere con la valutazione dei modelli ISDT e PoSTWITA.

Valutazione di ISDT

F1 Score	Cucina1 lazzariturco	Cucina 2 Centosessanta maniere	Cucina 3 Boni	Giornale 1 La Stampa	Giornale 2 Corriere
UPOS	92.67	95.83	98.98	99.07	96.61
XPOS	93.13	95.67	98.64	98.76	96.76
UFeats	92.37	92.89	98.47	99.23	97.64
AllTags	90.53	92.58	97.96	98.45	95.73
Lemmas	93.89	96.91	97.61	98.76	96.76
UAS	84.27	83.15	93.87	88.39	89.84
LAS	75.73	78.67	90.63	86.84	87.33
CLAS	64.87	72.73	86.25	83.77	81.74
MLAS	60.06	67.49	85.31	82.14	77.97
BLEX	60.06	69.42	83.44	82.47	76.81

tabella 6

Per i testi giornalistici i valori di accuratezza sono piuttosto vicini fra loro, sia per la morfosintassi che per la sintassi. Nei testi di cucina invece i valori sono sbilanciati: Cucina-3 ha dei risultati molto migliori degli altri due, in alcuni casi anche drasticamente. Infatti, gli indicatori CLAS, MLAS e BLEX sono del 30% peggiori in Cucina-1 e 20% peggiori in Cucina-2 rispetto a Cucina-3.

Valutazione di PoSTWITA

	Cucina1 lazzariturco	Cucina 2 Centosessanta maniere	Cucina 3 Boni	Giornale 1 La Stampa	Giornale 2 Corriere
UPOS	89.16	88.10	90.97	95.05	92.49
XPOS	87.48	86.55	90.12	94.28	91.61
UFeats	85.65	84.39	88.07	93.35	91.16
AllTags	81.53	80.68	85.18	91.04	87.92
Lemmas	86.41	89.03	88.42	93.35	92.05
UAS	74.35	76.97	77.51	75.27	78.65
LAS	65.19	69.71	69.85	71.41	73.93
CLAS	54.11	64.74	59.69	63.49	68.82
MLAS	45.33	51.52	51.56	57.24	58.82
BLEX	43.63	55.9	47.81	57.57	60.00

tabella 7

Infine, generalmente PoSTWITA ha una performance inferiore rispetto a ISDT, e ciò è più marcato nei testi di cucina, con la perdita di 10-15 punti percentuali comune a tutti gli indicatori.

Conclusioni

Questo progetto ci ha dato la possibilità di riflettere sia sulle trivialità operative che sui meccanismi teorici di un *parser data-driven*, ma tre sono gli aspetti mostratisi fondamentali, che senz'altro approfindirò in lavori successivi:

1. Le problematiche derivanti dal *Domain adaptation*
2. Le limitazioni poste dal criterio di prossimità
3. Il peso del contesto per l'accuratezza dell'analisi

Post-scriptum:

Ci siamo imbattuti alla fine del lavoro in clitici non splittati che erano stati sotto i nostri occhi tutto il tempo senza che li vedessimo, e a volte abbiamo dovuto rifare tutta una parte perché siamo stati disordinati con i file. A questo si accompagna il fatto che più si prende dimestichezza con il *tagset* e più si riesce ad essere coerenti e precisi. Infine, l'approccio quantitativo alle proprie osservazioni è fondamentale: annotando si ha l'impressione che praticamente nessun clitico sia stato correttamente diviso, mentre dopo aver fatto i calcoli ci si rende conto che circa la metà sono corretti.

Ma tutte queste cose le abbiamo scoperte in fieri e quindi è giusto che stiano nella conclusione.