Analisi degli errori di un parser di Universal Dependencies

Christian Colonna

Novembre 2019

Contents

1 Introduzione

In questo report vengono presentati i dettagli dell'analisi automatica degli errori di un parser di Universal Dependecy 2.0[1], frutto del lavoro di collaborazione con Valerio Di Carlo e Mario Caruso di BUP[2]. L'analisi prende spunto dalla metodologia proposta da Ryan McDonald e Joakim Nivre [1], per vedere il tipo di errori linguistici commessi dal parser nell'analisi sintattica della frase. La metodologia di analisi è stata applicata a partire dal dataset italiano ISDT UD[3]: il parser è stato allenato sulla porzione di training del dataset ed è stato eseguito sulla porzione di testing; l'analisi dei risultati è stata basata sul confronto tra le annotazioni gold del test-set e le predizioni prodotte dal parser su di esso. Da

[1] `https://universaldependencies.org`
[2] `https://www.bupsolutions.com`
[3] `https://universaldependencies.org/treebanks/it_isdt/index.html`

qui in avanti, il test set annotato correttamente da umani verrà chiamato *gold dataset*, mentre il test set con le predizioni del parser verrà chiamato *system dataset*.

Il dataset analizzato contiene 482 frasi, per un totale di 10417 parole. Sono state usate come metriche di analisi UAS (Unlabelled Attachment Score) e LAS (Labelled Attachement Score). Rispetto la prima metrica, il parser risponde correttamente se, relativamente ad una dipendenza, connette giustamente le due parole con un arco. Rispetto la metrica LAS, affinché la predizione sia considerata corretta, il parser deve anche individuare la giusta relazione di dipendenza tra le due parole (o nodi), assegnando la giusta etichetta all'arco. La metodologia seguita applica in sostanza queste due metriche a diverse partizioni delle parole contenute nel dataset, accomunate da una caratteristica comune (ad esempio, le parole partizionate rispetto alla lunghezza della frase di cui esse fanno parte, vedi Sez. 2.1).

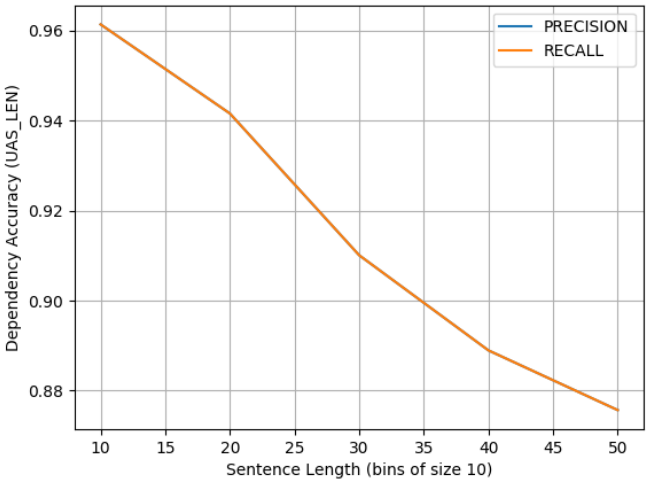
I tipi di errori analizzati possono essere divisi in tre classi: di lunghezza, di grafo, linguistici.

2 Fattori di lunghezza

2.1 Lunghezza delle frasi

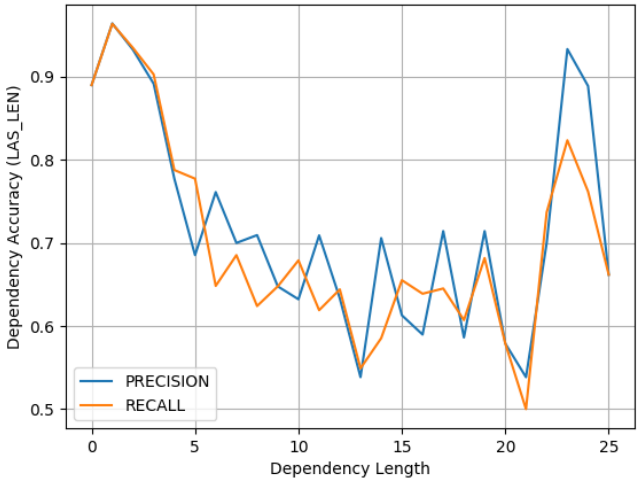
In questa analisi le parole sono state partizionate rispetto alla lunghezza della frase in cui esse appaiono. Le metriche sono state applicate suddividendo le frasi per lunghezza (bins da 10) e raggruppando tutte le frasi con lunghezza superiore a 50 (tutte le parole di frasi con lunghezza da 1 a 10, estremi inclusi, sono nel primo bin, tutte le parole di frasi con 51 o più parole sono nell'ultimo).

Come si può notare nella Figura 2.1, i valori di accuratezza diminuiscono all'aumentare della lunghezza delle frasi. Questo è un comportamento atteso, intrinseco al fatto che in frasi più brevi sono minori le decisioni che deve prendere il sistema e una decisione errata può ripercuotersi sulle quelle successive.



2.2 Lunghezza della dipendenza

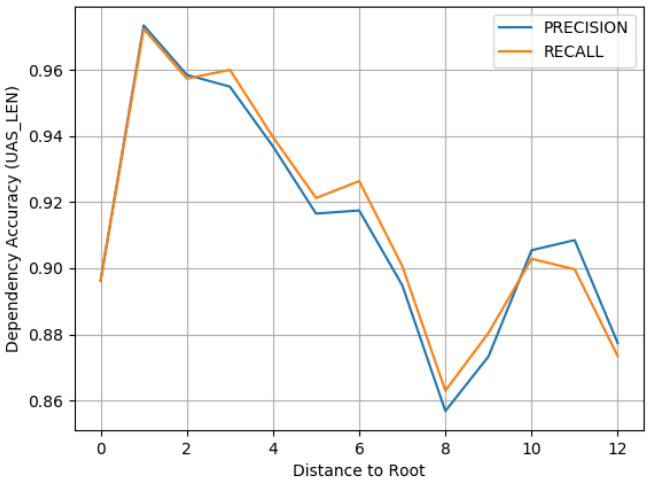
Con lunghezza di una dipendenza da una parola w_i ad una w_j intendiamo il valore assoluto $|i-j|$. Vediamo ancora che tanto più la parola "child" è lontana dalla parola "parent", tanto più è probabile per il parser risolva la dipendenza assegnandola in modo errato ad una delle parole compresa tra le due parole. Dai risultati nella Figura 2.2, si può inoltre notare che quando la lunghezza della dipendenza diventa molto alta le metriche tornino a crescere: questo potrebbe riflettere la presenza di elementi di punteggiatura assegnati correttamente alla root della frase.



3 Fattori del grafo

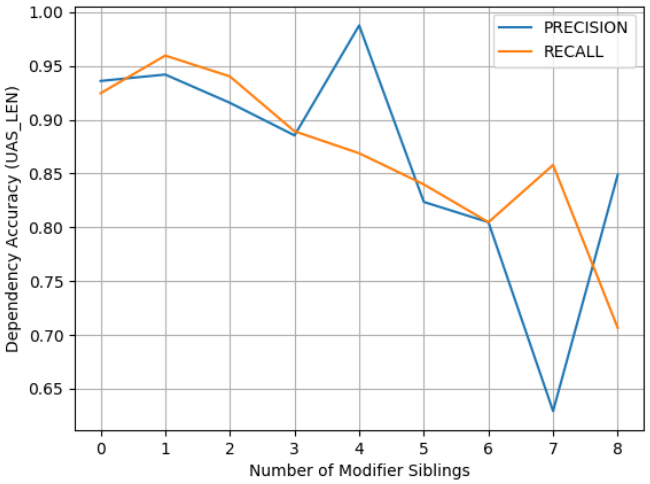
3.1 Distanza dalla root

Con questa analisi valutiamo le prestazioni del sistema nel predire correttamente gli archi in base alla distanza dalla root del grafo, laddove con d intendiamo il numero di archi presenti nel percorso inverso dal modificatore dell'arco alla root. Dai risultati della Figura 3.1, si nota come parole ad un livello gerarchico di dipendenza basso (distanti pochi archi dalla root del grafo) tendono ad ottenere punteggi migliori.


PRECISION
RECALL
Dependency Accuracy (UAS_LEN)
0.96
0.94
0.92
0.90
0.88
0.86
0
2
4
6
8
10
12
Distance to Root

3.2 Numero di modificatori fratelli

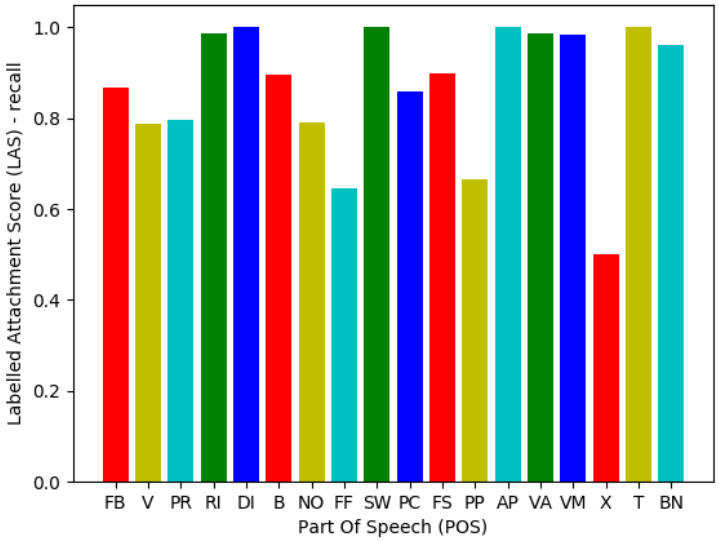
Dati due archi (w_i, w_j) e $(w_{i'}, w_{j'})$ gli archi sono considerati fratelli se rappresentano una dipendenza sintattica che parte dalla stessa parola, $i = i'$. Le parole vengono partizionate per numero di archi fratelli che si dipartono da esse.

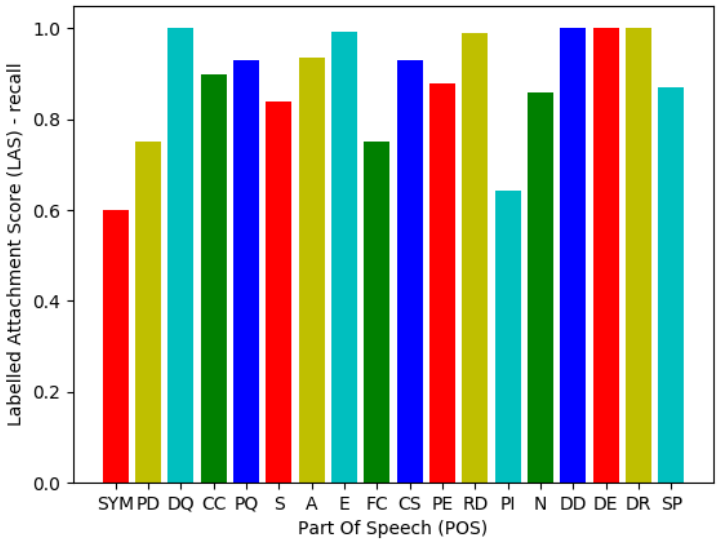


4 Fattori linguistici

4.1 Parti del discorso

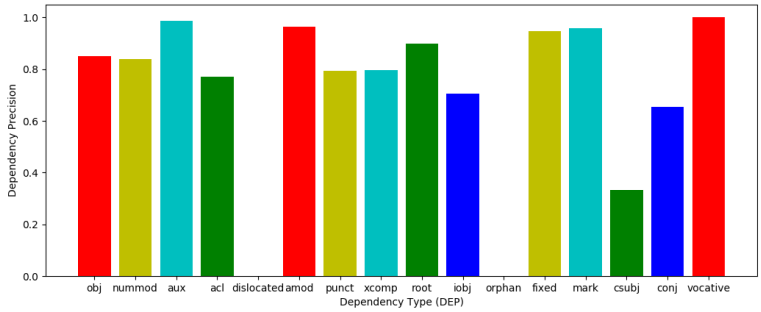
Le parti del discorso tendono ad essere sbagliate dal parser quando non sono canoniche. Con parti del discorso non canoniche intendiamo quelle categorie con funzioni solitamente riservate ad altre categorie. Esempio: l'aggettivo ha la funzione di modificatore (aggiunge qualità ad un nome). Participi passati sono usati in questa funzione al posto degli aggettivi.

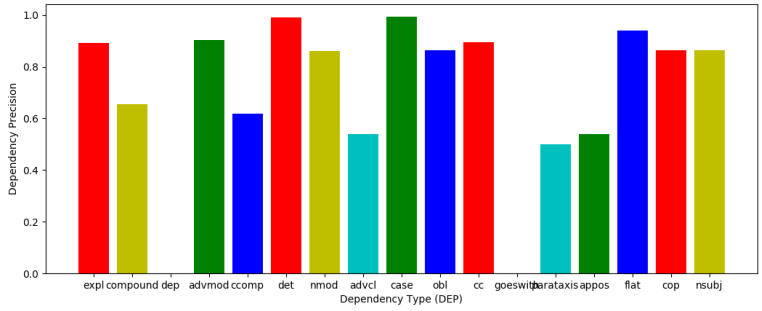



Labelled Attachment Score (LAS) - recall
1.0
0.8
0.6
0.4
0.2
0.0
SYM PD DQ CC PQ S A E FC CS PE RD PI N DD DE DR SP
Part Of Speech (POS)

4.2 Relazione di dipendenza

I dati sui fattori linguistici mostrano due classi di errori. Frase complessa e relazioni non canoniche. Con frase complessa intendiamo una frase che è composta da sottofrasi[4]: relative, avverbiali, temporali, soggettive, coordinazione. Un esempio è: *L'incidente è avvenuto quando la notte stava cadendo.* Qui si ha una relazione *advcl* - secondo la codifica delle UD - e un arco tra *cadendo* e *avvenuto* va tracciato. Altre relazioni di questo tipo sono: *ccomp*, *parataxis*, conj. Con relazioni non canoniche intendiamo quando parti del discorso non canoniche sono connesse da una relazione di dipendenza (vedi sezione 4.1). Un esempio é: *Sam, mio fratello, ... Mio fratello* è di solito *NP* ma qui viene usato per modificare o aggiungere dettagli su *Sam* comportandosi da aggettivo. Altre relazioni UD interessate da queste due classi di errori sono *csubj*, *appos*. La figura 4.2 e i dati sulle parti del discorso sembrano confermare queste ipotesi.





[4]Nel dominio linguistico non esistono sottofrasi. Si parla di frasi complete o indipendenti (possono sussistere da sole) e frasi dipendenti (completano frasi indipendenti o aggiungono dettagli). Esempi sono relative, soggettive, oggettive, temporali. Per astrarre dal dominio linguistico ci riferiamo a frasi complesse (es. una combinazioni di frase indipendente + dipendente) come somma di sottofrasi. Quindi chiameremo relative, temporal, ecc. generalmente sottofrasi.

5 Conclusioni

Le prestazioni del parser tendono a calare per vari motivi. In particolare quando si hanno:

- Frasi lunghe [2.1];
- Distanze lunghe tra due nodi connessi da un arco [2.2];
- Dipendenze lontane dalla root della frase [3.1];
- Nodi da cui si dipartono molte relazioni[3.2];
- Parti del discorso non canoniche[4.1];
- Relazioni che collegano sottofrasi [4.2];
- Relazioni non canoniche [4.2].

Certe dimensioni di errori si sovrappongono. Es: Relazioni che collegano sottofrasi comportano distanze lunghe tra due nodi connessi da un arco.

Fattori di lunghezza

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
UAS-len10	895	931	931	931	96.13	96.13	96.13	96.13
UAS-len20	2207	2344	2344	2344	94.16	94.16	94.16	94.16
UAS-len30	2176	2391	2391	2391	91.01	91.01	91.01	91.01
UAS-len40	1649	1855	1855	1855	88.89	88.89	88.89	88.89
UAS-len50	2536	2896	2896	2896	87.57	87.57	87.57	87.57

Table 1: Lunghezza della frase (UAS)

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len10	877	931	931	931	94.20	94.20	94.20	94.20
LAS-len20	2167	2344	2344	2344	92.45	92.45	92.45	92.45
LAS-len30	2143	2391	2391	2391	89.63	89.63	89.63	89.63
LAS-len40	1625	1855	1855	1855	87.60	87.60	87.60	87.60
LAS-len50	2502	2896	2896	2896	86.40	86.40	86.40	86.40

Table 2: Lunghezza della frase (LAS)

Appendice

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
UAS-len0	432	482	482	482	89.63	89.63	89.63	89.63
UAS-len1	4055	4170	4168	4170	97.29	97.24	97.27	97.24
UAS-len2	2385	2498	2504	2498	95.25	95.48	95.36	95.48
UAS-len3	1092	1194	1209	1194	90.32	91.46	90.89	91.46
UAS-len4	429	523	530	523	80.94	82.03	81.48	82.03
UAS-len5	227	283	321	283	70.72	80.21	75.17	80.21
UAS-len6	157	236	201	236	78.11	66.53	71.85	66.53
UAS-len7	99	143	140	143	70.71	69.23	69.96	69.23
UAS-len8	87	133	117	133	74.36	65.41	69.60	65.41
UAS-len9	58	88	88	88	65.91	65.91	65.91	65.91
UAS-len10	60	81	87	81	68.97	74.07	71.43	74.07
UAS-len11	39	63	55	63	70.91	61.90	66.10	61.90
UAS-len12	39	59	60	59	65.00	66.10	65.55	66.10
UAS-len13	30	51	52	51	57.69	58.82	58.25	58.82
UAS-len14	25	41	34	41	73.53	60.98	66.67	60.98
UAS-len15	20	29	31	29	64.52	68.97	66.67	68.97
UAS-len16	23	36	39	36	58.97	63.89	61.33	63.89
UAS-len17	21	31	28	31	75.00	67.74	71.19	67.74
UAS-len18	17	28	29	28	58.62	60.71	59.65	60.71
UAS-len19	15	22	21	22	71.43	68.18	69.77	68.18
UAS-len20	11	19	19	19	57.89	57.89	57.89	57.89
UAS-len21	7	14	13	14	53.85	50.00	51.85	50.00
UAS-len22	14	19	20	19	70.00	73.68	71.79	73.68
UAS-len23	14	17	15	17	93.33	82.35	87.50	82.35
UAS-len24	16	21	18	21	88.89	76.19	82.05	76.19
UAS-len25	91	136	136	136	66.91	66.91	66.91	66.91

Table 3: Lunghezza della dipendenza (UAS)

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len0	429	482	482	482	89.00	89.00	89.00	89.00
LAS-len1	4019	4170	4168	4170	96.43	96.38	96.40	96.38
LAS-len2	2335	2498	2504	2498	93.25	93.47	93.36	93.47
LAS-len3	1078	1194	1209	1194	89.16	90.28	89.72	90.28
LAS-len4	412	523	530	523	77.74	78.78	78.25	78.78
LAS-len5	220	283	321	283	68.54	77.74	72.85	77.74
LAS-len6	153	236	201	236	76.12	64.83	70.02	64.83
LAS-len7	98	143	140	143	70.00	68.53	69.26	68.53
LAS-len8	83	133	117	133	70.94	62.41	66.40	62.41
LAS-len9	57	88	88	88	64.77	64.77	64.77	64.77
LAS-len10	55	81	87	81	63.22	67.90	65.48	67.90
LAS-len11	39	63	55	63	70.91	61.90	66.10	61.90
LAS-len12	38	59	60	59	63.33	64.41	63.87	64.41
LAS-len13	28	51	52	51	53.85	54.90	54.37	54.90
LAS-len14	24	41	34	41	70.59	58.54	64.00	58.54
LAS-len15	19	29	31	29	61.29	65.52	63.33	65.52
LAS-len16	23	36	39	36	58.97	63.89	61.33	63.89
LAS-len17	20	31	28	31	71.43	64.52	67.80	64.52
LAS-len18	17	28	29	28	58.62	60.71	59.65	60.71
LAS-len19	15	22	21	22	71.43	68.18	69.77	68.18
LAS-len20	11	19	19	19	57.89	57.89	57.89	57.89
LAS-len21	7	14	13	14	53.85	50.00	51.85	50.00
LAS-len22	14	19	20	19	70.00	73.68	71.79	73.68
LAS-len23	14	17	15	17	93.33	82.35	87.50	82.35
LAS-len24	16	21	18	21	88.89	76.19	82.05	76.19
LAS-len25	90	136	136	136	66.18	66.18	66.18	66.18

Table 4: Lunghezza della dipendenza (LAS)

Fattori del grafo

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
UAS-len0	432	482	482	482	89.63	89.63	89.63	89.63
UAS-len1	879	904	903	904	97.34	97.23	97.29	97.23
UAS-len2	808	844	843	844	95.85	95.73	95.79	95.73
UAS-len3	720	750	754	750	95.49	96.00	95.74	96.00
UAS-len4	623	663	665	663	93.68	93.97	93.83	93.97
UAS-len5	538	584	587	584	91.65	92.12	91.89	92.12
UAS-len6	478	516	521	516	91.75	92.64	92.19	92.64
UAS-len7	417	463	466	463	89.48	90.06	89.77	90.06
UAS-len8	359	416	419	416	85.68	86.30	85.99	86.30
UAS-len9	324	368	371	368	87.33	88.04	87.69	88.04
UAS-len10	316	350	349	350	90.54	90.29	90.41	90.29
UAS-len11	278	309	306	309	90.85	89.97	90.41	89.97
UAS-len12	3291	3768	3751	3768	87.74	87.34	87.54	87.34

Table 5: Distanza dalla root (UAS)

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len0	429	482	482	482	89.00	89.00	89.00	89.00
LAS-len1	864	904	903	904	95.68	95.58	95.63	95.58
LAS-len2	782	844	843	844	92.76	92.65	92.71	92.65
LAS-len3	710	750	754	750	94.16	94.67	94.41	94.67
LAS-len4	614	663	665	663	92.33	92.61	92.47	92.61
LAS-len5	532	584	587	584	90.63	91.10	90.86	91.10
LAS-len6	472	516	521	516	90.60	91.47	91.03	91.47
LAS-len7	415	463	466	463	89.06	89.63	89.34	89.63
LAS-len8	355	416	419	416	84.73	85.34	85.03	85.34
LAS-len9	316	368	371	368	85.18	85.87	85.52	85.87
LAS-len10	310	350	349	350	88.83	88.57	88.70	88.57
LAS-len11	276	309	306	309	90.20	89.32	89.76	89.32
LAS-len12	3239	3768	3751	3768	86.35	85.96	86.16	85.96

Table 6: Distanza dalla root (LAS)

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
UAS-len0	1289	1394	1377	1394	93.61	92.47	93.04	92.47
UAS-len1	2048	2134	2174	2134	94.20	95.97	95.08	95.97
UAS-len2	2316	2463	2529	2463	91.58	94.03	92.79	94.03
UAS-len3	1647	1852	1860	1852	88.55	88.93	88.74	88.93
UAS-len4	1047	1205	1060	1205	98.77	86.89	92.45	86.89
UAS-len5	509	606	618	606	82.36	83.99	83.17	83.99
UAS-len6	338	420	420	420	80.48	80.48	80.48	80.48
UAS-len7	151	176	240	176	62.92	85.80	72.60	85.80
UAS-len8	118	167	139	167	84.89	70.66	77.12	70.66

Table 7: Modificatori Fratelli (UAS)

Metric	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len0	1280	1394	1377	1394	92.96	91.82	92.39	91.82
LAS-len1	2033	2134	2174	2134	93.51	95.27	94.38	95.27
LAS-len2	2294	2463	2529	2463	90.71	93.14	91.91	93.14
LAS-len3	1606	1852	1860	1852	86.34	86.72	86.53	86.72
LAS-len4	1018	1205	1060	1205	96.04	84.48	89.89	84.48
LAS-len5	493	606	618	606	79.77	81.35	80.56	81.35
LAS-len6	328	420	420	420	78.10	78.10	78.10	78.10
LAS-len7	146	176	240	176	60.83	82.95	70.19	82.95
LAS-len8	116	167	139	167	83.45	69.46	75.82	69.46

Table 8: Modificatori Fratelli (LAS)

Fattori linguistici

Metric	Pos	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len	SYM	3	5	5	5	60.00	60.00	60.00	60.00
LAS-len	RI	152	154	154	154	98.70	98.70	98.70	98.70
LAS-len	SW	11	11	11	11	100.00	100.00	100.00	100.00
LAS-len	DQ	43	43	43	43	100.00	100.00	100.00	100.00
LAS-len	B	289	323	323	323	89.47	89.47	89.47	89.47
LAS-len	RD	1344	1359	1359	1359	98.90	98.90	98.90	98.90
LAS-len	FC	21	28	28	28	75.00	75.00	75.00	75.00
LAS-len	PP	2	3	3	3	66.67	66.67	66.67	66.67
LAS-len	A	600	642	642	642	93.46	93.46	93.46	93.46
LAS-len	S	1738	2070	2070	2070	83.96	83.96	83.96	83.96
LAS-len	PR	97	122	122	122	79.51	79.51	79.51	79.51
LAS-len	DR	1	1	1	1	100.00	100.00	100.00	100.00
LAS-len	DE	1	1	1	1	100.00	100.00	100.00	100.00
LAS-len	N	147	171	171	171	85.96	85.96	85.96	85.96
LAS-len	DI	46	46	46	46	100.00	100.00	100.00	100.00
LAS-len	BN	75	78	78	78	96.15	96.15	96.15	96.15
LAS-len	FB	195	225	225	225	86.67	86.67	86.67	86.67
LAS-len	V	760	966	966	966	78.67	78.67	78.67	78.67
LAS-len	T	15	15	15	15	100.00	100.00	100.00	100.00
LAS-len	PQ	39	42	42	42	92.86	92.86	92.86	92.86
LAS-len	CC	236	263	263	263	89.73	89.73	89.73	89.73
LAS-len	DD	32	32	32	32	100.00	100.00	100.00	100.00
LAS-len	NO	30	38	38	38	78.95	78.95	78.95	78.95
LAS-len	VA	240	243	243	243	98.77	98.77	98.77	98.77
LAS-len	VM	58	59	59	59	98.31	98.31	98.31	98.31
LAS-len	SP	439	505	505	505	86.93	86.93	86.93	86.93
LAS-len	FF	300	465	465	465	64.52	64.52	64.52	64.52
LAS-len	PE	29	33	33	33	87.88	87.88	87.88	87.88
LAS-len	PC	141	164	164	164	85.98	85.98	85.98	85.98
LAS-len	PD	15	20	20	20	75.00	75.00	75.00	75.00
LAS-len	X	1	2	2	2	50.00	50.00	50.00	50.00
LAS-len	E	1632	1643	1643	1643	99.33	99.33	99.33	99.33
LAS-len	CS	92	99	99	99	92.93	92.93	92.93	92.93
LAS-len	FS	411	457	457	457	89.93	89.93	89.93	89.93
LAS-len	AP	61	61	61	61	100.00	100.00	100.00	100.00
LAS-len	PI	18	28	28	28	64.29	64.29	64.29	64.29

Table 9: Parti del Discorso (LAS)

Metric	DepRel	Correct	Gold	Pred	Alignd	Prec	Rec	F1	AligndAcc
LAS-len	mark	183	190	191	190	95.81	96.32	96.06	96.32
LAS-len	csubj	2	3	6	3	33.33	66.67	44.44	66.67
LAS-len	advcl	75	133	139	133	53.96	56.39	55.15	56.39
LAS-len	nmod	728	840	847	840	85.95	86.67	86.31	86.67
LAS-len	dep	0	1	1	1	0.00	0.00	0.00	0.00
LAS-len	goeswith	0	1	0	1	0.00	0.00	0.00	0.00
LAS-len	iobj	12	20	17	20	70.59	60.00	64.86	60.00
LAS-len	det	1695	1711	1713	1711	98.95	99.06	99.01	99.06
LAS-len	nsubj	455	535	526	535	86.50	85.05	85.77	85.05
LAS-len	appos	20	40	37	40	54.05	50.00	51.95	50.00
LAS-len	obj	304	336	358	336	84.92	90.48	87.61	90.48
LAS-len	acl	158	213	205	213	77.07	74.18	75.60	74.18
LAS-len	case	1532	1542	1543	1542	99.29	99.35	99.32	99.35
LAS-len	vocative	1	3	1	3	100.00	33.33	50.00	33.33
LAS-len	advmod	338	369	374	369	90.37	91.60	90.98	91.60
LAS-len	cop	96	105	111	105	86.49	91.43	88.89	91.43
LAS-len	conj	190	311	290	311	65.52	61.09	63.23	61.09
LAS-len	flat	140	143	149	143	93.96	97.90	95.89	97.90
LAS-len	nummod	104	114	124	114	83.87	91.23	87.39	91.23
LAS-len	expl	100	105	112	105	89.29	95.24	92.17	95.24
LAS-len	ccomp	26	38	42	38	61.90	68.42	65.00	68.42
LAS-len	cc	236	263	264	263	89.39	89.73	89.56	89.73
LAS-len	parataxis	5	14	10	14	50.00	35.71	41.67	35.71
LAS-len	orphan	0	1	6	1	0.00	0.00	0.00	0.00
LAS-len	xcomp	59	78	74	78	79.73	75.64	77.63	75.64
LAS-len	compound	21	23	32	23	65.62	91.30	76.36	91.30
LAS-len	punct	927	1175	1169	1175	79.30	78.89	79.10	78.89
LAS-len	amod	564	586	586	586	96.25	96.25	96.25	96.25
LAS-len	dislocated	0	1	0	1	0.00	0.00	0.00	0.00
LAS-len	aux	297	300	301	300	98.67	99.00	98.84	99.00
LAS-len	fixed	36	42	38	42	94.74	85.71	90.00	85.71
LAS-len	root	429	482	478	482	89.75	89.00	89.38	89.00
LAS-len	obl	581	699	673	699	86.33	83.12	84.69	83.12

Table 10: Relazione di Dipendenza (LAS)

References

[1] Ryan McDonald and Joakim Nivre. "Characterizing the errors of data-driven dependency parsing models". In: *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007.