Dependency Treebank Annotation and Null Elements: an experiment with VIT

Rodolfo Delmonte
Ca' Foscari University Venice

Overview

- Treebanks of Dependency Structures: what for?
- The Importance of Null Elements: related problems
- Are NEs genre/domain dependent? (they are already language dependent)
- An experiment with VIT: a semi-automatic algorithm to populate a surface level dependency treebank
- Open issues: many!
  - The treatment of relative pronouns
- Conclusions & Future work

Treebanks and Predicate-Argument Structures

- La ragione primaria per la creazione di treebank sintattici dovrebbe essere quella di permettere il mapping semantico in strutture predicato-argomentali e quindi indirettamente costituire un strumento per il training di parser profondi che siano in grado di costruire delle rappresentazioni in forma logica consistenti.
- Perlomeno questo era quanto Marcus ('94) aveva in mente nel dirigere i lavori per la creazione del Penn Treebank.

Treebanks and Predicate-Argument Structures

- I problemi però vennero messi presto in luce dalla difficoltà oggettiva di annotare i Null Elements e di definire in maniera chiara e precisa una strategia per la loro consistenza semantica, indicando ogni volta un antecedente che li potesse riempire adeguatamenmte in Logical Form.
- Successivamente sorsero altri problemi legati all'uso del treebank per scopi di apprendimento automatico. Come Gaizauskas, 1995 notò, la presenza di NEs "did not facilitate structural learning and could not prevent the probabilistic engine to postulate the existence of deprecated null elements everywhere".

State of the art treebanks (de facto)

- Penn Treebank la treebank de facto stato dell'arte nella sua seconda versione ha annotato i null elements. I numeri di riferimento sono i seguenti:
- - 36862 casi di null elements (66.3% Tot-utterances)
- - 8416 non sono coindicizzati (22.83%) per problemi strettamente legati alla difficoltà dell'annotazione
- - 12172, escludendo tutte le tracce sintattiche di tipo wh- e topicalizzazione e limitandoci alla categoria definita OTHER TRACES, che include i SUBJ inespressi
- Su un totale di 55600 enunciati, che si decompongono in 93532 frasi semplici o clausole

Treebanks of Dependency Structures

- La prima cosa che ci chiediamo è se la presenza di NEs è in funzione di qualche variabile, tipo il genere o il dominio
- Ho quindi considerato TUT e VIT confrontando diverse parti del corpus per evidenziare le differenze
- Ne è risultata una forte dipendenza dal dominio o genere che determina in maniera decisiva il numero di NE e anche il tipo
- In buona sostanza è lo stile nel suo complesso che porta a un incremento di NE

TUT Treebank: statistics for 2 subcorpora

Type	NEWS	CIV.COD	%
sents	700	1100	63.6
tokens	18046	28050	64.3
Conll V-V	2270	3519	64.5
Conll Subj	1936	3016	64.1
Tut V-subjs	2051	3924	52.2
Tut N-subj	129	134	96.2
Tut Pr-subj	210	189	111,1

TUT Treebank: statistics for 2 subcorpora

Type	NEWS	CIV.COD	%
sents	700	1100	63.6
tokens	18046	28050	64.3
traces	1090	2621	41.5
generic	291	1005	28.9
Pron-rel	280	517	54,1
Rel-subj	192	334	57.4
Rel-obj	27	60	45.0
Rel-obl	24	84	28.5
Rel-loc	21	3	700.
Rel-adj	13	39	33.3
visitor	6	7	85.7

Algorithm for NE generation

- Primo step: individuazione dei soggetti non-canonici classificati in tre tipologie; semi-automatico
- Secondo step: individuazione dei SI impersonali e coindicizzazione del pro_piccolo taggato s_impers – semi-automatico
- Terzo step: Controllo sintattico frasi relative – semi-automatico
- Quarto step: pro_piccolo a soggetti frasali dislocati (infinitive e completive soggetto) – automatico (difficile)

Algorithm for NE generation

- Quinto step: pro_piccolo di verbi impersonali, atmosferici ecc., su base strettamente lessicale
- Sesto step: pro_piccolo frasi a verbo flesso (coindicizzato solo nei casi semplici, cioè in presenza di coordinazioni; l'algoritmo di anaphora resolution non è stato ancora agganciato)
- Settimo step: pro_grande frasi di modo indefinito (coindicizzato automaticamente sulla base della regola di default introdotta da LFG per cui sono necessari gli argomenti core, cioè soggetto, oggetto e oggetto indiretto)
- Ottavo step: i nuovi items linguistici vengono aggiunti alla struttura con una marca distintiva nell'indice, '.10' per i pro_piccoli e '.11' per i pro_grandi

Canonical and Non-canonical Subjects in VIT

Type of .Struc.	Freq. occurr.	
SUBJ (lexical)	6166	
S_DIS	1037	
S_TOP	2165	
S_FOC	266	
Total NC	3468	36%
Total Subjects	9634	
Tensed verbs	9369	
Mood irrealis	2613	
Copulative verbs	3892	
Total Tensed	15874	66.5%

Statistics for NE in VIT and fragment

Type	Vit-frag	VIT-all	%
sents	500	10195	4.9
tokens	15355	275520	5.5
V-main	500	10166	4.9
V-subjs_lex	382	9634	3.9
V-subjs	524	15874	3.3
V-subjs:expl,imp	130=654	15874	4.2
PastPart	513	6748	7.6
Infinit	224	2766	8.1
Gerund	50	663	7.5
Total/Indef	787	10177	7.7
V-subjs:s_impl	425		
V-subjs, ant=	445		
V-subjs, ant=\d	233		

Some examples of VIT fragment

18 quando quando cosu(conjunction_subordinate) fs [] 20 fs par
19 si si clit(clitic_pronoun) ibar per=3|gen=m|num=sp 20 ibar nom
20 arriva arrivare vin(verb_intrans_tensed) ibar punt 30 ibar unac/posit
20.11 pro si pro(little_pro) sn per=3|gen=m|num=sp 19 s_impers-theme_unaff nom

0 Si si clit(clitic_pronoun) ibar - 1 ibar nil
1 tratta trattare vin(verb_intrans_tensed) cl(main) punt - ibar refl/exten
1.11 pro si pro(little_pro) nil num=s|gen=m ant=0 s_expl com
2 del di partd(preposition_di_plus_article) spd num=s|gen=m 1 obl det

Treatment of Relative Pronouns

- In shallow or surface dependency treebanks, relative pronouns are only visible if lexically expressed. So the case of implicit relative pronoun signalled by CHE complementizer does not exist – but it does in all deep dependency treebanks as we will show below. What is usually done, is the transformation of CHE itself into a relative pronoun like CHI, CUI or QUALE and others. We will discuss other cases below. However, even though this is what all shallow treebanks do, the treatment of CHE is not uniform.

Treatment of Relative Pronouns

- As with the complementizer of sentential complements, we have come up with two different approaches to the problem of treating CHE/QUE/ THAT and others relative pronouns
- linked to the governing Noun head (VIT)
  - the governed verb of the relative clause linked to CHE
- linked to the governed verb in the relative clause (ALL Other TBs)
  - the governed verb being an auxiliary if present (TUT, PTB)
  - the governed verb being the lexical semantic verb (ALL Other TBs)

Treatment of Relative Pronouns – ISST/TALN

21	produrre	produrre	V	V	mod=f	5	sub	_	_		
22	individui	individuo	S	S	num=p|gen=m	21	obj	_	_		
23	che	che	P	PR	num=n|gen=n	24	subj	_	_		
24	sanno	sapere	V	V	num=p|per=3|mod=i|ten=p	22	mod_rel	_	_		
25	fare	fare	V	V	mod=f	24	arg	_	_		
26	cose	cosa	S	S	num=p|gen=f	25	obj	_	_		
27	che	che	P	PR	num=n|gen=n	33	obj	_	_		
28	essi	essi	P	PE	num=p|per=3|gen=m	33	subj	_	_		
29	non	non	B	BN	_	33	neg	_	_		
30	potranno	potere	V	VM	num=p|per=3|mod=i|ten=f	33	modal	_	_		
31	mai	mai	B	B	_	33	mod_temp	_	_		
32	nemmeno	nemmeno	B	B	_	33	mod	_	_		
33	immaginare	immaginare	V	V	mod=f	26	mod_rel	_	_		

Treatment of Relative Pronouns - TUT

1	Nelle	IN	PREP	PREP	MONO	15	RMOD	_	_			
2	Nelle	IL	ART	ART	DEF|F|PL	1	ARG	_	_			
3	societa'	SOCIETÀ	NOUN	NOUN	COMMON|F|ALLVAL	2	ARG	_	_			
4	di	DI	PREP	PREP	MONO	3	RMOD	_	_			
5	Tirana	TIRANA	NOUN	NOUN	PROPER|F|SING|££CITY	4	ARG	_	_			
6	che	CHE	PRON	PRON	RELAT|ALLVAL|ALLVAL|LSUBJ+LOBJ	8	SUBJ	_	_			
7	hanno	AVERE	VERB	VERB	AUX|IND|PRES|TRANS|3|PL	8	AUX+TENSE	_	_			
8	truffato	TRUFFARE	VERB	VERB	MAIN|PARTICIPLE|PAST|TRANS|SING|M	3						
	RMOD+RELCL	_	_									
...												
14	c'	CI	PRON	PRON	LOC|ALLVAL|ALLVAL|LOC|CLITIC	15	RMOD	_	_			
15	e'	ESSERE	VERB	VERB	MAIN|IND|PRES|INTRANS|3|SING	0	TOP	_	_			

Treatment of Relative Pronouns - VIT

13 emergere emergere n(noun) sn num=s|gen=m 12 pobj com
14 di di pd(preposition_di) spd - 13 mod nil
15 una uno art(article) sn num=s|gen=f 17 sn ind
16 crescente crescente ag(adjective) sa num=s|per=fm 17 mod nil
17 concorrenza concorrenza n(noun) sn num=s|gen=f 14 pobj com
18 che che rel(relative) f2 - 17 subj-theme_aff nil
19 si si clit(clitic_pronoun) ibar per=3|gen=m|num=sp 22 ibar acc
20 è essere ause(auxiliary_essere_tensed) ibar punt 22 ibar aux
21 progressivamente progressivamente avv(adverb) ibar [] 22 adjv mn
22 spostata spostare vppin(verb_intrans_past_participle) ibar punt 18 ibar refl_in/posit

Tre Tipologie di controllo sintattico

- controllo diretto, il CHE o il pronome sono di tipo Soggetto o Oggetto o Oggetto Indiretto o Aggiunto locativo, temporale, modale (la dipendenza è con il verbo della relativa con la marca funzionale)

- controllo indiretto, il pronome è un modificatore o specificatore di un complemento della frase relativa, con verbi copulativi (la dipendenza è con il complemento della relativa con la marca funzionale)

Tre Tipologie di controllo sintattico

- controllo doppio, o pied piping, quando il pronome relativo modifica una testa locale, che a sua volta modifica un complemento della frase relativa. In questi casi una struttura a costituenti avrebbe messo in risalto la appartenenza del pronome relativo alla testa nominale interna. Questo caso non è rappresentabile con metodi superficiali (la dipendenza è con il nome di testa locale, il quale a sua volta dipende dal complemento della relativa di cui è modificatore)

Treatment of Relative Pronouns – VIT (direct)

19 Berlusconi Berlusconi nh(noun_human) sn propr 15 s_top-experiencer hum
20 che che rel(relative) f2 - 19 binder nil
21 è essere vc(verb_copulative) ibar punt 20 ibar cop/existence
21.11 pro pro pro(little_pro) sn num=s|per=3 ant=19 s_impl-tema_bound
22 industriale industriale n(noun) sn num=s 21 ncomp com

38 dell di partd(preposition_di_plus_article) spd num=s|per=fm 49 mod det
38.1 l il art sn num=s|per=fm 49 det def
39 ambiente ambiente n(noun) sn num=s|gen=m 38 pobj com
40 socio_economico socio_economico ag(adjective) sa num=s 39 mod nil
41 in in p(preposition) sp - 39 adj nil
42 cui cui relob(relative_oblique) sn [] 41 binder rel_obl
43 sono essere ause(auxiliary_essere_tensed) ibar punt 44 ibar aux
44 inserite inserire vppt(verb_trans_past_participle) ibar punt 39 ibar refl_in/into_hole
44.11 prep_relob in_ambiente prep_relob(prepositional_rel_oblique) sp num=s|gen=m ant=41_42 bindee com

Treatment of Relative Pronouns – VIT (indirect)

36 sulle su part(preposition_plus_article) sp num=p|gen=f 32 pcomp det
36.1 le il art sn num=p|gen=f 32 det def
37 due due num(numeral) sn [] 38 sn card
38 branche branca n(noun) sn num=p|gen=f 36 pobj com
39 operative operativo ag(adjective) sa num=p|gen=f 38 mod nil
40 , , punt(sentence_internal) sn punt 38 sn nil
41 di di pd(preposition_di) spd - 38 adj nil
42 cui cui relob(relative_oblique) sn [] 41 binder rel_obl
43 pure pure cong(conjunction) f2 [] 38 cong sum
44 è essere vc(verb_copulative) ibar punt 38 ibar cop/esistenza
45 nominalmente nominalmente avv(adverb) savv [] 46 adjm mn
46 responsabile responsabile ag(adjective) sa num=s|per=fm 44 acomp nil
46.11 prep_relob di_branca prep_relob(prepositional_rel_oblique) sp num=p|gen=f ant=41_42 bindee com

Treatment of Relative Pronouns – VIT (indirect)

21 fondi fondo n(noun) sn num=p|gen=m 19 obj com
22 di di pd(preposition_di) spd - 21 adj nil
23 cui cui relob(relative_oblique) sn [] 22 binder rel_obl
24 abbiamo avere vc(verb_copulative) ibar nil 21 ibar cop/stato
24.10 pro pro pro(little_pro) sn nil ant=7 s_impl-esperiente impL1p
25 bisogno bisogno n(noun) sn num=s|gen=m 24 ncomp com
25.11 prep_relob di_fondo prep_relob(prepositional_rel_oblique) sp num=p|gen=m ant=22_23 bindee com

8 commissione commissione n(noun) sn num=s|gen=f 6 pobj com
9 esteri estero ag(adjective) sa num=p|gen=m 8 mod nil
10 alla a part(preposition_plus_article) sp num=s|gen=f 8 adj det
10.1 la il art sn num=s|gen=f 8 det def
11 cui cui relob(relative_oblique) sp [] 10 sp rel_obl
12 presidenza presidenza n(noun) sn num=s|gen=f 10 pobj com
13 è essere vc(verb_copulative) ibar punt 8 ibar cop/esistenza
14 candidato candidato n(noun) sn num=s|gen=m 13 ncomp com
14.11 prep_relob alla_commissione prep_relob(prepositional_rel_oblique) sp num=s|gen=m ant=10_11 bindee com

CUI and similar pronouns in VIT

- There at least four different typologies of structure accompanying CUI oblique relative pronoun, and that we have found in VIT:
- 1. argument/adjunct of relative verb - it directly modifies the main verb of the relative clause
- 2. adjunct modifier of argument of relative verb - it modifies an argument of the verb of relative clause
- 3. adjunct modifier of a noun
- 4. adjunct modifier of the internal nominal head

Treatment of Relative Pronouns – VIT (indirect)

0 Non non neg(negation) ir_infl - 1 neg nil
1 sarà essere vcir(verb_copulative_mood_irrealis) cl(main) punt - ir_infl cop/esistenza
2 presente presente ag(adjective) sa num=s|per=fm 1 acomp nil
3 , , punt(sentence_internal) compc punt 1 compc nil
4 invece invece congf(conjunction_sentential) compc [] 1 cong av
5 , , punt(sentence_internal) compc punt 1 compc nil
6 l il art(article) sn num=s|gen=m 7 sn def
7 uomo uomo n(noun) sn num=m 1 s_top-tema_bound invar
...
19 la il art(article) sn num=s|gen=f 21 sn def
20 cui cui relob(relative_oblique) sn [] 7 sn rel_obl
21 posizione posizione n(noun) sn num=s|gen=f 24 subj-theme_unaff com
21.11 prep_relob di_uomo prep_relob(prepositional_rel_oblique) sp num=s|gen=m ant=7 bindee com
22 è essere ause(auxiliary_essere_tensed) ibar punt 24 ibar aux
23 stata essere ausep(auxiliary_essere_past_participle) ibar punt 24 ibar aux
24 stralciata stralciare vppt(verb_trans_past_participle) ibar punt 21 ibar tr/possess

Treatment of Relative Pronouns – TUT (direct)

1 Il (IL ART DEF M SING) [12;VERB-SUBJ]
2 punto (PUNTO_DI_VISTA NOUN COMMON M SING LOCUTION) [1;DET+DEF-ARG]
3 di (PUNTO_DI_VISTA NOUN COMMON LOCUTION) [2;CONTIN+LOCUT]
4 vista (PUNTO_DI_VISTA NOUN COMMON LOCUTION) [3;CONTIN+LOCUT]
5 da (DA PREP MONO) [7;VISITOR]
6 cui (CUI PRON RELAT LIOBJ+OBL) [5;PREP-ARG]
7 volevo (VOLERE VERB MOD IND IMPERF TRANS 1 SING) [2;VERB-RMOD+RELCL]
7.10 t [] (DEITT-T PRON PERS ALLVAL SING 1) [7;VERB-SUBJ]
8 raccontare (RACCONTARE VERB MAIN INFINITE PRES TRANS) [7;VERB+MODAL-INDCOMPL]
8.10 t [7.10p] (DEITT-T PRON PERS ALLVAL SING 1) [8;VERB-SUBJ]
8.11 t [5f] (DA PREP MONO) [8;PREP-RMOD-PERSPECT]
9 la (IL ART DEF F SING) [8;VERB-OBJ]
10 storia (STORIA NOUN COMMON F SING) [9;DET+DEF-ARG]
11 non (NON ADV NEG) [12;ADVB-RMOD-NEG]
12 era (ESSERE VERB MAIN IND IMPERF INTRANS 3 SING) [0;TOP-VERB]

Treatment of Relative Pronouns – TUT (indirect)

8 e` (ESSERE VERB MAIN IND PRES INTRANS 3 SING) [0;TOP-VERB]
9 il (IL ART DEF M SING) [8;VERB-PREDCOMPL+SUBJ]
10 coronamento (CORONAMENTO NOUN COMMON M SING CORONARE TRANS) [9;DET +DEF-ARG]
11 del (DI PREP MONO) [10;NOUN-OBJ]
11.1 del (IL ART DEF M SING) [11;PREP-ARG]
12 dialogo (DIALOGO NOUN COMMON M SING) [11.1;DET+DEF-ARG]
13 di (DI PREP MONO) [17;VISITOR]
14 cui (CUI PRON RELAT LIOBJ+OBL) [13;PREP-ARG]
15 oggi (OGGI ADV TIME) [17;ADVB-RMOD-TIME]
16 si (SI PRON REFL-IMPERS ALLVAL ALLVAL 3 LSUBJ+LOBJ+LIOBJ CLITIC) [17;VERB-SUBJ/VERB-SUBJ+IMPERS]
17 vedono (VEDERE VERB MAIN IND PRES TRANS 3 PL) [12;VERB-RMOD+RELCL]
18 i (IL ART DEF M PL) [17;VERB-OBJ]
19 risultati (RISULTATO NOUN COMMON M PL RISULTARE INTRANS) [18;DET+DEF-ARG]
19.10 t [13f] (DI PREP MONO) [19;NOUN-SUBJ]

Preliminary Results and Evaluation

- Overall we added 617 new fully annotated null elements. Then, we used this dataset as gold data to check the working of the algorithm: we ran the algorithm on the raw version of the dataset and matched the result with the gold augmented version of the dataset of the 500 sentences: we found 43 mistakes (that is 0.7% error rate), most of which (32, that is 0.5%) was a wrong antecedent for discourse bound little_pros. Results for the anaphora resolution – which are state of the art and average 75% accuracy – would require further improvements.