Annotazione di contenuti concettuali in un corpus italiano: I-CAB

Bernardo Magnini*, Amedeo Cappelli, Emanuele Pianta*, Manuela Speranza*, Valentina Bartalesi Lenzi**, Rachele Sprugnoli**, Lorenza Romano*, Christian Girardi*, Matteo Negri***

* ITC-irst, Povo (Trento) ** CELCT, Povo (Trento)

Abstract

In questo articolo presentiamo I-CAB (Italian Content Annotation Bank), un corpus di articoli in lingua italiana annotato semanticamente. L'attività di annotazione, realizzata in modo completamente manuale, prevede tre livelli: le espressioni temporali, le entità (cioè persone, organizzazioni, luoghi ed entità geo-politiche) e le relazioni tra entità (per esempio la relazione di affiliazione che collega una persona a un'organizzazione). I primi due livelli di annotazione sono completi, mentre il terzo livello è in fase di realizzazione. Avendo come scopo quello di fare di I-CAB un corpus di riferimento per diversi task di Estrazione automatica di Informazione, abbiamo seguito una politica di riutilizzo di linguaggi di annotazione già disponibili. In particolare, abbiamo adottato gli schemi di annotazione sviluppati per il task *ACE Entity Detection and Recognition* e per il task *Time Expression Recognition and Normalization*. Poiché le linee-guida di questi task sono state sviluppate originariamente per l'inglese, è stato necessario adattarle alle caratteristiche morfo-sintattiche dell'italiano; si è deciso inoltre di estenderle in modo tale da includere un insieme più ampio di entità, come ad esempio le congiunzioni.

1. Introduzione

Negli ultimi anni sono state avviate numerose iniziative volte alla realizzazione di corpora di riferimento per la valutazione di diversi task di elaborazione del linguaggio naturale. Recentemente, all'interno del Programma ACE[1] (*Automatic Content Extraction*), è stato sviluppato un insieme di schemi di annotazione per task nel campo dell'Estrazione di Informazione, prendendo in considerazione espressioni di tempo, menzioni di entità e relazioni tra entità. Sulla base delle risorse ottenute sono state organizzate con successo diverse campagne di valutazione (TERN 2004 e 2005, ACE 2000-2006). Queste esperienze hanno stimolato, nell'ambito della lingua inglese, sia la ricerca nel campo dell'Estrazione di Informazione sia lo sviluppo di risorse annotate dal punto di vista semantico, mentre poco è stato fatto per altre lingue, come l'italiano.

In questa prospettiva si colloca I-CAB (*Italian Content Annotation Bank*), un corpus italiano di articoli di giornale annotato semanticamente. In particolare, I-CAB contiene annotazioni relative a ESPRESSIONI_TEMPORALI, a entità PERSONA, ORGANIZZAZIONE, LUOGO e GEO-POLITICHE. Il corpus è accessibile on-line attraverso una specifica interfaccia utente[2].

I-CAB è stato annotato manualmente e vuole configurarsi come corpus di riferimento per diversi task di Estrazione di Informazione, tra cui il riconoscimento e la normalizzazione di espressioni relative al tempo, di entità e di relazioni tra entità. Seguendo una politica di riutilizzo di linguaggi di annotazione già disponibili, sono stati adottati i formalismi sviluppati all'interno del programma ACE; a causa delle notevoli differenze morfo-sintattiche tra l'inglese e l'italiano, tuttavia, si è rivelata necessaria una revisione delle linee-guida.

La creazione di I-CAB è parte del progetto triennale Ontotext[3] finanziato dalla Provincia Autonoma di Trento. Ontotext mira allo studio e allo sviluppo di tecnologie innovative per l'estrazione dell'informazione e della conoscenza richieste nell'ambito del Web Semantico. All'interno di questa nuova area di ricerca, che possiamo indicare come Estrazione di Conoscenza basata su Ontologie, Ontotext concentra la propria attenzione su tre obiettivi principali: (i) annotare documenti con informazione semantica e relazionale, (ii) facilitare la interoperabilità da tale informazione, (iii) aggiornare ed estendere le ontologie usate per le annotazioni del Web Semantico. Lo scenario concreto in cui gli algoritmi saranno testati mediante diversi esperimenti su larga scala, è rappresentato dall'acquisizione automatica di informazione da articoli di giornale.

L'articolo è strutturato come segue: nella Sezione 2 presentiamo gli standard ACE; nella Sezione 3 forniamo una descrizione del corpus, degli strumenti di annotazione e dei formati; nelle Sezioni 4 e 5 descriviamo rispettivamente l'annotazione delle espressioni di tempo e delle entità; nella Sezione 6 presentiamo i dati relativi all'accordo tra gli annotatori; la Sezione 7, infine, è dedicata alle conclusioni.

2. Linguaggi di annotazione

Il Programma ACE utilizza un linguaggio di annotazione flessibile che consente di identificare il contenuto informativo dei testi e di annotarli con informazioni di tipo sintattico. Lo scopo di ACE è quello di sviluppare tecnologie per l'Estrazione di Informazione al fine di supportare il trattamento automatico di dati linguistici. In particolare gli annotatori del Programma ACE lavorano su testi in inglese, cinese e arabo; per ogni lingua producono una sezione di *training* e una di *test*, rispettivamente per l'addestramento e per la valutazione dei sistemi.

Il Programma ACE è mosso dalle stesse motivazioni che hanno mosso precedentemente la *Message Understanding Conference* (MUC), ma ne rappresenta un'evoluzione in termini di complessità. In particolare, in MUC il task *Named Entity* considera solo tre tipi di entità (persone, organizzazioni e luoghi geografici) e prevede l'annotazione solo dei nomi propri e degli acronimi, mentre il task *Co-reference* prevede che vengano catturate e raggruppate tutte le espressioni che si riferiscono alla stessa entità.

ACE ha ampliato la lista delle tipologie di entità da considerare nell'annotazione aggiungendo, rispetto alle

[1] http://www.nist.gov/speech/tests/ace/
[2] http://ontotext.itc.it/webicab/
[3] http://tcc.itc.it/projects/ontotext/

precedenti, le entità di tipo geo-politico, le infrastrutture, le armi e i mezzi di trasporto. Il task relativo alla co-referenza è preservato, ma viene annotata una più vasta gamma di espressioni, tra le quali nomi comuni e pronomi. Infine, vengono definiti due livelli interconnessi di annotazione: il livello delle *entità*, che riguarda la rappresentazione di un oggetto nel mondo, e il livello delle *menzioni di entità*, che fornisce informazioni sulle realizzazioni testuali dello stesso. Per esempio, se George W. Bush è menzionato in due differenti frasi di un testo con l'espressione *Il presidente degli U.S.A.* e con il pronome *egli*, queste due espressioni vengono considerate come due menzioni della stessa entità.

Per raggiungere i nostri obiettivi, abbiamo deciso di adottare gli standard creati per il task *ACE Entity Detection and Recognition* e per il task *Time Expression Recognition and Normalization*, che consentono un arricchimento semantico e una normalizzazione di espressioni di tempo, entità e menzioni di entità.

Infine, per la nostra annotazione abbiamo deciso di tenere conto anche delle linee-guida sviluppate dal *Lingustic Data Consortium* (LCD) a supporto del programma ACE. Nel 2004 LDC ha distribuito il formato ALF (*ACE-LDC Format*), che differisce dal formato APF (*ACE Program Format*) per l'aggiunta di nuovi tipi di menzioni.

3. Descrizione del corpus e processo di annotazione

I-CAB è composto di 525 articoli del quotidiano locale "L'Adige"[4] distribuito nella Provincia di Trento. Gli articoli sono tratti da 4 differenti giornate (7-8 Settembre 2004 e 7-8 Ottobre 2004) e sono raggruppati in 5 categorie: Attualità (87 articoli), Cultura (72 articoli), Economia (54 articoli), Sport (123 articoli) e Trento (189 articoli).

I-CAB si divide in una sezione di *training* e in una sezione di *test*, contenenti rispettivamente 335 e 190 documenti. In totale il corpus è composto da circa 182.500 parole: 113.000 nella sezione di *training* (la lunghezza media di un articolo è di circa 339 parole) e 69.000 parole nella sezione di *test* (la lunghezza media di un articolo è di circa 363 parole).

Per la creazione di I-CAB abbiamo utilizzato il software di annotazione Callisto[5], sviluppato e distribuito gratuitamente dalla MITRE Corporation. Callisto supporta l'annotazione linguistica di testi scritti con caratteri codificati UTF-8 e US-ASCII; è scritto in Java così da consentirne una facile portabilità ed è stato progettato con un design modulare. Il software utilizza un'annotazione di tipo *stand-off* che si basa sulla separazione fisica tra il testo annotato e le annotazioni stesse e permette la realizzazione di moduli di annotazione indipendenti uno dall'altro: nel nostro caso, ad esempio, abbiamo utilizzato il modulo *TIMEX2* per l'annotazione delle espressioni temporali e il modulo *ACE Event* per l'annotazione delle entità.

In I-CAB l'annotazione manuale è unita ad un'annotazione automatica di livelli linguistici più bassi (tokenizzazione, lemmatizzazione, riconoscimento di unità polirematiche, assegnazione di parti del discorso) e tutti i livelli di annotazione sono salvati nel formato di annotazione Meaning (MEAF), un formato conforme alle indicazioni della *Text Encoding Initiative* (TEI) e sviluppato dall'ITC-irst nell'ambito del progetto Meaning (Bentivogli et al., 2003). MEAF è un formato basato su XML, in cui differenti livelli di annotazione sono contenuti in documenti separati o in sezioni diverse dello stesso documento e nel quale il livello base è costituito dai file in puro testo (*Hub*). Ciascun livello è collegato all'altro secondo una struttura gerarchica (cfr. Fig.1): il primo è il livello di annotazione ortografica, che rappresenta i *token*, implementato con puntatori alle posizioni dei caratteri ; il secondo è il livello di annotazione morfo-sintattica, con puntatori ai *token*; il terzo è il livello di annotazione di *multi-word*, con puntatori alle parole descritte al livello morfo-sintattico.

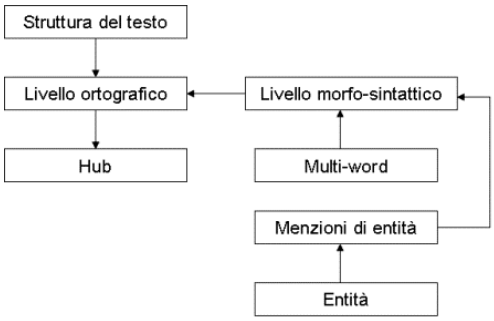

Figura 1: Livelli di annotazione in MEAF

In accordo con questa struttura gerarchica, le espressioni temporali e le menzioni di entità vengono rappresentate con puntatori al livello morfo-sintattico e le entità con puntatori alle menzioni di entità. Diversamente dalle espressioni temporali e dalle menzioni di entità in MEAF, tutte le annotazioni salvate con Callisto nel formato AIF puntano alle posizioni dei caratteri. Per questo motivo, nella trasformazione da AIF a MEAF i puntatori alle posizioni dei caratteri sono stati sostituiti con puntatori a oggetti morfo-sintattici.

4. Annotazione delle Espressioni Temporali

Per l'annotazione delle ESPRESSIONI_TEMPORALI (TEMPORAL_EXPRESSIONS, TE) abbiamo seguito lo standard di codifica TIMEX2 (Ferro et al. 2004), secondo il quale sono ritenute annotabili sia TE indicanti periodi di tempo (<*tre anni*>[6]) sia TE di tipo puntuale (<*6 maggio 2004*>, <*oggi*>). Queste ultime, a loro volta, possono essere assolute (<*6 maggio 2004*>) o relative, cioè anaforiche (<*oggi*>, <*tre giorni fa*>). Un'ulteriore categoria è quella delle TE *fuzzy* che si riferiscono in termini generici al passato, al presente o al futuro (<*recentemente*>, <*oggigiorno*>, <*nel futuro*>).

TIMEX2 consente di individuare le TE presenti all'interno di un testo e di normalizzarle, cioè di interpretarne il significato assegnando dei valori a un insieme predefinito di attributi. L'attributo principale è *VAL*, che si riferisce al valore delle TE (per esempio, *VAL* corrisponde a "2004-05-06" nel caso della data <*6 maggio*

[4] http://www.ladige.it/

[5] http://callisto.mitre.org/

[6] Negli esempi di questa sezione l'estensione delle TE è racchiusa in parentesi angolari.

2004> e a "P3Y" per il periodo <*tre anni*>). Nel caso di TE sotto-specificate, come <*per lungo tempo*>, non viene attribuito alcun valore. Gli altri attributi per la normalizzazione sono definiti come segue:

- MOD: cattura il significato espresso da alcuni modificatori temporali. Valori possibili sono APPROX (<*verso mezzanotte*>), MORE THAN (<*più di 3 ore*>) e START (<*i primi anni '70*>);
- ANCHOR VAL: contiene la forma normalizzata di una data (o di un orario) che funge da ancora temporale;
- ANCHOR DIR: cattura la direzione di una TE, come AFTER (*dopo*) e BEFORE (*prima*). Per esempio, prendendo il 6 Maggio 2004 come data di riferimento, la TE nella frase *sarò in vacanza per* <*due mesi*> è normalizzata come: VAL="P2M" ANCHOR_VAL= "2004-05-06" e ANCHOR DIR="AFTER" (dato che il periodo di due mesi è successivo alla data di riferimento);
- SET: identifica le TE relative a *set of times*, cioè espressioni che indicano il ripetersi di azioni a scadenza costante. Per esempio, <*ogni anno*> è annotata con SET="YES".

In totale sono state annotate circa 4.610 TE: 2.932 nella sezione di *training* e 1.678 nella sezione di *test* (cfr. Tab. 1). Si noti che in entrambe le sezioni del corpus le espressioni puntuali superano il 50% del totale. Per quanto riguarda la normalizzazione, ANCHOR_DIR e ANCHOR_VAL, che sono usati sempre in coppia, costituiscono gli attributi a cui più frequentemente viene assegnato un valore (23,7%). Sia a MOD che a SET è stato attribuito un valore solo nel 4% dei casi.

	Training		Test		TOT.	
Punti	1.851	63,1%	982	58,5%	2.833	61,5%
Periodi	507	17,3%	382	22,8%	889	19,3%
Fuzzy	428	14,6%	192	11,4%	620	13,4%
No VAL	146	5,0%	122	7,3%	268	5,8%
TOT.	8.761		4.964		13.725	

Tabella 1: Distribuzione delle TE per categoria.

L'adattamento dello standard TIMEX2 all'annotazione di testi italiani ha richiesto alcune modifiche (Lavelli et al., 2005). In particolare, come conseguenza delle caratteristiche specifiche dell'italiano, che ha una morfologia più ricca rispetto all'inglese, abbiamo introdotto alcuni cambiamenti riguardanti l'estensione delle TE. Secondo le linee-guida ACE-LDC, gli articoli determinativi e indeterminativi sono considerati parte delle realizzazioni testuali delle TE, mentre le preposizioni non lo sono (in inglese abbiamo *at* <*the end of March*>*/alla fine di Marzo*). Questa regola però non è adeguata al trattamento delle preposizioni articolate, quindi abbiamo deciso di includerle nell'estensione dell'annotazione (<*alla fine di marzo*>); lo stesso criterio è stato adottato per l'annotazione delle entità (si veda Sezione 5).

5. Annotazione delle entità

Il task di annotazione delle entità ACE richiede che queste siano individuate all'interno del testo e che il loro significato sia disambiguato, assegnando valori agli attributi definiti nelle linee-guida. In particolare, vengono etichettate tutte le menzioni di entità che compaiono in un documento e vengono raggruppate quelle che co-referiscono (cioè quelle che si riferiscono alla stessa entità).

Nelle linee-guida ACE-LDC sono previsti sette tipi di entità, mentre in I-CAB abbiamo ristretto tale insieme a cinque. Abbiamo infatti deciso di non annotare FACILITY (infrastrutture), VEHICLE (veicoli) e WEAPON (armi) e di aggiungere MIXED (MIX), che comprende gruppi di entità non omogenei per i quali è impossibile scegliere un singolo tipo (ad esempio in <*lui e l'azienda*> abbiamo un gruppo formato da un'entità PERSONA e un'entità ORGANIZZAZIONE).

I dati complessivi relativi alle entità e alle menzioni annotate in I-CAB sono presentati in Tab. 2. In totale, sono state identificate 13.725 entità (in media, circa 26 per documento) e 28.519 menzioni (54 per documento). La distribuzione tra I-CAB *training* and I-CAB *test* riflette le stesse proporzioni: 8.761 entità e 18.141 menzioni nel primo, 4.964 entità e 10.378 menzioni nel secondo. I dati relativi ai tipi di entità mostrano una netta prevalenza delle entità PERSONA (PER), che superano il 50% del totale; le entità ORGANIZZAZIONE (ORG), GEO-POLITICHE (GPE) e LUOGO (LOC) rappresentano rispettivamente il 24%, il 18% e il 5% del totale. Per quanto riguarda la co-referenza, si può facilmente osservare come un'entità venga menzionata, in media, più di due volte. Si noti però che la co-referenza è piuttosto alta per le PERSONE, mentre abbiamo poco più di una menzione per entità nel caso delle entità LOC e MIX.

		Training	Test	TOT.
PER	Entità	4.531	2.679	7.210
	Menzioni	10.136	6.174	16.310
ORG	Entità	2.235	1.047	3.282
	Menzioni	4.336	1.964	6.300
LOC	Entità	398	213	611
	Menzioni	575	310	885
GPE	Entità	1.466	955	2.421
	Menzioni	2.928	1.821	4.749
MIX	Entità	131	70	201
	Menzioni	166	109	275
TOT	Entità	8.761	4.964	13.725
	Menzioni	18.141	10.378	28.519

Tabella 2: Dati sulle annotazioni.

A ciascuna entità viene assegnata una delle quattro classi semantiche individuate nelle linee-guida ACE-LDC a seconda del tipo di riferimento che esiste tra essa e il corrispettivo oggetto nel mondo:

- *Specific referential* (SPC): l'entità si riferisce a un oggetto specifico o un gruppo di oggetti specifici (<*L'*[*avvocato*] *di Giovanni*> *ha vinto la causa*)[7].
- *Generic referential* (GEN): si riferisce a una categoria di entità e non a un particolare oggetto (o gruppo di oggetti) nel mondo (<*Gli* [*avvocati*]> *non lavorano gratis*).

[7] Negli esempi, l'estensione delle menzioni è racchiusa in parentesi angolari e la testa è tra parentesi quadre. Quest'ultima non è evidenziata nei casi in cui coincide con l'intera menzione.

- *Under-specified referential* (USP): nel caso di riferimenti non generici, non specifici. In particolare, questa categoria include quantificazioni non precise (<*molti*>/<*alcuni*>/ <*100.000* [*persone*]>), pronomi impersonali (<*si*> *dice che*), sintagmi nominali contenuti all'interno di periodi ipotetici, proposizioni interrogative e condizionali (*Chissà <chi> è arrivato!*), etc.
- *Negative* (NEG): si riferisce a insiemi vuoti (<*Nessun* [*avvocato*]>).

Come presentato nella Tab. 3, la maggior parte delle entità (quasi il 90% del totale) appartiene alla classe semantica SPC, mentre le entità NEG sono piuttosto rare.

	Training		Test		TOT.	
SPC	7.581	86,5%	4.416	89,0%	11.997	87,4%
GEN	570	6,5%	258	5,2 %	828	6,0%
USP	579	6,6%	280	5,6%	859	6,3%
NEG	31	0,4%	10	0,2%	41	0,3%
TOT.	8.761		4.964		13.725	

Tabella 3: Distribuzione delle entità per classe semantica.

L'estensione delle menzioni di entità coincide con l'intero sintagma nominale usato per riferirsi all'entità stessa, in modo da includere i modificatori (<*una grande* [*famiglia*]>), i sintagmi preposizionali (<*il* [*Presidente*] *della Repubblica*>) e le proposizioni relative (<*la* [*ragazza*] *che lavora in giardino*>). All'interno di ciascuna menzione viene infine evidenziata la testa sintattica.

Le menzioni sono classificate sulla base delle loro caratteristiche sintattiche. In particolare, ACE propone dieci tipi sintattici:

- NAM: nomi propri (<*Totti*>, <*UE*>);
- NOM: costrutti nominali (<*i* [*bambini*] *buoni*>, <*l'*[*azienda*]>);
- PRE: pre-modificatori (*il* <*brasiliano*> *Ronaldo*);
- BAR: costrutti nominali non introdotti da pre-modificatori ed articoli (<[*poliziotti*] *di quartiere*>);
- HLS: costrutti nei quali la testa nominale non è esplicitamente espressa (<*Il più* [*forte*] *di tutti*>);
- WHQ: pronomi interrogativi e relativi (<*Chi*> *è lì?*);
- PRO: pronomi, personali (<*tu*>) e indefiniti (<*qualcuno*>);
- PTV: partitivi (<[*alcune*] *delle scuole*>);
- APP: costruzioni appositive (<*il Po, fiume italiano*>);
- ARC: costruzioni appositive con una relativa (WHQ) adiacente (<*L'ex direttore, Rossi, che faceva spesso tardi*>).

Per l'annotazione di testi in italiano, è stato necessario aggiungere alcuni nuovi tipi di menzioni (Pianta et al., 2006). In particolare, sono state create due categorie specifiche, ENCLIT e PROCLIT, per annotare gli enclitici e i proclitici per i quali le convenzioni ortografiche dell'italiano impongono che elemento atono e parola precedente o seguente formino un'unica unità grafica (<*veder*[*lo*]>/<[*gli*]*elo*> *dico sempre*).

Per quanto riguarda i modificatori, inoltre, è stata aggiunta la categoria POST (oltre alla già esistente PRE) per annotare i modificatori che in italiano, a differenza dell'inglese, si trovano quasi sempre dopo il sostantivo (*un tessuto di fabbricazione <francese>*).

In ACE vengono annotate solo le congiunzioni di entità che hanno dei modificatori comuni (<*gli antichi filosofi e pensatori*>). Nel processo di estensione delle linee-guida, abbiamo deciso di annotare tutte le congiunzioni di entità, creando un nuovo tipo di menzione (CONJ) (<*la madre e il figlio*>)[8]. Ciò consente di annotare in modo sistematico la co-referenza con espressioni anaforiche, come <*loro*> e <*le due* [*persone*]>, che potrebbero seguire nel testo.

La Tab. 4 mette in evidenza, da un lato, la netta prevalenza di menzioni di tipo NAM e NOM (questi due tipi sintattici includono infatti più del 65% delle menzioni totali) e, dall'altro, il fatto che le due sezioni del corpus mostrano una distribuzione molto simile tra loro.

	Training		Test		TOT.	
NAM	7.423	40,9%	4.003	38,6%	11.426	40,0%
NOM	4.595	25,3%	2.652	25,6%	7.247	25,4%
PRE	95	0,5%	46	0,4%	141	0,5%
BAR	1.040	5,7%	687	6,6%	1.727	6,1%
POST	518	2,9%	350	3,4%	868	3,0%
HLS	222	1,2%	141	1,4%	363	1,3%
WHQ	720	4,0%	471	4,5%	1.191	4,2%
PRO	1.607	8,9%	900	8,7%	2.507	8,8%
PTV	55	0,3%	58	0,6%	113	0,4%
APP	817	4,5%	452	4,3%	1.269	4,4%
ARC	70	0,4%	62	0,6%	132	0,5%
ENCL.	100	0,6%	71	0,7%	171	0,6%
PROC.	1	0,0%	2	0,0%	3	0,0%
CONJ	878	4,8%	483	4,6%	1.361	4,8%
TOT.	18.141		10.378		28.519	

Tabella 4: Distribuzione delle entità per tipo sintattico.

5.1. Entità Persona

Secondo gli standard di ACE, ogni singola persona o insieme di persone menzionati in un documento si riferisce a un'entità di tipo PERSONA (PER).

Le entità PERSONA sono classificate secondo i seguenti sottotipi: (i) INDIVIDUAL: entità di tipo PER che si riferiscono a un singolo individuo (<*George W. Bush*>); (ii) GROUP: entità PER che si riferiscono a più di una persona (<*quei due* [*signori*]>, <*la tua* [*famiglia*]>, <*Alice e suo figlio*>); (iii) INDEFINITE: un'entità è classificata come INDEFINITE quando dal contesto non è possibile giudicare se si tratta di una più persone (*Mi chiedo <chi> arriverà*).

La Tab. 5 mostra una distribuzione bilanciata tra i due sottotipi più frequenti (47% di entità PER-Individual e 45% di entità PER-Group), con una piccola percentuale (l'8%) di entità PER-Indefinite.

[8] Le apposizioni e le congiunzioni sono considerate menzioni complesse. Secondo le linee-guida ACE-LDC non è necessario annotare la testa sintattica di questo tipo di menzioni. Tuttavia Callisto richiede che ogni menzione abbia la testa, quindi abbiamo deciso di fare coincidere la testa con l'intera estensione.

	Training		Test		TOT.	
Indiv.	2.073	45,7%	1.241	46,3%	3.314	46%
Group	2.056	45,4%	1.257	46,9%	3.313	46%
Indef.	402	8,9%	181	6,8%	583	8%
TOT.	4.531		2.679		7.210	

Tabella 5: Entità PER divise per sottotipo.

5.2. Entità Organizzazione

Come indicato dalle linee-guida ACE-LDC, le entità ORGANIZZAZIONE (ORG) sono state divise in 10 sottotipi: GOVERNMENT (<*I* [*Carabinieri*]>), COMMERCIAL (<*La* [*Microsoft*]>), EDUCATIONAL *(<l'*[*Università di Pisa*]>), MEDIA (<*National Geographic*>), RELIGIOUS (<*La* [*Chiesa Valdese*]>), SPORTS (<*La* [*Juventus*]>), MEDICAL-SCIENCE (<*Il* [*laboratorio*] *di analisi*>), NON-GOVERNMENTAL (<*La* [*Croce Rossa*]>) e ENTERTAINMENT. (<*La* [*compagnia*] *teatrale*>). È stato inoltre aggiunto un nuovo sottotipo, MIXED, per annotare entità costituite da gruppi di ORG con sottotipo diverso, come per esempio <*L'Università di Trento e la Microsoft*> (le università sono infatti ORG-EDUCATIONAL, mentre le aziende sono ORG-COMMERCIAL).

Tra questi sottotipi, i più frequenti sono SPORT, COMMERCIAL, NON-GOVERNMENTAL e GOVERNMENT che rappresentano in totale l'81% delle ORG (vedi Tab. 6).

	Training		Test		TOT.	
Govern.	331	14,8%	170	16,2%	501	15,3%
Comm.	478	21,4%	201	19,2%	679	20,7%
Educat.	159	7,1%	94	9,0%	253	7,7%
Media	47	2,1%	21	2,0%	68	2,0%
Relig.	30	1,3%	18	1,7%	48	1,5%
Sports	590	26,4%	360	34,4%	950	28,9%
Med.	54	2,4%	26	2,5%	80	2,4%
Non-Gov.	406	18,2%	121	11,6%	527	16,1%
Entert.	106	4,8%	18	1,7%	124	3,8%
Mixed	34	1,5%	18	1,7%	52	1,6%
TOT.	2235		1047		3282	

Tabella 6: Entità ORG divise per sottotipo.

Le menzioni di organizzazioni non italiane sono state annotate come nomi propri (NAM) quando vengono tradotte letteralmente; sono state annotate come nomi comuni (NOM) se si tratta di trasposizioni culturali del nome originale. Ad esempio, <*Dipartimento di Stato Americano*> è annotato come nome proprio in quanto è la traduzione letterale dell'inglese *U.S. Department of State*. Al contrario, <[*Polizia*] *francese*> è considerato nome comune traducendo il francese *Gendarmerie*.

5.3. Entità Luogo

Secondo la definizione ACE, le entità LUOGO (LOC) sono aree individuate su basi geografiche o astronomiche che non costituiscono soggetti politici. Sono divise in 8 sottotipi: ADDRESS (<*Via Nazionale 12*>), BOUNDARY (<*Il* [*confine*] *siriano*>), CELESTIAL (<*Il* [*sole*]>), WATER-BODY (<*Il* [*Po*]>, <*Il* [*mare*]>), LAND-REGION-NATURAL (<*Il* [*Monte Bianco*]>), REGION-INTERNATIONAL (<*L'*[*Africa*] *meridionale*>), REGION-GENERAL (<*una* [*parte*] *della città*>, <*Il* [*nord-est*]>) e MIXED (da noi aggiunto per annotare insiemi di ORG di sottotipo diverso, come <*il sole e il mare*>; cfr. Sezione 5.2).

Come mostrato nella Tab. 7, risaltano i dati riguardanti i sottotipi ADDRESS (in I-CAB compaiono infatti molti indirizzi di mostre e manifestazioni, ma anche vie in cui avvengono fatti di cronaca), LAND-REGION-GENERAL (soprattutto fiumi, montagne ed altipiani del Trentino) e REGION-GENERAL (vasta categoria che comprende i quartieri delle città e aree geografiche entro i confini nazionali).

	Training		Test		TOT.	
Address	95	23,9%	40	18,8%	135	22,1%
Boundary	4	1,0%	6	2,8%	10	1,6%
Celestial	24	6,0%	19	8,9%	43	7,0%
Water-B.	31	7,8%	14	6,6%	45	7,4%
Land-R-N	114	28,6%	50	27,2%	172	28,2%
Region-I.	8	2,0%	14	6,6%	22	3,6%
Region-G.	120	30,2%	61	28,6%	181	29,6%
Mixed	2	0,5%	1	0,5%	3	0,5%
TOT.	398		213		611	

Tabella 7: Entità LOC divise per sottotipo.

5.4. Entità Geo-Politiche

Le linee-guida ACE-LDC definiscono le entità Geo-Politiche come regioni geografiche caratterizzate dalla presenza di gruppi sociali e/o politici. Un'ulteriore classificazione prevede l'individuazione di 7 sottotipi: CONTINENT (<*Asia*>), NATION (<*Italia*>, <*Stati Uniti*>), STATE-OR-PROVINCE (<*Florida*>), COUNTY-OR-DISTRICT (<*Canton Ticino*>), POPULATION-CENTER (<*Trento*>), GPE-CLUSTER (gruppi di GPE che agiscono come entità geo-politiche, ad esempio <*Unione Europea*>), SPECIAL (GPE a cui è difficile applicare un'etichetta convenzionale, come <*La* [*Palestina*]>) e MIXED (<*Gli Stati Uniti e l'UE*>, cfr. Sezioni 5.2). Per adattare i sottotipi delle line-guida ACE-LDC alla nostra realtà nazionale, le regioni e le province sono state inserite nella categoria STATE-OR-PROVINCE, i comuni in COUNTY-OR-DISTRICT e le circoscrizioni in POPULATION-CENTER.

La Tab. 8 mostra la netta prevalenza del sottotipo POPULATION-CENTER (tutti i centri abitati, dalle grandi città alle frazioni comunali), seguito da NATION (molto frequente negli articoli di attualità nazionale ed internazionale) e STATE-OR-PROVINCE (in particolare nelle numerose notizie riguardanti la Provincia Autonoma di Trento).

	Training		Test		TOT.	
Continent	21	1,4%	10	1,0%	31	1,3%
Nation	298	20,3%	238	24,9%	536	22,1%
State-or-Pr.	254	17,3%	153	16,0%	407	16,8%
County-or-D.	71	4,9%	44	4,6%	115	4,8%
Pop-center	770	52,5%	477	50,0%	1.247	51,5%
GPE-Cluster	24	1,6%	12	1,3%	36	1,5%
Special	9	0,6%	8	0,8%	17	0,7%
Mixed	19	1,4%	13	1,4%	32	1,3%
TOT.	1.466		955		2.421	

Tabella 8: Entità GPE divise per sottotipo.

Rispetto alle entità descritte in precedenza, le GPE richiedono anche l'annotazione del *ruolo*, ovvero l'aspetto della GPE a cui ciascuna menzione fa riferimento: la localizzazione fisica della GPE (GPE.LOC, *Il G8 si riunisce in* <*Francia*>), la sua popolazione (GPE.PER, <*I* [*francesi*]> *attendono con ansia le elezioni*), il suo governo (GPE.ORG, <*La* [*Francia*]> *firmerà presto un accordo*) o un insieme non distinguibile di questi aspetti (GPE.GPE, <*La* [*Francia*]> *produce un ottimo vino*). Come indicato in Tab. 9, quest'ultimo ruolo si è rivelato essere il più diffuso nel corpus (più del 50%) mentre rari sono apparsi i riferimenti all'intera popolazione di una GPE (circa il 2%).

	Training		Test		TOT.	
GPE.LOC	983	33,6%	609	33,4%	1.592	33,5%
GPE.PER	66	2,3%	36	2,0%	102	2,2%
GPE.ORG	440	15,0%	141	7,7%	581	12,2%
GPE.GPE	1.439	49,1%	1.035	56,9%	2474	52,1%
TOT.	2.928		1.821		4.749	

Tabella 9: Entità GPE divise per ruolo.

5.5. Annotazione della metonimia

Le linee-guida ACE-LDC prevedono l'annotazione della *Nickname Metonymy*, un particolare tipo di metonimia che occorre quando il nome proprio di una GPE viene usato per riferirsi ad un'altra entità (di tipo ORG o GPE). Gli esempi più comuni di *Nickname Metonymy* sono i seguenti:

- la capitale di una nazione viene usata per riferirsi al governo della nazione stessa (<*Parigi*> *ha firmato l'accordo*);
- il nome di una GPE viene usata per indicare una squadra sportiva (<*La* [*Russia*]> *ha conquistato la medaglia d'oro*).

Negli esempi sopra citati, le due menzioni <*Parigi*> (che si riferisce ad un'entità GPE con sottotipo NATION e Ruolo ORG) e <*La* [*Russia*]> (che si riferisce ad un'entità ORG con sottotipo SPORT) vengono marcate come metonimiche. Il fenomeno della *Nickname Metonymy* occorre raramente nel corpus: sono state, infatti, riconosciute come metonimiche soltanto 360 menzioni nel training e 185 nel test.

6. Accordo tra annotatori

Al fine di valutare il livello di accordo tra gli annotatori, sono stati realizzati cinque test separati: per le TE e per le entità PER, ORG, LOC e GPE. Per ciascun test, dieci articoli estratti a caso da I-CAB sono stati annotati in maniera indipendente da due persone. I dati riportati nelle Sezioni 6.1 e 6.2 sono basati sul confronto tra le due versioni.

6.1. Espressioni temporali

La misura più comunemente usata è la *kappa statistic* (Cohen 1960), che misura l'accordo tra annotatori sulla base di giudizi di categoria, tenendo in considerazione anche la probabilità che tale accordo sia ottenuto per caso.

D'altra parte, volendo valutare l'accordo degli annotatori nell'individuazione dell'estensione delle TE, dovremmo considerare che teoricamente essi potrebbero scegliere di annotare come TE una qualunque sequenza di *token* adiacenti all'interno di una frase[9] e dunque si renderebbe necessario considerare ogni possibile sequenza come candidata all'annotazione. Se applicassimo la misura *kappa* per valutare l'accordo degli annotatori nell'individuare le sequenze di *token* che corrispondono a una TE, otterremmo dei risultati estremamente bassi, in quanto l'annotazione diventerebbe un problema di categorizzazione binaria con una distribuzione altamente asimmetrica (solo un numero ristretto di sequenze candidate sarebbero davvero espressioni temporali).

Per questa ragione, abbiamo usato la *kappa statistic* solo per calcolare l'accordo tra annotatori nel determinare se un *token* fa parte o no di una TE. Questa misura, tuttavia, non valuta l'accordo nell'assegnare una sequenza di *token* alla stessa espressione, perciò abbiamo confrontato le due versioni annotate usando anche il coefficiente di Dice. Questo è calcolato come da formula [1], dove *C* è il numero delle annotazioni comuni, mentre *A* e *B* sono rispettivamente il numero delle annotazioni fornite dal primo e dal secondo annotatore.

[1] $\text{Dice}=2C/(A+B)$ [10]

Su un corpus di dieci articoli di giornale (per un totale di 5.200 parole) abbiamo ottenuto k=0,958. Il coefficiente di Dice è risultato essere 0,955 per il task di individuazione delle TE e 0,931 per quello di identificazione della loro estensione.

L'accordo nella normalizzazione è stato misurato sulle TE uniformemente riconosciute. Riportiamo di seguito le percentuali dei casi in cui gli annotatori si sono trovati in accordo nell'assegnare o meno il valore a ciascun attributo: 92,2% per VAL, 92,2% per ANCHOR_VAL, 90,3% per ANCHOR_DIR, al 99,3% per MOD e 98,7 per SET. In merito agli attributi che ammettono un numero limitato di valori, abbiamo ottenuto *kappa statistic* pari a 0,749 per ANCHOR_DIR, 0,886 per MOD e 0,744 per SET.

6.2. Entità

Per quanto riguarda l'accordo nell'annotazione delle entità, abbiamo adottato i criteri utilizzati nel software di valutazione distribuito per la campagna di valutazione ACE 2005:

- un'entità è riconosciuta da entrambi gli annotatori se ciascuno annota almeno una menzione di tale entità;
- una menzione è riconosciuta da entrambi gli annotatori se la porzione di sovrapposizione reciproca delle teste delle menzioni è almeno del 30%;
- la massima differenza ammessa per avere accordo sull'estensione di una menzione è di quattro caratteri.

Di conseguenza, se un annotatore individua <*il grande* [*Savani*]> come una menzione mentre l'altro erroneamente limita l'estensione a <*Savani*>, abbiamo accordo nel riconoscimento della menzione (entrambi gli

[9] Questa considerazione vale anche per le entità.

[10] Si noti che il coefficiente di Dice ha lo stesso valore della misura F1 calcolata considerando uno qualunque dei due annotatori come punto di riferimento.

annotatori riconoscono infatti [*Savani*] come testa sintattica), ma non sull'estensione.

La definizione della *kappa statistic* proposta nella Sezione 6.1 (cioè in base all'appartenenza o meno di una parola a una ESPRESSIONE_TEMPORALE) non tiene conto delle annotazioni annidate. Poiché questo fenomeno è estremamente frequente nel caso delle entità, abbiamo deciso di calcolare invece il coefficiente di Dice e di limitare l'uso della *kappa statistic* all'assegnazione degli attributi.

L'accordo tra annotatori è stato valutato sulla base di un corpus costituito da 4.657 parole per le PER, 3.405 parole per le ORG, 4.868 parole per le LOC e 4.741 per le GPE.

Di seguito sono riportati i risultati:

- il coefficiente di Dice per l'individuazione delle entità è 0,906 per le PER, 0,857 per le ORG, 0,957 per le LOC e 1 per le GPE;
- limitatamente alle entità individuate da entrambi gli annotatori, il coefficiente di Dice per l'individuazione delle menzioni è 0,951 per le PER, 0,845 per le ORG, 0,938 per le LOC e 0,980 per le GPE;
- limitatamente alle entità individuate da entrambi gli annotatori, la *kappa statistic* per l'assegnazione dei sottotipi è 0,937 per le PER, 0,970 per le ORG, 1 per le LOC e le GPE;
- limitatamente alle entità individuate da entrambi gli annotatori, la *kappa statistic* per l'assegnazione delle classi è 0,734 per le PER, 1 per le ORG e le GPE[11];
- limitatamente alle menzioni individuate da entrambi gli annotatori, abbiamo un disaccordo nell'annotazione dell'estensione di 3,8% per le PER e le ORG mentre è dello 0% (cioè accordo totale) per le LOC e le GPE;
- la *kappa statistic* per l'assegnamento del ruolo delle GPE (sempre limitatamente alle menzioni identificate da entrambi gli annotatori) è 0,965.

7. Conclusioni e lavoro futuro

I-CAB è accessibile direttamente dal sito web di Ontotext attraverso un'apposita interfaccia che permette all'utente di effettuare ricerche all'interno del corpus secondo diverse modalità di ricerca (per esempio, ricerca per documento o per parola) e di visualizzare tutte le annotazioni o di selezionare soltanto specifici tipi o combinazioni di tipi.

Nel prossimo futuro, avvieremo l'annotazione delle RELAZIONI tra entità e degli EVENTI come definiti nei task *Relation Detection and Characterization* (RDC) e *Event Detection and Characterization* (EDC). Il corpus sarà distribuito gratuitamente per scopi di ricerca[12].

[11] Non abbiamo riportato il valore della *kappa statistic* per le LOC perché il dato non è significativo L'accordo tra gli annotatori è infatti risultato accettabile (su un totale di 11 entità, 10 sono state annotate SPC da entrambi e una è stata annotata in maniera diversa) ma il calcolo della k dà zero a causa del forte sbilanciamento nella distribuzione delle classi, cioè della netta prevalenza della classe SPC, che fa aumentare moltissimo la probabilità che l'accordo tra gli annotatori sia ottenuto per caso.

[12] Ringraziamenti: questo lavoro è stato parzialmente supportato dal progetto ONTOTEXT, finanziato dalla Provincia Autonoma di Trento nell'ambito del programma di ricerca FUP-2004.

8. Riferimenti

Bentivogli, L., Girardi, C., Pianta, E. (2003). The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster: UCREL, pp. 103-112.

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*. New York: Sage Publications, 20, pp. 37-46.

Di Eugenio, B., Glass, M. (2004). The kappa statistic: A second look. In *Computational Linguistics*. Boston: MIT Press, 30(1), pp. 95-101.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE. http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf

Fleischman, M. (2001). Automated Subcategorization of Named Entities. In *Proceedings of the 39th Annual Meeting of the ACL, Student Research Workshop*. Toulouse: CNRS, pp. 25-30.

Fleischman, M., Hovy, E. (2002). Fine Grained Classification of Named Entities. In *Proceedings of COLING 2002*. Taipei: Morgan Kauffmann, pp. 1-7.

Hearst, M. (1998). Automated Discovery of WordNet Relations. In *WordNet: An Electronic Lexical Database*. Boston: MIT Press, pp. 131-151.

Lavelli, A., Magnini, B., Negri, M., Pianta, E., Speranza, M., Sprugnoli, R. (2005). Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 1.0). Technical Report T-0505-12. Trento: ITC-irst.

Linguistic Data Consortium (2004). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1 2005.05.23. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf

Pianta, E., Bentivogli, L., Girardi, C., Magnini, B. (2006). Representing and Accessing Multilevel Linguistic Annotation using the MEANING Format. In *Proceedings of NLPXML-2006 Multi-dimensional Markup in Natural Language Processing* (Workshop EACL 2006). Trento: pp. 77-80.

Pianta E., Speranza M., Magnini B., Bartalesi Lenzi V., Sprugnoli R. (2006). Italian Content Annotation Bank (I-CAB): Person Entities (V. 1.1). Technical Report. Trento: ITC-irst.

Siegel, S., Castellan, N. J. (1988). *Non parametric statistics for the behavioral sciences*. New York: McGraw Hill.