
The Italian Content Annotation Bank

I-CAB

Valentina Bartalesi Lenzi
Manuela Speranza
Rachele Sprugnoli

bartalesi@celct.it
manspera@itc.it
sprugnoli@celct.it

Outline

- The ONTOTEXT project
- Description of the I-CAB corpus
- Temporal expression annotation
- Entity annotation
- Entity mention annotation
- Adaptations to Italian
- Inter-annotator agreement
- The I-CAB browser
- Conclusions

Outline

- **The ONTOTEXT project**
- **Description of the I-CAB corpus**
- **Temporal expression annotation**
- Entity annotation
- Entity mention annotation
- Adaptations to Italian
- Inter-annotator agreement
- The I-CAB browser
- Conclusions

Our Team



*Center for the Evaluation of Language and
Communication Technologies*

www.celct.it



FBK-irst

Center for Scientific and Technological Research

www.itc.it/irst/

The ONTOTEXT Project

<http://ontotext.itc.it/>

Three key research aspects:

1) Automatic text annotation

Creation of a benchmark → I-CAB

Development of automatic annotation systems

2) Knowledge Extraction

3) Ontology learning and population

Updating and extending the ontologies used for Semantic Web annotation

Evaluation scenario: automatic acquisition of information from local newspaper articles in order to be allowed to process the data and to make them available as a Web service through a dedicated portal

Annotations

Benchmark intended as a reference work for various automatic Information Extraction tasks, providing manual annotations of:

- temporal expressions (TEs)
- entities:
 - ◆ persons (PER)
 - ◆ organizations (ORG)
 - ◆ geo-political entities (GPE)
 - ◆ locations (LOC)
- entity mentions
- relations among entities (*work in progress*)

Description of the I-CAB corpus

- 525 news stories from the Italian local newspaper “L’Adige”
- 4 days
 - 7–8 September 2004
 - 7–8 October 2004
- 5 categories
 - News Stories
 - Cultural News
 - Economic News
 - Sports News
 - Local News
- Two sections: **training** (335 news stories) and **test** (190 news stories)

Number of words = 182.500
Average number of words per file = 348

Formalisms adopted

- ◆ ACE (*Automatic Content Extraction*) Program
 - The objective is to develop automatic content extraction technology to support automatic processing of human language in text form
 - <http://www.nist.gov/speech/tests/ace>
 - Guidelines provided by the *Linguistic Data Consortium* (LDC)
 - <http://projects ldc.upenn.edu/ace/annotation/2005Tasks.html>

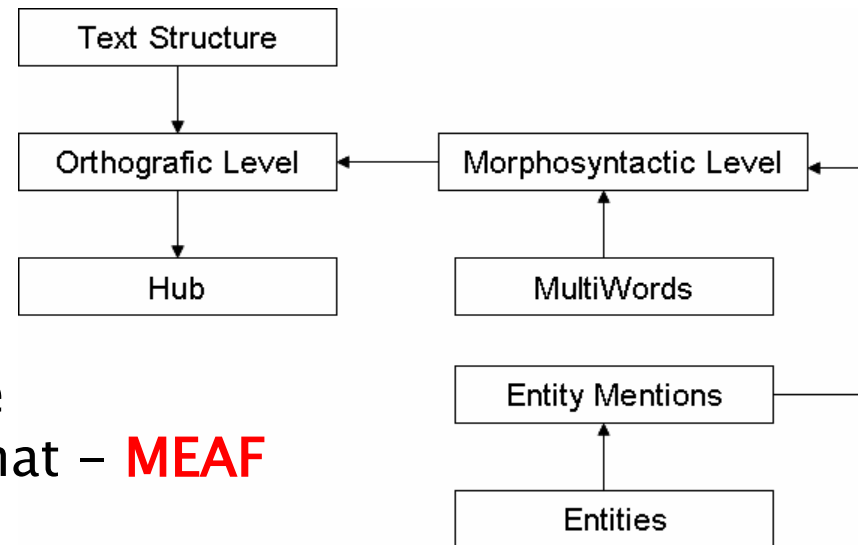
We adopted the annotation schemes developed for:

- *Time Expressions Recognition and Normalization Task (TERN)*
- *Entity Detection and Recognition Task*
- *Relation Detection and Recognition Task*

Annotation tool and formats

CALLISTO: <http://callisto.mitre.org/>

- It accepts files encoded as UTF-8 and US-ASCII
- It is written in Java
- It utilizes standoff-annotation
- Several *tasks* are available (E.g. **TIMEX2** for temporal expressions and **ACE Event Task** for entities and relations)




- I-CAB is distributed in the Meaning Annotation Format – **MEAF**

Temporal Expressions

- Annotation scheme: TIMEX2 (<http://timex2.mitre.org/>)

- Subtasks: Recognition and Normalization


finding the TEs within a text
(detection) and determining
their extension (bracketing)


interpreting the meaning of TEs

- Examples of markable temporal expressions:

- points (*31/05/2007, alle 23:00, oggi*)
- periods (*3 mesi, due settimane*)
- not gregorian dates (*stagione calcistica, anno accademico*)



are markable but they are NOT to be normalized

Temporal Expressions: Normalization Attributes

- ◆ VAL: value of a TE following the ISO-8601 standard
<15 maggio 2006> → VAL="2006-05-15"
- ◆ MOD: captures temporal modifiers
<verso mezzanotte> → MOD="APPROX"
- ◆ SET: identifies expressions denoting sets of time
<ogni anno> → SET="YES"
- ◆ ANCHOR_VAL: a normalized form of an anchoring date
- ◆ ANCHOR_DIR: captures the direction of a TE –
appears in combination with ANCHOR_VAL
sarò in vacanza per <due mesi > → VAL="P2M"
ANCHOR_VAL= "2007-05-31"
ANCHOR_DIR="AFTER"

Temporal Expressions: Example

KUALA LUMPUR – 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie dalla quale aveva divorziato nel 1958 .

Il matrimonio più breve è durato due giorni e il più lungo 20 anni con l'ultima moglie, una thailandese.

Temporal Expressions: Recognition

KUALA LUMPUR – 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie dalla quale aveva divorziato nel 1958.

Il matrimonio più breve è durato due giorni e il più lungo 20 anni con l'ultima moglie, una thailandese.

Temporal Expressions: Normalization (1)

KUALA LUMPUR – 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie dalla quale aveva divorziato

`<TIMEX2 val="1958">nel 1958</TIMEX2>` .

Il matrimonio più breve è durato due giorni e il più lungo 20 anni con l'ultima moglie, una thailandese.

Temporal Expressions: Normalization (2)

KUALA LUMPUR – 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie dalla quale aveva divorziato nel 1958.

Il matrimonio più breve è durato

`<TIMEX2 val="P2D">due giorni</TIMEX2>`

e il più lungo

`<TIMEX2 val="P20Y">20 anni</TIMEX2>`

con l'ultima moglie, una thailandese.

Temporal Expressions: Data

	Training	Test	Total
Points	2279	1174	3453 (75%)
Periods	507	382	889 (19%)
Not gregorian dates	146	122	268 (6%)
Total	2932	1678	4610

Outline

- The ONTOTEXT project
- Description of the I-CAB corpus
- Temporal expression annotation
- **Entity annotation**
- **Entity mention annotation**
- **Adaptations to Italian**
- Inter-annotator agreement
- The I-CAB browser
- Conclusions

Entities

- ◆ **Entity**: an object or a set of objects in the world
 - **Person (PER)**: are limited to humans. A person may be a single individual or a group.
Es. *Ciampi, i calciatori, la mia famiglia*
 - **Organization (ORG)**: are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
Es. *Microsoft, Università di Firenze, Polizia di Stato*
 - **Geo-political Entity (GPE)**: geographical regions defined by political and/or social groups.
Es. *Italia, Provincia Autonoma di Trento, Trento*
 - **Location (LOC)**: geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
Es. *Marte, il Caucaso, il Po*

Entity Mentions

- ◆ **Entity Mentions**: textual realizations of entities

Es. *Ciampi* -> *l'ex presidente della Repubblica*
-> *egli*
-> *il senatore a vita*

We annotated:

- the **extent**, the entire nominal phrase used to refer to an entity.

The extent includes:

- modifiers, “*Una grande famiglia*”
 - prepositional phrases, “*Il Presidente della Repubblica*”
 - dependent clauses, “*La ragazza che lavora in giardino*”
- the **syntactic head**
 - the **co-reference**

Example of ACE entity annotation (1)

KUALA LUMPUR - 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie.

Example of ACE entity annotation (2)

KUALA LUMPUR - 53 matrimoni: è il record personale di **un malese**.

Kamarudin Mohamad ha risposato infatti la **sua** prima moglie.

- 1) **Person entity** *Kamarudin Mohamad*
(mentioned three times: *un malese*, *Kamarudin Mohamad* and *sua*)

Example of ACE entity annotation (3)

KUALA LUMPUR - 53 matrimoni: è il record personale di un malese.

Kamarudin Mohamad ha risposato infatti la sua prima moglie.

2) **Person entity** *La sua prima moglie* (mentioned once)

Entity subtypes

PER: Individual (*Ciampi*), Group (*la mia famiglia*), Indefinite (*No so chi arriverà*)

ORG: Government (*i Carabinieri*), Commercial (*Microsoft*), Educational (*l'Università di Pisa*), Media (*National Geographic*), Religious (*la Chiesa valdese*), Sports (*il Milan*), Medical-Science (*il laboratorio analisi*), Non-Governmental (*la Croce Rossa*), Entertainment (*la compagnia teatrale*)

GPE: Continent (*l'Asia*), Nation (*l'Italia*), State-or-Province (*la Florida*), County-or-District (*il Canton Ticino*), Population-Center (*Trento*), GPE-Cluster (*l'Unione Europea*), Special (*La Palestina*)

LOC: Address (*Via Nazionale 12*), Boundary (*Il confine siriano*), Celestial (*The sun*), Waterbody (*il mare*), Lend-Region-Natural (*il Monte Bianco*), Region-International (*L'Africa meridionale*), Region-General (*il nord-est*)

Entity Mention types

Mentions of entities can be of different types, for example:

- **Proper names** → *Kamarudin Mohamad*
- **Nominal compounds** → *la sua prima moglie*
- **Pronouns** → *egli, sua*
- **Pre-modifier** → *il malese Kamarudin Mohamad*
- **Appositive constructions** → *Leopardi, famoso poeta*

Entities and Entity Mentions: Data

		Training	Test	TOTAL
PER	Entities	4.531	2.679	7.210 (53%)
	Mentions	10.136	6.174	16.310 (57%)
ORG	Entities	2.235	1.047	3.282 (24%)
	Mentions	4.336	1.964	6.300 (22%)
LOC	Entities	398	213	611 (4%)
	Mentions	575	310	885 (3%)
GPE	Entities	1.466	955	2.421 (18%)
	Mentions	2.928	1.821	4.749 (17%)
MIX	Entities	131	70	201 (1%)
	Mentions	166	109	275 (1%)
TOT	Entities	8.761	4.964	13.725
	Mentions	18.141	10.378	28.519

Italian vs. English: Morpho-syntactic adaptations

Adaptations to Italian morpho-syntactic features:

- articulated prepositions are included in the extent

at <the end of March> vs. <*alla* fine di marzo>

in <the United States> vs. <*negli* Stati Uniti>

- mention types ENCLIT and PROCLIT have been added
(respectively for enclitics and proclitics that are attached at the end
or the beginning of a word)

to see <*him*> vs. <*veder**lo*>

Can you talk to <*him*> *about that?* vs. <*gliene*> *puoi parlare?*

- mention type POST has been added
(to annotate modifiers that follow the head noun)

a <*French*> *stylist* vs. *uno stilista* <*francese*>

Italian vs. English: Extensions

Extensions:

- mention type CONJ

(to annotate conjunctions of entities allowing us to mark the co-reference with anaphoric mentions which might follow in the text: e.g. *essi* = *they*, *queste persone* = *these people*, etc.)

<*the woman*> *and* <*the child*> -> 2 entities

<<*la donna*> *e* <*il bambino*>> -> 3 entities

- entity type MIXED

(to annotate non-uniform entity groups for which it is impossible to choose a single semantic type)

<*il medico e l'ospedale*> (*the doctor and the hospital*)

Outline

- The ONTOTEXT project
- Description of the I-CAB corpus
- Temporal expression annotation
- Entity annotation
- Entity mention annotation
- Adaptations to Italian
- **Inter-annotator agreement**
- The I-CAB browser
- Conclusions

Inter-annotator agreement: Datasets

- ◆ 5 different subsets of I-CAB made of 10 news stories (chosen randomly)
- ◆ Matching criteria for TEs: TERN 2004 evaluation scorer
- ◆ Matching criteria for Entities: ACE 2005 evaluation scorer

	# Words	Annotator A	Annotator B	# Common annotations
TE	5,204	165	166	158
PER	4,657	153	167	145
ORG	3,405	46	52	42
GPE	4,741	46	46	46
LOC	4,868	11	11	11

- ◆ Dice coefficient (for detection of TEs, Entities and Mentions)
$$Dice = 2C / (A + B)$$
- ◆ Kappa statistic (for attribute and subtype assignment)

Inter-annotator agreement: TEs

- ◆ Detection: *Dice*=0.955
- ◆ Bracketing: *Dice*=0.931
- ◆ Normalization (on the TEs uniformly bracketed):

	Agreement*	Kappa Statistic	Value Range
VAL	92.2% (142/154)	-	-
ANCHOR_VAL	92.2% (142/154)	-	-
ANCHOR_DIR	90.3% (139/154)	0.749	6
MOD	99.3% (153/154)	0.886	12
SET	98.7% (152/154)	0.744	2

* Cases where the two annotators agreed in assigning or not assigning a value for the attribute

Inter-annotator agreement: Entities (1)

◆ Person Entities

- entity detection: *Dice* = 0.906
- mention detection (on uniformly detected entities): *Dice* = 0.951
- subtype assignment (on uniformly detected entities): *Kappa* = 0.937
- extent mismatch (on mentions detected by both annotators): 3.8%

◆ Organization Entities

- entity detection: *Dice* = 0.857
- mention detection (on uniformly detected entities): *Dice* = 0.845
- subtype assignment (on uniformly detected entities) : *Kappa* = 0.970
- extent mismatch (on mentions detected by both annotators): 3.8%

Inter-annotator agreement: Entities (2)

◆ Geo-Political Entities

- entity detection: $Dice = 1$
- mention detection (on uniformly detected entities): $Dice = 0.980$
- subtype assignment (on uniformly detected entities): $Kappa = 1$
- extent mismatch (on mentions detected by both annotators): 0%

◆ Location Entities

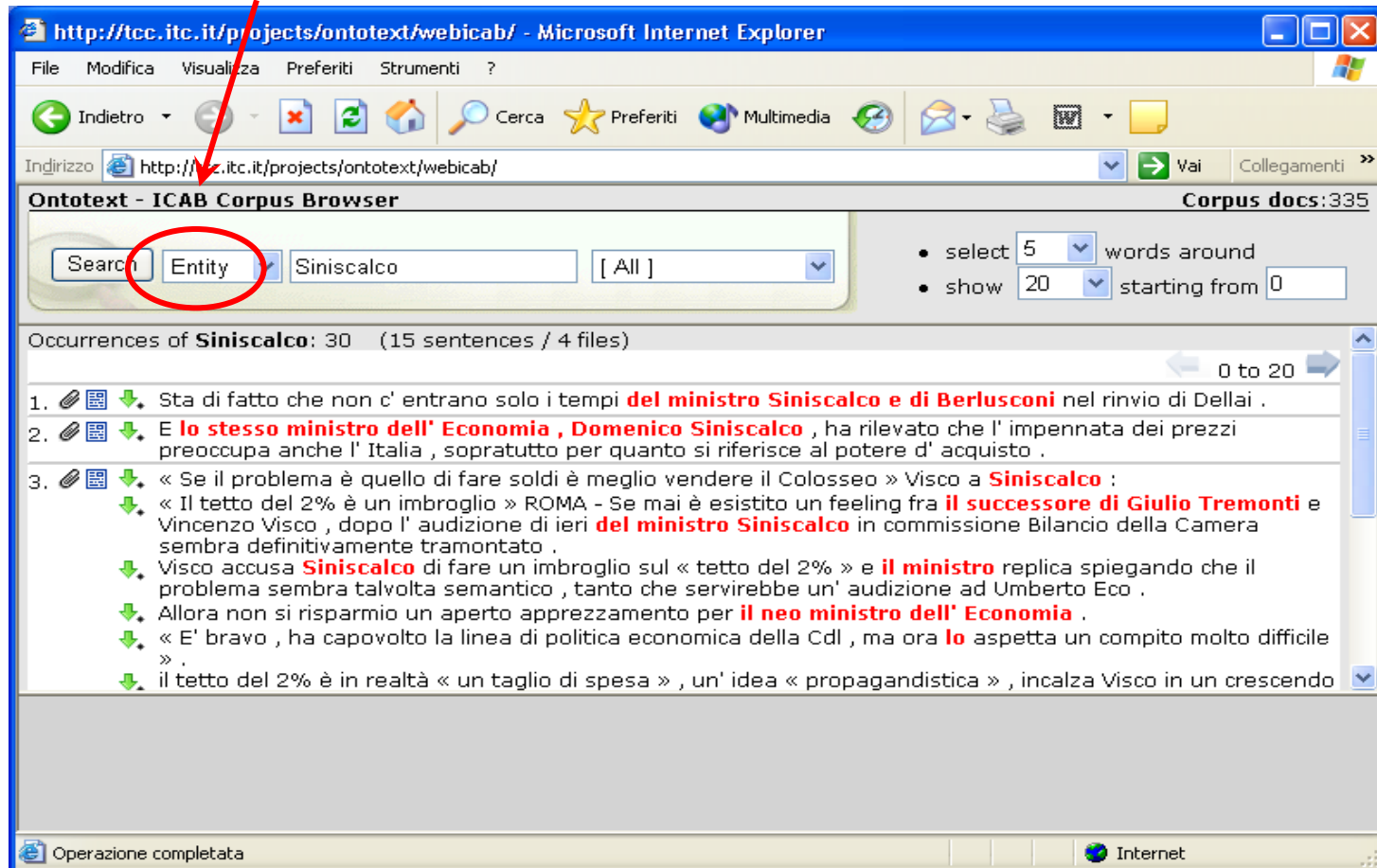
- entity detection: $Dice = 0.957$
- mention detection (on uniformly detected entities): $Dice = 0.938$
- subtype assignment (on uniformly detected entities): $Kappa = 1$
- extent mismatch (on mentions detected by both annotators): 0%

I-CAB browser (1)

- ◆ I-CAB is directly accessible from the Ontotext website at: <http://tcc.itc.it/projects/ontotext/webicab/>
- ◆ the I-CAB browser allows to search the corpus by token, by lemma, or by entity
- ◆ the I-CAB browser allows to select a specific news story and visualize the linguistic annotations (parts of speech and multiwords) and the semantic annotations (temporal expressions and entity mentions) it contains

I-CAB browser (2)

Corpus search (by entity)



The screenshot shows the Microsoft Internet Explorer browser window displaying the I-CAB Corpus Browser interface. The address bar shows the URL <http://tcc.itc.it/projects/ontotext/webicab/>. The browser title is "Ontotext - ICAB Corpus Browser". The search bar contains the text "Siniscalco" and the search type is set to "Entity". The search results show 30 occurrences of "Siniscalco" (15 sentences / 4 files). The results list includes:

1. Sta di fatto che non c' entrano solo i tempi **del ministro Siniscalco e di Berlusconi** nel rinvio di Dellai .
2. E **lo stesso ministro dell' Economia , Domenico Siniscalco** , ha rilevato che l' impennata dei prezzi preoccupa anche l' Italia , soprattutto per quanto si riferisce al potere d' acquisto .
3. « Se il problema è quello di fare soldi è meglio vendere il Colosseo » Visco a **Siniscalco** :
« Il tetto del 2% è un imbroglio » ROMA - Se mai è esistito un feeling fra **il successore di Giulio Tremonti e Vincenzo Visco** , dopo l' audizione di ieri **del ministro Siniscalco** in commissione Bilancio della Camera sembra definitivamente tramontato .
Visco accusa **Siniscalco** di fare un imbroglio sul « tetto del 2% » e **il ministro** replica spiegando che il problema sembra talvolta semantico , tanto che servirebbe un' audizione ad Umberto Eco .
Allora non si risparmia un aperto apprezzamento per **il neo ministro dell' Economia** .
« E' bravo , ha capovolto la linea di politica economica della Cdl , ma ora **lo** aspetta un compito molto difficile » .
il tetto del 2% è in realtà « un taglio di spesa » , un' idea « propagandistica » , incalza Visco in un crescendo

The status bar at the bottom indicates "Operazione completata" and "Internet".

I-CAB browser (3)

Visualization of (semantic) annotations

The screenshot shows a Microsoft Internet Explorer window titled "MEANING Corpus Annotation". The browser's address bar displays the URL "l'Adige - training/adige20041008_id413974". The page content is a news article snippet with several words and phrases highlighted in red and enclosed in boxes, indicating semantic annotations. The browser's interface includes a menu bar (File, Modifica, Visualizza, Preferiti, Strumenti), a toolbar with navigation buttons (Indietro, Avanti, Home, Cerca, Preferiti, Multimedia), and a search bar. Below the toolbar, there are checkboxes for "Linguistic annotation" (Noun, Verb, Adjective, Adverb, Multiword) and "Semantic annotation" (Time, Mention). The "Time" and "Mention" checkboxes are circled in red. The article text includes the following annotations: "Visco" (Mention), "Siniscalco" (Mention), "il successore di Giulio Tremonti" (Mention), "Vincenzo Visco" (Mention), "ieri" (Time), "del" (Mention), "ministro" (Mention), "Siniscalco" (Mention), "commissione Bilancio della Camera" (Mention), "Visco" (Mention), "Siniscalco" (Mention), "il ministro" (Mention), "Umberto Eco" (Mention), "Visco" (Mention), "lo stesso ex" (Mention), "ministro delle Finanze" (Mention), "lo scorso agosto" (Time), "il neo ministro dell' Economia" (Mention), "Ora" (Time), "lo" (Mention), "Ora" (Time), "si" (Mention), "si" (Mention), "si" (Mention), "ministro" (Mention).

Distribution: EVALITA 2007

Evaluation of NLP Tools for Italian: <http://evalita.itc.it/>

- Five tasks:
1. Part of Speech Tagging
 2. Parsing
 3. Word Sense Disambiguation
 4. Temporal Expression Recognition and Normalization
 5. Named Entity Recognition

Final workshop organized in conjunction with AI*IA 2007
10th September 2007, Rome

The proceedings of Evalita 2007 will be published as a special issue of *Intelligenza Artificiale*, the journal of AI*IA.

<i>Task</i>	<i>Registrations</i>
PoS-Tagging	15
Parsing	13
Word Sense Disambiguation	7
Temporal Expressions	6
Named Entities	15

Distribution: EVALITA 2007

Evaluation of NLP Tools for Italian: <http://evalita.itc.it/>

- Five tasks:
1. Part of Speech Tagging
 2. Parsing
 3. Word Sense Disambiguation
 4. Temporal Expression Recognition and Normalization
 5. Named Entity Recognition

Final workshop organized in conjunction with AI*IA 2007
10th September 2007, Rome

The proceedings of Evalita 2007 will be published as a special issue of *Intelligenza Artificiale*, the journal of AI*IA.

<i>Task</i>	<i>Registrations</i>
PoS-Tagging	15
Parsing	13
Word Sense Disambiguation	7
Temporal Expressions	6
Named Entities	15

Ongoing work

- Manual annotation of Relations

Relation Detection and Characterization (RDC) Task

ARG2 ARG1	GPE	LOC	ORG	PER
GPE	PHYSICAL–Near PART WHOLE–Geographical ORG AFFILIATION–Investor	PHYSICAL–Located PHYSICAL–Near PART WHOLE–Geographical	ORG AFFILIATION–Investor ORG AFFILIATION–Membership	
LOC	PHYSICAL–Near PART WHOLE–Geographical	PHYSICAL–Near PART WHOLE–Geographical		
ORG	PHYSICAL–Located PART WHOLE–Subsidiary ORG AFFILIATION–Founder ORG AFFILIATION–Investor GPE AFFILIATION–ORG Origin	PHYSICAL–Located GPE AFFILIATION–ORG Origin	PART WHOLE–Subsidiary ORG AFFILIATION–Founder ORG AFFILIATION–Investor ORG AFFILIATION–Membership ORG AFFILIATION–Customer	
PER	PHYSICAL–Located PHYSICAL–Near ORG AFFILIATION–Employment ORG AFFILIATION–Founder ORG AFFILIATION–Investor GPE AFFILIATION–Citizen	PHYSICAL–Located PHYSICAL–Near GPE AFFILIATION–Citizen	PHYSICAL–Located ORG AFFILIATION–Employment ORG AFFILIATION–Ownership ORG AFFILIATION–Founder ORG AFFILIATION–Student ORG AFFILIATION–Sport ORG AFFILIATION–Investor ORG AFFILIATION–Membership ORG AFFILIATION–Customer GPE AFFILIATION–Religion	GPE AFFILIATION–Citizen SOCIAL–Business SOCIAL–Family SOCIAL–Lasting Personal SOCIAL–Parent Child SOCIAL–Sibling SOCIAL–Married

Ongoing work

- Manual annotation of Relations

Relation Detection and Characterization (RDC) Task

ARG2 ARG1	GPE	LOC	ORG	PER
GPE	PHYSICAL-Near PART WHOLE-Geographical ORG AFFILIATION-Investor	PHYSICAL-Located PHYSICAL-Near PART WHOLE-Geographical	ORG AFFILIATION-Investor ORG AFFILIATION-Membership	
LOC	PHYSICAL-Near PART WHOLE-Geographical	PHYSICAL-Near PART WHOLE-Geographical		
ORG	PHYSICAL-Located PART WHOLE-Subsidiary ORG AFFILIATION-Founder ORG AFFILIATION-Investor GPE AFFILIATION-ORG Origin	PHYSICAL-Located GPE AFFILIATION-ORG Origin	PART WHOLE-Subsidiary ORG AFFILIATION-Founder ORG AFFILIATION-Investor ORG AFFILIATION-Membership ORG AFFILIATION-Customer	
PER	PHYSICAL-Located PHYSICAL-Near ORG AFFILIATION-Employment ORG AFFILIATION-Founder ORG AFFILIATION-Investor GPE AFFILIATION-Citizen	PHYSICAL-Located PHYSICAL-Near GPE AFFILIATION-Citizen	PHYSICAL-Located ORG AFFILIATION-Employment ORG AFFILIATION-Ownership ORG AFFILIATION-Founder ORG AFFILIATION-Student ORG AFFILIATION-Sport ORG AFFILIATION-Investor ORG AFFILIATION-Membership ORG AFFILIATION-Customer GPE AFFILIATION-Religion	GPE AFFILIATION-Citizen SOCIAL-Business SOCIAL-Family SOCIAL-Lasting Personal SOCIAL-Parent Child SOCIAL-Sibling SOCIAL-Married

Conclusions

- ◆ We have adapted to Italian the ACE–LDC guidelines, trying to create a standard for the annotation of Italian texts
- ◆ The ACE standards turned out to be flexible to be adapted to our needs
- ◆ Annotations are directly accessible through the I–CAB browser (<http://tcc.itc.it/projects/ontotext/webicab/>)
- ◆ I–CAB has been made freely available in the context of EVALITA 2007 (Evaluation of NLP Tools for Italian)

References

- ◆ Magnini, Pianta, Girardi, Negri, Romano, Speranza, Bartalesi Lenzi, Sprugnoli. *I-CAB: the Italian Content Annotation Bank*, in Proceedings of LREC 2006.
- ◆ Magnini, Negri, Pianta, Romano, Speranza, Serafini, Girardi, Bartalesi Lenzi, Sprugnoli. *From Text to Knowledge for the Semantic Web: the ONTOTEXT Project*, in Proceedings of SWAP 2005.
- ◆ Magnini, Cappelli, Pianta, Speranza, Bartalesi Lenzi, Sprugnoli, Romano, Girardi, Negri. *Annotazione di contenuti concettuali in un corpus italiano: I-CAB*, in Proceedings of SILFI 2006.
- ◆ Pianta, Speranza, Magnini, Bartalesi Lenzi and Sprugnoli. *Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 2.1)*, ITC-irst Technical Report.
- ◆ Pianta, Speranza, Magnini, Bartalesi Lenzi and Sprugnoli. *Italian Content Annotation Bank (I-CAB): Person Entities (V. 1.3)*, ITC-irst Technical Report.
- ◆ Pianta, Speranza, Magnini, Bartalesi Lenzi and Sprugnoli. *Italian Content Annotation Bank (I-CAB): Organization Entities (V. 2.0)*, ITC-irst Technical Report.

Thank you