I-CAB: the Italian Content Annotation Bank

**B. Magnini*, E. Pianta*, C. Girardi*, M. Negri*, L. Romano*,
M. Speranza*, V. Bartalesi Lenzi**, and R. Sprugnoli****

* ITC.irst, Povo (Trento), Italy

** CELCT, Povo (Trento), Italy

E-mail: magnini@itc.it, pianta@itc.it, cgirardi@itc.it, negri@itc.it, romano@itc.it, manspera@itc.it, bartalesi@celct.it, sprugnoli@celct.it

Abstract

In this paper we present work in progress for the creation of the Italian Content Annotation Bank (I-CAB), a corpus of Italian news annotated with semantic information at different levels. The first level is represented by temporal expressions, the second level is represented by different types of entities (i.e. person, organizations, locations and geo-political entities), and the third level is represented by relations between entities (e.g. the affiliation relation connecting a person to an organization). So far I-CAB has been manually annotated with temporal expressions, person entities and organization entities. As we intend I-CAB to become a benchmark for various automatic Information Extraction tasks, we followed a policy of reusing already available markup languages. In particular, we adopted the annotation schemes developed for the ACE Entity Detection and Time Expressions Recognition and Normalization tasks. As the ACE guidelines have originally been developed for English, part of the effort consisted in adapting them to the specific morpho-syntactic features of Italian. Finally, we have extended them to include a wider range of entities, such as conjunctions.

1. Introduction

In recent years there have been several initiatives for the realization of annotated resources for different tasks in Natural Language Processing, including Word Sense Disambiguation (e.g. Semcor), parsing (e.g. the PennTreebank) and Named Entity Recognition. More recently, the ACE (Automatic Content Extraction) program started developing a set of annotation schemas for higher level tasks in Information Extraction, addressing Temporal Expressions, mentions of Entities and mentions of Relations among entities. On the basis of the resulting resources a number of evaluation campaigns have been successfully organized (e.g. TERN 2004 and 2005, ACE 2002-2005) for Content Annotation tasks.

While such efforts have stimulated research in Information Extraction for the English language, little has been done for other languages; in particular, there are no content-annotated resources for Italian. This paper presents ongoing work aimed at the realization of I-CAB (Italian Content Annotation Bank), a corpus of semantically annotated documents for Italian containing annotations of PERSON ENTITIES, ORGANIZATION ENTITIES, LOCATION ENTITIES, GEO-POLITICAL ENTITIES and of a number of selected RELATIONS among such entities.

Following a policy of reusing already available markup languages, the annotation activity has been carried out adopting the formalisms developed within the American ACE program[1]. However, due to the differences between English and Italian, part of the work has been dedicated to the revision and adaptation to Italian of the annotation guidelines (Lavelli et al. 2005).

The main result of the manual annotation is represented by the first release of the Italian Content Annotation Bank (I-CAB) corpus. I-CAB is an Italian corpus of news stories (around 182,000 words) which at present contains annotations about PERSON ENTITIES and TEMPORAL EXPRESSIONS; it is accessible on the Web through a browser created specifically for this purpose.

The creation of I-CAB is part of the three-year project Ontotext[2] funded by the Autonomous Province of Trento. Ontotext focuses on the study and development of innovative knowledge extraction techniques to produce new or less noisy information to be made available for the Semantic Web. Within the new research area of Ontology-Based Knowledge Extraction, Ontotext addresses three key research aspects: annotating documents with semantic and relational information, providing an adequate degree of interoperability of such relational information, and updating and extending the ontologies used for Semantic Web annotation. The concrete evaluation scenario in which algorithms will be tested with a number of large-scale experiments is the automatic acquisition of information about people from newspaper articles.

The paper is structured as follows. In Section 2 we will present how mark-up languages are used in the framework of the ACE program. In Section 3 we will describe the corpus. In Section 4 and Section 5 we will report on the annotation of TEMPORAL EXPRESSIONS, PERSON ENTITIES and ORGANIZATION ENTITIES. Finally, in Section 6 we will draw some conclusions.

2. Content Mark-up Languages

The ACE formalisms have been chosen because they represent a flexible mark-up language to identify content information in a given source text, and annotate it with additional metadata providing a semantically rich and normalized description.

The aim of the ACE program is to develop extraction technology to support automatic processing of source language data. In particular ACE annotators tag English,

[1] http://www.nist.gov/speech/tests/ace

[2] http://ontotext.itc.it

Chinese and Arabic texts, producing both training and test sets for the evaluation of technologies that automatically detect and characterize the meaning conveyed by the data.

The ACE program is motivated by the same issues as the Message Understanding Conference (MUC) program that preceded it, but represents an evolution in terms of complexity. In particular, in the MUC Name Entity Task three entity types were considered (persons, organizations, and locations) and only proper names and acronyms were markable, while in the MUC Co-reference Task all co-referring expressions, i.e. all mentions of a given entity, were captured and grouped.

ACE modifies the list of entity types dividing locations into geo-political entities and facilities and by adding weapons, substances, and vehicles. Co-reference is preserved but a wider range of markable expressions, including common nouns and pronouns, is taken into account. Finally, two inter-connected levels of annotation are defined: the level of the *entity*, which provides a representation of an object in the world, and the level of the *entity mention*, which provides information about any textual references to that object. For instance, if George W. Bush is mentioned in two different sentences of a text as *the president of the U.S.A.* and as *he*, these two expressions are considered as two co-referring entity mentions (i.e. two mentions of the same entity).

For our purposes, the ACE standards developed for the Entity Detection and Recognition task and the Time Expression Recognition and Normalization task turned out to be adequate, as they allow for a semantically rich and normalized annotation of: (i) different types of entities (i.e. objects or set of objects in the world), (ii) different types of entity mentions (i.e. any textual reference to an entity), and (iii) different types of temporal expressions (e.g. absolute expressions, such as "Sunday, March 13 2005", and implicit expressions, such as "three days later").

We also follow the guidelines provided by the Linguistic Data Consortium (LDC) that develops linguistic resources to support the ACE program. In 2004 LDC distributed samples of ALF (the "ACE LDC Format"), that is the style of annotation they proposed for the ACE program. It differs from APF (the "ACE Program Format") by adding a number of new mention types to those proposed in APF.

3. Annotation Process

The Italian Content Annotation Bank (I-CAB) is a corpus of Italian news documents annotated with different kinds of semantic information.

3.1 Description of the Corpus

I-CAB consists of 525 news documents taken from the local newspaper 'L'Adige'[3]. The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004) and are grouped into five categories: News Stories, Cultural News, Economic News, Sports News and Local News (see Table 1).

	09/07	09/08	10/07	10/08	Total
News	23	25	18	21	87
Culture	20	18	16	18	72
Economy	13	15	12	14	54
Sport	29	41	27	26	123
Local	46	43	49	51	189
TOTAL	131	142	122	130	525

Table 1: Number of news stories per category

I-CAB is further divided into the training and the test sections, which contain 335 and 190 documents respectively. In total, I-CAB consists of around 182,500 words: 113,500 in the training part (the average length of a news story is around 339 words) and 69,000 words in the test part (with an average of 363 words per news stories).

The annotation of I-CAB is being carried out manually, as we intend I-CAB to become a benchmark for various automatic Information Extraction tasks, including recognition and normalization of TEMPORAL EXPRESSIONS, ENTITIES and RELATIONS between entities (e.g. the relation affiliation connecting a person to the organization with which he or she is affiliated).

The annotation of I-CAB is work in progress. So far, the whole corpus has been annotated with TEMPORAL EXPRESSIONS and PERSON ENTITIES, while only I-CAB Training has been annotated with ORGANIZATION ENTITIES (see Table 2). The work started in October 2004 and required 2.5 person/years.

		Training	Test	Total
TIME EXPRESS.	Tags	2901	1652	4553
PERS. ENTITIES	Entities	4459	2628	7087
	Mentions	9994	6065	16059
ORG. ENTITIES	Entities	2217	-	2217
	Mentions	4235	-	4235

Table 2: Annotation data

3.2 Annotation Tool and Formats

For the creation of I-CAB we have chosen the freely distributed annotation tool Callisto[4], developed at the MITRE Corporation. It supports linguistic annotation of textual sources for any Unicode-supported language and accepts files encoded as UTF-8, US-ASCII and several other character encodings. Callisto is written in Java, taking advantage of its portability and language support; it has been built with a modular design and utilizes standoff-annotation, allowing for unique tag-set definitions and domain dependent interfaces. Stand-off annotation support allows for many different annotation tasks to be represented. For the annotation of TEMPORAL EXPRESSIONS we have used the TIMEX2 task, whereas for the annotation of ENTITIES we are using the ACE2004 task.

All data annotated with Callisto are saved in the Atlas Interchange Format (AIF). The TIMEX2 task also allows exporting annotated files from the AIF into the SGML

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

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

format, whereas the ACE2004 task does not allow it. In I-CAB, the manual annotation of the corpus is merged with automatic annotation of lower linguistic levels (tokenization, lemma, PoS, multi-words). All the different levels of annotations are delivered in the Meaning Annotation Format, an XCES and TEI conformant scheme which was developed within the EU-funded MEANING project (Bentivogli et al., 2003) and has now been extended to represent TEMPORAL EXPRESSIONS and ENTITIES.

The Meaning Annotation Format (MAF) is a stand-off XML-based annotation scheme. Different representation levels are contained in separate documents, or document sections. Annotation levels are related to each other following a hierarchy of annotation levels: first, the orthographic annotation level, representing tokens, is implemented with pointers to the character positions in the hub corpus; second, the morpho-syntactic level contains pointers to the tokens; third, the multiword level points to the words described at morpho-syntactic level. According to this hierarchical approach, temporal expressions and entity mentions are represented with pointers to morpho-syntactic level entities and entities are represented with pointers to entity mentions (Pianta et al., 2006).

Unlike temporal expressions and entity mentions in MAF, all the annotations produced by Callisto in the AIF format point to character positions; as a consequence of this, in the transformation from AIF to MAF, pointers to character positions have been substituted with pointers to morpho-syntactic objects.

4. Time Expression Recognition and Normalization

For the annotation of TEMPORAL EXPRESSIONS (TEs) we have followed the TIMEX2 mark-up standard (Ferro et al. 2004), according to which markable expressions include both time durations (e.g. *three years*) and points (e.g. *July 17th 1999*, *today*). Time points can be either absolute expressions (e.g. *the 17th of July, 1999*) or relative, i.e. anaphoric expressions (e.g. *today*). Also markable are event anchored expressions (e.g. *two days before the departure*) and sets of times (e.g. *every month*).

The standards developed for the Time Expressions Recognition and Normalization tasks allow for a semantically rich and normalized annotation. TEs Recognition refers to the task of finding the TEs within a text (detection) and determining their extension (bracketing). TEs Normalization refers to the task of interpreting the TEs by assigning values to pre-defined normalization attributes.

Normalization attributes are described as follows:

- *VAL*: contains the value of a TE (e.g. VAL="2004-05-06" for the date *<6 maggio 2004>*/*May 6th, 2004* and VAL="P6D" for the period *<sei giorni>*/*six days*); no VAL is attributed to underspecified TEs (e.g. *<per lungo tempo>*/*for a long time*);
- *MOD*: captures temporal modifiers. Possible values are APPROX (*<verso mezzanotte>*/*around midnight*), MORE THAN (e.g. *<più di 3 ore>*/*more than 3 hours*) and START (e.g. *<i primi anni '70>*/*the early 70s*);
- *ANCHOR VAL:* contains a normalized form of an anchoring date/time and appears in combination with ANCHOR_DIR;
- *ANCHOR DIR*: captures the direction of a TE, e.g. AFTER and BEFORE. For instance, assuming May 6th, 2004 as the reference time, the TE in *<sarò in vacanza per due mesi>*/*I will be on holiday for two months* is normalized as: VAL="P2M" ANCHOR_VAL= "2004-05-06" and ANCHOR DIR="AFTER" (as the period of two months is after the reference date);
- *SET*: identifies expressions denoting sets of time. E.g. *<ogni anno>*/*every year* is annotated with SET="YES".

The adaptation of the TIMEX2 annotation scheme to the annotation of Italian texts required some extensions (Lavelli et al., 2005). In particular, as a consequence of the specific features of Italian, which has a far richer morphology than English, we have introduced some changes concerning the extension of TEMPORAL EXPRESSIONS. According to the guidelines, definite and indefinite articles are considered as part of their textual realization, while prepositions are not (e.g. *at <the end of March>*). As the annotation is word-based, this does not account for Italian articulated prepositions, where a definite article and a preposition are merged, as in *alla (a+la) fine di marzo/at the end of March*. We have decided that this type of preposition should be included, so as to consistently include all the articles (e.g. *<alla fine di marzo>/at <the end of March>*); the same criteria have been adopted for the annotation of entities (see Section 5).

As shown in Table 3, the total number of annotated TEMPORAL EXPRESSIONS is around 4,550 (2,901 and 1,652 in the training and test sections respectively); in both sections of the corpus, the number of time points is slightly higher that the number of time durations.

As to the normalization of TEs, the combination ANCHOR_DIR and ANCHOR_VAL is the most frequently used attribute, as about 23% of the TEs in the corpus are anchored durations (see Table 4).

	Training		**Test**		**Total**	
Points	1553	53.5%	796	48.2%	2349	51.6%
Durations	1207	41.6%	738	44.7%	1945	42.7%
Underspec.	141	4.9%	118	7.1%	259	5.7%
TOTAL	2901		1652		4553	

Table 3: Occurrences and percentage of points, durations and TEMPORAL EXPRESSIONS with no value

Attribute	**Training**		**Test**		**Total**	
VAL	2760	95,1%	1534	92,9%	4294	94,3%
ANCH._VAL	696	24%	362	21.9%	1058	23.2%
ANCH._DIR	696	24%	362	21.9%	1058	23.2%
MOD	112	3.9%	76	4.6%	188	4.1%
SET	121	4.2%	51	3.1%	172	3.8%

Table 4: Occurrences (in absolute numbers and percentages) of normalization attributes

Inter-annotator agreement has been evaluated on the dual annotation of a corpus of ten randomly chosen news

stories, for a total of about 5,204 words.

The most commonly used measure to characterize inter-annotator agreement is the kappa statistic (Cohen 1960), which measures pairwise agreement among a set of coders making category judgments taking into consideration agreement obtained by chance. In the case of TEs (and ENTITIES), however, annotators can theoretically choose to tag any sequence of adjacent tokens in a sentence. This makes it necessary to consider every possible sequence as a candidate, which would give exceedingly low results in terms of kappa, as annotation would become a binary categorization problem with an extremely skewed distribution (only a minimum number of the candidate sequences, in fact, are TEs).

For this reason we have used the kappa statistic to simply measure the agreement in determining whether each token is or is not part of any TE, and we have obtained k=0.958. However, this measure does not take into account the extent of the annotated TEs, so we have also compared the two annotated versions using the Dice coefficient. The Dice coefficient is computed as in [1], where C is the number of common annotations, while A and B are respectively the number of annotations provided by the first and the second annotator.

[1] Dice=2C/(A+B) [5]

The Dice coefficient is 0.955 for TE detection and 0.931 for TE bracketing. Agreement in normalization has been measured on the TEs uniformly bracketed. Table 5 reports, for each attribute, the cases where the two annotators agreed in assigning or not assigning a value for the attribute and, for the attributes which admit a restricted number of values, it also reports the kappa statistic.[6]

	agreement	kappa statistic
VAL	92.2% (142/154)	-
ANCH._VAL	92.2% (142/154)	-
ANCH._DIR	90.3% (139/154)	0.749
MOD	99.3% (153/154)	0.886
SET	98.7% (152/154)	0.744

Table 5: Agreement in attribute value assignment

5. Entity Detection and Recognition

As indicated in the ACE Entity Detection task, the annotation of ENTITIES (e.g. PERSONS, ORGANIZATION, LOCATIONS AND GEO-POLITICAL ENTITIES) requires that the entities mentioned in a text be detected, their syntactic head marked, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity.

Entity mentions, i.e. textual realizations of entities, can be intuitively described as portions of text; the extent of this portion of text is defined to be the entire nominal phrase used to refer to an entity, thus including modifiers (e.g. <*una grande* [*famiglia*]>/*a big family*)[7], prepositional phrases (e.g. <*il* [*Presidente*] *della Repubblica*>/*the President of the Republic*) and dependent clauses (e.g. <*la* [*ragazza*] *che lavora in giardino*>/*the girl who is working in the garden*).

ACE classifies entity mentions according to two dimensions: (i) the kind of reference they make to entities in the world and (ii) their syntactic features.

On the basis of the reference they make to entities in the world, we distinguish four types of entity mentions:

- Specific referential (SPC) are those where the entity being referred to is a unique object or set of objects (e.g. *John's lawyer won the case*).
- Generic referential (GEN) refer to a kind or type of entity and not to a particular object (or set of objects) in the world (e.g. *Lawyers don't work for free*).
- Under-specified referential (USP) are non-generic non-specific references, including imprecise quantifications (e.g. *many/some/15 thousand people*), quantified NP's in future, hypothetical, or question contexts (e.g. *I wonder who arrived*), etc.
- Negatively quantified (NEG) refer to the empty set of the mentioned type of object (e.g. *No lawyer*).

As for syntactic features, we distinguish between:

- NAM: proper names (e.g. <[*Totti*]>, <[*ONU*]>/*UN*);
- NOM: nominal constructions (e.g. <*i* [*bambini*] *buoni*>/*good children*, <*l'*[*azienda*]>/*the company*);
- PRO: pronouns, e.g. personal (<[*tu*]>/*you*) and indefinite (<[*qualcuno*]>/*someone*);
- WHQ: wh-words, such as relatives and interrogatives (e.g. <[*Chi*]> *è lì?*/*Who is there?*);
- PTV: partitive constructions (e.g. <[*alcune/una*] *delle scuole*>/one/*some of the schools*);
- APP: appositive constructions (e.g. *<[la Juventus, la squadra italiana]>/Juventus, the Italian club).*

Some new types of mentions have been added; for instance, we have created a specific tag, ENCLIT, to annotate the clitics whose extension can not be identified at word-level (e.g. <*veder*[*lo*]>/*to see him*). Some types of mentions, on the other hand, have been eliminated; this is the case of pre-modifiers, due to syntactic differences between English, where both adjectives and nouns can be used as pre-modifiers, and Italian, which only admits adjectives in that position.

In extending the annotation guidelines, we have decided to annotate all conjunctions of entities, not only those which share the same modifiers as indicated in the ACE guidelines, thus creating the new mention type CONJ, whose head corresponds to the entire mention (e.g. <[*la madre e il figlio*]>/*mother and son*)[8]. This allows us to mark the co-reference with anaphoric mentions, such as

[5] Notice that the Dice coefficient has the same value of the F_1 measure computed considering any of the two annotators as the reference.

[6] As observed in (Di Eugenio, Glass 2004) the kappa statistic could be affected by bias and prevalence problems. By also calculating kappa according to the (Siegel, Castellan 1988) definition we verified there are no bias problems (values are equal), but the natural skewing of the distribution of categories does affect kappa (e.g. for the SET attribute).

[7] In Italian examples, mentions are in angular brackets and heads are in square brackets.

[8] Appositive and conjoined mentions are complex constructions. Although LDC does not identify heads for such constructions, we have decided to annotate the whole extent as head.

they or *the two people*, which might follow in the text.

As a consequence of this, a second new mention type had to be created, namely MIX, which is used when the different parts of the CONJ do not have the same syntactic structure; for instance, <[*Maria e suo figlio*]>/*Mary and her child*, is annotated as MIX because neither NAM nor NOM would hold for the whole mention.

5.1 Person Entities

According to the ACE standards, each distinct person or set of people mentioned in a document refers to a PERSON ENTITY. For example, people may be specified by name (*John Smith*), occupation (*the butcher*), pronoun (*he*), etc., or by some combination of these.

PERSON ENTITIES (PEs) are further classified with the following subtypes:

- Individual: PEs which refer to a single person (e.g. *George W. Bush*);
- Group: PEs which refer to more than one person (e.g. *my parents*, *your family*, *Mary and her child*);
- Indefinite: a PE is classified as indefinite when it is not possible to judge from the context whether it refers to one or more persons (e.g. *I wonder who will arrive*).

A total of 7,087 PERSON ENTITIES (on average, 13.5 per document) and 16,059 PERSON ENTITY mentions (30.6 per document) have been identified. On average, an entity is mentioned 2.3 times in a document. The distribution between I-CAB Training and I-CAB Test is as follows: 4,459 entities and 9,994 entity mentions in the first, and 2,628 entities and 6,065 entity mentions in the latter.

As shown in Table 6, the majority of PERSON ENTITIES (almost 80% of the total) belong to the class referential. Table 7, on the other hand, shows a balanced distribution between the two most frequent subtypes (e.g. 47% of individual PEs and 45% of group PEs), with a small group of indefinite PEs (less than 8%).

	Training		**Test**		**Total**	
SPC	3474	77.9%	2142	81.5%	5616	79.2%
GEN	443	9.9%	213	8.1%	656	9.3%
USP	517	11.6%	263	10%	780	11%
NEG	25	0.6%	10	0.4%	35	0.5%
TOTAL	4459		2628		7087	

Table 6: Distribution of PERSON ENTITIES by entity class

	Training		**Test**		**Total**	
Indiv.	2067	46.4%	1256	47.8%	3323	46.9%
Group	1995	44.7%	1206	45.9%	3201	45.2%
Indef.	397	8.9%	166	6.3%	563	7.9%
TOTAL	4459		2628		7087	

Table 7: Distribution of PERSON ENTITIES by subtype

Inter-annotator agreement has been evaluated on the dual annotation of a subset of ten randomly chosen news stories for a total of 4,657 words.

We have adopted the matching criteria of the ACE 2005 distributed scorer:

- an entity is detected by both annotators if they detect at least a mention of that entity;
- a mention is detected by both annotators if the mutual fractional head overlap is at least 30%;
- the maximum extent difference allowed for mentions to be declared an extent match is 4 characters.

Therefore, if one annotates <[*Savani e Vujevic sempre meglio]>/Savani and Vujevic always better* as a mention while the other restricts the extent to *Savani e Vujevic*, we have agreement in mention detection, but no extent match.

The kappa statistic as computed for TEs (i.e. whether a token is or is not part of a TE) does not account for nested annotations. As this phenomenon is extremely frequent in the case of PEs, we have chosen to calculate the Dice coefficient instead (see Section 4) and limit the use of the kappa statistic to the assignment of attributes.

Results are as follows:

- the Dice coefficient for *person entity detection* is 0.906;
- limited to the entities detected by both annotators, the Dice coefficient for *mention detection* is 0.951;
- limited to the entities detected by both annotators, the kappa statistic is 0.937 for subtype assignment (i.e. Group, Individual or Indefinite) and 0.734 for class assignment (this relatively low value is due to the high prevalence of the SPC class and to some mismatches in the USP and GEN classes);
- limited to the mentions detected by both annotators we have a 3.7% of extent mismatch.

5.2 Organization Entities

As indicated in the ACE guidelines, ORGANIZATION ENTITIES are divided into ten different subtypes: Government (*The Navy*), Commercial (*Microsoft*), Educational (*University*), Media (*National Geographic*), Religious (*The Vatican*), Sports (*the Italian ski Club*), Medical-Science (*Massachusetts General Hospital*), Non-Governmental (*The Red Cross*) and Entertainment. (*Theatre Company*). The Mixed subtype has been added to support the annotation of conjunction made of two or more organizations with different subtypes. In the sentence *The University of Trento and Microsoft stipulated an agreement*, for instance, we have a conjunction between an organization of subtype Educational and a Commercial one and so we annotate it as Mixed.

Mentions of foreign organizations have been annotated as proper nouns ("type="NAM") if they were the literal translation of the original name, whereas they have been annotated as nominal constructions (type="NOM") if they were considered a cultural transposition of the concept expressed by the original word. Following this rule, *Dipartimento di Stato Americano* is annotated as NAM since it is the direct translation of *U.S. Department of State*. On the contrary *Polizia francese* is NOM because the official name of the French police is *Gendarmerie*.

The training section contains a total number of 2,217 ORGANIZATION ENTITIES (on average, 6.6 entities per document) and 4,235 mentions (12.6 mentions per document). On average, an entity is mentioned 1.9 times in a document.

Table 8 shows that, similarly to what we saw for PEs,

most of the ORGANIZATION ENTITIES are specific referential (more than 90% of the total). As far as subtypes are concerned, the most frequent are Sports, Commercial, Non-Governmental and Government (see Table 9).

	Training	
SPC	2082	93.9%
GEN	93	4.2%
USP	39	1.8%
NEG	3	0.1%
TOTAL	2217	

Table 8: ORGANIZATION ENTITIES by class

Subtypes	Training	
Government	326	14.7%
Commercial	486	21.9%
Educational	159	7.2%
Media	47	2.1%
Religious	32	1.4%
Sports	581	26.2%
Medical-Science	50	2.3%
Non-Governmental	397	17.9%
Entertainment	104	4.7%
Mixed	35	1.6%
TOTAL	2217	

Table 9: ORGANIZATION ENTITIES by subtype

Inter-annotator agreement has been evaluated on the dual annotation of a corpus of ten randomly chosen news stories for a total of 3,405 words.

Results are as follows:

- the Dice coefficient for *organization entity detection* is 0.857;
- limited to the entities detected by both annotators, the Dice coefficient for *mention detection* is 0.845;
- limited to the entities detected by both annotators, the kappa statistic is 0.970 for subtype assignment and 1 for class assignment;
- limited to the mentions detected by both annotators we have a 3.7% of extent mismatch.

6. Conclusion and Future Work

I-CAB is directly accessible from the Ontotext website through a specific browser[9] which enables the user to search the corpus according to different modalities (e.g. search by document or by token) and to visualize all the annotations or to select only specific types or combinations of types.

In the near future we will annotate I-CAB with GEO-POLITICAL ENTITIES and LOCATIONS. Contemporarily, we will start to annotate RELATIONS between entities and EVENTS as defined in the Relation Detection and Characterization (RDC) and Event Detection and Characterization (EDC) tasks. The corpus will be freely available for research purposes.

[9] http://ontotext.itc.it/webicab

7. Acknowledgements

This work has been supported by the ONTOTEXT project (From Text to Knowledge for the Semantic Web), funded by the Autonomous Province of Trento under the FUP-2004 research program.

8. References

Bentivogli, L., Girardi, C., Pianta, E. (2003). The MEANING Italian Corpus. In Proceedings of the Corpus Linguistics 2003 Conference. Lancaster, UK.

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

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

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE.

Fleischman, M. (2001). Automated Subcategorization of Named Entities. 39th Annual Meeting of the ACL, Student Research Workshop. Toulouse, FR.

Fleischman, M., Hovy, E. (2002). Fine Grained Classification of Named Entities. In Proceedings of COLING 2002. Taipei, TW.

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

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. ITC-irst, Trento

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. To appear in the EACL 2006 Workshop: NLPXML-2006 Multi-dimensional Markup in Natural Language Processing

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