From Text to Knowledge for the Semantic Web: the ONTOTEXT Project

Bernardo Magnini[1], Matteo Negri[1], Emanuele Pianta[1], Lorenza Romano[1], Manuela Speranza[1], Luciano Serafini[1], Christian Girardi[1], Valentina Bartalesi[2], Rachele Sprugnoli[2]

[1] ITC-irst
38050 Povo-Trento, Italy
{magnini, negri, pianta, romano, speranza, serafini, cgirardi}@itc.it
http://www.itc.it

[2] CELCT,
38050 Povo, Italy
{bartalesi, sprugnoli}@celct.it
http://www.celct.it

Abstract. This paper presents the general objectives of the ONTOTEXT project (From Text to Knowledge for the Semantic Web), and the activities carried out during the first year of its development cycle. First, the task of annotating huge amounts of textual data (e.g. those available on the Web or in local document collections) will be introduced, focusing on its importance in order to enhance the interoperability of such data through ontology-based reasoning. Then, the main issues related to the annotation task will be discussed. These include the choice of an adequate formalism to capture and describe different types of relevant information contained in a text, and the adaptation of existing language-specific markup formalisms to a new language (Italian in our case). Finally, the results of our experience in the concrete annotation of information about people and temporal expressions for the Italian Content Annotation Bank (I-CAB) being developed at ITC-irst and CELCT will be reported.

1 Introduction

ONTOTEXT is a three-year FU-PAT project, started in October 2004, which involves ITC-irst and ISTI-CNR (http://tcc.itc.it/projects/ontotext). The general objective of the project is to study and develop innovative knowledge extraction techniques for producing new or less noisy information to be made available to the Semantic Web.

The most common approach to the creation of the Semantic Web relies on annotations of Web resources with respect to concepts and relations defined in ontologies, which serve as a means for establishing conceptually concise bases for knowledge communication. The semantic annotation of Web Resources is meant to facilitate access to documents both by humans and artificial agents such as web robots.

Unfortunately, the information contained in annotated documents is prone to be inconsistent and very sparse. Moreover it can change over time. As a consequence, its exploitation by both human and artificial agents can be difficult. While adhering to the Semantic Web perspective, ONTOTEXT addresses the semantic annotation problem by integrating into the Semantic Web a new type of information source resulting from a process of ontology-driven knowledge extraction. In our view, knowledge contained in annotated resources is extracted and organized in a structured knowledge base, which allows for consistent representation and updating of the information, while guaranteeing its traceability with respect to the sources. The resulting repository of facts can then be used to refine and extend existing ontologies, which are also made available to the Semantic Web. The enabling power and potentiality for integration of the technologies being developed will be proven through the realization of PEOPLE ON-LINE, a Web portal devoted to information about people mentioned in Trentino's local newspapers, which will allow citizens to consult facts contained in the ONTOTEXT knowledge base through a user-friendly interface.

The first year of the project activities has been mainly focused on the first of the three above mentioned research directions. Our effort addressed the large scale annotation of Italian news documents with semantic information about temporal expressions, different types of entities present in the texts, and relations between such entities. In this direction, we are developing both a manually annotated benchmark, used for training and evaluation, and tools for automatic annotation. Starting from temporal expressions and entities belonging to the person category, one of the primary objectives of the first ONTOTEXT development cycle was to employ a flexible markup language to identify such information in a given source text, and annotate it with additional metadata providing a semantically rich and normalized description.

The availability of a benchmark and of automatic annotation tools is a fundamental asset, both to achieve the project's objectives and under the more general Semantic Web perspective, as they represent an enabling technology for ontology learning. Building on such components, novel ontology learning techniques can be developed to refine and extend an already existing ontology on the basis of the information contained in the annotated document collection. These refinements, which may consist of adding new relations between concepts (e.g. linguistic ontologies, such as WordNet, can be expanded with new IS-A relations automatically discovered), as well as new properties and slot restrictions, will serve to pinpoint new relevant information in the input texts in a process of iterative improvements.

Following a policy of reusing already available markup languages, the annotation activity has been carried out adopting the formalisms developed within the American ACE (Automatic Content Extraction, www.nist.gov/speech/tests/ace) program. At a glance, the ACE standards developed for the Entity Detection and the Time Expressions Recognition and Normalization tasks were perfectly adequate for our purposes, as they allow for a semantically rich and normalized annotation of:

- different types of entities (i.e. objects or set of objects in the world, including persons, organizations, and locations);
- different types of entity mentions (i.e. any textual reference to an entity);
- different types of temporal expressions (ranging from explicit expressions such as "Sunday, March 13 2005", to more implicit ones such as "three days later").

However, due to the differences between English and Italian, part of the work has been dedicated to the revision and the 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 (182,000 words, divided into 524 files with an average length of 384 words) which at present contains annotations about persons (PE) and temporal expressions (TE).

The paper is structured as follows. Section 2 provides relevant background context in Ontology learning and Ontology Population, defining the task addressed in the ONTOTEXT project. Section 3 describes the architecture of the system. Section 4 gives details on the benchmark for Ontology Population we are developing, while Section 5 addresses the automatic recognition and normalization of temporal expressions.

2 Ontology Learning and Population

Within the recent research area on *Ontology-Based Knowledge Extraction*, ONTOTEXT addresses three key research aspects: (i) annotating documents with semantic and relational information, *e.g.* properties and facts in which entities are involved (*Knowledge Markup*); (ii) providing an adequate degree of interoperability of such relational information, with particular attention to the temporal dimension (*Knowledge Extraction*); and (iii) updating and extending the ontologies used for Semantic Web annotation (*Ontology Learning and Population*). 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.

Automatic Ontology Population (OP) from texts has recently emerged as a new field of application for knowledge acquisition techniques (Buitelaar, Cimiano, Magnini 2005). Although there is no a univocally accepted definition for the OP task, a useful approximation has been suggested (Bontcheva and Cunningham, 2003) as Ontology Driven Information Extraction, where, in place of a template to be filled, the goal of the task is the extraction and classification of instances of concepts and relations defined in a Ontology. A similar task has been approached in a variety of similar perspectives, including term clustering (Lin, 1998 and Almuhareb and Poesio, 2004) and term categorization (Avancini et al. 2003).

A rather different task is Ontology Learning (OL), where new concepts and relations are supposed to be acquired, with the consequence of changing the definition of the Ontology itself (Velardi et al. 2005).

In this paper OP is defined in the following scenario. Given a set of terms $T=\{t_1, t_2, ..., t_n\}$ a document collection D, where terms in *T* are supposed to appear, and a set of predefined classes $C=\{c_1, c_2,..., c_m\}$ denoting concepts in an Ontology, each term t_i has to be assigned to the proper class in *C*. At the state of advancement presented in this

paper we assume that (i) classes in *C* are mutually disjoint and (ii) each term is assigned to just one class.

As we have defined it, OP shows a strong similarity with Named Entity Recognition and Classification (NERC). However, a major difference is that in NERC all occurrences of recognized terms have to be classified in one of the classes in C, while in OP it is the term, independently of the context in which it appears, that has to be classified.

While Information Extraction, and NERC in particular, have been addressed prevalently by means of supervised approaches, Ontology Population is typically attacked in an unsupervised way. As many authors have pointed out (e.g. Cimiano, 2005), the main motivation is the fact that in OP the set of classes is usually larger and more fine grained than in NERC (where the typical set includes Person, Location, Organization, GPE, and a Miscellanea class for all other kind of entities). In addition, by definition, the set of classes in C changes as a new ontology is considered, making the creation of annotated data almost impossible practically.

3 System Architecture

This section provides a general description of the main ontological assumptions underlying the ONTOTEXT architecture. The project is interested in extracting information, storing and reasoning about the following ontological categories: Entities, Temporal Objects, Relations, Events, Topics and Opinions. The main focus of the project stands in the relations between such entities and the way they are expressed in texts: as a consequence, we assume two levels of description, one textual, where ontological categories are *mentioned*, and the ontological level where such categories are represented in a knowledge base. The starting point for basic definitions is the work carried out within the ACE (Automatic Content Extraction) program.

Entities denote object or set of objects in the reference domain. ONTOTEXT covers the following entities: (i) Persons, Locations, Organizations, Geo political Entities (GPE). *Temporal entities* denote either points, intervals or durations in a model of time. Textual mentions of entities, including temporal ones, are introduced in detail in Section 4. A *relation* is an ordered pair of entities. Simple kinds of relations are the date of born of a person, the number of inhabitants of a city. Relations belong to a predefined hierarchy of relation types, including, for instance, Part-Whole, Physical Location and Membership. In addition relations are temporally bounded (e.g. Trento has 100.000 inhabitants on a certain date). *Events* are something that happen involving a number of participants and resulting in a change of state. A textual mention for an event is the sentence describing the event, with a trigger word for the event (usually the verb) marked.

Topics are collections of relations and events about the same subject. Examples of topics are *Giro d'Italia 2004*, *Dimissioni di Collina*, *September 11*, *Uragano Katrina*, *Bullismo a Trento*. Topics have a certain degree of activation at a certain time, depending on two factors: (i) the dimension of the topic (i.e. the number of Facts they consist of); (ii) the frequency of the topic. Topics may vary both in dimension (i.e. the

number of relations and events they consist of) and in extension (i.e. the period the topic is "active").

Opinions are subjective judgments expressed about Entities, Relations, Events and Topics. As a first approximation to a richer treatment of opinions we consider two kinds of opinions: positive orientations and negative orientations.

In ONTOTEXT we distinguish between two sources of information: textual sources, that is articles from local newspapers (L'Adige, Corriere del Trentino, Trentino, Vita Trentina); non textual sources, that is available databases with information about people, locations, etc. in the Trentino region (e.g. Annuario del Trentino). One of the objectives of the project is an investigation of the relations between these two sources in the context of an Ontology Population task.

As it is depicted in Figure 1, documents are firstly annotated with linguistic information at several levels (tokens, parts-of-speech, multiwords) and coded in a XML-like format (see Bentivogli et al. 2003 for more details). Then mentions (i.e. textual expressions referring to entities, relations, events, topics and opinions) are marked. Each mention is then associated to an instance of a concept defined in the ONTOTEX ontology. This process, i.e. the Ontology population, involves a *normalization* process, where different mentions denoting the same individual in the world are recognized as co-refering, either at the level of a single document, or at the level of the whole collection. For instance, the two mentions "A. Pacher" and "the first citizen of Trento" denote the same individual in the ONTOTEXT domain, whose first name is "Alberto" and whose family name is "Pacher". At this point, the concepts of the ontology related to the particular instance are available as semantic annotations which enrich the original document, allowing a number of semantic-based retrieval functionalities, including specific web services.

As for the ONTOTEXT ontology, it represents the main ontological categories mentioned at the beginning of this section, defines the assioms over such categories (e.g. the fact that a Person has one and only one date of birth, while she/he may have more than one profession during the life) and allows to reason about them. This is crucial in ONTOTEXT since knowledge is automatically extracted from documents and has to be validated under several respects. Each portion of knowledge (e.g. entity, relation, event, topic) has a degree of confidence, depending on the textual evidence found by the system, the frequency with which it is reported, and the extent to which it meets the ontological requirements.

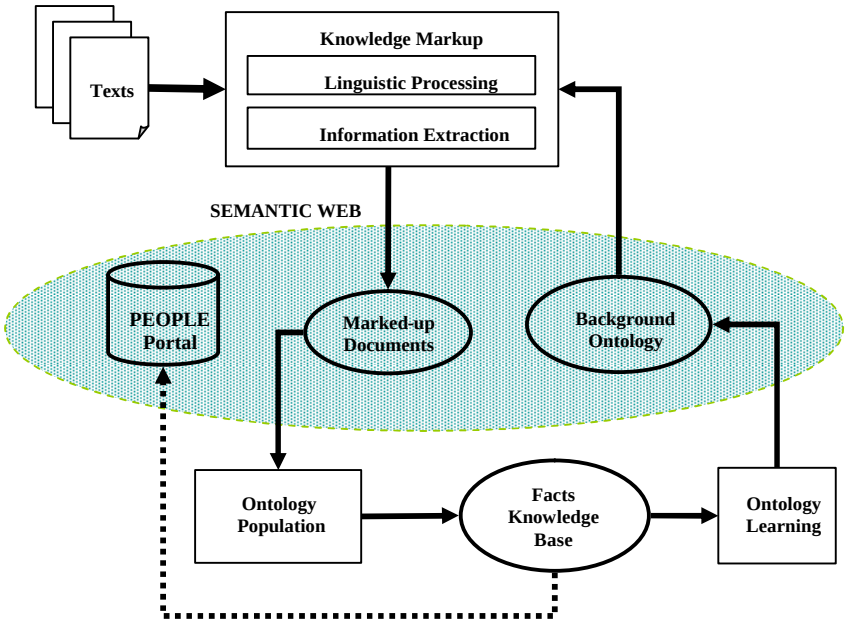

Fig. 1. ONTOTEXT Information flow.

4 The Italian Content Annotation Bank (I-CAB)

In this section we present work in progress for the creation of the Italian Content Annotation Bank (I-CAB), a corpus of Italian news stories annotated with different kinds of semantic information. The annotation is being carried out manually, as we intend I-CAB to become a benchmark for various automatic Information Extraction and Ontology Population tasks, including recognition and normalization of different types of entities, temporal expressions, relations between entities (e.g. the relation resident-in connecting a person to the place where they live), and relations between entities and temporal expressions (e.g. the relation date-of-birth connecting a person to a date).

For the annotation of I-CAB with semantic and relational information, we follow a policy of reusing already available markup languages and therefore we adhere to the formalisms developed within the American ACE program (ACE - Automatic Content Extraction, http://www.nist.gov/speech/tests/ace), which offer a flexible markup language to identify content information in a given source text, and annotate it with additional metadata providing a semantically rich and normalized description.

In particular, the ACE standards developed for the Entity Detection and the Time Expressions Recognition and Normalization tasks allow for a semantically rich and

normalized annotation of different types of entities (e.g. persons, organizations, locations, geo-political entities, etc.) and different types of temporal expressions (e.g. points, durations, etc.). We also follow the guidelines provided by the Linguistic Data Consortium (LDC) in 2004, which differ from those of the ACE Program, for example by providing a more detailed classification of the possible textual realizations of an entity.

So far I-CAB has been annotated with temporal expressions and entities of type person. Manual annotation will be finally delivered in the Meaning Format, a stand-off XML-based annotation scheme conformant to the XCES and TEI corpus annotation standards. The first version of the Meaning Format was developed within the EU-funded MEANING project (Bentivogli et al., 2003). It has now been extended with the aim of representing temporal expressions, entities and relations.

4.1 Modifications to the ACE Guidelines

The adaptation of the ACE annotation schemes to the annotation of Italian texts required some extensions (Lavelli et al., 2005). In particular, the revision of the ACE annotation guidelines has been performed in two main directions: on the one hand, we have modified the ACE guidelines to adapt them to the specific morpho-syntactic features of Italian; on the other hand, we have extended them to include a wider range of entities.

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 both entities and temporal expressions. According the ACE guidelines, in fact, definite and indefinite articles are considered as part of their textual realization, while prepositions are not. As the annotation is word-based, this does not account for Italian articulated prepositions, where a definite article and a preposition are merged. We have decided that this type of prepositions should be included, so as to consistently include all the articles.

Other modifications affected only the annotation of entities, whose structure is generally more language-dependent than the structure of temporal expressions: some new types of possible textual realizations have been introduced (for instance, we have created a specific tag to annotate the clitics whose extension can not be identified at word-level), while some others have been eliminated (for instance, due to the syntactic differences between English, where both adjectives and nouns can be used as pre-modifiers, and Italian, which only admits adjectives in that position, the tag used to mark the pre-modifiers has been removed from the classification).

In extending the range of entities to be annotated, we have decided to include all conjunctions of entities and not only those which share the same modifiers as indicated in the ACE guidelines, and to mark them using the specifically created tag CONJUNCTION. For example, in the phrase ‘George Bush and Bill Clinton’, we have the two person entities (of type PROPER_NAME) as indicated in the ACE guidelines, e.g. *George Bush* and *Bill Clinton*, but also a person entity of the type CONJUNCTION consisting of both individuals.

4.2 Time Expressions Recognition and Normalization

According to the TIMEX2 markup standard (Ferro et al. 2004), markable time expressions (TEs) 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*).

Recognition refers to the task of finding the TEs within a text (detection) and determining their extension (bracketing). Normalization refers to the interpretation of a TE by assigning values to pre-defined normalization attributes:

- *VAL*: contains the value of a temporal expression (e.g. VAL="2004-05-06" and VAL="P6D" for the date *<6 maggio 2004>*/*May 6th, 2004*, and the period *<sei giorni>*/*six days* respectively); no VAL is provided for underspecified TEs (e.g. *<per lungo tempo>*/*for a long time*);
- *MOD*: captures temporal modifiers. Possible value for the attribute MOD are APPROX (*<verso mezzanotte>*/*around midnight*), MORE THAN (e.g. *<più di 3 minuti>*/*more than 3 minutes*) and START (e.g. *<i primi anni '70>*/*the early 1970s*);
- *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 holidays for two months* is normalized as follows: 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 times (e.g. *<ogni giorno>*/*every day*).

4.3 Annotation of entities of type person

The annotation of entities of type person, as indicated in the ACE Entity Detection task for all the different types of 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.

As it often happens that the same entity is mentioned more than once in the same text, two inter-connected levels of annotation have been 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 the textual references to that object. For instance, the entity *George W. Bush* (e.g. the individual in the world who is the current president of the U.S.) can be referenced to in many different ways in the same text: e.g. *the president of the U.S.A.*, *Bush*, *the president*, *he*, *George W. Bush*, etc.

Four classes describes the kinds of reference entities make to something in the world: (i) we have specific referential entities when the entity being referred to is a unique object or set of objects (e.g. *<Il* [*presidente*] *della ditta> non è presente*/*The*

company president is not present); (ii) generic referential entities refer to a kind or type of entity and not to a particular object (or set of objects) in the world (e.g. *<Il* [*presidente*]*> viene eletto ogni 5 anni*/*The president is elected every 5 years*); (iii) under-specified referential entities are non-generic non-specific references, including imprecise quantifications (e.g. <[*tutti*]>/*everyone*) and estimates (e.g. *<oltre 10.000* [*persone*]>/*more than 10.000 people*); and (iv) negatively quantified entities refer to the empty set of the mentioned type of object (e.g. *<Nessun* [*avvocato*]>/*No lawyer*).

The different types of entities (e.g. persons, organizations, locations, geo-political entities, etc.) can be divided into subtypes. In the specific case of Person Entities (PE), we have three subtypes: (i) Individual PEs refer to a single person (*George W. Bush*), (ii) Group PEs refers to more than one person (*my parents*, *your family*, etc.), and (iii) a PE is classified as Indefinite when it is not possible to judge from the context whether it refers to one or more persons (*I wonder who came to see me*).

Textual realizations of entities, i.e. entity mentions, can be intuitively described as portions of text; the extent of this portion of text is defined as the entire nominal phrase used to refer to an entity, thus including modifiers (e.g. *<una grande* [*famiglia*]>/*a big family*), 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*).[1]

The classification of entity mentions is based on syntactic features; among the most significant LDC categories (including those created specifically for I-CAB) are:

- NAM: proper names (e.g. <[*Ciampi*]>)
- NOM: nominal constructions (e.g. *<i* [*bambini*] *buoni*>/*good children*)
- PRO: pronouns, e.g. personal (<[*tu*]>/*you*) and indefinite (<[*qualcuno*]>/*someone*)
- WHQ: wh-words, such as interrogatives (e.g. <[*Chi*]> *è lì?*/*Who is there?*)
- PTV: partitive constructions (e.g. <[*alcuni*] *di loro*>/*some of them*)
- APP: appositive constructions (e.g. <[*Dante, poeta famoso*]>/*Dante, famous poet)*
- CONJ: conjunctions (e.g. <[*la madre e il bambino*]>/*the mother and the child*)[2]
- ENCLIT: clitics (e.g. <veder[*lo*]>/*to see him*)

4.4 Quantitative Data

The main result of the manual annotation is represented by the first release of I-CAB, which consists of 525 news documents (182,564 words, with an average of 348 words per file) taken from local newspapers and grouped in five categories: News Stories, Cultural News, Economy News, Sports News and Local News. I-CAB is further divided into the training section (335 documents for a total of 113,634 words) end the test section (190 documents, 68,930 words).

The total number of annotated temporal expressions is 4,553 (2,901 and 1,652 in the test and training sections respectively). As shown in Table 1, the number of points is slightly lower that the number of durations. Among the normalization attributes, ANCHOR_DIR and ANCHOR_VAL are the most frequent (cfr. Table 2).

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

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

Table 1. Occurrences and percentage of points, durations and temporal expressions with no value

	Training	Test	Training + Test
Points	1553 (53.53 %)	796 (48.18 %)	2349 (51.59 %)
Durations	1207 (41.61 %)	738 (44.67 %)	1945 (42.72 %)
TEs with no value	141 (4.86 %)	118 (7.14 %)	259 (5.69 %)
TOTAL	2901	1652	4553

Table 2. Occurrences and percentages of set of times, TEs with a temporal modifier and anchored durations

	Training	Test	Training + Test
MOD attribute	112 (3.86 %)	76 (4.60 %)	188 (4.13 %)
SET attribute	121 (4.17 %)	51 (3.09 %)	172 (3.78 %)
ANCHOR_DIR	696 (24.00 %)	362 (21.91 %)	1058 (23.24 %)
ANCHOR_VAL	696 (24.00 %)	362 (21.91 %)	1058 (23.24 %)

As far as persons are concerned, the corpus contains a total number of 7,087 entities (on average, 13.5 entities per document) and 16,059 mentions (30.6 mentions per document). On average, an entity is mentioned 2.3 times in a document. The distribution between training and test is as follows: 4,459 entities and 9,994 mentions in the training, 2,628 entities and 6,065 mentions in the test.

As shown in Table3, the majority of person entities are referential (almost 80% of the total), whereas the distribution by subtypes is more balanced (see Table 4).

Table 3. Distribution of person entities by entity class

	Training	Test	Training + Test
SPC	3474 (77.89 %)	2142 (81.51 %)	5616 (79.24 %)
USP	517 (11.59 %)	263 (10.01 %)	780 (11.01 %)
GEN	443 (9.93 %)	213 (8.10 %)	656 (9.26 %)
NEG	25 (0.56 %)	10 (0.38 %)	35 (0.49 %)
TOTAL	4459	2628	7087

Table 4. Distribution of person entities by subtypes

	Training	Test	Training + Test
Individual	2067 (46.36 %)	1256 (47.79 %)	3323 (46.89 %)
Group	1995 (44.74 %)	1206 (45.89 %)	3201 (45.17 %)
Indefinite	397 (8.90 %)	166 (6.32 %)	563 (7.94 %)
TOTAL	4459	2628	7087

4.5 Inter-annotator Agreement

Inter-annotator agreement has been evaluated on the dual annotation of a corpus of randomly chosen news stories (ten for TEs and ten for PEs, for a total of about 5204 and 4657 words respectively).
For temporal expressions we have used the kappa statistic (Cohen 1960) to measure the agreement between the annotators in determining whether each token is or is not part of any TE, and we have obtained a k=0.958. However, as this measure does not take into account the extent of the annotated TEs, we have also compared the two annotated versions using the standard measures of recall and precision to compute the F-measure. In fact, unlike precision and recall which depend on which of the two versions is considered correct, F-measure has the same value if computed in terms of A against B or vice-versa. F measure for TE detection is 0.955 and F-measure for TE bracketing is 0.931.

Agreement in normalization (e.g. the assignment of attribute values) has been measured on the TEs uniformly bracketed. Table 5 reports, for each attribute, the percentage of cases where the annotators assign the same value and, for the attributes which admit a restricted number values, it also reports the kappa statistic.

As observed in (Di Eugenio, Glass 2004) the kappa statistic could be affected by *bias* and *prevalence* problems. Calculating also the statistic according to the (Siegel, Castellan 1988) definition we verified there is no *bias* problem (values are equal), but the natural skewing of the distribution of categories affect kappa statistic (i.e. for the SET attribute).

As far a person entities are concerned, the agreement in terms of F-measure is 0.912 for entity detection and 0.870 for the identification of the mention extent.

Table 5. Attribute value assignment agreement

	Percentage	kappa statistic
VAL	142/154 (92.2%)	-
MOD	153/154 (99.3%)	0.886
SET	152/154 (98.7%)	0.744
ANCHOR_VAL	142/154 (92.2%)	-
ANCHOR_DIR	139/154 (90.3%)	0.749

5 Automatic Annotation of Temporal Expressions: Ita- Chronos

As a first step towards the development automatic annotation tools, the ONTOTEXT activities have been focused on Temporal Expressions (TEs). The resulting system, CHRONOS has been designed for the twofold task of recognizing and normalizing a broad variety of TEs which can be found within an input written text according to the TIMEX2 markup standard.

5.1 System Architecture

Both the detection/bracketing and normalization subtasks have been addressed following a rule-based approach. The rule-based framework allows for achieving good performance results even in absence of vast amounts of annotated material, which represent the main shortcoming in the exploitation of machine learning algorithms for automatic annotation purposes. Moreover, the cost required for manually creating a large number of rules is compensated by the transparency of the resulting system, which can be easily modified and maintained due to the declarative structure of the rules.

The architecture of CHRONOS, depicted in Figure1, relies on three main components (i.e. the preprocessing component, the detection/bracketing component, and the normalization component), which sequentially carry out the linguistic analysis of the input text, the identification of all the TEs it contains, and their complete annotation in compliance with the TIMEX2 standard. The system has been implemented in a highly modular way, where each set of rules is specialized for dealing with a specific aspect of the annotation.

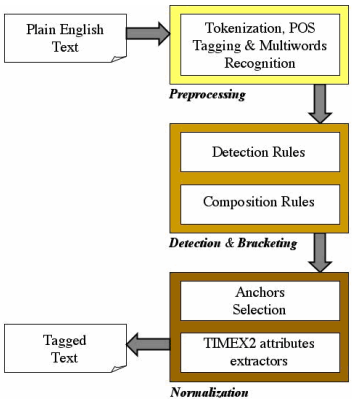

Fig. 2. Architecture of CHRONOS

5.1.1 Preprocessing

The preprocessing component is in charge of the linguistic analysis of the input text, which is tokenized and words are disambiguated with their lexical category by means of a statistical part of speech tagger. Also multiwords recognition is carried out at this stage: multiwords expressions are recognized considering a list of about five thousand multiwords (i.e. collocations, compounds, and complex terms) that have been automatically extracted from WordNet (Fellbaum, 1998).

5.1.2 Detection and Bracketing

All the TEs present in the text are pinpointed, and their extension determined by means of a set of approximately 300 handcrafted rules. These are regular expressions checking for different features of the input text, such as the presence of particular word senses, lemmas, parts of speech, symbols, or strings satisfying specific predicates.

As for detection, markable expressions are discovered considering the presence in the text of lexical triggers, i.e. words or particular configurations of numeric expressions that convey a meaning related to the concepts of time, date, and duration. Possible triggers considered by the system include: (i) nouns (e.g. *<giorno>*/*day* , *<stagione>*/*season*), (ii) proper names (e.g. *<agosto>*/*August*), (iii) adverbs (e.g. *<quotidianamente>*/*daily*), and (iv) numeric expressions (e.g. *<08-11-2004>*, *<12:30>*).

As for bracketing, extent recognition is carried out looking at the context surrounding the detected lexical triggers. To this aim, relevant information considered by the system is represented by: (i) nouns (e.g. *<inizio>*/*beginning*), (ii) adjectives (e.g. *<prossimo>*/*next*), (iii) adverbs (e.g. *<prima>*/*before*), (iv) prepositions (e.g. *<durante>*/*during*), and (v) numbers (e.g. *3*, *cinque*/*five*). Moreover, at this stage a set of composition rules is in charge of resolving conflicts that may arise between possible multiple taggings. Such conflicts may occur when a recognized TE contains, overlaps, or is adjacent to one or more other detected TEs. As an example, given the sentence *sabato scorso all'alba*/*Last Saturday, at dawn*, the basic rules application phase recognizes the following three time expressions: *<sabato>*/*Saturday*, *<sabato scorso>*/*Last Saturday*, *<sabato scorso all'alba>*/*Last Saturday, at dawn*. Simple composition rules considering the start/end position of the tags are used to deal with these conflicts.

5.1.3 Normalization

For a complete TIMEX2 annotation of each detected TE, the appropriate values are assigned to the TIMEX2 attributes by 8 different sets of extraction rules. Each set of rules is specialized to capture specific features of a TE. In particular:

- the MOD extractor checks for the presence of modifiers (e.g. *early*, *approximately*) within the boundaries of a TE, using this information to fill the MOD attribute;
- the SET extractor checks for the presence of expressions denoting sets of times (e.g. every, twice a) using this information to fill the SET attribute;
- the ANCHOR_DIR extractor checks for the presence of clues (e.g. *before*, *later*) concerning the most likely ANCHOR_DIR value for relative time expressions.

Often, however, the superficial form of a time expression does not provide enough information for a correct normalization. For instance, the VAL attribute of anaphoric time expressions such as *<la settimana scorsa>*/*last week* or *<domani all'alba>*/*tomorrow at dawn* cannot be determined simply by considering the triggers and their modifiers. In these cases, some degree of reasoning considering the information provided by the lexical context in which the expressions occur is necessary. For this reasoning purpose, the following extractors have been implemented:

- TYPE, which is in charge of determining if a TE is absolute (i.e. it can be normalized in virtue of its superficial form) or anaphoric (i.e. it refers to another date previously mentioned in the text). For this purpose, such extractor checks for the pres-

ence of words (e.g. <*scorso*>/*last*, <*prossimo*>/*next*) denoting anaphoric TEs, or particular sequences of words and numbers (e.g. *<1 Maggio 2005>*/*the first of May 2005*, *01/05/2005*) denoting absolute TEs.

- T-CAT, which determines the granularity of anaphoric time expressions. Possible values of the T-CAT attribute are obtained mapping the detected lexical triggers to the categories [*second*, *minute*, *hour*, *day*, *month*, ..., *millennium*]. This information is used to determine the correct temporal anchor of an anaphoric TE. For instance, the T-CAT attributes associated to <*tre anni fa*>/*three years ago* and <*sabato scorso*>/*last Saturday* will be year and day respectively. Using such information, the normalization component selects anchors with the same granularity to fill their VAL attributes.
- HEUR, which select the appropriate anchor selection strategy for the resolution of each anaphoric TE. The current version of the system carries out the anchors selection process following two main strategies: CR-DATE and PR-DATE. The CR-DATE heuristic associates to an anaphoric TE the document's creation date found at the beginning of the document. The PR-DATE heuristic takes as anchor the value of the nearest previous absolute time expression with a compatible granularity. According to this granularity constraint, the selected anchor must have the same or a higher degree of specificity with respect to the anaphoric expression.
- OP, which determines, for each anaphoric TE, the operator to be applied for the calculation of its final VAL. Such operator can be "+", "-", or "=". For instance, the OP value assigned to the relative time expressions tre anni dopo/three years after, due settimane fa/two weeks ago, and Oggi/today will be "+", "-", and "=" respectively.
- QUANT, which determines the quantity that has to be added or subtracted to the anchor for the calculation of the final VAL of an anaphoric time expression. Such quantity is expressed by an integer (n=0) assigned to the QUANT attribute. For instance, the QUANT attributes assigned to the time expressions reported in the previous examples will be filled with "3", "2", and "0" respectively.

5.2 Evaluation

CHRONOS has been evaluated over the I-CAB-*temp* test corpus; Table 6 reports the results achieved by the system, calculated with the scorer used in the framework of the TERN-2004 evaluation exercise (http://timex2.mitre.org/tern.html).

The first two columns, POSS and ACT, report the number of items in the reference (POSS= CORR + INCO + MISS) and the number of items in the system output (ACT= CORR + INCO + SPUR). The number of correct (CORR), incorrect (INCO), missing (MISS), and spurious (SPUR) items is also reported, both in terms of detection/bracketing (TIMEX2 and TIMEX2:TEXT rows), and in terms of normalization capabilities (all the other rows). Finally, the overall system's performance is summarized by the precision (PREC), recall (REC), and F-measure (F) scores reported in the last three columns of the table.

Table 6. System's performance calculated over the I-CAB test

TAG	POSS	ACT	CORR	INCO	MISS	SPUR	PREC	REC	F
TIMEX2	2638	2590	2396	0	242	194	0.925	0.908	0.917
TIMEX2 TEXT	2638	2590	2225	171	242	194	0.859	0.843	0.851
ANCHOR_DIR	522	479	351	56	145	72	0.733	0.636	0.681
ANCHOR_VAL	513	479	237	132	144	110	0.495	0.462	0.478
MOD	97	97	90	0	7	7	0.928	0.928	0.928
SET	106	86	53	0	53	33	0.616	0.500	0.552
VAL	2229	2359	1501	696	32	162	0.636	0.673	0.654

6 Conclusions

We have presented ONTOTEXT, a project aiming at investigating the relations between knowledge as it is expressed in documents and as it is coded in ontologies.

The starting point was work already carried out in the ACE program. However, we have extended the ACE work under two relevant aspects: first, we have adapted the task definition from English to Italian, introducing a number of modifications; second, we have extended the task itself which now includes the association between mentions and instances in the knowledge base, making it possible to evaluate not only the information extraction task (as defined within the computational linguistic community) but also the population of ontologies.

A relevant result at the current state of advancement of the project is the realization of a benchmark for Ontology Population from texts. As for our knowledge, this is the first resource which fully describes the whole task, from linguistic expressions in texts (i.e. mentions) to instances of concepts in the knowledge base. The availability of the benchmark makes it possible to compare different systems and to evaluate different aspects (e.g. precision and recall) of their performance.

Finally, we have reported on the automatic recognition and normalization of temporal expressions, as a first step towards recognition and normalization of other types of entities, including Persons, Organizations, Locations, and Geo-Political entities.

References

1. Almuhareb, A., Poesio, M.: Attribute-based and value-based clustering: An evaluation. In: Proceedings of EMNLP 2004, Barcelona (2004) 158-165
2. Avancini, H., Lavelli, A., Magnini, B., Sebastiani, F., Zanoli, R.: Expanding Domain-Specific Lexicons by Term Categorization. In: Proceedings of SAC 2003 (2003) 793-79
3. Bentivogli, L., Girardi, C., Pianta, E.: The MEANING Italian Corpus. In: Proceedings of the Corpus Linguistics 2003 conference, Lancaster (2003)
4. Bontcheva, K., Cunningham, H.: The Semantic Web: A New Opportunity and Challenge for HLT. In: Proceedings of the Workshop HLT for the Semantic Web and Web Services at ISWC 2003, Sanibel Island (2003)

5. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam (2005)
6. Cimiano, P., Pivk, A., Thieme, L.S., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam (2005)
7. Cimiano, P., Volker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Proceedings of RANLP'05, Borovets, Bulgaria (2005) 166-172
8. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, New York 20:37-46 (1960)
9. Di Eugenio, B., Glass, M., The kappa statistic: A second look. Computational Linguistics, 30(1):95-101 (2004)
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
11. Fleischman, M., Hovy, E.: Fine Grained Classification of Named Entities. In: Proceedings of COLING 2002, Taipei (2002)
12. Fleischman, M.: Automated Subcategorization of Named Entities, 39th Annual Meeting of the ACL, Student Research Workshop, Toulouse (2001)
13. Hearst, M.: Automated Discovery of WordNet Relations. In: WordNet: An Electronic Lexical Database, MIT Press (1998) 131-151
14. Lavelli, A., Magnini, B., Negri, M., Pianta, E., Speranza, M., Sprugnoli, R.: Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 1.0.). Technical Report T-0505-12, ITC-irst, Trento (2005)
15. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of COLING-ACL98, Montreal, Canada (1998)
16. Lin, D.: Dependency-based Evaluation of MiniPar. In: Proceedings of Workshop on the Evaluation of Parsing Systems, Granada (1998)
17. Schlobach, S., Olsthoorn, M., de Rijke, M.: Type Checking in Open-Domain Question Answering. In: Proceedings of ECAI 2004, Valencia (2004)
18. Sidney, S., Castellan, N. J.: Non parametric statistics for the behavioural sciences. McGraw Hill, Boston, MA. (1988)
19. Szpektor, I., Tanev, H., Dagan, I., Coppola, B.: Scaling Web-based Acquisition of Entailment Relations. In: Proceedings of EMNLP 2004, Barcelona (2004)
20. Velardi, P., Navigli, R., Cuchiarelli, A., Neri, F.: Evaluation of Ontolearn, a Methodology for Automatic Population of Domain Ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam (2005)