TRATTAMENTO DI ESPRESSIONI TEMPORALI IN ITALIANO: ITA-CHRONOS

DEALING WITH ITALIAN TEMPORAL EXPRESSIONS: THE ITA-CHRONOS SYSTEM

MATTEO NEGRI

SOMMARIO/ ***ABSTRACT***

Questo articolo descrive il sistema ITA-Chronos sviluppato da FBK-irst, riportando i risultati da esso ottenuti nel task "Temporal Expressions Recognition and Normalization" a EVALITA 2007.

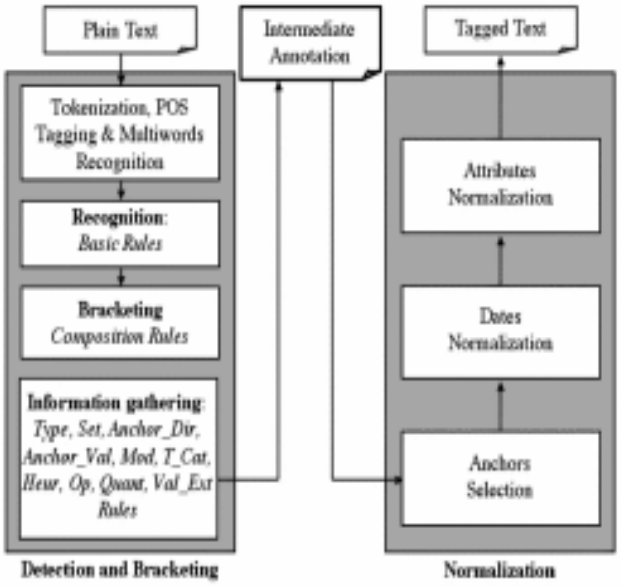
*This paper presents ITA-Chronos, the system developed at FBK-irst to participate in the "Temporal Expressions (TE) Recognition and Normalization Task" at EVALITA 2007. ITA-Chronos adopts a rule-based approach, with different sets of hand-crafted rules specialized to deal with different aspects of the problem. The system (**FBKirst_Negri_TIME**) achieved good results both in recognition (TERN-Value: 85,7%) and normalization (TERN-Value: 61,9%), ranking 1st in both the sub-tasks.*

1. Introduction

ITA-Chronos is the Italian extension of Chronos [2], a multilingual system written in Lisp, originally developed for the English language. The system is designed to recognize all the TEs occurring in a text, identify their extension, and normalize them according to the TIMEX2 standard [1].

Like all the other state-of-the-art systems[1] addressing the recognition/normalization task, Chronos relies on a rule-based approach. The architecture of the system, depicted in Figure1, is based on two main components: the *detection and bracketing* component, and the *normalization* component. The **detection and bracketing** component is in charge of *i)* performing the linguistic analysis of the input text, *ii)* recognizing the correct extension of the TEs it contains, and *iii)* gathering relevant information, to be used in the following normalization phases. Steps *ii)* and *iii)* are performed by applying different sets of rules specialized to deal with specific aspects of the problem. The output of the detection and bracketing component is an intermediate annotation, which is used by the **normalization** component to assign correct values to the TIMEX2 attributes of each detected TE.

Figure 1: System architecture



2. ITA-Chronos Rules

We distinguish between two levels of rules: the *tagging rules* and the *composition rules*. At a lower level, **tagging rules** are regular expressions that check for different features of the input text. These can be the presence of particular word senses, lemmas, parts of speech, symbols, or strings satisfying specific predicates. An example of rule is reported in Table1. This rule matches any POS tagged sequence of three words ("t1 t2 t3"), where:
- "t1" is recognized as a preposition;
- "t2" is a number;

1 The English version of the system, evaluated in the framework of the TERN 2004 evaluation exercise (http://timex2.mitre.org/tern.html), ranked 2nd. Our TERN-Value (76%) was very close to the top score achieved by the rule-based Lockheed Martin system (78%).

- "t3" is a *lexical trigger* satisfying the predicate "*TimeUnit-p*"[2].

For instance, it will match relative[3] time expressions and durations such as "*Fra tre giorni*", or "*Da sei anni*".

Table 1: A rule matching with "*Fra tre giorni*"

PATTERN	t1 t2 t3
T1	[pos = "E"]
T2	[pos = "N"]
T3	[pred = TimeUnit-p]
OUTPUT	<TIMEX2> t1 t2 t3 <TIMEX2>

At a higher level, **composition rules** are in charge of handling conflicts 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 "*Ogni sabato mattina*", the tagging rules application phase recognizes the following four TEs: "*Ogni sabato*", "*sabato*", "*sabato mattina*", and "*Ogni sabato mattina*". Simple composition rules considering the start/end position of the tags are used to deal with these situations.

3. Detection, Bracketing, and Normalization

While the total amount of composition rules used by ITA-Chronos is less than 10, the system makes a large use of tagging rules, which are 981 in total[4]. Of them, 409 are in charge of detecting all possible markable expressions present in the input text. Once TEs have been marked and the composition rules have been run to determine their actual extension, 572 rules divided into 10 specialized sets are run to determine the value of all the TIMEX2 attributes. For instance, for each detected TE, 26 *SET* rules are used to recognize expressions denoting sets of times (e.g. "*ogni*", "*tutti i*"), and 55 *MOD* rules are used to identify temporal modifiers (e.g. "*quasi*", "*circa*", "*fine*"). A crucial role is played by the sets of rules *TYPE* (115 rules), *HEUR* (35), OP (22), *GRANULARITY* (64), and *QUANT* (82), which provide the intermediate annotation with all the information required in the following normalization phase. *TYPE* rules are used to distinguish between durations, absolute, and relativeTEs. For each relative TE, the other sets of rules are respectively used to determine: *i*) the appropriate anchor (the document creation date or the nearest previously recognized TE), *ii*) the operator ("+", "-", "=") that has to be applied to calculate the value with respect to the selected anchor, *iii*) the time unit that has to be added or subtracted to the value of the anchor (e.g. "months", "years", "days"), and *iv*) the quantity that has to be added or subtracted to the anchor. For instance, given the recognized TE "*Tre mesi dopo*", the application of these rule sets will result in: [TYPE=*Relative*, HEUR=*PreviousDate*,OP=+, GRANULARITY=*Months*, and QUANT=*3*].

The normalization component takes as input the intermediate annotation, assigns values to each attribute of the detected TEs, and produces an output tagged text compliant with the TIMEX2 annotation standard.

4. Results

ITA-Chronos ranked 1st in both the sub-tasks of EVALITA 2007. Results are reported in Table2. They are obtained in 27'15" of computation time on a Sun Blade 1500 - 1.5 GHz UltraSPARC IIIi processor.

Table 2: ITA-Chronos (**FBKirst_Negri_TIME**) results

TASK	**Value**	**Prec.**	**Rec.**	**F-Measure**
Rec	85.7	95.7	89.9	92.6
Rec+Norm	61.9	68.5	66.3	67.4

Acknowledgements

This work has been partially supported by the three-year project ONTOTEXT (http://ontotext.itc.it), funded by the Provincia Autonoma di Trento. Many thanks to Lorenza Romano for her support in output format conversions.

REFERENCES

[1] L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical Report, MITRE, 2005.

[2] M. Negri, and L. Marseglia. Recognition and Normalization of Time Expressions. ITC-irst at TERN 2004. Technical Report, ITC-irst, Trento, 2004.

[2] Lexical triggers are words or particular configurations of numeric expressions that convey a meaning related to the concepts of time, date, and duration. Possible triggers satisfying the TimeUnit-p predicate are nouns such as "*minuto*", "*giorno*", "*mese*", "*anno*".

[3] Relative TEs are those expressions that have to be determined with respect to an *anchor date*, such as "*Tre mesi dopo*", or "*Il prossimo anno*".

[4] It's worth mentioning that, even though the number of rules is high, the required effort to write them is relatively moderate. A trained developer is in fact capable of writing 20-30 rules per hour.

CONTACT

MATTEO NEGRI
FBK-irst, Via Sommarive 18, Povo (Trento)
Email: negri@itc.it

MATTEO NEGRI graduated in Philosophy of Language at the University of Turin. Since 2000 he has been researcher at ITC-irst in the TCC Division. At present, his research is focused on the development of multilingual QA systems, within the EU funded project QALL-ME.