Automatic resolution rule assignment to multilingual Temporal Expressions using annotated corpora

E. Saquete P. Martínez-Barco R. Muñoz
gPLSI
DLSI. UA
Alicante, Spain
{stela,patricio,rafael}@dlsi.ua.es

M. Negri M. Speranza
ITC-irst
Povo (TN), Italy
{negri,manspera}@itc.it

R. Sprugnoli
CELCT
Trento, Italy
sprugnoli@celct.it

Abstract

The knowledge-based system TERSEO was originally developed for the Recognition and Normalization of temporal expressions in Spanish and then extended to other languages: to English first, through the automatic translation of the temporal expressions, and then to Italian, applying a porting process where the automatic translation of the rules was combined with the extraction of expressions from an annotated corpus.

In this paper we present a new automatic porting procedure, where resolution rules are automatically assigned to the temporal expressions that have been acquired in a new language, thus eliminating the need for automatic translation and consequently minimizing the errors produced. This is achieved by exploiting the rules of the temporal model, which are language independent, and the information extracted from the annotated corpus. Evaluation results of the updated version of TERSEO for English show a considerable improvement in recognition performance (+ 14% F-Measure) with respect to the original system.

1 Introduction

Recently, the Natural Language Processing community has become more and more interested in developing language independent systems, in an effort to break the language barrier hampering their application in real use scenarios. Such a strong interest towards multilinguality is demonstrated by the growing number of international conferences and initiatives putting systems' multilingual/cross-language capabilities among the hottest research topics. Among these, for instance, the European Cross-Language Evaluation Forum[1] (CLEF) is a successful evaluation campaign which aims at fostering research in different areas of multilingual information retrieval. Started in 2000, the CLEF initiative has rapidly grown over the years in terms of tasks, languages covered, and participating systems. At the same time, in the temporal expressions recognition and normalization field, systems featuring multilingual capabilities have been proposed. Among others, [7], [15] and [8] emphasized the potentialities of such applications in different information retrieval related tasks.

As in many other NLP areas, research in automated temporal reasoning has recently seen the emergence of machine learning approaches trying to overcome the difficulties of extending a language model to other languages [1, 4]. In this direction, the outcomes of the first Time Expression Recognition and Normalization Workshop (TERN 2004[2]) provide a clear indication of the state of the field. In spite of the good results obtained in the *recognition* task, *normalization* by means of machine learning techniques still shows relatively poor results with respect to rule-based approaches, and still remains an unresolved problem.

The difficulty of porting systems to new languages (or domains) affects both rule-based and machine learning approaches. With rule-based approaches [12, 3], the main problems are related to the fact that the porting process requires rewriting from scratch, or adapting to each new language, large numbers of rules, which is a costly and time-consuming process. Machine learning approaches [13, 5], on the other hand, can be extended with little human intervention through the use of language corpora. However, the large annotated corpora that are necessary to obtain high performance are not always available. In this paper we describe a new procedure to build temporal models for new languages, starting from previously defined ones. While still adhering to the rule-based paradigm, its main contribution is the proposal of a simple, but effective, methodology to automate the porting of a system from one language to another. To accomplish this, we take advantage of

[1] http://www.clef-campaign.org/

[2] http://timex2.mitre.org/tern.html

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

the architecture of an existing system developed for Spanish (TERSEO, see [11]) , where the recognition model is language-dependent but the normalizing procedure is completely independent. In this way, the approach is capable of automatically learning the recognition model, adjusting the set of normalization rules.

The paper is structured as follows: Section 2 provides a short overview of TERSEO; Section 3 describes the automatic extension of the system to other languages using automatic translation of the expressions; Section 4 presents the new procedure to automatically assign normalization rules to new temporal expressions using annotated corpora; Section 5 shows the results of evaluation experiments performed on this automatic assignment, and finally Section 6 compares the performance of TERSEO before and after adding the new procedure of using this automatic assignment of resolution rules with Chronos, which is a language-specific system.

2 Existing versions of TERSEO

2.1 TERSEO for Spanish: architecture

At first, TERSEO was developed in order to automatically recognize temporal expressions (TEs) appearing in a Spanish written text, and normalize them according to the temporal model proposed in [9], which is compatible with the ACE annotation standards for temporal expressions [2]. In this formalism TEs are annotated with a TIMEX2 XML tag (recognition) and, for each TE, values are assigned for a set of attributes (normalization). The meaning of the attributes, which will be evaluated in Section 6, is explained as:

- VAL: contains the value of a TE (e.g. VAL="2004-05-06" for "May 6th, 2004")
- ANCHOR_VAL: Contains a normalized form of an anchoring date-time.
- ANCHOR_DIR: Captures the relative direction-orientation between VAL and ANCHOR_VAL.
- MOD: Captures temporal modifiers (possible values are approximately, more than, less then)
- SET: Identifies expressions denoting sets of times (e.g. "every year").

The first step of the system (recognition) includes a preprocessing of the input texts, which are tagged with lexical and morphological information that are given as input to the temporal parser. The temporal parser is implemented using an ascending technique (chart parser) and is based on a temporal grammar.

Once the parser has recognized the TEs in an input text, in the second step they are resolved by the normalization unit, which updates the value of the reference according to the date they refer to, and generates the XML tags for each expression. The normalization unit uses an inference engine in order to resolve all the deleted temporal expressions. This inference engine exploits a centralized unit (TER-ILI unit) that contains a set of general resolution rules, as is shown in Figure 1. Unlike the rules used in the recognition phase, the resolution rules are language independent and will be common for all the sets of temporal expressions in any multilingual extension of TERSEO.

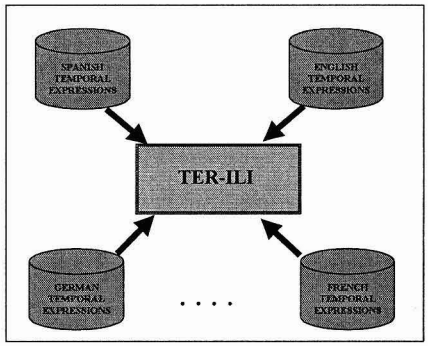

Figure 1. Multilingual TERSEO

Moreover, TERSEO has been extended to other languages with the automatic creation of temporal models for new languages starting from previously defined ones, so as to overcome the problems that are inherent in the rule-based paradigm. With the rule-based approach, the porting implies a big effort due to the necessity of rewriting rules from scratch.

2.2 Original English version of TERSEO

In a first experiment, for English language, the model was obtained automatically from the Spanish one, through the automatic translation of the Spanish temporal expressions into English. The development of the procedure for the automatic porting required 1 person during 1 week in order to developed the software necessary to perform it and less than an hour to obtain the new model using this new application. This platform can be reused to any language without any modification. The resulting system for the recognition and normalization of English TEs obtained good results both in terms of precision (P) and recall (R) [10].

2.3 Porting of TERSEO to Italian

In the case of Italian, we developed a new procedure which exploits both the Spanish and the English models al-

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

ready available and an Italian corpus annotated with temporal expressions. The software required to obtained new expressions from an annotated corpora was developed in less than 1 day and can be used to extract new temporal expressions in any language without any modification.

The reason for considering both models is the fact that they complement each other: on the one hand, the Spanish model was obtained manually and showed high precision values in detection (88%); on the other hand, although the English model showed lower precision results in detection (77%), the on-line translators from English to Italian form better than translators from Spanish to Italian. combined exploitation of two models works as follows

- An Italian corpus with temporal annotations is cessed in order to collect a set of Italian tempora pressions. The selected corpus belongs to the t ing part of the Italian Content Annotation Banl CAB)[6].

- Each single Italian TE is related to the appropriat ready existing normalization rule. In order to do all the expressions are first translated both into En; and into Spanish[3]. Then, the normalization rule lated to the translated expressions are taken into sideration. If both the Spanish and English express are found in their respective models in relation tc same normalization rule, then this rule is assigned to the Italian expression. When only one of the tı lated expressions is found in the existing model: normalization rule is assigned. In case of discre cies, i.e. if both translated expressions are found not coinciding in the same normalization rule, then it has been chosen to prioritize the Spanish rules, as the Spanish model was obtained manually and showed a higher precision. In other cases, the expression is reserved to a manual assignment.

- The set of Italian temporal expressions is augmented by automatically translating into Italian the Spanish and English TEs. In this case, a filtering module has been developed to guarantee the correctness of the new Italian TEs by searching the web with Google[4]: if the translated expression is not found by Google it is given up, otherwise it is included in the model, and related to the same normalization rule assigned to the Spanish or English temporal expression.

The entire translation process has been completed with an automatic generalization process, oriented to obtain generalized rules from the concrete cases that have been collected from the corpus. This generalization process has a double effect. On the one hand, it reduces the number of recognition rules. On the other hand, it allows to identify new expressions that were not previously learned; for instance, the expression *"Dieci mesi dopo"* (i.e. "Ten months later") can be recognized as a generalization of *"Nove mesi dopo"* (i.e. "Nine months later").

The multilingual extension procedure (figure 2) was carried out in three phases:

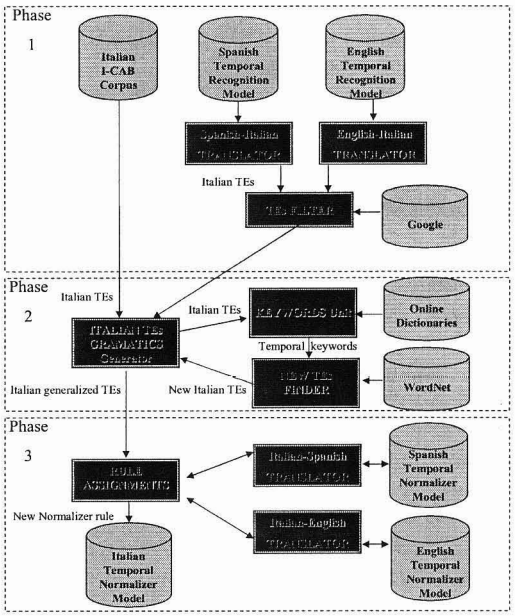

Figure 2. Multilingual extension procedure.

- Phase 1: TE Collection. The Italian temporal expressions are collected (i) from the I-CAB Corpus and (ii) with the automatic translation into Italian of the sets of Spanish and English TEs (and subsequent filter through Google).

- Phase 2: Phase 2: TE Generalization. In this phase, the TEs Gramatics Generator uses the morphological and syntactical information from the collected TEs to automatically generate the grammatical rules that generalize the recognition of the TEs. These grammatical rules allow TERSEO to ignore the information from the text that is not susceptible to be a temporal expression. In this step, the information that has to be analyzed and resolved is being reduced by the system. In addition to this, the keyword unit extracts a list of temporal keywords and, after augmenting them with their synonyms in EuroWordNet [14], uses them to extract

[3]The on-line machine translation systems used are InterTran (http://www.tranexp.com:2000/Translate/result.shtml) for Spanish-Italian and Altavista (http://world.altavista.com/) for English-Italian

[4]http://www.google.com/

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

new TEs from a non-annotated corpus in the target language.

There are two types of temporal keywords:

- IIigh temporality keywords: words that always a temporal meaning, such as "January", "day" and "days".
- Low temporality keywords: words with high probability of being part of a temporal expression, such as "next", "last" and "ago".

In order to extract new TEs, any sequence of keywords found in the corpus is considered as a possible temporal expression. For example, in "There were two accidents days ago", the keyword unit will first find a high temporality keyword ("days"), then look for more temporal words (high and low) preceeding or following this word, and finally return the expression "days ago". All the candidate temporal expression are checked for correctness using Google. In Phase 3, this new found expression should be translated in order to obtain the resolution rule for it.

- Phase 3: TE Normalizing Rule Assignment. In the last phase, the translators are used to relate the recognizing rule to the appropriate normalization rule. For this purpose, the system takes advantage of the previously defined Spanish and English temporal models.

3 A new porting procedure: automatic assignment of normalizing rules through annotated corpora

The problem with the automatic extension proposed in the previous section appears when the normalizing rules need to be assigned to the expressions that have been automatically extracted from the annotated corpus, as automatic translators are not completely accurate and a wrong assignment may happen as a consequence of an incorrect translation. The consequences of this are especially remarkable for languages other than English, for which less tools are available. In the porting to Italian, for example, a total of 2474 temporal expressions were extracted from the corpus. However, after the automatic translation and a manual debugging of the expressions, only 252 where left to be stored in the set of Italian temporal expressions.

To overcome this problem, a new procedure has been developed, which does not require the translation of the temporal expressions extracted from the corpus, but exploits more deeply the corpus annotations by taking into consideration also the information provided by the normalization attributes (see Section 2 for a description of such attributes).

In order to develop the procedure, all the temporal expressions and the values assigned to the different attributes have been extracted from the TERN 2004[5] training corpus of newspaper articles, improving the English database of TERSEO.

The new procedure has been carried out in four steps, that are described next, and it is shown in Figure 3.

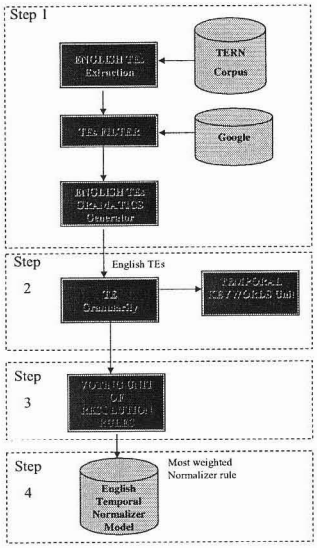

Figure 3. Automatic assignment of resolution rules

3.1 TE collection

As it was performed for the Italian extension of the system, all the annotated temporal expressions in English are extracted from an annotated corpus. However, not only the expressions are extracted and stored, but also the values of the normallization attributes of the TIMEX2 tags for each expression and the newspaper date (these values will be used for the automatic assignment of the resolution rule). After the TE-Collection phase, 2513 temporal expressions have been obtained. For example, the following annotation has been extracted: "<TIMEX2 val="2000-W44" mod="" set="" anchor_val="" anchor_dir="" comment="">this week< /TIMEX2>"; in this case, the temporal expression "this week" as well as the values of its attributes are extracted and the expression is resolved as "2000-W44". If this annotation is taken from the file "MNB20001102.2100.2766.tmx.sgml", the newspapers date, e.g. 20001102, is also extracted and attached to this

[5]http://timex2.mitre.org/corpora/timex2_corpora.html

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

expression as well as to every expression collected from the same article.

3.2 TE granularity

Due to the fact that the resolution rules in the TER-ILI unit are classified based on the granularity that the rule is referring to, it is necessary to determine the granularity of the new found expression to limit the search of the possible resolution rule for that expression. Some possible granularities stored in the database are: DAY, MONTH,YEAR, WEEK,...

The granularity is obtained using two factors associated with the expression:

- Temporal trigger words of the expression, such as, "day","months","January"
- The value assigned to the attribute VAL in the TIMEX2 tag of the expression.

For example, the TE "the last 10 years" has the attribute VAL="P10Y". This expression contains the temporal keyword "years" in the expression, and besides, the value of the VAL attribute contains "Y", that means year as well. So, the TE granularity unit is assigning "YEAR" for the expression of the example.

3.3 Voting unit of resolution rules

In this step, all the possible resolution rules related with the granularity of the temporal expression are obtained. Every possible rule is resolved for the temporal expression, using the newspaper date associated to the expression as a referent. The resolution obtained is compared with the values of the attributes of the TIMEX2 tag and a weight of similarity is assigned to every resolution rule. Finally, the resolution rule with the most similar results is assigned to the temporal expression. For example, for the expression "this week", the original tag extracted from the text was: "<TIMEX2 val="2000-W44" mod="" set="" anchor_val="" anchor_dir="" comment="">this week< /TIMEX2>". In this step, all the resolution rules related with weeks are obtained from the database and applied to the newspaper's date (20001102). After the resolution of this kind of rules, only a value for the VAL attribute was obtained, as it is shown in Table 1.

The weight is calculated as the sum of coincident attributes between the original TIMEX2 tag and the resulting tags after applying the different rules. So, in the example, for the expression "this week" the resolution rule "WEEK,PRESENT" is assigned.

Resolution Rule	Attribute values	Weight
WEEK,PAST	VAL=2000-W43	0
WEEK,PRESENT	VAL=2000-W44	1
WEEK, FUTURE	VAL=2000-W45	0

Table 1. Example of Step 3

3.4 TE-generalization

All the temporal expressions are transformed into patterns of temporal expression. For example, the expression "the last 10 years" is stored as "the last NUM years" in the database. The new found patterns are stored in the set of English temporal expressions, in order to be used by TERSEO system. The same process can be applied to other extensions of the language, such as Italian, without the necessity of the automatic translation of the new temporal expressions. In our experiments with the TERN 2004 training corpus, 1096 patterns of temporal expressions were obtained with its associated resolution rule.

This automatic assignment of the resolution rule will be evaluated in the next section.

4 Evaluation

In order to evaluate the automatic assignment of resolution rules for the new temporal expressions found in the TERN 2004 corpus, two different sets of expressions have been considered separately: expressions that were obtained from the annotated corpus and expressions that were obtained from the direct translation of Spanish expressions.

In the first case, the automatic assignment of the resolution rules resulted in a set of 994 expressions out of a total of 1096 expressions recognized by the system. These were considered as good new temporal expressions and could therefore be directly added to the English temporal expressions database.

In the case of the remaining 102 expressions, obtained from the direct translation of Spanish expressions, the resolution rule obtained automatically was compared with the one obtained by the translation and both were checked manually. Results are as follows:

- In 41 expressions out of a total of 102, both systems agreed in the resolution rule assigned to the expressions
- For the 61 expressions whose resolution rule differs from the one previously assigned, the results are: in 3 expressions, the new procedure assigns a better resolution rule than the older one; in 41 expressions, the old procedure assigns a better resolution rule than the new one; finally, there are 17 expressions whose assigned

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

	Original TERSEO			Updated TERSEO			ENG-Chronos		
Tag	P	R	F	P	R	F	P	R	F
timex2	0.673	0.728	0.699	0.780	0.924	0.846	0.976	0.880	0.926
anchor_dir	0.658	0.877	0.752	0.512	0.621	0.561	0.833	0.698	0.760
anchor_val	0.684	0.912	0.782	0.568	0.689	0.623	0.683	0.775	0.726
set	0.800	0.667	0.727	0.737	0.583	0.651	0.880	0.564	0.688
val	0.757	0.735	0.746	0.586	0.567	0.577	0.875	0.870	0.872

Table 2. Results obtained over TERN corpus by TERSEO and Chronos.

resolution rule is different but the annotation tag is the same after applying both rules.

The improvement that this new procedure adds to the automatic extension of the system is the possibility of obtaining temporal expressions that were not in the Spanish database. Besides, when the temporal expressions are extracted from the corpus, there is no necessity to automatically translate these expressions to other languages in order to obtain the resolution rule, which means avoiding the errors provoked by the lack of accuracy of the current automatic translators.

5 Comparative evaluation of the proposed approach

As a final evaluation experiment, we compared the automatic extension procedure described in the previous sections, with (i) the original English system based on the automatic translation of the Spanish rules, and (ii) a state of the art system developed for English.

Comparison with the original English version of TERSEO. When the original version of TERSEO participated in TERN 2004, the English temporal expressions handled by the system were only obtained through the automatic translation of the Spanish ones, sharing with them the corresponding resolution rule. The results obtained were high, taking into account that the English part of the system was automatically obtained from the existing knowledge for Spanish. However, with this methodology the number of TEs recognized by the system was limited to such existing (language specific) knowledge. In the new porting procedure, taking advantage of an annotated corpus, our primary goal was to increment the knowledge available to the system, thus increasing the number of TEs in the target language handled by TERSEO. In this direction, results reported in Table 2 show an improvement of more than 14% in recognition (first row) in the updated version of TERSEO, which confirms the viability of the proposed solution. It's worth noting that such improvement on recognition will be even more noticeable for languages other than English, where the poorer quality of the available translation tools represents an additional source of errors.

Comparison with Chronos. As a reference system for our comparative evaluation we adopted Chronos (Negri and Marseglia, 2004), a multilingual system for the recognition and normalization of time expressions in Italian and English. Like all the other state of the art systems addressing the recognition/normalization task, Chronos is a rule-based system. ¿From a design point of view, it shares with TERSEO a rather similar architecture which relies on different sets of rules. These are regular expressions that check for specific features of an input text, such as the presence of particular word senses, lemmas, parts of speech, symbols, or strings satisfying specific predicates (*e.g.* "Weekday-p" and "Time_Unit-p", which are respectively satisfied by strings such as "Monday", "Tuesday", ...,"Sunday", and "second", "minute", "hour", "day", ..., "century"). Each set of rules is in charge of dealing with different aspects of the problem. In particular, a set of around 130 rules is designed for TE recognition, and is capable of recognizing with high Precision/Recall rates both explicit and anaphoric TEs. Other sets of regular expressions, for a total of around 700 rules, are used in the normalization phase, and are in charge of handling each specific TIMEX2 attribute (*i.e.* VAL, SET, ANCHOR_VAL, and ANCHOR_DIR). The results obtained by the English version of Chronos over the TERN 2004 training corpus are shown in the last three columns of Table 2. As expected, the distance between the results obtained by TERSEO and Chronos is considerable. However, the great difference, both in terms of the required time and effort, in the development of the two systems should be taken into account. In fact, while the implementation of the manual one took several months, the porting procedure of TERSEO to English is a very fast process that can be accomplished in less than an hour. This makes the proposed procedure a viable solution which allows for a rapid porting of the system to other languages, while just requiring an annotated corpus.

6 Conclusions

In this paper, a complete automatic extension to other languages of a system that recognizes and normalizes temporal expression is presented. When these kind of knowl-

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 **IEEE**

edge based systems are implemented manually, the development could last several months. However, the porting procedure of TERSEO to other language (English, Italian) is a very fast process that can be accomplished in less than an hour. Moreover, even if an annotated corpus for a new language is not available, the automatic porting procedure we present still remains feasible. This makes the proposed approach a viable solution which allows for a rapid porting of the system to other languages, while just requiring an online translator (note that the Altavista Babel Fish translator[6] provides translations from English to 12 target languages) and/or an annotated corpora. In light of these considerations, the results obtained by TERSEO are encouraging. In addition, a set of different approaches to automatic extending the system are being performed in order to improve the automatic extension as much as possible.

References

[1] B. Carpenter. Phrasal Queries with LingPipe and Lucene. In *13th Text REtrieval Conference*, NIST Special Publication. National Institute of Standards and Technology, 2004.

[2] L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. Tides.2005 standard for the annotation of temporal expressions. Technical report, MITRE, 2005.

[3] E. Filatova and E. Hovy. Assigning time-stamps to event-clauses. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 88–95, Toulouse,France, 2001.

[4] A. Ittycheriah, L. Lita, N. Kambhatla, N. Nicolov, S. Roukos, and M. Stys. Identifying and Tracking Entity Mentions in a Maximum Entropy Framework. In ACL, editor, *Proceedings of the NorthAmerican Chapter Association for Computational Linguistic (NAACL) Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, 2003.

[5] G. Katz and F. Arosio. The annotation of temporal information in natural language sentences. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 104–111, Toulouse,France, 2001.

[6] A. Lavelli, B. Magnini, M. Negri, E. Pianta, M. Speranza, and R. Sprugnoli. Italian content annotation bank (i-cab): Temporal expressions (v. 1.0.): T-0505-12. Technical report, ITC-irst, Trento, 2005.

[7] T. Moia. Telling apart temporal locating adverbials and time-denoting expressions. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, Toulouse,France, 2001.

[8] M. Negri and L. Marseglia. Recognition and normalization of time expressions: Itc-irst at tern 2004. Technical report, ITC-irst, Trento, 2004.

[9] E. Saquete. *Temporal information Resolution and its application to Temporal Question Answering*. Phd, Departamento de Lenguages y Sistemas Informáticos. Universidad de Alicante, June 2005.

[10] E. Saquete, P. Martinez-Barco, and R. Munoz. Evaluation of the automatic multilinguality for time expression resolution. In *DEXA Workshops*, pages 25–30. IEEE Computer Society, 2004.

[11] E. Saquete, R. Munoz, and P. Martinez-Barco. Event ordering using terseo system. *Data and Knowledge Engineering Journal*, page (To be published), 2005.

[12] F. Schilder and C. Habel. From temporal expressions to temporal information: Semantic tagging of news messages. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 65–72, Toulouse,France, 2001.

[13] A. Setzer and R. Gaizauskas. On the importance of annotating event-event temporal relations in text. In LREC, editor, *Proceedings of the LREC Workshop on Temporal Annotation Standards, 2002*, pages 52–60, Las Palmas de Gran Canaria,Spain, 2002.

[14] P. Vossen. EuroWordNet: Building a Multilingual Database with WordNets in 8 European Languages. *The ELRA Newsletter*, 5(1):9–10, 2000.

[15] G. Wilson, I. Mani, B. Sundheim, and L. Ferro. A multilingual approach to annotating and extracting temporal information. In ACL, editor, *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*, pages 81–87, Toulouse,France, 2001.

[6] http://world.altavista.com/

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE