Evaluating Knowledge-based Approaches to the Multilingual Extension of a Temporal Expression Normalizer

Matteo Negri
ITC-irst
Povo - Trento, Italy
negri@itc.it

Estela Saquete, Patricio Martínez-Barco, Rafael Muñoz
DLSI, University of Alicante
Alicante, Spain
{stela,patricio,rafael}@dlsi.ua.es

Abstract

The extension to new languages is a well known bottleneck for rule-based systems. Considerable human effort, which typically consists in re-writing from scratch huge amounts of rules, is in fact required to transfer the knowledge available to the system from one language to a new one. Provided sufficient annotated data, machine learning algorithms allow to minimize the costs of such knowledge transfer but, up to date, proved to be ineffective for some specific tasks. Among these, the recognition and normalization of temporal expressions still remains out of their reach. Focusing on this task, and still adhering to the rule-based framework, this paper presents a bunch of experiments on the automatic porting to Italian of a system originally developed for Spanish. Different automatic rule translation strategies are evaluated and discussed, providing a comprehensive overview of the challenge.

1 Introduction

In recent years, inspired by the success of MUC evaluations, a growing number of initiatives (*e.g.* TREC[1], CLEF[2], CoNLL[3], Senseval[4]) have been developed to boost research towards the automatic understanding of textual data. Since 1999, the Automatic Content Extraction (ACE) program[5] has been contributing to broaden the varied scenario of evaluation campaigns by proposing three main tasks, namely the recognition of *entities*, *relations*, and *events*. In 2004, the Timex2 Detection and Recognition task[6] (also known as TERN, for Time Expression Recognition and Normalization) has been added to the ACE program, making the whole evaluation exercise more complete. The main goal of the task was to foster research on systems capable of automatically detecting temporal expressions (TEs) present in an English text, and normalizing them with respect to a specifically defined annotation standard.

Within the above mentioned evaluation exercises, the research activity on monolingual tasks has gradually been complemented by a considerable interest towards multilingual and cross-language capabilities of NLP systems. This trend confirms how portability across languages has now become one of the key challenges for Natural Language Processing research, in the effort of breaking the language barrier hampering systems' application in many real use scenarios. In this direction, machine learning techniques have become the standard approach in many NLP areas. This is motivated by several reasons, including *i)* the fact that considerable amounts of annotated data, indispensable to train ML-based algorithms, are now available for many tasks, and *ii)* the difficulty, inherent to rule-based approaches, of porting language models from one language to new ones. In fact, while supervised ML algorithms can be easily extended to new languages given an annotated training corpus, rule-based approaches require to redefine the set of rules, adapting them to each new language. This is a time consuming and costly work, as it usually consists in manually rewriting from scratch huge amounts of rules.

[1] http://trec.nist.gov
[2] http://clef-campaign.org
[3] http://www.cnts.ua.ac.be/conll
[4] http://www.senseval.org
[5] http://www.nist.gov/speech/tests/ace
[6] http://timex2.mitre.org

In spite of their effectiveness for some tasks, ML techniques still fall short from providing effective solutions for others. This is confirmed by the outcomes of the TERN 2004 evaluation, which provide a clear picture of the situation. In spite of the good results obtained in the TE *recognition* task (Hacioglu et al., 2005), the *normalization* by means of ML techniques has not been tackled yet, and still remains an unresolved problem.

Considering the inadequacy of ML techniques to deal with the *normalization* problem, and focusing on portability across languages, this paper extends and completes the previous work presented in (Saquete et al., 2006b) and (Saquete et al., 2006a). More specifically, we address the following crucial issue: how to minimize the costs of building a rule-based TE recognition system for a new language, given an already existing system for another language. Our goal is to experiment with different automatic porting procedures to build temporal models for new languages, starting from previously defined ones. Still adhering to the rule-based paradigm, we analyse different porting methodologies that automatically learn the TE recognition model used by the system in one language, adjusting the set of normalization rules for the new target language.

In order to provide a clear and comprehensive overview of the challenge, an incremental approach is proposed. Starting from the architecture of an existing system developed for Spanish (Saquete et al., 2005), we present a bunch of experiments which take advantage of different knowledge sources to build an homologous system for Italian. Building on top of each other, such experiments aim at incrementally analyzing the contribution of additional information to attack the TE normalization task. More specifically, the following information will be considered:

- The output of online translators;
- The information mined from a manually annotated corpus;
- A combination of the two.

2 The task: TE recognition and normalization

The TERN task consists in automatically *detecting*, *bracketing*, and *normalizing* all the time expressions mentioned within an English text. The recognized TEs are then annotated according to the TIMEX2 annotation standard described in (Ferro et al., 2005). Markable TEs include both *absolute* (or *explicit*) expressions (*e.g.* "April 15, 2006"), and *relative* (or *anaphoric*) expressions (*e.g.* "three years ago"). Also markable are durations (*e.g.* "two weeks"), event-anchored expressions (*e.g.* "two days before departure"), and sets of times (*e.g.* "every week"). Detection and bracketing concern systems' capability to recognize TEs within an input text, and correctly determine their extension. Normalization concerns the ability of the system to correctly assign, for each detected TE, the correct values to the TIMEX2 normalization attributes. The meaning of these attributes can be summarized as follows:

- VAL: contains the normalized value of a TE (*e.g.* "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 include: "approximately", "more_than", "less_than")
- SET: identifies expressions denoting sets of times (*e.g.* "every year").

2.1 The evaluation benchmark

Moving to a new language, an evaluation benchmark is necessary to test systems performances. For this purpose, the temporal annotations of the Italian Content Annotation Bank (I-CAB-temp[7]) have been selected.

I-CAB consists of 525 news documents taken from the Italian newspaper L'Adige (http://www.adige.it), and contains around 182,500 words. Its 3,830 temporal expressions (2,393 in the *training* part of the corpus, and 1,437 in the *test* part) have been manually annotated following the TIMEX2 standard with some adaptations to the specific morpho-syntactic features of Italian, which has a far richer morphology than English (see (Magnini et al., 2006) for further details).

[7] I-CAB is being developed as part of the three-year project ONTOTEXT funded by the Provincia Autonoma di Trento, Italy. See http://tcc.itc.it/projects/ontotext

3 The starting point: TERSEO

As a starting point for our experiments we used TERSEO, a system originally developed for the automatic annotation of TEs appearing in a Spanish written text in compliance with the TIMEX2 standard (see (Saquete, 2005) for a thorough description of TERSEO's main features and functionalities).

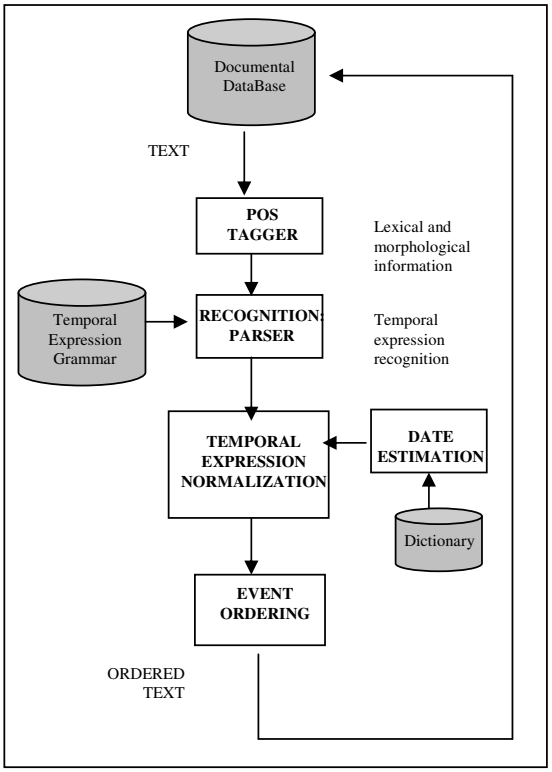

Figure 1: System's architecture.

Basically (see Figure 1), the TE recognition and normalization process is carried out in two phases. The first phase (recognition) includes a pre-processing of the input text, which is tagged with lexical and morphological information that will be used as input to a temporal parser. The temporal parser is implemented using an ascending technique (chart parser) and relies on a language-specific temporal grammar. As TEs can be divided into absolute and relative ones, such grammar is tuned for discriminating between the two groups. On the one hand, absolute TEs directly provide and fully describe a date. On the other hand, relative TEs require some degree of reasoning (as in the case of anaphora resolution).

In the second phase of the process, in order to translate these expressions into their normalized form, the lexical context in which they occur is considered. At this stage, a normalization unit is in charge of determining the appropriate reference date (anchor) associated to each *anaphoric* TE, calculating its value, and finally generating the corresponding TIMEX2 tag.

¿From a multilingual perspective, an important feature of TERSEO is the distinction between recognition rules, which are language-specific, and normalization rules, which are language-independent and potentially reusable for any other language. Taking the most from the modular architecture of the system, a first multilingual extension has been evaluated over the English TERN 2004 test set. In that extension, the English temporal model was automatically obtained from the Spanish one, through the automatic translation into English[8] of the Spanish TEs recognized by the system (Saquete et al., 2004). The resulting English TEs were then mapped onto the corresponding language-independent normalization rules, with good results (compared with other participants to the competition) both in terms of precision and recall. These results are shown in Table 1.

	Prec	**Rec**	**F**
timex2	0.673	0.728	0.699
anchor_dir	0.658	0.877	0.752
anchor_val	0.684	0.912	0.782
set	0.800	0.667	0.727
text	0.770	0.620	0.690
val	0.757	0.735	0.746

Table 1: Evaluation of *English*-TERSEO over the TERN 2004 test set

The positive results of this experience demonstrated the viability of the adopted solutions, and motivate our further investigation with Italian as a new target language.

4 Porting TERSEO to Italian

Due to the separation between language-specific recognition rules and language-independent normalization rules, the bulk of the porting process relies on the adaptation of the recognition rules to the new target language. Taking advantage of different knowledge sources (either alone or in combination), an incremental approach has been adopted, in order to determine the contribution of additional information on the performance of the resulting system for Italian.

[8] Altavista Babel Fish Translation has been used for this purpose (http://world.altavista.com).

4.1 Using online translators

As a first experiment, the same procedure adopted for the extension to English has been followed. This represents the simplest approach for porting TERSEO to other languages, and will be considered as a baseline for comparison with the results achieved in further experiments. The only minor difference with respect to the original procedure is that now, since two aligned sets of recognition rules (*i.e.* for Spanish and for English) are available, both models have been used. 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 perform better than translators from Spanish to Italian.

The process is carried out in the following four steps.

1. Eng-Ita translation. All the English TEs known by the system are translated into Italian[9]. Starting English, the probability of obtaining higher quality translations is maximized.

2. Spa-Ita translation. For each English TE without an Italian translation, the corresponding Spanish expression is translated into Italian. Also the Spanish TEs that do not have an English equivalent are translated from Spanish[10] into Italian. This way, the coverage of the resulting model is maximized, becoming comparable to the hand-crafted Spanish model.

3. TE Filtering. A filtering module is used to guarantee the correctness of the translations. For this purpose, the translated expressions are searched in the Web with Google. If an expression is not found by Google it is given up; otherwise it is considered as a valid Italian TE. The inconvenience of adopting this simple filtering strategy occurs in case of ambiguous expressions, *i.e.* when a correct expression is obtained through translation, and Google returns at least on document containing it, but the expression is not a temporal one. In these cases the system will erroneously store in its database non-temporal expressions. In this experiment the results returned by Google have not been analyzed (only the number of hits has been taken into account), nor the impact of these errors has been estimated. A more precise analysis of the output of the web search has been left as a future improvement direction.

4. Normalization rules assignment. Finally, the resulting Italian translations are mapped onto the language-independent normalization rules associated with the original English and Spanish TEs.

The development of this first automatic porting procedure required one person/week for software implementation, and less than an hour to obtain the new model for Italian. The performance of the resulting system, evaluated over the test set of I-CAB, is shown in table 2.

	Prec	**Rec**	**F**
timex2	0.725	0.833	0.775
anchor_dir	0.211	0.593	0.311
anchor_val	0.203	0.571	0.300
set	0.152	1.000	0.263
text	0.217	0.249	0.232
val	0.364	0.351	0.357

Table 2: Porting to Italian based on translations

The results achieved by the translation-based approach are controversial. On the one hand, we observe a *detection* performance in line with the English version of the system. The **timex2** attribute, which indicates the proportion of detected TEs[11], has even higher scores, both in terms of precision (+5%) and recall (+11%), with respect to the English system. On the other hand, both *bracketing* (see the **text** attribute, which indicates the quality of extent recognition) and *normalization* (described by the other attributes) show a performance drop. Unfortunately, the reasons of this drop are still unclear. One possible explanation is that, due to the intrinsic difficulties presented by the Italian language, the translation-based approach falls short from providing an adequate coverage of the many possible TE variants. While

[9] Also for English to Italian translation, Altavista Babel Fish Translation has been used

[10] Using the Spanish-Italian translator available at http://www.tranexp.com:2000/Translate/result.shtml

[11] At least one overlapping character in the extent of the reference and the system output is required for tag aligment.

the presence of *lexical triggers* denoting a TE appearing in a text (*e.g.* the Italian translations of "years", "Monday", "afternoon", "yesterday") can be easily captured by this approach, the complexity of many language-specific constructs is out of its reach.

4.2 Using an annotated corpus

In a second experiment, the annotations of the training portion of I-CAB have been used as a primary knowledge source. The main purpose of this approach is to maximize the coverage of the Italian TEs, starting from language-specific knowledge mined from the corpus. The basic hypothesis is that a *bottom-up* porting methodology, led by knowledge in the target language, is more effective than the *top-down* approach based on knowledge derived from models built for other languages. The former, in fact, is in principle more suitable to capture language-specific TE variations. In order to test the validity of ths hypothesis, the following two-step process has been set up:

1. TE Collection and translation. The Italian expressions are collected from the I-CAB training portion, and translated both into Spanish and English.

2. Normalization rules assignment. Italian TEs are assigned to the appropriate normalization rules. For each Italian TE mined from the corpus, the selection is done considering the normalization rules assigned to its translations. If both the Spanish and English expressions are found in their respective models, and are associated with the same normalization rule, then this rule is assigned also to the Italian expression. Also, when only one of the translated expressions is found in the existing models, the normalization rule is assigned. In case of discrepancies, *i.e.* if both expressions are found, but are not associated to the same normalization rule, then one of the languages must be prioritized. Since the manually obtained Spanish model has shown a higher precision, Spanish rules are preferred.

As the corpus-based approach is mostly built on the same software used for the translation-based porting procedure, it did not require additional time for implementation. Also in this case, the new model for Italian has been obtained in less than one hour. Performance results calculated over the I-CAB test set are reported in Table 3.

	Prec	**Rec**	**F**
timex2	0.730	0.839	0.781
anchor_dir	0.412	0.414	0.413
anchor_val	0.339	0.340	0.339
set	0.030	1.000	0.059
text	0.222	0.255	0.238
val	0.285	0.274	0.279

Table 3: Porting based on corpus annotations

These results partially confirm our working hypothesis, showing a performance increase in terms of the Italian TEs correctly recognized by the system. In fact, both the **timex2** attribute, which indicates the coverage of the system (detection), and the **text** attribute, which refers to the TEs extent determination (bracketing), are slightly increased. This may lead to the conclusion that automatic porting procedures can actually benefit from language-specific knowledge derived from a corpus.

However, looking at the other TIMEX2 attributes, the situation is not so clear due to the less coherent behaviour of the system on normalization. While for two attributes (**anchor_dir** and **anchor_val**) the system performs better, for the other two (**set** and **val**) a performance drop is observed. A possible reason for that could be related to the limited number of TE examples that can be extracted from the Italian corpus (whose dimensions are relatively small compared to the annotated corpora available for English). In fact, compared to the sum of English and Spanish examples used for the translation-based porting procedure, the Italian expressions present in the corpus are fewer and repetitive. For instance, with 131, 140, and 30 occurrences, the expressions "*oggi*" ("today"), "*ieri*" ("yesterday"), and "*domani*" ("tomorrow") represent around 12.5% of the 2,393 Italian TEs contained in the I-CAB training set.

4.3 Combining online translators and an annotated corpus

In light of the previous considerations, a third experiment has been conducted combining the *top-down* approach proposed in Section 4.1 and the *bottom-up* approach proposed in Section 4.2. The underlying hypothesis is that the induction of an effective temporal model for Italian can benefit from the combination of the large amount of examples coming from translations on the one

side, and from the more precise language-specific knowledge derived from the corpus on the other.

To check the validity of this hypothesis, the process described in Section 4.2 has been modified adding an additional phase. In this phase, the set of TEs derived from I-CAB is augmented with the expressions already available in the Spanish and English TE sets. The new porting process is carried out in the following steps:

1. TE Collection and translation. The Italian expressions are collected from the I-CAB training portion, and translated both into Spanish and English.

2. Normalization rules assignment. With the same methodology described in Section 4.2 (step 2), the Italian TEs mined from the corpus are mapped onto the appropriate normalization rules assigned to their translations.

3. TE set augmentation. The set of Italian TEs is automatically augmented with new expressions derived from the Spanish and English TE sets. As described in Section 4.1, these expressions are first translated into Italian using on-line translators, then filtered through Web searches. The remaining TEs are included in the Italian model, and related to the same normalization rules assigned to the corresponding Spanish or English TEs.

Also this porting experiment was carried out with minimal modifications of the existing code. The automatic acquisition of the new model for Italian required around one hour. Evaluation results, calculated over the I-CAB test set are presented in Table 4.

	Prec	**Rec**	**F**
timex2	0.726	0.834	0.776
anchor_dir	0.578	0.475	0.521
anchor_val	0.516	0.424	0.465
set	0.182	1.000	0.308
text	0.258	0.296	0.276
val	0.564	0.545	0.555

Table 4: Porting based on corpus annotations and online translators

As can be seen from the table, the combination of the two approaches leads to an overall performance improvement with respect to the previous experiments. Apart from a slight decrease in terms of *detection* (**timex2** attribute), both bracketing and normalization performance benefit from such combination. The improvement on *bracketing* (**text** attribute) is around 4% with respect to both the previous experiments. On average, the improvement for the *normalization* attributes is around 15% with respect to the translation-based method (ranging from +4,5% for the **set** attribute, to +20% for the **val** attribute), and 20% with respect to the corpus-based method (ranging from +11% for the **anchor_dir** attribute, to +30% for the **set** attribute). These performance improvements are summarized in Table 5, which reports the F-Measure scores achieved by the three porting approaches.

	F-Tran.	**F-Corpus**	**F-Comb.**
timex2	0.775	**0.781**	0.776
anchor_dir	0.311	0.413	**0.521**
anchor_val	0.300	0.339	**0.465**
set	0.152	0.059	**0.308**
text	0.263	0.238	**0.276**
val	0.232	0.279	**0.555**

Table 5: F-Measure scores comparison

These results confirm the validity of our working hypothesis, showing that:

- taken in isolation, both the knowledge derived from models built for other languages, and the language-specific knowledge derived from an annotated corpus, have a limited impact on the system's performance;

- taken in combination, the *top-down* and the *bottom-up* approaches can complement each other, allowing to cope with the complexity of the porting task.

5 Comparing TERSEO with a language-specific system

For the sake of completeness, the results achieved by our combined porting procedure have been compared with those achieved, over the I-CAB test set, by a system specifically designed for Italian. The *ITA*-Chronos system (Negri and Marseglia, 2004), a multilingual system for the recognition and normalization of TEs in Italian and English, has been used for this purpose. Up to date, being among the two top performing systems at TERN 2004, Chronos represents the state-of-the-art with respect to the TERN task. In addition, to the best of our knowledge, this is the only system effectively dealing with the Italian language.

Like all the other state-of-the-art systems addressing the recognition/normalization task, *ITA*-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 the input text, such as the presence of particular word senses, lemmas, parts of speech, symbols, or strings satisfying specific predicates[12]. Each set of rules is in charge of dealing with different aspects of the problem. In particular, a set of around 350 rules is designed for TE recognition and is capable of recognizing with high Precision/Recall rates a broad variety of 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 normalization attribute. The results obtained by the Italian version of Chronos over the I-CAB test set are shown in Table 6.

	Prec	Rec	F	F-Comb
timex2	0.925	0.908	**0.917**	0.776 (-14%)
anchor_dir	0.733	0.636	**0.681**	0.521 (-16%)
anchor_val	0.495	0.462	**0.478**	0.465 (-1.3%)
set	0.616	0.500	**0.552**	0.308 (-24%)
text	0.859	0.843	**0.851**	0.276 (-57%)
val	0.636	0.673	**0.654**	0.555 (-10%)

Table 6: Evaluation of *ITA*-Chronos over the I-CAB test set

As expected, the distance between the results obtained by *ITA*-Chronos and the best Italian system automatically obtained from TERSEO (F-Comb) is considerable. On average, in terms of F-Measure, the scores obtained by *ITA*-TERSEO are 20% lower, ranging from -1.3% for the **anchor_val** attribute, to -57% for the **text** attribute. However, going beyond the raw numbers, a comprehensive evaluation must also take into account the great difference, in terms of the required time, effort, and resources deployed in the development of the two systems. While the implementation of the manual one took several months, the automatic porting procedure of TERSEO to Italian (in all the three modalities described in this paper) is a very fast process that can be accomplished in less than an hour. Considering the trade-off between performance and effort required for system's development, the proposed methodology represents a viable solution to attack the porting problem.

[12]For instance, the predicates "Weekday-p" and "Time_Unit-p" are respectively satisfied by strings such as "Monday", "Tuesday", ..., "Sunday", and "second", "minute", "hour", "day", ..., "century". Of course, this also holds for the Italian equivalents of these expressions

6 Conclusions

In this paper, the problem of automatically extending to new languages a rule-based system for TE recognition and normalization has been addressed. Adopting an incremental approach, different porting strategies, for the creation of an Italian system starting from an already available Spanish system, have been evaluated and discussed. Each experiment has been carried out considering the contribution of different knowledge sources for rules translation. Firstly, the contribution given by the output of online translators has been evaluated, showing detection performances in line with a previously developed English extension of the system, but a performance drop in terms of normalization performance. Then, the contribution of knowledge mined from an annotated corpus has been considered. Results show a performance increase in terms of detection and bracketing, but a less coherent behaviour in terms of normalization. Finally, a combined approach has been experimented, resulting in an overall performance increase. System's performance is still far from the results obtained by a state-of-the-art system for Italian but, considering the trade-off between performance and effort required for system's development, results are encouraging.

References

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

K. Hacioglu, Y. Chen, and B. Douglas. 2005. Time Expression Labeling for English and Chinese Text. In *Proceedings of CICLing 2005*, pages 548–559.

B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, and R. Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006*. To appear.

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

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

E. Saquete, R. Muoz, and P. Martnez-Barco. 2005. Event ordering using TERSEO system. *Data and*

Knowledge Engineering Journal, page (To be published).

E. Saquete, P. Martinez-Barco, R. Munoz R., M. Negri, M. Speranza, and R. Sprugnoli R. 2006a. Automatic resolution rule assignment to multilingual temporal expressions using annotated corpora. In *Proceedings of the TIME 2006 International Symposium on Temporal Representation and Reasoning*. To Appear.

E. Saquete, P. Martinez-Barco, R. Munoz R., M. Negri, M. Speranza, and R. Sprugnoli R. 2006b. Multilingual Extension of a Temporal Expression Normalizer using Annotated Corpora. In *Proceedings of the EACL Workshop on Cross-Language Knowledge Induction*.

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