TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2

Hector Llorens, Estela Saquete, Borja Navarro
University of Alicante
Alicante, Spain
{hllorens,stela,borja}@dlsi.ua.es

Abstract

This paper presents TIPSem, a system to extract temporal information from natural language texts for English and Spanish. TIPSem, learns CRF models from training data. Although the used features include different language analysis levels, the approach is focused on semantic information. For Spanish, TIPSem achieved the best F1 score in all the tasks. For English, it obtained the best F1 in tasks B (events) and D (event-dct links); and was among the best systems in the rest.

1 Introduction

The automatic treatment of time expressions, events and their relations over natural language text consists of making temporal elements explicit through a system that identifies and annotates them following a standard scheme. This information is crucial for other natural language processing (NLP) areas, such as summarization or question answering. The relevance of temporal information has been reflected in specialized conferences (Schilder et al., 2007) and evaluation forums (Verhagen et al., 2007).

We present a system to tackle the six different tasks related to multilingual temporal information treatment proposed in TempEval-2. Particularly, in this evaluation exercise, TimeML (Pustejovsky et al., 2003) is adopted as temporal annotation scheme. In this manner, the tasks require participating systems to automatically annotate different TimeML elements. Firstly, task A consists of determining the extent of time expressions as defined by the TimeML TIMEX3 tag, as well as the attributes "type" and "value". Secondly, task B addresses the recognition and classification of events as defined by TimeML EVENT tag. Finally, tasks C to F comprise the categorization of different temporal links (TimeML LINKs). Figure 1 illustrates the TimeML elements in a sentence.

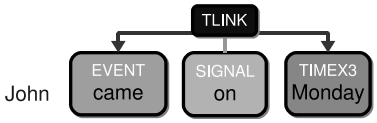

Figure 1: TimeML example

In the context of TempEval-2, we tackle all tasks for English and Spanish with a data-driven system. This consists of CRF models inferred from lexical, syntactic and semantic information of given training data.

Our main approach, TIPSem (Temporal Information Processing based on Semantic information), is focused on semantic roles and semantic networks. Furthermore, we present a secondary approach, TIPSem-B (TIPSem-Baseline), which contrary to the former does not consider semantic information.

The main objectives of this paper are (1) evaluating the performance of TIPSem comparing it to other participating systems and (2) measuring the contribution of semantic information to different TempEval-2 tasks though the comparison between our systems: TIPSem and TIPSem-B.

This paper is structured as follows. Our approach to address the TempEval-2 tasks is motivated in Section 2 and described in Section 3. The results obtained in the evaluation are shown and analyzed in Section 4. Finally, conclusions are drawn in Section 5.

2 Approach motivation

The next two subsections describe the two main characteristics of our approach, CRFs and semantic roles, and the reasons why we think they could be useful to tackle TimeML annotation.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 284–291,
Uppsala, Sweden, 15-16 July 2010.

2.1 CRF probabilistic model

Conditional Random Fields is a popular and efficient ML technique for supervised sequence labeling (Lafferty et al., 2001). In the recognition problem raised by TempEval-2 tasks A and B, assume X is a random variable over data sequences to be labeled, and Y is a random variable over the corresponding label sequences, being all Y components (Y_i) members of a finite label alphabet γ. X might range over the sentences and Y range over possible annotations of those sentences, with γ the set of possible event IOB2[1] labels. The following example illustrates the problem.

(1)

X	Y		
That	?		B-TIMEX3
was	?		B-EVENT
another	?	? =	I-TIMEX3
bad	?		I-EVENT
week	?		O

Then, CRFs construct a conditional model from paired observations and label sequences: $p(Y|X)$.

To extend the problem to classification, X is replaced with elements to be classified and γ is replaced with the possible classes, for instance, in task A $X = \{\text{TIMEX3 instances in text}\}$ and $\gamma = \{\text{DATE, DURATION, SET, TIME}\}$.

From our point of view, CRFs are well suited to address TempEval-2 tasks. Firstly, TimeML elements depend on structural properties of sentences. Not only the word sequence, but morphological, syntactic and semantic structure is related with them. Secondly, some TIMEX3 and EVENT elements are denoted by sequences of words, therefore the CRFs are very appropriate.

2.2 Semantic roles

Semantic role labeling (SRL) has achieved important results in the last years (Gildea and Jurafsky, 2002; Moreda et al., 2007). For each predicate in a sentence, semantic roles identify all constituents, determining their arguments (agent, patient, etc.) and their adjuncts (locative, temporal, etc.). Figure 2 illustrates a semantic role labeled sentence.

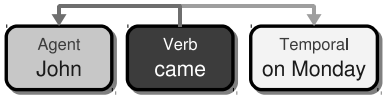

Figure 2: Semantic roles example

Semantic roles provide structural relations of the predicates in which TimeML elements may participate. Beyond syntactic relations expressed by means of the different types of phrases, semantic roles give further information about semantic relations between the arguments of a predicate. Due to the fact that roles represent high level information, they are more independent from word tokens. Hence, roles may aid in learning more general models that could improve the results of approaches focused on lower level information.

3 Our approach: TIPSem

As defined in previous section, this paper proposes CRF as learning method to infer models to face the TempEval-2 tasks. Specifically, CRF++ toolkit[2] was used for training and testing our approach. The learning process was done using the parameters: *CRF-L2* algorithm and hyperparameter C=1.

In order to set out the approach architecture and select the features for learning, we divided the tasks proposed in the evaluation exercise into four groups: recognition, classification, normalization and link-categorization. Each group represents a kind of problem to be resolved. Recognition problem is present in TIMEX3 and EVENT bounding (tasks A and B). Classification problem appears in TIMEX3 type and EVENT class attributes (tasks A and B). Normalization arises in TIMEX3 value attribute (task A). And link-categorization is applied to different kind of link relations (tasks C to F). Each group uses a particular feature set to learn an annotation model. The features of these sets are grouped in two subsets. On the one hand, general features, which are widely used in different NLP fields and represent lower language analysis levels. On the other hand, semantic features, which are a novelty in the task and our main focus.

TIPSem system uses all the features defined above. However, to measure the influence of semantic information in temporal information treatment, TIPSem-B system was implemented excluding the semantic features.

3.1 Recognition

In recognition, the features are obtained at token level, that is to say, each token has its own set of features.

Regarding each language analysis level, the general features used to train our CRF model are:

[1] IOB2 format: (B)egin, (I)nside, and (O)utside

[2] http://crfpp.sourceforge.net/

- **Morphological**: The lemma and part-of-speech (PoS) context, in a 5-window (-2,+2), was employed due to the good results it achieved in other NLP tasks. Tokenization, PoS and lemmatization were obtained using TreeTagger (Schmid, 1994) for English, and were got from AnCora (Taulé et al., 2008) for Spanish.

- **Syntactic**: Different TimeML elements are contained in particular types of phrases. This feature tries to capture this fact by considering phrase level syntactic information. The syntactic tree was obtained using Charniak parser (Charniak and Johnson, 2005) for English, and AnCora for Spanish.

- **Polarity, tense and aspect**: These were obtained using PoS and a set of handcrafted rules (e.g., will+verb → future).

The semantic level features used to enhance the training framework of the CRF model are:

- **Role**: For each token, we considered the role regarding the verb the token depends on. To get semantic roles, CCG SRL tool (Punyakanok et al., 2004) was used for English, and AnCora for Spanish.

- **Governing verb**: The verb to which the current token holds a particular role. This may distinguish tokens appearing under the influence of different verbs.

- **Role+verb combination**: The previous two features were combined to capture the relation between them. This introduces additional information by distinguishing roles depending on different verbs. The importance of this falls especially on the numbered roles (A0, A1, etc.) meaning different things when depending on different verbs.

- **Role configuration**: This feature is only present in verb tokens heading a sentence or sub-sentence. This consists of the set of roles depending on the verb. This may be particularly useful for distinguish different sentence settings.

- **Lexical semantics**: WordNet (Fellbaum, 1998) top ontology classes have been widely used to represent word meaning at ontological level, and demonstrated its worth in many tasks. TIPSem uses the top four classes for each word. For Spanish, EuroWordNet (Vossen, 1998) was used.

In this manner, given a list of tokens and its features, the trained recognition model will assign to each token one of the valid labels. For instance, in the case of TIMEX3 recognition: B-TIMEX3, I-TIMEX3 or O.

3.2 Classification

Classification features, used to get TIMEX3 types and EVENT classes, are basically the same as the ones used for recognition. However, the main difference is that the features are not obtained at token level but at TIMEX3 or EVENT level. This implies that the word context is set to the extent of each element (TIMEX3 and EVENT), as well as all the features have as many values as tokens comprises the element (e.g., element-tokens="next Monday", PoS-feature="JJ+NNP"). Hence, following this description, the classification models will assign to each element one of the valid classes. For example, in the case of TIMEX3 typing: DATE, DURATION, SET or TIME.

3.3 Normalization

As in classification the features are obtained at TIMEX3 level. Furthermore, word-spelled numbers contained in the TIMEX3 extent are translated to their numerical value (e.g., "three days" → "3 days").

Normalization process consists of two main steps: (1) obtain the normalization type and (2) apply the corresponding normalization rules.

The first step applies a CRF model that uses the same features as the previous two plus TIMEX3 pattern. This new feature consists of the tokens comprised by the TIMEX3 but replacing numbers by NUM, temporal units, such as years or days, by TUNIT, months by MONTH, and weekdays by WEEKDAY. In other words, "next Monday" would result in "next WEEKDAY" and "June 1999" in "MONTH NUM". Once the model is trained, for each new TIMEX3 it assigns a normalization type. We define seven normalization types: Period, ISO, ISO_set, ISO_function, present_ref, past_ref and future_ref.

The second step uses as input the output of the first one. Each normalization type has its own normalization rules.

- **Period**: Apply rules to convert period-like TIMEX3 (“3 days”) into P_NUM_TUNIT normalized period (“P3D”).

- **ISO**: Apply rules to convert any-format explicit date or time into a valid ISO 8601 standard date.

- **ISO_set**: Apply rules to get a valid ISO-like set from a TIMEX3 set (“monthly” → XXXX-XX).

- **ISO_function**: This is the most complex type. The system applies different functions to get a valid ISO date or time in a valid granularity from DCT[3] dates. Here, time direction indicators like “next” or “previous”, as well as verbal tenses are used.

- **Present_ref, past_ref and future_ref**: these are already normalized.

3.4 Link-categorization

Each one of link-related tasks (C to F) has its own link-categorization features. Nevertheless, all link types share some of them.

- **Task C**: For categorizing the relation between an EVENT and a TIMEX3, the system takes into account the following features:
  - *Heading preposition* if the event or the TIMEX3 are contained by a prepositional phrase as in “before the meeting”, where “meeting” is the event and “before” the heading preposition.
  - *Syntactic relation* of the event and the TIMEX3 in the sentence. This feature may be evaluated as: same sentence, same subsentence or same phrase.
  - *Time position*. If the event is not directly linked with the relation TIMEX3 but related to another TIMEX3, the time position represents whether the event is before, overlap or after the relation TIMEX3.
  - *Interval*. This feature is 0 unless there appears some interval indicator token near the TIMEX3. This is useful to identify overlap-and-after and overlap-and-before categories.
  - *TIMEX3 type*.
  - *Semantic roles* if the event or the TIMEX3 are contained by a temporal subordination (labeled with temporal role), for example, in “after he left home”, “left” is the event and “after” the subordinating element (role feature).

- **Task D**: To determine the relationship between an event and the DCT, TIPSem uses the same features as in task C except *interval*. In addition, all the features related to TIMEX3 are now related to the closer TIMEX3 (if exists) in the event sentence. In this manner, the *time position* is calculated by comparing DCT and that TIMEX3.

- **Task E**: Relations between two main events are categorized using only four features: the *tense and aspect* of the two events, the *syntactic relation* between them, and the *time position*, calculated using the closer TIMEX3 to each event.

- **Task F**: For categorizing subordinated events, TIPSem uses the subordinating element of temporal *roles* containing each event (if present), the *heading preposition* of a prepositional phrases containing each event (if present), as well as the *tense and aspect*.

To illustrate the system architecture, Figure 3 summarizes the strategies that TIPSem follows to tackle the tasks proposed in the TempEval-2 framework.

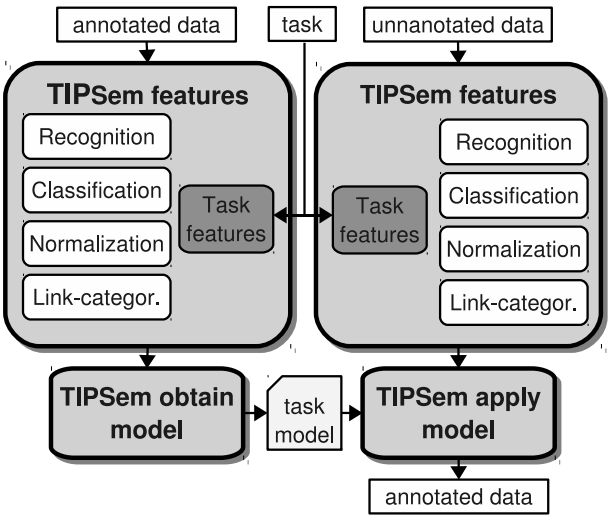

Figure 3: TIPSem architecture

[3]Date Creation Time

4 Evaluation

The test corpus consists of 17K words for English and 10K words for Spanish, in which approximately a half part correspond to tasks A and B, and the other half to tasks C, D, E and F. The performance is measured using precision, recall and $F_{\beta=1}$ metrics. A scoring script is provided. This counts correct instances at token level for tasks A and B, and at temporal link level for the rest.

Next subsections show the results obtained by TIPSem system in each one of the TempEval-2 tasks for English (EN) and Spanish (ES). Moreover, a final subsection illustrates the $F_{\beta=1}$ results in three comparative graphs. In tasks A and B, precision, recall and $F_{\beta=1}$ are given. In tasks C to E, links tasks precision, recall and $F_{\beta=1}$ are the same because our system does not consider the NONE value. Hence, only $F_{\beta=1}$ is given. Tasks E and F were not considered for Spanish in TempEval-2 evaluation and thus Spanish is excluded from those subsections.

For each task, scores in which our system obtained the first place in the evaluation exercise are in bold. Furthermore, in all cases the best score obtained by participating systems is reported. Finally, the influence of semantic information in terms of improvement is indicated and analyzed through the comparison with TIPSem-B system, which exclude the features related with semantics.

4.1 Task A: TIMEX3

Table 1 shows the results obtained by our approaches in TIMEX3 recognition, typing and ISO 8601 normalization (value).

System *lang*	**Prec.**	**Rec.**	**$F_{\beta=1}$**	**type**	**value**
TIPSem *EN*	**0.92**	0.80	0.85	0.92	0.65
TIPSem *ES*	0.95	**0.87**	**0.91**	**0.91**	0.78
TIPSem-B *EN*	0.88	0.60	0.71	0.88	0.59
TIPSem-B *ES*	**0.97**	0.81	0.88	0.99	0.75

Table 1: Task A - English and Spanish

As shown in results, TIPSem obtains the best results for Spanish in all measures except for "value" attribute, in which the best system obtained a 0.83. Another system obtained the same recall (0.87) but a lower precision (0.90), and thus a $F_{\beta=1}$ of (0.88) below TIPSem score (0.91). For English, our main approach obtained the best precision. However, another system obtained the best recall (0.91). The best $F_{\beta=1}$ was 0.86. Regarding type attribute, TIPSem obtained values closer to best system (0.98). Finally, in normalization, which is the only attribute that is not annotated by a purely data-driven process, best system surpassed TIPSem in 0.20.

These results indicate that CRFs represent an appropriate ML technique to learn models for annotating TIMEX3. Furthermore, they show that normalization process used by TIPSem could be improved using other techniques.

Specifically, the usage of semantic information improved the capability of learned models to generalize rules. For instance in time expressions, if an unseen instance is contained by a temporal role is a clear candidate to be a time expression. Hence, they improve system recall (33% EN, 7% ES).

4.2 Task B: EVENT

Table 2 shows the results obtained by our approaches in recognizing and classifying events.

System *lang*	**Prec.**	**Recall**	**$F_{\beta=1}$**	**class**
TIPSem *EN*	0.81	0.86	**0.83**	**0.79**
TIPSem *ES*	0.90	**0.86**	**0.88**	**0.66**
TIPSem-B *EN*	**0.83**	0.81	0.82	0.79
TIPSem-B *ES*	**0.92**	0.85	0.88	0.66

Table 2: Task B - English and Spanish

In this tasks, TIPSem obtained the best results in TempEval-2 for Spanish and English in both recognition and classification. Although for English another system achieved the best recall (0.88), it obtained a lower precision (0.55); and thus a 0.68 $F_{\beta=1}$. This indicates that our approach obtains the best $F_{\beta=1}$ (0.83) with a well-balanced precision and recall.

Again, the usage of semantic information improves the capability of learned models to generalize, which improves the recall (6% EN, 1% ES). For events, the improvement is lower than for TIMEX3 because, contrary to TIMEX3, they are not clearly defined by specific roles. In this case, features like role configuration, semantic classes, or role-governing verb are more useful.

Other attributes present in events such as polarity, mood and tense obtained values of about 90%. However, to get the values for these attributes the system applies a set of handcrafted rules and then the results are not relevant for our approach.

4.3 Task C: LINKS - Events and TIMEXs

Table 3 shows the results obtained by our approaches in categorizing EVENT-TIMEX3 links.

System *lang*	$\mathbf{F}_{\beta=1}$
TIPSem *EN*	0.55
TIPSem *ES*	**0.81**
TIPSem-B *EN*	0.54
TIPSem-B *ES*	0.81

Table 3: Task C - English and Spanish

TIPSem was the only system participating in this task for Spanish. Nevertheless, 0.81 is a high score comparing it to English best score (0.63). Our system, for English, is 8 points below top scored system.

In this task, the application of semantic roles introduced an improvement of 2% in $F_{\beta=1}$.

4.4 Task D: LINKS - Events and DCTs

Table 4 shows the results obtained by our approaches in categorizing events with respect to the creation time of a document.

System *lang*	$\mathbf{F}_{\beta=1}$
TIPSem *EN*	**0.82**
TIPSem *ES*	**0.59**
TIPSem-B *EN*	0.81
TIPSem-B *ES*	0.59

Table 4: Task D - English and Spanish

Task D is successfully covered by TIPSem obtaining the best results in the evaluation.

It seems that the relation of events with document creation time strongly depends on tense and aspect, as well as the event position in time with respect to DCT when defined by neighboring TIMEX3.

Furthermore, the learned CRF models take advantage of using temporal semantic roles information. Specifically, the usefulness of semantic roles in this task was quantified to 2%.

4.5 Task E: LINKS - Main events

Table 5 shows the results obtained by our approaches in categorizing main events relations in text.

System *lang*	$\mathbf{F}_{\beta=1}$
TIPSem *EN*	0.55
TIPSem-B *EN*	0.55

Table 5: Task E - English

In this task, our system obtains the second place. However, the top scored achieved a 0.56. Again, the tense and aspect features, as well as the events position in time resulted useful to tackle this task. In this case, semantic roles information is not used so TIPSem and TIPSem-B are equivalent.

4.6 Task F: LINKS - Subordinated events

Table 6 shows the results obtained by our approaches in categorizing events relations with the events they syntactically govern.

System *lang*	$\mathbf{F}_{\beta=1}$
TIPSem *EN*	0.59
TIPSem-B *EN*	0.60

Table 6: Task F - English

Categorizing subordinated events TIPSem obtained the second place. Best score was 0.66. In this task, the application of roles did not help and decreased the score in one point. The cause may be that for this task roles are not relevant but noisy. In this case, some extra information extending semantic roles is needed to turn them into a useful feature.

4.7 Comparative graphs

This subsection presents the TIPSem $F_{\beta=1}$ results in three graphs. Figure 4 illustrates the results for English indicating the higher and lower scores achieved by TempEval-2 participating systems. Figure 5 shows the same for Spanish but, due to the fact that TIPSem was the only participant in tasks B, C and D, the graph includes English min. and max. scores as indirect assessment. Finally, Figure 6, compares the TIPSem results for English and Spanish.

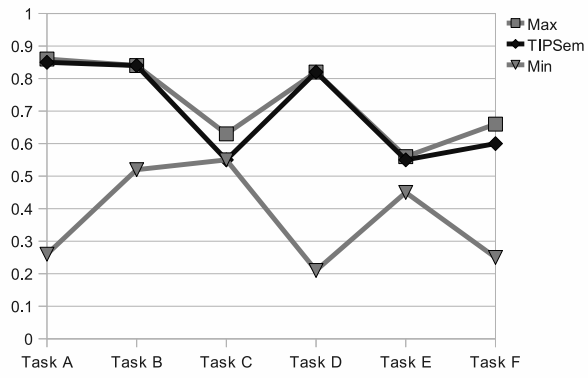

Figure 4: English $F_{\beta=1}$ comparative

Figure 4 shows how TIPSem achieved, in general, a high performance for English.

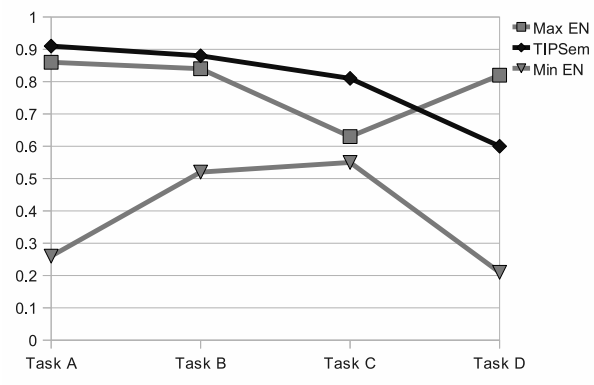

Figure 5: Spanish $F_{\beta=1}$ indirect assessment

For Spanish we can only report indirect assessment comparing the results to English scores. It can be seen that the quality of the results is similar for tasks A and B, but seems to be inverted in tasks C and D.

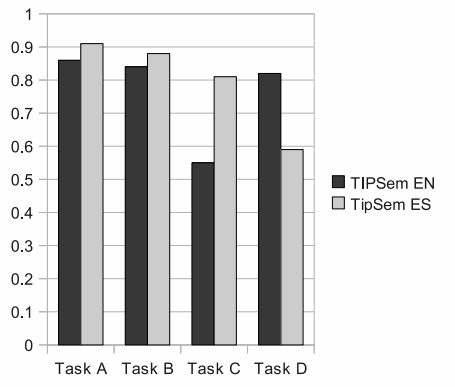

Figure 6: TIPSem EN - ES $F_{\beta=1}$ comparative

Finally, in this graph comparing TIPSem results, we observe that our approach achieved similar performance for both languages in tasks A and B. This indicates that for this tasks, the approach is valid for both languages. However, as in the previous graph, it seems that for English TIPSem performs worse in task C and better in task D while for Spanish it does right the opposite.

The train and test corpora were reviewed to analyze this fact. On the one hand, the reason for the high performance in task C for Spanish was the high amount of "overlap" instances in both corpora. This trained the CRF model for categorizing event-timex links as "overlap" in most of cases. On the other hand, the cause of the Spanish low performance in task D is "vague" links. The features defined in TIPSem cannot distinguish between "overlap" and "vague". Due to the fact that "vague" links are quite popular in Spanish test set, the results decreased. This did not affect to English results because of the spareness of "vague" links.

5 Conclusions and Further Work

This paper presented a system for automatically treating temporal information of natural language texts as required in the TempEval-2 evaluation exercise, in particular, following TimeML specifications.

Our system, TIPSem, is a data-driven approach and consists of different CRF models learned using semantic information as main feature. CRFs were used taking into account that data-driven approaches have obtained good results in many NLP tasks, and due to their appropriateness in sequence labeling problems and problems in which structural properties are relevant, as those proposed in TempEval-2. Furthermore, the models were enhanced using semantic information. Roles have been applied in other NLP fields with successful results, but never employed before for this purpose. With these two main characteristics, we designed a complete learning environment selecting, in addition to roles, different language analysis level properties as features to train the models.

The results obtained for English and Spanish in the evaluation exercise were satisfactory and well-balanced between precision and recall. For Spanish, TIPSem achieved the best $F_{\beta=1}$ in all tasks. For English, it obtained the best $F_{\beta=1}$ in event recognition and classification (task B), and event and document creation time links categorization (task D). Furthermore, in general, all the results of TIPSem were very competitive and were among the top scored systems. This verifies that our approach is appropriate to address TempEval-2 tasks.

Regarding multilinguality, the approach was proven to be valid for different languages (English and Spanish). This was also verified for Catalan language by earlier versions of TIPSem (Llorens et al., 2009). In fact, the data-driven part of the system could be considered language independent because it has been applied to different languages and could be applied to other languages without adaptation, provided that there are tools available to get the morphosyntactic and semantic information required by the approach. It has to be high-

lighted that to apply TIPSem-B only morphosyntactic information is required. Only the normalization of time expressions is a language dependent process in our system and requires the construction of a set of rules for each target language.

The contribution of semantic information to temporal information treatment was more significant in recall and the improvement was concentrated in tasks A and B (approx. 12% recall improvement). Although, TIPSem-B achieved lower results they are high enough to confirm that that most of temporal elements strongly depends on lexical and morphosyntactic information.

The main errors and difficulties of our approach in this evaluation exercise are related to TIMEX3 normalization (value attribute). A pure ML approach for solving this problem is not trivial, at least, using our approach philosophy. The treatment of normalization functions is an inherently complex task and requires many training data to be automatically learned. This required us to include in the system some handcrafted rules to enable the system for this task.

As further work we propose improving the TIMEX3 normalization by replacing handcrafted normalization rules with machine learned ones by combining statistic techniques and multilingual temporal knowledge resources (ontologies). Furthermore, link-categorization will be analyzed in more detail in order to include more features to improve the models. Finally, the suggested language independence of the approach will be tested using TempEval-2 available data for other languages.

Acknowledgments

This paper has been supported by the Spanish Government, projects TIN-2006-15265-C06-01, TIN-2009-13391-C04-01 and PROMETEO/2009/119, where Hector Llorens is funded under a FPI grant (BES-2007-16256).

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *43rd Annual Meeting of the ACL*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th ICML*, pages 282–289. Morgan Kaufmann.

Hector Llorens, Borja Navarro, and Estela Saquete. 2009. Detección de Expresiones Temporales TimeML en Catalán mediante Roles Semánticos y Redes Semánticas. In *Procesamiento del Lenguaje Natural (SEPLN)*, number 43, pages 13–21.

Paloma Moreda, Borja Navarro, and Manuel Palomar. 2007. Corpus-based semantic role approach in information retrieval. *Data Knowledge Engineering*, 61(3):467–483.

Vasin Punyakanok, Dan Roth, W. Yih, D. Zimak, and Y. Tu. 2004. Semantic role labeling via generalized inference over classifiers. In *HLT-NAACL (CoNLL)*, pages 130–133. ACL.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *IWCS-5*.

Frank Schilder, Graham Katz, and James Pustejovsky. 2007. *Annotating, Extracting and Reasoning About Time and Events (Dagstuhl 2005)*, volume 4795 of *LNCS*. Springer.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Mariona Taulé, M. Antonia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In ELRA, editor, *LREC*, Marrakech, Morocco.

Marc Verhagen, Robert Gaizauskas, Mark Hepple, Frank Schilder, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, Prague. ACL.

Piek Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, MA, USA.