Information Processing and Management 49 (2013) 179–197

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language

Hector Llorens *, Estela Saquete, Borja Navarro-Colorado

Natural Language Processing and Information Systems Group, University of Alicante, P.O. Box 99, E-03080 Alicante, Spain

ARTICLE INFO

Article history:
Received 22 December 2010
Received in revised form 6 February 2012
Accepted 27 May 2012
Available online 18 June 2012

Keywords:
Temporal information processing
Semantic roles
Semantic networks
TimeML

ABSTRACT

This paper addresses the problem of the automatic recognition and classification of temporal expressions and events in human language. Efficacy in these tasks is crucial if the broader task of temporal information processing is to be successfully performed. We analyze whether the application of semantic knowledge to these tasks improves the performance of current approaches. We therefore present and evaluate a data-driven approach as part of a system: *TIPSem*. Our approach uses lexical semantics and semantic roles as additional information to extend classical approaches which are principally based on morphosyntax. The results obtained for English show that semantic knowledge aids in temporal expression and event recognition, achieving an error reduction of 59% and 21%, while in classification the contribution is limited. From the analysis of the results it may be concluded that the application of semantic knowledge leads to more general models and aids in the recognition of temporal entities that are ambiguous at shallower language analysis levels. We also discovered that lexical semantics and semantic roles have complementary advantages, and that it is useful to combine them. Finally, we carried out the same analysis for Spanish. The results obtained show comparable advantages. This supports the hypothesis that applying the proposed semantic knowledge may be useful for different languages.

1. Introduction

This paper focuses on the temporal information processing task. Given a text in natural language, the objective of this task is the automatic anchoring and ordering of its events in time. This task can be divided into three steps: (i) recognizing (bounding in text) and classifying temporal expressions (*timex*es) and events; (ii) normalizing each timex using a structured format – that is independent of its linguistic expression; and (iii) obtaining the temporal relations between events and timexes in order to anchor and order the events in time.

For example (1) contains a timex (*October 1999*) and an event (*went*); the normalized value for *October 1999* is 1999–1910; and there is a temporal relation – the time of the event *went* is included in the 1999–1910 time-span.

(1) In [October 1999 $_{timex\ (class:date)}$], John [went $_{event\ (class:\ occurrence)}$] to Australia.

The interest in the temporal aspects of natural language is not new in the field of artificial intelligence (Allen, 1983). However, it is still a hot research topic, as is reflected in recent specialized evaluation forums such as TempEval-2 (Verhagen, Saurí, Caselli, & Pustejovsky, 2010).

* Corresponding author. Tel.: +34 965903400.
E-mail address: hllorens@dlsi.ua.es (H. Llorens).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.05.005

Temporal information processing has various applications. These include, but are not limited to, the following. In question answering, temporal reasoning is required to allow complex questions to be answered, such as "*Who was the president of US ten years ago?*" and "*What happened to world oil prices after the Iraqi annexation of Kuwait?*" (Saquete, González, Martínez-Barco, Muñoz, & Llorens, 2009). In text summarization, the ability to allocate events in time aids in obtaining better summaries when these have to be focused on a particular time period, or on specific events (Daniel, Radev, & Allison, 2003). In information retrieval, the value of the temporal dimension has been emphasized both in research (Alonso, Gertz, & Baeza-Yates, 2007) and in real applications (e.g., Google Timeline).

When applying temporal information processing to other tasks the performance of the computational approaches must be high. The subtask of recognizing and classifying timexes and events is crucial. If this first step fails, the performance of the whole process is negatively affected. The computational approaches to this subtask are currently based principally on morphosyntactic knowledge. These approaches have difficulty in coping with semantic ambiguity and generalization. Our hypothesis is that approaches using semantic knowledge will be (i) capable of solving the morphosyntactic ambiguity of some timexes and events; (ii) more general than those based solely on morphosyntax owing to the higher abstraction level of semantics as compared to morphosyntax; and (iii) useful for multiple languages.

The aim of this paper is to apply semantic knowledge as additional information in order to improve the efficacy of current temporal information processing techniques. We therefore present a data-driven approach: *TIPSem*(Temporal Information Processing using Semantics). This is based on semantic knowledge (semantic networks and semantic roles) plus morphosyntactic knowledge. In order to measure the influence of semantics in a precise manner, our approach and a baseline (TIPSem-B), which is based only on morphosyntactic knowledge, are evaluated and compared. Furthermore, in order to test the validity of the hypothesis in different languages, the approaches are developed and evaluated for English and Spanish.

This paper completes two previous works in this line of research: (i) our earlier version of TIPSem evaluated in TempEval-2 (Llorens, Saquete, & Navarro-Colorado, 2010b) and (ii) our initial data-driven implementation for English event processing (Llorens, Saquete, & Navarro-Colorado, 2010a). With regard to (i), this paper presents a new version which tackles timexes and events separately. This permits the semantics of each entity to be captured better, and in fact, improves upon the performance obtained in TempEval-2. With regard to (ii), we present an improved event processing approach which we evaluate over the new TempEval-2 dataset. TempEval-2 represents a fair benchmark for the comparison of different approaches. Moreover, we analyze lexical semantics and semantic roles independently, and from a multilingual perspective.

The following section includes the background to and the motivation for the use of semantics. Section 3 provides a detailed description of our proposal. Section 4 includes its evaluation and a comparative analysis of the results for English and Spanish. Finally, in Section 5, we draw conclusions and make suggestions for further studies.

2. Background and motivation

This section includes (i) a description of the schemes that are available for the annotation of temporal information in natural language, (ii) a review of the work related to the computational recognition and classification of timexes and events, and (iii) a discussion concerning our motivation for using specific semantic knowledge which differentiates our approach from those reviewed in the related work.

2.1. Temporal information annotation in natural language

The aim of studying temporal information in natural language has led the scientific community to define different temporal entities such as timexes and events.

Several efforts have been made to define standard ways in which to annotate the temporal information in texts. The main objective of annotation is to make temporal information explicit through standard schemes. Since temporal information extraction was first included in the context of Message Understanding Conference (MUC) (Grishman & Sundheim, 1996), there have been three important annotation schemes: TIDES, STAG, and TimeML. TIDES (Ferro, Mani, Sundheim, & Wilson, 2000) emerged in order to annotate only temporal expressions and their normalized values in text. STAG (Setzer & Gaizauskas, 2000) was defined with the aim of identifying events in news and their relationship with points on a temporal line. Finally, TimeML (Pustejovsky et al., 2003a) combined and extended features of both the aforementioned schemes.

This paper is focused on the TimeML[1] annotation scheme because it has been adopted as a standard by a large number of researchers owing to its completeness and the comprehensive improvements it makes on the previous schemes. Furthermore, TimeML is available for various languages including those tackled in this paper: English and Spanish. Example (2) shows a sentence annotated with TimeML temporal expressions (`TIMEX3`), events (`EVENT`), and their classes.

(2) John <EVENT class="OCCURRENCE">came</EVENT> on <TIMEX3 type="DATE">Monday</TIMEX3>

[1] TimeML annotation scheme official website: http://timeml.org/.

2.2. Related Work

Several computational approaches address the automatic recognition and classification of temporal expressions and events. In order to summarize the sate-of-the-art, we analyze the current temporal information processing systems as regards their strategies, the temporal entities covered, and the linguistic knowledge used (see Table 1). Furthermore, in order to provide a broader comparison, we analyze the performance results that these approaches obtained in the last international evaluation (see Table 2).

In Table 1, we consider *morphosyntactic (ms)* knowledge to be the use of all or part of the following linguistic information: sentence segmentation, tokenization, lemmatization, PoS tagging, syntactic parsing, and word-triggers[2]. *Lexical semantics* refers to the usage of resources related to the meaning of individual words and their lexical semantic relations (e.g., hypernymy), such as semantic networks like WordNet (Fellbaum, 1998). Finally, *semantic-logical form* and *semantic roles* refer to compositional semantics at sentence level and will be discussed later.

In Table 2, we detail the results obtained by the best approaches in the TempEval-2 evaluation exercise (Verhagen et al., 2010) in order to compare their efficacy.

As is shown in Table 2, the best scores for timex recognition were obtained by the HeidelTime system, closely followed by TIPSem and TRIOS. For event processing, the best scores were obtained by TIPSem, followed by the Edinburgh-LTG system.

With regard to the strategies used, many of the systems are rule-based, particularly those which only address timex processing. Rule-based and data-driven approaches obtained a very close performance in timex processing, while data-driven and hybrid systems were more popular among those approaches that tackled event processing, and show better results than rule-based systems in this task.

In particular, among the data-driven and hybrid approaches, the most popular machine learning techniques are: (i) support vector machines (SVM) which are used by March et al. and STEP and (ii) conditional random fields (CRF) which are used by TRIOS and TIPSem. SVMs (Cortes & Vapnik, 1995) are widely known classifiers which have been successfully applied to various natural language processing (NLP) tasks. CRF (Lafferty, McCallum, & Pereira, 2001) is a graph-based technique that is designed to learn sequence labeling problems. This makes it useful for recognizing timexes and events, because they appear in word sequences.

Regarding the linguistic knowledge used, most systems are based on morphosyntax. Some of them include lexical semantics and only two of them apply compositional semantics (TRIOS and TIPSem). The results obtained by those systems that use only morphosyntax are reasonably high for timex processing (HeidelTime), while their performance is lower for event processing.

Finally, with regard to the target language, the approaches reviewed were only developed for English, with the exception of our approach (TIPSem) which tackles English and Spanish. For Spanish, TIPSem obtained the best results for timex and event processing. Only one other system participated in TempEval-2 for Spanish, UC3M (Vicente-Díez, Moreno-Schneider, & Martínez, 2010). UC3M is a rule-based system which uses morphosyntactic knowledge. It addresses timex processing, but not event processing. Our approach is therefore the most complete approach that is currently available for Spanish.

2.3. Motivation for applying semantic knowledge

As described previously, our proposal is focused on the use of semantic knowledge. The following subsections report on each type of semantic knowledge applied to timex and event processing. We first describe the application of lexical semantics and, we then explain the application of semantic roles, which are our main focus.

2.3.1. Lexical semantics

As previously mentioned, many of the approaches that tackle the proposed tasks are based on morphosyntactic knowledge. Furthermore, they achieved a high performance when using morphosyntax, particularly in the timex processing task. We attribute the high performance obtained to the inclusion of word-triggers. These predefined lists of keywords that are likely to appear within timexes or events have proven to be useful for recognition.

From our point of view, the application of word-triggers could be considered a naive form of domain-oriented lexical semantics. However, the application of general lexical semantic resources such as semantic networks has two advantages over word-triggers. First, a general resource like WordNet (Fellbaum, 1998) captures not only the lexical semantics of a word (as a char sequence) in a specific domain (e.g., time/eventuality) but the semantics of the word (as a specific concept or sense) within a general domain, encoded in a lexicon with a semantic network structure. Secondly, general semantic networks tend to be standard and complete, and they are maintained and improved by third parties independently of our system. Their use increases the modularity of the approaches and avoids the costly task of developing a complete trigger-list.

WordNet and EuroWordNet (Vossen, 1998) are the most popular semantic networks for English and European languages respectively. Semantic networks have been widely used in the temporal expression identification task. Negri and Marseglia (2004) used WordNet to create a list of named time expressions such as "Bastille Day" and "Hanukkah", by collecting all the hyponyms of the *calendar_day* synset. In Saquete, Martínez-Barco, and Muñoz (2004), semantic networks were used to expand a list of temporal triggers by using their synonyms. Finally, Kolomiyets, Bethard, and Moens (2011) explored the use of synonyms of timex words from WordNet and also from the Latent Words Language Model, which predicts synonyms in context using an unsupervised approach. These authors also show that the combination of both resources provides more stable results.

[2] A predefined list of keywords that are likely to appear within a temporal expression or event (e.g., "year", "war").

Table 1
TimeML timex and event processing approaches. Abbreviations: r (recognition), c (classification).

Strategy	System	Tasks	Linguistic knowledge
Rule-based	JU_CSE_TEMP (Kolya et al., 2010)	both (r & c)	morphosyntactic (ms)
	HeidelTime (Strötgen & Gertz, 2010)	timex (r & c)	ms
	USFD2 (Derczynski & Gaizauskas, 2010)	timex (r & c)	ms
	TERSEO (Saquete Boro, 2010)	timex (r & c)	ms
	Edinburgh-LTG (Grover et al., 2010)	both (r & c)	ms, lexical semantics
	TTK (Verhagen et al., 2005)	both (r & c)	ms, lexical semantics
Data-driven	(Boguraev & Ando, 2005)	both (r & c)	ms
	(March & Baldwin, 2008)	event (r)	ms
	STEP (Bethard & Martin, 2006)	event (r & c)	ms, lexical semantics
	TIPSem (Llorens et al., 2010b)	both (r & c)	ms, lexical& semantic roles
Hybrid	KUL (Kolomiyets & Moens, 2010)	timex (r & c)	ms, lexical semantics
	TRIOS (UzZaman & Allen, 2010)	both (r & c)	ms, semantic-logical form

[a]TTK is mainly rule-based but uses some statistical techniques for disambiguation in event recognition.

Table 2
TempEval-2 top-systems for English (*norm = recall∗class_score*).

Element	System	Recognition			Classification score (norm)
		Precision	Recall	$F_{\beta=1}$	
Timex	HeidelTime	0.90	0.82	0.86	0.96 (0.79)
	TRIOS	0.85	0.85	0.85	0.94 (0.80)
	TIPSem	0.92	0.80	0.85	0.92 (0.74)
Event	TIPSem	0.81	0.86	0.83	0.79 (0.68)
	Edinburgh-LTG	0.75	0.85	0.80	0.76 (0.61)

Various proposals use WordNet (Saurí, Knippen, Verhagen, & Pustejovsky, 2005; Bethard & Martin, 2006; Grover, Tobin, Alex, & Byrne, 2010) in event recognition and classification. These proposals derive features or trigger-lists from *event* and *state* hyponyms.

2.3.2. Semantic roles

Semantic roles capture the meaning of a sentence as regards how their arguments are related. A semantic role is the semantic relationship held between a syntactic constituent and a predicate. For each predicate, all the constituents are labeled as arguments (agent, patient, etc.) or adjuncts (locative, temporal, etc.). Example (3) shows some sentences labeled with semantic roles[3] (in square-brackets) in which the timexes and events are underlined. In (3a), with regard to the verb heading the predicate (to write), AM-TMP refers to the temporal adjunct (*when* it took place), A0 represents the agent argument (writer), A1 is the patient argument (thing written), AM-MNR is the manner adjunct (*how* was it done), and AM-LOC the locative adjunct (*where*). In (3d), AM-CAU represents the causative adjunct (*why*).

(3)
a. [In July 1277 $_{AM-TMP}$], [John $_{A0}$] wrote [a letter $_{A1}$] [with a quill $_{AM-MNR}$] [in his chamber $_{AM-LOC}$].
b. [The tax rates $_{A1}$] were [lower $_{A2}$].
c. [The company $_{A0}$] completed [the transaction $_{A1}$] [yesterday $_{AM-TMP}$].
d. [Iraq $_{A0}$] invaded [Kuwait $_{A1}$] [because of the disputes over oil $_{AM-CAU}$].

With regard to the relationship between semantic roles and temporal entities, in general lines, it will be noted that main verbs representing events and temporal expressions tend to be contained by temporal adjuncts. Furthermore, the verb "to be" does not represent an event but may describe stative-events in its A2 role (3b). Finally, non-verbal events are more likely to appear under specific adjuncts (3d), or to play the A1 role of certain verbs (3c).

The first modern proposals concerning semantic roles were enumerated by Fillmore (1968), Fillmore (1971). These proposals have led to the proliferation of different sets of semantic roles. Proposals exist which vary from the very specific, with many different roles (e.g., eater, eaten), to the very abstract, consisting of few roles (e.g., agent, temporal). Some references (Dowty, 1991) argue that a hierarchy of roles can be established from the more detailed ones to only two *proto-roles*: proto-agent and proto-patient.

[3] Labeled using PropBank roleset (Palmer, Gildea, & Kingsbury, 2005).

In NLP, the most widely used semantic role sets are the set developed in the FrameNet (FN) project (Baker, Fillmore, & Lowe, 1998) and the set defined in the Proposition Bank (PB) project (Palmer et al., 2005). Each proposal focuses on a different role granularity. FN defines a detailed representation of situations including a set of very specific roles (frames), while PB consists of a limited set of abstract roles. FN avoids the problem of defining a small set of abstract roles by defining as many roles as necessary with minimal information loss. However, the PB set offers a wider lexical coverage than FN, which only covers most general English verbs and nouns.

Semantic role labeling (SRL) has achieved important results in recent years (Gildea & Jurafsky, 2002). Many studies concerning the application of semantic roles have demonstrated that this information is very useful for different NLP purposes (Melli, Shi, Wang, & Popowich, 2006; Moreda, Llorens, Saquete, & Palomar, 2011). With regard to the application of roles to timex and event processing, they are only used by one other system apart from ours: the TRIOS system. TRIOS uses a logical form representation of semantic roles based on a variation of FrameNet. The variation reduces the FrameNet roleset to approximately 2500 semantic types and 30 semantic roles.

In this paper, we cover the following issues not tackled in the state-of-the-art. The few approaches that include semantics are focused on the use of lexical semantics and only one of them (TRIOS) includes semantic roles (FrameNet roles). The influence of semantics on performance has not, therefore, been analyzed in depth. We analyze the PropBank role set which has never been used for temporal information processing. Furthermore, we analyze the effects of combining lexical semantics and semantic roles.

3. Our proposal: TIPSem

We tackle the automatic recognition and classification of timexes and events in natural language. This is done by following the TimeML annotation scheme and involves bounding them in the text using XML tags and assigning each one a class from the TimeML taxonomy. Example (4) illustrates a raw text input and the expected output after the processing described.

(4) **Input**: John came on Monday. After, he left US. Then, he told us he was born in 1980.
Output:`John <event class="OCCURRENCE">came</event> on <timex3 type="DATE">Monday</timex3>.After, he <event class="OCCURRENCE">left</event> US. Then, he <event class="REPORTING">told</event> us he was <event class="OCCURRENCE">born</event> in <timex3 type="DATE">1980</timex3>.`

The objectives of our proposal are: (i) the application of lexical semantics and semantic roles in addition to morphosyntax to these tasks in order to analyze the possible advantages of this linguistic information and (ii) the realization of a multilingual study of the proposal, for English and Spanish, in order to ascertain its degree of language independence.

Prior to developing the approach, we considered two possible strategies for the application of semantics: rule-based and data-driven. The rule-based strategy can be applied to timex recognition owing to the relationship that timexes have with temporal semantic classes and semantic roles (Llorens, Navarro, & Saquete, 2009a; Llorens, Saquete, & Navarro, 2009b). However, the relatedness of semantics with events is more complex, particularly with regard to semantic roles. The manual construction of high coverage rules is costly and is limited to our prior knowledge of the task addressed.

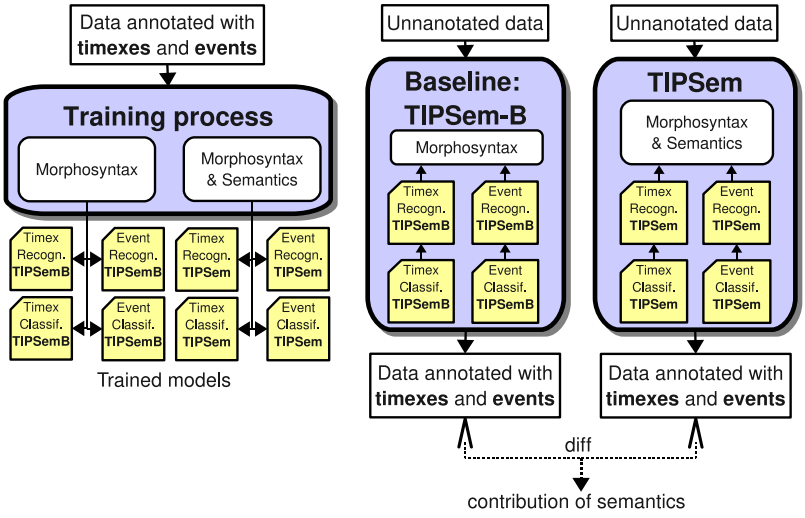

Fig. 1. Proposed architecture.

The data-driven strategy is becoming common among related works since corpora are available (e.g., TempEval-2 dataset). This strategy prevents us from coding rules based on our prior knowledge, which may be incomplete or inconsistent with the real examples – especially when the task is complex, as in the case of event processing.

Given this analysis, we chose to develop TIPSem as a data-driven approach. We then analyzed two possible ways in which to carry out recognition and classification: classify-after-recognition and recognition-of-classes. The former is a two step architecture: first, bound the element (e.g., timex) and second classify it (e.g., duration). The latter merges recognition and classification by directly recognizing the element as its class (e.g., timex-duration). After carrying out various experiments, we noted that the two step strategy performs better given the small size of the available corpora which makes it more appropriate for the construction of good general recognizers and the subsequent classification of the recognized entities. TIPSem has therefore been developed as a classify-after-recognition approach.

The resulting TIPSem architecture is summarized in Fig. 1. This is multilingual, and is equivalent for English and Spanish.

In the proposal, first, given annotated examples in one of the languages, four models are trained: timex recognition, timex classification, event recognition and event classification. Once the models have been generated, the system uses them to annotate any input raw text. We trained two approaches, TIPSem (semantics + morphosyntax) and TIPSem-B (morphosyntax). The difference in the correctness of their outputs will reflect the contribution of semantics.

The external tools and resources used to obtain the morphosyntactic and semantic features for each language are the following. For English, the morphological features are obtained from TreeTagger (Schmid, 1994), lexical semantics from WordNet (Fellbaum, 1998), and syntactic and semantic roles (PropBank) features from the CCG SRL tool (Punyakanok, Roth, Yih, Zimak, & Tu, 2004). For Spanish, all the features are extracted from the AnCora corpus (Taulé, Martí, & Recasens, 2008), with the exception of lexical semantics, which are extracted from EuroWordNet (Vossen, 1998).

3.1. Learning models for the recognition and classification of temporal expressions (Timexes)

We tackled timex recognition and timex classification separately as two sequential steps.

3.1.1. Timex recognition

A timex is a linguistic representation of a time point or period. It is normally denoted by a noun, adjective, adverb, or a noun phrase, adjective phrase, or adverb phrase. The timexes shown in example (5) are underlined.

(5)
a. John came on Monday and he will leave on October 9th, 2012.
b. She will be here for two weeks.
c. The show starts at 8 p.m.
d. I visit my parents monthly.

Recognition implies the bounding of TimeML time expressions in the text. This problem can be viewed as a sequence labeling problem. Given an input text (token sequence), each token must be labeled with an IOB2 value (label sequence) as being the beginning of a timex (B-timex), inside a timex (I-timex), or outside any timex (O). Example (6) illustrates the recognition task.

(6)

input text	solution
There	O
was	O
no	O
profit	O
this	B-timex
year	I-timex
.	O

The aforementioned problem is tackled with the following **morphosyntactic features**.

- **Morphological**: The lemma and PoS context, in a 5-window (−2,+2), is employed. This proved to have a better performance than considering only the information of each token individually (without context), or a smaller (3-window) or larger context (7-window).
  The lemma is crucial for the recognition of time expressions, such as that in (7).

(7) The conference ended [$yesterday_{TIMEX3}$].

As is shown in the example, some lemmas, such as "yesterday", are very likely to represent timexes[4]. However, example (8) shows a temporal expression which may be ambiguous at a lexical level, "May[5]".

(8) *May* you find what you are looking for.
She will be back in [*May* $_{TIMEX3}$].

Here, the information concerning the word class (PoS) solves the ambiguity owing to the fact that the first "May" is a modal verb while the second is a proper noun (a month).
The context window is also very important for the recognition of multi-token expressions, because some tokens only represent a temporal expression if their neighboring tokens hold specific lemmas and PoS – see example (9).

(9) This is the chief *of* the company.
Today is the [3rd *of* June $_{TIMEX3}$].

In this example, the token "of" only becomes part of a temporal expression if, for instance, its left neighboring word is an ordinal and its right neighboring word is a proper noun with a specific lemma (a month), but not if its left neighboring word is "chief" and its right neighboring word is "the". The 5-window also acts as contextual disambiguation information. In example (10), we can see how a number is more likely to be part of a temporal expression if it is preceded by a preposition and not followed by a noun.

(10) He owns 1999 properties.
He was born in [1999 $_{TIMEX3}$].

Although morphological information is very useful, it is not always sufficient to allow temporal expressions to be distinguished from the rest of the text.

- **Syntactic**: Time expressions are contained in specific types of phrases (i.e., NP, ADJP, or ADVP). In timex recognition, we include a feature whose value indicates whether a token belongs to one of these phrases. This is useful for identifying which tokens might be part of a timex. In the example (11a), this feature indicates that "stayed" and "during" cannot participate in a timex because they do not belong to any of these types of phrases.

(11) a. (S (NP She) (VP stayed (PP during (NP [the second half of the year $_{TIMEX3}$])))). (timex)
b. (S (NP She) (VP stayed (PP behind (NP the wall)))). (not timex)

Furthermore, if an NP is governed by a PP, the heading prepositions may be useful to increase the probability of the NP containing a timex, and it is therefore also included as a feature. In the example (11), the PP headed by "during" introduces a timex in the NP it governs (11a), while the PP headed by "behind" does not (11b). However, examples (12a) and (12b) show the same syntactic structure, but only (12a) represents a timex.

(12) a. (S (NP She) (VP won (PP in (NP April))). (*timex*)
b. (S (NP She) (VP won (PP in (NP Africa)))). (*not timex*)

This motivates the use of higher level linguistic information like semantics. Lexical semantics indicate that the concept "April" is related to time and semantic roles show that "in April" plays a temporal role.

The **semantic features** used to enhance the timex recognition model are:

- **Lexical Semantics**: The top ontology classes of WordNet have been widely used to represent word meaning at the ontological level. Most of the nouns contained in timexes are hyponyms of **time**, **time_period** or **time_unit**. These concepts are distributed between the third and the forth level from the top concept, *entity*. Our approach therefore considers the top four hypernyms (*top4hypers*) of the most common sense for every noun and verb as a feature, excluding the top concept – see example (13).

[4] Except when they do not have a temporal meaning, as with The Beatles' song title.
[5] It can be also ambiguous at morphosyntactic level if it refers to a person's name.

(13) The *top4hypers* for "decades" is:
decade: **time_period** => fundamental_quantity => measure => abstraction

Since many of the timexes contain tokens with time-related values for *top4hypers*, this feature will increase the probability of representing timexes for tokens that obtain such values, even if their lemmas do not appear in training, which favors generalization.

- **Semantic roles**: Semantic roles provide the structural semantic relations of the predicates in which TimeML elements might participate. The temporal semantic role (***AM-TMP***) represents the temporal adjunct and often contains a time expression. For each token in a predicate, we consider the *role* it plays as a feature. Our hypothesis is that the *role* feature will increase the probability of all the tokens belonging to the AM-TMP role being timexes, regardless their lemmas. Furthermore, the combination *role* + *lemma* is included as a feature. This will specifically increase the probability of being able to represent a timex of the combinations of timex-related lemmas (e.g., April, week, etc.) with an AM-TMP role. Semantic roles are also useful for tackling morphosyntactic ambiguity –see example (14).

(14)		
	[April $_{A0}$] likes apples.	(female proper name)
	I went to Canada [in April $_{AM-TMP}$].	(month, TIMEX3)
	The Iraqis have resisted attempts to inspect [such quarters $_{AM-LOC}$].	(location)
	[The noun "Monday" $_{A0}$] is derived from Middle English *Monenday*.	(entity name)
	["Last week" $_{A0}$] is my favorite song.	(song title)

The first two sentences contain the noun "April". However, in the first sentence "April" is a person named entity and in the second it is a TIMEX3. At the semantic role level, the expression "April" represents a numbered argument role of the verb "*to like*" in the first sentence while it represents a temporal role in the second. Moreover, in the third sentence the presence of a locative role prevents us from considering "quarters" as a temporal expression. Finally, the expressions "Monday" and "Last week" do not necessarily represent temporal expressions, as is shown in the example. The semantic role labeling provide valuable information by marking them in a non-temporal semantic role (i.e., A0). In short, the approach must learn from the annotated examples which semantic roles are more likely to contain temporal expressions.

Fig. 2 illustrates the feature-vector described.

3.1.2. Timex classification

According to TimeML scheme, timexes can be classified in four classes:

- *Date*. Expressions with date granularities (e.g., "Monday").
- *Duration*. Periods of time (e.g., "two weeks").
- *Time*. Expressions with time granularities (e.g., "8 p.m.").
- *Set*. Recurring patterns in time (e.g., "monthly").

Classification involves the selection of one type for each recognized timex – see example (15).

(15)	TIMEX	solution
	1999	date
	8 a.m.	time
	monthly	set
	two years	duration
	next Monday	date

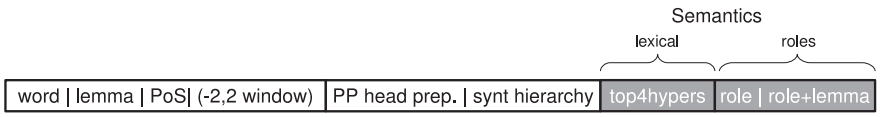

Fig. 2. Timex recognition features.

Since, in this case, we have a set of timexes rather than a set of tokens, the features previously defined have to be adapted. In order to translate token-based features to element-based features, we clustered the recognition features in the following manner. For timexes spanning one token (e.g., yesterday), the features remain the same. For timexes spanning more than one token (e.g., "two years"), the feature values for each token are concatenated by "_" (e.g., lemma = two_year,PoS = CD_NNS), and if the values are shared by all the tokens (e.g., role (two) = AM-TMP, role (years) = AM-TMP) then they are clustered in just one value (e.g., role (two years) = AM-TMP).

3.2. Learning models for the recognition and classification of events

Event recognition and classification tasks are defined in the same way as the timex tasks.

3.2.1. Event recognition

An event is a linguistic expression of something that *happens, occurs, or takes place*, or something that obtains or holds the truth for a time-interval, also called a state. Events can be punctual or can last for a period of time. In natural language, events are generally expressed by verbs, nominalizations, adjectives, predicative clauses or prepositional phrases. Some events are underlined in example (16).

(16) a. John came on Monday.
b. He will attend the conference.
c. It's turning out to be another bad financial week, he said.

We address event recognition as illustrated in example (17).

(17)

input text	solution
He	O
said	B-event
she	O
likes	B-event
him	O

This problem is tackled by defining the following **morphosyntactic features**.

- **Morphological**: The lemma and PoS context, in a 5-window (−2,+2), is again employed. Events are generally expressed by verbs, nominalizations, adjectives, predicative clauses or prepositional phrases (e.g., "born", "said", "war", "match"). Given this definition, PoS may assist in the recognition of verbs, and specific lemmas may identify certain nouns that represent events such as "war" or "conference". However, the performance in detecting unseen nominal events may be weaker. Example (18) illustrates some cases.

(18) He went to the house.
He went to the [meeting $_{EVENT}$].
He went to the [conference $_{EVENT}$].

In this example we find three sentences in which the sole difference at the morphosyntactic level is the last lemma. Imagine that "meeting" was unseen in the training examples while "conference" and "house" were seen. The sole use of the lemma and the PoS, does not therefore make it possible to decide whether or not "meeting" should be an event.

- **Syntactic**: The syntactic structure aids in the recognition of events. Example (19) shows sentences which may benefit from the consideration of syntactic information.

(19) a. (S (NP America) (VP [has $_{EVENT}$] (NP the tanks))).
b. (S (NP America) (VP has (VP been (VP sending (NP troops))))).

In (19), "has" is either an event or simply an auxiliary verb depending on the VP hierarchy.

The **semantic features** used are:

- **Lexical Semantics**: Most verbs in a text denote events. However, only some nouns denote events in particular semantic settings. Semantic networks provide useful information with which to recognize nominal events. Many nominal events are hyponyms of **event**, **act** or **state**. These concepts are distributed between the fourth and the fifth level from the top concept (*entity*) in the WordNet hyponym hierarchy. Our approach therefore considers the top four hypernyms – excluding the top hypernym *entity* – for the most common sense for every word as a feature – see example (20).

(20) "Ritter led his team in a 10-h tour".
The *top4hypers* value for "tour" is:
tour => act => **event** => psychological_feature => abstraction

Since many of the events contain tokens that obtain the same values for *top4hypers*, this feature will increase the probability of representing events for tokens that obtain these values, even if their lemmas do not appear in training.

- **Semantic roles**: The relation between semantic roles and events is two fold: (i) many events are denoted by main verbs which will have arguments that play different roles and (ii) events also appear under specific roles for particular verbs. We therefore consider that semantic roles can be used to recognize and classify events.
  We used four role-based features in our approach:
  - *role* represents the semantic role each token plays in a predicate. Apart from the main verbs (those that govern arguments in predicates) that represent many events, this feature learns the probability of events appearing in different roles. (21) shows three sentences labeled with semantic roles with the events underlined:

(21) a. [The tax rates $_{A1}$] were [lower $_{A2}$].
b. [The company $_{A0}$] completed [the transaction $_{A1}$].
c. [It $_{A0}$] [might $_{AM-MOD}$] drop out [the effort against Iraq $_{A4}$].

In the example, (21a) shows a event "lower" appearing in the A2 role. In (21b), the A1 role of the verb "to complete" (task, action coming to an end) indicates a nominal event "transaction". Finally (21c), shows a sentence in which none of the arguments (A0, AM-MOD, A4) contain an event.
This shows that events normally appear as main verbs, when they are verbal, and under the influence of certain roles when they are represented by other word-classes such as a noun or an adjective. Obviously, the fact of finding an A2 role or an A1 role does not guarantee that an event will be found, and the fact of finding an A0, AM-MOD or A4 does not mean that it is impossible to find an event. However, the probability of finding an event within A1 or A2 is higher.

  - *role + word* is the combination of the token and the semantic role it plays:

(22) a. [He $_{A0}$] made [an offer $_{A1}$].
b. [He $_{A0}$] made [a cake $_{A1}$].
c. [He $_{A0}$] completed [the transaction$_{A1}$].
d. [He $_{A0}$] completed [the puzzle $_{A1}$].

In the example sentences (22a) and (22b), both "offer" and "cake" are nouns playing the A1 role. However, only "offer" represents an event. Similarly, in sentences (22c) and (22d), "transaction" and "puzzle" play the A1 role but only "transaction" represents an event. This is because not all the words can represent events, but only those signifying actions, processes or states. This feature merges the word and the role it plays to capture this information.

  - *role + verb* is the combination of the semantic role and the governing verb of the predicate. This distinguishes roles depending on specific verbs. This is particularly important in numbered roles (A0, A1, etc.), which might signify different things when governed by different verbs:

(23) a. [The tax rates $_{A1}$] were [lower $_{A2}$].
b. [The tax rates $_{A0}$] raised [10% $_{A2}$] [to 11 dollars $_{A4}$] [from 10 $_{A3}$].
c. [He $_{A0}$] carried out [an experiment $_{A1}$].
d. [He $_{A0}$] ate [an apple $_{A1}$].

As is shown in sentences (23a) and (23b), the A2 role of the verb "to be" is more likely to contain an event than the A2 role of the verb "to rise". In the same manner, the A1 role of "to carry out" is more likely to contain an action or process than the A1 role of "to eat" which represents the thing eaten.

– *role + top4hypers* is the combination of the role and the *top4hypers* feature of the token. This feature generalizes the *role + word*, thus obtaining the same value for all the tokens sharing the top four hypernyms and appearing in the same role. For example, the words: "assembly", "mobilization", "convocation", "meeting", "congregation" and "convention" appear within the A0 role, they obtain different values for the feature *role + word*, while they obtain the same value `A0 + group_action=> event=> psychological_feature=> abstraction` for *role + top4hypers*. This favors the model in generalizing that event-related *top4hpers* are likely to represent events when they appear in specific roles (e.g., A1).

Fig. 3 illustrates the features described.

3.2.2. Event classification

According to TimeML scheme, there are seven classes of events:

- *Reporting*. Action of declaring or narrating an event (e.g., "say", "report").
- *Perception*. Physical perception of another event (e.g., "said", "hear").
- *Aspectual*. Aspectual predication of another event (e.g., "start", "continue").
- *I_Action*. Intentional action (e.g., "try","attempt").
- *I_State*. Intentional state (e.g., "feel", "hope").
- *State*. Circumstance in which something holds the truth (e.g., "rainy", "bad").
- *Occurrence*. Events that describe things that happen (e.g., "came", "conference").

Event classification, as illustrated in (24), consists of assigning a TimeML event-class to each event. The same strategy used for timex has been followed to translate the token-based features defined for event recognition into element based features for event classification.

(24)

event	solution
went	occurrence
said	reporting
meeting	occurrence
started	aspectual
tried	i_action

3.3. Machine learning techniques

As described previously, TIPSem implements a two-step architecture. It performs first the recognition and then the classification of timexes and events.

On the one hand, the recognition step is defined as a sequence labeling problem. Of the techniques that are available, our approach employs conditional random fields (CRF) (Lafferty et al., 2001) to infer models with which to address these tasks. CRFs are undirected graphical models, a special case of conditionally-trained finite state machines. This is a popular and efficient supervised learning technique for sequence labeling and is thus appropriate for our purpose. Not only the word sequence, but also the morphological, syntactic and semantic structure, which are present in the previously described features of our approach, benefit from the use of this learning technique. CRF has also been used to learn timex processing models in works related to temporal information processing (Ahn, van Rantwijk, & de Rijke, 2007; UzZaman & Allen, 2010; Kolya, Ekbal, & Bandyopadhyay, 2010).

In the tasks tackled in this paper, we assume that X is a random variable over the data sequences to be labeled, and Y is a random variable over the corresponding label sequences, where all Y components (Y_i) are members of a finite label alphabet γ. X might range over the sentences and Y might range over possible annotations of these sentences, with γ being the set of

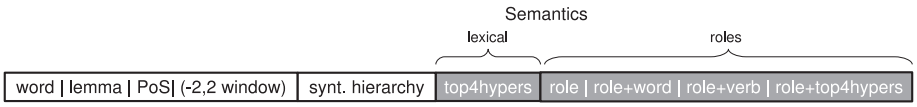

Fig. 3. Event recognition features.

possible IOB2 labels. The random variables X and Y are jointly distributed, and CRFs construct a conditional model from paired observation and label sequences: $p(Y|X)$. The CRF++ toolkit[6] was used in the implementation, and the learning process was carried out using the parameters: *CRF-L2* algorithm and hyper-parameter $C = 1$.

On the other hand, the classification step is defined as the classification of those elements which have already been recognized. In this case, we have a set of elements to be classified but, this time, the sequence is not as relevant as in the previous case. For example, in timex classification if there is a *date* followed by a *period* and then there is another *date* in the training data, this does not signify that this situation will always occur in a real text, and it is for that reason, that the sequence labeling properties of CRFs are not useful in this task. We have therefore used support vector machines (SVMs) (Cortes & Vapnik, 1995), which is a popular technique that achieved good results in classification tasks, and is one of the techniques most widely used in the related works (see Section 2.2).

SVMs are defined as binary classifiers. Training examples of each class are represented graphically by points in space and an SVM constructs a hyperplane in an n-dimensional space, which separates the points pertaining to one category from the others with the largest margin. The problems tackled in this paper are multi-class classification problems. SVMs may be extended to these problems, thus reducing the multi-class problem to multiple binary classification problems, which predict whether or not an example pertains to a single class. In particular, TIPSem uses the one-vs-one SVM classifier included in YamCha[7] (parameters: $C = 1$ and polynomial_degree = 1).

The parameter configurations described were selected because they obtained the best results in a previous comparative experiment.

4. Multilingual evaluation and discussion

The objective of this evaluation is to measure and analyze the effect of applying semantic knowledge to TimeML timex and event processing. We shall therefore evaluate the proposal described in this paper, a data-driven approach based on semantic features plus morphosyntactic features (TIPSem), and compare it with a baseline based on morphosyntactic features only (TIPSem-B). Since two types of semantic knowledge are used, we shall study the contribution of each kind of semantics by including two intermediate approaches using only one type, TIPSem-LS (lexical semantics) and TIPSem-SR (semantic roles).

In order to provide a multilingual perspective of the problem tackled in this paper, the approaches presented are evaluated for English and Spanish. In our research we are specifically interested in analyzing whether the inclusion of semantic knowledge in the feature set of the learning framework improves our approach in a similar way for different languages.

4.1. Evaluation framework

In this section, we describe the corpora and the evaluation criterion used to measure the performance of the approaches presented.

4.1.1. Corpora for English and Spanish

The approaches presented have been trained and tested using TempEval-2 datasets for English and Spanish. These are our main focus because they enable our results to be compared with those obtained in this recent evaluation exercise.

- **English Corpus**: We used the TempEval-2 data for English, which supposes a revision of the previous TimeML gold standard: TimeBank (Pustejovsky et al., 2003b). The data text originates from a variety of news articles from the Automatic Content Extraction program (ACE) and PropBank texts (TreeBank2). Articles from the former are transcribed broadcast news and those supplied by PropBank are from the Wall Street Journal. Table 3 shows a summary of the statistics of the TempEval-2 data for English.
- **Spanish Corpus**: We used the TempEval-2 data for Spanish, which originates from AnCora corpus (Saurí, Saquete, & Pustejovsky, 2009). It consists of 500 K words mainly taken from newspaper texts. This corpus is annotated and manually reviewed at: morphological level, and syntactic level, and semantic level (Navarro, Civit, Martí, Marcos, & Fernández, 2003). Table 4 shows a summary of the dataset statistics.

4.1.2. Criterion and measures

We applied the criterion used in TempEval-2 (SemEval-2010). The recognition was evaluated using the following metrics:

$$\textbf{precision} = \frac{true\_positives}{true\_positives + false\_positives} \quad \textbf{recall} = \frac{true\_positives}{true\_positives + false\_negatives} \quad \mathbf{F}_{\beta=1} = \frac{2 * precision * recall}{precision + recall}$$

[6] http://crfpp.sourceforge.net/.
[7] http://chasen.org/~taku/software/YamCha/.

Table 3
English data (TempEval-2).

Set	Documents	Words (K)	Element: classes
Training	162	53	TIMEX (1052): date (870), duration (153), time (15) set (12) EVENT (5688): occurrence (3220), reporting (838), i_action (530), i_state (440), state (426), aspectual (196), perception (38)
Test	9	5	TIMEX (81): date (69), duration (8), time (4) set (0) EVENT (498): occurrence (291), reporting (66), i_action (53), i_state (27), state (26), aspectual (31), perception (4)

where *true_positives* are the number of tokens that are part of an extent in the key file and the approach output, *false_positives* are the number of tokens that are part of an extent in the approach output but not in the key, and *false_negatives* are the number of tokens that are part of an extent in the key but not in the approach output.

For the classification, accuracy (*class_score*) is calculated for the correctly recognized timexes and events. The value is computed by dividing the correct classifications, which match the key, by the total amount of correctly recognized instances.

Since the *class_score* depends on the amount of instances which are correctly recognized, two systems recognizing a different set of timexes or events cannot be easily compared. We include a normalized score: *norm_class_score* = *recall*∗*class_score*, which makes this comparison easier.

4.2. Results

The results of timex and event processing are presented separately. Moreover, they have been analyzed by considering the peculiarities of the recognition and the classification task. The best $F_{\beta=1}$ and *norm_class_score* results are highlighted in bold type to simplify their interpretation.

4.2.1. Results for temporal expressions (timexes)

Table 5 shows the results obtained in timex processing for English and Spanish.

4.2.1.1. Analysis of timex recognition results. For English, the application of lexical semantics, semantic roles, and their combination obtained an error reduction in $F_{\beta=1}$ of 26%, 35%, and 59% with regard to the baseline. The combination of both types of semantic knowledge obtained the greatest error reduction (59%), which indicates that their contribution is complementary.

For Spanish, the error reduction was 33%. The contribution of both types of semantic knowledge was also complementary and then the combination obtained the highest results.

The significance of the $F_{\beta=1}$ improvement was tested by applying a paired *t*-test, supporting the hypothesis that the error reduction is statistically significant with a confidence of 95% for both languages.[8]

For English, and with regard to lexical semantics, the presence of time-related semantic classes (*top4hypers* feature) in timexes contributed to their correct recognition. Example (25) shows a timex that was missed by the baseline but recognized by the TIPSem-LS approach.

(25) `TempEval-2 test set – file: VOA19980501.1800.0355 – sentence: 1`
He and Palestinian leader Yasser Arafat meet separately Monday with US Secretary of State.

The baseline missed the timex "Monday" because its 5-window context is not common in the training data. The probability of the model for "Monday" being preceded by the adverb "separately" and followed by the preposition "with" is not therefore sufficiently high. However, by using lexical semantics, "Monday" obtains the value *time_period* => *fundamental_quantity* => *measure* => *abstraction* for the *top4hypers*, which is very common among the timexes in the training, and, therefore, the probability of it being a timex increases.

Although, the application of lexical semantics has been mostly positive for the task in some cases it decreased the performance. For example (26).

(26) `TempEval-2 test set – file: WSJ900813-0157 – sentence: 2`
James Baker, speaking in ABC News "This Week" said "the Kuwaiti request gives [...]".

[8] For English, $t(9) = 2.10$ (one-tail $p = 0.04$), and for Spanish $t(9) = 1.97$ (one-tail $p = 0.04$). p represents the probability of obtaining the improvement by chance.

Table 4
Spanish data (TempEval-2).

Set	Documents	Words (K)	Elements: classes
Training	175	58	TIMEX (1094): date (727), duration (274), time (56) set (37) EVENT (10449): occurrence (3428), state (3281), i_state (1625), i_action (1375), reporting (541), aspectual (153), perception (46)
Test	17	4	TIMEX (92): date (65), duration (24), time (1) set (2) EVENT (825): occurrence (200), state (318), i_state (149), i_action (87), reporting (57),aspectual (13), perception (1)

Table 5
Temporal expression results. Abbreviations: norm (*norm_class_score*).

Language	Approach	Precision	Recall	$F_{\beta=1}$	class_score (norm)
English	TIPSem-B	0.89	0.68	0.77	0.93 (0.63)
	TIPSem-LS	0.91	0.79	0.85	0.94 (0.74)
	TIPSem-SR	0.91	0.76	0.83	0.93 (0.71)
	TIPSem	0.93	0.84	**0.88**	**0.94 (0.79)**
Spanish	TIPSem-B	0.97	0.81	0.88	0.99 (0.80)
	TIPSem-LS	0.96	0.85	0.90	0.96 (0.82)
	TIPSem-SR	0.96	0.88	0.91	0.93 (0.83)
	TIPSem	0.96	0.89	**0.92**	**0.94 (0.84)**

"This Week" was not recognized as a timex by the baseline because "week" followed by a double-quote is not found as a timex in the training. However, if the lexical semantics is concerned, "week" holds the *time_period* => ... semantic class, which increments the probability of it being a timex, as discussed above. This is normally useful but in this example "This Week" is the name of an ABC News section and is not a timex. Hence, lexical semantics can be noisy, and insufficient in the recognition of ambiguous cases.

The contribution of semantic roles is clearly focused on the temporal role. In the sentence analyzed above (26), semantic roles lead to "This Week" being discarded as a timex. Since it plays the A1 role, which is not common in the training timexes, the probability of recognizing it as a timex decreases. The error of TIPSem-LS is therefore solved by the combined approach.

Semantic roles also helped in (27). The number "1360" was marked as a timex by the baseline and the problem was solved by using semantic roles.

(27) `TempEval-2 test set – file: wsj_0505 – sentence: 6`
[The number of stores $_{A1}$] would increase [to 1450 stores $_{A4}$] [from 1360 $_{A3}$].

The non-semantic role approaches failed because it is highly probable that the sequence "from + *number*" represents a timex. However, semantic roles for the main verb "increase" indicate that "the number of stores" is the thing that is increasing (A1 role), "to 1450 stores" is the end point of the increase (A4 role), and "from 1360" is the start point (A3 role), but that there is nothing related to time.

Another advantage of semantic roles is that they bound the temporal arguments for a predicate. This is useful in the detection multi-token temporal expressions. The timex in (28) was incorrectly bounded by the baseline and TIPSem-LS, and only "the week" was recognized. However, it was correctly bounded by the approaches including semantic roles because "the end of this week" was signaled by the AM-TMP role.[9]

(28) `TempEval-2 test set – file: wsj_0073 – sentence: 0`
[...] and the company will begin mailing materials to shareholders at the end of the week.

In some cases, such as that of (29), the application of semantic roles reduced the performance.

(29) `TempEval-2 test set – file: APW19980306.1001 – sentence: 13`
[...] these sites could only be visited by diplomats as laid down by the Feb. 23 accord signed by [...]

[9] Moreover, "at" is within AM-TMP but the model learns that the heading prepositions are excluded from the timexes.

In (29), "Feb. 23 (accord)" does not hold the temporal role and TIPSem thus failed to recognize it as a timex. Timexes as nominal modifiers are difficult to recognize. For example, in the sentence "I live on November 6th street", "November 6th" does not hold any temporal meaning and considering it as a timex would lead to a wrong temporal interpretation. The recognition of a noun-modifier timex depends on the noun modified, for instance, "accords" are likely to be modified by real timexes but "streets" are not.

Assuming that every annotated corpus may contain errors, we also found some false positives caused by human annotation errors. Examples of timexes that were not annotated in the gold corpora are, among others, "recent years", "10-year", "each July", "20th century", "two decades".

For Spanish, the analysis leads to the same conclusions. Semantic knowledge aided in the recognition of real timexes missed by the baseline, as is the case of (30a). Furthermore, semantic roles improved the bounding of long timexes, as in (30b).

(30)

TempEval-2 test set – file: l08_20000601_d.txt – sentence: 3

Es la mayor caída en 7 meses y supone [...] (EN: It is the greatest fall in 7 months and supposes [...])

TempEval-2 test set – file: l08_20000601_a.txt – sentence: 10

El caso, que dura ya más de dos años, [...] (EN: The case, lasting for more than two years, [...])

For both languages, the contribution of semantics can be summarized in two ideas. Both lexical semantics and semantic roles are an advantage for the approaches' generalization capabilities. The application of lexical semantics increases the probability of being able to represent a timex for all the concepts that are semantically related to time (even if they do not appear in the training). By using semantic roles, the approach increases the probability to represent a timex for all the arguments playing a certain role (e.g., temporal role), which is useful in ambiguous cases, such as that of "ABC News *This Week*". The contribution of semantics principally affected the recall. Improving the precision to over 90% maintaining a high recall is very difficult if we consider that even the agreement between human annotators in the dataset is 83%.

4.2.1.2. Analysis of timex classification results. In timex classification, the contribution of semantics has been much more limited. For English, the score ranges between 0.93 (0.63) for the baseline to 0.94 (0.79) for TIPSem. For Spanish, it ranges from 0.99 (0.80) to 0.94 (0.84). This could be interpreted as an improvement, because TIPSem classified more timexes correctly. However, this could also mean that only the new expressions recognized by using semantics are also classified correctly.

Since we had no evidence as to whether they would also be correctly classified by the baseline, in order to compare the approaches on classification when isolated from the recognition, the following experiment was carried out: the baseline and TIPSem classifiers were evaluated over the same set of timexes (i.e., those correctly recognized by TIPSem).

For English, the results obtained were the same for the baseline and the TIPSem classifiers. This signifies that the application of semantics in timex classification had no effect.

For Spanish, the same experiment was carried out and it was verified that semantic information was not discriminative. For instance, with regard to lexical semantics, dates (e.g., this year) and durations (e.g., 15 years) may contain lemmas (e.g., year) that share the top four hypernyms, and with regard to semantic roles, both timex classes may appear under a temporal argument (e.g., [this year $_{AM-TMP}$], [for 15 years $_{AM-TMP}$]).

Most errors appeared as a result of the wrong classification of poorly represented classes. *Time*s represent 1% of timexes, and the *set*s proportion is even lower. This prevented our approach from learning a good classifier. This problem can be dealt with by extending the available corpora with documents containing more instances of *time* and *set* timexes. As regards semantics, roles are not discriminative but lexical semantics can be useful. Timex classification requires a lexical resource that is capable of distinguishing, for example, dates (e.g., Monday) from times (e.g., morning). We used a general resource (WordNet), in which *Monday* and *morning* are hyponyms of *period* (i.e., equivalent). This can be improved using a resource distinguishing different granularities, i.e., finding one more detailed than WordNet.

4.2.2. Results for events

Table 6 shows the results obtained for event processing for English and Spanish.

4.2.2.1. Analysis of event recognition results. For English, the application of lexical semantics, semantic roles, and their combination obtained an error reduction in $F_{\beta=1}$ of 11%, 16%, and 21% with regard to the baseline. As in the case of timex, the application of lexical semantics and semantic roles is complementary. For Spanish, the highest error reduction was also obtained by the combination (17%). We used a paired t-test to verify that the improvements are significant with a confidence above 95%.[10]

For English, lexical semantics improved the baseline in cases such as that of (31).

[10] For English, t (9) = 3.16 (one-tail p = 0.006), and for Spanish t (9) = 3.16 (one-tail p = 0.006).

(31) TempEval-2 test set – file: APW19980306.1001 – sentence: 0
An American leader of a UN weapons inspection team resumed work in Iraq Friday.

(32) This is your work.
There are opportunities for work.

The noun "work" is an event in (31) but not in other situations such as (32). Since there are more instances of this noun that are not events in the training, the baseline has a high probability of discarding them as events. With lexical semantics, the approach correctly recognized "work" as an event in (31), owing to the fact that its value for *top4hypers* is *act => event => psychological_feature => abstraction*. This value is very popular among the nominal events of the training corpus, and the probability of recognizing tokens holding this value thus increases. This favors the generalization properties of the approach for many nominal events (e.g., inspection, occupation, negotiations, invasion).

The application of semantic roles increased the number of correctly recognized events. In (33a) and (33b), the events "shakeup" and "inspection" were missed by the baseline and TIPSem-LS but were correctly recognized by using semantic roles.

(33) a. TempEval-2 test set – file: wsj_0586 – sentence: 38
British shakeup was widely cited by the declines.
b. TempEval-2 test set – file: APW19980306.1001 – sentence: 6
They had been allowed to carry out an inspection.

The baseline fails because these lemmas do not appear in the training. Although lexical semantics increases the probability of their being events, this is not sufficiently high for the model to be able to recognize them. However, when semantic roles are concerned, another sign adds probability to the model. Since they hold the A1 role, and many nominal events appearing in the training play this role, the combined approach recognizes the events in (33). Particularly, in (33b), "inspection" holds the A1 role for the verb "to carry out", which normally defines events (e.g., carry out + action/process).

With Spanish, since the properties of the evaluation data signified that the baseline had already obtained an $F_{\beta=1}$ of 0.88, there was less room for improvement. However, the performance was still improved by using semantics, for instance in (34).

(34) TempEval-2 test set – file: l08_20000301_b.txt – sentence: 8
Valero Ribera no disimuló su preocupación por el sorteo y [...]
EN: Valero Ribera did not hide his concern for the draw and [...]

In this sentence, the noun "*sorteo*" (EN: draw) is an event that was missed by the baseline. However, the combination of lexical semantics (*top4hypers* = event) and semantic roles (*role* = A1) assisted in the correct recognition.

The contribution of semantics to event recognition is focused on recall. The precision obtained is comparable to that obtained by human annotators (agreement of 81%), and its improvement is thus difficult if a high recall is to be maintained.

4.2.2.2. Analysis of event classification results. As occurred with the timex classification, the contribution of semantics to event classification has been more limited in both languages. For English, the score ranges between 0.79 (0.63) for the baseline to 0.80 (0.70) for TIPSem. For Spanish, it ranges from 0.66 (0.57) to 0.68 (0.60).

Table 6
Event processing results. Abbreviations: norm (*norm_class_score*).

Language	Approach	Precision	Recall	$F_{\beta=1}$	class_score (norm)
English	TIPSem-B	0.82	0.80	0.81	0.79 (0.63)
	TIPSem-LS	0.81	0.86	0.83	0.80 (0.69)
	TIPSem-SR	0.82	0.86	0.84	0.80 (0.69)
	TIPSem	0.82	0.87	**0.85**	**0.80 (0.70)**
Spanish	TIPSem-B	0.90	0.86	0.88	0.66 (0.57)
	TIPSem-LS	0.93	0.87	0.90	0.66 (0.57)
	TIPSem-SR	0.92	0.87	0.89	0.66 (0.57)
	TIPSem	0.92	0.88	**0.90**	**0.68 (0.60)**

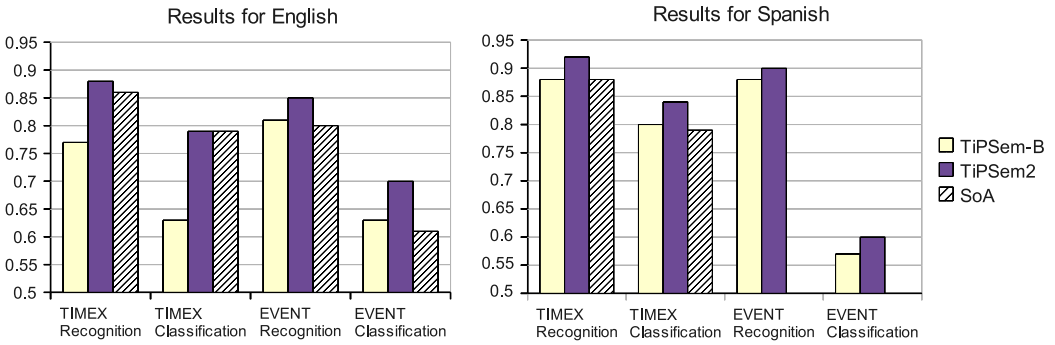

Fig. 4. Results and comparison with state-of-the-art.

In order to verify whether this increase in the score was caused by a real improvement to the classifiers as a result of the application of semantics, we carried out the classification of the total events recognized by the best recognition approach TIPSem as we did for timex. The results obtained for English and Spanish were the same for the baseline and TIPSem classifiers, and the semantic features used are not therefore discriminative. Events such as "maintain", which may represent different classes (e.g., OCCURRENCE, REPORTING), obtain the same value for the *top4hypers* feature because our approach uses the first sense (most frequent). Moreover, verbal events like "to complete" may represent an OCCURRENCE (e.g., [He $_{A0}$] completed [the puzzle $_{A1}$]) or an ASPECTUAL event (e.g., [He $_{A0}$] completed [the transaction $_{A1}$]). However, both instances of "complete" obtain the same values for the role-based features.

Although the current semantic features are not discriminative, we noted that the semantic features could aid in classification with the following updates. With regard to lexical semantics, the application of word sense disambiguation (WSD) techniques prior to obtaining the *top4hypers* feature (e.g., maintain-sense1 and maintain-sense2) may improve the classification. With regard to semantic roles, a feature considering whether the A1 role of a verbal event (e.g., complete) contains another event (e.g., transaction) may aid the classification.

If we observe specific classes, some were classified as having a very high performance (i.e., perception 1.0, reporting 0.90, occurrence 0.90), and others as having a lower performance (i.e., aspectual 0.66, i_action 0.56, i_state 0.43 and state 0.33). This performance distribution do not match the training example distribution shown in Table 3, and the results obtained instead indicate the complexity of each different class.

For Spanish, perception (1.0), reporting (0.86) and occurrence (0.77) again obtained the highest results, while i_state (0.55), aspectual (0.54) and i_action (0.46) obtained a lower performance. However, state events obtained a score of 0.72.

Although the datasets for English and Spanish are different in content, and the results are not directly comparable, they both are built from news articles, and a score of 0.72 in the classification of state-events is much greater than the score obtained in the English experiment (0.33). Moreover, if we observe the training example distribution, the proportion of state events in Spanish is much greater than in English. We analyzed both datasets and came to the conclusion that this difference was caused by an annotation difference in the Spanish corpus: all the instances of the verb "to be" are annotated as state-events, which does not occur in the English dataset.

4.3. Comparing results with the baseline, the state of the art, and between English and Spanish

This subsection presents the results previously analyzed in a comparative chart. Fig. 4 illustrates the results for English and Spanish, in which TIPSem-B is shown in the lighter color and TIPSem is shown in the darker color, whilst the best state-of-the-art (SoA) appears in hatched white. The scores shown for the classification correspond to its normalized value (*norm_class_score = recall∗ class_score*). Note that since there are not SoA results for Spanish event recognition and classification, these columns are empty.

TIPSem outperforms the baseline (TIPSem-B) in all tasks for both languages. This indicates that semantic knowledge is useful for timex and event recognition and classification. However, as discussed in previous subsections, there is only a significant improvement in recognition tasks. For timexes, this improvement occurs as a result of the direct relation between these elements with temporal roles and with temporal semantic classes in WordNets. For events, the semantic knowledge aided learning in more general models based on the event-related semantic classes, the detection of main verbs, and the identification of semantic roles that often contain events.

The state-of-the-art is outperformed by TIPSem. This indicates that the approach presented herein addresses the proposed problem in a highly competent manner. Our approach was considerably better than SoA in event recognition, thus indicating that semantic knowledge is needed to achieve a high-performance in this task. TIPSem results are also an improvement on those obtained in TempEval-2 with its earlier version, indicating that the individualized treatment of timexes and events with specific feature sets is useful. Recently, Kolomiyets et al. (2011) applied a Latent Words Language Model (LWLM) to timex recognition. Their approach (0.84 F1) does not outperform TIPSem, but it showed that LWLM is

beneficial. It would be interesting to add this new semantic knowledge to TIPSem in order to verify whether it improves the results.

The results for English and Spanish cannot be compared because, although both consist of news articles, the datasets are not parallel.[11] The fact that the results obtained are slightly higher for Spanish cannot therefore be assessed. However, two factors might have influenced this: (i) the English approach uses automatic tools to obtain the features while the Spanish one extracts them from a manually reviewed corpus and (ii) for Spanish, the "to be" verb is always annotated as an event and the training data size is slightly greater than that used for English.

5. Conclusions and further work

This paper analyzes the application of semantic knowledge to the automatic recognition and classification of temporal expressions and events in English and Spanish. The performance level of these tasks is crucial for high performance in the broader task of temporal information processing. Our aim is to ascertain whether the use of semantics leads to an improvement on the results obtained by state-of-the-art approaches which are principally based on morphosyntax.

The specific goals of this paper are: (i) to analyze the advantages that lexical semantics (semantic networks) and semantic roles (PropBank roles) may offer in addition to morphosyntax and (ii) to carry out a multilingual study of the advantages for English and Spanish, evaluating the degree of language independence of these advantages.

We have therefore presented a data-driven approach (TIPSem), which uses semantic knowledge in addition to morphosyntax. We have evaluated TIPSem for English and Spanish over the datasets employed in the last international evaluation exercise on temporal information processing (TempEval-2). In order to measure the influence of semantics, the results have been compared with a baseline (TIPSem-B) which only uses morphosyntactic information, and with the state-of-the-art.

The results demonstrate that semantic knowledge is particularly useful for recognition tasks, showing an F1 error reduction of 59% for timexes and of 21% for events. This improvement has been proven to be statically significant with a confidence of 95%. The contribution of lexical semantics and semantic roles has been found to be complementary, so that their combination is useful. More specifically, semantics lead to the learning of more general models that improve the recognition of timexes and events which only share semantic properties. Semantics also helped in the recognition of timexes and events that were ambiguous at lower language analysis levels. With regard to classification, the application of semantics did not lead to a remarkable improvement, since the semantic features applied are not sufficiently discriminative. Although our approach did not show conclusive improvements, after analyzing the causes, we have proposed more advanced semantic features that might be discriminative, as is discussed in Section 4.

As compared with the state-of-the-art, TIPSem outperforms other approaches, particularly in event recognition. This supports the hypothesis that semantic knowledge is needed to achieve high performance. These results are also an improvement on those obtained in TempEval-2 by the earlier version of TIPSem. This indicates that the individualized treatment of timexes and events with specific feature sets is useful.

Although we applied a combination of semantic features to timex and event extraction, these features may be useful for other information processing tasks, particularly those that face morphosyntactic ambiguity, or those in which the training dataset is too small to learn sufficiently general models with the sole use of morphosyntax. For instance, semantic features have been applied to Information Retrieval and Question Answering (Moreda, Navarro, & Palomar, 2007; Moreda et al., 2011).

As further work, we shall add to TIPSem other kinds of semantic knowledge (i.e., LWLM), which have been shown to be beneficial in recent works (Kolomiyets et al., 2011). Moreover, we shall improve classification performance using semantics. Firstly, we will confront the problem of the lack of training examples that causes the low performance in timex classification, by using a larger annotated dataset. A 1 million corpus will be released for the next international evaluation exercise (TempEval-3[12]) and this will be used to retrain our models. Secondly, as suggested in Section 4.2.2, event classification will be improved by using word sense disambiguation techniques, in addition to intra-sentential syntactic and semantic dependencies between events.

Acknowledgements

This paper has been supported by the Spanish Government (Project TIN-2009-13391-C04-01), Generalitat Valenciana (Project PROMETEO/2009/119 and ACOMP/2010/286), and the University of Alicante (Project GRE09-31). Furthermore, we wish to thank the anonymous reviewers for their comments and suggestions which have allowed us to improve the quality of this paper.

References

Ahn, D., van Rantwijk, J., & de Rijke, M. (2007). A cascaded machine learning approach to interpreting temporal expressions. In *NAACL. ACL* (pp. 420–427).
Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of ACM, 26*(11), 832–843.

[11] One is not the translation of the other.
[12] http://www.cs.york.ac.uk/semeval-2013/task1/.

Alonso, O., Gertz, M., & Baeza-Yates, R. (2007). On the value of temporal information in IR. *SIGIR, 41*(2), 35–41.
Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *COLING-ACL* (pp. 86–90).
Bethard, S., & Martin, J. H. (2006). Identification of event mentions and their semantic class. In *EMNLP: Proceedings of the Conference on Empirical Methods in NLP. ACL* (pp. 146–154).
Boguraev, B., & Ando, R. K. (2005). Effective use of TimeBank for TimeML analysis. In *Annotating, extracting and reasoning about time and events 05151*.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
Daniel, N., Radev, D., & Allison, T. (2003). Sub-event based multi-document summarization. In *HLT-NAACL text summarization workshop. ACL* (pp. 9–16).
Derczynski, L., & Gaizauskas, R. (2010). Usfd2: Annotating temporal expresions and tlinks for tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 337–340).
Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*(3).
Fellbaum, C. (1998). *WordNet: An electronic lexical database (language, speech, and communication)*. MIT Press.
Ferro, L., Mani, I., Sundheim, B., & Wilson, G. (2000). *TIDES temporal annotation guidelines*, Draft v.1.0. MITRE technical report MTR00W0000094, MITRE.
Fillmore, C. (1971). Types of lexical information. In D. Steinberg & L. Jacobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*. Cambridge: Cambridge University Press.
Fillmore, C. J. (1968). Universals in linguistic theory. In *The case for case* (pp. 1–88). NY: Holt, Rinehart & Winston.
Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245–288.
Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING* (pp. 466–471).
Grover, C., Tobin, R., Alex, B., & Byrne, K. (2010). Edinburgh-ltg: Tempeval-2 system description. In: *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 333–336).
Kolomiyets, O., Bethard, S., & Moens, M. -F. (2011). Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th annual meeting of the ACL-HLT. ACL*, pp. 271–276.
Kolomiyets, O., & Moens, M. -F. (2010). Kul: Recognition and normalization of temporal expressions. In *Proceedings of the 5th international workshop on semantic evaluation. ACL*, pp. 325–328.
Kolya, A. K., Ekbal, A., & Bandyopadhyay, S. (2010). Ju_cse_temp: A first step towards evaluating events, time expressions and temporal relations. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 345–350).
Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th ICML* (pp. 282–289). Morgan Kaufmann.
Llorens, H., Navarro, B., & Saquete, E. (2009a). Using semantic networks to identify temporal expressions from semantic roles. In *VI RANLP* (pp. 219–224).
Llorens, H., Saquete, E., & Navarro, B. (2009b). Temporal expression identification based on semantic roles. In *14th International conference on applications of natural language to information systems (NLDB). LNCS*, pp. 230–242.
Llorens, H., Saquete, E., & Navarro-Colorado, B. (2010a). TimeML events recognition and classification: Learning CRF models with semantic roles. In *Proceedings of the 23rd COLING* (pp. 725–733).
Llorens, H., Saquete, E., & Navarro-Colorado, B. (2010b). TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 284–291).
March, O., & Baldwin, T. (2008). Automatic event reference identification. In *ALTA 2008. Australia* (pp. 79–87).
Melli, G., Shi, Z., Wang, Y., & Popowich, Y. L. (2006). Description of SQUASH, the SFU question answering summary handler for the DUC-2006 summarization task. In *DUC*.
Moreda, P., Llorens, H., Saquete, E., & Palomar, M. (2011). Combining semantic information in question answering systems. *Information Processing & Management, 47*(6), 870–885.
Moreda, P., Navarro, B., & Palomar, M. (2007). Corpus-based semantic role approach in information retrieval. *Data Knowledge Engineering, 61*(3), 467–483.
Navarro, B., Civit, M., Martí, M. A., Marcos, R., & Fernández, B. (2003). Syntactic, semantic and pragmatic annotation in Cast3LB. In *Corpus linguistics (SProLaC)*.
Negri, M., & Marseglia, L. (2004). *Recognition and normalization of time expressions: ITC-irst at TERN 2004*. Technical report, Information Society Technologies.
Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31.
Punyakanok, V., Roth, D., Yih, W., Zimak, D., & Tu, Y. (2004). Semantic role labeling via generalized inference over classifiers. In *HLT-NAACL (CoNLL). ACL* (pp. 130–133).
Pustejovsky, J., Castaño, J. M., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., et al. (2003a). TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5*.
Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., et al. (2003b). The TIMEBANK corpus. In *Corpus linguistics* (pp. 647–656).
Saquete, E., González, J. L. V., Martínez-Barco, P., Muñoz, R., & Llorens, H. (2009). Enhancing QA systems with complex temporal question processing capabilities. *Journal of Artificial Intelligence Reresarch (JAIR), 35*, 775–811.
Saquete, E., Martínez-Barco, P., & Muñoz, R. (2004). Automatic multilinguality for time expression resolution. In *MICAI. Vol. 2972 of LNCS* (pp. 458–467).
Saquete Boro, E. (2010). Id 392:terseo + t2t3 transducer. A systems for recognizing and normalizing timex3. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 317–320).
Saurí, R., Knippen, R., Verhagen, M., & Pustejovsky, J. (2005). Evita: A robust event recognizer for QA systems. In *HLT/EMNLP. ACL*.
Saurí, R., Saquete, E., & Pustejovsky, J. (2009). *Annotating time expressions in spanish. timeml annotation guidelines*. Technical report, Barcelona Media – Innovation Center.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the international conference on new methods in language processing* (pp. 44–49).
Setzer, A., & Gaizauskas, R. (2000). Annotating events and temporal information in newswire texts. In *LREC*.
Strötgen, J., & Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 321–324).
Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *ELRA (Ed.), LREC. Marrakech, Morocco*.
UzZaman, N., & Allen, J. F. (2010). Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 276–283).
Verhagen, M., Mani, I., Saurí, R., Knippen, R., Jang, S.B., Littman, J., et al. (2005). Automating temporal annotation with TARSQI. In *ACL. ACL, NJ, USA* (pp. 81–84).
Verhagen, M., Saurí, R., Caselli, T., & Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation. ACL* (pp. 57–62).
Vicente-Díez, M. T., Moreno-Schneider, J., & Martínez, P. (2010). Uc3m system: Determining the extent, type and value of time expressions in tempeval-2. In *Proceedings of the 5th SemEval. ACL* (pp. 329–332).
Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. MA, USA: Kluwer Academic.