Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 115-118 recibido 21-12-11 revisado 06-03-12 aceptado 18-03-12

A Semantic Approach to Temporal Information Processing [*]

Aproximación Semántica al Procesamiento de la Información Temporal

Hector Llorens
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
hllorens@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada por Héctor Llorens en la Universidad de Alicante (UA) bajo la dirección de la Dra. Estela Saquete (UA) y del Dr. Borja Navarro-Colorado (UA). El acto de defensa de la tesis tuvo lugar en Alicante el 11 de Julio de 2011 ante el tribunal formado por los doctores Patricio Martínez-Barco (UA), Paloma Moreda (UA), Roser Saurí (Barcelona Media), Horacio Rodríguez (UPC) y Tommaso Caselli (ILC-CNR, Italia). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad (con mencion europea).

Palabras clave: Extracción de la Información Temporal, TimeML, Semántica

Abstract: Ph.D. Thesis in Computer Science, in the field of Computational Linguistics, written by Hector Llorens at the University of Alicante (UA), under the supervision of Dr. Estela Saquete (UA) and Dr. Borja Navarro-Colorado (UA). The author was examined on July 11th 2011, by a panel formed by the doctors Patricio Martínez-Barco (UA), Paloma Moreda (UA), Roser Saurí (Barcelona Media), Horacio Rodríguez (UPC) and Tommaso Caselli (ILC-CNR, Italy). The grade obtained was *Excellent Cum Laude* unanimously (with European mention).

Keywords: Temporal Information Extraction, TimeML, Semantics

1 Introduction

The huge amount of digitalized information and its wide diffusion through the Internet has named the present era as the “Information Society”. The amount of information has surpassed human tractable limits. Currently, research focuses on improving information access and management to take advantage from this valuable knowledge source.

Within natural language processing (NLP), this thesis tackles the processing of the temporal information. The aim of this task is obtaining the temporal location and ordering of the events expressed in text, which requires the automatic extraction and interpretation of temporal expressions (timexes) such as “May 1995” or “yesterday”; events such as “born” or “war”; and their temporal relations. Access to processed temporal information is crucial for many NLP tasks including question answering (Saquete et al., 2009), text summarization (Daniel, Radev, and Allison, 2003), and information retrieval (Alonso, Gertz, and Baeza-Yates, 2007).

The majority of current approaches are based on morphosyntactic knowledge. However, temporal entities are often ambiguous at that language analysis level. Example (1) underlines some ambiguous expressions that often represent timexes or events but they do not in the shown cases.

(1) April is watching a news program, Last Week. The number of control towers increased from 1995 to 2000 towers.

Our hypothesis is that the linguistic expression of time is a semantic phenomenon and therefore, to achieve a better extraction performance, temporal information must be processed using semantics.

To prove this hypothesis, we have developed and evaluated a semantic approach to temporal information processing: *TIPSem*. This is an automated system that includes features based on lexical semantics, and semantic roles, in addition to features based on morphosyntax.

Furthermore, to evaluate the proposal extrinsically, TIPSem has been applied to develop an innovative graphical interface which

[*] Este trabajo ha sido co-financiado por los proyectos TEXT-MESS 2.0 (TIN2009-13391-C04-01) y MESOLAP (TIN2010-14860).

ISSN 1135-5948 © 2012 Sociedad Española para el Procesamiento de Lenguaje Natural

brings users time-based access to information: *Time-Surfer*.

Next we describe the related work, our investigations, and their evaluation. Finally, we summarize the conclusions obtained.

2 Related Work

The research on temporal information processing evolved from different rationalist formal strategies to the empiricist corpus-based strategy, which consists of the annotation of corpora following a temporal annotation scheme (Grishman and Sundheim, 1996; Wilson et al., 2001; Setzer and Gaizauskas, 2000). Currently, the standard temporal annotation scheme is TimeML (Pustejovsky et al., 2003) due to its completeness and the improvements this introduces to its predecessors.

Although temporal information processing requires linguistic knowledge at semantic, discourse and pragmatic level, the majority of state-of-the-art computational approaches only use morphosyntactic information. Few approaches use lexical semantics and only one of them includes FrameNet semantic roles (Verhagen et al., 2007; Verhagen et al., 2010).

The influence of semantics in the performance of the approaches has not been analyzed in depth. Our proposal adds to the state-of-the-art the following aspects: it applies lexical semantics and semantic roles and analyzes specifically its contribution to this task multilingually for English, Spanish, Italian, and Chinese; and it uses the PropBank role set, which provides valuable information about the temporal entities, and has never been used for this purpose.

3 Contributions of this thesis

We analyzed the effect of semantics in temporal information processing through the development and evaluation of an information extraction system, TIPSem; and a graphical interface for exploring temporal information, Time-Surfer. Both are available on-line[1].

3.1 TIPSem System

Since its central element is semantics, we named our proposal *TIPSem*: *T*emporal *I*nformation *P*rocessing using *Sem*antics. TIPSem is a hybrid approach (mainly data-driven but also rule-based), which exploits semantic roles and lexical semantics (semantic networks), in addition to morphosyntax. The system considers temporal information as defined by the *TimeML* annotation scheme. The architecture of TIPSem is divided into:

- *Timex processing.* This task implies the recognition in text, and the classification and normalization of such entities.
- *Event processing.* This includes the recognition in text, and the classification of such entities.
- *Temporal relation processing.* This implies the categorization of different types of relations between the previous entities. Namely, relations between events and timexes, between events and the creation time of the document, between events syntactically subordinated (intra-sentential), and between main events across sentences (inter-sentential).

Figure 1 illustrates the architecture of TIPSem system showing its different parts.

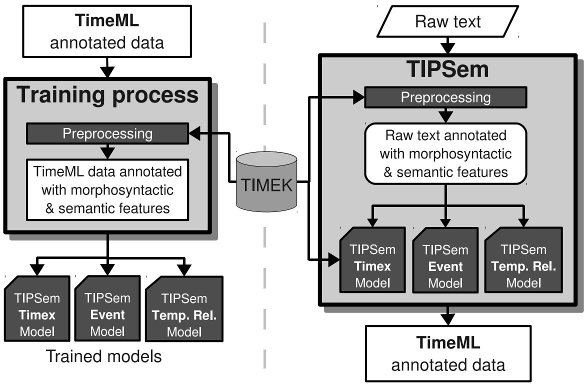


Figure 1: TIPSem architecture

For the English version, the features used in the data-driven part can be summarized as:

- *Morphosyntactic features.* These are tokens, lemmas and part-of-speech (PoS), which are obtained using TreeTagger; and the syntactic tree (Treebank-like) obtained using Charniak parser.
- *Semantic features.* These are divided into two groups:
  - *Lexical semantics*: semantics at word level, obtained from WordNet.

[1] http://gplsi.dlsi.ua.es/demos/TIMEE/

The main feature of this set is the four top hypernyms of each token. This is useful for generalize tokens by its semantic class.

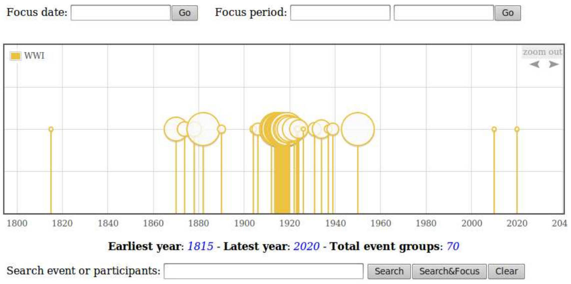
- *Semantic roles*: semantics at predicate level following PropBank semantic role labeling, obtained using CCG tool. The main role features consist of the role that each token plays, and the role regarding which verb. These are useful to differentiate ambiguous tokens.

In order to learn models for each subtask from these features, we applied two widely known supervised machine learning techniques: conditional random fields (CRFs) and support-vector machines (SVMs). CRFs are useful for classifying elements whose classification is dependent to adjacent elements class and features: the sequence is relevant. We applied them in timex and event recognition, and relation categorization. SVMs are useful when the classification of an element is independent from the adjacent elements. We used them for timex and event classification, and timex normalization-type classification.

3.2 Time-Surfer Interface

Time-Surfer is a graphical interface that interactively shows the temporal information extracted by TIPSem. It brings users a dynamic picture of the temporal distribution of events within a document through a timeline. Event groups are represented by bubbles whose position depends on the events' time, and whose radius depends on the number of events they contains. If a bubble is hovered the information about the events contained and their participants is shown.

Figure 2 shows Time-Surfer view for the First World War Wikipedia article.



Earliest year: *1815* - **Latest year**: *2020* - **Total event groups**: *70*

Figure 2: Time-Surfer Interface

In addition to navigation facilities such as time-zooming, Time-Surfer includes forms to search dates, periods, events, or event participants. If a query is introduced, bubbles containing relevant events are colored.

4 Evaluation Results

To demonstrate our hypothesis, the performance of TIPSem has been measured, and the effect of applying the proposed semantic features to temporal information processing has been analyzed in different languages: English, Spanish, Italian and Chinese.

The experimentation has been divided into intrinsic and extrinsic evaluation.

4.1 Intrinsic Evaluation

For this purpose, TIPSem has been evaluated intrinsically through the participation in the SemEval TempEval-2 evaluation exercise (Verhagen et al., 2010). The results have been compared with those obtained by a non-semantic baseline, TIPSem-B, which only uses morphosyntactic features. Table 1 summarizes the results obtained for English.

Element-Task		**Score**	**IMPR. (SSC)**
TIMEX	recognition	0.88 F1	26% (.99)
	classification	0.98 AC	19% (.97)
	normalization	0.84 AC	19% (.99)
EVENT	recognition	0.85 F1	11% (.99)
	classification	0.79 AC	0% (-)
TLINK	event-timex	0.68 AC	5% (.90)
	event-DCT	0.82 AC	3% (.90)
	main-events	0.61 AC	3% (.99)
	sub-events	0.66 AC	2% (.95)

Table 1: TIPSem result summary for English. Abbrev: AC is accuracy, IMPR. is relative error reduction introduced by semantics in a 10-fold experiment, and SSC is the statistical significance confidence of the improvement.

The results obtained in the different elements and tasks show that the proposed semantic features are appropriate to tackle the problem. The significant performance improvement obtained in general over a morphosyntactic baseline firmly support the presented hypothesis: semantics are useful for temporal information processing. Specifically, semantic features aid in handling morphosyntactic ambiguity and favour generalization capabilities.

As compared with the state-of-the-art, TIPSem obtains a competitive performance and introduces a remarkable improvement in event recognition (85% vs. 80% F1).

We also carried out evaluations for other languages which confirmed that our semantic approach is also useful for Spanish, Italian, and Chinese.

Regarding the errors that TIPSem leaves unsolved, these are mainly the cases that require a language analysis level higher than semantics.

4.2 Extrinsic Evaluation

To extrinsically evaluate TIPSem, we tested Time-Surfer. The output quality of Time-Surfer depends on the performance of TIPSem to extract the temporal information.

The positive results obtained from evaluating if users take advantage of using Time-Surfer for answering temporal questions quicker than using Wikipedia, support that TIPSem's performance is satisfactory to extract temporal information from an extrinsic standpoint –for human users.

5 Conclusions and Further Work

This paper summarizes a Ph.D. thesis addressing the task of temporal information processing, within the area of NLP.

Starting from the hypothesis that semantics is beneficial for this task, and current state-of-the-art has not analyzed its effects in depth, this thesis presents a semantic approach to temporal information processing, and an application which uses the extracted temporal information to create a dynamic graphical timeline.

To verify the hypothesis, the approach is evaluated and compared with a morphosyntactic baseline and with the state-of-the-art for different languages. Furthermore, we tested the presented application to evaluate it extrinsically.

The results support that adding the proposed semantic features: (i) improves the performance of morphosyntactic approaches, specifically, in handling morphosyntactic ambiguity and favouring generalization capabilities, (ii) leads to a high performance as compared to the state-of-the-art, (iii) benefits different languages: English, Spanish, Chinese, and Italian, and (iv) is also beneficial from an extrinsic viewpoint.

As further work, we propose addressing the efficacy issues of the approach using higher language analysis level information, improving the approach efficiency, and building a bigger annotated dataset to extend the evaluation of the approach.

References

Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. 2007. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum*, 41(2):35–41.

Daniel, Naomi, Dragomir Radev, and Timothy Allison. 2003. Sub-event based multi-document summarization. In *HLT-NAACL Text summarization workshop*, pages 9–16. ACL.

Grishman, Ralph and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *COLING*, pages 466–471.

Pustejovsky, James, 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*.

Saquete, Estela, José Luis Vicedo González, Patricio Martínez-Barco, Rafael Muñoz, and Hector Llorens. 2009. Enhancing QA Systems with Complex Temporal Question Processing Capabilities. *Journal of Artificial Intelligence Reresarch (JAIR)*, 35:775–811.

Setzer, Andrea and Robert Gaizauskas. 2000. Annotating Events and Temporal Information in Newswire Texts. In *LREC 2000*, pages 1287–1294.

Verhagen, Marc, 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.

Verhagen, Marc, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. ACL.

Wilson, George, Inderjeet Mani, Beth Sundheim, and Lisa Ferro. 2001. A multilingual approach to annotating and extracting temporal information. In *Proceedings of the workshop on Temporal and Spatial information processing*, pages 81–87, NJ, USA. ACL.