Data-Driven Approach Using Semantics for Recognizing and Classifying TimeML Events in Italian

Tommaso Caselli
ILC-CNR, Pisa
tommaso.caselli@ilc.cnr.it

Hector Llorens
DLSI - Universidad de Alicante
hllorens@dlsi.ua.es

Borja Navarro-Colorado
DLSI - Universidad de Alicante
borja@dlsi.ua.es

Estela Saquete
DLSI - Universidad de Alicante
stela@dlsi.ua.es

Abstract

We present a data-driven approach for recognizing and classifying TimeML events in Italian. A high-performance state-of-the-art approach, TIPSem, is adopted and extended with Italian-specific semantic features from a lexical resource. The resulting approach has been evaluated over the official TempEval2 Italian test data. The analysis of the results shows a positive impact of the semantic features both for event recognition and classification. Moreover, the presented data-driven approach has been compared with an existing rule-based prototype over the same data set. The results are directly comparable and show that the machine learning strategy better deals with the complexity of the tasks.

1 Introduction

Recognizing and classifying events is a strategic task in order to improve the performance of many NLP applications such as automatic summarization and question answering (Q.A.). In NLP, different definitions of event can be found regarding the target application. Recently, TimeML (Pustejovsky et al., 2003a) introduced a rich specification language for annotating and classifying events and it has been applied to English documents (the TimeBank corpus (Pustejovsky et al., 2003b)). The SemEval TempEval-1 and TempEval-2 international evaluation excercises (Verhagen et al., 2007; Verhagen et al., 2010), have provided the NLP community with gold standard resources for comparative evaluations of different systems. In addition to this, TempEval-2 made available TimeML annotated data in languages other than English, namely Italian, French, Spanish, Chinese and Korean. Unfortunately, there were only participants for English and Spanish.

This paper focuses on the recognition and classification of TimeML events in Italian by means of a state of the art data driven approach, TIPSem (Llorens et al., 2010), which obtained competitive results in the TempEval tasks for English and Spanish. To the best of our knowledge this is the first data-driven approach which is developed for this language and differs from state-of-the-art approches developed for English for the use of specific lexical semantic features. In particular, the Italian-specific semantic features have been obtained from a semi-automatically built *event lexicon* which has been derived from the SIMPLE/CLIPS lexicon (Ruimy et al., 2003), following the proposal in Caselli (2009). The objectives of this paper are (i) evaluating TIPSemIT over the official TempEval-2 data for Italian and assessing the impact of the semantic resource, and (ii) comparing the performance of data-driven to rule-based approaches in Italian over the same data. Section 2 reports a short background on the TimeML specifications. Section 3 describes the adaptation of TIPSem approach to Italian. Section 4 is devoted to the different evaluation experiments introduced above and, finally, section 5 focuses on conclusions and future work.

2 TimeML specifications for events

In TimeML an event is defined as something that happens or holds true. Natural language (NL) offers a variety of means to realize events, such as verbs (*andare* [to go]), complex VPs (light verb constructions, *fare una doccia* [to have a shower], or idioms), nouns (nominalizations - *volo* [flight],

Proceedings of Recent Advances in Natural Language Processing, pages 533–538,
Hissar, Bulgaria, 12-14 September 2011.

costruzione [building] - second order nominals - *assemblea* [meeting] - and type-coercions), predicative constructions (*essere ricco* [to be rich]), prepositional phrases (*a bordo* [on board]) or adjectival phrases (*dormiente* [dormant]). Two innovative aspects introduced by TimeML with respect to event recognition and classification concern (i) the extent of the text span to be annotated and (ii) the classes. As for the text span of the <EVENT> tag, TimeML implements the notion of minimal chunk, i.e. only the head of the constituent(s) realizing an event must be annotated and not the entire phrase(s). This distinction is of utmost importance, since phrases can include more than one event instance. To clarify, consider example 1, where the extent of the event phrase is in bold and the event elements are marked with the <EVENT> tag.

(1) *Marco* ***deve andare a casa***. [Marco has to go home]

```
Marco <EVENT id="001"
...>deve</EVENT> <EVENT id="002"
...>andare</EVENT> a casa .
```

Events's classes are established by means of criteria that characterize their nature as irrealis, factual, possible, reported, intensional and so forth, thus departing from theoretical linguistic approaches (Vendler, 1967). In this way, seven classes have been identified, namely:

- REPORTING: the action of a person, an organization declaring something or informing about an event (e.g. say, tell...);
- PERCEPTION: events which involve the physical perception of another event (e.g. see, hear...);
- I_ACTION: events which give rise to an intensional relation with their event argument (e.g. try...);
- I_STATE: events which give rise to an intensional state with their event argument (e.g. love, want...);
- STATE: temporally bound circumstances in which something obtains (e.g. peace, be in love...);
- OCCURRENCE: events which describe thighs that happen in the world (e.g. happen, come...);
- ASPECTUAL : events which describe an aspectual predication of another event (e.g. start, finish...).

Notice that the same event item may belong to different classes according to the linguistic context in which it occurs. To clarify, consider the following examples where the event *pensare* [to think] is classified both as OCCURRENCE and I_STATE:

(2) *Marco pensa.* [Marco thinks.]

```
Marco <EVENT id="001"
class="OCCURRENCE" ...>pensa</EVENT>
```

(3) *Marco pensa di andare a casa.* [Marco thinks to go home]

```
Marco <EVENT id="001"
class="I_STATE" ...>pensa</EVENT> di
<EVENT id="002" class="OCCURRENCE"
...>andare</EVENT> a casa .
```

3 TIPSemIT: Adapting TIPSem to Italian

TIPSem is a state-of-the-art data-driven approach which uses conditional random fields (CRF) (Lafferty et al., 2001) and semantic features.

We address the problem of event detection as a sequence labeling problem, which can be also seen as a classification problem. In this bounding task, we use IOB2 labels to classify all the tokens. Given an input text, each token must be classified as being the beginning of an event, inside an event, or outside an event. The resulting IOB2 alphabet consists of B-*event*, I-*event* and O. Example 4 illustrates the event recognition problem for the sentence in example 3.

(4)

input text	problem	solution
Marco	(B-event | I-event | O)	O
pensa	(B-event | I-event | O)	B-event
di	(B-event | I-event | O)	O
andare	(B-event | I-event | O)	B-event
a	(B-event | I-event | O)	O
casa	(B-event | I-event | O)	O
.	(B-event | I-event | O)	O

The classification problem is similarly defined but restricted to those tokens which are assigned the labels B-*event* or I-*event*; e.g.:

(5)

input text	problem	solution
pensa - B-event	(TimeML Classes)	I_ACTION
andare - B-event	(TimeML Classes)	OCCURRENCE

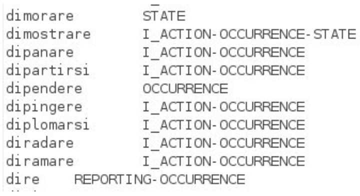
One of the most challenging part of our work is represented by the extent of the data set. As a matter of fact, the TempEval-2 data set for Italian is not very large, containing 27,152 tokens for training and 4,995 for test[1]. Our proposal maintains TIPSem machine learning environment and the general morphological features, but, in order to reduce the impact of data sparseness, we have integrated the learner with an additional semantic resource, a derived event lexicon from the SIMPLE/CLIPS lexicon (Ruimy et al., 2003; Caselli, 2009).

The tasks of event recognition and classification are tackled in a two-step approach. First, events are recognized and then the recognized events are classified. In recognition the features are obtained at the token level. The *morphological features* used are:

- lemma
- Treebank-like PoS obtained by a statistical tagger (Dell'Orletta, 2009);
- token (word)

In the development of the models we have combined the morphological features in contexts of different window sizes.

The *semantic features* are obtained from the event derived lexicon. This lexicon has been created from a mapping between the TimeML event classes and the SIMPLE/CLIPS entries at the ontological level and it is composed by 8,721 lemmas (1,068 for adjectives, 4,614 for nouns, 3,390 for verbs). The mapping has been realized in a semi-automatic way. The SIMPLE/CLIPS ontology is a multidimensional type system based on both hierarchical and non-hierarchical conceptual relations. The Event top node has seven subtypes (Perception, Aspectual, State, Act, Psychological Event, Change, Cause Change) which can be associated to one or more TimeML classes. Semantic information plays a primary role in the assignment of the TimeML classes. However, the semantic information is not always a necessary and sufficient condition for its classification. Other levels of linguistic information, such as the argument structure, may influence the class assignment. The mapping provides each event denoting expression with one or more default TimeML classes. The assignment of the right class is strictly dependent on the occurrence of each token in the text/discourse. The availability of this knowledge to the system should be useful for improving event classification. Its use in event recognition has been tested as well. Figure 1 illustrates a short portion of the lexicon for verb entries.



```
dimorare        STATE
dimostrare      I_ACTION-OCCURRENCE-STATE
dipanare        I_ACTION-OCCURRENCE
dipartirsi      I_ACTION-OCCURRENCE
dipendere       OCCURRENCE
dipingere       I_ACTION-OCCURRENCE
diplomarsi      I_ACTION-OCCURRENCE
diradare        I_ACTION-OCCURRENCE
diramare        I_ACTION-OCCURRENCE
dire    REPORTING-OCCURRENCE
```

Figure 1: Verb entries of the event lexicon.

4 Evaluation

Evaluation is divided in two experiments that correspond to the objectives of this paper. The Italian TempEval-2 data contains 4,543 events in the training set and 834 in the test set. In Table 1, we report the distributions of the event tokens in the seven TimeML classes for training and test.

We set as baseline for the evaluation a previous realization of a TimeML event detector and classification system for Italian, the TimeML TULE Converter[2] (Robaldo et al., 2011). The Converter takes as input the syntactic trees of the sentences in a document built by the TULE parser (Lesmo and Lombardo, 2002). The TULE Converter implements two different sets of rules: a group for event recognition which takes into account morphological features (PoS) and dependency relations with a set of "event trigger expressions" and a group for event classification. In particular for classification, the TULE converter exploits the derived event lexicon for having access to the TimeML class(es) associated with each event lemma and then integrates this information with syntactic information.

We have developed three data driven models to capture, incrementally, the influence of the features. The basic model, TIPSemIT_basic uses only the basic morphological features, namely lemma, token and PoS without any context window combination. The other two best performing models differ from the basic one for the combination of morphological features and presence of semantic features. In particular, TIPSemIT_FPC5 has

[1] Available at http://timeml.org/site/timebank/timebank.html

[2] The reported results differ from those published in the referred paper as the TempEval test set was not used for the evaluation. At the time of writing this article a new version of the TULE Converter has been developed only for event detection (Robaldo et al., in press). New experiments and comparisons will be performed when the Converter will be finalized also for event classes.

been obtained by adding a five window size context for lemma, token and fine-grained PoS together with bigrams for lemma and PoS. Finally, TIPSemIT_FPC5Sem adds semantic features to the previous model.

Event Classes	# training set	# test set
OCCURRENCE	2,360	456
STATE	1,089	166
I_ACTION	288	58
I_STATE	502	88
REPORTING	216	47
PERCEPTION	13	1
ASPECTUAL	75	18
Total events	4,543	834

Table 1: Event classes in TempEval-2 data

4.1 Event recognition

Table 2 reports the results for event recognition obtained by the described models.

Models	P	R	F1
TULE Converter	0.84	0.74	0.79
TIPSemIT_basic	0.90	0.77	0.83
TIPSemIT_FPC5	0.89	0.81	0.85
TIPSemIT_FPC5Sem	0.91	0.83	0.87

Table 2: Event recognition - TempEval-2 data

Although we have a very reduced corpus at disposal, TIPSemIT_basic obtains a better result with respect to the baseline in terms of precision (0.90 *vs.* 0.84) while the recall is not satisfactory (only +2%). A relative low number of events is recognized and it is close to that of the baseline system (644/834 *vs.* 624/834). It is interesting to notice that this model is not able to correctly identify 12 verb token realized by past participle forms. This is due to the PoS tagger which considers absolute past participle forms as adjectives when they are not followed by specific complement phrases (e.g. "*PP_da* + *NP*") making their identification as events more challenging. The TULE TimeML Converter does not suffer from this kind of issues, since the tagging approach adopted is different. In particular, we have observed that all events realized by verbs were correctly annotated.

The similarity of the results with respect to the recall is not surprising. The low recall of the baseline system (TimeML TULE Converter) is due to the fact that the system is not able to identify items, words and constructions which have not been implemented in the rules. Similarly, TIPSemIT_basic suffers from data sparseness. The reduced dimensions of the training set and the features used are not sufficient enough to identify previously "unseen" event instances nor to generalize information about the linguistic contexts of occurrence. The precision obtained by TIPSemIT_basic is higher than that of the TimeML TULE Converter, showing that the data-driven approach has a lower number of false positives with respect the rule-based system (72 *vs.* 117). This difference suggests that better recognition rules are to be developed, taking into account more complex features (both morphosyntactic and semantic, i.e. word-sense disambiguation). As for TIPSemIT_FPC5, the precision is slightly lower than the previous model (0.89), but the model is well balanced (recall=0.81). The increase in recall is +7% with respect to the baseline and +4% with respect to TIPsemIT_basic. The combination of PoS appears as a good strategy for approaching WSD of events realized by PoS other than verbs, especially for nouns as previously demonstrated by (Mohammad and Pedersen, 2004) (+26 nouns; +15 adjectives; +5 prepositional phrases). This model can detect instances of events which are out of range for the TimeML TULE Converter, in particular for nouns. For instance, the noun "*fuga*" [escape/flight] in example 6 is not recognized by the TimeML TULE Converter because the rules are not able to identify the causative construction realized by the presence of the preposition "*per*" [for/due to].

(6) *[...] evacuta per una* ***fuga*** *di gas.* [evacuated due to a flight of gas.]

Finally, TIPSemIT_FPC5Sem shows the highest recall (689/834 = 83%). The use of the event lexicon appears to be useful for the recognition of event nouns (+36 tokens) and adjectives (+13 tokens). One of the main contribution of the event lexicon is the reduction of data sparseness. The dimension of the training corpus is small and in order to obtain generalizations on event readings of lexical items such as nouns and adjectives a relevant number of instances are necessary. The presence of the event lexicon overcomes this limitation.

We carried out a 10-fold cross validation experiment to check if the improvement over the

TIPSemIT basic model is significant. With the results obtained, we performed a one-tailed paired t-test which showed that the mean F1 relative error reduction (21%) is statistically significant with a confidence of 99.5% ($p = 0.005$).

4.2 Event classification

The classification approaches have been evaluated over the events recognized by the best recognition model (i.e. TIPSemIT_FPC5Sem). Table 3 shows the results obtained.

Models	Accuracy
TULE Converter	0.65
TIPSemIT_basic	0.74
TIPSemIT_FP	0.74
TIPSemIT_FPC5	0.74
TIPSemIT_FPC5Sem	0.77

Table 3: Event classification - TempEval-2 data

For event classification, the TULE Converter exploits the derived event lexicon for having access to the default TimeML class and then integrates this information with syntactic information. The Converter's accuracy is lower (-12%) than that obtained by TIPSemIT_C5Sem. The primary source of errors for the Converter is due to parsing errors which prevent the activation of the corresponding rule(s), thus decreasing the number of correctly classified events. The performance improvement of the TIPSemIT_FPC5Sem with respect to the other models is due to the contribution of the event lexicon. In particular, we register an improvement in the classification of less frequent classes in the data such as I_STATE (52% *vs.* 73%), ASPECTUAL (41% *vs.* 65%) and REPORTING (53% *vs.* 68%), with the exception of I_ACTION.

In terms of number of event tokens correctly classified, the access to the event lexicon improves the classification of 42 tokens with respect to TIPSemIT_basic and TIPSemIT_FPC5. It is worth noticing that the context windows differentiating TIPSemIT_FPC5 from TIPSemIT_basic do not contribute at all to an improvement in classification, while this feature has a positive impact on event recognition.

A detailed error analysis of the event classes shows that the access to the default class information is clearly an advantage for reducing the impact of data sparseness. For instance, the verb "RAFFORZARE" [strenghten] occurs twice in the training set, and both occurrences belong to two different classes, namely OCCURRENCE and I_STATE. In the test set, this verb appears twice, once classifed as STATE, as it is realized by a past participle form, and another as I_STATE. Both TIPSemIT_basic and TIPSemIT_FPC5 can correctly classify the STATE instance thanks to the PoS information but fail in the classification of the I_STATE one. On the contrary, TIPSemIT_FPC5Sem correctly classify both cases. The correct classification of the I_STATE instance is due to the event lexicon. A 10-fold experiment has been performed to check if the improvement over TIPSemIT basic for event classification is significant. A one-tailed paired t-test showed that the mean accuracy relative error reduction (7%) is statistically significant with a confidence of 99.5% ($p = 0.005$).

However, the event lexicon is not perfect. In particular, we have observed that the coverage of the lexicon, i.e. the number of entries, must be extended especially for nouns an adjectives.

5 Conclusions and future works

This paper focuses on the adaptation to Italian of a data-driven state of the art approach based on CRF for event recognition and classification, TIPSem. Our proposal, TIPSemIT_FPC5Sem includes an Italian-specific semantic resource and has been evaluated over the available gold-standard Italian data.

The results obtained are satisfactory and show an overall improvement of 0.08% for event recognition and 0.12% in classification accuracy with respect to the baseline, i.e. the TimeML TULE Converter. This suggests that the proposed semantic features are useful for learning both event recognition and classification models.

In event recognition, the semantic features help to improve the recall without introducing too many false positives (689 events *vs.* 624 in the baseline and 644 in TIPSemIT_basic) and with a positive impact for the most difficult cases such as eventive nouns and adjectives. The results obtained from the TIPSemIT_basic and the TIPSemIT_FPC5 models are very interesting. Apparently the combination of context windows as features provides necessary information for improving event recognition even with a relative poor set of training data.

In event classification, more complex features are required. These rely on a combination of semantic and syntactic information. In addition to this, the class variability that each event lemma may give rise to requires a relatively large set of data for training. However, the results of TIPSemIT_FPC5Sem have proved that the issue of data sparseness can be dealt with *ad hoc* lexical resources, such as the derived event lexicon, which can be obtained from existing ones with a small effort.

It is worth noticing that the TIPSemIT models have a better performance with respect to the rule-based system. For instance, TIPSemIT_basic outperforms the TimeML TULE Converter in terms of precision with a reduced number of false positives. In general, the better performance of the data-driven models both in recognition and classification is due to the limitations of a rule-based approach to model complex cases. Implementing handcrafted rules for recognizing and classifying the eventive reading of nouns, adjectives and prepositional phrases is not easy and a machine learning solution appears to better deal with the complexity of the tasks.

As future work, we are planning to run a different set of experiments with more training and test data for Italian in order to assess the value of the data-driven approach, and the contribution of the semantic resource for event processing. Also, different context-window sizes will be compared. Moreover, we propose to experiment the impact of syntactic dependencies as a feature, which may facilitate the recognition and classification of events.

References

T. Caselli. 2009. *Time, events and temporal relations: an empirical model for temporal processing of Italian texts*. Ph.D. thesis, Dept. of Linguistics, University of Pisa.

T. Caselli. 2010. It-timeml: Timeml annotation scheme for italian - version 1.3.1. Technical report, ILC-CNR, Pisa.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *EVALITA 2009 - Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*.

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

L. Lesmo and V. Lombardo. 2002. Transformed subcategorization frames in chunk parsing. In *In Proc. of the 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, pages 512–519, Las Palmas.

H. Llorens, E. Saquete, and B. Navarro-Colorado. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. ACL.

S. Mohammad and T. Pedersen. 2004. Combining lexical and syntactic features for supervised word sense disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 25–32.

J. Pustejovsky, J. Castao, R. Saurì, R. Ingria, R. Gaizauskas, A. Setzer, and G. Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*.

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003b. The TIMEBANK Corpus. In *Corpus Linguistics*, pages 647–656.

L. Robaldo, T. Caselli, I. Russo, and M. Grella. 2011. From italian text to timeml document via dependency parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 177–187. Springer Berlin / Heidelberg.

L. Robaldo, T. Caselli, and M. Grella. in press. Rule-based creation of timeml documents from dependency trees. In *Proceedings of the 12th Conference of the Italian Association for Artificial Intelligence*.

N. Ruimy, M. Monachini, E. Gola, N. Calzolari, M.C. Del Fiorentino, M. Ulivieri, and S. Rossi. 2003. A computational semantic lexicon of italian: *SIMPLE. Linguistica Computazionale XVIII-XIX, Pisa*, pages 821–64.

Z. Vendler, 1967. *Linguistics and philosophy*, chapter Verbs and times, pages 97–121. Cornell University Press, Ithaca, NY.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, June.

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