MulTiSEX - A Multilanguage Timex Sequential Extractor

Stefan Rigo[†‡] and Alberto Lavelli [‡]
[‡]*Human Language Technology Research Unit, Fondazione Bruno Kessler*
[†] *DISI, University of Trento*
Trento, Italy
{rigo,lavelli}@fbk.eu

***Abstract*—In this paper we present a light-weighted Machine Learning based approach to the recognition and semantic classification of temporal expressions in different languages. We applied the proposed approach to English, Italian and Spanish with limited porting efforts. The experimental results show that our system produces state-of-the-art performance on all the corpora used and in some cases outperforms available systems.**

Keywords-Natural Language Processing; Temporal Processing; Temporal Expressions; Machine Learning

I. INTRODUCTION

Time is a critical dimension of our information space. Hence, it is not surprising that the Natural Language Processing (NLP) community is highly interested in exploiting the time dimension in text/discourse. The automatic recognition of temporal information has become an area of intense research both in Computational Linguistics (CL) and Artificial Intelligence (AI).

NLP-related systems could substantially benefit from temporal processing capabilities. Applications for Information Retrieval (IR), Information Extraction (IE), Question Answering (QA), summarization, and, last but not least, the actual development of the Semantic Web, all exploit temporal related information. Advanced content processing systems start integrating temporal reasoning in their operational architecture. For instance, a complex QA system needs to be able to operate on more information than can be extracted from temporal markers. To produce a timeline, a text summarization system needs to extract and chronologically order temporal entities.

The recognition of temporal expressions is an essential task to achieve such processing skills. Temporal expressions (henceforth, *timexes*) are the basic, time-related constituents in a temporal-information space, and they are needed to anchor events on a timeline, or to temporally structure a text. In general, the recognition and classification of events and temporal expressions, and the temporal ordering over these entities, are gaining momentum in the NLP community. Take the following example:

Example 1: *17-10-1997: Italian chemical giant Montedison S.p.A., through its Montedison Acquisition N.V. indirect unit, began its $37-a-share tender offer for all the common shares outstanding of Erbamont N.V., a maker of pharmaceuticals incorporated in the Netherlands. The offer, advertised in today's editions of The Wall Street Journal, is scheduled to expire at the end of November. Montedison currently owns about 72% of Erbamont 's common shares outstanding.*

The hermeneutic effort that humans produce in order to understand the temporal information in Example 1 is fairly small. On the other hand, a computer can hardly process questions like "When was Montedison's tender offer published?", "When does Montedison's offer expire?" or "How much of Erbamont's shares does Montedison actually own?" without accessing information about the temporal structure of a text, which, in turn, can not be produced without the recognition, classification and normalization of temporal expressions.

This paper addresses the recognition and semantic classification of temporal expressions, i.e. the identification and interpretation of phrases that convey temporal expressions in sentences. A light-weighted Machine Learning (ML) based system built using off-the-shelf components is presented. The intuition we are following is that the performance of ML-driven approaches for *timex* recognition can be improved by taking into account lexical peculiarities of *timexes* and of the context they are occurring in.

Our system produces state-of-the-art results employing a small set of generic morpho-syntactic and surface features, augmented by a database of time-related lexical items. Conditional Random Fields [11] are employed as underlying learning paradigm. We address a cross-linguistic setting as test case for our system, using the English, Spanish and Italian datasets delivered for the TempEval-2 challenge [27]. In order to be able to compare our results on a multilingual basis, and to the most recent state-of-the-art systems, we chose the TempEval-2 scorer as reference benchmark. Nevertheless, we will also evaluate our system using a stricter metric, namely the CoNLL chunking-task scorer [25].

The experiments show that our system can be ported to different languages with little effort while maintaining high accuracy both on the recognition and the classification of *timexes*. Our system performs better than the first ranking systems in the English and Spanish *timex*-recognition tasks at TempEval-2 (task A). For Italian we obtain high performance as well (comparison for Italian is not possible since there were no participants to TempEval-2 Italian task A). Moreover, we extend the evaluation to the English TimeBank [19] and the Italian ICAB corpus [13].

1530-1311/11 $26.00 © 2011 IEEE
DOI 10.1109/TIME.2011.13

The remainder of this paper is structured as follows. In Section II we give a short survey about what *timexes* are. Section III provides an overview on temporal annotation standards and corpora. In Section IV relevant works are discussed. Section V presents our system architecture. In Section VI we illustrate and discuss evaluation and results. Finally, we draw some conclusion and outline future work in Section VII.

II. WHAT IS A *TIMEX*?

Temporal expressions are natural language phrases that refer to time points or intervals. They convey temporal information on their own, but operate also as anchors for temporally locating events referred to in a text. The semantics of *timexes* can express duration, points in time, sets, direction in time (future, past). We define a *timex* as a chunk of text denoting an explicit or implicit (i.e. inferable) temporal information. *Timexes* can occur in the following forms:

- Fully specified *timexes*, like dates or times: *November 27th, 2006, the eight century,11/27/2006 at 12:00*
- Anaphoric *timexes* anchored to an expression in the local context: *three days after the meeting, the previous month, two weeks since the exam*
- Deictic *timexes* (indexical), anchored to the time an expression was written (e.g. the document creation time, DCT): *today, this day, currently, now, last five weeks, tomorrow*
- Durations or intervals: *a week, five days, three semesters*
- Frequencies: *weekly, every other day, once a month, every first day of the week*
- Culturally dependent *timexes*: *Christmas, Easter, World AIDS Day*
- Fuzzy (quantified) *timexes*: *the past, some day*

The recognition of a *timex* consists firstly in bracketing the extension of the chunk representing it, and successively in assigning it a type-value. Finally, *timexes* have to be normalized, i.e. an ISO-8601 value has to be assigned to each recognized *timex*. The value is constrained by the type of the temporal expression. The normalization step is not discussed in this paper.

III. BACKGROUND

Work on temporal annotation of English content started in the middle of the nineties (Message Understanding Conference - MUC-6, 1995), gradually evolving in annotation schemes of increasing expressive power. In MUC-6 the recognition of temporal expressions was part of the Named Entity Recognition (NER) task, in which tokens had to be classified with labels from a given set. In 2004 the Automated Content Extraction (ACE) conference proposed a competition for Temporal Expression Recognition and Normalization (TERN 2004), which extended the complexity of the task: temporal expressions had to be recognized in free text and normalized using an ISO-based value. TERN 2004 was the first challenge where *timex* recognition and normalization were tackled as a task distinct from NER. Moreover, TERN 2004 was important for the fact that the TIDES TIMEX2 annotation guidelines [6] were established.

More recent efforts produced the specification language TimeML [20], an annotation scheme for events, temporal expressions and temporal relations in text/discourse. In this paper we refer to the TIMEX3 guidelines, which are defined in the TimeML specifications. TIMEX3 entities are classified as DATE, TIME, DURATION and SET.

The development of TimeML led to the creation of TimeBank [19], a temporally annotated corpus of English documents. The latest release is TimeBank 1.2.1, containing 186 articles with approximately 68,000 tokens. The content is taken from a variety of different sources, including newswire and transcribed broadcast news.

The relevant corpus for our purposes is the TempEval-2 [27] multilingual training and test data, which we will use for our cross-linguistic experiments. TempEval-2 data was annotated with a simplified version of TimeML. The type and value attributes of *timexes* are annotated. The English data sets are derived from TimeBank.

While the TempEval-2 corpora and scoring script are our primary benchmarking reference, we evaluate our system also on TimeBank and on the Italian ICAB corpus [13] annotated according to the It-TimeML [4] specifications.

IV. RELEVANT WORK

The best performing system at TERN 2004 English Full Task, the rule-based CHRONOS [16] achieved F_1 results of 87% for bracketing and normalization. After the TERN challenge, [1], [8] and [18] used sequential approaches for the classification of temporal expressions and used the TERN dataset as test bed. F_1 results for bracketing range from 77% [18] to 88% [8]. The system described in [2], while being mainly a ML approach, uses a small set of rules to compute first an underspecified representation for recognized *timexes* and subsequently to assign a ISO-conform value. They report an F_1 measure of 90% for the recognition and bracketing of *timex*-chunks. Using TimeBank as dataset, [3] built a *timex*-classifier based on a cascaded finite-state grammar. They report an F_1 measure around 82% for bracketing and around 68% for simultaneous bracketing and typing. [10] used a Maximum Entropy classifier and TimeBank for training and test, reporting an F_1 measure of around 82% for bracketing. One of the best performing TempEval-2 systems [12] uses a sequential ML approach together with semantic roles. Results are F_1 of 85% for bracketing, and accuracy of 92% for the classification (cf. Table V for all TempEval-2 English results).

As for Italian, to the best of our knowledge, the only TIMEX3-compliant system is the rule-based TETI [5]. TETI was evaluated on a manually annotated subset of the Italian syntactic-semantic Treebank [14] and is limited to the bracketing of *timexes*, i.e. they are neither type-classified nor normalized. Results are around 86% (F_1) for the recognition of temporal expressions. Another rule-based

system (adhering to the TIMEX2 specifications and evaluated on a TIMEX2-compliant version of ICAB), ITA-Chronos [15], has an F_1 performance of 92% for the recognition (i.e. bracketing), and of 67% for the recognition and normalization of temporal expressions (i.e. bracketing and assigning a value).

The first Spanish TIMEX2-compliant system was TERSEO [22], a knowledge-based architecture, which performs with an F_1 of 77% on a manually annotated corpus of Spanish news. The same system was automatically adapted to English and Italian [21], and evaluated on the English TERN 2004 corpus and on a TIMEX2-adhering version of the Italian ICAB corpus. Results for English are comparable to other systems, while for Italian performance is considerably lower compared to ITA-Chronos. The rule-based system that participated to the ACE 2007 TERN pilot-task for Spanish [28] was evaluated on a TIMEX2-annotated corpus of Spanish news. Results range around F_1 = 62% for recognition and bracketing, and accuracy over 95% for the classification of recognized *timexes*. The best performing system at TempEval-2 task A for Spanish is also one of the best for English [12]. The ML-based architecture performs with F_1 of 91% for recognition and bracketing, and with an accuracy of 91% for classification. The other participating system was UC3M [29], a rule-based architecture that performed with comparable results (cf. Table VI for all TempEval-2 Spanish results).

V. SYSTEM ARCHITECTURE

Most state-of-the-art systems for the classification of temporal expressions recognize and label *timexes* in two steps, i.e. first bracketing the extension of the TIMEX3-chunk and then classifying it. On the contrary, we tackle the recognition of *timexes* as a classification problem in a word-chunking paradigm, i.e. as a sequential labeling task where chunk information is encoded into token tags. For the classification of *timexes*, we use an I-O-B tagging convention, i.e. the tokens are labeled either as (B)eginning, (I)nside or (O)utside a TIMEX3 chunk. The *timex*-phrase is classified with one of the four values for the type-attribute of the TIMEX3 tag (or with the OTHER label). Example 2 shows an annotated fragment:

Example 2:

Several	B-DATE
years	I-DATE
ago	I-DATE
it	O
was	O
hot	O
as	O
this	B-DATE
year	I-DATE
.	O

A. Conditional Random Fields

Timexes are connected with the structural properties of sentences. They are related to the syntactic structure of phrases and with morphological, semantic and syntactic properties of the words in their context. Furthermore, *timexes* are often expressed by sequences of words. We argue that Conditional Random Fields (CRF) [11] are well suited to tackle *timex* recognition and classification in a sequential chunking task.

CRFs ground on exponential models where probabilities are computed from the values of a set of features derived from both the observation and the label sequences. For sequence labeling problems linear chain CRFs are widely used. They are expressed in the following form:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1} \sum_{k} \lambda_k f_k(t, y_{t-1}, y_t, x) \right) \quad (1)$$

where Z(x) is the normalization factor, $X=\{x_1,...,x_n\}$ represents the observation sequence, $Y=\{y_1,...,y_t\}$ represents the label sequence, and f_k and λ_k represent the feature functions and their respective weights.

CRFs have been employed for PoS tagging, shallow parsing, and named entity recognition. We use the CRF++ toolkit for our experiments.[1]

B. Feature engineering

We use a small set of features, based on morpho-syntactic and surface properties. *Timexes* exhibit various textual properties, ranging from patterns that can be recognized using simple regular expressions to complex linguistic forms (phrases). While *timexes* are realized in different phrase types, the lexical items expressing *timexes*, and their modifiers, quantifiers, adverbs, adjectives etc, form a restricted set. Grounding on this consideration, two vectors that generalize the occurrence of different classes of temporally related items in a given context-window are assembled. To be able to do this, we classify lexical items using a set of 22 categories. These are:

1. Items matching duration patterns (e.g. *20", eight-minute*)
2. Items matching year-patterns (e.g. *2011, '45*)
3. Items matching time-patterns (e.g. *11:19, 01.30*)
4. Items matching ordinals (e.g. *first [of July]*)
5. Items matching digits
6. Items matching numbers expressed through strings (e.g. *five [weeks])*
7. Weekdays
8. Months
9. Part of day (e.g. *morning*, *noon*)
10. Seasons
11. Festivities
12. Items referring to the past (e.g. *earlier*, *last*)
13. Items referring to the present (e.g. *current*, *now*)
14. Items referring to the future (e.g. *next*, *coming*)
15. Fuzzy quantifiers (e.g. *few*, *some*)
16. Modifiers (e.g. *long*, *short*)
17. Temporal adverbs (e.g. *daily*, *earlier*)

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

18. Temporal adjectives (e.g. *early*)
19. Temporal conjunctives (e.g. *when*, *meanwhile*)
20. Temporal prepositions (e.g. *during*, *for*)
21. Time units (e.g. *seconds*, *minutes*)
22. Temporal co-references (e.g. *time*, *period*)

The first vector (p1) summarizes both the category of the processed token, and the categories of the surrounding items together with their relative position in an eight token context window. We do this by assigning a letter to each category (and to the case the item does not match any of the defined categories) and assembling a string where the rightmost character expresses the category of the forth token after the parsed one, the leftmost the category of the first token in the context window (-3 to 4). Lexical items that belong to different categories (e.g. *early* can be an adverb or an adjective) are currently not disambiguated with regard to the PoS tag.

The second vector is a simple sequence of binary flags (p2) expressing the properties of the processed token, i.e. we are encoding all categories a given token belongs to. The feature set was developed over the ICAB corpus applying 10-fold cross-validation, and is listed in Table I.

We first built our system for Italian and successively adapted it to English and Spanish. All we needed were a POS-Tagger and a stemmer for these languages, and a database of language-specific temporally related items. TreeTagger [23] was used for PoS tagging and the Porter Stemmer for stemming.[2]

TABLE I. FEATURE SET

Feature	Description
token	Baseline system – lowercase token
PoS	Part of Speech of token
p1	Context and category aware vector
p2	Vector of binary flags
3pos	PoS 3-gram (previous, current, next)
3stm	Stem 3-gram (previous, current, next)
abs	Surface properties abstractions
3-gr	Character 3-grams of token

For the purpose of this paper the items in the lexical database have been translated manually. However, we argue that Machine Translation (MT) could be employed with limited effort, since we are handling single tokens. The choice to manually translate the database was motivated by the fact that we wanted to i) empirically investigate differences in temporal semantics among languages, and ii) create a gold standard to evaluate the automatic translation (which we plan to implement in the near future).

[2] We used the Python implementation of the Snowball project, http://snowball.tartarus.org.

VI. EVALUATION

While the English and Spanish training sets delivered for TempEval-2 have approximately the same size (around 67,000 tokens), the Italian data set is smaller (approximately 28,000 tokens). Nevertheless, results for all languages reach state-of-the-art performance, showing that our system is easily portable among languages.

Moreover, we discuss results obtained using 10-fold cross validation over the Italian ICAB corpus (approx. 200,000 tokens) and the English TimeBank (approx. 68,000 tokens).

A. Metrics

As previously stated, we use the TempEval-2 data sets and scoring script as reference benchmark. This allows us to compare our architecture to the most recent state-of-the-art systems for the processing of temporal expressions. However, since the TempEval-2 scorer is "relaxed" (i.e. it computes the score on a token-basis and not on a phrase-basis) we will also evaluate our system using a stricter scoring method (i.e. based on exact matches between the recognized and the gold-standard chunks), namely the CoNLL chunking task scoring script.

The TempEval-2 scorer computes precision, recall and F_1-measure of recognized and bracketed *timex*-extents on a token-basis, using the following formulas:

$$Precision = tp/(tp + fp) \quad (2)$$

$$Recall = tp/(tp + fn) \quad (3)$$

$$F_1 = 2 *(Precision * Recall)/(Precision + Recall) \quad (4)$$

where *tp* is the number of tokens that are part of both the recognized and the gold-extent, *fp* is the number of tokens that are part of a recognized extent but not of the gold-standard, and *fn* is the number of tokens that are part of the gold-extent but not in the recognized extent. The semantic class of a time expression is computed on the recognized *timexes*, using a simple metric: the number of correct answers divided by the number of answers (last column in Table II).

On the other hand, the CoNLL chunking task scorer computes performance on a phrase-basis. The performance is measured with three figures: i) the percentage of detected phrases that are correct (precision), ii) the percentage of phrases in the test-data that were correctly recognized (recall), and iii) the F_1 rate, which is computed with the same formula as in the TempEval-2 scoring script.

Moreover, the CoNLL scoring script gives an F_1 measure of the combined recognition *and* classification task. Hence, while we use the TempEval-2 scorer as reference benchmark, we will also use the CoNLL scorer since it provides us a more expressive measure. We exemplify the difference between the two scripts using the fragment in Example 2. Let us assume a system recognizes and classifies the chunk *years/B-DATE ago/I-DATE* (hence, missing the initial token of the gold-extent: *Several*), and correctly found and classified the temporal expression *this/B-DATE year/I-DATE*. For the first chunk, the TempEval-2 scorer would

compute a precision of 2/(2+0) = 1, recall of 2/(2+1) = 0.66, and F1 = 2*(1*0.66)/(1+0.66) = 0.79. For the second chunk we get precision = recall = F_1 = 1. The final TempeEval-2 figures on the fragment in example 2 are precision = 1, recall = 0.83, and F_1 = 0.89. On the contrary, the CoNLL scorer would yield precision = recall = F_1 = 0.50.

B. Results

We trained and evaluated our TIMEX3-classifier on the TempEval-2 multilingual data sets using the features in Table I. Figures 1 and 2 show the results obtained by subsequently adding the listed features.

We observed differences related to the employed PoS tagger. While the choice of TreeTagger was motivated also by the fact that it was available for all languages used in our experiments, it did not seem to be the best choice for every language. For instance in the case of English, if we use the PoS tagger that is part of the TextPro suite [17], we gain 2% for the F_1 measure on both scoring methods, while for Italian results remain the same (TextPro does not provide a PoS tagger for Spanish).

Figures 3, 4 and 5 show the incremental results for precision, recall, and F_1 for the combined recognition *and* classification of *timexes* for the languages in our test case. We used TreeTagger and the CoNLL scorer on the TempEval-2 data sets.

Concerning PoS tagging, differences in accuracy emerged also on the Italian ICAB corpus. Performing a 10-fold cross-validation, we gain 3% in F_1 results on exact match and 1% on token-based match when using TextPro instead of TreeTagger.

As can be seen in Figure 4, the PoS-3-gram feature seems to penalize the performance on the Spanish data. Training a model without this feature would yield a 2% F1 gain on exact match, while differences for the token-based scoring are negligible. Similar performance penalties can be observed also for Italian (surface abstraction) and English (p1 and stem-3-gram). However, in both cases subtracting these features from the set did lower the overall performance. We argue that the performance gain observed on the Spanish test data is peculiar to this corpus and that it cannot be generalized over other corpora and/or languages. The low recall seen on the Italian TempEval-2 corpus is clearly an indication of its reduced size.

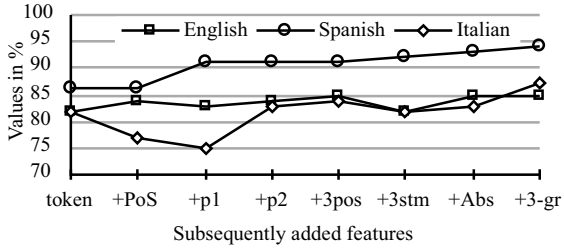

Figure 1. F1 – TempEval-2 scorer (TempEval-2, TreeTagger)

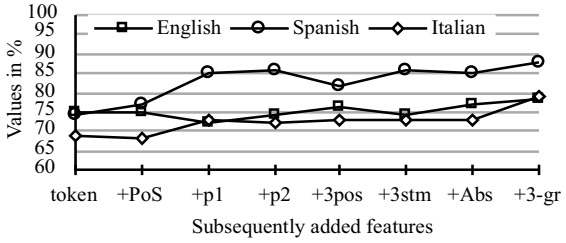

Figure 2. F1 – CoNLL scorer (TempEval-2, TreeTagger)

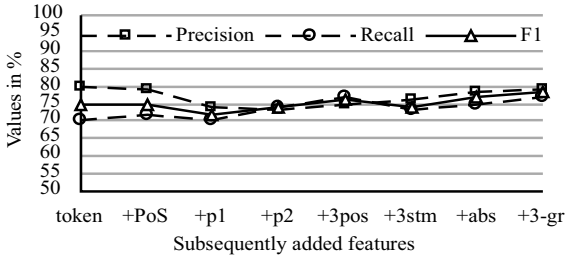

Figure 3. English F1 exact match (TempEval-2, TreeTagger)

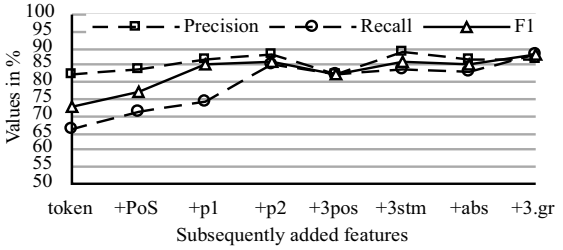

Figure 4. Spanish F1 exact match (TempEval-2, TreeTagger)

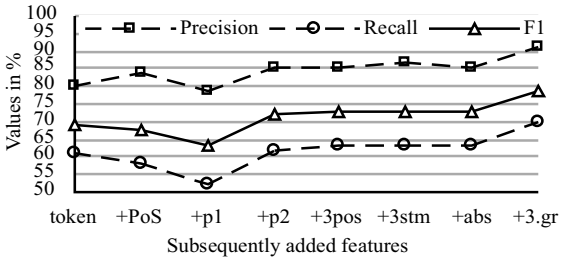

Figure 5. Italian F1 exact match (TempEval-2, TreeTagger)

Table II resumes the results obtained with the TempEval-2 scoring script, Table III those obtained using the CoNLL scorer. Finally, Table IV shows the results obtained for the sole bracketing of *timexes* (using the CoNLL scorer). Figures 6 and 7 show the results obtained respectively on the English TimeBank (TB) and the Italian ICAB. TextPro was used on ICAB, and on the English TempEval-2 (TE-2) corpus, TreeTagger for all other corpora.

The experiments showed that we could gain additional accuracy using different PoS taggers (as for English). For the purpose of comparing our system to others, on the TempEval-2 data sets we used TreeTagger for PoS-tagging on the Spanish and Italian corpora, and TextPro for English. This is the only adaption we made. We used the feature set in Table I for all languages.

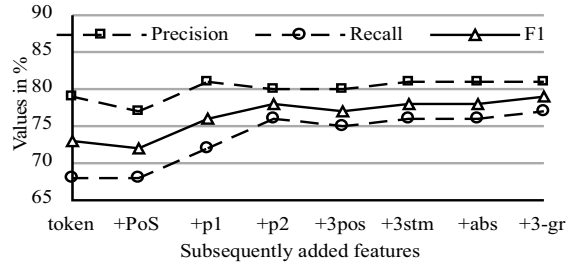

Figure 6. TimeBank F1 exact match (TreeTagger)

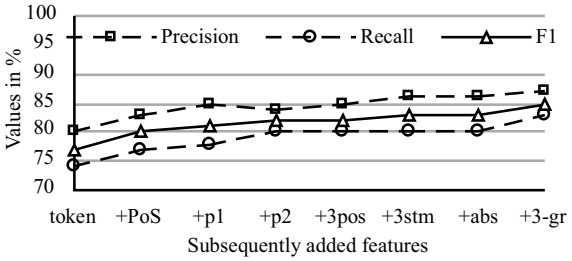

Figure 7. ICAB F1 exact match (TextPro)

For the subtasks of recognizing the extent and type-value of a TIMEX3, our classifier outperforms all systems (English and Spanish) submitted for task A at TempEval-2 [27]. Tables 5 and 6 show our results compared to the English and Spanish systems that participated to TempEval-2 task A. As for Italian, we obtain a precision of 95%, recall of 79% and F_1 of 87%, and our system correctly classified 96% of the recognized TIMEX3 chunks. Although this result is satisfactory, it reflects the fact that the used training corpus has half the size of the other corpora. There is no system we can use for direct comparison of the results obtained using the *timex*-classifier on the Italian TempEval-2 data sets.

VII. CONCLUSION AND FUTURE WORK

We have presented a state-of-the-art ML-classifier for temporal expressions and have shown that it can be adapted to different languages without major efforts. A small set of generic, morpho-syntactic features is employed. The feature representation is augmented using a lexical database of time-related items. We evaluated our architecture across different languages and corpora. Our classifier outperforms all systems that participated to the English and Spanish task A of TempEval-2, and reaches similar high performance on the Italian TempEval-2 data set. State-of-the-art performance is obtained also on the TimeBank and ICAB corpora. While we are getting satisfactory results, we argue there is still room for improvement, particularly with regard to feature engineering.

We are currently working on a normalization module for both English and Italian, and on PoS-based disambiguation of items that belong to different categories of our database of time-related lexical items. Finally, we will explore the use of Machine Translation to translate database of temporally related items.

TABLE II. RECOGNITION & CLASSIFICATION - TEMPEVAL-2 SCORER

Dataset	Precision	Recall	F1	Class-acc.
TE-2 ENG	0.89	0.85	0.87	0.93
TE-2 SPA	0.94	0.93	0.94	1
TE-2 ITA	0.95	0.79	0.87	0.96
TB (ENG)	0.92	0.87	0.90	0.92
ICAB (ITA)	0.96	0.90	0.93	0.95

TABLE III. RECOGNITION & CLASSIFICATION - CONLL SCORER

Dataset	Precision	Recall	F1 rec+class
TE-2 ENG	0.82	0.78	0.80
TE-2 SPA	0.87	0.88	0.88
TE-2 ITA	0.91	0.70	0.79
TB (ENG)	0.81	0.77	0.79
ICAB (ITA)	0.87	0.83	0.85

TABLE IV. BRACKETING RESULTS - CONLL SCORER

Dataset	Precision	Recall	F1 bracketing
TE-2 ENG	0.84	0.81	0.82
TE-2 SPA	0.87	0.88	0.88
TE-2 ITA	0.93	0.73	0.82
TB (ENG)	0.86	0.83	0.85
ICAB (ITA)	0.90	0.88	0.89

TABLE V. TEMPEVAL-2 TASK A ENGLISH RESULTS

TempEval-2 English	Precision	Recall	F1	type
HEIDELTIME1 [24]	0.90	0.82	0.86	0.96
HEIDELTIME2 [24]	0.82	0.91	0.86	0.92
TIPSEM [12]	0.92	0.80	0.85	0.92
TRIOS & TRIPS [26]	0.85	0.85	0.85	0.94
EDINBURGH [7]	0.85	0.82	0.84	0.84
KUL-3 [9]	0.85	0.84	0.84	0.91
MULTISEX ENGLISH	**0.89**	**0.85**	**0.87**	**0.93**

TABLE VI. TEMPEVAL-2 TASK A SPANISH RESULTS

TempEval-2 Spanish	Precision	Recall	F1	type
TIPSEM [12]	0.95	0.87	0.91	0.91
TIPSEM -B [12]	0.97	0.81	0.88	0.99
UC3M [29]	0.90	0.87	0.88	0.91
MULTISEX SPANISH	**0.94**	**0.93**	**0.94**	**1**

ACKNOWLEDGMENTS

This work has been partially supported by the LiveMemories project (Active Digital Memories of Collective Life - http://www.livememories.org) funded by the Autonomous Province of Trento under the call "Major Project 2006". We would like to thank Matteo Negri and Emanuele Pianta for support and insights.

REFERENCES

[1] Ahn, D., Adafre, S.F., Rijke, M.D. Towards Task-Based Temporal Extraction and Recognition. In *Proceedings of the Dagstuhl Workshop on Annotating, Extracting, and Reasoning about Time and Events*, Dagstuhl, Germany, 2005.

[2] Ahn, D., van Rantwijk, J., de Rijke, M. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In *Proceedings of the Annual Conference of the North American Chapter of the ACL NAACL-HLT 2007*, 2007, 420-427.

[3] Boguraev, B., Ando, R.K. TimeBank-Driven TimeML Analysis. In *Proceedings of the Dagstuhl Workshop on Annotating, Extracting, and Reasoning about Time and Events*, Dagstuhl, Germany, 2005.

[4] Caselli, T. It-TimeML Annotation Scheme for Italian. Linee guida per l'applicazione di uno schema di annotazione. Version 1.3.1, ILC-PISA, Pisa (Italy), 2010.

[5] Caselli, T. *Time, Events and Temporal Relations: an Empirical Model for Temporal Processing of Italian Texts.* University of Pisa, PhD Thesis, 2009.

[6] Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. TIDES. 2003 Standard for the Annotation of Temporal Expressions, 2004.

[7] Grover, C., Tobin, R., Alex, B., Byrne, K. Edinburgh-LTG: TempEval-2 system description. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010, 333-336.

[8] Hacioglu, K., Chen, Y., Douglas, B. Automatic Time Expression Labeling for English and Chinese Text. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'05)*, 2005, 548-559.

[9] Kolomiyets, O., Moens, M.-F. KUL: Recognition and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010, 325-328.

[10] Kolomiyets, O., Moens, M.-F. Meeting TempEval-2: Shallow Approach for Temporal Tagger. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, Boulder, Colorado, 2009, 52-57.

[11] Lafferty, J.D., McCallum, A., Pereira, F.C.N. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In *Proceedings of the ICML '01*, 2001, 282-289.

[12] Llorens, H., Saquete, E., Navarro, B. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010, 284-291.

[13] Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06),* Genova, Italy, 2006.

[14] Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzar, O., Lenci, A., Pirelli, V., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R. "The syntactic-semantic treebank of Italian. An overview". *Linguistica Computazionale, Computational Linguistics in Pisa, special Issue*, vol. XVIII-XIX, 2003, 461-493.

[15] Negri, M. Dealing with Italian temporal expressions: The ITA-Chronos system. In *Proceedings of EVALITA 2007. Workshop held in conjunction with AI*IA*, 2007.

[16] Negri, M., Marseglia, L. Recognition and normalization of time expressions: Itc-irst at TERN 2004. Technical Report, ITC-irst, Trento, 2004.

[17] Pianta, E., Girardi, C., Zanoli, R. The TextPro tool suite. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech (Morocco), 2008.

[18] Poveda, J., Surdeanu, M., Turmo, J. A Comparison of Statistical and Rule-Induction Learners for Automatic Tagging of Time Expressions in English. In *Proceedings of the 14th International Symposium on Temporal Representation and Reasoning (TIME'07)*, 2007, 141-149.

[19] Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M. The TIMEBANK Corpus. In *Proceedings of the Corpus Linguistics*, 2003, 647-656.

[20] Pustejovsky, J., Ingria, R., Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. The Specification Language TimeML. The Language of Time: A Reader. Oxford University Press (2004).

[21] Saquete, E., Martínez-Barco, P., Muñoz, R., Negri, M., Speranza, M., Sprugnoli, R. Multilingual extension of a temporal expression normalizer using annotated corpora. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*, Trento, Italy, 2006, 1-8.

[22] Saquete, E., Muñoz, R., Martínez, P. "Event ordering using TERSEO system". *Data & Knowledge Engineering - Special issue: Application of Natural Language to Information Systems (NLDB'04)*, vol. 58-1, 2006, 70-89.

[23] Schmid, H. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the First International Conference on New Methods in Language Processing*, Manchester, England, 1994, 44-49.

[24] Strötgen, J., Gertz, M. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)* , Uppsala, Sweden, 2010, 321-324.

[25] Tjong Kim Sang, E.F., Buchholz, S. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the CoNLL-2000*, Lisbon, Portugal, 2000, 127-132.

[26] UzZaman, N., Allen, J. TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010, 276-283.

[27] Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010.

[28] Vicente-Diéz, M.T., de Pablo-Sanchez, C., Martínez, P. "Evaluación de un sistema de reconocimiento y

normalización de expresiones temporales en español". *Procesamiento del Lenguaje Natural*, vol. 39-2007, 113-120.

[29] Vicente-Diéz, M.T., Schneider, J.M., Martínez, P. UC3M system: Determining the Extent, Type and Value of Time Expressions in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 2010, 329-332.