Meeting TempEval-2: Shallow Approach for Temporal Tagger

Oleksandr Kolomiyets
Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A, Heverlee, Belgium
oleksandr.kolomiyets
@cs.kuleuven.be

Marie-Francine Moens
Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A, Heverlee, Belgium
sien.moens@cs.kuleuven.be

Abstract

Temporal expressions are one of the important structures in natural language. In order to understand text, temporal expressions have to be identified and normalized by providing ISO-based values. In this paper we present a shallow approach for automatic recognition of temporal expressions based on a supervised machine learning approach trained on an annotated corpus for temporal information, namely TimeBank. Our experiments demonstrate a performance level comparable to a rule-based implementation and achieve the scores of 0.872, 0.836 and 0.852 for precision, recall and F1-measure for the detection task respectively, and 0.866, 0.796, 0.828 when an exact match is required.

1 Introduction

The task of recognizing temporal expressions (sometimes also referred as time expressions or simply TIMEX) was first introduced in the Message Understanding Conference (MUC) in 1995. Temporal expressions were treated as a part of the Named Entity Recognition (NER) task, in which capitalized tokens in text were labeled with one of the predefined semantic labels, such as Date, Time, Person, Organization, Location, Percentage, and Money. As the types of temporal entities identified in this way were too restricted and provided little further information, the Automated Content Extraction (ACE) launched a competition campaign for Temporal Expression Recognition and Normalization (TERN 2004). The tasks were to identify temporal expressions in free text and normalize them providing an ISO-based date-time value. Later evaluations of ACE in 2005, 2006 and 2007 unfortunately did not set new challenges for temporal expression recognition and thus the participation interest in this particular task decreased.

TempEval-2 is a successor of TempEval-2007 and will take place in 2010. The new evaluation initiative sets new challenges for temporal text analysis. While TempEval-2007 was solely focused on recognition of temporal links, the TempEval-2 tasks aim at an all-around temporal processing with separate evaluations for recognition of temporal expressions and events, for the estimation of temporal relations between events and times in the same sentence, between events and document creation time, between two events in consecutive sentences and between two events, where one of them syntactically dominates the other (Pustejovsky et al., 2009). These evaluations became possible with a new freely available corpus with annotated temporal information, TimeBank (Pustejovsky et al., 2003a), and an annotation schema, called TimeML (Pustejovsky et al., 2003b).

For us all the tasks of TempEval-2 seem to be interesting. In this paper we make the first step towards a comprehensive temporal analysis and address the problem of temporal expression recognition as it is set in TempEval-2. Despite a number of previous implementations mainly done in the context of the ACE TERN competition, very few,

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 52–57,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

and exclusively rule-based methods were reported for temporal taggers on TimeBank developed by using the TimeML annotation scheme. As a main result of the deep analysis of relevant work (Section 2), we decided to employ a machine learning approach for constituent-based classifications with generic syntactic and lexical features.

The remainder of the paper is organized as follows: in Section 2 we provide the details of relevant work done in this field along with corpora and annotations schemes used; Section 3 describes the approach; experimental setup, results and error analysis are provided in Section 4. Finally, Section 5 gives an outlook for further improvements and research.

2 Related Work

For better understanding of the performance levels provided in the paper we first describe evaluation metrics defined for the temporal expression recognition task and then the methods and datasets used in previous research.

2.1 Evaluation metrics

With the start of the ACE TERN competition in 2004, two major evaluation conditions were proposed: Recognition+Normalization (full task) and Recognition only (TERN, 2004).

Detection (Recognition): Detection is a preliminary task towards the full TERN task, in which temporally relevant expressions have to be found. The scoring is very generous and implies a minimal overlap in the extent of the reference and the system output tags. As long as there is at least one overlapping character, the tags will be aligned. Any alignment of the system output tags are scored as a correct detection.

Sloopy span: Spans usually refer to strict match of both boundaries (the extent) of a temporal expression (see Exact Match). "Sloopy" admits recognized temporal expressions as long as their right boundary is the same as in the corresponding TimeBank's extents (Boguraev and Ando, 2005). The motivation was to assess the correctness of temporal expressions recognized in TimeBank, which was reported as inconsistent with respect to some left boundary items, such as determiners and pre-determiners.

Exact Match (Bracketing or Extent Recognition): Exact match measures the ability to correctly identify the extent of the TIMEX. The extent of the reference and the system output tags must match exactly the system output tag to be scored as correct.

2.2 Datasets

To date, there are two annotated corpora used for temporal evaluations, the ACE TERN corpus and TimeBank (Pustejovsky et al., 2003a). In this section we provide a brief description of the temporal corpora and annotation standards, which can substantially influence recognition results.

Most of the implementations referred as the state-of-the-art were developed in the scope of the ACE TERN 2004. For evaluations, a training corpus of 862 documents with about 306 thousand words was provided. Each document represents a news article formatted in XML, in which TIMEX2 tags denote temporal expressions. The total number of temporal expressions for training is 8047 TIMEX2 tags with an average of 10.5 per document. The test set comprises 192 documents with 1828 TIMEX2 tags (Ferro, 2004).

The annotation of temporal expressions in the ACE corpus was done with respect to the TIDES annotation guidelines (Ferro et al., 2003). The TIDES standard specifies so-called markable expressions, whose syntactic head must be an appropriate lexical trigger, e.g. "*minute*", "*afternoon*", "*Monday*", "*8:00*", "*future*" etc. When tagged, the full extent of the tag must correspond to one of the grammatical categories: nouns (NN, NNP), noun phrases (NP), adjectives (JJ), adjective phrases (ADJP), adverbs (RB) and adverb phrases (ADVP). According to this, all pre- and postmodifiers as well as dependent clauses are also included to the TIMEX2 extent, e.g. "*five days after he came back*", "*nearly four decades of experience*". Such a broad extent for annotations is of course necessary for correct normalization, but on the other hand, introduces difficulties for exact match. Another important characteristic of the TIDES standard are the nested temporal expressions as for example:

```
<TIMEX2>The<TIMEX2 VAL = "1994">1994
</TIMEX2> baseball season </TIMEX2>
```

The most recent annotation language for temporal expressions, TimeML (Pustejovsky et al., 2003b), with an underlying corpus TimeBank (Pustejovsky et al., 2003a), opens up new possibilities for processing temporal information in text. Besides the specification for temporal expressions, i.e. TIMEX3, which is to a large extent inherited from TIDES, TimeML provides a means to capture temporal semantics by annotations with suitably defined attributes for fine-grained specification of analytical detail (Boguraev et al., 2007). The annotation schema establishes new entity and relation marking tags along with numerous attributes for them. This advancement influenced the extent for event-based temporal expression, in which dependent clauses are no longer included into TIMEX3 tags. The TimeBank corpus includes 186 documents with 68.5 thousand words and 1423 TIMEX3 tags.

2.3 Approaches for temporal processing

As for any recognition problem, there are two major ways to solve it. Historically, *rule-based systems* were first implemented. Such systems are characterized by a great human effort in data analysis and rule writing. With a high precision such systems can be successfully employed for recognition of temporal expressions, whereas the recall reflects the effort put into the rule development. By contrast, *machine learning methods* require an annotated training set, and with a decent feature design and a minimal human effort can provide comparable or even better results than rule-based implementations. As the temporal expression recognition is not only about to detect them but also to provide an exact match, machine learning approaches can be divided into *token-by-token classification* following **B**(egin)-**I**(nside)-**O**(utside) encoding and *binary constituent-based classification*, in which an entire chunk-phrase is under consideration to be classified as a temporal expression or not. In this case, exact segmentation is the responsibility of the chunker or the parser used.

Rule-based systems: One of the first well-known implementations of temporal taggers was presented in (Many and Wilson, 2000). The approach relies on a set of hand-crafted and machine-discovered rules, which are based upon shallow lexical features. On average the system achieved a value of 83.2% for F1-measure against hand-annotated data. The dataset used comprised a set of 22 New York Times articles and 199 transcripts of Voice of America taken from the TDT2 collection (Graff et al., 1999). It should be noted that the reported performance was provided in terms of an exact match. Another example of rule-based temporal taggers is Chronos described in (Negri and Marseglia, 2004), which achieved the highest scores (F1-measure) in the TERN 2004 of 0.926 and 0.878 for recognition and exact match.

Recognition of temporal expressions using TimeBank as an annotated corpus, is reported in (Boguraev and Ando, 2005) based on a cascaded finite-state grammar (500 stages and 16000 transitions). A complex approach achieved an F1-measure value of 0.817 for exact match and 0.896 for detecting "sloopy" spans. Another known implementation for TimeBank is an adaptation of (Mani and Wilson, 2000) from TIMEX2 to TIMEX3 with no reported performance level.

Machine learning recognition systems: Successful machine learning TIMEX recognition systems are described in (Ahn et al., 2005; Hacioglu et al., 2005; Poveda et al., 2007). Proposed approaches made use of a token-by-token classification for temporal expressions represented by B-I-O encoding with a set of lexical and syntactic features, e.g., token itself, part-of-speech tag, label in the chunk phrase and the same features for each token in the context window. The performance levels are presented in Table 1. All the results were obtained on the ACE TERN dataset.

Approach	*F1* (detection)	*F1* (exact match)
Ahn et al., 2005	0.914	0.798
Hacioglu et al., 2005	0.935	0.878
Poveda et al., 2007	0.986	0.757

Table 1. Performance of Machine Learning Approaches with B-I-O Encoding

Constituent-based classification approach for temporal expression recognition was presented in (Ahn et al., 2007). By comparing to the previous work (Ahn et al., 2005) on the same ACE TERN dataset, the method demonstrates a slight decrease in detection with F1-measure of 0.844 and a nearly equivalent F1-measure value for exact match of 0.787.

The major characteristic of machine learning approaches was a simple system design with a minimal human effort. Machine-learning based recognition systems have proven to have a comparable recognition performance level to state-of-the-art rule-based detectors.

3 Approach

The approach we describe in this section employs a machine-learning technique and more specifically a binary constituent based classification. In this case the entire phrase is under consideration to be labeled as a TIMEX or not. We restrict the classification for the following phrase types and grammatical categories: NN, NNP, CD, NP, JJ, ADJP, RB, ADVP and PP. In order to make it possible, for each sentence we parse the initial input line with a Maximum Entropy parser (Ratnaparkhi, 1998) and extract all phrase candidates with respect the types defined above. Each phrase candidate is examined against the manual annotations for temporal expressions found in the sentence. Those phrases, which correspond to the temporal expressions in the sentence are taken as positive examples, while the rest are considered as negative ones. Only one sub-tree from a parse is marked as positive for a distinct TIMEX at once. After that, for each candidate we produce a feature vector, which includes the following features: head phrase, head word, part-of-speech for head word, character type and character type pattern for head word as well as for the entire phrase. Character type and character type pattern[1] features are implemented following Ahn et al. (2005). The patterns are defined by using the symbols X, x and 9. X and x are used for character type as well as for character type patterns for representing capital and lower-case letters for a token. 9 is used for representing numeric tokens. Once the character types are computed, the corresponding character patterns are produced. A pattern consists of the same symbols as character types, and contains no sequential redundant occurrences of the same symbol. For example, the constituent "*January 30th*" has character type "Xxxxxxx 99xx" and pattern "X(x) (9)(x)".

On this basis, we employ a classifier that implements a Maximum Entropy model[2] and performs categorization of constituent-phrases extracted from the input.

4 Experiments, Results and Error Analysis

After processing the TimeBank corpus of 183 documents we had 2612 parsed sentences with 1224 temporal expressions in them. 2612 sentences resulted in 49656 phrase candidates. We separated the data in order to perform 10-fold cross validation, train the classifier and test it on an unseen dataset. The evaluations were conducted with respect to the TERN 2004 evaluation plan (TERN, 2004) and described in Section 2.1.

After running experiments the classifier demonstrated the performance in detection of TIMEX3 tags with a minimal overlap of one character with precision, recall and F1-measure at 0.872, 0.836 and 0.852 respectively. Since the candidate phrases provided by the parser do not always exactly align annotated temporal expressions, the results for the exact match experiments are constrained by an estimated upper-bound recall of 0.919. The experiments on exact match demonstrated a small decline of performance level and received scores of 0.866, 0.796 and 0.828 for precision, recall and F1-measure respectively.

Putting the received figures in context, we can say that with a very few shallow features and a standard machine learning algorithm the recognizer of temporal expressions performed at a comparable operational level to the rule-based approach of (Boguraev and Ando, 2005) and outperformed it in exact match. A comparative performance summary is presented in Table 2.

Sometimes it is very hard even for humans to identify the use of obvious temporal triggers in a specific context. As a result, many occurrences of such triggers remained unannotated for which TIMEX3 identification could not be properly carried out. Apart of obvious incorrect parses, inexact alignment between temporal expressions and candidate phrases was caused by annotations that occurred at the middle of a phrase, for example "*eight-years-long*", "*overnight*", "*yesterday's*". In total there are 99 TIMEX3 tags (or 8.1%) misaligned with the parser output, which resulted in 53 (or 4.3%) undetected TIMEX3s.

[1] In literature such patterns are also known as shorttypes.

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

	P	*R*	*F1*
Detection			
Our approach	0.872	0.836	0.852
Sloopy Span			
(Boguraev and Ando, 2005)	0.852	0.952	0.896
Exact Match			
Our approach	0.866	0.796	0.828
(Boguraev and Ando, 2005)	0.776	0.861	0.817

Table 2. Comparative Performance Summary

Definite and indefinite articles are unsystematically left out or included into TIMEX3 extent, which may introduce an additional bias in classification.

5 Conclusion and Future Work

In this paper we presented a machine learning approach for detecting temporal expression using a recent annotated corpus for temporal information, TimeBank. Employing shallow syntactic and lexical features, the performance level of the method achieved comparable results to a rule-based approach of Boguraev and Ando (2005) and for the exact match task even outperforms it. Although a direct comparison with other state-of-the-art systems is not possible, due to different evaluation corpora, annotation standards and size in particular, our experiments disclose a very important characteristic. While the recognition systems in the TERN 2004 reported a substantial drop of F1-measure between detection and exact match results (6.5 – 11.6%), our phrase-based detector demonstrates a light decrease in F1-measure (2.4%), whereas the precision declines only by 0.6%. This important finding leads us to the conclusion that most of TIMEX3s in TimeBank can be detected at a phrase-based level with a reasonably high performance.

Despite a good recognition performance level there is, of course, room for improvement. Many implementations in the TERN 2004 employ a set of apparent temporal tokens as one of the features. In our implementation, the classifier has difficulties with very simple temporal expressions such as "*now*", "*future*", "*current*", "*currently*", "*recent*", "*recently*". A direct employment of vocabularies with temporal tokens may substantially increase the F1-measure of the method, however, it yet has to be proven. As reported in (Ahn et al., 2007) a precise recognition of temporal expressions is a prerequisite for accurate normalization.

With our detector and a future normalizer we are able make the first step towards solving the TempEval-2 tasks, which introduce new challenges in temporal information processing: identification of events, identification of temporal expressions and identification of temporal relations (Pustejovsky et al., 2009). Our future work will be focused on improving current results by a new feature design, finalizing the normalization task and identification of temporal relations. All these components will result in a solid system infrastructure for all-around temporal analysis.

Acknowledgments

This work has been partly funded by the Flemish government (through IWT) and by Space Applications Services NV as part of the ITEA2 project LINDO (ITEA2-06011).

References

Ahn, D., Adafre, S. F., and de Rijke, M. 2005. Extracting Temporal Information from Open Domain Text: A Comparative Exploration. *Digital Information Management*, 3(1):14-20, 2005.

Ahn, D., van Rantwijk, J., and de Rijke, M. 2007. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In *Proceedings NAACL-HLT 2007*.

Boguraev, B., and Ando, R. K. 2005. TimeBank-Driven TimeML Analysis. In *Annotating, Extracting and Reasoning about Time and Events*. Dagstuhl Seminar Proceedings. Dagstuhl, Germany

Boguraev, B., Pustejovsky, J., Ando, R., and Verhagen, M. 2007. TimeBank Evolution as a Community Resource for TimeML Parsing. *Language Resource and Evaluation*, 41(1): 91–115.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. 2003. TIDES 2003 Standard for the Annotation of Temporal Expressions. Sept. 2003. timex2.mitre.org.

Ferro, L. 2004. TERN Evaluation Task Overview and Corpus, <http://fofoca.mitre.org/tern_2004/ferro1_TERN2004_task_corpus.pdf> (accessed: 5.03.2009)

Graff, D., Cieri, C., Strassel, S., and Martey, N. 1999. The TDT-2 Text and Speech Corpus. In *Proceedings of DARPA Broadcast News Workshop*, pp. 57-60.

Hacioglu, K., Chen, Y., and Douglas, B. 2005. Automatic Time Expression Labeling for English and Chinese Text. In *Proceedings of CICLing-2005*, pp. 348-359; Springer-Verlag, Lecture Notes in Computer Science, vol. 3406.

Mani, I. and Wilson, G. 2000. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (Hong Kong, October 03 - 06, 2000). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, pp. 69-76.

Negri, M. and Marseglia, L. 2004. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Technical Report, ITC-irst, Trento.

Poveda, J., Surdeanu, M., and Turmo, J. 2007. A Comparison of Statistical and Rule-Induction Learners for Automatic Tagging of Time Expressions in English. In *Proceedings of the International Symposium on Temporal Representation and Reasoning*, pp. 141-149.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Day, D., Ferro, L., Gaizauskas, R., Lazo, M., Setzer, A., and Sundheim, B. 2003a. The TimeBank Corpus. In *Proceedings of Corpus Linguistics 2003*, pp. 647-656.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. 2003b. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*.

Pustejovsky, J., Verhagen, M., Nianwen, X., Gaizauskas, R., Hepple, M., Schilder, F., Katz, G., Saurí, R., Saquete, E., Caselli, T., Calzolari, N., Lee, K., and Im, S. 2009. TempEval2: Evaluating Events, Time Expressions and Temporal Relations. <http://www.timeml.org/tempeval2/tempeval2-proposal.pdf> (accessed: 5.03.2009)

Ratnaparkhi, A. 1999. Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*, 34(1): 151-175.

TERN 2004 Evaluation Plan, 2004, <http://fofoca.mitre.org/tern_2004/tern_evalplan-2004.29apr04.pdf> (accessed: 5.03.2009)