KUL: A Data-driven Approach to Temporal Parsing of Documents

Oleksandr Kolomiyets
KU Leuven
Celestijnenlaan 200A
Heverlee 3001, Belgium
Department of Computer Science
oleksandr.kolomiyets
@cs.kuleuven.be

Marie-Francine Moens
KU Leuven
Celestijnenlaan 200A
Heverlee 3001, Belgium
Department of Computer Science
sien.moens@cs.kuleuven.be

Abstract

This paper describes a system for temporal processing of text, which participated in the Temporal Evaluations 2013 campaign. The system employs a number of machine learning classifiers to perform the core tasks of: identification of time expressions and events, recognition of their attributes, and estimation of temporal links between recognized events and times. The central feature of the proposed system is temporal parsing – an approach which identifies temporal relation arguments (event-event and event-timex pairs) and *the semantic label* of the relation as a single decision.

1 Introduction

Temporal Evaluations 2013 (TempEval-3) is the third iteration of temporal evaluations (after TempEval-1 (Verhagen et al., 2007) and TempEval-2 (Verhagen et al., 2010)) which addresses the task of temporal information processing of text. In contrast to the previous evaluation campaigns where the temporal relation recognition task was simplified by restricting grammatical context (events in adjacent sentences, events and times in the same sentences) and proposed relation pairs, TempEval-3 does not set any context in which temporal relations have to be identified. Thus, for temporal relation recognition the challenges consist of: first, detecting a pair of events, or an event and a time that constitutes a temporal relation; and, second, determining what semantic label to assign to the proposed pair. Moreover, TempEval-3 proposes the task of **end-to-end** temporal processing in which events and times, their attributes and relations have to be identified from a raw text input.

In this paper we present a data-driven approach to all-around temporal processing of text. A number of machine-learning detectors were designed to recognize temporal "markables" (events and times) and their attributes. The key feature of our approach is that argument pairs, as well as relations between them, are jointly estimated without specifying in advance the context in which these pairs have to occur.

2 Our Approach

2.1 Timex Processing

2.1.1 Timex Recognition and Normalization

The proposed method for timex recognition implements a supervised machine learning approach that processes each chunk-phrase derived from the parse tree. Time expressions are detected by the model as phrasal chunks in the parse with their corresponding spans. In addition, the model is bootstrapped by substitutions of temporal triggers with their synonyms learned by the Latent Words Language Model (Deschacht et al., 2012) as described in (Kolomiyets et al., 2011). We implemented a logistic regression model that makes use of the following features:

- the head word of the phrase and its POS tag;
- all tokens and POS tags in the phrase as a bag of words;
- the word-shape representation of the head word and the entire phrase, e.g. `Xxxxx 99` for the expression *April 30*;

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 83–87, Atlanta, Georgia, June 14-15, 2013.

- the condensed word-shape representation for the head word and the entire phrase, e.g. `X(x)(9)` for the expression *April 30*;
- the concatenated string of the syntactic types of the children of the phrase in the parse tree;
- the depth in the parse tree.

In addition, we considered a special label for single tokens of time expressions. In this way, we detect parts of temporal expressions if they cannot be found in the chunk-based fashion. In detail, if a token is recognized as part of a timex and satisfies the pre-condition on its POS tag, we employ a "look-behind" rule for the phrasal chunk to match the begin token of the temporal expression. The legitimate start POS tags are determiners, adjectives, and cardinals. Another set of rules specifies unsuitable timexes, such as single cardinals with values outside predefined ranges of *day-of-month*, *month-of-year* and *year* numbers.

Normalization of temporal expressions is a process of estimating standard temporal values and types for temporal expressions. Due to a large variance of expressions denoting the same date and vagueness in language, rule-based approaches are usually employed for the normalization task, and our implementation is a rule-based system. The normalization procedure is the same as described in (Kolomiyets and Moens, 2010), which participated in TempEval-2.

2.2 Event Processing

The proposed method to event recognition implements a supervised machine learning approach that classifies every single token in the input sentence as an event instance of a specific semantic type. We implemented a logistic regression model with features largely derived from the work of Bethard and Martin (2006):

- the token, its lemma, coarse and fine-grained POS tags, token's suffixes and affixes;
- token's hypernyms and derivations in WordNet;
- the grammatical class of the chunk, in which the token occurs;
- the lemma of the governing verb of the token;
- phrasal chunks in the contextual window;
- the *light* verb feature for the governing verb;
- the polarity of the token's context;
- the determiner of the token and the sentence's subject;

In addition, we classify the tense attribute for the detected event by applying a set of thirteen handcrafted rules.

2.3 Temporal Relation Processing

Temporal relation recognition is the most difficult task of temporal information processing, as it requires recognitions of argument pairs, and subsequent classifications of relation types. Our approach employs a shift-reduce parsing technique, which treats each document as a dependency structure of annotations labeled with temporal relations (Kolomiyets et al., 2012). On the one hand, the advantage of the model is that the relation arguments and the relation between them are extracted as a single decision of a statistical classification model. On the other hand, such a decision is local and might not lead to the optimal global solution[1]. The following features for deterministic shift-reduce temporal parsing are employed:

- the token, its lemma, suffixes, coarse and fine-grained POS tags;
- the governing verb, its POS tag and suffixes;
- the sentence's root verb, its lemma and POS tag;
- features for a prepositional phrase occurrence, and domination by an auxiliary or modal verb;
- features for the presence of a temporal signal in the chunk and co-occurrence in the same sentence;
- a feature indicating if the sentence root verb lemmas of the arguments are the same;
- the temporal relation between the argument and the document creation time (DCT) (see below);
- a feature indicating if one argument is labeled as a semantic role of the other;
- timex value generation pattern (e.g. `YYYY-MM` for `2013-02`, or `PXY` for `P5Y`) and timex granularity (e.g. `DAY-OF-MONTH` for *Friday*, `MONTH-OF-YEAR` for *February* etc.);

[1]For further details on the deterministic temporal parsing model we refer the reader to (Kolomiyets et al., 2012).

Training	Test	P	R	F_1
TimeBank	TimeBank 10-fold	0.907	0.99	0.947
	AQUAINT	0.755	0.972	0.850
	Silver	0.736	0.963	0.834
AQUAINT	TimeBank	0.918	0.986	0.951
	AQUAINT 10-fold	0.795	0.970	0.874
	Silver	0.746	0.959	0.851
Silver	TimeBank	0.941	0.976	0.958
	AQUAINT	0.822	0.955	0.883
	Silver 10-fold	0.798	0.944	0.865

Table 1: Results for timex detection in different corpora.

As one of the features above provides information about the temporal relation between the argument and the DCT, we employ an interval-based algebra to classify relations between timexes and the DCT. In case the argument is an event, we use a simple logistic regression classifier with the following features:

- the event token, its lemma, coarse and fine-grained POS tags;
- tense, polarity, modality and aspect attributes;
- the token's suffixes;
- the governing verb, its POS tag, tense and the grammatical class of the chunk, in which the event occurs;
- preceding tokens of the chunk;

3 Results

3.1 Pre-Evaluation Results

The following results are obtained by 10-fold cross-validations and corpus cross-validations with respect to the evaluation criteria and metrics used in TempEval-2. Tables 1 and 2 present the results for the timex recognition and normalization tasks (Task A), and, Tables 3 and 4 present the results for the event recognition task (Task B).

As can be seen from the pre-evaluation results, the most accurate classification of timexes on all corpora in terms of F_1 score is achieved for the model trained on the Silver corpus. As for timex normalization, the performances on TimeBank and the Silver corpus are not very different for type and value accuracies. Similarly, we observe the tendency for a better performance on larger datasets with an exception for 10-fold cross-validation using the AQUAINT corpus.

Test Corpus	Type Acc.	Value Acc.
TimeBank	0.847	0.742
AQUAINT	0.852	0.714
Silver	0.853	0.739

Table 2: Results for normalization in different corpora.

Training	Test	P	R	F_1
TimeBank	TimeBank 10-fold	0.82	0.641	0.72
	AQUAINT	0.864	0.649	0.741
	Silver	0.888	0.734	0.804
AQUAINT	TimeBank	0.766	0.575	0.657
	AQUAINT 10-fold	0.900	0.776	0.836
	Silver	0.869	0.755	0.808
Silver	TimeBank	0.827	0.717	0.768
	AQUAINT	0.906	0.807	0.854
	Silver 10-fold	0.916	0.888	0.902

Table 3: Results for event detection in different corpora.

Training	Test	Class Acc.
TimeBank	TimeBank 10-fold	0.691
	AQUAINT	0.717
	Silver	0.804
AQUAINT	TimeBank	0.620
	AQUAINT 10-fold	0.830
	Silver	0.794
Silver	TimeBank	0.724
	AQUAINT	0.829
	Silver 10-fold	0.900

Table 4: Results for event classification in different corpora.

3.2 Evaluation Results

For the official evaluations we submitted three runs of the system, one of which addresses Tasks A and B (timex and event recognition)[2], one (KUL-

[2]During the official evaluation period, this run was resubmitted with no changes in the output together with KULTE3RunABC, which led to duplicate evaluation results known

Run	Relaxed Evaluation			
	P	R	F_1	Rank
KULRun-1	0.929	0.769	0.836	21/23
KUL-TE3RunABC	0.921	0.754	0.829	22/23
Run	Strict Evaluation			
	P	R	F_1	Rank
KULRun-1	0.77	0.63	0.693	22/23
KUL-TE3RunABC	0.814	0.667	0.733	15/23

Table 5: Results for the timex detection task.

TE3RunABC) provides a full temporal information processing pipeline (Task ABC), and the one for Task C only (KUL-TaskC). For KULRun-1 we employed the recognition models described above, all trained on the aggregated corpus comprising all three available training corpora in the evaluations. For KUL-TE3RunABC we also trained the markable recognition models on the aggregated corpus, but the event recognition output was slightly changed in order to merge multiple consequent events of the same semantic class into a single multi-token event. The temporal dependency parsing model was trained on the TimeBank and AQUAINT corpora only, with a reduced set of relation labels. This decision was motivated by the time constraints and the training time needed. The final relation label set contains the following temporal relation labels: `BEFORE`, `AFTER`, `DURING`, `DURING_INV`, `INCLUDES` and `IS_INCLUDED`. Below we present the obtained results for each task separately. The results for Task A are presented in Tables 5 and 6, for Task B in Tables 7 and 8, and, for Task ABC and Task-C-only in Table 9. It is worth mentioning that for Task B the aspect value was provided as `NONE`, thus this evaluation criterion is not representative for our system.

4 Conclusion

For TempEval-3 we proposed a number of statistical and rule-based approaches. For Task A we employed a logistic regression classifier whose output was augmented by a small number of hand-crafted rules to increase the recall. For the temporal ex-

as KULRun-1 and KULRun-2. Further in the paper, we refer to this run as simply to KULRun-1.

Run			Rank
KULRun-1	F_1		18/23
	Value	Type	
	0.629	0.741	
	Accuracy		14/23
	Value	Type	
	0.752	0.886	
KUL-TE3RunABC	F_1		19/23
	Value	Type	
	0.621	0.733	
	Accuracy		15/23
	Value	Type	
	0.750	0.885	

Table 6: Results for the timex normalization task.

Run	P	R	F_1	Rank
KULRun-1	0.807	0.779	0.792	5/15
KUL-TE3RunABC	0.776	0.765	0.77	12/15

Table 7: Results for the event detection task.

Run				Rank
KULRun-1	F_1			3/15
	Class	Tense	Aspect	
	0.701	n.a.	n.a.	
	Accuracy			3/15
	Class	Tense	Aspect	
	0.884	n.a.	n.a.	
KUL-TE3RunABC	F_1			5/15
	Class	Tense	Aspect	
	0.687	0.497	0.632	
	Accuracy			1/15
	Class	Tense	Aspect	
	0.891	0.644	0.82	

Table 8: Results for the event attribute recognition task.

Run	P	R	F_1	Rank
KUL-TE3RunABC	0.18	0.202	0.191	8/8
KUL-TaskC	0.234	0.265	0.248	10/13

Table 9: Results for Tasks ABC (end-to-end processing) and C (gold entities are given).

pression normalization subtask we employed a rule-based system which estimates the attribute values for the recognized timexes. For Task B we proposed a logistic regression classifier which processes input tokens and classifies them as event instances of particular semantic classes. The optional tense attribute was estimated by a number of manually designed rules. For the most difficult tasks, Task ABC and Task C, we proposed a dependency parsing technique that jointly learns from data what arguments constitute a temporal relation and what the temporal relation label is. Due to evaluation time constraints and the time needed to model training, we reduced the set of relation labels and trained the model on two small annotated corpora.

The evaluations evidenced that the use of larger annotated data sets did not improve the timex recognition performance as it was expected from the pre-evaluations. Interestingly, we did not observe the expected improvement in terms of recall, as it was the case in the pre-evaluations. Yet, the timex normalization performance levels in the official evaluations were slightly higher than in the pre-evaluations. In contrast to timex recognition, the use of a large annotated corpus improved the results for event recognition. The pilot implementation of a temporal parser for newswire articles showed the lowest performance in the evaluations for Task ABC, but still provided decent results for Task C. One of the advantages of the proposed temporal parser is that the parser selects arguments for a temporal relation and classifies it at the same time. The decision is drawn by a statistical model trained on the annotated data, that is, the parser does not consider any particular predefined grammatical context in which the relation arguments have to be found. Another weak point of the parser is that it requires a large volume of high-quality annotations and long training times. The last two facts made it impossible to fully evaluate the proposed temporal parsing model, and we will further investigate the effectiveness of the model.

Acknowledgments

The presented research was supporter by the TERENCE (EU FP7-257410) and MUSE (EU FP7-296703) projects.

References

Steven Bethard and James H Martin. 2006. Identification of Event Mentions and their Semantic Class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154. Association for Computational Linguistics.

Koen Deschacht, Jan De Belder, and Marie-Francine Moens. 2012. The Latent Words Language Model. *Computer Speech & Language*.

Oleksandr Kolomiyets and Marie-Francine Moens. 2010. Kul: Recognition and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 325–328. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-Portability Experiments for Textual Temporal Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting Narrative Timelines as Temporal Dependency Structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, 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. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, 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. Association for Computational Linguistics.