Systems' Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task

Tommaso Caselli∗, Roser Morante
∗University of Groningen, VU Amsterdam
The Netherlads
∗t.caselli@gmail.com, r.morantevallejo@vu.nl

Abstract

In this article we review Temporal Processing systems that participated in the TempEval-3 task as a basis to develop our own system, that we also present and release. The system incorporates high level lexical semantic features, obtaining the best scores for event detection (F1-Class 72.24) and second best result for temporal relation classification from raw text (F1 29.69) when evaluated on the TempEval-3 data. Additionally, we analyse the errors of all TempEval-3 systems for which the output is publicly available with the purpose of finding out what are the weaknesses of current approaches. Although incorporating lexical semantics features increases the performance of our system, the error analysis shows that systems should incorporate inference mechanisms and world knowledge, as well as having strategies to compensate for data skewness.

Keywords: temporal processing, error analysis, written corpora

1. Introduction

Any discourse, spoken or written, contains temporally connected linguistic mentions, such as events and temporal expressions (i.e. timexes). Relations between these mentions can be meaningfully interpreted by using models of time, which allow to connect events on a timeline (i.e. temporal anchoring) and to understand complex sequences of events (i.e. temporal ordering). Temporal relations (TRs) provide such a model, and a set of properties, to account for the connections between pairs of entities.

Temporal Processing (TP) is a task consisting in automatically identifying and classifying basic entities and their relations, such as event-event (**e-e**), and event-timex (**e-t**). Temporally aware Natural Language Processing (NLP) systems are crucial not only to generate timelines and storylines (Vossen et al., 2015), but also in decision support systems, summarization and textual entailment applications, question answering systems, and document archiving, among others. Since the availability of the TimeBank corpus (Pustejovsky et al., 2003) there has been a renewed interest in the area of TP, which has resulted in the celebration of several evaluation campaigns[1] and in the creation of corpora and tools in languages other than English. [2]

This paper focuses on analysing errors of TP systems that have participated in the TempEval-3 competition (UzZaman et al., 2013) in order to find out what are the main limitations of the systems. Our study builds on the work by Derczynski (2013), who proposes a classification of TR errors as a result of analysing the output of systems participating in the TempEval-2 competition . However, the focus of our analysis is different because the data of TempEval-2 is a subset of the TimeBank corpus, participating systems produced a simplified set of TRs, and were not evaluated on an end-to-end approach.

Two are the contributions of this paper: first, we review state-of-the-art TP systems to identify their properties (i.e. features and learning algorithm), common characteristics, and limitations. Based on that we have developed our own system, which we release to the public [3] [4]. Secondly, we have conducted an extensive error analysis by comparing the output of different systems, including our own, to provide a better understanding of the limitations and issues that still need to be addressed in this task.

The remainder of the paper is structured as follows: Section 2. explains the TP task in general and as formulated in the TempEval-3 evaluation exercise. Section 3. reviews TP systems that participated in the TempEval-3 competition whose output is publicly available. The results of the error analysis are presented in Section 4. and Section 5. for event trigger detection and temporal relation classification, respectively. Finally, Section 6. puts forward conclusions and future work.

2. Task Description

TP is a concatenation of 4 subtasks: identification and classification of linguistic mentions that denote events (ES); detection and normalization of timexes (TE); identification of **e-e** and **e-t** pairs (TD); and classification of valid temporal relations according to a predefined set of values (TC).

TempEval-3 is a follow-up of two previous evaluation exercises (TempEval and TempEval-2), with the difference that the task of TP is evaluated from an end-to-end perspective, i.e. systems should produce full temporally annotated documents starting from raw text. The TempEval-3 datasets are compliant with the TimeML Annotation Guidelines. In particular, an event is defined as any linguistic mention, including verbs, nouns, adjectives and preposi-

[1] TempEval (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), Clinical TempEval (Bethard et al., 2016; Bethard et al., 2017), Q-A TempEval (Llorens et al., 2015).

[2] For an extended list of available TimeBanks see (Caselli and Sprugnoli, 2017).

[3] `https://github.com/cltl/TimeMLEventTrigger`

[4] At the moment of writing of this abstract, we are still completing the linked repository.

tional phrases, which denotes something that happens, occurs, or describes states/circumstances in which something obtains or holds true. Each event mention is further characterized by a set of 5 attributes such as: class, tense, aspect, polarity, and modality. Timexes are defined as lexical items which denote a time, a date, a duration, or a set (e.g. *noon*, *yesterday*, *two days ago*, *yearly*), extending previous annotation initiatives such as TIDES (Ferro et al., 2002) and STAG (Setzer, 2001). Finally, the set of possible TRs is based on Allen's temporal intervals consisting of a total of 14 possible values: `BEFORE`, `AFTER`, `INCLUDES`, `IS_INCLUDED`, `BEGINS`, `ENDS`, `BEGUN_BY`, `ENDED_BY`, `SIMULTANEOUS`, `IAFTER`, `IBEFORE`, `DURING`, `DURING_INV`, `IDENTITY`. Additionally, for the TempEval-3 evaluation extra training data were provided, automatically annotated for event, timexes, and TRs; a new evaluation metric was used to assess the *temporal awareness* of end-to-end systems; and a new test set was used.

3. Temporal Processing Systems

In order to develop our own out-of-competition TP system, we analyzed first the best systems from TempEval-3 that targeted either the event extraction and classification subtask only (Task B in the TempEval-3 guidelines) or the end-to-end temporal relation identification and classification subtask (Task C in the TempEval-3 guidelines, which includes Task B as well). This results in a total of 6 unique systems (5 for event detection and classification only and 4 for the full TP).

The event detection and classification task was addressed by all systems using supervised discrete machine learning classifiers such as Conditional Random Fields (CRFs) (Kolya et al., 2013; Bethard, 2013), Logistic Regression (Kolomiyets and Moens, 2013), and Maximum Entropy (Chambers, 2013; Jung and Stent, 2013). Most of the systems (4 out of 5) adopted the same learning model also for event classification. Overall, 17 features were represented in the learning models, which can be aggregated in 5 groups:

- Basic morpho-syntactic features (e.g. token, lemma, stem, parts-of-speech (POSs), token's affix and/or suffix, among others);
- Syntactic features (e.g. constituency/dependency parsing; governing verb lemma, verb chunks);
- Contextual features (e.g. context windows of token, lemma, POS; and tokens polarity, among others);
- Semantic features, limited to semantic roles;
- Lexical semantic features, limited to WordNet synsets and hypernyms.

For the event detection task the learning models were more complex in terms of features used than for the classification task, where a lower number of features is selected. Semantics and lexical semantics features were used by less systems (2 systems for event detection and only 1 system for classification).

The temporal relation detection and classification task was addressed as a supervised multi-class classification task. Systems used either a single classifier (Maximum Entropy (Chambers, 2013); CRFs (Kolya et al., 2013)) or a combination of two classifiers (SVM and Logistic Regression (Kolomiyets and Moens, 2013); SVM and Maximum Entropy (Bethard, 2013)).

3 of 4 systems solve the task in a two-step approach: recognition of eligible temporal relations and assignment of the temporal values. Only 1 system (Bethard, 2013) uses a single step approach, introducing the value NORELATION for negative examples. All systems incorporate different classifiers for different subsets of relations (**e-e**, **e-t**, and event-document creation time (DCT) pairs (**e-dct**)). Only 2 systems (Chambers, 2013; Kolya et al., 2013) incorporate classifiers for intra- and inter-sentence relations, while the others deal only with intra-sentence relations. Finally, 2 systems (Bethard, 2013; Kolomiyets and Moens, 2013) use a reduced set of temporal values, while the others adopted the full 14 temporal values.

The feature set for classification of TRs is larger than for event detection and classification, up to 29 features per system, and scattered. There are specific features for some sub-types of TRs (e.g. syntactic path between **e-t** pairs, timex tokens, and linear order in the text, among others). Most of the features fall into the same categories of the event detection and classification task, although some extra features are used: tense and aspect values, order of presentation of the events, presence of temporal prepositions/adverbs, and type of timexes, which are grounded in linguistic theories of time (Reichenbach, 1947; Comrie, 1985; Declerck, 1986). Features which account for discourse structure and world knowledge are either missing or simplified (e.g. only WordNet synsets).

3.1. A New TP System

Based on our study of participating systems, we developed a new end-to-end TP system. Similarly to previous works, we used a single learner and we split the task in multiple subtasks. The system is based on a cascade of 7 CRF classifiers. It shares with existing systems the use of morpho-syntactic features, contextual features ([+/-2] context window, which has proven to be optimal), and semantic features such as semantic roles and WordNet synsets. To extract the features the data are processed with state-of-the art tools, such as the Stanford CoreNLP (Manning et al., 2014) and the NewsReader NLP pipeline (Agerri et al., 2014). Additionally, the system uses lexical semantics features such as VerbNet classes and FrameNet frames computed from the alignments in the Predicate Matrix (Lacalle et al., 2014). This allow to access high level lexical semantic features which have a role in the identification of event mentions and TRs.

We used the TempEval-2 test data as a development set, to tune the features and conduct an ablation study to verify the impact of the extended lexical semantics features. We observed two things: i) the quality of the pre-processing tools has an impact on the final results; ii) the semantic and syntactic features have the biggest impact on the system's performance: using only morpho-syntactic and context window features gives an F1 of 82.1 for event detection, which increases to 88.2 when adding lexical semantics features

only, and reaches 90.9 when lexical semantic are combined with syntactic information.

As for the TR task, to overcome the lack of connectivity between all possible **e-e** and **e-t** pairs, we assumed that in the test data all possible pairs of entities are temporally connected. Such a decision is inspired by the solution adopted in the TimeBank Dense corpus (Cassidy et al., 2014). This model has been developed to better evaluate the completeness of the test data, by identifying pairs which are correct but not annotated. Furthermore, we have used all 14 temporal values, rather than simplifying the set to the most frequent ones in the training data [5].

Table 1 contains the results of our TP system and of the reviewed systems.

System	F1	P	R
Our system	29.69	23.86	**39.29**
(Bethard, 2013)	**30.98**	**34.08**	28.40
(Chambers, 2013)	27.28	31.25	24.20
(Kolya et al., 2013)	24.61	19.17	34.36
(Kolomiyets and Moens, 2013)	19.01	17.94	20.22

Table 1: Results for TempEval-3 Task C (Temporal Processing from raw text).

Our system qualifies as the second best system on this task (F1 29.69). The higher recall reflects two aspects: firstly, a good performance of the system in detecting the basic entities, especially event mentions (F1 80.3), and, secondly, the assumption of a full temporal connection of the entities, which tends to over-generate TRs. Breaking down these results per type of entity pairs in a TR, our system has the best F1 for **e-e** pairs (25.69), while the best score for the other participating system is obtained by Chambers (2013) (F1 19.01). Things are different when looking at the results for **e-t** and **e-dct** pairs. In both cases Bethard (2013) obtains the best scores, with an F1 of 41.41 for **e-t** and 24.75 for **e-dct**. Our system, on the other hand, scores only 27.59 F1 for **e-t**, and a competitive 23.48 for **e-dct**.

4. Event Triggers: What is it wrong?

We analyzed the errors made by all systems presented in the previous section for the event detection and classification subtask. As for event detection, of the 749 gold events, 64% are correctly detected by all systems, 10% by 5, 3.6% by 4, 4% by 3, 4.4% by 2, 4.4% by 1, and 9.4% by 0 systems. From the events that all systems correctly detect 91.87% are verbs, 7.08% nouns, 0.83% adjectives, and 0.20% other. From the events that no system correctly detects 67.60% are nouns, 18.30% are adjectives, 11.26% other, and 2.81% prepositions. This indicates that systems are well trained to detect events expressed by prototypical POS (i.e. verbs), while events with less prototypical POS are more challenging. This is coherent with the statistics of the training data, where 80.5% of events are verbs. By looking at cases in which systems disagree, we find that more systems agree for events with POS verb, than for events with POS noun. In this sense we are confronted with a very standard characteristic of NLP gold data sets, namely class imbalance. As in many other NLP tasks, a good system will have to be able to deal with the sparse examples that belong to the long tail of data distribution. From the events with POS noun that no system detects, 3 are proper nouns. We found that these are cases of metonymy, which would require a system to apply inference mechanisms, as in the example 1 where *Everest* is a proper noun that refers to the event 'climbing the Everest'. No system was able to make this inference.

1. He said: "[...] 60 years on from *Everest* his achievements deserve wider recognition."

Additionally, for all events that no system detects correctly we checked if they occur in the training corpus. We found that out of 71, 4 occur less than 5 times and 2 around 40 times, but with a different POS (verb, adjective and noun). The rest do not occur in the training corpus. This suggests that unseen events are difficult to detect using discrete models and features because systems can not generalize enough.

As for event classification, 43.95% of the examples are correctly classified by all systems, 22.12% by 5, 7.96% by 4, 5.75% by 3, 5.16% by 2, 6.19% by 1, and in 8.84% of the cases no system finds the right solution. The events that all systems correctly classify belong mostly to the classes OCCURRENCE (74.16%) and REPORTING (21.81%), which are the most frequent classes in the training set (61.71% and 14.36%). The distribution of classes where all systems fail is as follows: STATE (43.33%), ASPECTUAL (23.33%), I STATE (10%), OCCURRENCE (8.33%), I ACTION (6.66%), REPORTING (5%), and PERCEPTION (3.33%). This indicates again that the most difficult cases belong to low-represented classes in the training data.

5. Temporal Relations: When is it wrong?

For the error analysis of the temporal relation subtask we look at three aspects: i) what type of relations are incorrectly classified by all systems; ii) what type of processing requirements are needed to solve cases where all systems fail; and, iii) to which extent False Positives (FP) identified by our own system are correct.

As for the incorrectly classified relations, we observe that all systems commit errors for **e-t** and **e-e** relations. This indicates that anchoring to the right timex (**e-t**) and ordering relations (**e-t**) are both complex and difficult.

With the exception of one system (Bethard, 2013), the temporal values that the systems output are as skewed as in the training data, since systems tend to predict the most frequent classes. In particular, systems tend to predict the values BEFORE, AFTER, and SIMULTANEOUS for **e-e** pairs, IS INCLUDED for **e-t** pairs, and BEFORE, INCLUDES, IS INCLUDED, and AFTER for **e-dct** pairs.

The errors committed by the systems can be grouped into the 6 categories listed below.[6] Excluding *Error*, the categories refer to the processing requirement that the system should have fulfilled in order to correctly identify and classify a TR.

- **Iconicity**: the system should interpret the linear order of presentation of the entities;

[5] The final version of the paper will provide more details about the system.

[6] 4 classes: Iconicity, Inference, Signaled and World Knowledge have been proposed in Derczynski (2013).

- **Signaled**: the system needs to process an explicit temporal signal (e.g. *before*, *since*, and similar) that connects the elements in the pair;
- **Inference**: the relation can be identified and classified through inference via other existing relations (e.g. two events linked to two different timexes can be order by means of the comparison of the values of timexes only);
- **Grammar**: the system needs to infer the relation via grammatical information (e.g. tense and aspect values), and/or syntactic dependencies between the elements in a pair;
- **World Knowledge**: the relation can be classified by applying knowledge about event semantics, discourse structure, factuality profiling of the events, and external information concerning commonsense knowledge;
- **Error**: the gold temporal relation is either wrong or in dispute.

We summarize our findings in Table 2. The error analysis is based on all cases of **e-t** and **e-dct** pairs and is limited to 50% of the **e-e** pairs.

Processing requirement	**[e-dct]**	**[e-t]**	**[e-e]**
Iconicity	0 (0%)	0 (0%)	4 (4.49%)
Signaled	0 (0%)	6 (31.57%)	9 (10.11%)
Inference	6 (30%)	5 (26.31%)	22 (24.71%)
Grammar	3 (15%)	2 (10.52%)	21 (23.59%)
World Knowledge	10 (50%)	6 (31.57%)	32 (35.95%)
Error	1 (5%)	0 (0%)	1 (1.12%)

Table 2: TLINK error classification of the cases where all systems fail.

With the exception of the **e-dct** pairs, world knowledge plays a less prominent role than expected. The classification of both **e-t** and **e-e** pairs would improve by training the systems with more densely annotated data such as TimeBank Dense. At the same time, the results would improve if inference mechanisms were applied to keep track of the TRs in which events are involved: around 25% of errors in these relations could be avoided by means of inferences.
Grammatical information plays an important role for **e-e** relations. A subset of errors in the Grammar class could be avoided by computing the contextual values of the tense and aspect rather than using the values of the their superficial form (e.g. a superficial present tense which is used to described past events should have value PAST for tense rather then PRESENT). These cases are hard to solve because in TimeML only the superficial values of tense and aspect are annotated. We also looked at whether it is easier to identify and classify **e-e** pairs that have different tense and aspect values. We observe that this is not the case as the error rate in **e-e** pairs with the same tense and aspect values (40%) does not differ from the error rate of pairs with different values (39%).
Additional errors are found in relations between events across sentences. Overall, they represent 26.96% of the analyzed data. Interestingly, the correct processing of intersentential **e-e** pairs cannot be related to a particular processing requirement. Another cause of errors is unproper processing of pairs where one of the elements is a reporting verb. These cases require careful processing of the previous linguistic context in order to identify the correct relation between the events.

What concerns the errors of our system, we have measured the impact of the semantic features by removing the lexical semantic features. On the test data, the scores drop 3.63 points of temporal awareness. [7] As for the False Positives (FP), our system produces 555 FP for **e-dct** relations; 301 for **e-t**, and 571 for **e-e**. We manually checked 15% of the test files to establish if the temporal links predicted by the system are correct. Out of 81 **e-dct** links, 74 of them are valid, which results in 48 (64.86%) links correctly classified. The same applies to the **e-e** pairs, where out of 54 system output links, 44 are valid with 26 (59.09%) correctly classified. As for **e-t** pairs, we have identified 40 possible links, with 33 valid links. Contrary to the other cases, only 9 (27.27%) **e-t** links are correct due to an over-generation of the IS_INCLUDED value.

This preliminary error analysis [8] has shown that wrong results for TRs are mainly dependent on: i.) inference, ii.) contextual interpretation of grammatical devices which encode TRs (i.e. tense and aspect), and iii.) lack of world knowledge. Full connectivity among entity pairs and more data can benefit the task, but they will not solve all the issues identified.

6. Conclusions and Future Work

In this paper we have focused on TP in the framework of TempEval-3. We have reviewed the 5 top performing systems to gain insights into their architectures and features. We found that no system has used rich lexical semantic information as a means to encode world knowledge information. We developed a new TP system that, by incorporating rich lexical semantic information, outperforms all systems in Task B (F1-Class 72.24) and qualifies second in Task C (F1 29.69). Additionally, we performed an error analysis by comparing the output of all the systems and detecting the easy and difficult cases for Tasks B and C.

The results of the error analysis can be summarized as follows: i) training data are skewed and unbalanced thus making it hard for current machine learning methods to deal with rare and low frequent cases; ii) inference phenomena and world knowledge have a prominent role in resolving complex semantic tasks such as TP.

In the final version of the paper we will elaborate on this and propose future directions on how to improve the current systems both by using distributed features (e.g. word embeddings) and different learning algorithms for the different tasks (e.g. Bi-LSTM for temporal classification).

[7] The official scorer computes the temporal awareness including inferred temporal link, but it does not output them.

[8] In the final version of the paper we will provide more details and examples.

7. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Bethard, S., Savova, G., Palmer, M., and Pustejovsky, J. (2017). Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada, August. Association for Computational Linguistics.

Bethard, S. (2013). Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

Caselli, T. and Sprugnoli, R. (2017). In *It-TimeML and the Ita-TimeBank: Language Specific Adaptations for Temporal Annotation*, pages 969–988. Springer.

Cassidy, T., McDowell, B., Chambers, N., and Bethard, S. (2014). An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland, June. Association for Computational Linguistics.

Chambers, N. (2013). NavyTime: Event and Time Ordering from Raw Text. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 73–77, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Choubey, P. K. and Huang, R. (2017). A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1803, Copenhagen, Denmark, September. Association for Computational Linguistics.

Comrie, B. (1985). *Tense*, volume 17. Cambridge University Press.

Declerck, R. (1986). From Reichenbach (1947) to Comrie (1985) and beyond. *Lingua*, 70(4):305 – 364.

Derczynski, L. (2013). *Determining the Types of Temporal Relations in Discourse*. Ph.D. thesis, University of Sheffield.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G., (2002). *Instruction Manual for the Annotation of Temporal Expressions*. MITRE, Washington C3 Center, McLean, Virginia.

Jung, H. and Stent, A. (2013). Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 20–24, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Kolomiyets, O. and Moens, M.-F. (2013). Kul: Data-driven approach to temporal parsing of newswire articles. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 83–87, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Kolya, A. K., Kundu, A., Gupta, R., Ekbal, A., and Bandyopadhyay, S. (2013). JU_CSE: A CRF based approach to annotation of temporal expression, event and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 64–72. Citeseer.

Lacalle, M. L. D., Laparra, E., and Rigau, G. (2014). Predicate matrix: extending semlink through wordnet mappings. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Llorens, H., Chambers, N., UzZaman, N., Mostafazadeh, N., Allen, J., and Pustejovsky, J. (2015). Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado, June. Association for Computational Linguistics.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. pages 55–60.

Mirza, P. and Tonelli, S. (2014). An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003). The TimeBank corpus. 2003:40.

Reichenbach, H. (1947). *Elements of symbolyc logic*. the Free Press.

Setzer, A. (2001). *Temporal information in newswire articles: an annotation scheme and a corpus study*. Ph.D. thesis, University of Sheffield.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Pro-*

ceedings of the 4th International Workshop on Semantic Evaluations, pages 75–80. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Vossen, P., Caselli, T., and Kontzopoulou, Y. (2015). Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49, Beijing, China, July. Association for Computational Linguistics.

8. Language Resource References

Predicate Matrix Working Group. (2017). *Predicate Matrix*. http://adimen.si.ehu.es/web/PredicateMatrix, Unspecified, ISLRN 264-387-270-241-5.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen , R., and Setzer, A. (2014). *TimeBank 1.2*. Linguistic Data Consortium: Web Download, 1.0, ISLRN 717-712-373-266-4.