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

Tommaso Caselli†, **Roser Morante**‡
†CLCG – University of Groningen, ‡CLTL Lab – VU Amsterdam
The Netherlands
†t.caselli@gmail.com|rug.nl, ‡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 (*timexes*). Relations between these mentions can be meaningfully interpreted by using models of time, which allow to connect events on a timeline (temporal anchoring) and to understand complex sequences of events (temporal ordering). Temporal relations (TRs) provide 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 release 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] The TempEval-3 campaign (UzZaman et al., 2013) is the latest campaign on open-domain TP in English. The major contribution of TempEval-3 is the release of a platform for the development and evaluation of end-to-end TP systems based on the TimeML mark-up language (Pustejovsky et al., 2003).

After the TempEval-3 evaluation a new dataset has been released, the TimeBank-Dense corpus (Cassidy et al., 2014). This corpus has been developed to address one of the major shortcomings of the TempEval-3 dataset, namely lack of connectivity between all possible **e-e** and **e-t** pairs, by completing the transitive closure. The outcome of this approach is a very dense temporal graph consisting of 12,713 annotated TRs, with a 6.3 ratio of relations to events and timexes, which is much higher than the 0.8 ratio of the TempEval-3 data, where 11,098 TRs were manually annotated. Density has been obtained by forcing the annotators to always provide an answer. Additionally, the set of TRs has been simplified with respect to the one used in TempEval-3. Only 6 values are used instead of the original 14: `BEFORE`, `AFTER`, `INCLUDES`, `IS_INCLUDED`, `SIMULTANEOUS`, and `VAGUE`. The value `VAGUE` applies to cases where there is no TR between **e-e** and **e-t** pairs, and to cases where a TR exists but the specific value cannot be reliably determined. Not surprisingly, the `VAGUE` value is the most frequently used (selected 5910 times).

This paper focuses on analysing errors of several TP systems that participated in the TempEval-3 campaign, to identify their limitations. 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 campaign. However, our error analysis is different due to differences in the setting of the two competitions: the test data of TempEval-2 is a subset of the TimeBank corpus, whereas in TempEval-3 the evaluation is conducted on a new set of manually annotated documents and the entire TimeBank corpus is made available for training; systems participating in TempEval-2 produced a simplified set of TRs, whereas systems participating in TempEval-3 are asked to provide the full set of fine-grained TRs which have been proposed in TimeML and annotated in the TimeBank corpus; and in TempEval-2 systems were not evaluated on an end-to-end approach, whereas in TempEval-3 they were.

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

[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).

system, which we release to the public.[3] 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 the TP systems that participated in the TempEval-3 competition whose output is publicly available, highlighting commonalities, differences, and limitations. The results of the error analysis are presented in Section 5. and Section 6. for event trigger detection and temporal relation classification, respectively. Finally, Section 7. 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 campaigns (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 (Saurı et al., 2006). In particular, an *event* is defined as any linguistic mention, including verbs, nouns, adjectives and prepositional 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: 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`. The value `IDENTITY` is actually non-temporal, but it is used to identify coreference relations between event mentions.

For the TempEval-3 campaign extra training data were provided, by automatically annotating almost 600,000 tokens for event, timexes, and TRs. Additionally, a new test data was released with manual gold annotations (20 articles, 8,000 ca. tokens). Evaluation was conducted by means of a new evaluation measure, aimed at assessing the *temporal awareness* of end-to-end systems (UzZaman et al., 2013). Temporal awareness measures the ability of a system to identify and classify TRs. This includes the correct identification and classification of the temporal entities participating in the TR, i.e. event mentions and timexes.

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

Table 1 contains the number of events in the TempEval-3 data splitted by part-of-speech (POS), and Table 2, the distribution per value of TRs for the manually annotated (Training-Gold and Test) and the automatically annotated (Training-Auto) training data .

Event POS	Training-Gold	Training-Auto	Test
Verb	5837	65813	539
Noun	2450	13489	169
Adjective	202	473	23
Preposition	10	0	1
Other	19	879	15
Overall	11108	80654	749

Table 1: Events per POS in the TempEval-3 dataset.

TLINK Value	Training Gold	Training Auto	Test
BEFORE	3701	45581	330
AFTER	1361	35241	200
INCLUDES	1523	5062	91
IS_INCLUDED	2287	16029	177
BEGINS	110	7	2
ENDS	77	0	3
BEGUN_BY	70	5	3
ENDED_BY	137	0	2
SIMULTANEOUS	580	10894	93
IAFTER	49	84	10
IBEFORE	65	8	8
DURING	280	0	2
DURING_INV	0	0	1
IDENTITY	858	0	15
Overall	11098	112911	937

Table 2: TRs per temporal value in the TempEval-3 dataset.

3. Temporal Processing Systems: a Review

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). In total we review 6 unique systems (5 for event detection and classification only and 4 for the full TP).

3.1. Event detection and classification

The event detection and classification task is 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 are represented in the learning models, which can be aggregated in 5 groups:

- Basic morpho-syntactic features: token, lemma, stem, parts-of-speech (POSs), token's affix and/or suffix, among others.
- Syntactic features: constituency/dependency syntax relations; governing verb lemma, verb chunks.
- Contextual features: 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 are more complex in terms of features than for the classification task. Semantics and lexical semantics features are used by less systems (2 systems for event detection and only 1 system for classification), and they are limited to WordNet data only. F1 scores range from 78.2 to 81.05 for event detection, but decrease for event class (from 52.69 to 71.88), and other attributes, such as tense (from 49.7 to 61.63) and aspect (from 63.2 to 73.5). The analysis of systems shows that:

- The identification of event attributes benefits from the use of event detection features only when a rich feature set is used that includes extended syntactic (constituent and dependency parsing) and semantic information (semantic roles).
- Reducing the feature complexity for event attribute identification does not help a machine learning system to generalize enough to beat a baseline based on the most frequent values (Bethard, 2013), but it allows machine learning systems to perform better than rule-based solutions, at least for this dataset (Kolya et al., 2013).
- Varying the composition of the training data (i.e., only gold data (Chambers, 2013) vs. gold data plus silver data (Jung and Stent, 2013)) affects the final results of the specific problem addressed, by either improving the results, like for the Event Trigger and Class subtask, or by downgrading them, as in the case of the TR substask.

3.2. Temporal relation detection and classification

The temporal relation detection and classification task is addressed as a supervised multi-class classification task. Systems use 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)).
Three of four systems solve the task in a two-step approach: recognition of eligible temporal relations and assignment of the temporal values. Only one 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 two 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, two 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).
The best system obtains 30.98 F1 for global temporal awareness (Bethard, 2013). It uses a reduced set of temporal relations (3 for **e-dct** and 2 for **e-t** and **e-e**), and models only intra-sentence relations.

4. CRF4TimeML: A New TP System

Based on our study of participating systems, we developed a new end-to-end TP system, CRF4TimeML. Similarly to previous work, we used a single learner and we split the task in multiple subtasks. The system is based on a cascade of 7 CRF classifiers. At this stage of development, we have chosen to use a discrete classifier (CRF), relying more on feature engineering than on distributed feature representations, such as word embeddings, and neural network architectures. This choice is motivated by two main reasons: i) we aim at understanding both the fitness and the limitations of the selected features for a high-level semantic task such as TR identification and classification; ii) we intend to establish a relation between the errors made by the systems, including ours, and the features used.
CRF4TimeML has been designed taking as reference efficient existing systems that incorporate discrete classifiers. In particular, all classifiers we developed share with previous systems basic morpho-syntactic features, such as token, lemma, POS, and dependency relations. However, we have added lexical semantic information by using not only WordNet synsets, but also VerbNet classes and FrameNet frames, obtained from the alignments in the Predicate Matrix (Lacalle et al., 2014). The pre-processing of data is performed with state-of-the-art tools, such as the Stanford CoreNLP (Manning et al., 2014) and the NewsReader NLP pipelines (Agerri et al., 2014). [4]

4.1. Event Detection and Classification

The event detection and classification task is performed by 4 different classifiers share the basic morpho-syntactic and

[4]The timex detection and normalization task is performed using a state-of-the-art system (Bethard, 2013), available at `https://bitbucket.org/qwaider/textpro-en`.

lexico-semantic features: one for the identification of event triggers, and 3 for the identification of attributes, namely Class, Tense, and Aspect. Each classifier has then its own set of specific features:

- The Event Trigger classifier is extended with semantic roles, context windows of [+/-2] for token, lemmas, and POS.
- The Class classifier uses the typed dependency from an event token to the root token of the sentence, plus semantic roles.
- The Tense classifier is extended with context windows of [-2] tokens, and POS.
- The Aspect classifier employs (predicted) tense, a combination of tense and POS, and context windows of [-4] tokens, and POS.

We used the TempEval-2 test data as a development set to fine-tune the context windows and features. This allowed us to observe that:

- The quality of pre-processsing tools concerning lemmatization, POS tagging, and syntactic dependencies affects the quality of the Event Trigger classifier, especially for Recall.
- The best scores for Event Trigger and Class are obtained by using manually and automatically created training data. Although training with automatically generated data causes replication of errors, for these 2 subtasks, with high variability of realizations, it can positively affect the learning process by providing more positive examples for the sparse classes. Nevertheless, for the other attributes, which have less variability and are more dependent on specific combinations of co-textual data, automatically generated data lower the quality of the classifiers (e.g. the values PRESENT and PERFECTIVE for tense and aspect can only be correctly derived by identifying the relationship between an auxiliary and the main verb).
- Syntactic and semantic features have the biggest impact on the classifiers performance: using only morpho-syntactic and context window features gives an F1 of 82.1 for Event Trigger detection, which increases to 88.2 when adding lexical semantics features only, and reaches 90.9 when lexical semantics information is combined with syntactic information.

The best results of the CRF4TimeML system are obtained with the following configuration: [5]

- Pre-processing was performed with NewsReader NLP for semantic roles, and Stanford CoreNLP for tokenization, morpho-syntax, and dependency parsing.
- Adding the automatically generated training data to the manually created training data only for the Event Trigger and Class classifiers.
- Adding additional semantic information: VerbNet classes and FrameNet frames.

Table 3 provides the results of the CRF4TimeML system for Event Detection and Classification, as well as the results of the other TempEval-3 systems.

System	Trigger	Class	Tense	Aspect
CRF4TimeML	**81.87**	**72.24**	60.87	73.18
(Jung and Stent, 2013)	81.05	71.88	59.47	73.5
(Bethard, 2013)	78.81	67.87	**61.63**	71.61
(Chambers, 2013)	80.3	67.48	60.86	**73.28**
(Kolya et al., 2013)	78.62	52.69	58.62	72.14

Table 3: F1 system results on the TempEval-3 Event Detection and Classification Task (Event Trigger, Class, Tense, and Aspect).

CRF4TimeML obtains the best F1 scores for Event Trigger and Class, while the results for Tense and Aspect are slighlty lower than the scores of the best systems. According to the TempEval-3 ranking, CRF4TimeML would be first. However, differences in results are not statistically significant after performing the McNemar's test ($p>0.05$).

4.2. Temporal Relation and Classification

The TR task is addressed by means of 3 multi-class CRF classifiers, one for each pair of temporal entities (**e-dct**, **e-t**, and **e-e** pairs), which predict the 14 TimeML temporal values. Similarly to existing systems, we target only intra-sentence relations for **e-t** and **e-e** pairs, given that the number of cross-sentence relations in the training data is low. The classifiers are trained with the gold data set only, plus additional relations from Bethard et al. (2014). Different pairs have been normalized with respect to the directionality of the relation in order to reduce the variability of the temporal values. This resulted in the following ordering of pairs: i) relations involving an event and a timex, including the DCT, have been represented as **e-t/e-dct** pairs; ii) relations involving event pairs have been normalized according to the linear order of presentation of the events in the sentences.

In this task the system uses predicted event triggers, which in the learning model are represented with the morpho-syntactic and lexical-semantic features used in the Event Detection and Classification task plus the predicted values for class, tense, and aspect.

Each TP classifier uses additional specific features, namely:

- **e-dct**: TimeML type of the DCT (DATE or TIME).
- **e-t**: Additional features for each each timex: the TimeML type, token(s), dependency relation, head, lemma, POS, and a combination of the POS, dependency relation, and head POS. In addition to this, we have included: the dependency path connecting the event and the timex, the token(s) of any temporal signal between the event and the timex, the token of a

[5] Details for replicating the experiments and results are available at `https://github.com/cltl/TimeMLEventTrigger`

temporal signal at the beginning of the sentence, and the distance between the two elements in the pairs.

- **e-e**: Features about the connection of the two events (typed dependency paths; typed dependency paths and POS; typed dependency paths, tokens, and POS); a combination of the tense values of the two events; a combination of the tense and aspect values of the two events; the distance between the two events in the sentence; the token of a temporal signal before the first event in the pair; the token(s) of any temporal signal between the two events; presence of any other events between the elements in the pair; a combination of the events classes.

One the known problems of the TempEval-3 dataset is that the annotation of TRs is not complete because the temporal closure of the relations between all temporal entities is not available. Following the solution proposed in the TimeBank Dense corpus (Cassidy et al., 2014), we have assumed that in the test data the temporal closure of all events and timexes has been calculated, so that all events are temporally connected to each other. With this biased model we aim at better evaluating the completeness of the test data, by identifying pairs which are correct but not annotated.
Finally, we use the 14 temporal relation values, rather than simplifying the set to the most frequent ones in the training data only.
Table 4 contains the results of our TP system, compared to the results of the reviewed systems.

System	F1	P	R
CRF4TimeML	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 4: Results for TempEval-3 Task C (Temporal Processing from raw text).

Our system qualifies as the second best on this task (F1 29.69). The higher recall and lower precision is a consequence of assuming a full temporal connection of the entities. Systems that are designed based on this assumption tend to over-generate TRs. Breaking down these results per type of entity pairs (see also Table 5), our system has the best F1 for **e-e** pairs (25.51), while the best score for the other participating system is obtained by Chambers (2013) (F1 19.01). Results are different 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**. The different scores for the **e-t** pairs and the rest indicates that more (and better) annotated data are needed on specific pairs of temporal entities, namely **e-dct** and **e-e**.

5. 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. These numbers indicate that more systems agree for events with POS verb, than for events with POS noun and that events with POS noun are more difficult to detect.
The high number of verbs being correctly detected might reflect the annotation decisions stated in the TimeML annotation guidelines, which are strongly leaning towards the verb category (Pustejovsky et al., 2003). This impacts the results of the systems, which are well trained to detect events expressed by prototypical POS (i.e. verbs), while detecting events with less prototypical POS remains a challenge. This is coherent with statistics from the training data, where 80.5% of events are verbs. 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 affected by metonymy, as in the Example 1 where *Everest* is a proper noun that refers to the event 'climbing the Everest'. Solving this cases would require a system to apply inference mechanisms.

1. He said: "Lowe was a brilliant, kind fellow who never sought the limelight ... and 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. The rest do not occur in the training corpus. This raises the question of how reliable systems are when confronted to previously unseen data. It also confirms the well-known dependence on training data of discrete models and their limitations to generalize.
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.

6. Temporal Relations: When is it wrong?

For the error analysis of the TR subtask we look at three aspects: i) how many and what type of relations are incorrectly classified by all systems; ii) what type of processing

	[e-dct]			[e-t]			[e-e]		
System	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
CRF4TimeML	21.95	13.66	**55.80**	27.71	20.72	41.81	**25.51**	22.40	**29.61**
(Bethard, 2013)	**24.75**	16.54	49.17	**41.41**	**38.46**	**44.84**	16.57	**35.62**	10.80
(Chambers, 2013)	19.72	**25.66**	16.02	40.65	37.18	**44.84**	19.01	25.52	15.15
(Kolya et al., 2013)	20.76	13.04	50.82	27.05	21.32	36.96	18.50	15.10	23.86
(Kolomiyets and Moens, 2013)	21.15	13.17	53.59	1.66	2.66	1.21	9.88	13.92	7.66

Table 5: Results for Temporal Relations Detection and Classification per type of TLINK.

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.

Table 6 presents the errors of systems per type of TR. For a case to be considered correct, both the pair of temporal entities and the TR value have to be correct with respect to the gold standard data.

Correct	**[e-dct]**	**[e-t]**	**[e-e]**
All systems	16 (10.70%)	0 (0%)	2 (0.45%)
4 systems	45 (30.20%)	35 (30.07%)	9 (2.04%)
3 systems	31 (20.80%)	31 (27.19%)	28 (6.34%)
2 systems	12 (8.05%)	8 (7.01%)	79 (17.91%)
1 system	23 (15.43%)	18 (15.78%)	150 (34.01%)
No system	22 (14.76%)	22 (19.29%)	173 (39.22%)

Table 6: Number of correct solutions per number of systems and type of relation in the TLINK task.

The first observation is that there is no pair of **e-t** where all systems "agree" in providing a correct answer. A similar observation holds for the **e-e** pairs, although, in this case, only in 2 instances all systems provide a correct answer. These figures, especially for the **e-t** and **e-e** relations, clearly point out that both anchoring and ordering TRs are complex tasks when performed from raw text data.

Furthermore, the complexity of both tasks is increased, in this case, by divergences in how the training and test data have been annotated. [6] For instance, when looking at the percentage of events realized by nouns and verbs in the manually annotated TempEval-3 training data, we find that 43.78% of events are not temporally connected, while in the test data they amount to 20.11%. The lack of annotated TRs in the training data makes the two-step learning algorithm weak, since the success of TR classification depends on having found first the right pairs of entities which stand in a TR. This is also an additional motivation for having assumed, in this work, the temporal closure of **e-e** and **e-t** pairs.

With the exception of one system (Bethard, 2013) that uses a restricted set of temporal values, all system outputs are as skewed as in the training data. 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.

Concerning **e-e** relations, we have analysed also the impact of tense and aspect values. According to tense and aspect semantics (Comrie, 1985), differences in tense forms between sequences of events are primary hints for correct TR identification and classification. We look at the cases where all systems fail, but we found that the percentage of errors in sequences with different tense and aspect values is similar to the percentage of errors in sequences of events with the same tense and aspect values (38.38% and 40.76%, respectively). This raises questions about choices in the TimeML Annotation Scheme concerning i) the granularity of tense and aspect attributes, and ii) their annotation methodology. For instance, having more fine-grained values for the Past temporal dimension, thus allowing to differentiate between a simple past and a past perfect, may positively affect the task. Furthermore, the TimeML surface-based annotation philosophy should not be applied to the tense and aspect values because rather than surface forms, what is actually needed is a contextual interpretation of these values.

We have analyzed the **e-e** pairs in terms of parts-of-speech. Apparently, TRs between events with different parts-of-speech is less prone to errors than pairs with the same parts-of-speech values (25.26% and 38.15%, respectively, when looking at the cases where all systems fail to correctly classify the TR).

As a result of a detailed error analysis, we provide a classification of errors into the 6 categories listed below.[7] Except for *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.
- **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 ordered 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 (including *aktionsaart*), discourse structure, factuality profiling of events, script and frame knowledge, and general world knowledge.

[6] On this point see also (Cassidy et al., 2014; Orasmaa and Kaalep, 2017)

[7] 4 classes: Iconicity, Inference, Signaled and World Knowledge have been proposed in Derczynski (2013).

- **Error**: the gold temporal relation is either wrong or in dispute.

We summarize our findings in Table 7. The following 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 (4.59%)
Signaled	0 (0%)	9 (40.9%)	9 (10.34%)
Inference	6 (27.27%)	5 (22.72%)	21 (24.13%)
Grammar	4 (18.18%)	1 (4.5%)	20 (22.98%)
World Knowledge	11 (50%)	3 (13.63%)	32 (36.78%)
Error	1 (4.5%)	4 (18.18%)	1 (1.14%)

Table 7: TLINK error classification of errors for cases where all systems fail.

With the exception of the **e-dct** pairs, world knowledge has a less prominent role than expected, especially for **e-e** pairs. Previous studies (Van Der Meer et al., 2002; McRae and Matsuki, 2009; Caselli and Prodanof, 2009) have shown that event knowledge is highly salient and, actually, plays a preeminent role in the identification and classification of temporal relations by humans. Nevertheless, as our data show, the impact of world knowledge is limited to 36.78% of the cases for **e-e** pairs and only 13.63% for **e-t** ones. Theses figures seem to be in-line with other studies on error analysis for high-level semantic tasks. For instance, Clark et al. (2007) in their error analysis of the RTE-3 dataset show that world knowledge has an impact on the correct resolution of text/hypothesis pairs only for 36% of the cases. We report below some examples of temporal entity pairs which could be correctly solved only by taking into account world knowledge.

2. (**DCT:2013-03-21**) **Lodged** in his brain is an untreatable and inoperable cancerous tumor [...]. [WSJ_20130321_1145.tml] `TLINK e-dct:INCLUDES`

3. He won the Gusher Marathon, **finishing** in **3:07:35**$_{DURATION}$. [WSJ_20130321_1145.tml] `TLINK e-t:ENDS`

4. Over 200,000 Belfast customers were **affected** by a blackout but power is starting to be **restored**. [bbc_20130322_1600.tml] `TLINK e-e:INCLUDES`

An average of circa 24% for TRs between the three types of entity pairs could be addressed and (possibly) correctly solved if inference mechanisms were applied. Being able to keep track of which entities have been already temporally connected and, most importantly, with which values, may positively impact the resolution of the task. In example 5, the temporal relation between the event pair **fail**-**creates** can be solved if the TR between the event **last year** and the timex *fail* is also taken into account and not just, for instance, the tenses of the two events.

5. Greece may well have been too big to **fail** *last year*, but Cyprus, which **creates** less than one-half percent [...] [nyt_20130321_cyprus.tml] `TLINK e-e:BEFORE`

Grammatical information has an important role for **e-e** relations. The following cases could have been solved if grammatical structure (Example 6), and tense and aspect values (Example 7) were correctly processed and taken into account.

6. [...] it would consider **keeping** a tower, *if* the airport **convinces** the agency [...] [CNN_20130322_248.tml] `TLINK e-e:AFTER`

7. "We are **growing**$_{present:progressive}$ in number", **said**$_{past:none}$ Senator Amy Klobuchar [...] [nyt_20130321_women_senate.tml] `TLINK e-e:INCLUDES`

A subset of errors could have been avoided by having access to the contextual interpretation of the tense and aspect values, as illustrated in Example 7. For both events, we report the TimeML values of tense and aspect.

8. The season **started**$_{past:none}$ a month ago, **sparking**$_{present:none}$ concerns [...] [AP_20130322.tml] `TLINK e-e:SIMULTANEOUS`

Another cause of errors is incorrect processing of pairs where one of the elements is a reporting verb. These cases require careful processing of the linguistic context in order to identify the correct TR between the reported events and the reporting verbs. To correctly process event pairs that contain a reporting verb, it seems important to distinguish the type of reported speech, i.e. whether it is direct, indirect, or mixed. Reported speech tends to maintain an internal coherence and a specific temporal dimension which may differ with respect to the one signaled by the reporting verb and, in some cases, may also not fit the expected sequence of tenses.

9. [...] gale force wind would **blow** snow ...they **added**. [bbc_20130322_1600.tml] `TLINK e-e:AFTER`

10. "We **appear** to be getting close [...]" Jhung **added**. [AP_20130322.tml] `TLINK e-e:INCLUDES`

Additional errors are found in inter-sentential **e-e** relations. Overall, they represent 26.96% of the analyzed data. Interestingly, the correct processing of inter-sentential **e-e** pairs cannot be related to a particular processing requirement, since errors are distributed in the 6 categories reported in Table 7.

Focusing on 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. [8] 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

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

test files to establish whether 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.

7. 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 end-to-end TP system that, by incorporating rich lexical semantic information, performs better than all systems for the Event Detection and Classification task, (F1-Class 72.24) and qualifies second on the TR Identification and Classification task (F1 29.69). Additionally, we performed an error analysis by comparing the output of all the systems focusing on the cases where no system can give a correct answer.

A detailed error analysis shows that there are easy and difficult cases both for event trigger detection and TR processing. Summing up, for event detection and classification problems arise when non-prototypical POS and polysemous lexical items are involved, while for TRs the difficulty lies in the creation of the pairs.

Concerning the classification errors of temporal relations, we observe that inference phenomena and world knowledge have a prominent role. As for inference, the analysis of data suggests that a two step strategy should be followed: first, provide a temporal anchor to the events by addressing first the **e-t** and **e-dct** pairs, and then use this information to enrich models to learn **e-e** pairs. Sieve-based architectures expanded with transitivity rules like the one proposed by Chambers et al. (2014) are addressing this problem in the right way but they require "densely" annotated data, or else the transitivity rules fail.

We showed that rich lexical semantic information is beneficial for the TP task but not enough. Recent work (Mirza and Tonelli, 2016; McDowell et al., 2017) has shown on a different dataset (TimeBank-Dense corpus) that word embeddings positively contribute[9] to the classification of TRs between **e-e** pairs both when occurring in the same sentence and in different sentences.

As a result of our error analysis, we would like to stress the following aspects as possible future directions. Firstly, we need more data, systematically annotated, in the line of the TimeBank-Dense corpus. At the same time, we think that it is important to densely annotate only those temporal entities (namely events) which are *actually* in a temporal relation, avoiding to introduce temporal chains with events which are not temporally connected. This would make the task more difficult but also more natural with respect to how humans reconstruct a document's timeline. Secondly, systems for TP should be able to "keep track" of the incremental nature of the document. So far, almost all approaches (also those that recently targeted the task via Neural Networks (Choubey and Huang, 2017)) have assumed that each sentence is a discourse universe of its own, while sentences occur in a text, which is a unitary and coherent message. Finally, the granularity of temporal relations should be reduced as most of the fine-grained values are very rare (and in some cases very hard to annotate), thus having a negative impact in the development of robust machine learning models.

[9] McDowell et al. (2017) report an increase of 5% in F1 in the TimeBank-Dense corpus

8. Acknowledgements

The work presented in this paper was funded by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines".

9. 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.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Bethard, S., Chambers, N., McDowell, B., and Cassidy, T. (2014). An annotation framework for dense event ordering. In *52nd Annual Meeting of the Association for Computational Linguistics*, 52, Baltimore, MD, USA, June. Association for Computational Linguistics.

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 Prodanof, I. (2009). Robust temporal processing: from model to system. *Research in Computing Science. Special Issue: Natural Language Processing and its Applications*, pages 29–40.

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., Cassidy, T., McDowell, B., and Bethard, S. (2014). Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

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.

Clark, P., Murray, W. R., Thompson, J., Harrison, P., Hobbs, J., and Fellbaum, C. (2007). On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 54–59. 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.

Christiane Fellbaum, editor. (1998). *WordNet: an electronic lexical database*. MIT Press.

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. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 55–60.

McDowell, B., Chambers, N., Ororbia II, A., and Reitter, D. (2017). Event ordering with a generalized model for sieve prediction ranking. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 843–853.

McRae, K. and Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, 3(6):1417–1429.

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.

Mirza, P. and Tonelli, S. (2016). On the contribution of word embeddings to temporal relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2818–2828.

Orasmaa, S. and Kaalep, H.-J. (2017). Can we create a tool for general domain event analysis? In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 192–201.

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.

Saurı, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). Timeml annotation guidelines version 1.2. 1.

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.

Van Der Meer, E., Beyer, R., Heinze, B., and Badel, I. (2002). Temporal order relations in language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):770.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (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.

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.

10. 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.

VV.AA. (2017a). *VerbNet*. unspecified, ISLRN 606-348-320-105-2.

VV.AA. (2017b). *WordNet*. unspecified, ISLRN 379-473-059-273-1.