Chapter 4
Relation Labelling Analysis

Felix, qui potuit rerum cognoscere causas
Fortunate was he, who was able to determine the causes of things

Georgica (II, v.490)
VIRGIL

4.1 Introduction

In Chap. 3, we discovered that automatic temporal relation typing is a difficult problem. This motivates an investigation into potential ways of improving performance in relation typing. This chapter details an attempt to discover potential ways of improving performance at the task. As humans are readily able to identify the nature of temporal links, one may a priori draw the conclusion that the information required to do so must be available somewhere. This knowledge is in a given document or in information known by the reader before encountering that document (referred to as **world knowledge**). Following the tradition of performing post-hoc analyses on temporally annotated corpora [1, 2], we attempt to characterise and enumerate the in-document knowledge used to support temporal link labelling. In later chapters, we will use some of these types of knowledge to improve automatic temporal relation labelling.

Firstly, this chapter reports on an attempt to identify a common set of challenging temporal links in the TempEval-2 evaluation task. This includes re-examination of the surface information available in TempEval-2 data and an analysis of its distribution in difficult links. Secondly, finding that the surface information presents no clear paths for investigation (as suggested by the performance cap of previous work using surface information discussed in Sect. 3.5.6), a manual investigation of difficult links is undertaken. This comprises a qualitative characterisation of the information used to label the links and motivates our later experimental investigations.

© Springer International Publishing AG 2017

L.R.A. Derczynski, *Automatically Ordering Events and Times in Text*,
Studies in Computational Intelligence 677, DOI 10.1007/978-3-319-47241-6_4

4.2 Survey of Difficult TLINKs

Our hypothesis is that there may be temporal relations that are consistently difficult to classify correctly. That is, some meta-system using an agglomerative approach (e.g. voting) will still have problems with the relation typing problem. It has been difficult to conduct a thorough error analysis of the temporal relation typing task, as authors often do not or cannot make their attempted labellings available, instead publishing more concise overall performance figures. Further, there are many different corpora and corpora-versions used, which hampers comparability.

This section introduces a source of data on attempts at the relation labelling task, followed by a method for grading temporal links in terms of difficulty, reports on the measured proportions of the degrees of difficulty found in typing various temporal relations, defines what constitutes a difficult link and finally presents a data-driven analysis of difficult links based on their surface features.

4.2.1 The TempEval Participant Dataset

As mentioned in Sect. 3.4.4.4, the TempEval exercises strive to produce comparable results over a fixed and agreed dataset, using pre-annotated events, timexes and TLINK arguments, which constrains the scope for variation in systems outside the task focus – temporal labelling methods.

The second TempEval exercise took place in 2010, as part of SemEval [3]. This exercise included four temporal link labelling exercises, in multiple languages, over a purpose-built corpus. Many teams participated in the evaluation and attempted to label these temporal links. As a result, from their submissions we gained a snapshot of the state of the art of temporal link labelling, all on the same data, with multiple approaches. Some teams were prepared to share their submitted results, which, when compared with the correct answer data and the original corpus, could be merged. From this, we were able to measure a "success rate" for each temporal link, determined by the proportion of systems that managed to label it correctly. We then can build a list of links that are difficult for most (or all) of the systems to annotate automatically.

Fortunately, the TempEval-2 organisers released a full dataset of not only source but also evaluation data.[1] Data concerning the distribution of features over events are contained in Figs. 4.1 and 4.2, of features over timexes in Fig. 4.3.

After contacting teams participating in temporal relation labelling tasks, many were kind enough to donate their submitted labels [4–7]. This data was used to conduct a data-driven failure analysis of four separate temporal linking tasks undertaken

[1] Downloadable from http://timeml.org/site/timebank/tempeval/tempeval2-data.zip. It is important to note that this contains more data than was in the tasks set; evaluating systems using this release as-is will not give accurate figures.

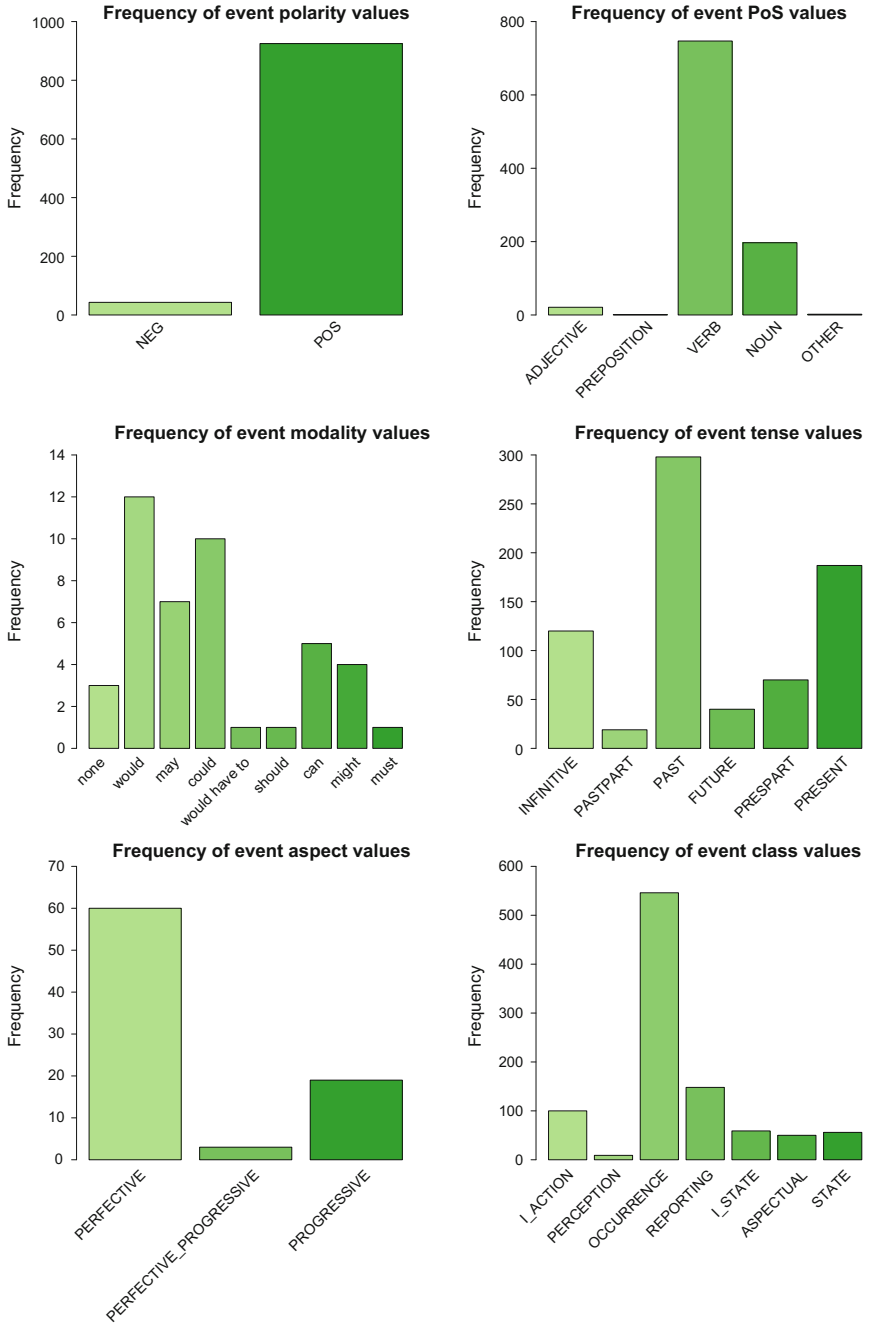

Fig. 4.1 Frequencies of event attribute values in the TempEval-2 English test data

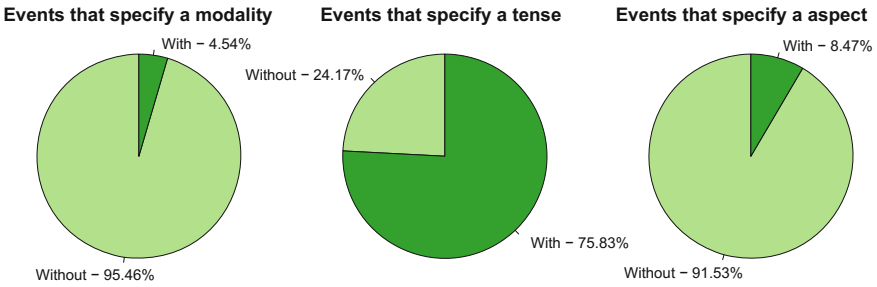

Fig. 4.2 Proportions missing events attribute values in the TempEval-2 English test data

by directly comparable systems. The analysis continues the work on TempEval-1 by [8] and incorporates data from many individual teams.

Given the apparent performance ceiling of systems that use only the annotated TimeML/TempEval-2 feature:value pairs (surface information), clear directions for further investigation are not expected from a formal analysis using these feature:value pairs. However to omit an analysis of difficult links in terms of their arguments' TempEval-2 descriptions would be to ignore a potentially useful and readily available information source and so results are included below.

4.2.2 Defining What Constitutes "difficult" Temporal Links

We start by measuring the "difficulty" of each link, calculating the proportion of attempting labelling systems that generated a correct response. The measurements have values ranging from "all systems correct" (an easy link) to "no systems correct" (a difficult link). This gives a discrete set of difficulty categories for each task. We then count the number of links in each difficulty category as a proportion of the whole and present the data graphically. The results are shown in Fig. 4.4 and Table 4.1.

- **Task C** – Linking events and timexes in the same sentence. For example, *The day$_t$ before Raymond Roth was pulled$_e$ over …*
- **Task D** – Linking events with the document creation time. For example, *11/01/89$_t$ … As part of the agreement, Cilcorp said$_e$ it will co-operate.*
- **Task E** – Linking main events in adjacent sentences. For example, *There are 12 flood warnings in the South West, with Met Office warnings for snow covering$_{e1}$ much of the UK. This comes$_{e2}$ just over a week before the start of British Summer Time.*
- **Task F** – Linking main events with subordinate events. For example, *He said$_{e1}$ he discussed$_{e2}$ the issue with Mr. Netanyahu.*

This information permits a brief overall analysis of the relative complexity of the different relation tasks. Task E (Table 4.4) has a fairly stable difficulty gradient,

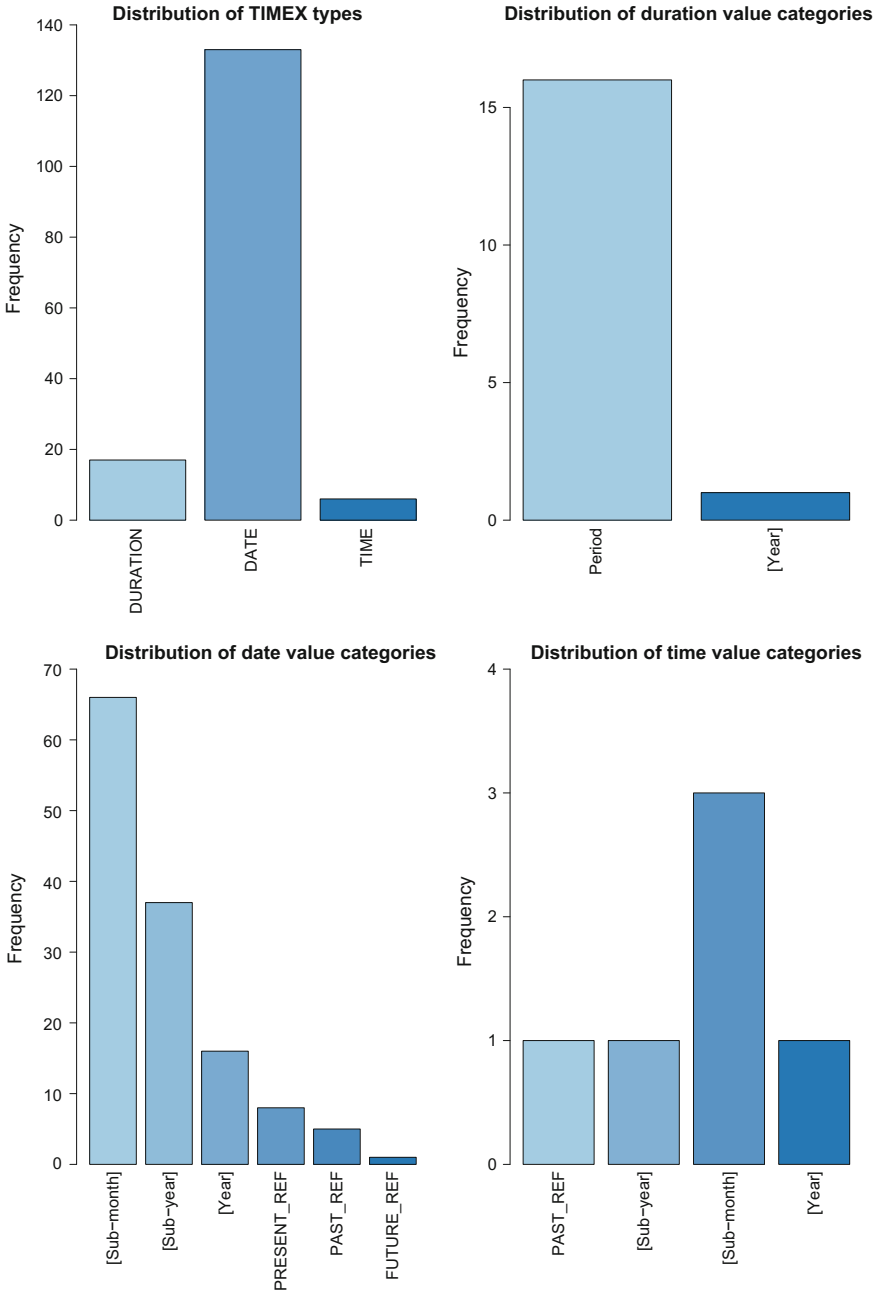

Fig. 4.3 Frequencies of timex attribute values in the TempEval-2 English test data

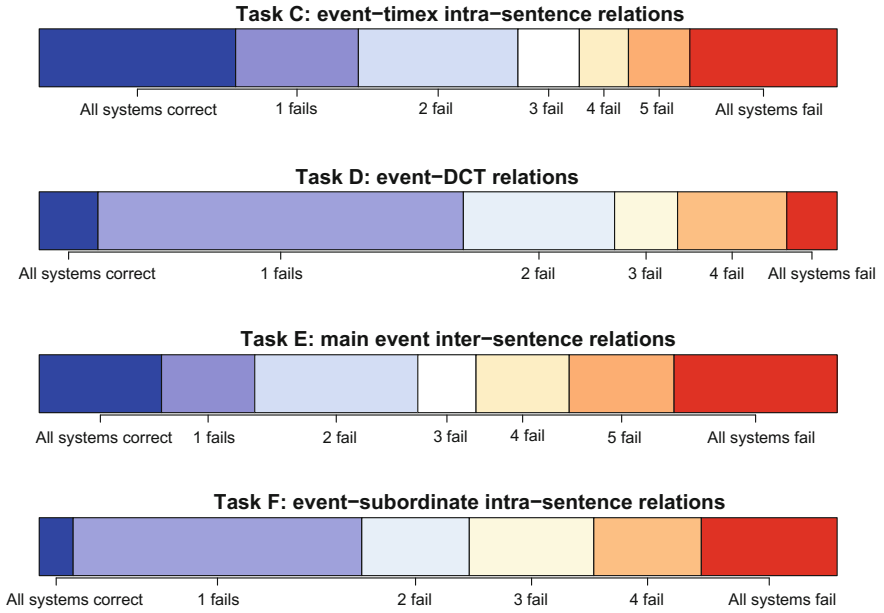

Fig. 4.4 TempEval-2 relation labelling tasks, showing the proportion of relations organised by number of systems that failed to label them correctly. Six systems attempted tasks C and E; five attempted tasks D and F

Table 4.1 Proportion of difficult links in each TempEval-2 task

Task	Difficult links	Difficult proportion (%)	Best score (%)
C	22	8.59	65
D	39	18.4	82
E	62	44.3	58
F	44	46.8	66

with the least deviation between category sizes. Task D (Table 4.3) is easiest. Task C (Table 4.2) has a very tough set; when compared to task E (Table 4.4), although a greater proportion of the links are successfully labelled, the size of the "all fail" group is the same in absolute terms and relatively dominates the set of harder links. Finally, it can be seen that event-event labelling (tasks E+F, Tables 4.4 and 4.5) is harder than event-timex labelling (C+D, Tables 4.2 and 4.3).

Data was available for five or six systems, depending on the task. One system only attempted two of the four tasks, so its absence should not unduly undermine the quality of overall observations. Difficult links are defined as those wrongly labelled by all systems or wrongly labelled by all-but-one system. Given this threshold, we can define a set of difficult links for further analysis. The composition of this set is given in Table 4.1 and shown in Fig. 4.5.

Table 4.2 Error rates in TempEval-2 Task C, event-timex linking

Systems in error	Number of TLINKs	% of TLINKs (%)
No faults	16	24.6
1 fault	10	15.4
2 faults	13	20.0
3 faults	5	7.69
4 faults	4	6.15
5 faults	5	7.69
All fail	12	18.5

Table 4.3 Error rates in TempEval-2 Task D, event-DCT linking

Systems in error	Number of TLINKs	% of TLINKs (%)
No faults	14	7.37
1 fault	87	45.8
2 faults	36	18.9
3 faults	15	15.8
4 faults	26	21.1
All fail	12	6.32

Table 4.4 Error rates in TempEval-2 Task E, linking main events of subsequent sentences

Systems in error	Number of TLINKs	% of TLINKs (%)
No faults	21	15.3
1 fault	16	11.7
2 faults	28	20.4
3 faults	10	7.30
4 faults	16	11.7
5 faults	18	13.1
All fail	28	20.4

Figure 4.6 shows the proportion of links within each task that are difficult and reinforces the earlier observation that event-event links are tougher than event-times links. In the figures, event-timex tasks (C and D) are shown in blue and event-event tasks (E and F) in green. Event-event tasks are comparatively hard, with higher proportions of difficult TLINKs.

Table 4.5 Error rates in TempEval-2 Task F, linking events to events that they subordinate

Systems in error	Number of TLINKs	% of TLINKs (%)
No faults	6	4.26
1 fault	51	36.2
2 faults	19	13.5
3 faults	22	16.1
4 faults	19	13.5
All fail	24	17.5

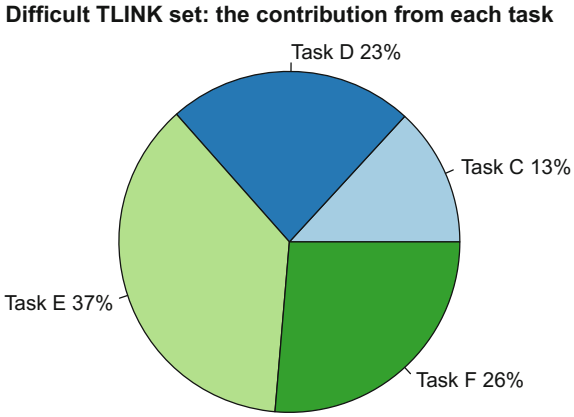

Fig. 4.5 Composition of the set of difficult links. Event-event tasks (E and F) in *green*, event-timex tasks (C and D) in *blue* (color figure online)

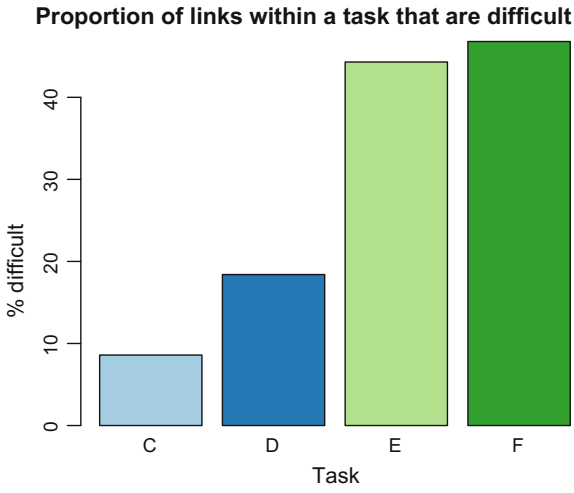

Fig. 4.6 Proportion of each TempEval-2 task's links that are difficult

4.2.3 Comparative Distribution Analysis

Given a set of gold-standard event annotations and gold-standard temporal link annotations, one can conduct a survey of features and values for temporal links. Given also a set of difficult links, one may determine which particular attribute combinations are difficult or easy to automatically label. This is demonstrated in Fig. 4.7, which may be read as follows. Each row corresponds to all events *related to* a given event having a particular property. For example, one row may detail the statistical properties of all other events that are linked to a verb event (e.g. having `pos.VERB`). The columns in this row show the distribution of feature/value pairs in the related event for all relations surveyed. So, continuing the example, in the `pos.VERB` row, the colour represents the likelihood of other argument in the temporal link having a particular feature/value pair. More saturated colours represent higher frequencies. Reds indicate relatively high presence in difficult links (e.g., a "hard" feature combination); blues indicate a low frequency in difficult links (e.g., that the feature combination is "easy").

One could imagine that graph 1 minus graph 2 is graph 3 and that the reds correspond to negative values. Let **A** be a matrix of feature:value co-distributions and **B** be feature co-distributions in the set of difficult links. If comparison $\mathbf{O} = \mathbf{A} - \mathbf{B}$, then negative values in **O** correspond to feature combinations that occur more frequently in **B** than **A**; that is, combinations that are more likely than average to be occur in difficult relations.

4.2.3.1 Difficult Event-Event Link Attribute Distribution

Following this, the Fig. 4.7 presents three saturation maps. The first shows the feature:value co-distribution matrix for all relations. The second shows the matrix just for the difficult relations in that task. By subtracting the second from the first, we can derive the difference between all relations' feature:value distribution and just the difficult relation's distributions. That is, we can identify feature:value pairings that are easier or harder to classify. The harder examples are in red, the easier in blue. Where the distribution varies little between all links and just difficult links, the tone tends to white (unsaturated). Thus, a red cell (for example, where an event of `class.I_STATE` is related to a different event which has `aspect.PERFECTIVE`) represents a frequently difficult combination. Conversely, a dark blue cell (e.g. when an adjective is linked with a present-tense event) shows an easy combination; that is, a pairing which, though frequent, is rarely found in the difficult set. The graphs should not exhibit symmetry, because each row represents a different prior assertion, and is the distribution of other features given that assertion, whereas columns do not represent priors.

This information for Task E, linking main events in successive sentences, is in Fig. 4.7, and for Task F, that of linking events where on linguistically subordinates the other, is presented in Fig. 4.8.

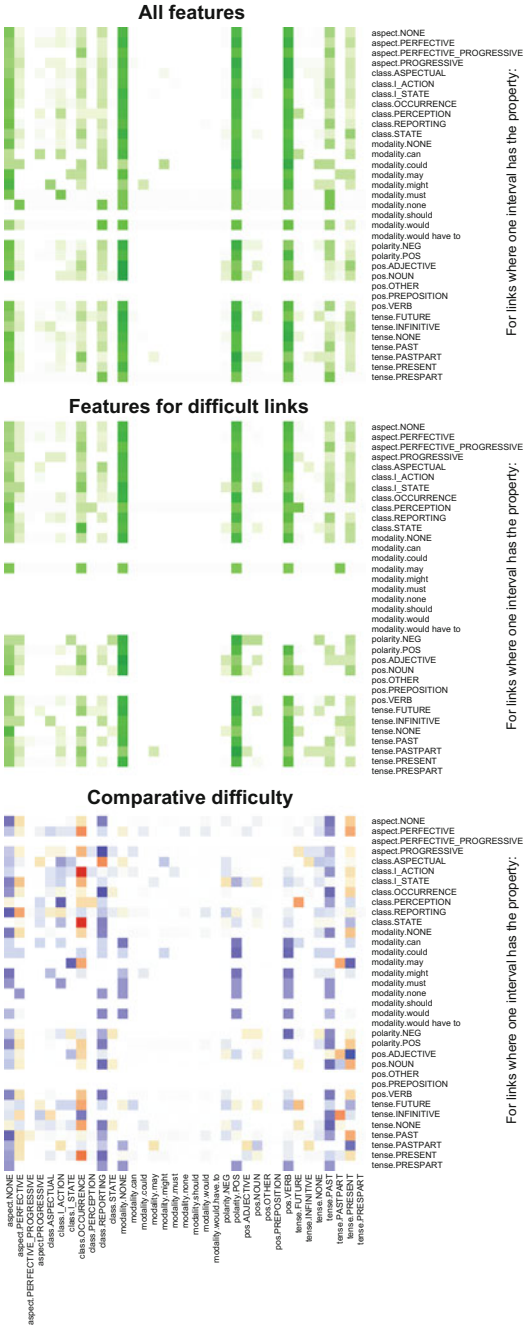

Fig. 4.7 Comparative analysis of features for TempEval-2 task E

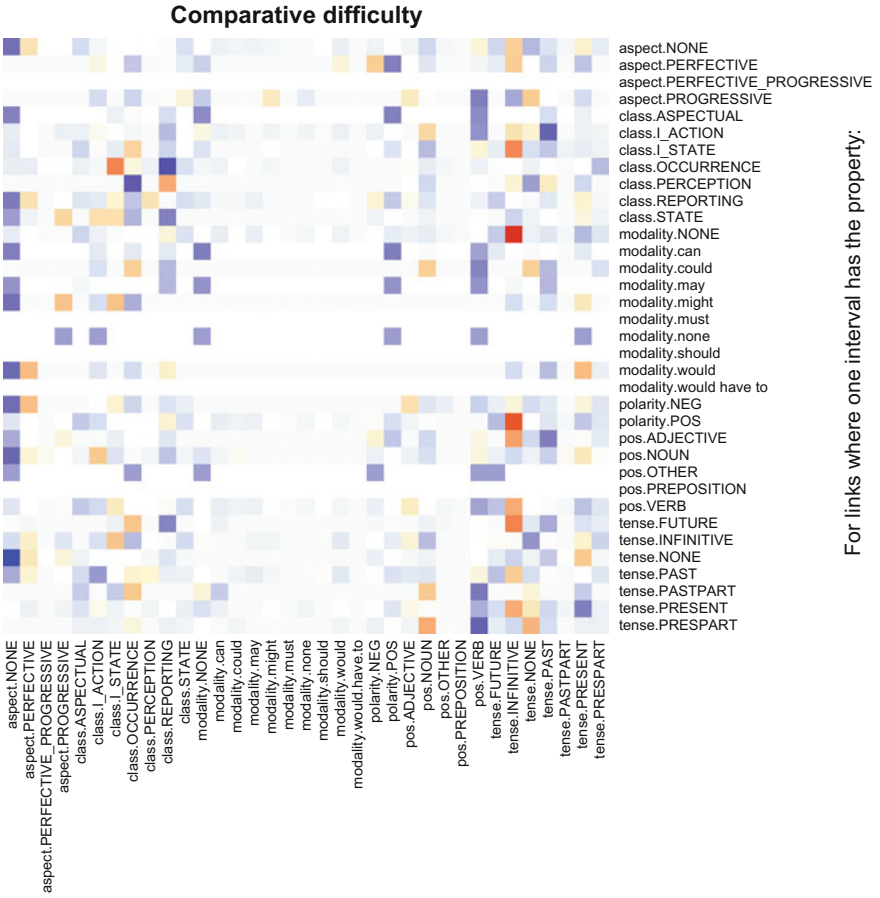

Fig. 4.8 Comparative analysis of features for task F, relating events with their subordinate events

For Task E, from the vertical red stripe in the differential diagram, it can be seen that links to *occurrence*-class events were particularly difficult to label, especially when the other event is of class state or intentional action. However, links to *reporting*-class events were generally easier than average. This could perhaps be due to better consistency in annotations leading to better supervised models, or that a reporting event is typically after the events that are reported but before DCT, giving inherent constraints to this event class. Aside from links with reporting events, particularly easy were links between perceptions and intensional actions (perhaps with perceptions encouraging a reaction?) and links between adjectives and present-tense verbs (perhaps because these always overlap – e.g. *"He says it's hot out there."*).

As for Task F (Fig. 4.8), links with verbs that have no aspect seem to be consistently easier than most. There is less variation in difficulty between certain feature pairings when compared to Task E, as evidenced by the comparatively less saturated graph.

Links to infinitive or un-tensed arguments (e.g. non-verbs) seem to present more difficult than other parts of speech. Of note for being difficult are cases where there is no modality specified in one event and the other is infinitive, possibly due to a reduced number of amodal training examples in a set dedicated to subordination; with links between an occurrence and a state; and with links between future-tense verbs and infinitive verbs.

4.2.3.2 Difficult Event-Timex Link Attribute Distribution

The corresponding data for Tasks C and D are shown in Figs. 4.9 and 4.10 respectively. The colour scheme for event data in green and timex data in blue is continued here, with the exception of comparative difficulty graphs, which use a red/blue divergence colour scheme. In these cases, deep reds indicate very difficult combinations and blue blues very easy ones. Note that the data for task D is only for date-type timexes of granularity less than a month, because in all cases the timex refers to a specific date – DCT – in the data.

For Task C, times, dates and duration appear to be difficult with different sets of event features. Dates and times are difficult to relate correctly to nouns, whereas durations are heard to link to occurrences and present tense verbs. Interestingly, year-sized timexes are very difficult to correctly link to progressive verbs, but very easy to relate to events with no aspect information.

In Task D, we do not have much information. This may be due to a small number of timexes being present in this task's difficult set; the task turned out to be relatively easy. Of these, they are easier to relate correctly to past tensed verbs, and harder to link to occurrence-type events.

4.2.4 Attribute Distribution Summary

It was consistently found that temporal relations between two events are harder to classify than relations between an event and a time. This should direct future research efforts, and was the focus of the latter part of the section, which related a more detailed investigation into the properties of the intervals coupled in difficult links.

Regarding patterns in attribute values over difficult links, although some specific situations of high frequency of difficult links are identified, no clear overall picture emerges. A few specific cases were identified as consistently difficult or easy, but these generally comprised a small proportion of all links. For example, perfect aspect events were had to relate to timexes lasting a year or more; occurrence-class events were difficult to relate with other events, and reporting-class events were easier to relate with other events; and adjective events were easy to relate to present-tense events.

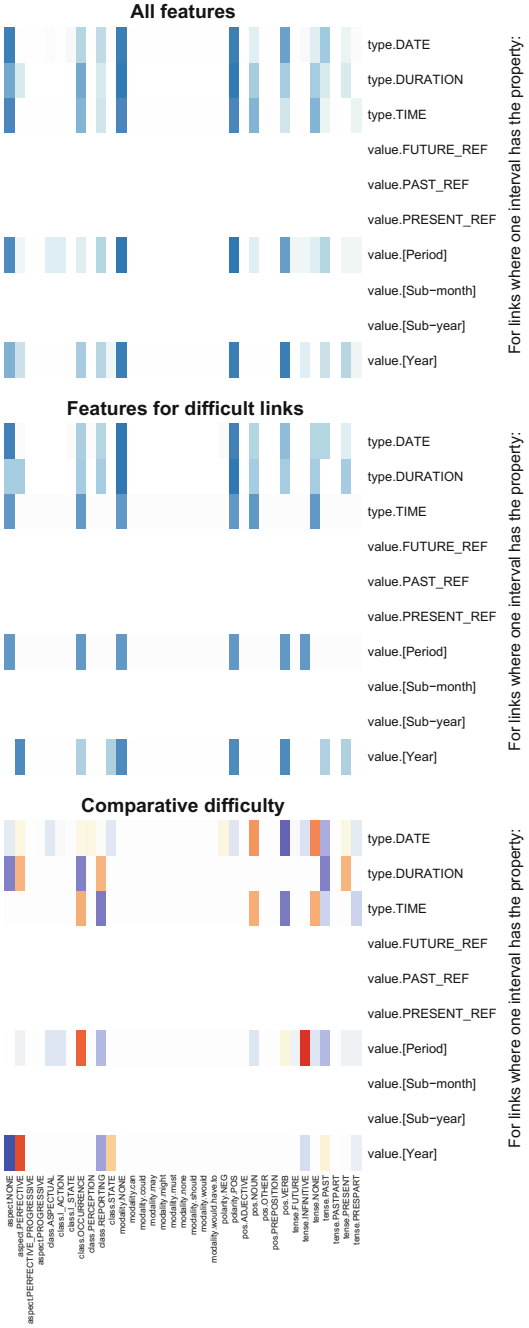

Fig. 4.9 Comparative analysis of features for TempEval-2 task C

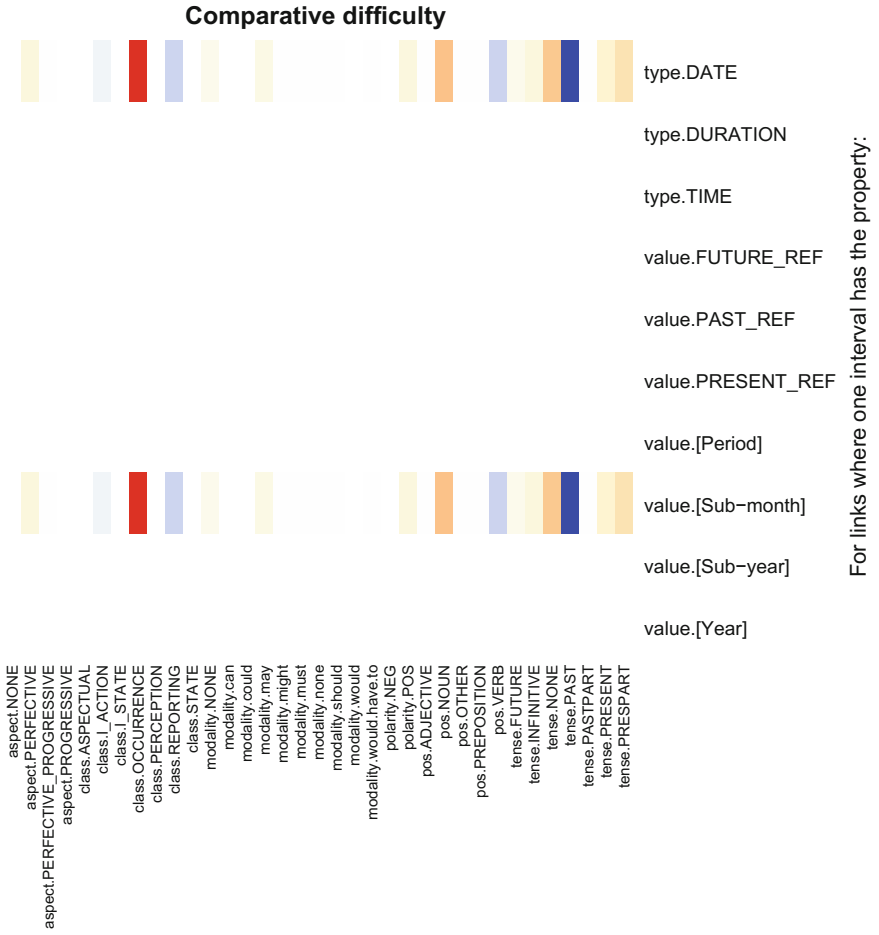

Fig. 4.10 Comparative analysis of features for task D, relating events to DCT

We lead in the next section to a more qualitative approach, taking phenomena contained elsewhere in annotations or not in annotations at all and examining their prevalence in difficult links.

4.3 Extra-Feature Analysis

The overall goal is to determine linguistic sources of temporal ordering information. Because the annotated features do not appear to contain enough information to automatically label links (Sect. 4.2, Chap. 3), other sources of information must be considered. Formal analysis of the surface data does not present immediate clues.

This section presents the results of a survey of each link in the TempEval-2 "difficult set" in terms of the type(s) of information required to determine the temporal relation, aside from that given in TimeML annotations. The resulting information is then used in the next section to attempt to characterise information that temporal links may draw upon, based on prior knowledge about linguistic representations of time.

This analysis was conducted independently of available models and tools, focusing instead on linguistic phenomena. This is to reduce bias from existing methods for and knowledge of the problem. To this end, no TimeML annotation features, tense models or linguistic processing tools were used to construct criteria for characterisation.

4.3.1 Characterisation

It is useful to analyse the difficult TLINKs in a manner that allows identification of common traits. While one can qualitatively express what information is used express a temporal ordering in discourse, to feed into a computational approach one requires quantifiable or at least discrete measures that can be taken consistently from all links. To this end, a set of readily-identifiable linguistic phenomena were determined that could provide temporal information beyond those expressable in TimeML. Each difficult TLINK is then examined and a record made of whether or not each of these phenomena is in place. The result is a survey of types of information used to support temporal orderings for the set of TempEval-2 difficult TLINKs.

The set of phenomena is listed below. Each link may use any number of phenomena. The set is broken into two types: information about the relation and the ordering and information about the interaction between arguments in text.

Relation Information

- **Signalled** - the relation intervals is explicitly expressed by a co-ordinating temporal conjuction or phrase (such as *before*).
- **Inference** - the relation can be easily inferred by reasoning involving other relations in the document
- **From world knowledge** - external information about the general structure of complex events can help determine this relation
- **Iconicity** - temporal order of relation arguments matches the order of their appearance in the source text
- **Disagree** - the annotated relation type is in dispute

Arguments in Text

- **Same sentence** - the relation's arguments are in the same sentence
- **Same clause** - the relation's arguments are in the same clause
- **Tense shift** - there is a shift of tense from one argument to the other
- **Differing modalities** - the arguments do not have the same modality or are not in the same conditional world

- **Differing progression** - one argument is progressive or signifies a culmination or has another aspectual difference from the other
- **Causal** - one argument causes the other and this is critical to the ordering

A "world knowledge" category is therefore included in the above list, in an attempt to roughly estimate how often extra-discourse information is required to resolve difficult links. Also, a "not clear" category is present, for cases where one disagrees with the gold standard.

4.3.2 Analysis

The proportion of difficult links that use each of these phenomena as part of their temporal ordering information is shown in Table 4.6.

Overall, 11.2 % of all TLINKs in TimeBank are annotated as using an explicit temporal signal. It seems that a greater-than-average proportion of difficult intra-sentence event-time links rely on signals (task C), but that difficult subordinated relations (task F) use them less often than is typical.

World knowledge rarely supported difficult links. The task that it helped in most was linking main events in adjacent sentences.

Iconicity – that is, when temporal order follows discourse mention order – was generally not observed within the difficult links set. No task had more than 40 % of its difficult links in the same textual and temporal order. The prevalence of iconicity

Table 4.6 Temporal ordering phenomena and their occurrence in difficult links

	Task			
Description	C	D	E	F
Total instances	21	38	62	43
Signalled	33.33 %	13.16 %	11.29 %	6.98 %
Inference	61.90 %	42.11 %	30.65 %	9.30 %
World knowledge	9.52 %	2.63 %	14.52 %	9.30 %
Iconicity	19.05 %	0.00 %	37.10 %	34.88 %
Unclear/Disagree	14.29 %	18.42 %	4.84 %	4.65 %
Same sentence	100.00 %	0.00 %	0.00 %	97.67 %
Same clause	19.05 %	0.00 %	0.00 %	30.23 %
Tense shift	0.00 %	0.00 %	37.10 %	34.88 %
Differing modalities	47.62 %	34.21 %	8.06 %	51.16 %
Differing progression	0.00 %	0.00 %	16.13 %	11.63 %
Causal	0.00 %	0.00 %	9.68 %	4.65 %

was higher in difficult event-event links than event-timex. This may be because it is somewhat redundant in the case of DATE and TIME timexes, because the timex provides an explicit temporal reference point, and one has less need to rely on implicit factors in order to situate link arguments. Nevertheless, it is interesting to observe that times earlier than events tended to be mentioned in text *after* the events, for the difficult link set. It may also be the case that general discourse follows the principle of iconicity [9] and that having made this observation, automatic temporal relation systems run into difficulties when the principle does not apply.

For event-event links (tasks E and F), a notable proportion of difficult links employ a tense shift. This is where the tense dominating one event is different from that dominating the other. Of the difficult set, this phenomenon occurs 37.1 % of the time in adjacent sentence main event links and 34.9 % of the time in links where one event subordinates another. This suggests that further investigation may be fruitful. There is comparatively very little change of tense in the event-time linking tasks; none in same-sentence event-timex linking and only 5.3 % for event-DCT links.

Differing modalities are very common in in task F's difficult set, as expected for cases where some events subordinate others (this is the category that *if-event-then-event* constructions typically go in), but not common at all for task E.

It is interesting to note the relative lack of shifts in dominant tense in difficult timex-event links when compared to difficult event-event links. This reflects the findings of [10], that temporal adverbs bolster the cognitive role of verb tenses. From these observations, one could suggest that when times are known, a qualifying temporal adverb can be used in place of the information provided by a shift of tense. Validation of this hypothesis remains for future work.

Poor annotation is a potential difficulty source. TempEval-2 data is based on TimeBank, which has an IAA of only 0.77 for TLINK relTypes. The TempEval-2 relation set is simpler than TimeBank's, so 0.77 is a minimum IAA. Investigation of the difficult set showed that the frequency of annotation disagreement was in line with what one might expect. The rate of disagreement with the relation type annotation among links in the difficult set was between 4.6 and 18.5 %. This disagreement rate was consistently higher for event-time links than event-event links, but never higher than average IAA accounts for (23 %), so the difficult links are probably not hard due solely to poor annotation.

4.3.3 Signals Versus Tense Shifts

Signals and tense shift are prevalent in the difficult set. It may be useful to investigate both these phenomena. To avoid redundant investigation, one must first establish some degree of independence between the two; if e.g. solving the relation labelling problem for links with tense shifts also solves the problem for those with signals, then it is not worth investigating both.

It has been proposed that both tense shifts and temporal adverbs provide temporal ordering cues [10]. Further, it is suggested that lexicalised temporal markers and

tense shifts provide information independently – that is to say, there is no overlap in the information provided by either one of these. Temporal information conveyed by tense shift is independent of that provided in signals. We investigate this using empirical data and briefly test the hypothesis that they are exclusive with regard to the temporal information they provide.

Exploring further the idea of explicit temporal qualification (such as with a temporal adverbial) as an alternative to tense shifts, a brief investigation into the overlap between temporal signals and tense shifts is worthwhile. The data has been gathered and, while not excessive, 105 records (total difficult links from tasks E and F) is enough to estimate the degree of overlap. Results are shown in Table 4.7.

In the case of the difficult event-event links, there was no overlap between links where tense shifted between arguments and links that used an explicit temporal signal. The two categories were in fact mutually exclusive. This was a significant deviation from the overlap that would occur if the two phenomena were mutually exclusive (which would be ~6.3 TLINKs).

Looking at all event-event links in TimeBank 1.2 (difficult and non-difficult), the data is different from TempEval. The overlap between signalled and tense-shifted links is as if these phenomena are almost independent (Table 4.8). This can be demonstrated as follows. The global probability of an event-event link using a signal, $P(S)$, is 7.76 %. Similarly, that of such a link using a tense shift $P(T)$ is 40.6 %. If these variables are independent, $P(S \cap T) = P(S) \cdot P(T)$. We know that in the general case, $P(S \cap T) = 3.30\,\%$; further, $P(S) \cdot P(T) = 3.15\,\%$. This is close to suggesting independence.

Another test is to look for prior probabilities with Bayes' theorem. If independent of T, S with not affect $P(T)$ and vice versa. From the data, $P(T|S) = 42.6\,\%$ which is only 4.9 % out from $P(T)$ and $P(S|T) = 8.11\,\%$ is even closer to $P(S)$ with a 4.5 % difference.

However, for the difficult links, despite $P(S)$ and $P(T)$ having roughly similar values, $P(S \cap T) = 0$, which is significantly different from what one would expect, even after taking into account the size of the dataset. Therefore, we might say that

Table 4.7 Co-occurrence frequencies for temporal signals and tense shifts in event-event difficult links

		Tense shift		
		No	Yes	Total
Signal	No	57	38	95
	Yes	10	0	10
	Total	67	38	105

Table 4.8 Co-occurence frequencies for temporal signals and tense shifts in all TimeBank v1.2's event-event links

		Tense shift		
		No	Yes	Total
Signal	No	1908	1303	3211
	Yes	155	115	270
	Total	2063	1418	3481

having both a tense shift and a signal present makes a link relatively easy to automatically label. Certainly in cases where neither a tense shift not a signal appear, the relation is likely to be difficult to classify.

4.3.4 Extra-Feature Analysis Summary

Certain properties were observed in large proportions of difficult links. Difficult event-time relations (tasks C and D) often employed a temporal signal, relied on global inference, or had differing modalities. Difficult event-event relations (tasks E and F) often relied on inference, exhibited iconicity, involved a tense or aspect shift, or had differing modalities. A large proportion of relations have explicit signal or tense/aspect annotations. As this data is directly available and affects a notable proportion of observed TLINKs, these two phenomena were selected for future investigation.

4.3.5 Next Directions

This section provided a data-driven analysis of difficult TLINKs in a well-known dataset using non-surface criteria. A set of commonly-difficult links was identified for each task. Further, a set of potential temporal information sources was identified in terms of linguistic phenomena and these phenomena monitored for each difficult link. This leads to a set of candidate information types for further investigation. What remains to be done is to outline a framework for working with temporal links using these types of temporal phenomena, so that we have experimental and evaluation methods to use in investigation.

4.4 Analysing TLINKs Through Dataset Segmentation

Our approach is to first identify the type of information used to link two entities and then to classify a relation. This section describes the core approach and then enumerates the various special situations of links to be explored in later experimental chapters.

We are not concerned with determining which entities should be temporally linked in a discourse. We constrain our problem, as in the majority of previous work, to providing the relation type of a given entity pair.

4.4.1 Core Approach

The temporal relation labelling experiments in this book adopt a machine-learning approach, based on that of [11]. Experiments are split into "situations", each of which applies to a subset of temporal links. The identification of links in a particular situation is automatic and a method given for each. Additional features are then added to the core set and a classifier learned and evaluated on the links in a situation. Performance is compared with a classifier learned over the same data but without the additional features.

The base set of features is derived directly from the TimeML attribute values, and is as follows:

- event/timex text;
- TimeML tense for each event;
- TimeML aspect for each event;
- modality for each event;
- cardinality for each event;
- polarity for each event;
- part-of-speech for each event;
- class for each event;
- document function for each timex;
- quantisation for each timex;
- frequency for each timex;
- timex value for each timex;
- temporal function for each timex;
- "mod" for each timex;
- type for each timex;
- are both relation arguments in the same sentence?;
- are both relation arguments in adjacent sentences?;
- if events, do both relation arguments have the same TimeML aspect?;
- if events, do both relation arguments have the same TimeML tense?;
- does argument 1 textually precede argument 2?

4.4.2 Theoretical Assumptions

This analysis expects that expressions conveying temporal relation type are present in discourse. Also, even though each relation may be expressed in many way, we assume that it is not. If every available device above is always used to indicate a temporal relation, the analysis' results would be meaningless, as it would show that all types of information are used for all links.

Instead, the approach outlined above makes the assumption that only the minimum amount of language is used to express temporal information. That is, that information theory [12] concepts such as the minimum description length (MDL) [13] will apply

to languages also (as also posited by e.g. [14]). In this context, the MDL principle suggests that unexpected deviations from how time is described require the addition syntactic or lexical information, given a standard "temporal model" of discourse.

Examples of the principle being present in time-relation language are not difficult to come by. One may observe it in phenomena such as temporal signals, tense shifts or temporal expressions. Temporal signals are connectives that explicitly describe a certain ordering but are not required for the majority of relations (they only signal about 12 % of TimeBank's links, for example). Tense shifts require a different term of expression, which may come from the insertion of auxiliary verbs or a change of inflection, and yield a new reference time, event time or even temporal relation. Each shift carries information. Finally, the length and complexity of a temporal expression can correlate to its precision or its distance from the current timeframe; "*At 8.56 am on the 19th August, 2006*" is long, complex and highly specific – "*last week*" serves only to shift the timeframe for anchoring day names backwards. Changing the nominal structure of a sentence is required to express temporal phenomena again. It is this extra information, describing temporal relations, that we are attempting to identify and exploit.

4.5 Chapter Summary

This chapter used a set of empirical data to determine what constitutes a difficult temporal link, and an investigation into linguistic phenomena that occur frequently in the relations that are hardest to automatically label. For each category of relation in TempEval-2 (i.e. Tasks C–F), between 8 and 47 % of temporal relations in documents were difficult for the majority of automatic systems. Event-event relations were consistently the most difficult to type: where 44–47 % of event-event links were difficult, in contrast to event-time links, for which only 8–19 % were difficult.

After an analysis of temporal relations that are difficult to label automatically, themes common in these difficult temporal relations were identified. It was found that two linguistic phenomena were particularly more prevalent in difficult relations than in the general case. First, difficult links often incorporated an explicit co-ordinating temporal signal (a word like *simultaneously* or *thereafter*). Second, shifts of tense and aspect between arguments were often present in difficult links. Other contributing factors were implicit temporal relations discoverable through inference, and changes in modality, though these were less prevalent.

Based on this analysis, the remainder of this book comprises two major parts: an investigation into temporal signals, and another into a framework of tense and aspect. Signals have been found to be useful. We demonstrate how they may be used for temporal relation labelling and then investigate the automatic annotation of temporal signals in Chap. 5. Models of tense can account for a whole group of situations, including reported speech, tense shifts and the use of timexes to shift the frame of reference. Such situations are detailed in Chap. 6.

References

1. Boguraev, B., Pustejovsky, J., Ando, R., Verhagen, M.: TimeBank evolution as a community resource for TimeML parsing. Lang. Resour. Eval. **41**(1), 91–115 (2007)
2. Tissot, H., Roberts, A., Derczynski, L., Gorrell, G., Del Fabro, M.D.: Analysis of temporal expressions annotated in clinical notes. In: Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11), p. 93. Association for Computational Linguistics (2015)
3. Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J.: SemEval-2010 task 13: TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 57–62. Association for Computational Linguistics (2010)
4. Ha, E., Baikadi, A., Licata, C., Lester, J.: NCSU: modeling temporal relations with markov logic and lexical ontology. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pp. 341–344. Association for Computational Linguistics (2010)
5. Derczynski, L., Gaizauskas, R.: USFD2: annotating temporal expressions and TLINKs for TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 337–340. Association for Computational Linguistics (2010)
6. UzZaman, N., Allen, J.: TRIPS and TRIOS system for TempEval-2: extracting temporal information from text. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 276–283. Association for Computational Linguistics (2010)
7. Llorens, H., Saquete, E., Navarro, B.: TIPSem (English and Spanish): evaluating CRFs and semantic roles in TempEval-2. In: Proceedings of SemEval-2010, pp. 284–291. ACL (2010)
8. Lee, C., Katz, G.: Error analysis of the TempEval temporal relation identification task. In: SEW-2009 Semantic Evaluations: Recent Achievements and Future Directions, pp. 138–145 (2009)
9. Diessel, H.: Iconicity of sequence: a corpus-based analysis of the positioning of temporal adverbial clauses in English. cognit. linguist. **19**(3), 465–490 (2008)
10. Harris, R., Brewer, W.: Deixis in memory for verb tense. J. Verbal Learn. Verbal Behav. **12**(5), 590–597 (1973)
11. Mani, I., Wellner, B., Verhagen, M., Pustejovsky, J.: Three approaches to learning TLINKS in TimeML. Technical Report. CS-07-268, Brandeis University, Waltham, MA, USA (2007)
12. Shannon, C.: Communication theory of secrecy systems. Bell Syst. Tech. J. **28**(4), 656–715 (1949)
13. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5), 465–471 (1978)
14. Grünwald, P.: A minimum description length approach to grammar inference. Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, pp. 203–216 (1996)